UNIVERZITET U BEOGRADU
E L E K T R O T E H N I Č K I F A K U L T E T

Draško V. Furundžić

OCENA KVALTETA ARTIKULACIJE GLASOVA SRPSKOG JEZIKA PRIMENOM NEURONSKIH MREŽA

doktorska disertacija

Beograd 2018

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

Draško V. Furundžić

EVALUATION THE QUALITY OF ARTICULATION PHONEMES OF SERBIAN LANGUAGE USING NEURAL NETWORKS

Doctoral Dissertation

BELGRADE, 2018.

MENTOR:

Prof. Dr Srđan Stanković, Profesor Emeritus, Univerzitet u Beogradu, Elektrortehnički Fakultet

ČLANOVI KOMISIJE:

Prof. Dr Srđan Stanković, Profesor Emeritus, Univerzitet u Beogradu, Elektrortehnički Fakultet

Dr Željko Đurović, redovni profesor, Univerzitet u Beogradu Elektrotehnički fakultet

Dr Zoran Šarić, naučni savetnik, Centar za Unapređenje Životnih Aktivnosti, Beograd

Prof. Dr Milo Tomašević, redovni profesor, Univerzitet u Beogradu Elektrotehnički fakultet

Dr Miško Subotić, naučni saradnik, Institut za eksperimentalnu fonetiku i patologiju govora, Beograd

ZAHVALNOST

Želim da izrazim zahvalnost svojim kolegama i saradnicima, koji su na razne načine doprineli izradi ove disertacije. Najveću zahvalnost dugujem profesoru Srđanu Stankoviću, svom dugogodišnjem profesoru i mentoru za strpljenje i angažovanost. Zahvalnost dugujem i nastavnom osoblju katedre za signale i sisteme Elektrotrhničkog Fakulteta Beograd, kao i osoblju Instituta za Eksperimentalnu Fonetiku i patologiju Govora. Takođe želim da se zahvalim svim bliskim i dragim ljudima čije postojanje za mene predstavlja veliki motiv svakodnevnog angažovanja.

Naslov: OCENA KVALTETA ARTIKULACIJE GLASOVA SRPSKOG JEZIKA PRIMENOM NEURONSKIH MREŽA

REZIME

Glavni zadatak istraživanja prikazanog u disertaciji je modeliranje složenog procesa logopedske procene kvaliteta artikulacije fonema srpskog jezika, zasnovane na inteligentnim „*data driven*“ učećim modelima. Multidisciplinarna priroda i složenost zadatka determinisala je sledeći niz metodoloških koraka za njegovo izvršenje: a) deskriptivna analiza procesa artikulacije, kao najvažnijeg i najsloženijeg aspekta u psiho-fiziološkom procesu produkcje govornog izraza, u svetlu definicije i komparativne analize njegovih tipičnih i atipičnih realizacija; b) karakterizacija akustičke manifestacije procesa artikulacije govornog izraza, kao pojave pogodne za posrednu analizu kvaliteta artikulacije preko instrumentalnih eksperimentalnih metoda; c) deskripcija procesa auditivne percepcije i evaluacije kvaliteta artikulacije od strane eksperta (logopeda) zasnovane na treniranom slušanju, odnosno, na sinhronizovanoj analizi skupa relevantnih artikulacionih i akustičkih atributa govornog signala, u skladu sa standardnim testovima; d) deskripcija procesa formiranja logopedskog ukupnog akustičkog utiska, odnosno akustičke slike artikulacije govormog izraza i njegove sublimacije u odgovarajuću vrednost na standardnoj numeričkoj skali ocena; e) uspostavljanja različitih modela algoritamske korespondencije između vektora akustičkih obeležja i numeričkih indikatora klasa različitog kvaliteta artikulacije što rezultuje računarskim modeliranjem procesa logopedske ocene kvaliteta artikulacije zasnovanim na inteligentnim učećim prediktorima, gde se računar nalazi u sličnom informacionom okruženju kao i logoped.

Ključne reči: Patologija govora, veštačke neuronske mreže, računarsko modeliranje, Bayesov klasifikator, kvalitet artikulacije

Naučna oblast: elektrotehnika

Uža naučna oblast: Signali i sistemi

UDK broj: 621.3

Title: EVALUATION THE QUALITY OF ARTICULATION PHONEMES OF SERBIAN LANGUAGE USING NEURAL NETWORKS

SUMMARY

The main task of the research presented in the dissertation is the modeling of the complex process of logopedic assessment of the quality of the articulation of Serbian phonemes, based on intelligent "data driven" learning models. The multidisciplinary nature and complexity of the task determined the following set of methodological steps for its execution: a) a descriptive analysis of the articulation process, as the most important and complicated aspect of the psycho-physiological process, the production of speech expression, in the light of the definition and comparative analysis of its typical and atypical realizations; b) the characterization of an acoustic manifestation of the process of articulation of speech expression, as phenomena suitable for the indirect analysis of articulation through instrumental experimental acoustic methods; c) Describing the process of audible perception and evaluating the quality of articulation by the expert (logoped) based on a trained listening, or, in a synchronized analysis of a set of relevant articulation and acoustic attributes of the speech signal, in accordance with standard tests; d) the description of the process of forming the logopedic total acoustic impression, that is, the acoustic image of the articulation of the speech term and its sublimation in the corresponding value on the standard numerical scaling scale; e) the establishment of different models of algorithmic correspondence between the acoustic feature vectors and the numerical indicators of the classes of different articulation quality, which results in computer modeling of the process of logopedic articulation quality assessment based on intelligent learning predictors, where the computer is located in a similar information environment as well as speech therapist.

Keywords: Pathology of speech, artificial neural network, computer modeling, Bayesian classifier, quality of articulation
Scientific area: electrical engineering
Scientific subarea: signals and systems
UDC number: 621.3

SADRŽAJ

SPISAK SLIKA:

SPISAK TABELA:

SPISAK SKRAĆENICA

Skraćenica:	Puni naziv:
ANN	Artificial Neural Netvork
AT	Analitički Test
GAT	Globalni Artikulacioni Test
KNN	K Nearest Neighbour
LPC	Linear Predictive Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multileyer Perceptron
NB	Naive Bayes

NM	Neuronske Mreže
SMOTE	Synthetic Majority Oversampling TEchnique
SOM	Self Organizing Map
VAD	Voice Activity Detection
VUS	Voiced/Unvoiced/Silence

1. UVOD

Aktuelna epidemiološka situacija u domenu govorne komunikacije ukazuje na širenje različitih formi govorne patologije koja se najčešće manifestuje padom kvaliteta artikulacije, što se ne retko dovodi u vezu sa naglim tehnološkim razvojem u oblasti telekomunikacija, interneta i mobilne telefonije. Pored nespornog značaja za razvoj društva, moderne tehnologije generišu niz negativnih uticaja na opšte zdravstveno stanje mlađe populacije a posebno na razvoj govora. Naime, favorizacija neverbalne i indirektne komunikacije na štetu direktne govorne komunikacije bitno utiče na pad kvaliteta artikulacije glasova i kvaliteta govora generalno. Egzistencijalni značaj govora i jezika implicira da pravilno razvijen govor predstavlja neophodan uslov zdravog razvoja i kvalita svih aspekata života jedinke. Shodno tome, u logopediji su već uspostavljene standardne metode ocene kvaliteta artikulacije, naročito tokom razvoja govora, zasnovane na komparativnoj auditivno-perceptivnoj analizi artikulaciono-akustičkih karakteristika govora u tipičnoj i atipičnoj realizaciji.

Kvantitativna i kvalitativna računarska evaluacija artikulacije glasova, aktuelna još od sredine prošlog veka, zbog prirode govornog signala je predmet multidisciplinarnog pristupa nekoliko manje ili više srodnih naučnih diciplina. Takav pristup uslovljen je složenošću odnosa između neuro-fiziološkog procesa artikulacije i govornog signala kao njegove direktne akustičke manifestacije, sa jedne strane, i složenošću auditivno-perceptivnog procesa obrade produkuvanog artikulaciono-akustičkog sadržaja koji je takođe neuro-fiziološke pirode, sa druge strane. Zbog velikog broja i jako izražene variabilnosti nosilaca kvaliteta govora i složenosti procesa artikulacije, do sada nije pronađeno jedinstveno rešenje sistema za njegovu pouzdanu i objektivnu računarsku procenu, zasnovano na utvrđenom skupu pouzdanih artikulaciono-akustičkih indikatora. Ova činjenica ukazuju na potrebu definisanja optimalnog skupa relevantnih obeležja kao preduslova za dizajniranje odgovarajućeg računarskog modela. Ovako rešenje bi doprinelo objektivizaciji i povećanju pouzdanosti u proceni, karakterizaciji i kategorizaciji kvaliteta govora, što bi značajno unapredilo postojeće logopedske procedure u radu sa decom i ostalim pacijentima.

Poslednje decenije prošlog veka karakteriše nagli razvoj veštačke inteligencije, interneta i informacionih tehnologija generalno je omogućio razvoj efikasnih alata i formiranje velikog broja lako dostupnih baza podataka, između ostalog, i u domenu prepoznavanja i evaluacije kvaliteta govora. Postojeći uslovi daju mogućnost za dizajniranje fleksibilnih računarskih modela i softverskih alata za objektivnu ocenu kvaliteta govora sa širokom primenom u logopediji koji se mogu koristiti i putem interneta. Takvi alati bi znatno smanjili subjektivni

uticaj logopeda i omogućili simultano testiranje većih grupa pacijenata koje se može izvoditi bez fizičkog prisustva logopeda, što bi povećalo dostupnost i efikasnost s jedne strane, i smanjilo troškove procedure sa druge strane. Takvi alati bi se takođe mogli koristiti prilikom edukacije i treninga logopeda povećavajući prostor za njihovo kreativnije i delotvornije angažovanje. Potencijalni pacijenti bi ih mogli koristiti za samostalni trening i rehabilitaciju ili detekciju ozbiljnijih poremećaja koji zahtevaju stručni tretman. Takvi sistemi bi se mogli koristiti u cilju automatizovane komparativne ocene efikasnosti primene različitih terapeutskih procedura. Novi zaključci o proceni kvaliteta artikulacije, koja u osnovi predstavlja finu diskriminaciju između atributa unapred definisanog i ograničenog skupa srodnih i sličnih fonemskih i subfonemslih struktura, mogu imati značaj za unapređenje metodologije prepoznavanja govora koji se fokusira na diskriminaciju instanci velikog skupa govornih segmenata sa izraženim diverzitetom atributa.

Učeći modeli, a posebno Neuronske Mreže, kao moderni alati iz domena veštačke inteligencije, imaju strukturu koja u datom informacionom okruženju omogućava računarsku identifikaciju funkcionalne zavisnosti kvaliteta artikulacije od pojedinih artikulaciono-akustičkih obeležja, što može poslužiti za rangiranje stepena relevantnosti pri izboru potencijalnih obeležja i interni uvid u netransparentne kauzalne relacije složenog procesa artikulacije i njegove akustičke manifestacije. Performanse učećih prediktora, poznataih kao „*data driven*“ modeli, po definiciji su determinisane kako kvalitetom (reprezentativnost) i kvantitetom (broj instanci) raspoloživih baza podataka, tako i stepenom relevantnosti, odnosno, informativnosti reprezentativnih obeležja. Obe kategorije, reprezentativnost i informativnost, su predmet opsežnih istraživanja i primene velikog broja efikasnih metoda i manje ili više poznatih alata. Reprezentativnošću obučavajućeg uzorka se bavi novija informaciona tehnologije poznate kao Imbalanced Learning (učenje u neizbalansiranim uslovima), koja obuhvata veliki broj metoda i algoritama za izbor skupa instanci obučavajućeg uzorka, bilo da se radi o realnom, sintetički generisanom ili kombinovanom skupu instanci. Informativnost odabranih obeležja je predmet istraživanja širokog spektra dostupnih alata i metoda poznatih pod imenima extrakcija obeležja, redukcija dimenzija, data mining. Zbog značajnog uticaja pomenutih kategorija na performanse učećih prediktora neophodno je primeniti neku od navedenih regulatornih metoda u fazi pripreme obučavajućeg uzorka.

Koristeći prednosti metoda za izbor relevantnog vektora odabranih artikulaciono-akustičkih obeležja govornog signala, prednosti alata za smanjenje redundanse, povećanje reprezentativnoti i balansa raspoloživog uzorka govornih stimulusa, kao i mogućnost izabora tipa i strukture fleksibilnih i robustnih učećih modela, sa razlogom smo anticipirali uspeh u

dizajniranju objektivnog i efikasnog računarskog modela postupka logopedske procene kvaliteta artikulacije glasova srpskog jezika.

1.1. Predmet i cilj istraživanja

Predmet prikazanog istraživanja se odnosi na potrenbe uslove za definiciju formalnih modela algoritamske korespondencije između vektora akustičkih mera kvaliteta artikulacije glasova i numeričkih indikatora njihovih ocena, dobijenih auditivno-perciptivno od strane tima logopeda; Krajnji cilj istraživanja je definicija pouzdanog i objektivnog računarskog modela stadardnog logopedskog postupka procene kvaliteta artikulacije glasova srpskog jezika prihvatljive tačnosti i visokog stepena automatizacije. Realizacija primarnog cilja implicirala je nekoliko posebnih, sukcesivnih koraka, kao njegovih parcijalnih realizacija:

- definicija opšteg modela za ocenu kvaliteta artikulacije analizom mera odstupanja artikulaciono-akustičkih obeležja izgovornih glasova u tipičnom i atipičnom izgovoru;
- definicija i karakterizacija skupa relevantnih artikulaciono-akustičkih indikatora kvaliteta artikulacije, definicija njihove metrike i njihovih tipičnih odstupanja;
- balansiranje neravnomernosti raspodele zastupljenosti tipičnog i atipičnog izgovora u raspoloživom uzorku instance u cilju povećanja njegove reprezentativnosti;
- komparacija različitih algoritama klasifikacije vektora akustičkih obeležja izgovornih glasova i izbor optimalnog modela klasifikatora za računarsku ocenu kvaliteta artikulacije;
- analiza funkcionalne zavisnosti kvaliteta artikulacije glasova od pojedinih odstupanja;
- Provera ispravnosti osnovnih hipoteza ostvarivosti i opravdanosti predloženog sistema koji je krajnji cilj istraživanja.

1.2. Polazne pretpostavke

Naglašen subjektivistički karakter, nedovoljna efikasnost i pouzdanost kao i drugi nedostaci metoda za procenu kvaliteta artikulacije glasova srpskog jezika, zasnovanih na tradicionalnoj iskustvenoj evaluaciji logopeda, su motiv za simplifikaciju i objektivizaciju ovog procesa. Dosadašnji pristupi problemu se zasnivaju na parcijalnim rešenjima baziranim samo na artikulacionom aspektu kvaliteta artikulacije ili analizi pojedinačnih akustičkih manifestacija njegove variabilnosti. Uočavanjem nedostataka ovih rešenja, došlo se do ideje o povećanju dimenzija i diverziteta komponenti vektora obeležja i primeni modernih alata za njihovu klasifikaciju, u cilju kreiranja pouzdanog, objektivnog i efikasnog računarskog modela sa

visokim stepenom automatizacije koji će unaprediti postojeći način ocene kvaliteta artikulacije. Generalna pretpostavka ostvarivosti, opravdanosti i potrebnih performansi planiranog modela morala je zadovoljiti nekoliko preduslova i proći proveru kroz proces dokazivanja sledećih pretpostavki manjeg stepena opštosti u cilju potvrde njene ostvarivosti i opravdanosti:

P1 Višedimenzionalni prostor artikulaciono-akustičkih atributa izgovornog glasa omogućava pouzdanu distinkciju između njegovih tipičnih i atipičnih realizacija.

P2 Ne postoji značajna razlika u tačnosti ocene kvaliteta artikulacije između logopeda i izabranog algoritma za automatsku ocenu kvaliteta artikulacije.

P3 „Direktne mere balansa raspodele instanci različitih kvaliteta artikulacije u prostoru obeležja obučavajućeg uzorka (Entropija i Standardna devijacija) stoje u jakoj pozitivnoj korelaciji sa indirektnim merama njegove reprezentativnosti (Tačnost predikcije involviranih prediktora)".

P4 Od planiranih modela za klasifikaciju vektora akustičkih obeležja, najbolje performanse se očekuju od ansambla višeslojnih perceptrona sa balansiranim obučavajućim skupom.

1.3. Metodološki aspekt istraživanja

Ostvarivost predloženog objektivnog računarskog modela za ocenu kvaliteta izgovora glasova srpskog jezika uslovljena je potvrdom valjanosti definisanih polaznih pretpostavki a realizacija modela zahtevala je istraživanje i korišćenje raspoloživih relevantnih znanja i alata sa jedne strane i kreiranje neophodnih originalnih rešenja, kao očekivanih doprinosa, sa druge strane.

Od primarnog značaja za uspešno rešenje uočenog problema bilo je razumevanje samog procesa artikulacije glasova kao složenog mehanizma zasnovanog na naučenim paradigmama koordinisane aktivacije organa govornog aparata i čula sluha, koji zavisi kako od iskustva i sposobnosti govornika tako i od uticaja psiho-emotivnih faktora verbalne eksperesije u realnim uslovima.

Sledeća, vrlo značajna premisa rešenja definisanog zadatka, bila je analiza i razumevanje mehanizma naučene percepcije kvaliteta izgovorenog sadražaja, od strane logopeda, pre svega u cilju identifikacije relevantnih artikulaciono-akustičkih manifestacija izgovorenih fonema, odnosno skupa obelažja na osnovu kojih logoped procenjuje kvalitet izgovora. Artikulaciono-akustička obeležja, kao korelati kvaliteta artikulacije govora koji su u velikoj meri perceptabilni, tehnički detektabilni i merljivi, mogu se predstaviti u numeričkoj ili simboličkoj formi, što omogućava računarsko modeliranje procesa evaluacije kvaliteta artikuilacije i

prezentaciju rezultata. Ovaj korak, koji predstavlja teoretsku karakterizaciju i procenu distinktivnih obeležja kvaliteta produkovanih fonema, za nas je od velikog značaja jer determiniše tok daljeg istraživanja, međutim, za razliku od ostalih, ovaj korak je naglašeno subjektivne prirode pošto zavisi od eksperta – logopeda i kao takav je zahtevao posebnu pažnju u kontekstu pojma objektivizacije modela kao jedne od premisa prihvatljivog rešenja.

Jako značajan momenat istraživanja se odnosi na vrlo aktuelni problem informativnosti raspoloživog i obučavajućeg skupa instanci koji direktno utiče na performanse korišćenih učećih prediktora. Dakle, obučavajući skup instanci različitog kvaliteta artikulacije bilo je potrebno podvrgnuti proceduri ciljanog povećanja reprezentativnosti kroz smanjenje redundanse i balansiranje razlike u broju predstavnika različitih klasa. Ova originalna procedura se zasniva na maksimizaciji entropije raspoloživog uzorka putem povećanja ravnomernosti raspodele instanci u prostoru obeležja primenom adekvatnih metoda odabiranja.

Sledeći istraživački korak se odnosio na izbor tipa i strukture predloženog matematičkog modela prediktora koji treba na osnovu raspoložive baze podataka ograničene veličine da uspostavi prihvatljiv algoritamski model korespondencije između vektora obeležja i analiziranog kvaliteta artikulacije fonema. Kroz komparativnu analizu performansi nekoliko predloženih tipova i struktura modela izvršili smo izbor optimalnog računarskog modela. Ovaj korak je imao izuzetan značaj zato što koriguje eventualne ili očekivane nedostatke prethodnih koraka a naročito izbora artikulaciono-akustičkih obeležja.

Ostvarenje postavljenog zadatka podrazumevalo je dakle sledeće važne, planirane, metodološke korake: a) određivanje i karakterizacija artikulaciono-akustičkih indikatora kvaliteta artikulacije, definicija njihove metrike i njihovih tipičnih odstupanja; b) analiza i povećanje nivoa informativnosti i reprezentativnosti izabranog vektora obeležja i obučavajućeg uzorka primenom odgovarajućih metoda; c) uspostavljanje različitih formalnih modela algoritamske korespondencije između vektora akustičkih mera i numeričkih indikatora, klasa različitog kvaliteta artikulacije, apriori ocenjenih auditivno-perciptivno od strane tima logopeda; d) izbor optimalnog modela kroz komparativnu analizu performansi nekoliko predloženih tipova i struktura modela; e) verifikaciju validnosti predikcije ocean kvaliteta artikulacije izabranih modela kroz njihovu komparaciju sa individualnim logopedskim ocenama, gde je komparacija obavljena posredno preko etalona kvaliteta artikulacije određenog većinskim odlučivanjem grupe od pet iskusnih logopeda.

Priprema istraživanja zahtevala je obezbeđivanje uslova za izvođenje eksperimenata kao osnovne za dalji rad. To podrazumeva formiranje govorne baze stimulusa u formi uzorka adekvatne veličine i reprezentativnosti u skladu sa definisanim kriterijumima izbora ispitanika.

Istraživanja su koncipirana i realizovana tako da obezbede prikupljanje statistički reprezentativnih podataka čijom će se analizom omogućiti razumevanje relacije između artikulacionih, akustičkih i auditivnih fenomena pri izgovoru glasova sa jedne strane i relevantne numeričke ocene tog izgovora, sa druge strane. Baza je korišćena za ekstrakciju relevantnih obeležja fonema, proveru relevantnosti pojedinih parametara pri segmentaciji fonemskih i subfonemskih struktura analiziranih reči, trening, validaciju i verifikaciju performansi klasifikatora, kao i komparaciju različitih tipova i arhitektura klasifikatora. Istraživanje uključuje dve grupe ispitanika: kontrolna sa korektnim izgovorom i eksperimentalna sa različitim vrstama i nivoima odstupanja u izgovoru glasova. Pripadnici obe kategorije su podeljene na obučavajući uzorak i test uzorak približnih kardinalnih vrednosti. Pored dve grupe govornika, čiji izgovor će poslužiti za formiranje govorne baze, predviđene je i jedna grupa treniranih slušalaca - eksperata, koja će poslužiti za objektivizaciju logopedskog procesa ocene kvaliteta artikulacije i, samim tim, objektivizaciju računarskog modela procesa, što implicira poboljšanje performansi našeg algoritma.

Definisani ciljevi, hipoteze i sam predmet planiranog istraživanja determinisali su multidisciplinarni pristup koji će iskoristiti informacije, metode i tehnike sledećih oblasti: informacione nauke, veštačke inteligencije, obrade govornog signala, neurofoziologije govora, elektro-akustike, logopedije, eksperimentalne fonetike. U istraživanju su korišćeni adekvatni standardni testovi primenjeni u logopediji i relevantno ekspertsko znanje. Za obradu podataka, modeliranje procesa evaluacije i prikaz rezultata korišćeni su softverski alati Matlab i Praat. Za statističku analiza su korišćeni softverski paketi Microsoft Excel i Statistica. Zbog specifičnosti problema obuhvaćenih tezom ne postoje adekvatne i sveobuhvatne programske procedure ili funkcije namenjene analizi kvaliteta artikulacije, pa su za ovu svrhu kreirane originalne programske funkcije i alati koji su korišćeni u navedenom softverskom okruženju. Metodologija rešenja definisanog problema u okviru predviđenih istraživanja ima nekoliko praktičnih aspekata više ili manje međusobno povezanih. Polazna pretpostavka je ostvarivost prihvatljivog modela ekspertskog pristupa oceni kvaliteta izgovora glasova, zasnovanog na auditivnoj i vizuelnoj percepciji artikulaciono akustičkih manifestacija procesa artikulacije konkretno izgovorenog glasa. U domaćoj logopedskoj praksi postoje dva osnovna pristupa tom problem i zasnovani su na: a) globalnom artikulacionom testu (GAT) kojim se detektuje prisustvo odstupanja (narušavanja opšte akustičke slike) u odnosu na normalnu artikulaciju i vrši njegova kategoritzacija i evaluacija, i b) analitičkom testu (AT) pomoću kog se vrši ocena kvaliteta artikulacije na osnovu sinteze ocena većeg broja pojedinih akustičkih obeležja konkretnog glasa. Grupe pomenutih obeležja su specifične za različite kategorije glasova što je

posledica različitih načina artikulacije. Auditivno-perceptivnom evaluacijom odstupanja posmatranih relevantnih obeležja pri izgovoru glasova, logopedi formieraju globalnu ocenu kvaliteta artikulacije u numeričkoj formi. Izborom grupe merljivih, numerički kodiranih, artikulaciono akustičkih karakteristika za svaki od izgovorenih glasova iz govorne baze, formiramo jedinstveni vektor obeležja kom se pridružuje odgovarajuća ekspertska numerička vrednost ocene. Koristeći raspoloživu bazu zvučnih stimulusa formiramo dva skupa vektora obeležja i njima korespodentnih ekspertskih vrednosti ocena. Prvi skup koristimo za obuku različitih prediktora-klasifikatora, dok drugi skup y neizmewebnoj formi služi za verifikaciju performansi obučenih klasifikatora. Jedan od predloženih računarskih modela za jednostavnu i pouzdanu ocenu kvaliteta artikulacije od kog se očekuju najbolje prefromanse je zasnovan na optimalnom ansamblu višelojnih perceptrona, iz razloga dokazanih prednosti u odnosu na standardne induktivne modele. Generalno, neuronske mreže kao softverske arhitekture za simulirano paralelno procesiranje signala, nastale modeliranjem principa morfološko-funkcionalne organizacije nervnog sistema, imaju ključne i ekskluzivne sledeće karakteristike: robustnost karakterističnu za paralelno distribuirane modela, plastičnost, adaptabilnost i sposobnosti apstrakcije i generalizacije. Navedene favorizujuće osobine su opredeljujući razlog pri izboru vrste modela za modeliranje složenog procesa ocene kvaliteta artikulacije. Planirana je i komparativna analiza rezultata primene različitih tipova klasifikatora u cilju pronalaženja optimalnog rešenja.

Važnniji eksperimentalni koraci ka rešenju su:

- Istraživanje i prikaz postojećih rezultata u domenu kvaliteta artikulacije glasova i.
- Selekcija uzorka ispitanika i govornog korpusa odabranih reči.
- Predobrada govorne baze, odnosno detekcija govorne aktivnosti (VAD) na osnovu odabranih parametara primenom poznatih algoritama.
- Segmentacija govorne baze na fonemske i subfonemske segmente primenom odgovarajućih algoritama.
- Formiranje baze vektora izabranih obeležja iz baze segmenata i prateće baze ekspertskih ocena kvaliteta artikulacije.
- Određenje arhitekture i obuka predloženih tipova učećih prediktora.
- Obrada i analiza dobijenih rezultata računarske evaluacije izgovora i poređenje sa ekspertskim-logopedskim rezultatima.
- Komparacija i analiza performansi različitih računarskih algoritma za ocenu kvaliteta govora.

- Komparacija tačnosti računarskih algoritama sa tačnostima pojedinih logopeda.

1.3.1. Očekivani doprinos

-Određivanje vektora akustičkih obeležja govornog signala glasova u cilju njegove bolje formalne reprezentacije, kvantifikacije i kvalifikacije sa aspekta kvaliteta artikulacije,
-Originalan pristup važnom problemu učenja sa neizbalansiranim skupovima i njegova primena u analizi kvaliteta artikulacije,
-Originalan pristup oblasti izbora i primene optimalnog ansambla neuronskih mreža poboljšanih performansi u obradi govornog signala,
-Doprinos rešavanju problema ocene relevantnosti aktuelnih obeležja kroz analizu osetljivosti neuronskih mreža,
-Izbor optimalnog algoritma kroz komparativnu analizu nekoliko različitih algoritama za računarsku ocenu kvaliteta artikulacije glasova srpskog jezika.
-Doprinos boljem razumevanju kompleksnog problema auditivne percepcije akustičkih manifestacija tipične i atipične artikulacije glasova srpskog jezika.
- Prikaz aktuelnog stanja u bitnim domenima istraživanja ovog problema i pratećih problema.

Dobro definisanje distinktivnih akustičkih karakteristika glasova tipične i atipične realizacije u auditivno-perceptivnom prostoru, doprinosi boljem razumevanju i rešavanju široke klase problema iz oblasti govorne patologije i govora i jezika uopšte.
Ostvarenje planiranih istraživačkih aktivnosti omogućava utvrđivanje jasnih pokazatelja odstupanja u izgovoru glasova i doprinosi pouzdanoj i objektivnoj automatizovanoj proceni kvaliteta artikulacije glasova. Problematika iz domena ovog istraživanja je, iako vrlo značajna, ipak je nedovoljno istražena što dodatno ukazuje na njegov značaj.

1.3.2. Plan istraživanja i struktura rada

Pronalaženje pouzdanog i objektivnog modela za ocenu kvaliteta artikulacije glasova kao ispunjenje glavnog cilja podrazumeva sledeće korake:
-Artikulaciono-akustičku analizu tipičnih odstupanja i njihovu karakterizaciju,
-Identifikaciju i analizu akustičkih mera variranja kvaliteta izgovora,
-Uspostavljanje formalnih algoritamskih metoda za klasifikaciju skupa vektora akustičkih mera odstupanja u klase različitog kvaliteta artikulacije definisane od strane logopeda auditivnom percepcijom,

-Primenu efikasnog novog načina izbora ansambla neuronskih mreža kod problema klasifikacije u cilju povećanjna robustnosti i tačnosti modela,
-Novi pristup obuci neuronskih mreža pri naglašeno neizbalansiranim obučavajućim uzorcima kako među klasama tako i unutar klasa sa ciljem povećanja njihove reprezentativnosti.

1.4. Skraćeni pregled sadržaja disertacije

Aktuelni sadržaj disertacije je koncipiran i prikazan u poglavljima čiji je skraćeni sadržaj prikazan u ovom potpoglavlju.

U drugom poglavlju data je karakterizacija pojma kvaliteta artikulacije izgovornoh glasova u funkciji vrednosti mera njihovih artikulaciono-akustičkih atributa, koje definišu domen višedimenzionalnog prostora standarnih zadatih granica koji razgraničava tipične i atipične realizacije određene kategorije govornih segmenata. Prikazane su osnovne artikulaciono-akustičkih karakteristike različitih grupa fonema u kontekstu zavisnosti frekvencije pojavljivanja atipičnosti njihovog izgovora od složenosti procesa njihove produkcije. Ovde su prikazana dva testa koji sadrže detaljan opis atributa izgovornih glasova i načina za logopedsku ocenu kvaliteta njihove artikulacije, na osnovu kojih je razvijen autonomni računarski model. Pregled relevantne literature, prikazan je u svakom poglavlju i kompatibilan je sa saržajem odgovarajućih poglavlja.

U trećem poglavlju prikazan je sistem za ocenu kvaliteta artikulacije glasova srpskog jezika sa svim važnijim modulima i njihovom interakcijom datim na Slci 3.1. Početni modul (100) služi za akviziciju zvučnog signala. Drugi modul (200) predstavlja „Voice Activity Detection (VAD)“ algoritam za detekciju i diskriminaciju govornih zvučnih segmenata (reči) i segmenata tišine, od kojih se formira uzorak izgovornih reči i uzorak tišine. U trećem modulu (300) se vrši segmentacija govornog signala detektovanih reči na fonemske i subfonemske segmente (frejmove) i formira baza fonemskih segmenata. Izdvojeni segmenti su nosioci različitih distinktivnih obeležja koja karakterišu tipične i atipične realizacije fonema. Izdvajanjem pogodnih vektora obelžja iz ovih segmenata, u četvrtom modulu (400), dobijamo njihovu kondenzovanu formu očuvane informativnosti o pripadnosti klasama određenog kvaliteta na osnovu čega formiramo obučavajući i test uzorak vektora ulaznih atributa. Peti modul (500) predstavlja modul ekspertske – logopedske numeričke ocene kvaliteta artikulacije fonemskih segmenata primenom globalnog atrikulacionog testa. Ovako dobijene vrednosti se koriste kao vektori željenih izlaznih vrednosti za obučavajući i test uzorak. U šestom modulu (600) se vrši povećanje reprezentativnosti primenom originalnog Distance Besed Balancing (DBB)

algoritama za povećanje uniformnosti raspodele instanci u prostoru obeležja. U sedmom modulu (700) se obavlja obuka i verifikacija performansi četiri sledeća prediktora: Naive Bayes (NB), k najbližih suseda (kNN), samoorganizujuće mape (SOM) i ansambl višeslojnih perceptrona (MLP). Konačni izlaz iz sedmog modulaa je predikcija ocene kvaliteta artikulacije prezentovanih fonema, što je i konačni cilj ovog postupka.

U četvrtom poglavlju predstavljena je akvizicija i segmentacija zvučnog zapisa signala koja je izvedena kako ekspertski tako i algoritamski na osnovu ekstrakcije i klasifikacije karakterističnih distinktivnih obeležja artikulisanih fonema. Ručna segmentacija je neophodna inicijalna faza za formiranje incijalnog obučavajućeg uzorka paradigmi na osnovu ekspertskog razlikovanja karakteristika zvučnog signala, koji će poslužiti za formiranje algoritamskog metoda segmentacije. Detaljno su prikazana dva nivoa segmentacije: 1) ekstrakcija govornih segmenata koji sadrže izgovorne reči i segmenata tišine primenom VAD algoritma; 2) ekstrakcija fonema iz govornih segmenata reči čiji će kvalitet artikulacije treba oceniti. Reprezentativni vektor izabranih obeležja za VAD sadrži vrednosti sledećih komponenti izmerenih na subfonemskim segmentima - frejmovima: energija zvučnog signala, broj promena znaka amplitude talasnog oblika (ZCR), vrednosti autokorelacionih koeficijenata C1u granicama -1, +1, Linear Predictive Codding (LPC) koeficijenti i energija greške predikcije.

Vektor obeležja za segmentaciju fonema iz segmenata reči dobijenih rimenom VAD metoda, pored navedenih pet komponenti sadrži još 12 MFCC vrednosti. Na ovaj način je formiran trening uzorak zvučnih stimulusa i korespondentnih vektora obeležja. Obučavanjem klasifikatora pomoću inicijalnog trening uzorka vektora obeležja vršimo estimaciju parametara prediktora koji će poslužiti za dalju algoritamsku segmentaciju i proširenje baze govornih stimulusa reči i fonema različitih kvaliteta artikulacije.

Ovako proširena baza predstavlja relevantan skup stimulusa za ocenu kvaliteta artikulacije. Selekcijom reprezentativnih obeležja izgovornih fonema formiramo obučavajući i test uzorak vektora obeležja za klasifikatore koji će vršiti pouzdanu ocenu kvaliteta artikulacije. Svaka instanca predstavljena je vektorom izabranih atributa dužine 19. U ovom poglavlju takođe su prikazani rezultati tačnosti VAD ekstrakcije govornih segmenata iz kontinuiranog zvučnog signala kao i rezultati segmentacije karakterističnih fonema iz ekstrahovanih izgovornih reči primenom različitih algoritama. Ručno segmentirani govorni signali izvedeni od strane eksperata poslužili su kao referentni za ocenu tačnosti korišćenih alata. Prikazana automatizacija procesa priprema zvučnih stimulusa omogućava uvećanje baze stimulusa fenomena a shodno tome i relevantnosti obučavajućeg uzorka.

Peto poglavlje se odnosi na učenje u uslovima nieizbalansiranih podataka u smislu nejednake zastupljenosti klasa u uzorku. Data je procedura analize reprezentativnosti raspoloživog uzorka vektora obeležja. Pošto su učeći prediktori poznati kao „*data driven"* modeli, što znači prediktori rukovođeni podacima, neophodno je uzeti u obzir reprezentativnost korišćenih podataka i koristiti metode za njeno povećanje. Dva osnovna pristupa problemu su interni pristup zasnovan na intervenciji na raspodeli raspoloživih podataka (resampling) i eksterni ili algoritamski pristup koji se zasniva na kreiranju novih i korišćenju postojećih algoritama koji se prilagođavaju postojećoj raspodeli klasa. Postoje dve osnovne kategorije internih pristupa problemu balansiranja podataka izvorno poznatie kao: *oversampling* i *undersampling*, što bi se moglo prevesti povećanje ili smanjenje broja instanci u različitim klasama uzoraka. Najčešće uzorci sadrže različit broj primeraka u različitim klasama što dovodi do favorizacije frekvencije pojavlivanja dominantne klase u odnosu na manjinsku klasu, tako da se često deševa da svi primerci test uzorka budu klasifikovani u većinsku klasu. Ova pojava se može sprečiti ili smanjenjem broja instanci u većinskoj klasi ili povećanjem broja u manjnskoj klsai. I smanjenje i povećanje broja instanci se može vršiti na slučajan ili dirigovan način. Posebna oversampling metoda je SMOTE (Synthetic majority Oversampling Tenique). Ova metoda generiše novi sintetički skup primeraka manjinske klase. Sintetički uzorci su interpolirani u prostor obeležja manjinske klase u cilju simulacije novih realnih instanci i balansiranja ravnoteže broja predstavnika različitih klasa. U ovom poglavlju je prezentovana originalna, za potrebe ovog rada dizajnirana „Distance Based Balancing" (DBB) metoda. Ova metoda se zasniva na generalnom pristupu balansiranja podataka u cilju povećanja njihove reprezentativnosti kroz maksimizaciju entropije. Polazna pretpostavka tumači svaki primerak (instancu) raspoloživog uzorka kao tačku u višedimanzionalnom prostoru karakterističnih obeležja tog uzorka. Pošto uniformna raspodela instanci uzorka garantuje njegovu maksimalnu entropiju i shodno tome maksimalnu reprezentativnost, ovaj algoritam generiše nove instance u porostoru njihove male gustine i smanjuje broj instanci u prostoru velike gustine na način da se postiže kvazi uniformna raspodela. Time se povećava reprezentativnost uzorka. Ovo poglavlje prikazuje detaljnu komparativno analizu trinaest aktuelnih algoritama za balansiranje podataka u svetlu prikaza prednosti i nedostataka DBB algoritma.

U šestom poglavlju su prikazane teorijske osnove četiri tipa klasifikatora, metode njihove obuke i komparativna analiza njihovih performansi, prednosti i nedostaka. Opisani su k najbližih suseda (kNN), Naive Bayes (NB), samoorganizujuće mape (SOM) i ansambl višeslojnih perceprtona (MLP).

Jedna od najjednostavijih i najstarijih metoda za klasifikaciju je kNN klasifikator. On klasifikuje nepoznate realizacije (instance) u klasu kojoj pripada većina od k njima najbližih suseda. Blizina je definisana merom rastojanja, najčešće euklidsko rastojanje, raspoloživog trening skupa instanci iz kog se odabira predefinisani broj susednih instanci. Uprkos imanentnoj jednostavnosti, kNN metoda daje rezultate uporedive sa mnogo složenijim metodama. Međutim, na kNN često negativno utiču neinformativne variable u podacima, što je slučaj sa visoko dimenzionalnim podacima. Kao standardna metoda, kNN klsifikator je uvršten u aktuelni skup prediktora.

Jednostavni Bayes, NB klasifikator, je u stvari primenjena Bayesova teorema koja se neposredno zasniva na računanju uslovnih verovatnoća događaja. Na dobro dizajniranom trening skupu, ovaj algoritam daje rezultate komparabilne sa ostalim standardnim agoritmima, što je razlog njegove primene u istraživanju.

Samoorganizujuće Mape predstavljaju uopšteni model lateralne interakcije neurona u slojevima (laminama) kore velikog mozga, u cilju karakterizacije i diskriminacije spoljašnjih stimulusa. Perceptroni su bihevioristički modeli naučenih relacija stimulus – odgovor na višem nivou organizacije. Najveći nedostatak perceptrona je u tome što se podrazumeva da svaka neuronska ćelija operiše uglavnom samostalno, mada u paralelnom radu sa drugim ćelijama, jer se apstrahuje njihova lateralna interakcija unutar sloja. Na osnovu znanja o organizaciji velikog mozga, izuzetno važan tip organizacije struktura sa povratnim granama je takozvani *lateralni povratni model* ili Kohonenov *laminarni model mreža*. Ovaj model podrazumeva interakciju neurona istog sloja koja rezultuje promenom odziva jedinica, odnosno kao ekscitacija ili kao inhibicija u odnosu na primarni signal tokom niza vremenski diskretnih koraka. Evidentno je da se fenomen *klasterovanja* može ostvariti primenom različitih formi lateralne povratne funkcije, pa je matematička formalizacija ove pojave osnova računarkog modeliranja samoorganizacije neurona. Ovi modeli su naročito primenjivi za klasifikaciju uzoraka kada ne postoji definisan indikator klasa. Ovde su primenjeni kao jedna od standardnih metoda klasifikacije. U ovom poglavlju takođe su prikazani Višeslojni Perceptroni, kao računarske strukture za obradu informacija zasnovane na generalizovanim matematičkim modelima principa morfološko funkcionalne organizacije centralnog nervnog sistema. Višeslojni perceptron spada u neuronske mreže sa propagacijom signala unapred, čija se obuka odvija pod nadzorom, odnosno u prisustvu signala željenog odziva na predefinisani ulaz. MLP koristi generalizovano delta pravilo učenja, odnosno pravilo povratne propagacije signala greške. Jedna od najvažnijih primena perceptrona, pored aproksimacije funkcija, je klasifikacija uzoraka. Pored niza operativnih prednosti poput fleksibilnosti i robustnosti perceptroni imaju

poznate nedostatke poput lokalnih minimuma i overfiting situacije koji su posledica slučajnog izbora početnih parametara i uticaja disbalansa predstavnika različitih klasa. Efikasna procedura za kompenzaciju ovih problema podrazumeva primenu ansambla perceptrona sa većinskim odlučivanjem pri klasifikaciji koji je pored ostalih, korišćen klasifikator zasnovan na MLP ansamblu. U ovom istraživanju prikazan je originalan način izabora optimalnog ansambla perceptrona u svrhu povećanja prediktorskih performansi, i tako dizajniran klasigikator je korišćen za ocenu kvaliteta artikulacije glasova i poređen sa ostalim prediktorima.

U sedmom poglavlju prikazani su rezultati primene svih vrsta klasifikatora za ocenu kvaliteta artikulacije i izvršeno međusobno poređenje njihovih performansi. Poređenje tačnosti klasifikatora je izvršeno posredno preko etalona za tačnost ocena kvaliteta artikulacije koji je formiran kao većinska odluka grupe od pet iskusnih logopeda. Na osnovu rangiranja mera tačnosti klasifikatora zaključeno je da najbolje rezultate pokazuje optimalni MLP ansambl obučen na balansiranim uzorcima uz primenu DBB algoritma. Drugi po rangu sa neznatno manjom tačnošću je kNN klasifikator. Takođe su uporedno prikazani rezultati komparativne korelacije vektora ocena primenjenih klasifikatora i pojedinih logopeda u odnosu na vektor etalon, kojom prilikom je utvrđeno da je tačnost klasifikatora na nešto većem nivou u odnosu na individualnu tačnost logopeda. Detaljno su prikazani rezultati primene novog DBB algoritma za povećanje reprezentativnosti trening uzoraka i rezultati komparcije ovog algoritma sa standardnim algoritmima za neizbalansirano učenje. Takođe su dati grafički prikazi funkcionalne zavisnosti kvaliteta artikulacije od određenih obležja na osnovu osetljivosti posmatranih MLP struktura na perturbacije ulaznih variabli. Ove funkcije sadrže informacije relevantnosti uključenih obležja. Prikazani rezultati primene upotrebljenih alata na raspoloživim uzorcima srpskih fonema, predstavljaju opšti pregled prednosti i mogućnosti primene i eventualnih korekcija sistema za računarsku ocenu kvaliteta artikulacije glasova srpskog jezika. Tokom prikazanih istraživanja izvedeni su dokazi početnih pretpostavki kao ključnih parcijalnih uslova ostvarivosti disertacijom definisanog cilja. Ovi dokazi koji podrazumevaju ostvarenje nekoliko anticipiranih istraživačkih rezultata, istovremeno predstavljaju i originalan doprinos prikazane disertacije.

U osmom poglavlju prikazan je pregled postignutih rezultrata istraživanja, generalni doprinos istraživanja oblasti računarske evaluacije kvaliteta artikulacije i pozicija prikazanih istraživanja u relevantnom informacionom prostoru i potencijalni uticaj na primenu i buduće istraživanje u logopedskoj praksi. Posebno su istaknuti rezultati razvoja novih metoda za poboljšanje performansi učećih prediktora kako u algoritamskom domenu tako i u domenu načina odabiranja obučavajućih uzoraka iz raspoloživih podataka.

2. KVALITET ARTIKULACIJE, MANIFESTACIJA I REPREZENTACIJA

Govorna komunikacija, koja po sebi predstavlja razmenu informacija između sagovornika putem govora kao složenog zvučnog signala, je najvažnija govorna funkcija i odvija se sukcesivno na sledeća tri nivoa realizacije: neuro-fiziološkom (artikulacionom), fizičkom (akustičkom) i auditivnom, takođe neuro-fiziološkom nivou realizacije. Svaki od ovih nivoa predstavlja poseban složeni proces a poremećaji ili prekidi na bilo kom od njih degradiraju kvalitet ili onemogućavaju proces komunikacije. Artikulacioni nivo se u suštini odnosi na mehanizam pravilnog kodiranja poruke u složeni govorni signal posredstvom naučenih paradigmi koordinisane aktivacije komponenti govornog aparata od strane govornika, kao najkompleksnije procedure ne samo na nivou govorne komunikacije već generalno na nivou ljudskih neuro-fizioloških aktivnosti. Akustički nivo predstavlja akustičku manifestaciju procesa artikulacije govora kroz interakciju artikulisanog govornog signala sa lokalnim ambijentom, njegov prenos do slušaoca kroz medij za propagaciju koji je u realnim uslovima najčešće izložen uticaju šuma. U opštem slučaju podrazumevamo da je taj medij vazduh a šum koji potiče od različitih izvora zvuka u ambijentu ima relativno malu energiju u odnosu na korisni govorni signal. Auditivni nivo se odnosi na recepciju i percepciju govornog signala kao važan uslov komunikacije, tokom koje je auditivna pažnja slušaoca usmerena na strukturu njegovih akustičkih atributa kao nosilaca kodirane poruke, odnosno na adekvatan prijem signala i pravilno dekodiranje ili razumevanje poruke do nivoa njegove razumljivosti. Ovaj nivo podrazumeva i aferentnu regulatornu artikulaciono-akustičku povratnu spregu za prenos signala od uha do govornih struktura mozga govornika.

2.1. Fiziološki aspekt produkcije govora

Tokom produkcije govora, sistem govornih organa se nalazi u stanju koordinisane aktivnosti koja rezultuje produkcijom, potiskivanjem, propouštanjem, oblikovanjem i oslobađanjem vazdušne fonacione struje. Sa aspekta uloge u ovim aktivnostima, ovaj sistem se deli na generatore i modulatore govora. Generatore čine induktori (pluća i dušnik) i fonatori (grkljan i glasnice). Modulatore čine rezonatori (ždrelo, nosna i usna duplja) i artikulatori (jezik, usne, vilice, zubi, nepca i resica).

Pluća kao induktori snabdevaju govorni trakt izdahnutim (egresivnim) vazduhom sa indukovanom kinetičkom energijom koji se preko traheje prenosi do grkljana (larinksa) i

glasnica, koji ga kao fonatori transformišu u usmerenu fonacionu struju i time ostvaruju svoju osnovnu funkciju u govoru.

Glasnice su par mišićnih nabora u srednjem suženom delu larinksa koje se odlikuju vrlo složenim mehnizmom vibracija i kao takve predstavljaju neophodnu kariku u procesu produkcije govora. Prostor u larinksu ograničen glasnicama naziva se glotis i on predstavlja granicu između subglotalnog i supraglotalnog prostora koji se odlikuju različitim pritiscima vazduha tokom procesa fonacije, bilo da je u pitanju tonalna ili atonalna fonacija. Najvažnija manifestacija tonalne fonacije je pojava tonalne fonacione struje, odnosno zujnog zvuka poznatog kao osnovni (laringealni) ton koji je čujna vibracija egresivne vazdušne struje tokom prolaska kroz glotis pod dejstvom gradijenta pritiska. Svaki impuls vibracija laringealnog tona je manifestacija jednog ciklusa rastavljanja i sastavljanja glasnica, pri čamu se otvara i zatvara glotis. Učestanost otvaranja i zatvaranja glotisa određuje frekvanciju laringealnog tona koja se kreće od 70 – 80 Hz kod dubokuh muških glasova (basova) do 1200-1400 hz kod najviših ženskih vokala (soprana). Amplituda horizontalnih pomeranja glasnica tokom fonacije određuje intenzitet govora. Intenzitet govora se može regulisati na dva načina: putem pojačane aktivnosti disajnih mišića i kontrolisanim zatvaranjem glotisa. Tonalna (zvučna) fonacija u larinksu predstavlja početak uobličavanja složenog zvuka u sistemu rezonantnih prostora, odnosno početak produkcije oko dve trećine konkretnih realizacija fonema, odnosno glasova. Ovim procesom nastaje zvučnost kao zajednički kvalitet vokala, sonanata i zvučnih šumnih konsonanata.

U larinksu se takođe odvija i atonalna (bezvučna) fonacija koja se manifestuje nastankom atonalne fonacione struje, na osnovu gradijenta pritiska u glotisu, i to sužavanjem glotisa bez dodira i vibracija glasnica, što predstavlja u stvari propuštanje neozvučene vazdušne struje koja će se tek u rezonatorima transformisati u bezvučne šumne konsonante. Ovi konsonanti čine oko jedne trećine konkretnih realizacija fonema u glasovnom sistemu Srpskog jezika. Pored fonacije kao osnovne govorne funkcije, lariks i glasnice imaju važnu regulatornu ulogu u variranju frekvencijskih karakteristika govora koja se manifestuje kroz promenu frekvencije vibracija glasnica pri ostvarenju raznih akcentskih i intonacionih obrazaca i naročito pri pevanju. Treba istaći još jednu ulogu larinksa u produkciji govora koja se ispoljava pri pomeranju larinksa naviše i naniže, pri čemu se manjaju rezonantne karakteristike faringealnog rezonantnog prostora što uzrokuje promenu strukture pratećih tonova kod nekih govornih segmenata (artikulacija vokala u i vokala i).

Rezonatori čine jedinstvenu celinu supraglotalnih šupljina pozicioniranih između larinksa i usana, koje se odlikuju inherentno različitim zapreminama i oblicima kao i velikim

mogućnostima promene zapremine i oblika karakterističnim za biološke adaptivne sisteme. Zahvaljujući ovakvoj morfološko-funkcionalnoj organizaciji, rezonatori imaju velike mogućnosti modulacije laringealne fonacione struje (tonalne i atonalne), u smislu pojačanja i prigušenja prispelih tonova i šumova u cilju stvaranja govornog izraza kao složenog zvuka i modeliranja govornih segmenata. Važna odlika rezonatora je produkcija harmonika osnovnog tona koja se kod govora manifestuje pojavom pratećih tonova (formanata) i koncentrata šuma različitih frekvencijskih i intenzitetskih karakteristika, koji prate osnovni laringealni ton ili atonalnu fonacionu struju. Kompozicija različitih karakteristika ovih pratećih tonova i šumova pri produkciji govornih segmenata predstavlja originalni skup jedinstvenih, distinktivnih akustičkih kvaliteta, odnosno jedinstvenu akustičku strukturu tih govornih segmenata. Jedinstveni akustički kvaliteti koje govorni segmenti stiču u rezonatorima govornog trakta predstavljaju jednu grupu neophodnih uslova ostvarivosti procesa jezičke komunikacije u smislu kodiranja i dekodiranja sadržaja poruke, dok drugu grupu uslova predstavljaju distinktivni kvaliteti koje determinišu artikulatori.

Supraglotičke šupljine rezonatornog sistema su: ždreona (faringealna), nosna (nazalna) i usna (oralna) duplja. Farinks predstavlja nastavak laringealnog prostora i pretkomoru usne i nosne duplje pa se u skladu sa tim deli na laringofarinks, nazofarinks i orofarinks. Između farinksa i nosne duplje se nalazi meko nepce koje po potrebi može preusmeravati fonacionu struju ka nosnoj duplji. Artikulaciona interakcija jezika i farinksa dovodi do promena obima i oblika faringealnog prostora što znatno menja rezonantne karakteristike ove duplje i frekventne karakteristike produkovanih govornih segmenata. Nosna duplja ima nepromenjive inherentne rezonatorske osobine i koristi se povremeno, tokom artikulacije nazalnih i nazalizovanih glasova, tako što se meko nepce u određenoj meri spušta i delimično propušta fonacionu struju iz farinksa kroz nosne otvore.

Oralni rezonatorski prostor determiniše finalnu akustičku strukturu najvećeg broja govornih segmenata. Suštinski rezonatorski artikulacioni potencijal usne duplje potiče od velikih mogućnosti promene zapremine i oblika rezonatorskog prostora, determinisane velikom pokretljivošću donje vilice, usana i jezika kao najvažnijih artikulatora. Vertikalni pokreti vilice menjaju širinu rezonantnog prostora i veličinu usnog otvora dok pokreti usana menjaju dužinu ovog prostora kao i veličinu i oblik usnog otvora. Pokreti jezika utiču na brojne promene inherentnih karakteristika i konfiguracije oralnog rezonatora. Dimenzije ukupnog rezonantnog prostora govornog aparata od larinksa do usana kod odraslih muškaraca kreću se u proseku oko 17 cm dužine i 5cm^2 poprečnog preseka. Morfološko funkcionalna struktura ovog prostora u

potpunosti determiniše domen frekventnog opsega glasova koji se u njemu formiraju u granicama od 80Hz do 10000Hz.

Artikulatori svojom pozicijom i pokretima tokom dinamičkog procesa produkcije govora menjaju oblik i vlumen rezonantnog prostora kroz koji teče fonaciona struja i na taj način daju konačne kvalitete govornim segmentima i izrazima. Anatomske i funkcionalne karakteristike (predispozicije) artikulatora čine ih najvažnijim delom govornog trakta odgovornim za produkciju razlikovnih kvaliteta govornih segmenata neophodnih za kodiranje jezičkih poruka. Dve kinematički različite kategorije artikulatora uključene u proces artikulacije su nepokretni artikulatori (delovi gornje vilice – zubi, alveolarni greben i tvrdo nepce) i pokretni artikulatori (meko nepce sa resicom, donja vilica, jezik i usne). Vremenski promenljivi relativni odnos (pozicija) pokretnih i nepokretnih artikulatora determiniše niz trajektorija artikulatora u rezonantnom prostoru koje korespondiraju akustičkim kvalitetima produkovanog govornog izraza. Promena oblika rezonatora u interakciji sa artikulatorima može se odvijati bez prekida toka fonacione struje, može delimično prekidati i usmeravati tok ostatka fonacione struje, može stvarati tesnac kroz koji se struja probija a može kratkotrajno potpuno prekinuti tok struje formiranjem pregrade.

Nepokretni artikulatori - Zubi gornje vilice (sekutići), uz pomoć pokretnih artikulatora, učestvuju u stvaranju pregrade tokom partikulacije ploziva (t, d), tesnaca pri artikulaciji frikativa (c, z, f), kompleksa pregrada-tesnac kod afrikata (c, dz) i delimičnog usmeravanja struje pri artikulaciji sonanta (v). Na gornjim kutnjacima se vrši usmeravanje fonacione struje kod vokala i stvaranje jednog dela pregrade u usnoj duplji kod nazala (n, nj) i dela pregrade kod ploziva (d, t).

Alveolarni greben je odgovoran za stvaranje delimčne pregrade kod formiranja laterala l, stvaranje pregrade u usnoj duplji kod formiranja nazala n i vibranta r.

Tvrdo nepce (palatum durum) je gornji prednji deo usne duplje gde pokretni artikulatori delimično usmeravaju fonacionu struju tokom artikulacije sonanata (j, lj), stvaraju pregradu tokom produkcije nazala (nj), tesnac kod artikulacije frikativa (š, ž) i kombinaciju tesnaca i pregrade prilikom artikulacije afrikata (č, dž, ć i đ). Gore navedene realizacije glasova se odnose glasovni sistem artikulacione baze govornika srpskog jezika.

Pokretni artikulatori – Meko nepce (palatum molle) čini gornji zadnji mišićni deo usne duplje koji se završava resicom. Anastomsko – morfološke karakteristike nepca omogućavaju njegovo podizanje, spuštanje i zatezanje tokom procesa disanja i govora. Tokom produkcije govora, meko nepce usmerava fonacionu struju ka oralnoi ili nazalnoj rezonantnoj duplji čime determiniše nazalnost ili oralnost kao razlikovne karakteristike dve grupe glasova. Kada je

meko nepce podignuto tada je isključen nosni rezonator i rezonantnog sistema što rezultuje produkcijom oralnih glasova kojih ima najviše u glasovnim sistemima svih jezika. Pri spuštenom mekom nepcu za vreme artikulacije produkuju se nazalni konsonanti odnosno (sonanti m, n, nj). Prilikom artikulacije ploziva (k, g) na mekom nepcu se stvara pregrada a prilikom artikulacije frikativa (h) stvara se tesnac.

Donja vilica svojim pokretima kontroliše vilični ugao, veličinu otvora između zuba, utiče na položaj usana, definiše oblik i zapreminu usne duplje kao rezonatora i tako utiče na konačnu formu nekih zvučnih segmenata.

Jezik je mišićni organ koji se odlikuje izuzetnom pokretljivošću u svim pravcima i velikim mogućnostima promene oblika, što omogućava pokrete naviše-naniže, napred-nazad kao i kombinacije ovih pokreta sa jedne strane a sa druge strane skraćivanje, produživanje, sužavanje, proširivanje, savijanje i promnene oblika u bilo kom položaju. Navedeni pokreti i variabilnost oblika jezika menjaju i determinišu oblik, volumen i rezonantne karakteristike oralnog i faringealnog rezonantnog prostora. Ove promene karakteristika rezonatora rezultuju odgovarajućim frekvencijskim pozicijama pratećih harmonika i formanata u strukturi vokala. Upravo je diverzitet rasporeda formanata generator distinktivne razlike akustičkih kvaliteta među vokalima koja ih čini različitim o d drugih. Jezik učestvuje u delimičnom usmeravanju fonacione struje tokom artikulacije sonanata: kod nazala (n, nj) učestvuje u stvaranju pregrade u usnoj duplji, kod laterala (l, lj) učestvuje u formiranju delimične pregrade, formira žleb kod poluvokala j, i zajedno sa alveolama generiše brzu naizmeničnu izmenu režima slobodnog protoka i prekida fonacione struje kod vibranta r. Ovakvim pokretima jezik omogućava formiranje i oblikovanje pratećih tonova (sonantskih formanata) u delovima rezonantnog prostora i sinhronizovano uobličavanje originalnih koncentrata šuma koji karakterišu pomenute glasove. Primicanjem drugim artikulatorima praćenim promenom oblika, jezik formira tesnace u kojima se formiraju karakteristični šumni energetski koncentrati energije frikativa (s, z, š, ž, x) distribuirani u različitim domenima frekventnog opsega. Jezik takođe učestvuje u formiranju pregrada koje uzrokuju potpune kratkotrajne prekide protoka fonacione struje kroz rezonantni prostor, omogućavajući na taj način pojavu naglih akustičkih šumnih efekata (eksplozija) tokom artikulacije određenih ploziva (t, d, k, g). Mogućnost brzih promena položaja i oblika jezika u interakciji sa pasivnim artikulatorima omogućava koordinirano sukcesivno formiranje pregrade i tesnaca na putu fonacione struje što rezultuje određenim frekventnim rasporedom energetskih koncentrata šuma i tišine koji karakteriše afrikate (c, dz, ć, đ, č, dž).

Usne u koordinaciji sa ostalim artikulatorima tokom produkcije govora deluju sa jedne strane na promenu volumena i oblika rezonatornog sistema a sa druge strane mogu delimično

usmeravati, delimično zaustavljati i potpuno prekidati tok fonacione struje. Usne čini jedan kružni mišić i sitem ostalih mišića koji ih povezuju sa mišićima za mimiku lica. Isturanjem i zaobljavanjem usana pri produkciji vokala (o, u) produžava se rezonantni prostor i snižava frekvencija pratećih tonova. Isturanjem i zaokruživanjem usana tokom artikulacije frikativa (š, ž) i afrikata (č, dž) dobija se dodatni rezonantni prostor koji uzrokuje snižavanje frekvencije pratećih koncentrarta šumne energije frikcije ovih glasova. Priljubljvanje usana rezultuje prekidom toka fonacione struje kroz usnu duplju tokom artikulacije nazala m potpunom blokadom toka fonacione struje prilikom artikulacije ploziva p i b, čime se determiniše konačna akustička slika ovih glasova. Blago približavanje donje usne gornjim sekutićima uslovljava delimično usmeravanje fonacione struje i omogućava uporedno formiranje pratećih sonantskih formanata i energhetskih koncentrata šuma određenih frekvencija kod sonanta v. Približavanjem usana gornjim sekutićima uz potpuni ali vrlo mekani dodir, formira se tesnac u kom se generiše vrtložna fonaciona struja što rezultuje pojavom niskofrekvencijskih koncentrata šumne energije pri artikulaciji frikativa f.

2.2. Akustički aspekt produkcije govora

Generator glasa su glasnice koje se odlikuju inherentnom sposobnošću vibriranja prilikom prolaska vazdušne fonacine struje kroz glotis. Fonacija je proces produkcije glasa u larinksu zasnovan na modulaciji kinetičke energije vazdušne sruje koja uzrokuje vibracije glasnica, čija je manifestacija glas. Ffrekvencija glasnica je uslovljena anatomsko funkcionalnim karakteristikama i zavisi od mase, dužine, elastičnosti i tenzije glasnica. Vibracije glasnica su posledica međudejstva kinetičke energije vazdušne struje nastele pod dejstvom subglotičkog nadpritiska i elastične energije mišićnog tkiva glasnica. Obe pomenute energije su generisane sinergijom složenih neurofizioloških mehanizama. Današnji pristup procesu fonacije ima utemeljenje u radovima Van de Berg-a (1958), Njihovi modeli vibracija, glasnica pri analizi različitih procesa (načina) fonacije uključuju u razmatranje aerodinamčke karakteristike vazdušne struje i biomehaničke karakteristike glasnica i larinksa, Oni su rezultat mišićne djelatnosti koja usklađuje respiraciju i fonaciju. Veliki progres u razumevanju teoretske osnove fonacija predstavljajaja dvodelni model vibtacija glasnica prikazan u radu Ishizaka i Flanagan (1972). Ovde predstavljeni model objašnjava takozvani vertikalni talas koji se uoćava pri kontrakciji glasnica i širi se odozdo naviše u larinksu.

Za izražavanje putem govora Kašić (2000a), upotrebljava sintagmu "govorni izraz", koji smatra osnovnom, najprirodnijom i najpotpunijom formom konkretizacije i realizacije jezika.

Autorka govorni izraz definiše kao zvučnu signalnu supstancu kojom se prenosi sadržaj poruke upućene sagovornicima. Ova signalna supstanca nastaje aktivnošću fiziološke baze govora transformacijom kinetičke energije indukovanog vazduha u akustičku energiju. Visok nivo anatomsko funkcionalne organizacije i specijalizacije govorne fiziološke baze čoveka najvećim delom determiniše akustičke karakteristike i kvalitet govora generalno, pod uslovom normalnih individualnih predispozicija govornika i normalnih okolnosti za razvoj govora. Ista autorka u analizi akustičkog aspekta govora, definiše govor kao složeni zvuk proizveden fonatornim mehanizmom u čiji sastav ulaze osnovni ton, viši harmonični tonovi (formanti) i viši neharmonični koncentrati šuma. Jovičić (1999) za govor koristi sintagmu "govorni signal" kao akustički signal nastao sekvencijalnim povezivanjem glasova, na bazi lingvističkih pravila, koji nosi govornu informaciju.

Osnovne karakteristike akustičke strukture ili kvaliteta zvučnih segmenata u kontinualnom govoru su trajanje, frekvencija i intenzitet. Ove osobine su produkt koordinirane aktivnosti sistema govornih organa i pojavljuju se u kontrolisanom promenljivom odnosu u govornim segmentima, što omogućava formiranje i varijacije akustičkih kvaliteta kao nosilaca informacionog sadržaja govornih segmenata.

Trajanje glasova zavisi od sledećih faktora: individualnih karakteristika govornika, uslova komunikacije, vrste i uloge glasa u segmentu kom pripada, relevantne suprasegmentne strukture ... Kod neakcentovanih slogova glasovi u proseku traju 80 do 100 milisekundi. Ljudske perceptivne mogućnosti postavljaju donju perceptabilnu granicu trajanja glasa na 20 milisekundi. Gornja granica nije tačno utvrđena ali se uzima kao granica izobličenja koja definiše granicu poremećene strukture i prepoznatljivosti. Istraživanja su pokazala da najduže trajanje imaju vokali, zatim frikativi, pa sonanti, plozivi i afrikati. Variranje trajanja glasova služi kao suprasegmentna izražajna osobina.

Frekvencija leži u osnovi većine distinktivnih karakteristika glasova. Kao fizička pojava, frekvencija govornog signala je determinisana brojem punih oscilacija glasnica u jedinici vremena. Frekvencijski opseg ljudskog govora je određen karakteristikama sistema govornih organa, prevashodno rezonantnih prostora, i realizuje se između 80 i 10000 Hz.

U domenu frekvencijskih tkarakteristika glasova postoje permanentna (inherentna) i varijabilna obeležja. Inherentna obeležja su permanentni, neodvojivi kvaliteti glasa bez kojih glas ne može biti prepoznat niti razlikovan u opoziciji na druge glasove. Suštinu inherentnih obeležja čine specifični tip, raspored i relativni odnos pojačanih frekvencijskih oblasti, odnosno formanata. Navedimo nekoliko generalnih inherentnih frekvencijskih karakteristika glasova. Svi zvučni glasovi (vokali i sonanti) se odlikuju osnovnim laringealnim tonom,

takozvanom osnovnom učestalošću F_0 ili fundamentalnom frekvencijom, koji se prostire u zoni niskih frekvencija. Osnovni ton se produkuje oscilacijom glasnica i definiše različite inherentne karakteristike i igra različitu ulogu u različitim grupama glasova. U slučaju vokala i sonanata laringealni ton je pokretač produkcije harmoničnih tonova koji se mogu pojačati u odgovarajućim frekventnim zonama. Ta pojačana područja su nosioci formanata i za razliku od osnovnog tona (F_0) imaju razlikovnu funkciju. Osnovni ton je zvuk bez boje slabog intenziteta koji svoje akustičke kvalitete dobija filtriranjem u rezonantnim šupljinama vokalnog trakta. Ovaj ton se ne čuje tokom govora ali njegovi viši harmonici postižu adekvatnu glasnoću tokom rezonancije u vokalnom traktu. Uklanjanjem osnovnog tona iz govornog spektra, dobija se rezidualni ton ali se glas i dalje normalno čuje, što je slučaj kod telefonskog signala koji ne prenosi područje frekvencija ispod 300 Hz, gde se upravo nalazi osnovni ton i mi bez problema prepoznajemo boju i ostale kvaliteta glasa naših sagovornika. Na osnovu istraživanja, Titze i Talkin (1979) su došli do nekoliko važnih zaključaka o osnovnom tonu:

- Fundamentalna frekvencija osnovnog tona uslovljena je veličinom larinksa (dužina i širina), dužinom glasnica i samo donekle debljinom glasnica;
- Kod konkretnog larinksa određene veličine, osnovna frekvencija je prevashodno uslovljena uzdužnom tenzijom mišićnih vlakana glasnica. Subglotički pritisak vazdušne struje kao i promene dužine glasnica takođe dovode do promene zategnutosti tkiva što dodatno simultano utiče na frekvencijsku visinu tona;
- Intenzitet osnovnoga tona takođe zavisi od dimenzija larinksa, preciznije od dužine glasnica;
- Povećanje subglotičkog pritiska vazdušne struje i abdukcija glasnica, povećavaju intenzitet osnovnog tona, dok povećanje tenzije tkiva najčešće umanjuje njegov intenzitet;
- U komparaciji uticaja dimenzija larinksa i subglotičkog pritiska na intenzitet, pokazalo se da je intenzitet zvuka na izlazu iz usne duplje proporcionalan dužini glasnica ali je subglotički pritisak najpouzdaniji parametar za ocenu intenziteta zvuka.

Bezvučni glasovi nemaju osnovnu frekvenciju ali se odlikuju koncentratima šuma. Bez obzira na individualni frekvencijski opseg nultog formanta za svaki glas je karakterističan očuvani utvrđeni odnos u rasponu između osnovnog tona i pojačanih frekvencijskih područja formanata u zonama viših frekvencija. Takođe je sačuvan odnos raspona između prvog i drugog formanta, drugog i trećeg, trećeg i četvrtog i tako dalje. Isto pravilo važi i za koncentrate šuma kao inherentne karakteristike bezvučnih konsonanata koji su pandani formanata. Termin formant se odnosi na pojačana frekventna područja pratećih tonova (harmonika) kod vokala i sonanata, a termin koncentrati šuma se odnosi na pojačane

frekvancije raznih tipova pratećih šumova kod konsonanata. Vokalski formanti su pojačana tonska područja sa jako naglašenim intenzitetom u zoni frekvencija ispod 3000Hz koja se odlikuju energetski istaknutim pojasevima sa grebenima. Osnovna inherentna odlika vokala je očuvanje frekvencijskog raspona između prvog i drugog formanta, ali i odnos među ostalim formantima je takođe bitan. Sonantski formanti, za razliku od vokalskih, se odlikuju pojačanim tonskim područjima sa manje izraženim intenzitetskim razlikama i primesama nenaglašenih šumnih koncentrata, a takođe su locirani u zoni nižih frekvencija. Koncentrati šuma su naglašena frekventna područja aperiodičnih oscilacija generisanih vrtloženjem fonacione struje u rezonatorima pri artikulacije konsonanata. Energetski koncentra šuma prema dinamičkim karakteristikama se dele na postepene odnosno frikativne (turbulentne) i nagle odnosno eksplozivne (abruptne). Prema distribuciji koncentrati šuma se mogu naći u zonama viših i nižih frekvencija ali takođe mogu biti difuzno raspoređeni po celom frekventnom opsegu. Jedna od inherentnih osobina konsonanata odnosi se na trajanje koncentrata šuma gde se frikativi odlikuju pojačanom energijom na celom vremenskom domenu glasa, kod afrikata pojačana energija se detektuje na drugoj polovini glasa dok se kod ploziva pojačana koncentracija energije manifestuje na samom kraju produkcije. Kod konsonanata osnovni ton ima distinktivnu ulogu u uodnosu na slične glasove koji ga nemaju. Promena frekvencije osnovnog tona tokom govora produkuje obeležje melodije i intonacije kao izražajnih sredstava u komunikaciji. Prema Kašić (2003a), promenljiva frekvencijska obeležja segmenata klasifikuju se na dve grupe: individualna i poziciona. Individualne varijacije obeležja su lična, neponovljiva govorna svojstva svakog čoveka prema kojima se on razlikuje od svih ostalih govornika, isključujući blizance gde su ove razlike upadljivo manje. Individualne polne i uzrasne varijacije osnovnog tona manifestuju se na pomeranje naviše i naniže u frekventnom području naglašenih pratećih tonova, tako što nizak ili visok osnovni ton uslovljava individualno nižu ili višu frekvenciju vokalskih formanata, sonantnih formanata i energetskih koncentrata šuma kod konsonanata. Kod većine muškaraca osnovni ton se produkuje u području između 80 i 180 Hz, kod većine žena ovaj raspon je između 180 i 230 Hz, a kod dece je u zoni 230 do 300Hz.

Pozicione varijacije su posledica varijacija pozicije govornih segmanata u odnosu na druge segmente u govornom izrazu što se manifestuje pojavom i objašnjava pojmom koartikulacije. Tokom realnog govora proces artikulacije pojedinačnih glasova ili slogova prati brza racionalizovana optimalna promena položaja govornih organa neophodna za atrikulaciju sledećeg glasa. Ove brze kontinualne promene položaja artikulatora dovode do takozvanih pozicionih varijacija, što se definiše kao prelazno nestabilno stanje tranzicije glasova koje se

odlikuje pojasom velike varijabilnosti frekvencijskih karakteristika između dva stabilna pojasa, koji odgovaraju stabilnom režimu artikulscije uključenih glasova. Pozicione varijacije akustičkih kvaliteta govornih segmenata su posledica transpozicije govornih organa koje se mogu predstaviti skupom njihovih trajektorija. Nestabilni režim artikulacije odgovara naglim promenama u ponašanju trajektorija govornih organa. Svaki akustički realizovan glas se odlikuje početnom i završnom tranzicijom čije frekventne karakteristike zavise od glasova involviranih u koartikulaciju. Zone tranzicije kao posledica koartikulacije se najbolje mogu analizirati na spektrogramima gde se uočava da početnim i završnim tranzicijama u normalnom govoru odgovara veliki deo trajanja svakog glasa što determiniše izrazitu varijabilnost akustičke strukture istih glasova u različitom koartikulacionom okruženju kod istog govornika. Koartikulacija se može posmatrati kao uticaj produkcije (artikulacije) govornih segmenata (glasova ili slogova) na kvalitet artikulacije bližih i daljih susednih segmenata, kako prethodnih tako i narednih. Artikulacija govornih segmenata se može smatrati dinamičkim procesom koji se karakteriše pojavom inercije govornih organa. Ova inercija pokazuje remanentno dejstvo na promenu kvaliteta predstojećih segmenata, tako da oni imaju različite akustičke kvalitete u poređenju sa kvalitetima istog segmenta u inicijalnoj poziciji u nekom drugom govornom izrazu. Koartikulacija se takođe odlikuje psihološkim nivoom interakcije, u smislu anticipacije predstojećih segmenata iz pozicije aktivnog segmenta što rezultuje pokretanjem naučenih paradigmi aktivacije relevantnih govornih organa, u kojima iskustvo ima veliki značaj.

Intenzitet ili jačina zvuka (I) je fizička veličina koja predstavlja emitovanu energije talasa zvučnog izvora u 1 sekundi kroz površinu od 1 m^2, generišući pri tome pritisak, a koja se izražava u W/m^2. Percepcija zvuka posredstvom uva, kada je u pitanju intenzitet, je ograničena na domen između praga čujnosti i granice bola. Pragom čujnosti se ovde definiše minimalna jačina zvuka koju judsko uho može da registruje kao zvuk i ona je zavisna od frekvencije zvuka pa za frekvencije između 1000 Hz i 4000 Hz, gde je uvo najosetljivije, prag čujnosti normalnog uva iznosi: $I_0=10^{-12}$ W/m². Korespondentne vrednosti fizičkih karakteristika zvučnog signala na pragu čujnosti su: amplituda otklona čestica koja iznosi oko 10^{-11} m, dok amplituda akustičkog pritiska iznosi oko $2 \cdot 10^{-5}$ Pa.

Jačina zvuka oko 10 W/m² izaziva bol u ušima i ova jačina se označava kao granica bola. Za zvuk na granici bola amplituda otklona čestica je 10^{-5} m, a akustički pritisak 30 Pa. Iznad ove granice uvo nije u mogućnosti da percipira zvuk već samo registruje bol.

Intenzitet glasa, koji percipiramo kao glasnoću, direktno je proporcionalna amplitudi vibracija glasnice i subglotičkom pritisku vazdušne struje. Što su veće amplitude vibracija glasnica, intenzitet glasa je veći i obrnuto (Stemple, 1992). Iz praktičnih razloga jačina glasa se

izražava u decibelima (dB), što je logaritamska mera odnosa između aktuelnog i referentnog intenziteta (I_0). Maksimalni (dinamički) raspon intenziteta glasa (od najtišeg do najjačeg) kreće se obično do 70 dB. Srednji intenzitet glasa tokom govora u normalnim uslovima okruženja iznosi oko 60 dB. Intenzitet glasa tokom govora je promenljiv i uslovljen je sa više faktora: rastojanje sagovornika, ambijetalna buka, način govora ... Buka je čest uzrok pojačavanja intenziteta glasa u cilju njenog nadglasavanja. Neretko, naročito kod needukovanih govornika, povećanje jačine glasa povlači povećanje njegove frekvencije i takav glas se naziva povišenim glasom. Intenzitet govornog izraza najčešće varira iz dva razloga. Individualne razlike među govornicima u smislu ukupnog intenziteta su prvi razlog, dok drugi razlog leži u aktuelnim komunikativnim zahtevima. Intenzitet glasova je njihvo inherentno obeležje po kome se razlikuju. Izračunati prosečne decibelske vrednosti za svaki pojedinačni glas nekog jezika i dovesti ih u međusobnu relativnu intenzitetsku relaciju predstavlja dobru osnovu za formiranje egzaktne metrike. Za engleski jezik postoji takva tabela gde najnižu referentnu vrednost (0 dB) ima bezvučni dentalni frikativ (th), dok otvoreni vokali imaju najvišu vrednost (oko 30 dB).

2.3. Auditivni aspekt govora

Auditivna percepcija govora je termin zastupljen u fonetici i predstavlja psihofiziološki proces primanja i dekodiranja govornog signala. Dakle ovaj pojam se odnosi na psihološku reakciju slušaoca na simultanu pojavu zvuka i značenja poruke inkorporirane u taj zvuk (glas). Adekvatna percepcija zahteva od slušaoca primijem zvučnih signala i poznavanje ustaljenih glasovnih obrazaca jezičkog izražavanja koji služe za kodiranje poruka u različitim situacijama. Za normalnu komunikaciju, poterebno je poznavanje svih ostalih nivoa jezičke strukture i kompleknih odnosa među jezičkim znacima što znači poznavanje jezika. Prenošenje zvučnih vibracija kroz vazdušnu sredinu predstavlja akustičku fazu dok je auditivna percepcija slušaoca predstavlja fiziološku fazu komunikacije (Kašić, 2003a). Ljudski mozak kao računar za obradu podataka prima zvučne talase i u vrlo kratkim vremenskim intervalima obrađuje kodirane informacije. Ali još uvek nije razjašnjen način kako mozak percipira elementarne informacije iz govornog signala. Definisan je pojam perceptivnih jedinica koja predstavlja minimalni segment govornog signala koji se razume kao najmanja jezička informacija . Istraživanja govore da ulogu ovih jedinica imaju vokali i slogovi sastavljeni od vokala i konsonanata u proizvoljnom redosledu (Jovičić, 1999). Percepcije traje toliko da omogućava odvijanje dva procesa: detektovanje akustičkih karakteristika signala i akumulacija i integracija u perceptivnu jedinicu. Na osnovu ovih postavki zaključujemo da se govor sastoji od niza

ovakvih jedinica, koje se u višim neurofiziološkim nivoima pretvaraju u jezičku poruku. Ova činjenica navodi na zaključak da se percepcija jedne perceptivne jedinice završava pre nego što počne obrada sledeće. Ovi procesi se odvijaju u takozvanoj *preperceptivnoj memoriji* čija aktivnost traje koliki je vremenski period produkcije perceptivne jedinice i to iznosi oko 250 ms (Jovičić, 1999). Oblast auditivne fonetike je zbog nemogućnosti primene direktnih metoda istraživanja kasnila za oblastima akustuike i artikulacije. Ipak, i ovde postoje značajna istraživanja uticaja akustičkih karakteristika govornog singnala na auditivno-perceptivnom domenu. U eksperimentu koji je izveo Horga (1988) ispitivan je stepen prepoznatljivosti izgovornog glasa u različitim opsezima frekvencije. Istraživanja su pokazala da trajanje frikativa utiče na percepciju mesta artikulacije (Hughes and Halle, 1956).

2.3.1. Logopedska percepcija kvaliteta artikulacije

Razumevanje logopedskog pristupa analizi karakteristika signala na fonemskom nivou je vrlo važan aspekt u dizajniranju računarskog modela tog pristupa. Logopedski pristup analize karakteristika fonema je u najvećoj meri determinisan standardnim testovima i to Globalni Artikulacioni Test (GAT) i Analitički Test (AT) (Kostić i sar. 1983, Vlaisavljević 1981). Oba testa se zasnivaju na auditivno perceptivnom detektovanju i prepoznavanju vrste odstupanja od tipičnog izgovora fonema srpskog jezika. Analitički test (Dodatak II) zbog velikog broja različitih odstupanja karakterističnih za različite kategorije glasova zahteva veliko logopedsko iskustvo pri artikulaciji opšteg utiska u konkretnu ocenu kvaliteta izgovora. Sva odstupanja definisana u AT testu su posledica varijabilnosti osnovnih artikulaciono akustičkih parametara ali je ta variabilnost generator velikog broja odstupanja i zato detaljno modeliranje svih ovih odstupanja pojedinačno ili u celini nema velikog smisla. Zato se u modeliranju logopedskog rada oslanjamo potpuno na GAT test koji u direktnom obliku ne sadrži pojedinačne kvalitete ili nedosrtatke izgovornih glasova već se logopedu prepušta da iznese ocenu u vidu globalnog stava o kvalitetu artikulacije. Bez obzira na taj globalni pristup iskusni logopedi se tokom formiranja GAT ocene oslanjaju na brzu analizu odstupanja definisanih upravo u AT testu. Pošto nam u ovakvoj situaciji nedostaje detaljna logopedska karakterizacija akustičke strukture (skup odstupanja) testiranih glasova, moramo pronaći način da detektujemo i izmerimo akustičke parametre koji prate kvalitet artikulacije glasova koji su ocenjeni od strane logopeda, kako bi smo uspostavili formalnu korespondenciju između parametara i ocene.

Pri ocenjivanju glasa GAT testom koristi se numerička skala sa vrednostima ocena od **1** do **7**. Ocene **1**, **2** i **3** predstavljaju tipičan izgovor pri tom se ocene **1** i **2** koriste za procenu

izgovora odraslih govornika (profesionalaca) dok se kod dece i prosečnih govornika koristi ocena **3**. Ocena **7** označava zamenu (supstituciju) ili izostavljanje (omisiju). Ocenama **4**, **5** i **6** predstavljaju se razni nivoi atipičnosti (distorzije) glasaova, koji mogu ići od najnižeg (lakog), do teške atipiočnosti. Ovi nivoi su memorisani u svesti logopeda koji ih tokom obrade detektuje i poredi sa paradigmama tipičnih izgovora i konvertuje u vrednost ocene. Na kraju procesa obrade stimulusa, logoped formuliše ocene od **3** do **6** u zavisnosti od opšteg utiska.

Osnovna pretpostavka u radu sa ovim testovima je logopedsko poznavanje kvaliteta tipičnih artikulaciono akustičkih karakteristika analiziranih glasova. Svako od detektovanih obeležja pri atipičnom izgovoru nekog glasa može se javiti na različitom nivou. Broj obeležja, njihov odnos, postojanje dominantnog obeležja i stepena njegovog uticaja predstavlja se na ocenama od **4** do **6**, gde veća ocene znači veći stepen odstupanja. Postoje istraživanje (Kašić, 2003) u kom se iznosi stav o tome da se kvalitet glasova ne percipira na osnovu ukupne akustičke strukture već na osnovu takozvanih akustičkih naputa, odnosno određenih diskriminativnih obeležja koja su presudna prilikom procene kvaliteta govornih segmenata. Ove činjenice o načinu obrade informacija pri ispitivanju prikazanim testovima ukazuju na složenost procesa ocenjivanja na osnovu slušanja logopeda ali bez obzira na to, ova tehnika procene akustičkih karakteristika govora tokom dijagnostike i tretmana, još uvek je u praksi nezamenljiva.

Najstarija tehnika analize artikulaciono-akustičkih karakteristika govora koja se očuvala do danas je auditivno-perceptivno procenjivanje sluhom (Kašić, 2003a). Ocenom kvaliteta artikulacije se bave logopedi na osnovu slušanja direktnog govora ili preslušavanja odgovarajućih audio zapisa. Suština metode se zasniva na komparaciji kompleksne opšte akustičke slike analiziranog govornog segmenta sa njegovom apstraktnom paradigmom formiranom na osnovu iskustva u svesti logopeda. Ova tehnika je poznata kao ekspertsko ili trenirano slušanje i može se koristiti za auditivnu procenu kvaliteta izgovora glasova kao osnovnih komponenti govornog izraza. Metoda u simplifikovanoj formi može poslužiti za jednostavno razgraničenje tipične i atipične realizacije glasova. Procedura ocene kvaliteta artikulacije glasova je početni i ključni momenat u tretmanu ispitanika sa govornom patologijom jer prethodi eventualnom terapeutskom procesu koji ima nemerljiv značaj za pacijenata i kao takav može biti dugotrajan i skup. Fundamentalni problemi artikulacije govora se odnose na kvalitet artikulacije osnovnih govornih entiteta, odnosno glasova i slogova, koja je neophodan ali ne i dovoljan uslov visokog kvaliteta artikulacije govora i same govorne komunikacije. Nepravilna artikulacije glasa tokom govora ugrožava njegovu akustičku sliku a u najgorem slučaju i njegovu distinktivnu ili kontrastivnu ulogu u govoru što dovodi do pogrešnog kodiranja sadržaja govornog signala. Istaknuta odlika govora je njegova

varijabilnost koja se manifestuje pojavom promenljivih vrednosti kvantitativnih i kvalitativnih mera artikulaciono-akustičkih atributa govornih segmenata i promenljivog odnosa među tim merama.

Varijabilnost govornih atributa je kategorija koja je uzrok fenomena „neponovljivosti" identične reprodukcije govornog izraza od strane istog govornika. Vrednosti pomenutih mera atributa definišu domen višedimenzionalnog prostora standarnih zadatih granica u koji se projektuju tipične realizacije određene kategorije govornih segmenata. Ovaj prostor je u fonetici poznat kao varijaciono polje bilo da se radi o artikulacionom ili akustičkom domenu. Svaka projekcija govornog segmenata iz iste kategorije u prostor van ovog područja je percipira se kao izvesna degradacija kvaliteta njegove artikulacije, odnosno, kao atipičnost njegove realizacije. Ova važna činjenica, naizgled prilično opšteg karaktera, je krucijalna polazna osnova matematičkog pristupa kategorizaciji kvaliteta izgovorenih glasova.

Varijabilnost govora je posledica kompleksnog procesa artikulacije, odnosno, složene interakcije velikog broja artikulacionih faktora čije je funkcionisanje zasnovano na naučenim paradigmama koordinisane aktivacije komponenti govornog aparata i čula sluha, koja direktno zavisi kako od iskustva i sposobnosti govornika tako i od uticaja psiho-emotivnih faktora verbalne eksperesije u realnoj situaciji. Kriterijume za ocenu kvaliteta artikulacije u praksi definišu eksperti-logopedi na osnovu iskustva u auditivnoj i vizuelnoj, odnosno, sinestetičkoj percepciji, kako artikulacionih i akustičkih korelata procesa artikulacije, tako i prateće spoljašnje facijalne ekspresije govornika. Kriterijumi kvaliteta artikulacije su utvrđeni i definisani u obliku standardnih artikulacionih testova u kojima je prisustvo unapred definisanih atributa indikator stepena atipičnosti izgovora glasova srpskog jezika (Kostić i sar., 1983). Na osnovu iskustvom stečene sposobnosti ocene kvaliteta artikulacije, odnosno opšte akustičke slike percipiranih glasova, logopedi vrše ocenu kvaliteta izgovorenih glasova u skladu sa standardnim testovima. Ipak ovakav način procene kvaliteta izgovora glasova još uvek ima naglašen subjektivistički karakter jer se zasniva na iskustvenoj evaluaciji i zavisi od uticaja različitih okolnosti i stanja eksperta. Pored toga, ova metoda zahteva dugotrajnu proceduru treninga logopeda što utiče na problem kadrova a sama procedura je dugotrajna i naporna.

2.4. Praktični aspekt ocene kvaliteta izgovora

Prethodno navedene činjenice su motiv za automatizaciju, simplifikaciju i objektivizaciju procesa ocene kvaliteta artikulacije primenom algoritma prihvatljive pouzdanosti. Još uvek ne postoji pouzdan opšti postupak za sveobuhvatnu automatsku ocenu kvaliteta izgovora glasova

srpskog jezika iako se u zadnje vreme pojavljuju predlozi rešenja specifičnih problema u oblasti. Uzrok ovakve situacije, leži pre svega u multidisciplinarnoj prirodi, kompleksnosti i nedovoljnom poznavanju kako procesa artikulacije tako i auditivne percepcije, pa samim tim i njihove neadekvatne formalne reprezentacije. Kao logičan pristup rešenju ovog problema nameće se potreba determiniacije optimalnog višedimenzionog skupa relevantnih kvantitativnih artikulaciono-akustičkih obeležja u vremenskom, amplitudskom, frekventncijskom i parametarskom domenu kao i obeležja u kvalitativnoj-kategorijalnoj formi koji determinišu kvalitet artikulacije. Ovaj vektor obeležja treba da pokrije što širu lepezu akustičkih kvaliteta koji su relevantni za logopedsku percepciju stimulusa, formiranje njihove mentalne akustičke slike i formiranje ocene kvaliteta njihove produkcije. Kompozicijom računarski determinisanog vektora obeležja zvučnih stimulusa i numeričkih vrednosti logopedskih ocena kvaliteta njihove produkcije, dobijamo hibridni obučavajući uzorak potreban za formalnog računarskog modela za procenu kvaliteta izgovora.

Rešenju ovog zadatka, zbog prirode govornog signala i bioloških osnova artikulacije pristupamo koristeći fleksibilne modele zasnovane na učećim prediktorima poznatijh pod opštim imenom „*data driven models*“, čija se struktura determiniše kroz interakciju sa raspoloživim podacima za obuku koji reprezentuju proces. Kao indikator potrebe i potencijalnog značaja predloženog istraživanja mogu poslužiti naučni radovi na polju automatske detekcije variranja kvaliteta govora (Hadjitodorov i sar., 2000; Manfredi., 2001; Godino-Llorente, i sar., 2004). Pojedini radovi su usmereni ka generalnoj detekciji poremećaja govora (Wallen i sar., 1996; Hadjitodorov i sar.,2002) dok se drugi bave specijalnim vrstama patologije (Maier i sar., 2009; Hossein i sar., 2009). U aktuelnoj literaturi uočava se parcijalnost pristupa u rešavanju problema automatske evaluacije kvaliteta izgovora na osnovu pojedinih akustičkih karakteristika govornog signala , kako u spektralnom (Markaki, Stylianou, 2011; Bilibajkić i sar. 2016) tako i onih definisanih u vremenskom domenu (Markaki, Stylianou, 2011; Jovičić i sar. 2010; Vasilakis, Stylianou, 2009; Paulraj i sar., 2009). Navedene reference se odnose na skorašnja istraživanja u oblasti što ukazuje na to da još uvek ne postoje optimalna rešenja aktuelnog problema pre svega u smislu definicije homogenog skupa relevantnih i pouzdanih parametara za ocenu kvaliteta govora. Kao posledica povećanja dimenzija skupa aktuelnih parametara aktuelizuje se kurs dimenzionalnosti koji uslovljava izvesna ograničenja za rešenje problema u smislu porasta kompleksnosti modela, relevantnosti (redundanse) uključenih parametara i „overfitting“ problema neuronske mreže koja je često , kao i ovde, korišćeni alat u analizi govornog signala (Haykin, 1998; Rabiner, Juang. 1993). Generalni problem konvergencije procesa obuke kod neuronskih mreža je pojava lokalnog

minimuma koja je uslovljena strukturom modela, početnim stanjem parametara i dimenzijom skupa za obuku, a za posledicu ima niži nivo tačnosti i kvaliteta generalizacije, odnosno pouzdanosti predikcije. Ovaj problem se rešava primenom ansambala neuronskih mreža (Hansen, Salamon, 1990; Sharkey i sar., 2002) i primenom algoritama obuke u neizbalansiranim uslovima (Japkowicz, Stephen, 2002; Haibo, Garcia, 2009). Koristeći prednost povećane informativnosti višedimenzionalnih vektora akustičkih atributa govornog signala, prednost povećane prediktabilnosti ansambla neuronskih mreža sa balansiranim procesom obuke pristupićemo dizajniranju pouzdanog modela za automatizovanu globalnu ocenu kvaliteta artikulacije (opšte akustičke slike) glasova srpskog jezika kao polaznom i ključnom koraku u dijagnostici govorne patologije.

Za valjanu ocenu kvaliteta izgovora upoređuju se rezultati grupe logopeda i formira jedinstvena ocena za svaki slučaj posebno. Dakle, radi se o zahtevnoj i skupoj proceduri koja uključuje subjektivni pristup problemu i donekle narušava objektivnost metode. Automatizacija ove procedure treba da zameni direktno učešće logopeda odnosno, isključi subjektivnost i pojednostavi i pojeftini ovaj proces. Pošto se radi o vrlo složenoj logopedskoj aktivnosti ocene slušanjem od strane grupe eksperata, neophono je definisati takav model koji će na prihvatljiv način formalizovati sve njene bitne aspekte.

Naš prvi zadatak u tom pravcu predstavlja razjašnjenje mehanizma adekvatne naučene percepcije kvaliteta izgovorenog sadražaja, od strane logopeda, u cilju detekcije skupa relevantnih artikulaciono-akustičkih manifestacija izgovorenih fonema, odnosno skupa obelažja kao ulaznih veličina na osnovu kojih logopedi ocenjuju kvalitet izgovora. Ovaj prvi korak, koji predstavlja teoretsku karakterizaciju relevantnih distinktivnih kvaliteta produkovanih fonema i njihovu evaluaciju, od velikog je značaja jer determiniše smer daljeg istraživanja, ali on, za razliku od sledećih koraka, najviše zavisi od eksperata logopeda. Na ovom stadijumu logopedi definišu ocenu kvaliteta izgovora u obliku celih brojeva od 1 do 7, na osnovu opšte akustičke slike stečene kompleksnom iskustvenom analizom pomenutih ulaznih obeležja. Iskusniji logopedi tokom evaluacije u analizu uključuju veći broj obeležja. Ipak, bez obzira na iskustvo logopeda, broj relevantnih obeležja analiziranog govornog signala je ograničen, a njihova pojava i varijacija su u velikoj meri perceptabilni, tehnički detektabilni i merljivi, i mogu se predstaviti u numeričkoj ili simboličkoj formi, što će nam pomoći pri automatizaciji njihove računarske obrade kao preduslova modeliranja.

Drugi, možda i najbitniji korak, se odnosi na observabilnu fizičku manifestaciju, relevantnih obeležja, njihovu detekciju, transformaciju i akviziciju u formi podataka pogodnih za automatizovanu računarsku obradu. Ovaj korak predstavlja parametrizaciju procesa i poznat je

kao ekstrakcija relevantnih obeležja, koji služe kao ulazi modela pri računarskom modeliranju procesa logopedske evaluacije kvaliteta artikulacije. Predloženi inicijalni skup obeležja takođe je podložan procesu ocene relevantnosti njegovih elemenata u cilju redukcije dimenzija i optimizacije finalnog modela.

Treći, finalni korak, predstavlja izbor tipa i strukture predloženog matematičkog modela klasifikatora koji će na osnovu ograničene raspoložive baze ulaznih vrednosti obeležja i odgovarajućih izlaznih celobrojnih vrednosti ocena izgovorenih fonema uspostaviti opšte važeći algoritamski model korespondencije između vektora obeležja i kvaliteta artikulacije fonema. Pošto složeni govorni signal kao akustička manifestacija govora, ima stohastičku prirodu, kao takav, upućuje nas na statistički pristup korišćenjem takozvanih „*data driven"* učećih modela, odnosno ekspertskih sistema zavisnih od podataka, kao najbolji izbor za modeliranje i evaluaciju njegovih karakteristika. U ovoj kategoriji algoritama treba tražiti one koji se odlikuju visokim stepenom fleksibilnosti praćenim potrebnom robustnošću kao međusobno oprečnim zahtevanim kvalitetima modela. Treba skrenuti pažnju da u ovom istraživanju, nismo direktno uključili vizuelni aspekt, odnosno informacije o vizuelnoj ekspresiji ispitanika, koja je za logopede vrlo bitna za ocenu artikulacije, a čiji se uticaj indirektno manifestuje u formiranju finalne ocene logopeda u Globalnom Artikulacionom Testu. Dakle, informacije o vizuelnom utisku logopeda se ne prezentuju učećim prediktorima kroz logopedsku globalnu akustičku sliku već se potenciraju isključivo akustički kvaliteti govora. Struktura pojedinih klasifikatora omogućava računarsku identifikaciju funkcionalne zavisnosti kvaliteta artikulacije od pojedinih artikulaciono-akustičkih obeležja. Pouzdani zaključci o kvalitetu artikulacije koji u osnovi predstavljaju finu diskriminaciju između atributa ograničenog skupa srodnih i vrlo sličnih fonemskih i subfonemslih govornih segmenata, mogu imati značaj za problem prepoznavanja govora koji se fokusira na diskriminaciju velikog skupa govornih segmenata sa izraženim diverzitetom atributa.

Složenost procesa ekstrakcije obeležja ukazuje na veliki značaj optimalnog izbora tipa i strukture klasifikatora i određivanje obučavajućeg uzorka podataka visokog nivoa reprezentativnosti kao bitnih uslova ispunjenja postavljenog zadatka. Zato je ovim koracima posvećena potrebna pažnja u metodološkom i praktičnom smislu i dat odgovarajući prostor u disertaciji. U slučaju optimalnog izbora atributa, grupe tipičnih i atipičnih realizacija glasova mogu biti potpuno razdvojene u višedimenzionalnom prostoru atributa, pa će dobro dizajniran klasifikator sa velikom verovatnoćom tačno klasifikovati uzorke u odgovarajuće klase.

3. SISTEM ZA OCENU KVALITETA ARTIKULACIJE

Analiza govornog signala se danas sve više koristi u medicini u svrhu dijagnostike različitih obolenja među koje spada i patologija govora koja se manifestuje insuficijencijom artikulaciono akustičkih kvaliteta govora i rezultuje nepravilnim govorom sa odstupanja od tipičnih normi izgovora u okviru nekog jezika. U zavisnosti od nivoa i težine ove devijacije se mogu javiti na nivou segmentne strukture govora (pojedinačni glasovi ili grupe glasova) i na nivou suprasegmentne strukture govora (melodija, tempo, ritam, akcenat). Detekcija odstupanja pri izgovaranju glasova podrazumeva različite metode koje se oslanjaju na oceni raznih vrasta distanci između normalnog i atipičnog izgovora. Svaka konkretna realizacija foneme (izgovorni glas) se definiše skupom vrednosti akustičkih obeležja, odnosno vektorom obeležja. U prostoru obeležja, sve varijacuije izgovora glasa nalaze se unutar ograničene višedimenzione oblasti. Ove oblasti su definisane kako za tipične tako i za atipične izgovore glasova bilo kog jezika a determinisane su lingvističkim i paralingvistčkičm parametrima. Iingvistički faktori su semantika, sintaksa, gramatika a paralingvistički su pol, uzrast, psihofizioloski status (emocije, bolest,...). Kao što vidimo, vektor parametara koji definišu granice varijacionog polja akustičkih obe1ežjaje može mati veliki broj komponenti. Odstupanje konkretnih realizacija glasa od normalne forme izgovora mogu biti lakša (mala ekskurzije akustičikih obe1ežja van oblasti tipičnog varijacionog po1ja), teža patološka forma (izgovorni glas je toliko izmenjen da ga je teško prepoznati kao fonemu koju predstavlja) i najteži oblik odstupanja kada se javlja supstitucija (zamena) i1i omisija (izostavljanje) izgovornih glasova.

Sistem za automatsku ocenu kvaliteta artikulacije treba da reši nekoliko važnih problema: određenje skupa relevantnih akustičkih obe1ežja koji reprezentuju određene foneme srpskog jezika u smislu njihove diskriminacije u odnosu na ostale, definisanje granične hiperpovrši u visedimenzionom prostoru izmedu tipične realizacije foneme i njenog odstupanja, definisanje adekvatnog obučavajućeg prediktora koji će na osnovu raspoloživog uzorka tipičnih i atipičnih realizacija foneme uspostaviti algoritam korespondencije između vektora obeležja i logopedske ocene kvaliteta artikulacije posmatranih fonema.

Najstarija tehnika analize artikulaciono-akustičkih karakteristika govora koja se očuvala do danas je auditivno-perceptivno procenjivanje sluhom. Ocenom kvaliteta artikulacije se bave logopedi na osnovu slušanja direktnog govora ili preslušavanja odgovarajućih audio zapisa. Suština metode se zasniva na komparaciji kompleksne opšte zvučne slike analiziranog govornog segmenta sa njegovom apstraktnom paradigmom formiranom na osnovu iskustva u svesti logopeda. Ova tehnika je poznata kao ekspertsko ili trenirano slušanje i može se koristiti

za auditivnu procenu kvaliteta izgovora glasova kao osnovnih komponenti govornog izraza. Metoda u simplifikovanoj formi može poslužiti za jednostavno razgraničenje tipične i atipične realizacije glasova. Procedura ocene kvaliteta artikulacije glasova je početni i ključni momenat u tretmanu ispitanika sa govornom patologijom jer prethodi eventualnom terapeutskom procesu koji ima nemerljiv značaj za pacijenata i kao takav može biti dugotrajan i skup. Fundamentalni problemi artikulacije govora se odnose na kvalitet artikulacije osnovnih govornih entiteta, odnosno glasova, koja je neophodan ali ne i dovoljan uslov visokog kvaliteta artikulacije govora i same govorne komunikacije. Nepravilna artikulacije glasa tokom govora ugrožava njegovu akustičku sliku a u najgorem slučaju i njegovu distinktivnu ili kontrastivnu ulogu u govoru što dovodi do pogrešnog kodiranja sadržaja govornog signala. Istaknuta odlika govora je njegova varijabilnost koja se manifestuje pojavom promenljivih vrednosti kvantitativnih i kvalitativnih mera artikulaciono-akustičkih atributa govornih segmenata i promenljivog odnosa među tim merama.

Varijabilnost govornih atributa je kategorija koja je uzrok fenomena „neponovljivosti“ identične reprodukcije govornog izraza od strane istog govornika. Vrednosti pomenutih mera atributa definišu domen višedimenzionalnog prostora standarnih zadatih granica u koji se projektuju tipične realizacije određene kategorije govornih segmenata. Ovaj prostor je u fonetici poznat kao varijaciono polje bilo da se radi o artikulacionom ili akustičkom domenu. Svaka projekcija govornog segmenata iz iste kategorije u prostor van ovog područja je indikator izvesne degradacije kvaliteta njegove artikulacije, odnosno, atipičnosti njegove realizacije. Ova važna činjenica, naizgled prilično opšteg karaktera, je krucijalna polazna osnova matematičkog pristupa kategorizaciji kvaliteta izgovorenih glasova.

Varijabilnost govora je posledica kompleksnog procesa artikulacije, odnosno, složene interakcije velikog broja artikulacionih faktora čije je funkcionisanje zasnovano na naučenim paradigmama koordinisane aktivacije komponenti govornog aparata i čula sluha, koja direktno zavisi kako od iskustva i sposobnosti govornika tako i od uticaja psiho-emotivnih faktora verbalne eksperesije u realnoj situaciji. Kriterijume za ocenu kvaliteta artikulacije u praksi definišu eksperti-logopedi na osnovu iskustva u auditivnoj i vizuelnoj, odnosno, sinestetičkoj percepciji, kako artikulacionih i akustičkih korelata procesa artikulacije, tako i prateće spoljašnje facijalne ekspresije govornika. Kriterijumi kvaliteta artikulacije su utvrđeni i definisani u obliku standardnih artikulacionih testova u kojima je prisustvo unapred definisanih atributa indikator stepena atipičnosti izgovora glasova srpskog jezika (Kostić i sar., 1983). Na osnovu iskustvom stečene sposobnosti ocene kvaliteta artikulacije, odnosno opšte akustičke slike percipiranih glasova, logopedi vrše ocenu kvaliteta izgovorenih glasova u skladu sa

standardnim testovima. Ipak ovakav način procene kvaliteta izgovora glasova još uvek ima naglašen subjektivistički karakter jer se zasniva na iskustvenoj evaluaciji i zavisi od uticaja različitih okolnosti i stanja eksperta. Navedene činjenice su motiv za automatizaciju, simplifikaciju i objektivizaciju procesa ocene kvaliteta artikulacije primenom algoritma prihvatljive pouzdanosti. Još uvek ne postoji pouzdan opšti postupak za sveobuhvatnu automatsku ocenu kvaliteta izgovora glasova srpskog jezika iako se u zadnje vreme pojavljuju predlozi rešenja specifičnih problema u oblasti. Uzrok ovakve situacije, leži pre svega u multidisciplinarnoj prirodi, kompleksnosti i nedovoljnom poznavanju kako procesa artikulacije tako i auditivne percepcije, pa samim tim i njihove neadekvatne formalne reprezentacije. Kao logičan pristup rešenju ovog problema nameće se ideja determiniacije optimalnog višedimenzionalnog skupa relevantnih kvantitativnih artikulaciono-akustičkih obeležja u vremenskom, amplitudskom, frekventncijskom i parametarskom domenu koji determinišu kvalitet artikulacije. Rešenju ovog zadatka pristupamo koristeći fleksibilne modele zasnovane na učećim prediktorima, čija se struktura determiniše kroz interakciju sa raspoloživim podacima za obuku koji reprezentuju proces. Kao indikator potrebe i potencijalnog značaja predloženog istraživanja mogu poslužiti naučni radovi na polju automatske detekcije variranja kvaliteta govora (Hadjitodorov i sar., 2000; Manfredi., 2001; Godino-Llorente, i sar., 2004). Pojedini radovi su usmereni ka generalnoj detekciji poremećaja govora (Wallen i sar., 1996; Hadjitodorov i sar.,2002) dok se drugi bave specijalnim vrstama patologije (Maier i sar., 2009; Hossein i sar., 2009). U aktuelnoj literaturi uočava se parcijalnost pristupa u rešavanju problema automatske evaluacije kvaliteta izgovora na osnovu pojedinih akustičkih karakteristika govornog signala , kako u spektralnom (Markaki, Stylianou, 2011; Bilibajkić i sar. 2016) tako i onih definisanih u vremenskom domenu (Markaki, Stylianou, 2011; Jovičić i sar. 2010; Vasilakis, Stylianou, 2009; Paulraj i sar., 2009). Navedene reference se odnose na skorašnja istraživanja u oblasti što ukazuje na to da još uvek ne postoje optimalna rešenja aktuelnog problema pre svega u smislu definicije homogenog skupa relevantnih i pouzdanih parametara za ocenu kvaliteta govora. Kao posledica povećanja dimenzija skupa aktuelnih parametara aktuelizuju se izvesna ograničenja za rešenje problema u smislu porasta kompleksnosti modela, relevantnosti (redundanse) uključenih parametara i „overfitting“ problema neuronske mreže koja je često , kao i ovde, korišćeni alat u analizi govornog signala (Haykin, 1998; Rabiner, Juang. 1993). Generalni problem konvergencije procesa obuke kod neuronskih mreža je pojava lokalnog minimuma koja je uslovljena strukturom modela, početnim stanjem parametara i dimenzijom skupa za obuku, a za posledicu ima niži nivo tačnosti i kvaliteta generalizacije, odnosno pouzdanosti predikcije. Ovaj problem se rešava

primenom ansambala neuronskih mreža (Hansen, Salamon, 1990; Sharkey i sar., 2002) i primenom algoritama obuke u neizbalansiranim uslovima (Japkowicz, Stephen, 2002; Haibo, Garcia, 2009).

Koristeći prednost visoke informativnosti višedimenzionalnih vektora akustičkih atributa govornog signala, prednost povećane prediktabilnosti ansambla neuronskih mreža sa balansiranim procesom obuke pristupamo dizajniranju pouzdanog modela za automatizovanu globalnu ocenu kvaliteta artikulacije glasova srpskog jezika kao polaznom i ključnom koraku u dijagnostici govorne patologije.

Za valjanu ocenu kvaliteta izgovora upoređuju se rezultati većinskog odlučivanja grupe logopeda i formira jedinstvena ocena za svaki slučaj posebno. Dakle, radi se o zahtevnoj i skupoj proceduri koja uključuje subjektivni pristup problemu i donekle narušava objektivnost metode. Automatizacija ove procedure treba da zameni direktno učešće logopeda odnosno, isključi subjektivnost, pojednostavi i pojeftini ovaj proces. Pošto se radi o složenoj logopedskoj aktivnosti ocene slušanjem od strane grupe eksperata, neophono je definisati takav model koji će na prihvatljiv način formalizovati sve njene bitne aspekte.

Naš prvi zadatak u tom pravcu predstavlja razjašnjenje mehanizma adekvatne naučene percepcije kvaliteta izgovorenog sadražaja, od strane logopeda, u cilju detekcije skupa relevantnih artikulaciono-akustičkih manifestacija izgovorenih fonema, odnosno skupa obelažja kao ulaznih veličina na osnovu kojih logopedi ocenjuju kvalitet izgovora.

Algoritam za ocenu kvaliteta izgovora glasova, koji bi se koristio u dijagnostici i terapiji treba da se oslanja na standardne procedure logopedske prakse gde spadaju dobro razvijeni testovi (Vladisavljavić, 1981; Kostić i sar. 1983).

Prvi korak, koji predstavlja teoretsku karakterizaciju relevantnih distinktivnih kvaliteta produkovanih fonema i njihovu evaluaciju, od velikog je značaja jer determiniše smer daljeg istraživanja, ali on, za razliku od sledećih koraka, najviše zavisi od eksperata logopeda. Iskusniji logopedi tokom evaluacije u analizu uključuju veći broj obeležja. Ipak, bez obzira na iskustvo logopeda, broj relevantnih obeležja analiziranog govornog signala je ograničen, a njihova pojava i varijacija su u velikoj meri perceptabilni, tehnički detektabilni i merljivi, i mogu se predstaviti u numeričkoj ili simboličkoj formi, što će nam pomoći pri automatizaciji njihove računarske obrade kao preduslova modeliranja.

Drugi, možda i najbitniji korak, se odnosi na observabilnu fizičku manifestaciju, relevantnih obeležja, njihovu detekciju, transformaciju i akviziciju u formi podataka pogodnih za automatizovanu računarsku obradu. Ovaj korak predstavlja parametrizaciju procesa i poznat je

kao ekstrakcija relevantnih obeležja, koja služe kao ulazi modela pri računarskom modeliranju procesa logopedske ocene kvaliteta artikulacije.

Treći korak predstavlja izbor tipa i strukture predloženog matematičkog modela prediktora koji će na osnovu ograničene raspoložive baze ulaznih vrednosti obeležja i odgovarajućih izlaznih celobrojnih vrednosti ocena izgovorenih fonema uspostaviti opšte važeći algoritamski model korespondencije između vektora obeležja i kvaliteta artikulacije fonema. Pošto složeni govorni signal kao akustička manifestacija govora, ima stohastičku prirodu, kao takav, upućuje nas na statistički pristup korišćenjem takozvanih „*data driven"* učećih prediktora, odnosno ekspertskih sistema zavisnih od podataka, kao najbolji izbor za modeliranje i evaluaciju njegovih karakteristika. U ovoj kategoriji algoritama treba tražiti one koji se odlikuju visokim stepenom fleksibilnosti praćenim potrebnom robustnošću kao međusobno oprečnim zahtevanim kvalitetima modela. Treba skrenuti pažnju da u ovom istraživanju, modeliranje procesa logopedske evaluacije kvaliteta izgovora nismo direktno uključili vizuelni aspekt, odnosno informacije o vizuelnoj ekspresiji ispitanika, koja je za logopede vrlo bitna za ocenu artikulacije, a čiji se uticaj indirektno manifestuje u formiranju finalne ocene logopeda u Globalnom Artikulacionom Testu.

Osamdesetih godina prošlog veka, došlo je do naglog razvoja veštačke inteligencija, interneta i informacionih tehnologija generalno, koji je omogućio razvoj efikasnih alata i formiranje velikog broja lako dostupnih baza podataka, između ostalog, i u domenu prepoznavanja govora i govorne patologije. Mogućnost pouzdane automatske evaluacije kvaliteta govora primenom potencijalnih alata dostupnih putem interneta znatno bi smanjila subjektivni uticaj logopeda i omogućlo simultano testiranje većih grupa pacijenata koje se može izvoditi bez direktnog prisustva logopeda, što bi znatno povećalo dostupnost i efikasnost rehabilitacije s jedne strane, i smanjilo troškove procedure sa druge strane. Ovakvi modeli se mogu koristiti u procesu edukacije i treninga logopeda povećavajući prostor za njihovo kreativnije i delotvornije angažovanje. Potencijalni pacijenti bi mogli koristiti ovakve alate za efikasan skrining stanja artikulacije, za samostalni trening, samostalnu rehabilitaciju ili detekciju ozbiljnijih poremećaja koji zahtevaju stručni tretman. Još uvek je rano očekivati da ovakvi sistemi potpuno zamene terapeute ali je evidentna njihova uloga u izvršenje velikog broja pojedinačnih operacija zahtevnih u smislu resursa. Ovakav sistem bi se mogao koristiti u cilju automatizovane komparativne ocene efikasnosti različitih terapeutskih procedura primenjenih na grupama pacijenata, kao i za praćenje stanja pacijenata tokom perioda rehabilitacije. Struktura pojedinih prediktora omogućava računarsku identifikaciju funkcionalne zavisnosti kvaliteta artikulacije od pojedinih artikulaciono-akustičkih obeležja što

može imati veliki značaj za rangiranje stepena relavantnosti posmatranih obeležja i interni uvid u eventuelne kauzalnosti složenog procesa artikulacije. Pouzdani zaključci o kvalitetu artikulacije koji u osnovi predstavljaju finu diskriminaciju između atributa ograničenog skupa srodnih i vrlo sličnih fonemskih i subfonemskih govornih segmenata, mogu imati značaj za problem prepoznavanja govora koji se fokusira na diskriminaciju velikog skupa govornih segmenata sa izraženim diverzitetom atributa.

Veliki broj reprezentativnih obeležja govornih segmenata i veliki broj metoda za ekstrakciju tih obeležja razvijenih za potrebe prepoznavanja govora i ocene kvaliteta artikulacije, uzrok su konfuzije pri izboru relevantnog skupa obeležja prihvatljive pouzdanosti i kardinalnosti, pa je optimalna određinanje ovih parametara jedan od glavnih problema ovog istraživanja. Složenost procesa ekstrakcije obeležja ukazuje na veliki značaj optimalnog izbora tipa i strukture klasifikatora i definisanje obučavajućeg uzorka podataka visokog nivoa reprezentativnosti kao bitnih uslova ispunjenja postavljenog zadatka. Zato je ovim koracima posvećena potrebna pažnja u metodološkom i praktičnom smislu i dat odgovarajući prostor u disertaciji.

Odabrani artikulaciono-akustički atributi formiraju višedimenzionalni prostor obeležja u kom se svaki primerak izgovorenog glasa projektuje u tačku Euklidskog prostora, sa tendencijom grupisanja tipičnih i atipičnih realizacija glasova u dve korespondentne, interno manje ili više homogene oblasti ovog prostora koje se međusobno mogu ali ne moraju presecati. U slučaju optimalnog izbora atributa, grupe tipičnih i atipičnih realizacija glasova mogu biti potpuno razdvojene u višedimenzionalnom prostoru atributa, pa će dobro dizajniran klasifikator sa velikom verovatnoćom tačno klasifikovati uzorke u odgovarajuće klase. Ovim želimo da kažemo, da i u slučaju potpunog razgraničenja različitih klasa u višedimenzionom prostoru atributa, nije garantovana njihova potpuna tačnost klasifikacije jer se od klasifikatora zahteva velika fleksibilnost pri generisanju vrlo složene granične hiperpovrši između klasa, i zato svaki klasifikator ne može u potpunosti da odgovori ovom zahtevu. U realnim slučajevima najčešće se dešava da se domeni raspodela tipičnih i atipičnih realizacija presecaju u prostoru obeležja tako da izbor učećeg prediktora i trening skupa uzoraka može imati presudan značaj za pouzdanost klasifikacije i ta vrsta problema će biti poseban domen ovog istraživanja. Pojedini analizirani atributi su dati u kategorijalnoj formi, u smislu prisustva ili odsustva određene mere atipičnosti, dok su ostali definisani celobrojnim ili racionalnim vrednostima. Efikasna estimacija domena i pronalaženje optimalnih i graničnih vrednosti obeležja, koje karakterišu tipične i atipične grupe realizacija glasova, će olakšati njihovo razgraničenje, odnosno proces klasifikacije.

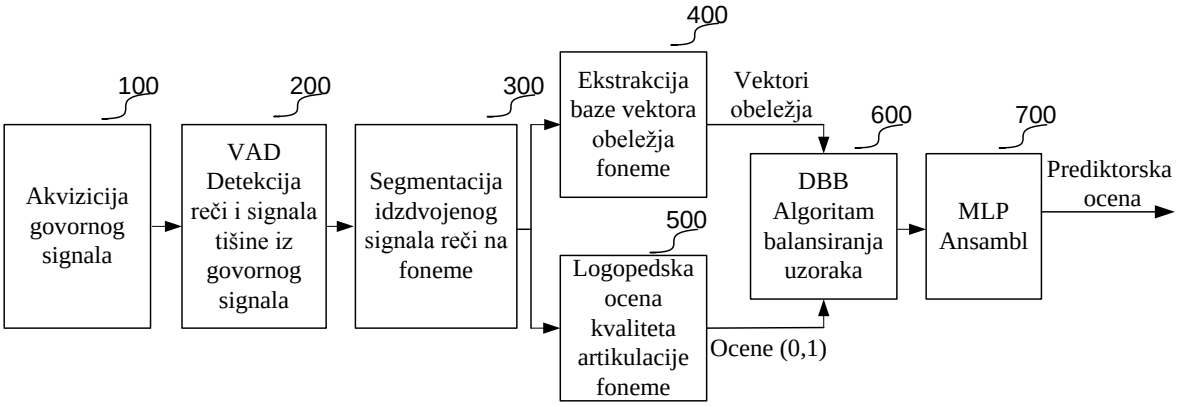
Pošto su performanse učećih prediktora uslovljene reprezentativnošću obučavajućeg uzorka, posebna pažnja je posvećena analizi problema narušene reprezentativnosti raspoloživog obučavajućeg uzorka, koja je posledica neravnomernosti njegove raspodele u prostoru obeležja i/ili disbalansa u zastupljenosti pojedinih klasa u uzorku. Uniformna raspodela uzorka u prostoru obeležja je indikator njegove maksimalne entropije koja korespondira sa maksimalnim stepenom njegove reprezentativnosti u odnosu na target populaciju. Dakle, naš glavni cilj pri izboru trening uzorka iz raspoloživog uzorka populacije će biti povećanje njegove reprezentativnosti putem maksimizacija vrednosti njegove entropije što se postiže adekvatnim metodama odabiranja, odnosno, balansiranja uzoraka.

Analiza govornog signala sa namerom da se detektuju artikulaciona odstupanja izgovora glasova u govornom signalu i dijagnostika različitih patologija govora i glasa se ozbiljno istražuje od polovine prošlog veka (Lieberman, 1961). Razvoj računarske tehnike, različitih metoda digitalne obrade signala, algoritama za prepoznavanje oblika (pattern recognition), prepopznavanja govora i govornika, pratili su i razvoj procedura za analizu kvaliteta govora na osnovu govornog signala, identifikaciju kvaliteta artikulacije i govorne patologije (Miehaelis i sar., 1998, Cesar i sar., 2000, Maguire i sar., 2003, Maguire i sar., 2003a, Moran i sar., 2004, Godino-Llorente i sar. 2004) itd. Glavni cilj ovih istraživanja leži u ekstrakciji vektora obeležja u segmentima govornog signala tipičnog i atipičnog (patološkog) glasa i detekcija tih odstupanja na osnovu različitih kriterijuma i različitih primenjenih algoritama klasifikacije. Postoji veliki broj vrlo različitih pristupa od izbora analiziranog govornog segmenta, do određinanja skupa akustičkih obeležja za analizu, korišćenja različitih matematičkih modela za klasifikaciju i kategorizaciju, tumačenje rezultata, itd. Pored obimnog istraživačkog rada, još uvek nije dovoljno istražena veza relevantnih akustičkih obeležja u govoru sa iskustvenom percepcijom artikulacionih odstupanja terapeuta.

3.1. Modeliranje procesa logopedske ocene kvaliteta artikulacije

Osnovni cilj istraživanja je projektovanje računarskog sistema za ocenu kvaliteta artikulacije glasova srpskog jezika. Blok dijagram ovog sistema prikazan na Slici 3.1 reprezentuje pojednostavljeni algoritamski model složenog logopedskog procesa ocenjivanja kvaliteta izgovora stimulusa izgovornih glasova. Prikazani računarski model ima sedam prikazanih modula organizovanih u pet funkcionalnih celina: moduli 100, 200 i 300 čine funkcionalnu celinu za predobradu govornih segmenata, modul 400 predstavlja proceduru za ekstrakciju vektora obeležja, modul 500 koji predstavlja proces ekspertske - logopedske analize

akustičkih kvaliteta govornih stimulusa i formiranja numeričke vrednosti odgovarajuće ocene, modul 600 predstavlja algoritam uz balansiranje inherentnog disbalansa uzorka u smislu reprezentativnosti klasa, modul 700 predstavlja potencijalne učeće klasifikatore gde je kao primer dat ansambl neuronskih mreža (MLP).



Slika 3.1 Blok dijagram Sistema za ocenu kvaliteta artikulacije glasova

3.2. Predprocesing snimljenog govornog signala

Preprocesing govornog signala se odnosi na obradu integralnog signala koji sadrži signal izgovornih reči po tačno definisanom redosledu i signal tišine naizmenično pozicionirane i podrazumeva sledeće: odvajanje efektivnog stimulusa izgovorenih reči od signala tišine, odnosno voice activity detection (VAD), segmentaciju izdvojenih reči na fonemske segmente (foneme) i subfonemske segmente (frejmove) i ekstrakciju relevantnih diskriminatornih obeležja kako na oba nivoa. Segmentacija signala na fonemskom nivou je najvažniji aspekt predobrade jer se algoritam zasniva na logopedskom pristupu analize karakteristika fonema u skladu sa standardnim testovima i to Globalni Artikulacioni Test (GAT) i Analitički Test (AT) (Kostić i sar. 1983, Vlaisavljević 1981). Svako izdvajanje segmenata podrazuzmeva nihovo razlikovanje po definisanom kriterijumu sličnosti, tako da je od samog početka neophodno uvesti parametrizaciju govornih segmenata u cilju diskriminacije signala reči i tišine. Metodološki pristup predviđa da ispitanici izgovaraju reči i rečenice po unapred tačno utvrđenom redosledu što uvodi dodatni determinizam olakšavajući celu proceduru modeliranja i apriori povećavajući tačnost. Korišćeni su različite metode parametrizacije signala tako da reprezentativni vektor obeležja za VAD sadrži vrednosti sledećih komponenti izmerenih na subfonemskim segmentima: energija freejma (E_f), broj promena znaka amplitude talasnog oblika (ZCR), vrednosti autokorelacionih koeficijenata C1, C2,..., u granicama -1, +1, Linear Predictive Codding (LPC) koeficijenti i energija greške predikcije. Vektor obeležja za

segmentaciju reči na fonemske segmente dobijenih primenom VAD metoda, pored navedenih pet komponenti sadrži još 12 MFCC vrednosti. Na ovaj način se formira trening uzorak zvučnih stimulusa i relevantnih vektora obeležja koji predstavljaju obučavajući uzorak za klasifikatore (Furundžić i sar. 2009, 2012b, 2013b, 2017c).

3.3. Ekstrakcija vektora obeležja kvaliteta izgovora fonema

U cilju razjašnjenja i računarske interpretacije mehanizma adekvatne naučene percepcije kvaliteta izgovorenog sadražaja, od strane logopeda, potrebno je detektovati skup relevantnih akustičkih manifestacija izgovorenih fonema, odnosno skup karakterističnih distinktivnih obelažja kao polaznih veličina na osnovu kojih logopedi ocenjuju kvalitet izgovora. Ovaj korak, od velikog je značaja jer usmerava tok daljeg istraživanja, ali on, za razliku od sledećih koraka, ima subjektivistički karakter jer najviše zavisi od logopeda. Iskusniji logopedi vremenom pri evaluaciji u analizu uključuju sve više obeležja. Na sreću, broj relevantnih obeležja govornog signala je ograničen, a njihova pojava i varijacija su u velikoj meri perceptabilni, tehnički detektabilni i merljivi. Observabilna fizička manifestacija, relevantnih obeležja, njihova detekcija, transformacija i akvizicija u formi podataka pogodnih za automatizovanu računarsku obradu predstavlja ekstrakciju relevantnih obeležja.

Naši laboratorijski uzorci sadrže skup akustičnih karakteristika govornog signala fonema srpskog jezika koje omogućavaju pouzdanu kasifikaciju (Furundžić i sar. 2012b, 2013b, 2017c). Posmatrani fonemi su izgovoreni u početnoj poziciji reči u GAT testu gde su pozicionirani. Svaka instanca predstavljena je vektorom izabranih atributa dužine 19. Prva grupa od dvanaest atributa predstavlja Mel Frekventnih Kepstralnih Coefficienata (MFCC) uzet iz svakog frejma. Tokom procesa obuke svakom okviru aktuelnog fonema pridružena je odgovarajuća labela klase. Tokom procesa testiranja, izlazi iz svakog ulaznog okvira svake foneme se sabiraju i usrednjavanjem se dobija procena pripadnosti nekoj od datih klasa kvaliteta artikulacije. Druga grupa od tri atributa datih koji se odnose na dužinu talasnog signala, sastoji se od broja semplova (nw) talasnog oblika signala reči koja u inicijalnoj poziciji sadrži aktuelne foneme, broj uzoraka talasnog oblika signala aktuelne foneme (nph), i odnos dužine ($nq = nph/nw$). Na kraju, treća grupa sadrži četiri sledeća atributa koja se odnose na distribuciju energije frejmova: energija frejmova foneme (Efph), energija frejmova odgovarajuće reči (Efv), ukupna energija fonema (tEph) i ukupna energija aktuelne reči (tEv). Glavni problem analitičkog sistema za procenu kvaliteta artikulacije je adekvatan izbor akustičkih parametara dovoljno pouzdanih za finu diskriminaciju unutar iste kategorije fonema.

3.4. Logopedski pristup oceni kvaliteta izgovora

U ovom modulu se nalaze relevantne opservacije logopeda o kvalitetu izgovora zvučnih stimulusa fonema u inicijalnoj poziciji odabranih reči u formi ocene (1, 0) gde 1 predstavlja indikator patološkog izgovora a 0 se dnosi na tipičan izgovor. Ovaj modul se sastoji od praktično dva nivoa logopedske analize: Analitički Test koji sadrži detaljnu analizu artikulaciono akustičkih karakteristika analiziranog izgovornog glasa u skladu sa Testom datim u dodatku II. U AT testu, svaku realizovanu fonemu karakteriše veći skup odstupanja od kvaliteta njene tipične realizacije, pa je broj i veličina stepena odstupanja od tipične realizacije u obrnutoj korelaciji sa kvalitetom artikulacije. Pošto se ovde radi o iskustvenom ekspertskom znanju ovaj segment neće biti važan predmet interesovanja osim u smislu naše svesti o načinu logopedskog odlučivanja. Logopedska ocena kvaliteta artikulacije na osnovu odstupanja artikulaciono-akustičkih obeležja je apstraktna naučena sposobnost formiranja relevantne akustičke slike foneme i uspostavljanja pouzdane audio-perceptivne korespondencije između te slike i kvaliteta artikulacije fonema. Drugi nivo analize predstavlja Globalni Artikulacioni Test prikazan u dodatku I. Ovaj test je relevantan za naš proces modeliranja evaluacije izgovora glasova jer rezultuje krajnjom ocenom kvaliteta artikulacije date celobrojnim vrednostima od 1 do 7 i celobrojnim aproksimacijama ovih ocena u binarnoj formi (1,0) koje su za nas relevantne. Dakle, ovaj deo modula na ulazu ima govorni segment foneme u inicijalnoj poziciji korespondentne reči koji treba oceniti. Na izlaznom delu se odvodi celobrojna logopedska ocena odstupanja, odnosno kvaliteta artikulacije. Ovaj modul (500) dakle formira izlazni signal logopedske ocena kvaliteta artikulacije dok modul 600 definiče korespondentni ulazni vektor obeležja i na taj način nastaje hibridbni obučavajući uzorak za učeće prediktore. Ovako definisani uzorci su klasifikovani primenom različitih prediktora, a najčešće preko veštačkih neuralnih mreža (Furundžić i sar.,2006, 2007, 2012b, 2017c, Bilibajkić i sar. 2014).

Detaljnija istraživanja u domenu upotrebe veštačkih neuralnih mreža za ocenu kvaliteta artikulacije dati su u poglavlju o klasifikatorima.

3.5. DBB algoritam balansiranja reprezentativnosti uzoraka

Ovaj računarski modul je odgovoran za povećanje reprezentativnostzi trening uzorka pre njegove prezentacije učećim prediktorima.

U ovom modulu je i primenjen novi opšti pristup neizbalansiranom učenju kao jednom od glavnih izazova u oblasti klasifikacije uzoraka. Preciznije, ovde je objedinjen data i

eksperimentalna potvrda prednosti novog algoritma za balansiranje neizbalansiranih klasa, zasnovanog na principu maksimizacije entropije uzoraka. Metod se zasniva na detekciji distribucionih karakteristika idealno izbalansiranog uzorka datog u formi pravilne rešetke i transferu ovih karakteristika na proizvoljni neizbalansirani uzorak, koristeći tehniku uzorkovanja kojom se menja struktura neizbalansiranog uzorka u pravcu porasta njegove reprezentativnosti i balansa. Predložena procedura podrazumeva uklanjanje postojećih instanci iz oblasti velike gustine raspodele verovatnoće (undersampling) u kombinaciji sa sintetičkom generacijom novih instanci u oblastima male gustine (oversampling). Ova procedura se primenjuje na svaku klasu pojedinačno u cilju redukcije unutrašnjeg imbalansa svake klase koje rezultuje značajnim poboljšanjem performamnsi involviranih klasifikatora.

Opservabilna manifestacija ovog algoritma je povećanje entropije uzoraka (klasa) koja implicira redukciju tendencije induktivnih klasifikatora da favorizuju većinsku klasu ili klaster tokom treninga. Prikazana teoretska osnova metoda je verifikovana na adekvatnom sintetičkom skupu uzoraka a njegova praktična primenjivost je potvrđena na komparativnoj klasifikaciji velikog skupa raspoloživih empirijkih podataka. Standardni induktivni algoritmi učenja su generalno dizajnirani da rade sa "dobro izbalansiranim" klasama, ali u realnosti oni uglavnom rade u uslovima neravnomerne distribucije podataka, kako unutar klasa tako i između klasa. U okviru klasa unutrašnji imbalans klasa utiče na performanse klasifikatora, dok je neravnoteža među klasama predstavlja mali problem kada postoji prihvatljivi balans unutar klasa, u smislu reprezentativnosti podataka (Japkovicz, 2003). Krajnji cilj primene ove metode je poboljšanje performansi klasifikatora pri klasifikaciji i oceni kvaliteta artikulacije fonema srpskog jezika (Furundžić i sar. 2017c, 2015).

3.6. Modeli za ocenu kvaliteta artikulacije

U ovom modulu su primenjena četiri tipa klasifikatora: k najbližih suseda (kNN), Naive Bayes (NB), samoorganizujuće mape (SOM) i ansambl višeslojnih perceprtona (MLP) koji je prikazan na Slici 3.1 kao jedan od prediktora. Jedna od najjednostavijih i najstarijih metoda za klasifikaciju je kNN klasifikator. On klasifikuje nepoznate realizacije (instance) u klasu kojoj pripada većina od k njima najbližih suseda. Blizina je definisana merom Euklidskog rastojanja, raspoloživog trening skupa instanci iz kog se odabira predefinisani broj susednih instanci.

Sledeći klasifikator u ovom modulu je jednostavni Bayes (NB) klasifikator, koji u stvari predstavlja primenjenu Bayesovu teoremu neposredno zasnovanu na računanju uslovnih verovatnoća događaja na sledeći način:

$$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c_i)P(c_i)}{P(\boldsymbol{x})}.$$

Gde $P(c_i|\boldsymbol{x})$ predstavlja aposteriornu verovatnoću dobijenu na trening uzorku, i služi za ocenu verovatnoće da neka test instanca $\boldsymbol{x}$, pripada klasi c_i. $P(c_i)$ je parametar prethodne verovatnoće klase c_i u odnosu na ostale klase dobijen na trening uzorku. $P(\boldsymbol{x})$ je evidentna prediktorska prethodna verovatnoća. $P(\boldsymbol{x}|c_i)$ je *likelihood* parametar, odnosno verovatnoća pojave test instance $\boldsymbol{x}$ pod uslovom realizacije date klase c_i. Prethodna prediktorska verovatnoća $P(\boldsymbol{x})$ ima konstantnu vrednost i ne utiče na vrednosti aposteriornih verovatnoća pa se ona može zanemariti što gornjoj jednačini za proračun aposteriorne verovatnoće test instance daje sledeću jednostavnu formu:

$$P(c_i|\boldsymbol{x}) \propto P(\boldsymbol{x}|c_i) \times P(c_i) = P(x_1|c_i) \times P(x_2|c_i) \times \ldots \times P(x_n|c_i) \times P(c_i).$$

Konačna odluka o klasi $c_i, i + 1,2, \ldots, L$ se donosi u skladu sa maksimumom od svih L vrednosti verovatnioća. $\underset{i}{argmax}(P(c_i|\boldsymbol{x}))$. Na dobro dizajniranom trening skupu, ovaj algoritam daje rezultate komparabilne sa ostalim standardnim agoritmima, što je razlog njegove primene u istraživanju.

Samoorganizujuće Mape predstavljaju uopšteni model lateralne interakcije neurona u slojevima (laminama) kore velikog mozga, u cilju karakterizacije i diskriminacije spoljašnjih stimulusa. Na osnovu znanja o organizaciji velikog mozga, izuzetno važan tip organizacije struktura sa povratnim granama je takozvani *lateralni povratni model* ili Kohonenov *laminarni model mreža*. Ovaj model podrazumeva interakciju neurona istog sloja koja rezultuje promenom odziva jedinica, odnosno ekscitacijom ili inhibicijom primarnog signala tokom niza vremenski diskretnih koraka. Evidentno je da se fenomen *klasterovanja* može ostvariti primenom različitih formi lateralne povratne funkcije, pa je matematička formalizacija ove pojave osnova računarkog modeliranja samoorganizacije neurona. Ovi modeli su naročito primenjivi za klasifikaciju uzoraka kada ne postoji definisan indikator klasa. Ovde su primenjeni kao jedan od standardnih metoda za predikciju klase.

U ovom modulu su primenjene veštačke neuronske mreže, kao računarske strukture za obradu informacija zasnovane na generalizovanim matematičkim modelima principa morfološko funkcionalne organizacije centralnog nervnog sistema. Višeslojni perceptron spada u neuronske mreže sa propagacijom signala unapred, čija se obuka odvija pod nadzorom, odnosno u prisustvu signala željenog odziva na predefinisani ulaz. MLP koristi generalizovano delta pravilo učenja, odnosno pravilo povratne propagacije signala greške. Jedna od najvažnijih primena perceptrona, pored aproksimacije funkcija, je klasifikacija uzoraka za šta

je ovde i upotrebljen. Pored niza operativnih prednosti poput fleksibilnosti i robustnosti perceptroni imaju poznate nedostatke poput problema lokalnih minimuma i overfiting situacije koji su posledica slučajnog izbora početnih parametara i uticaja disbalansa u brojnosti predstavnika različitih klasa respektivno. Efikasna procedura za rešavanje ovih problema podrazumeva primenu ansambla perceptrona sa većinskim odlučivanjem pri klasifikaciji je pored ostalih, korišćen klasifikator zasnovan na MLP ansamblu (Furundžić i sar. 2012a, 2017c, Bilibajkić i sar. 2014). U poglavlju o klasifikatorima ovog istraživanja prikazan je originalan način izabora optimalnog ansambla perceptrona i tako dizajniran klasigikator je korišćen za ocenu kvaliteta artikulacije glasova i poređen sa ostalim prediktorima.

3.7. Izbor baze govornika za ocenu kvaliteta artikulacije

Govorna baza i govorni stimulusi za analizu kvaliteta artikulacije glasova sačinjeni su dobro aproksimiraju distribucione karakteristike pojave glasova u normalnom neprekidnom govoru. Govornicima su prezentovani zadati stimulusi koje su ponavljali za logopedom, u vidu izolovanih reči iz GAT testa. Na taj način je formirana govorna baza stimulusa koja je korišćena za istraživanje. Uzorak govornika činila su 400 desetogodišnjaka i jedanaestogodišnjaka (po 200 dečaka i devojčica) koji su po pretpostavci imali automatizovanu artikulacionu bazu srpskog jezika kao maternjeg. Govornici se odlikuju: prosečnim intelektualnim sposobnostima, urednoim statusom govornih organa, urednim slušnim statusom, monolingvalnošću, pripadaju govornicima ekavskog izgovora štokavskog narečja, time što su rođeni i nastanjeni u Beogradu. Kao standard pravilne artikulacije uzet je postojeći standard, tj. opis akustičko-artikulacionih karakteristika izgovornih glasova srpskog jezika (Kostić i sar., 1964; Stevanović 1981). Govornici su izgovarali po 30 zadatih reči iz tabele GAT teesta, i time je kompiliran skup od 12000 reči od kojih su u istraživanju korišćene reči koje u inicijalnoj poziciji sadrže glasove sa najvećom frekvencijom odstupanja u kvalitetu, uglavnom frikative i afrikate. Govornici su snimani u akustički adaptiranoj prostoriji (tiha soba), pomoću mikrofona, sa frekvencijom odabiranja 44,1 kHz i 16 bitnom AD konverzijom. Snimci su memorisani u *wav* formatu. Segmentaciju ove govorne baza su izveli eksperti audo-vizuelnom metodom uz primenu softverskog alata *Praat* (Boersma, Weenink, 2010).

4. SEGMENTACIJA I EKSTRAKCIJA OBELEŽJA GOVORNOG SIGNALA

Izdvajanje i segmentacija aktivnog govornog signala predstavlja predobradu integralnog govornog signala koji sadrži naizmenično pozicionirane sekvance signala izgovornih reči i signala tišine i podrazumeva sledeće korake: razdvajanje efektivnog stimulusa izgovorenih reči od signala tišine, što je proces poznat kao „voice activity detection“ (VAD), segmentaciju izdvojenih reči na fonemske segmente (foneme) i subfonemske segmente (frejmove) kao i ekstrakciju relevantnih diskriminatornih obeležja na oba nivoa. Segmentacija signala na fonemskom nivou je najvažniji aspekt predobrade jer se naš računarski algoritam za ocenu kvaliteta izgovora fonema zasniva na logopedskom pristupu analize karakteristika fonema u skladu sa standardnim testovima i to Globalni Artikulacioni Test (GAT) i Analitički Test (AT) (Kostić i sar. 1983, Vlaisavljević 1981). Sa druge strane, segmentacije predstavlja prvi korak u pripremi vektora obeležja i zato bi potencijalna greška na ovom nivou negativno uticala na njegovo određivanje.

Svako izdvajanje govornih segmenata podrazuzmeva nihovo razlikovanje po definisanom kriterijumu sličnosti, tako da je od samog početka neophodno uvesti parametrizaciju govornih segmenata u cilju diskriminacije signala reči i tišine. Metodološki pristup predviđa da ispitanici izgovaraju reči i rečenice po unapred tačno utvrđenom redosledu što uvodi dodatni determinizam olakšavajući celu proceduru modeliranja i apriori povećavajući tačnost.

4.1. Detektor aktivnog govora u signalu (VAD)

Detekcija govornog signala iz integralnog audio signala ja poznata kao detekcija glasovne aktvnosti (Voice Activity Detection) skraćeno VAD. Ova opšta paradigma podrazumeva kombinaciju širokog spektra metoda za automatsko prepoznavanje govornih sekvenci u integralnom zvučnom signalu. Pošto se radi o prirodno generisanom slučajnom signalu koji je nestastacionaran, preciznije rečeno, kvazistacionaran sa konstantnim frekvencijama tokom krtatkih vremenskih intervala (10 do 25 ms) gde se frekvencije kao i amplitude brzo menjaju od intervala do intervala, nemoguće je napraviti jedinstven i jednostavan sistem za pouzdanu ekstrakciju govornih sekvenci i najčešće se pribegava različitim vrstama klasifikatora zasnovanih na komparaciji sličnosti analiziranih uzoraka signala sa unapred definisanim paradigmama karakterističnih klasa kako u vremenskom tako i u spektralnom i amplitudskom domenu. Većina VAD algoritama datih u literaturi koriste diskriminatorna obeležja govora u

različitim domenima, među kojima su najčešće zastupljena obeležja zasnovana na energiji, obeležja u spektralniom domenu, karakteristike kepstralnog domena i karakteristike dugotrajnih signala.

Osnovna funkcija VAD je diskriminacija govornog signala od signala tišine ili signala ambijentalnog šuma. Blok dijagram VAD detektora je prikazan na Slic 4.1. Svi VAD algoritmi se zasnivaju na smislenoj kombinatorici opštih diskriminativnih karakteristika govora kao što su vremenske varijacije energije, periodičnost i spektar. Zadatak detekcije aktivnog govora nije uopšte jednostavan jer pojava ili porast nivoa buke u okolini smanjuje efektivnost većine klasifikatora. Osnovni princip funkcionisanja VAD uređaja je taj što on izdvaja određene merljive karakteristike iz ulaznog signala i poredi ove vrednosti sa unapred postavljenim vrednostima praga. Ukoliko skup određenih izmerenih vrednosti pređe vrednost praga tada VAD signalizira aktivnost govra (1).

Ambijetalni šum je ozbiljna prepreka za većinu sistema za obradu govora zbog štetnog dejstva šuma na operabilnost sistema. Primer takvih sistema su govorne usluge novih bežičnih komunikacija ili digitalni uređaji koji služe kao pomoćno sredstvo kod oštećenja sluha. Zbog svoje kompleksnosti govor je još velika prepreka za sisteme zasnovane na prepoznavanju govora i govornika. Postoji veliki broj tehnika za eliminaciju šuma kako bi se sistem zaštitio od njenog uticaja, pre svega u cilju poboljšanja performansi. Ove tehnike se oslanjaju na estimaciju karakteristika signala šuma čija ekstrakciju zahteva efikasne VAD detektore govorne aktivnosti. VAD algoritam visoke pouzdanosti i efikasnosti još uvek je nedostupan što predstavlja veliki nerešeni problem i utiče na brojne aplikacije kao što su: pouzdano prevođenje govora (Ramirez i sar., 2003; Karrai i sar. 2003), prenos govora putem interneta u realnom vremenu (Sangvan i sar., 2002) sistemi za uklanjanje šuma i poništavanje eha u oblasti telefonije (Basbug i sar., 2004; Gustafsson i sar., 2002). Čest uzrok otkaza VAD algoritama povećanje nivoa pozadinskog šuma. U poslednje vreme brojni istraživači rade na razvoju različitih strategije za ekstrakciju govora iz zašumljenog signala (Cho i sar., 2001; Gazor i sar., 2003; Armani i sar., 2003). VAD algoritmi su sve više interesantni sa aspekta uticaja njihove efikasnosti na performanse sistema za obradu govora (Boukuin-Jeannes and Faucon, 1995). VAD algoritami su usmerani na detekciju robustnih diskriminatornih obeležja signala buke u cilju uspostavljanja pouzdanih pravila odlučivanja o prirodi analiziranog signala (Li i sar., 2002, Marzinzik i sar., 2002).

Detekcija govora u zašumljenom signalu se koristi kao algoritam za pred procesiranje za skoro sve ostale metode obrade govornog signala. Pri kodiranju govornog signala, VAD se koristi za određivanje vremenskog trenutka kada se prenos govora može prekinuti u cilju

redukcije količine prenesenih podataka. Kod prepoznavanja govora, VAD se koristi za ekstrakciju delova signala koje treba preneti na uređaj za prepoznavanje. Pošto je prepoznavanje govora računarski vrlo zahtevna operacija, ignorisanje negovornih delovi signala štedi procesorsku snagu. Pri poboljšanju govora, kada želimo da umanjimo ili potpuno otklonimo šum iz govornog signala, na osnovu estimacije karakteristika šuma iz delova koji predstavljaju govorne pauze, putem učenja i adaptacije možemo ukloniti šum iz delova zašumlenog govora. Sve ove operacije pre svega se koriste za uštedu resursa. U slučaju procene kvaliteta govora ovakvi algoritmi prave veliku uštedu vremena logopeda.

U višedimenzionalnom prostoru navedenih obeležja klasa sekvenci aktivnog govora i klasa sekvenci signala tišine su lokalizovane u domenima koji nisu potpuno razgraničeni već se delimično presecaju. Diskriminacioni potencijal obeležja predstavlja meru separabilosti između raspodela sekvenci aktivnog govora sekvenci tišine. Teoretski posmatrano, dobro izabrana obelažja trebala bi da onemoguće preklapanje vrednosti između klasa govora i tišine.
Na osnovu histograma aktuelnih klasa se može uočiti jasna linija razgraničenja u domenu nekih obeležja dok se druga obeležja međusobno preklapaju što ukazuje na delimično presecanje domena raspodele klasa. Ova činjenica ukazuje na potrebu za fleksibilnim MLP klasifikatorima koji treba da minimiziraju uticaj presecanja klasa i povećaju tačnost segmentacije. Vektor obeležja za dalju segmentaciju reči na fonemske segmente dobijenih primenom VAD metoda, pored navedenih pet komponenti sadrži još 12 MFCC vrednosti. Na ovaj način se formira trening uzorak zvučnih stimulusa i relevantnih vektora obeležja koji predstavljaju obučavajući uzorak za MLP klasifikator koji treba da automatizuje proceduru segmentacije (Furundžić i sar. 2009, 2012b, 2013b, 2017c). Opšti blok dijagram za ekstrakciju vektora obeležja prikazan je na Slici 4.2 a sama obeležja, za VAD detektor su prikazani po vrsti i nameni u odeljku 4.4.1.

4.2. Klasifikacija zvučnog zapisa na zvučne, bezvučne i segmente tišine

Klasifikacija govornog signala na zvučne (**V**oiced), bezvučne (**U**nvoiced) i segmente tišine (**S**ilence) (VUS) predstavlja osnovnu segmentaciju govornog signala zasnovanu na akustičkim obeležjima, koja je važna za prepoznavanje govora, govornika i analizu govora u smuislu njegovih akustikih kvaliteta. Zvučnost i bezvučnost su inherentne detektabilne karakteristike zvučnih i bezvučnih fonema. Smisao ove klasifikacije je u tome da utvrdi da li je u audio signalu prisutan govorni signal i, ako jeste, da li su u njegovoj produkciji uključuje vibracije glasnica. Vibracije glasnica generišu periodičnu ili kvazi-periodičnu pobudu vokalnog trakta pri produkciji zvučnog govornog signala, dok prelazni i/ili turbulentni šum fonacione

struje predstavljaju aperiodične pobude vokalnog trakta koje produkuju bezvučni govorni signal (bezvučne foneme). Kada su i kvazi-periodična i šumna (tonalna i atonalna) pobuda prisutne istovremeno (mešovite pobude), tada se produkuje govorni signal koji se klasifikuje kao zvučni, jer se vibracija glasnica smatra integrativnim delom govornog akta. Klasifikacija govornog signala na ove tri kategorije se može vršiti pomoću jednog, dva ili više njegovih parametariziranih karakterističnih obeležja kao što su vrednosti energije, autokorelacionih koeficijenata ili broja tranzicija signala između pozitivnih i negativnih vrednosti definisanih za govorni signal tokom kratkotrajnih vremenskih intervala (frejmova). U slučaju korišćenja samo pojedinačnih obeležja pri parametrizaciji a u svrhu klasifikacije signala može se postići samo ograničena tačnost zato što se raspodele bilo kog pojedinačnog parametra najčešće preklapaju između analiziranih kategorija, posebno kada govor nije detektovan u stabilnim uslovima. Klasifikacija na VUS se takođe uobičajeno koristi pri određivanje fundamentalne frekvencije govora Atal i Hanuer (1971). Međutim, pošto vibracije glasnica moraju nužno uvek proizvesti periodični signal, greška u detektovanju fundamentalne frekvencije za zvučni govorni signal, rezultirala bi greškom klasifikacije na VUS kategorije. Atal i Rabiner (1976) su za detekciju govora u zvučnom signalu predložili pristup prepoznavanja oblika, gde su koristili više karakterističnih obeležja govora za klasifikaciju na VUS kategorije. Izveli su klasifikaciju bez određivanja fundamentalne frekvencije koja je u osnovi bila Bajesovski proces odlučivanja, u kojem su pretpostavke o nepoznatoj raspodeli razmatranih obeležja i procene vrednosti parametara tih raspodela bile od suštinskog značaja. U takvim situacijama, pre kreiranja modela klasifikacije, neophodni su veliki trening skupovi podataka za pouzdanu estimaciju relevantnih statističkih parametara. U radu Atal i Rabiner (1976), pretpostavlja se da je raspodela relevantnih obeležja višedimenzionalna Gausovska raspodela. U nameri da ne koristi nesigurne pretpostavke o nepoznatoj statistici distribucija relevantnih obelžja, Siegel (1979) je predložio alternativu pristup za izradu Zvučno - Bezvučne klasifikacije. Primenjena metoda je zasnovana na prepoznavanja uzoraka pomoću linearne funkcije diskriminacije (Duda i Hart 1973). Funkcija diskriminacije predstavlja matricu težina koja je linearno preslikava svaki vektor obeležja na jednu od strana multidimenzionalnog prostora uzoraka razgraničenog pomoću hiperravni. Funkcija diskriminacije i hiperravan se definišu kroz algoritam minimizacije funkcije greške tokom obuke na trening uzorku. Ovo predstavlja neparametarski pristup problemu klasifikacije i rezultati klasifikacije su uporedivi sa onima dobijenim pomoću statističkog parametarskog metoda. Procedura obuke je prilično komplikovana delom zbog toga što diskontinuitet diskriminacione funkcije sprečava direktno analitičko izvođenje algoritma obuke. Siegel i Bessei (1982) su uključili mešovitu (periodično-šumnu) pobudu kao

posebnu, treću kategoriju u klasifikaciji koristeći ovaj neparametarski pristup. Vektor obeležja u parametarskim i neparametarskim metodama sastojao se od odabranih akustičkih obeležja čije raspodele se odlikuju izvesnim stepenom separabilnosti između zvučnih kategorija koje reprezentuju. Na primer, brzina promene znaka ili nultog prelaza (ZCR) je jedan od tipičnih parametara u vektoru obeležja koji ima malu vrednost za zvučni govorni signal i veliku vrednost za bezvučni govor zbog šumne prirode bezvučnog govora. Ovakav vektor obeležja je kao celina deluje prilično veštački kompiliran. Neke komponente u ovom vektoru obeležja su bile čak međusobno dobro korelisane Atal i Rabiner 1976, što nije preporučljivo. Nezadovoljni stepenom tačnosti klasifikacije u prethodnim radovima, Rabiner i Samuer (1977) koriste spektralne distance za poboljšanje klasifikacije. VUS klasifikaciju su izveli na osnovu spektralne blizine između ulaznog i tipičnog primera. Iako je značajno poboljšana tačnost klasifikacije, postupak klasifikacije je i dalje ostao probabilistički proces odlučivanja. Za dizajn pouzdanog klasifikatora neophodan je bio veliki obučavajući uzorak. Kao što ističu autori, "glavni nedostatak metode je potreba za obukom algoritma da bi se dobila skromna spektralna reprezentacija za tri klase signala." Nedostatak efikasnih metoda obuke prediktora je u stvari glavni nedostatak za sve algoritama klasifikacije koje su gore opisani. Primene ovih metoda su, ograničene jer su adaptivne modifikacije klasifikatora često nezaobilazne u praktičnim situacijama. Adaptivno formiranje diskriminatorne funkcije za klasifikaciju uzoraka lako se postiže korišćenjem višeslojnog perceptrona zahvaljujući razvoju u oblasti teorije neuronskih mreža (Rumelhart i sar. 1986). U studiji Qi i Hunt 1992, razvijena je procedura za klasifikaciju VUS primenom MLP. Vektor obeležja za klasifikaciju je nastao kombinacijom cepstralnih koeficijenata i karakteristika talasnog oblika signala. Cepstralni koeficijenti daju potrebne informacije o spektralnim karakteristikama pogodne za klasifikaciju. U cilju poboljšanja razgraničenja među klasama u prostoru obeležja, kada spektralne informacije nisu dovoljne za konstrukciju klasifikatora uključena su dodatna obeležja talasnog oblika signala. Pretpostavka za korišćenje neuronske mreže za VUS klasifikaciju jeste da vremenski ili kontekstualni podaci govora nisu bitni prilikom klasifikacije Fant (1981).

4.3. Priprema baze tipičnih i atipičnih zvučnih stimulusa reči

Proces snimanja niza govornih stimulusa, odnosno reči koje ispitanici izgovaraju ponavljajući za logopedom, produkuje zvučni zapis koji pored govornog signala sadrži i komponentu ambijetalnog šuma nastalog u sobi za snimanjne koja je inkorporirana kako u sam govorni signal tako i u intervale tišine između izgovorenoh reči. Pošto je naš krajnji cilj

pouzdana ocena kvaliteta izgovora fonema sadržanih u govornom delu audio zapisa, odnosno rečima, neophodno je ispuniti sledeće zahteve:

a) Izvršiti izbor vektora karakterističnih obeležja za parametrizaciju govornih signala koji će imati visok diskriminacioni potencijal za VAD detekciju, segmentaciju izdvojenih reči na foneme i razgraničenje tipičnih i atipičnih artikulacija izdvojenih fonema u prostoru njihovih obeležja;

b) Pripremiti bazu fonema posredstvom sledeće dve procedure:

- detekcija i ekstrakcije korisnih delova audio zapisa (VAD), odnosno govornih sekvenci koji sadrže reči iz integralnog audio zapisa,
- segmantacija reči u svrhu izdvajanja sekvenci signala analiziranih fonema;

c) Odabrati pouzdan klasifikator za automatizaciju procesa ocene kvaliteta artikulacije fonema.

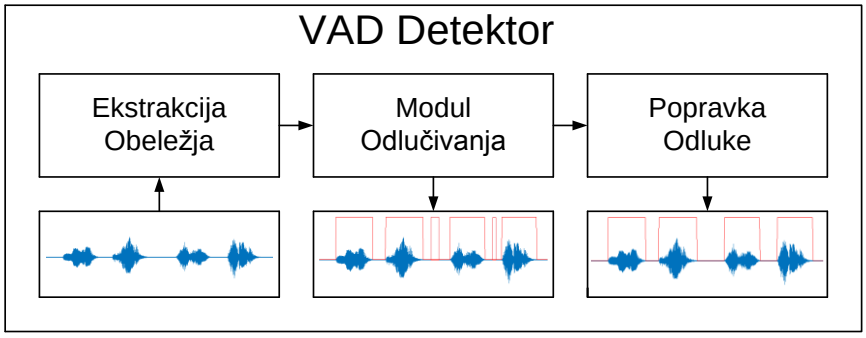
U ovom poglavlju će biti razmatrane zahtevi a) i b) dok će izbor klasifikatora biti razmatran u šestom i sedmom poglavlju. Izdvajanje govornih segmenata je vrlo bitan momenat za dizajn različitih sistema za obradu govornog signala kod prepoznavanja govora, prepoznavanja govornika ili dijagnostiku govorne patologije (Rabiner i Jung, 1993).

Za pouzdanu računarsku segmentaciju reči na konstitutivne foneme a) i diskriminaciju govornog signala i signala tišine b) i neophodna je njihova komparativna karakterizacija i parametrizacija koja se zasniva na poznavanju prirode ovih signala. Kao kriterijum tačnosti modula za ekstrakcije reči i njihovu segmentaciju uzeti su ručno segmentirani uzorci. Za ručnu definiciju granica, odnosno segmentaciju govornih stimulusa, potrebno je veliko ekspertsko iskustvo u analizi govornih segmenata u vremenskom, spektralnom i amplitudskom domenu. U ovom istraživanju smo u incijalnoj fazi procesa automatizovanog određivanja govornih stimulusa koristili manji skup ručno segmentiranih stimulusa govornih sekvenci kao paradigmu trening uzorka koji će poslužiti za obuku induktivnih prediktora i ocenu tačnosti primenjenih algoritama. Za dizajn pouzdanog algoritma računarske segmentacije signala potreban je iterativni postupak postepenog proširenja baze govornih stimulusa od inicijalne do znatno veće finalne veličine. Govor pripada kategoriji stohastičkih procesa što uslovljava metodološki prustup njegovoj analizi. Govor spada u procese sa vremenski promenljivim karakteristikama (nestacionarne) što je posledica različitih inercija i promenljivih rezonantnih frekvencija komponenti složenog vokalnog trakta. Preciznija karakterizacija govora ga definiše kao kvazistacionarni odnosno parcijalno stacionarni proces, što znači da se njegove sekvence u vremenskim intervalima od 5 do 40 ms mogu tretirati kao stacionarne na celoj vremenskoj osi tokom analize njegovih karakteristika. Ovo se pre svega odnosi na konstantnost frekvencija na tim intervalima kao jedne od esencijalnih karakteristika govornog signala uopšte. Opšte

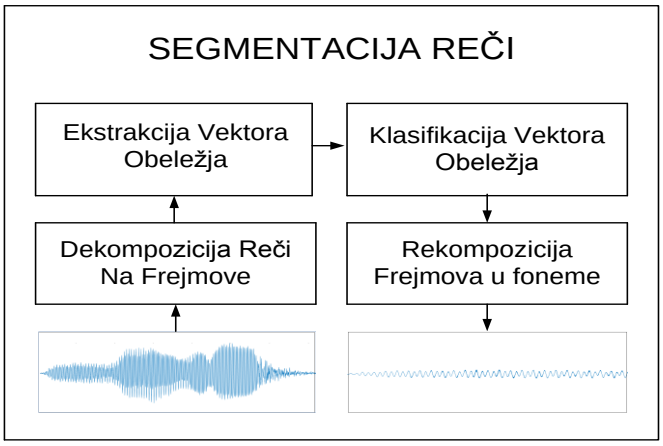
poznata akustička manifestacija govora je njegov talasni oblik odnosno kontinualni zvučni signal promenljive amplitude (intenzitet), frekvencije i definisanog vremenskog trajanja (dužine). Za segmentaciju baze govrnih signala izdvojenih reči zasnovanu na „Pattern Recognition“ pristupu, neophodno je izvršiti ekstrakciju skupa relevantnih obeležja prikazanu u sledećem potpoglavlju.

4.4. Ekstrakcija karakterističnih obeležja govornih signala

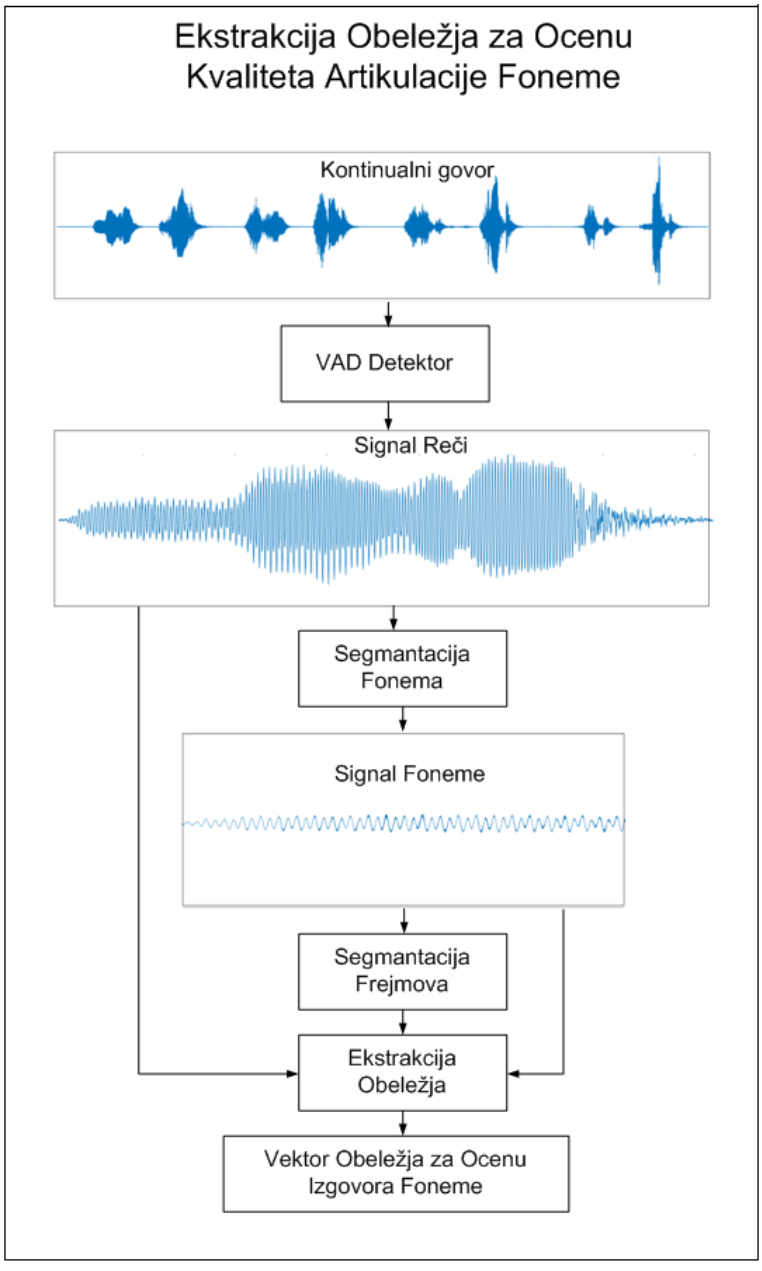
U skladu sa prethodno iznesenim činjenicama o ekstrakciji različitih obeležja govornog signala, predložena su tri skupa relevantnih akustičkih obeležja. Prvi skup se odnosi na modul Ekstrakcije Obeležja VAD detektora (Slika 4.1), drugi skup se odnosi na modul za Ekstrakciju Obeležja iz bloka za segmentaciju signala reči na fonemske segmente (Slika 4.2) a treći skup na modul Vektor Obeležja za Ocenu Izgovora Foneme iz bloka za ekstrakciju ukupnog vektora obeležja za predikciju kvaliteta artikulacije glasova (Slika 4.3). Opšti blok dijagram za ekstrakciju vektora obeležja prikazan je na Slici 4.3 a sama obeležja, njih 7 vrsta, su prikazani po vrsti i nameni u odeljku 4.4.1. Zbog značaja i nešto složenije sheme, na slici 4.4 je posebno prikazan blok dijagram ekstrakcije MFCC vektora obeležja.



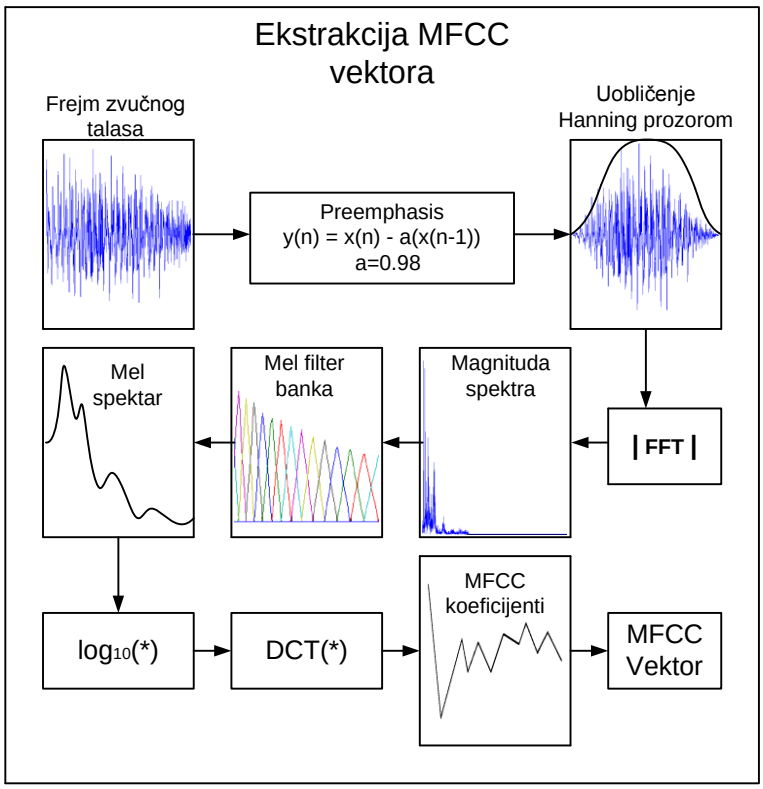
Slika 4.1 Blok dijagram VAD detektora.



4.2 Modul Segmentacije Reči.



Slika 4.3 Algoritam za ekstrakciju vektora obelažja foneme.



Slika 4.4 Shema algoritma za determinaciju MFCC vektora.

Postupak ekstrakcije i segmentacije kao i postupak ocene kvaliteta artikulacije, primenjen u ovom istraživanju, u celini je zasnovan na „Pattern Recognition“ pristupu koji podrazumeva ekstrakciju karakterističnih obeležja i estimaciju parametara induktivnih prediktora.

4.4.1. Izabrana relevantna obeležja

1) Energija segmenata S govornog signala (frejma).

$$E_s = 10 * log_{10}\left(\in + \frac{1}{N}\sum_{n=1}^{N} x(n)^2\right),$$

Gde x predstavlja amplitudu segmenta, $\in = 10^{-5}$ predstavlja malu pozitivnu vrednost koja ima ulogu da spreči pojavu logaritma od negativne vrednosti. Vrednost treba da zadovolji sledeću nejednačinu: $\in \ll MS(s)$, kako nebi uticala na vrednosti energije segmanta E_s, gde $MS(s)$ predstavlja srednju kvardatnu vrednost segmenta S.

2) Broj nultih prelaza u frejmu govora.

$Z_s = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|\, w(n-m)$, Gde je:

$$sgn[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases},$$

Dok, $w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, \ \ 0 \geq n \geq N-1 \end{cases}$, predstavlja prozor dužine N.

3) Normalizovani autokorelacioni koeficijent za jedinično kašnjenja signala.
Autokorelacija je korelacija između vremenske serije ili signala i njegove replike sa kašnjenjm i to u funkciji tog kašnjenja. Glavno pitanje koje se postavlja kod vremenskih serija je kako detektovati postojeću regularnost (periodičnost) maskiranu šumom u smislu prediktabilnosti sledećih vrednosti odbiraka na osnovu vrednosti prethodnih odbiraka. Govorni signal se odlikuje segmentima koji sa visokim stepenom regularnosti (zvučne govorne segmete) i sekvance koje se ponašaju poput nizova slučajnih vrednosti (bezvučni segmenti) sa slabom autokorelacijom. Period dekorelacije kod niza slučajnih vrednosti je vrlo mali jer koeficijenti vrlo brzo padnu na malu vrednost i tu dalje ostaju. Normalizovani autokorelacioni koeficijenti su dobri pokazatelji inherentne korelisanosti signala. Zato smo kao relevantno obeležje govornog signala uzeli koeficijent korelacije za jedinično kašnjenje signala. Autokorelacija sa jediničnim kašnjenjem signal je visoka (blizu +1) za signale čija je energija distribuirana u domenu niskih frekvencija, gde spadaju zvučni govorni signali poput vokala, a niska (blizu 0) za signale sa energetskim koncentratima u domenu visokih frekvencija gde spadaju bezvučni glasivi i neretko signal šuma.
Korišćeni koeficijent je definisan na sledeći način:

$$C_1 = \frac{\sum_{n=1}^{N} x(n)x(n-1)}{\sqrt{(\sum_{n=1}^{N} x^2(n))(\sum_{n=0}^{N-1} x^2(n))}}$$

U gornjem izrazu $x(n)$ predstavlja n-ti odbirak zvučnog signala.

4) Koeficijenti LPC prediktora.
Linearno Prediktivno Codiranje (LPC) je standardna tehnika u za reprezentaciju govornih signala. Linearna predikcija je matematička operacija obrade signala u kojoj su vrednosti diskretne vremenske serije determinisane kao linearna funkcija vrednosti prethodnih uzoraka. U digitalnoj obradi signala, linearna prediokcija se naziva linearno prediktivno kodiranje (LPC) i zato spada u domen teorije filtera.
$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$, gde $\hat{x}(n)$ predstavlja predviđenu vrednost signala, $x(n-i)$ predstavlja neku prethodnu observaciju dok a_i predstavlja prediktorske koeficijente. Ovi prediktorski sistemi po pravilu generišu grešku predikcije koja se manifestuje razlikom između observirane $x(n)$ i estimirane $\hat{x}(n)$ vrednosrti signala: $e(n) = x(n) - \hat{x}(n)$. U slučaju višedimenzionalnih signala greška je definisana u formi norme vektora: $e(n) = \|x(n) -$

$\hat{x}(n)\|$. LPC analiza prevodi svaki frejm u skup vektora dimenzije od 12 do 16. U prikazanom istraživanju jedna od korišćenih procedura za ekstrakciju obeležja signala je LPC analiza. Optimizacija parametara u cilju minimizacije greške predikcije se izvodi na dva standardna pristupa: RMS kriterijum ili autokorelacioni kriterijum i Levinson Durbinov rekurzivni algoritam. U našem istraživanju smo koristili samo prvi od izračunatih LPC koeficijenata, kao nosioca važnih informacija za svaki analizirani frejm.

5) Energija greške linearne predikcije.

Vrednost energije signala greške predikcije $e(n) = x(n) - \hat{x}(n)$ takođe spada u karakteristična obeležja pri obradi govornog signala, i kao takva njena normalizovana logaritamska vrednost je uključena u istraživanje: $E_{p=} E_s - 10 * log_{10}\left(\in + \frac{1}{N}\sum_{n=1}^{N} \hat{x}(n)^2\right)$, gde, $\in = 10^{-5}$ predstavlja malu pozitivnu vrednost koja ima ulogu da spreči pojavu logaritma od negativnih vrednosti.

6) MFCC koeficijenti

Mel frekvencijski kepstralni koeficijenti (MFCC) se poslednjih decenija često koriste u domenu obrade govornih signala (Rabiner i Jung, 1993). Ovi koeficijenati se tretiraju kao jedan od standarda za ekstrakciju obeležja govora. Motivacije za njihovu primenu leži u načinu na koji auditorni sistem čoveka izdvaja karakteristična obeležja govora. Kohlea, kao najbitnija organela u unutrašnjem uvu, ima funkciju banke filtara koja iz zvučnih signala ekastrahuje bitna obeležja. Korišćena Banka filtara je skup filtara propusnog opsega koji deli ulazni signal u više komponenti od kojih svaka sadrži odabrani opseg frekvencija originala. Neke frekvancije iz originala mogu biti izostavljene kao što su u našem slučaju izostavljene frekvencije niže od 133 Hz jer ne nose važne informaciju govornog signala već šum, dok su neke druge, naročito visoke ferekvencije, naglašene i uključene u aktuelni skup za analizu kao reprezentativne. MFCC koeficijenti sadrže informacije o raspodeli energetskih koncentrata na celom frekventnom opsegu i informacije o kretanju formanata po celoj vremenskoj osi, tako da praktično pokrivaju široku lepezu spektralnih karakteristika signala među kojima se nalaze formantske karakteristike važne za diskriminaciju glasova i karakteristike važne za finu diskriminaciju kvaliteta artikulacije u okviru istih grupa glasova. MFCC koeficijenti se izračunavaju primenom banke od oko 40 filtara propusnog opsega na apsolutnim vrednostima furijeovog spektara uz primenu sledećeg algoritma:

$c(n) = DCT\{\log_{10}|X(k)|\}$,

gde je $X = fft(x)$ a x je signal, frejm dužine 25 ms uobličen na način gore opisan, $c(n)$ je $n - ti$ kepstralni koeficijent dok je $IDFT$ je simbol za inverznu diskretnu furijeovu transformaciju. Pri korišćenju MFCC koeficijenata treba obratiti pažnju na proces dekompozicije signala foneme na niz frajmova relativno male dužine. Naime, Prvih 12 (dvanaest) *MFCC* vrednosti su uzeti iz svakog frejma a broj frejmova nije bio konstantan u svim fonemima raznih ispitanika već je varirao od 7 do 32 za fonem Š na primer. Dakle MFCC karakteristika za svaki fonem je u stvari matrica dimenzije Z x 12, gde je Z broj frejmova u fonemu. Pri obuci klasifikatora, svaki frejm dobija odgovarajuću identičnu vrednost željenog izlaza (ocene), koja odgovara integralnom signalu fonema, zato pri tumačenju rezultata na test uzorku treba izvršiti usrenjavanje izlaznih vrednosti za sve frejmove određenog fonema kako bi se dobila ocena kvaliteta artikulacije fonema.

7) Ostala akustička obeležja foneme

Trajanje, kao bitna akustička karakteristika govornog signala ima izuzetno veliku važnost pri analizi kvaliteta izgovora fonema. Zato je formirana grupa od tri atributa datih u formi realne vrednosti koje se odnose na dužine talasnog oblika govornog signala a sastoji se od:

nw - broja semplova talasnog signala koji se odnosi na celu reč koja u početnoj poziciji sadrži aktuelne foneme,

nph - broj odbiraka signala koji se odnose na aktuelne foneme,

$nq = nph/nw$ - odnos prethodnih dužina.

Srednja energija frejma foneme (Ef),

Srednja energija frejma odgovarajuće reči (Er),

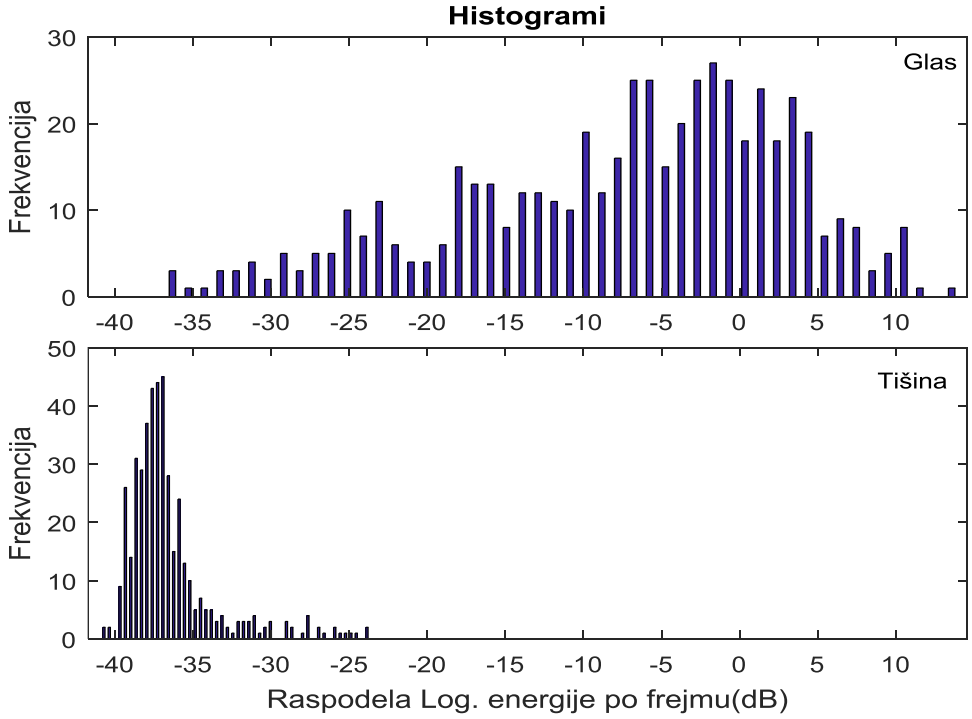
Ukupna energija fonema (Efu),

Ukupna energija aktuelne reči (Eru).

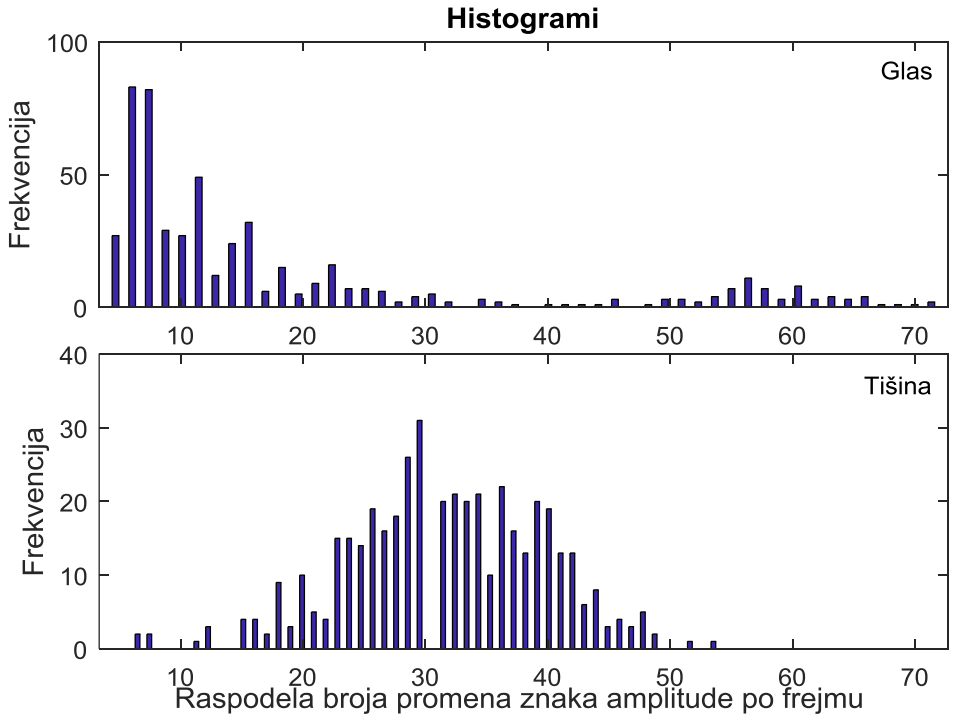
Ovde smo naveli sva reprezentativna obeležja zvučnih stimulusa korišćena u različitim kombinacijama za rešavanje raznih problema tokom istraživanja (VAD, Segmentacija i Ocena kvaliteta artikulacije). Zato je izvršeno grupisanje ovih obeležja u zavisnosti od situacije u kojoj se primenjuju i dato u odeljku 4.4.2.

Grafička prezentacija distribucija navedenih obeležja, u formi histograma, dobijenih na kontinuiranom govornom signalu (aktivni govor + tišina) GAT testa koji sadrži 30 reči data je na slikama 4.5 do 4.10. Na svim slikama gornji deo slike odgovara signalu glasa a donji signalu tišine. Suštinska vrednost ekstrahovanih obeležja je sadržana u njihovom diskriminacionom potencijalu koji se meri stepenom separabilnosti različitih klasa ostvarenim na osnovu raspodela vrednosti tog obeležja za pomenute klase. Dakle, stepen preklapanja

raspodela posmatranog obeležja kod suprotstavljenih klasa obrnuto je proporcionalan njegovom diskriminacionom potencijalu. Svako od prikazanih obeležja ima evidentan diskriminacioni potencijal koji je kod nekih obeležja vrlo naglašen dok je kod drugih manji.



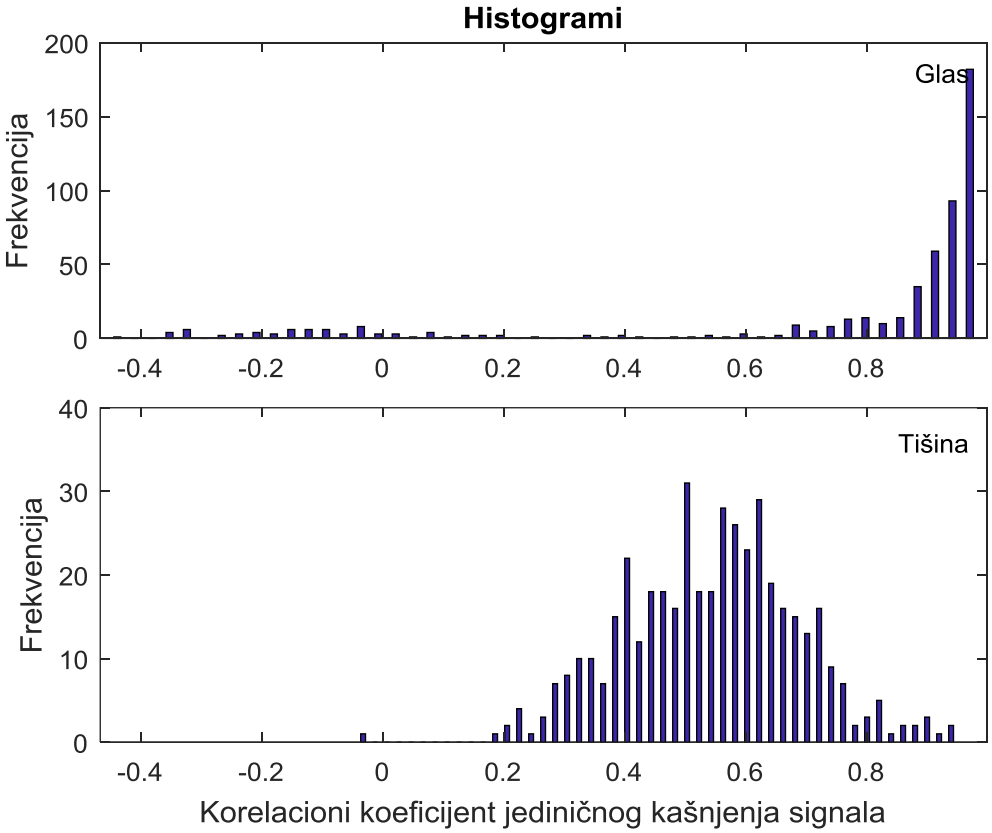
Slika 4.5 Raspodela logaritamskih vrednosti energije frejmova za signal glasa i tišine.



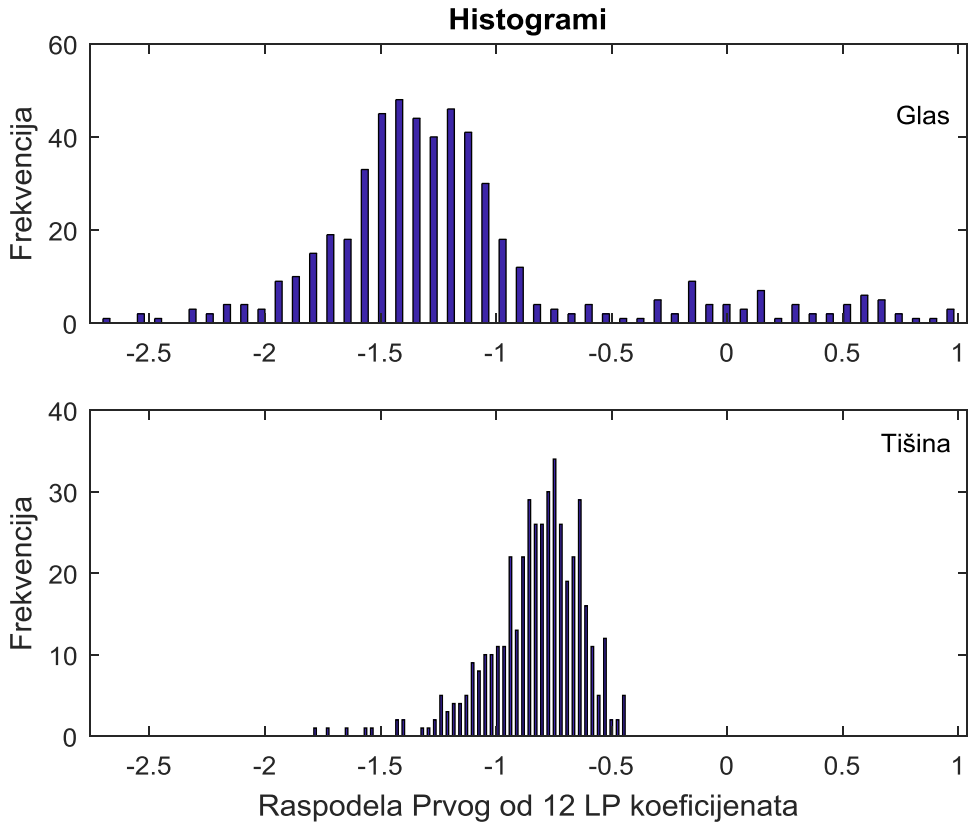
Slika 4.6 Raspodela vrednosti nultih prelaza kod frejmova za signal glasa i tišine.

Ako posmatramo primer raspodele vrednosti logaritma energije frejmovaa na Sl. 4.4, lako uočavamo da su energije glasa široko raspodeljene po srednjem i desnom delu prostora dok su energije tišine pomerene i usko koncentrisane u krajnjem levom delu.

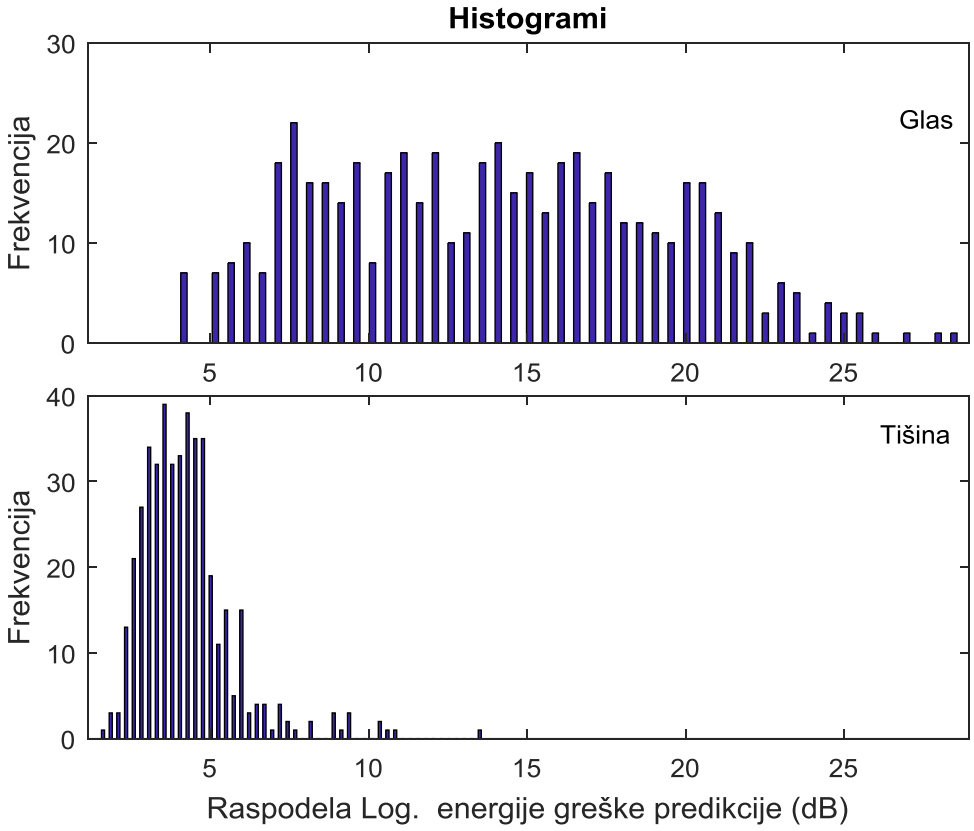
Dakle, postoji izvesno preklapanje ali je ono za ovakvu vrstu signala malo što znači da ovo obeležje daje veliki doprinos pouzdanom razgraničenju između frejmova glasa i tišine. Koncentracija energija tišine u levom uskom pojasu govori o maloj energiji signala tišine kojim se odlikuje tiha soba za snimanje. Energija glasa široko raspodeljena po celom prostoru vrednosti obeležja govori o spektru različitih enargije koje potiču od različitih glasova. Zvučni glasovi su koncentrisani u desnom delu prostora dok su bezvučni koncentrisani u prostoru manjih energija (levo). Ovo obeležje, iako ima veliki potencijal, samo nije dovoljno za pouzdano razgraničenje glasa i tišine u zvučnom signalu sa pozadinskim šumom. Na osnovu histograma ostalih obeležja možemo zaključiti da imamo relevantan skup informacija za rešenje problema VAD detekcije, segmentacije i ocene kvalteta artikulacije (klasifikacije) glasova srpskog jezika. Višedimenzionalni prostor obeležja povećava mogućnost razgraničenja među suprotstavljenim klasama signala ali zahteva fleksibilne klasifikatore za formiranje složene granične hiperpovrši.



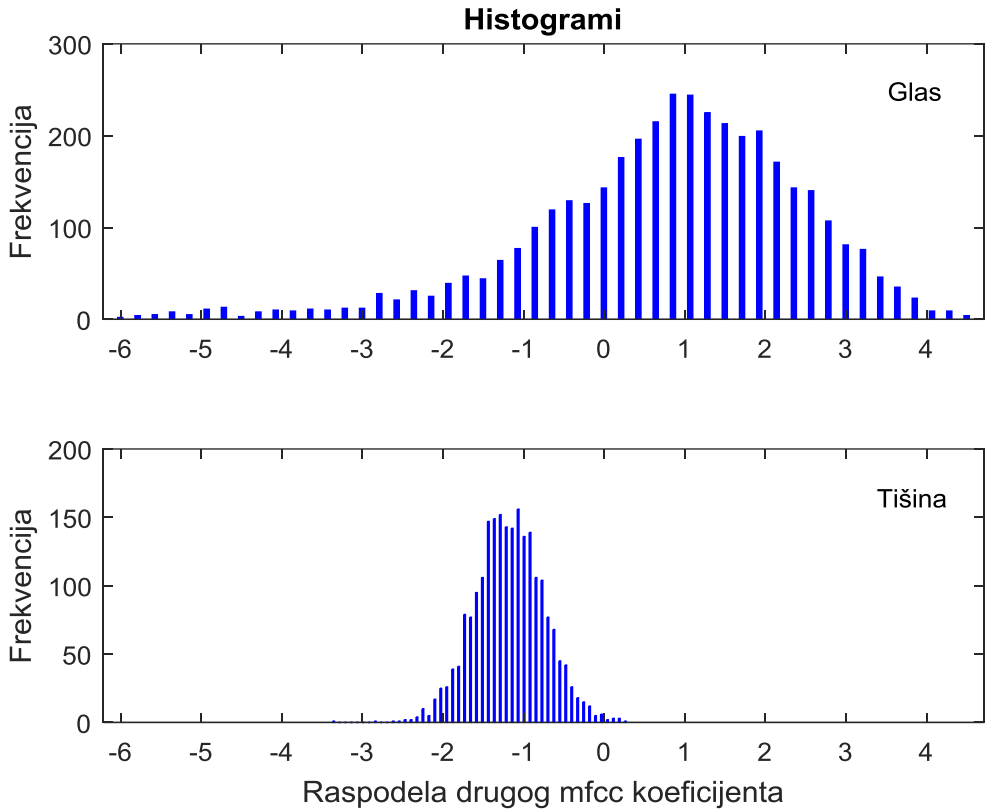
Slika 4.7 Raspodela vrednosti drugog autokorelacionog koeficijenta po frejmovima za signal glasa i tišine.



Slika 4.8 Raspodela vrednosti prvog od 12 koef. Lin. Prediktora za frejmove glasa i tišine.



Slika 4.9 Raspodela Log. vrednosti energije greške predikcije za frejmove glasa i tišine.



Slika 4.10 Raspodela vrednosti drugog MFCC koeficijenta za frejmove signala glasa i tišine.

4.4.2. Grupe karakterističnih obeležja

Obeležja za VAD

Za potrebe detekcije aktivnog govora u integralnom zvučnom signalu (VAD) korišćene su vrednosti sledećih obeležja:

Logaritam Energije Frejma (obeležje 1),

Broj nultih prelaza u frejmu (obeležje 2),

Koeficijent autokorelacije za jedinično kašnjenja signala Frejma (obeležje 3),

Prvi LPC koeficijent (obeležje 4),

Logaritam Energija greške linearne predikcije (obeležje 5).

Dužina ovog vektora obeležja je 5.

Obeležja za Segmentaciju

Za potrebe dekompozicije reči foneme (Segmentacija) korišćene su vrednosti sledećih obeležja:

Logaritam Energije Frejma (obeležje 1),

Broj nultih prelaza u frejmu (obeležje 2),
Koeficijent autokorelacije za jedinično kašnjenja signala Frejma (obeležje 3),
Prvi LPC koeficijent (obeležje 4),
Logaritam Energija greške linearne predikcije (obeležje 5),
Vrednosti niza od dvanest MFCC koeficijenata sledećih indeksa (2 do 13).
Dužina ovog vektora obeležja je 17.

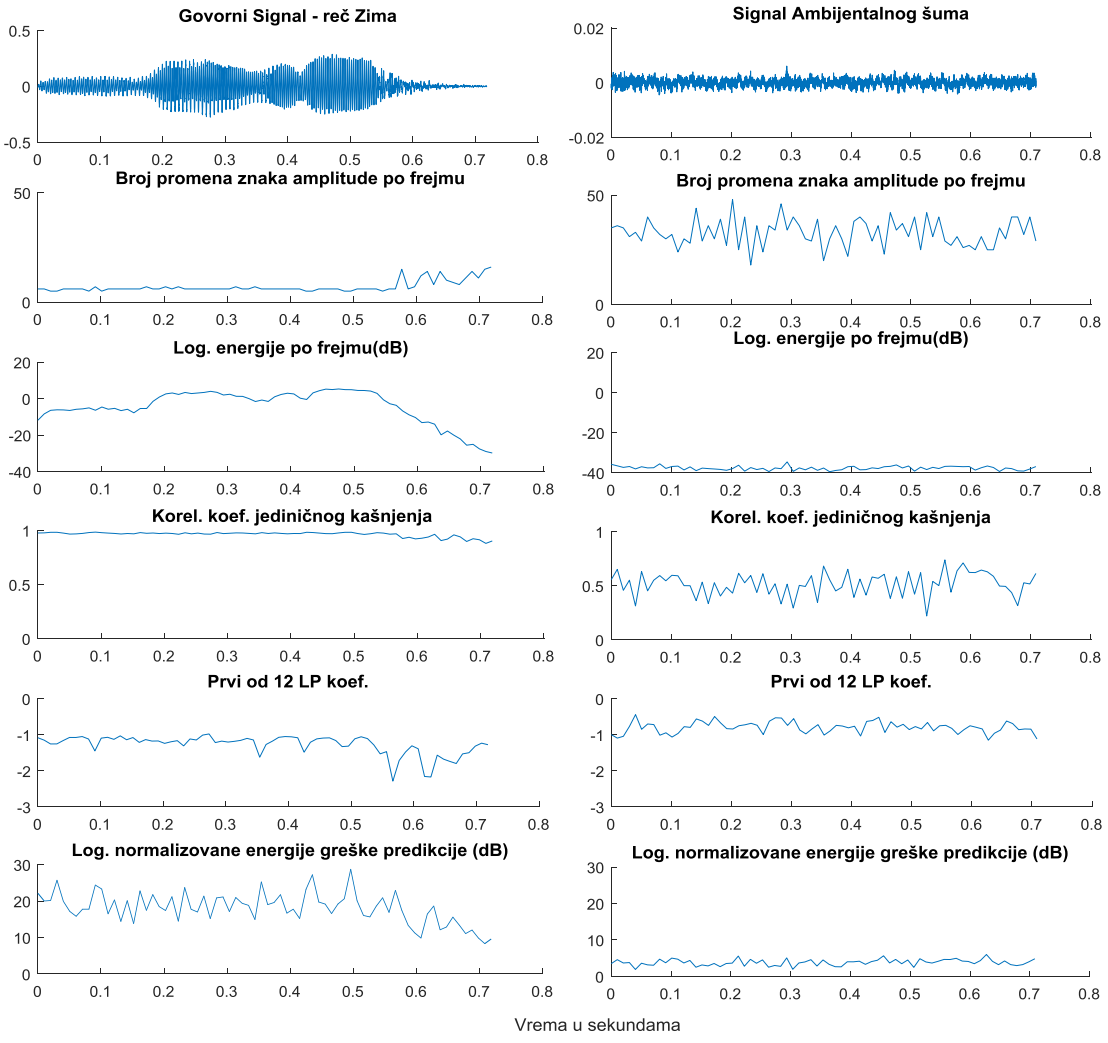
Obeležja za Klasifikaciju Kvaliteta Artikulacije Glasova

Za potrebe klasifikacije foneme po kvalitetu artikulacije (Klasifikacija) korišćene su vrednosti sledećih obeležja:
nw - broja semplova talasnog signala koji se odnosi na reč,
nph - broj odbiraka signala koji se odnose na aktuelne foneme,
$nq = nph/nw$ - odnos prethodnih dužina.
Srednja energija frejma foneme (Ef),
Srednja energija frejma odgovarajuće reči (Er),
Ukupna energija fonema (Efu),
Ukupna energija aktuelne reči (Eru).
Vrednosti niza od dvanest MFCC koeficijenata sledećih indeksa (2 do 13).
Dužina ovog vektora obeležja je 19.

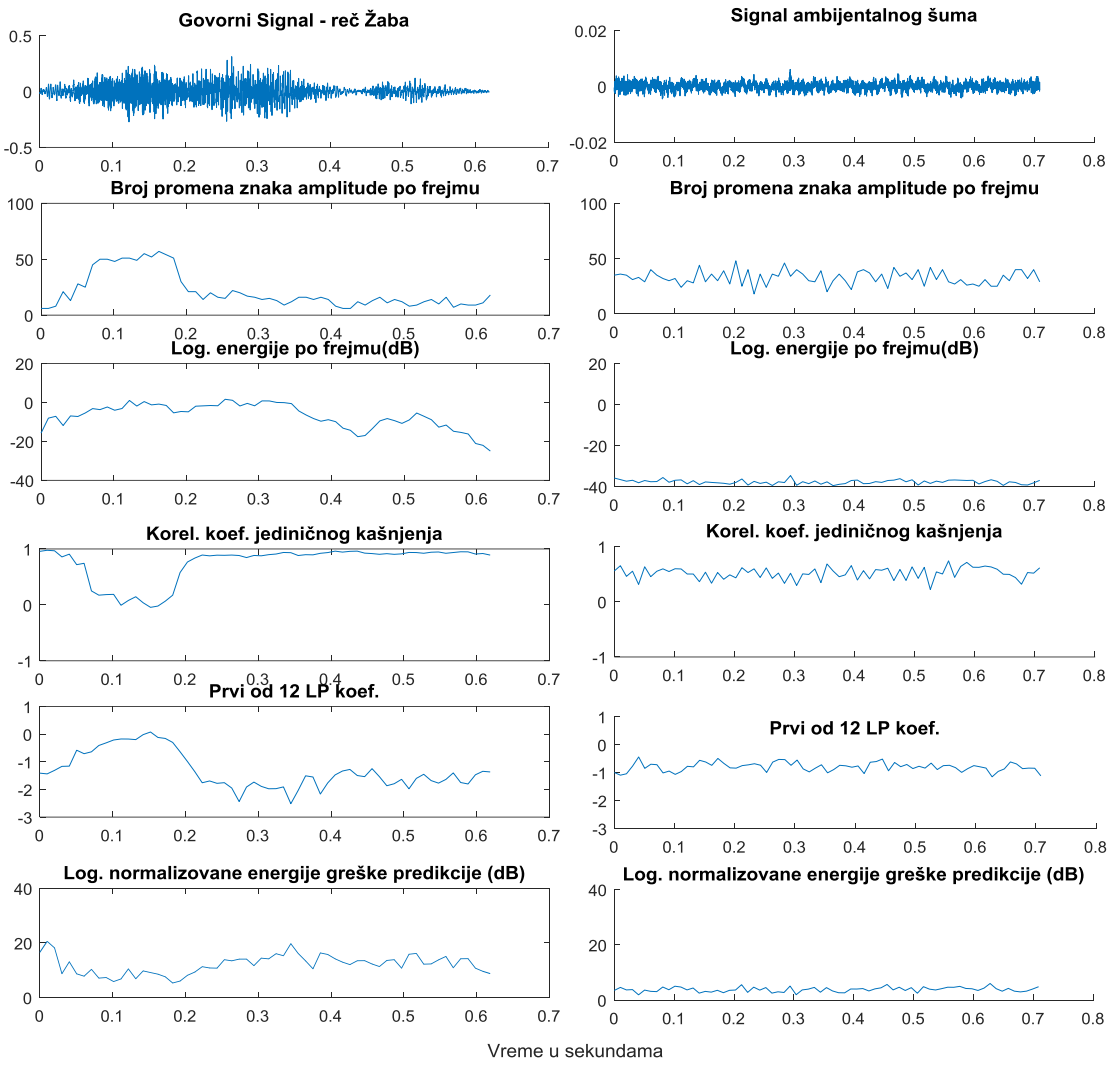
Glavni problem sistema za računarsku procenu kvaliteta artikulacije je adekvatan izbor akustičkih parametara koji su dovoljno pouzdani za finu diskriminaciju unutar iste kategorije fonema i izbor klasifikatora koji je u stanju da u višedimenzionalnom prostoru atributa definiše graničnu površ između klasa različitih kvaliteta. Rezultati nastali primenom diskutovanih obeležja u algoritmima za računarsku ocenu kvaliteta artikulacije prikazani su u sedmom poglavlju.

Na slikama 4.11 do 4.13 prikazani su signali karakterističnih reči i signali tišine kao i vrednosti za pet odabranih obeležja za ekstrakciju signala reči iz kontinuiranog zašumljenog zvučnog signala. Signali reči i tišine su prikazani u talasnom obliku. Vrednosti obeležja koje se odnose na frejmove su prikazane na ordinati po celoj vremenskoj osi od početka do kraja signala reči i tišine. Uočavaju se razlike u raspodeli vrednosti obeležja za signale reči i tišine u skladu sa njihovim akustičkim karakteristikama. Ove reči sadrže u inicijalnoj poziciji frikative s, š, z i ž, kao foneme sa velikom frekvencijom odstupanja u kvalitetu izgovora i kao takvi su predmet daljih analiza i računarske procene kvaliteta artikulacije. Na slici 4.11 je očigledna homogenost svih odabranih obeležja po celoj dužini reči Zima isključujući mali završni deo signala. Svi

glasovi ove reči su zvučni, karakterisani korelisanošću i prediktabilnošću. Broj prelazaka nule ordinate je gotovo nepromenjiv tokom izgovora svih fonema. Energija ima visoku i malo variabilnu vrednost do samog završetka glasa a. Prvi koeficijent koralacije C_1 signala po celoj dužini reči signlu reči ima vrednost vrlo blizu 1, što ukazuje na visoku vrednost autokorelacije. Prvi koeficijent Linearne predikcije ima konstantnu vrednost sa manjim povremenim oscilacijama na mestima koartikulacije i završetku reči. Logaritam energije greške predikcije kreće se u relativno malom pojasu oko srednje vrednosti. U poređenju sa vrednostima signala tišine, odnosno ambijentalnog šuma tihe sobe, uočava se velika razlika kako u vrednostima tako i u prirodi signala obeležja. Ovo je indikator velikog diskriminacionog potencijala odabtranih obeležja za razdvajanje najvećeg dela signala reči (od početka foneme z do blizu kraja foneme a) i signala tišine.

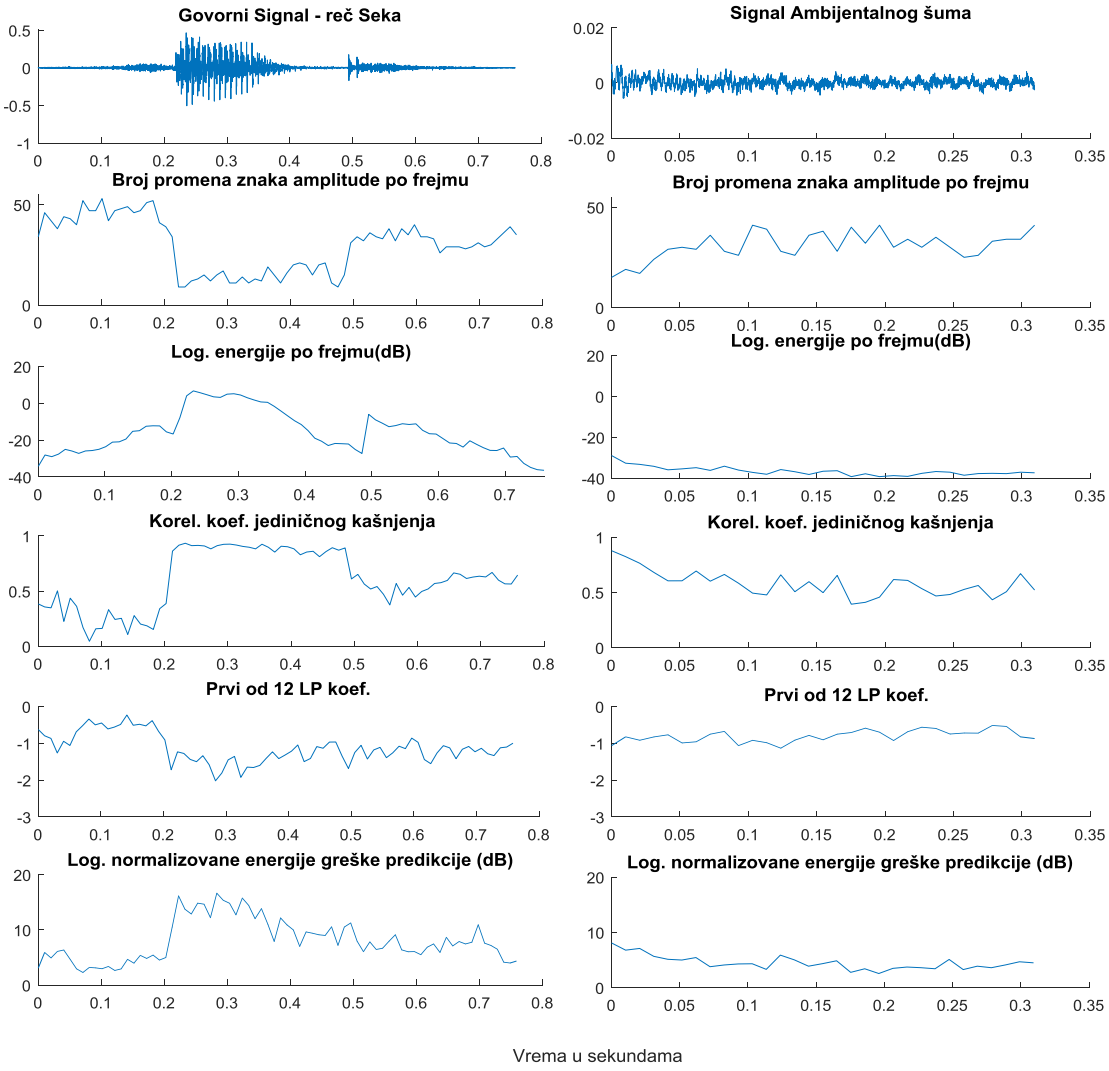


Slika 4.11 Uporedna slika VAD obeleža signala reči Zima i Tišine.

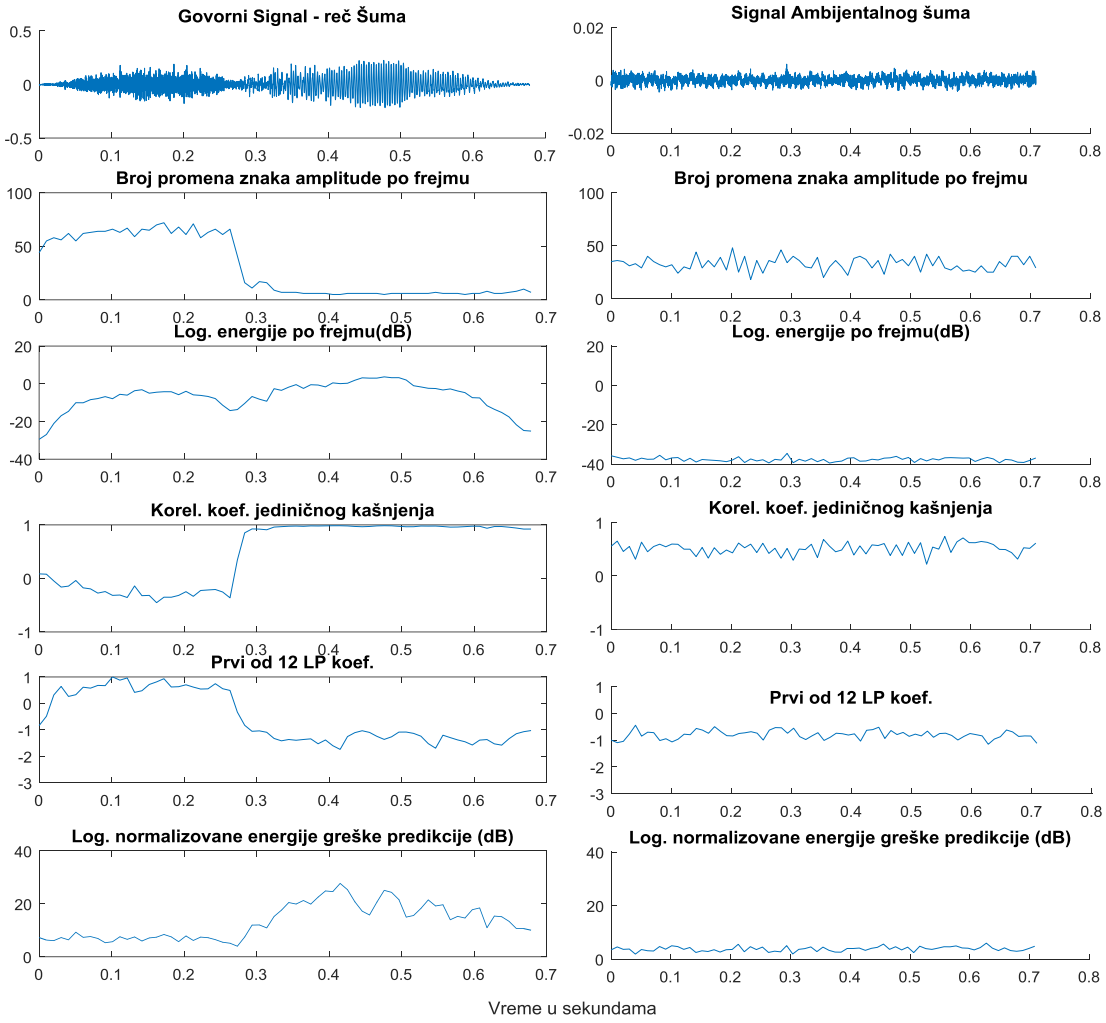


Slika 4.12 Uporedna slika VAD obeleža signale reči Žaba i Tišine.

Treba potsetiti da se akustička slika izgovorenih glasova, koju logopedi formiraju na osnovu percepcije i analize velikog broja odstupanja od tipičnog izgovora, zasniva na kombinatorici i varijabilnosti osnovnih inherentnih akustičkih karakteristika čije su manifestacije upravo obeležja koja mi koristimo. Sa obzirom na činjenicu da je do sada definisano u preko sto (100) različitih odstupanja u kvalitetu izgovora glasova, modeliranje svakog od njih pojedinačno ne bi dalo rezultate u dogledno vreme. To je jadan od glavnih razloga naše odluke da na osnovu opštih akustičkih karakteristika glasova koji su u GAT testu predstavljeni ocenom, obučimo induktivne prediktore da klasifikuju glasove u dve osnovne kategorije kvaliteta, tipičan (0) i atipičan (1).



Slika 4.13 Uporedna slika VAD obeležja za signale reči Seka i Tišine.



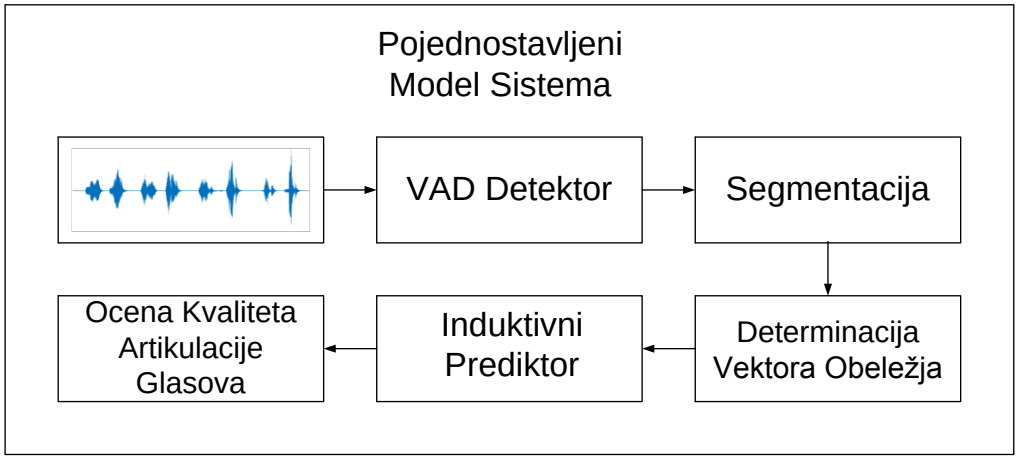
Slika 4.14 Uporedna slika VAD vrednosti obeležja za signale reči Šuma i Tišine.

4.5. Eksperimentalni rezultati segmantacije i VAD detekcije

U ovom potpoglavlju su prikazani eksperimentalni rezultati istraživanja koja se odnose na najvažnije konstitutivne elemanate sistema za ocenu kvaliteta izgovora čija je uprošćena shema data na Slici 4.15. Ti rezultati se odnose na ocenu tačnosti VAD modula za ekstrakciju reči iz kontinuiranog signala, modula za segmentaciju izdvojenih reči na foneme. Algoritam modula za predikciju kvaliteta artikulacije prikazan je na Slici 4.16. Algoritmi svih pomenutih modula su zasnovani na četiri induktivna prediktora koji su prikazani pojedinačno u poglavlju 6 a čija uloga je od suštinskog značaja za predloženi sistem ocene kvaliteta artikulacije. Ovi prediktori su prikazani kao nezavisni sastavni elemanti modula za ocenu kvaliteta artikulacije (Slika 4. 15). Iz tog razloga su rezultati sortirani po modulima i po tipovima prediktora. Komparacijom

rezultata na osnovu korišćenih prediktora dolazimo do rešenja postavljenog zadatka u smislu izbora algoritma sa najmanjom greškom predikcije.

Procedura za izbor relevantnih obeležja segmanata izgovornih glasova u cilju njihove parametrizacije u formi vektora obeležja ima tri važna dela. Ova procedura je prikazana na Slici 4.2, i podrazumeva a) primenu VAD algoritma za ekstrakciju signala reči, b) segmentaciju signala reči na foneme, podelu na subfonemske jedinice (frejmove) i c) proračun vektora obeležja. U svim fazama predobrade signala, kao inicijalne stimuluse koristili smo ekspertski ručno segmentirane sekvance kao osnovni trening uzorak skromnih dimanzija neophodan za inicijalnu obuku prediktora. Svi ostali koraci predstavljaju automatizovane računarske procedure.



Slika 4.15 Uprošćena shema modela za ocenu kvaliteta izgovora glasova.

4.5.1. Rezultati ekstrakcije reči iz govornog signala primenom VAD algoritma

Rezultati ocena tačnosti ekstrakcije reči primenom VAD algoritma za diskriminaciju aktivnog govornog signala i signala tišine, zasnovana na „Pattern recognition“ pristupu (Atal i Rabiner 1976), prikazani su u Tabeli 4.1. VAD koristi četiri različite metode klsaifikatora za razgraničenje govornih stimulusa i stimulusa tišine na osnovu istog skupa diskriminativnih obeležja iz prve grupe.

Kao objektivne (tačne) smatrane su srednje ocene, za svaku izdvojenu reč i pauzu, grupe od 5 logopeda dobijene na osnovu njihovog većinskog odlučivanja. Kada se govori o pojmu objektivizacije procene kvaliteta artikulacije misli se na računarsdko modeliranje logopedske većinske odluke o oceni.

Frekvancija odabiranja analiziranih govornih signala bila je 11.025 kHz dok je dužina frejmova bila 10 mm, što je zaokruženo na 110 odbiraka po frejmu. Tačnu poziciju aktuelnih granica su

odredili trenirani logopedi sa velikim iskustvom u oblasti analize govora u vremenskom i spektralnom domenu kao i auditivne percepcije karakterističnih obeležja. Uveden je kriterijum tačnost predikcije po kom je dopuštena greška pri određivanju granice fonema bila u dužini od 30 ms, odnosno 3 frejma ili 330 odbiraka signala. Ovo je u skladu sa činjenicom da razlike među logopedima pri ručnom određivanju ovih granica variraju u opsegu dužina od 10ms do 30ms. Pošto se VAD algoritmi zasnivaju na metodi prepoznavanja oblika, za obuku prediktora korišteni su kontinuirani govorni signali iz GAT testa (Prilog I) produkovani od strane 13 ispitanika što znači da imamo bazu od 390 segmenata aktivnog govora (reči) i isto toliko segmenata tišine. Broj prikazanih reči Seka, Šuma, Zima i Žaba u obučavajuém uzorku je iznosio po 13 govornih sekvenci i 13 sekvenci tišine za svaku reč. Teastiranje performansi obučenih prediktora za određivenje granica reči iz kontinuiranog govornog signala iz GAT testa, izvedeno je na test uzorku produkovanom od strane 137 ispitanika i rezultati prikazani u Tabeli 4.1 odnose se na ovaj test uzorak.

Ispred i iza svih reči nalazi se signal tišine koji ima relativno stabilne karakteristike tokom celog perioda snimanja, kada je jedan ispitanik u proceduri. Međutim, uslovi se menjaju tokom sledećih ispitivanja što izaziva varijacije kvaliteta signala tišine i samim tim otežava proces VAD detekcije. Sa druge strane, karakteristike signala reči su vrlo variabilne unutar korpusa reči što utiče na diskriminativni potencijal tih reči u odnosu na signal tišine. Naime, najveći uticaj na ovaj potencijal imaju granični fonemi reči u smislu sličnosti njihovh akustičkih karakteristika sa osobinama signala tišine (Slike 4.1.1 do 4.14). Sa tog aspekta reči koje počinju i završavaju se zvučnim fonemom imaju najveći očekivani nivo separabilnosti. Reči koje počinju i završavaju se bezvučnim fonemom imaju znatno manje šanse za razgraničenje sa signalom tišine. Reči koje samo sa jedne strane imaju zvučni fonem imaju srednje izglede za uspešno razgraničenje sa tišinom. Signal zvučnih fonema (Z i Ž) se odlikuje regularnošću i harmoničnošću što implicira visok nivo energije, mali broj nultih prelaza, veliki korelacioni koeficijent i malu grešku predikcije. Bezvučni fonemi pokazuju veliku razliku inherentnih akustičkih kvaliteta u odnosu na zvučne (manja vrednost energije, veći broj nultih prelaza, manji korelacioni koeficijenti i veća greška predikcije), ali su neretko u monogome slični sa signalom tišine pa je očekivana greška njihovog razgraničenja sa signalom tišine znatno veća u odnosu na zvučne foneme. Ovi faktori nisu jedini uticajni faktori na tačnost separacije aktivnog govora jer pored njih mogu delovati stepen ravnomernosti raspodele trening skupova i tipovi primenjenih prediktora. Tačnost separacije aktivnog govora (reči) od signala tišine kao nosilaca vektora obeležja tipičnosti izgovora fonema je vrlo bitan korak predobrade jer dobar obučavajući uzorak mnogo utiče na tačnost predikcije, pa se u tu svrhu razvijaju alati za

naknadnu proveru graničnih tačaka. Prikaz rezultata ekstrakcije izgovornih reči dat je u Tabeli 4.1. Očigledna je razlika u greškama predikcije različitih prediktora. Razlog prednosti Ansambla MLP u odnosu na ostale prediktore leži u većoj fleksibilnosti i robustnosti u odnosu na druge prediktore. SOM prediktor koji tokom obuke ne dobija informaciju o vrednosti željenog izlaznog signala, vrši klastrovanje na dva klastera tražeći pri obuci centre okupljanja instanci u višedimenzionalnom prostoru obeležja zvučnih stimulusa i klasifikuje test instance u zavisniosti od rastojanja u prostoru obeležja od tih centara okupljanja. Ovaj prediktor pokazuje najmanju tačnost predikcije jer ne dobija precizne instrukcije pri formiranju granične hiperpovrši u prostoru obeležja. Greška u predikciji granica segmenata reči Zima i Žaba sa zvučnim frikativima (z i ž) u inicijalnoj poziciji respektivno je primetno manja u odnosu na grešku predikcije granica reči Seka i Šuma sa bezvučnim frikativima (s i š) u inicijalnoj poziciji. To je posledica veće sličnosti između akustičkih karakteristika bezvučnih fonema isignala tišine. Sa druge strane, sve analizirane reči u finalnoj poziciji imaju isti vokal (a) koji se odlikuje definisanim formantnim strukturama iako nešto oslabljenim zbog završetka reči. Pošto se vokal a nalazi u svim pomenutim rečima njegov uticaj na tačnost segmentacije reči je zanemariv. Ovaj stav ima očiglednu grafičku potvrdu na spektrogramima prikazanim na slikama 4.19 - 4.21.

Tabela 4.1 Greška segmentacije VAD algoritma.

VAD Postupak	Ulaz-Kontinuirani govorni signal iz GAT testa			
	Greška pri detekciji aktivnog govornog signala [%]			
	Seka	Šuma	Zima	Žaba
KNN Ansambl	8.0292	10.2190	6.5693	7.2993
Bayes	15.3285	18.2482	13.8686	16.7883
SOM	24.0876	22.6277	21.8978	19.7080
MLP Ansambl	7.2993	8.7591	5.1095	6.5693

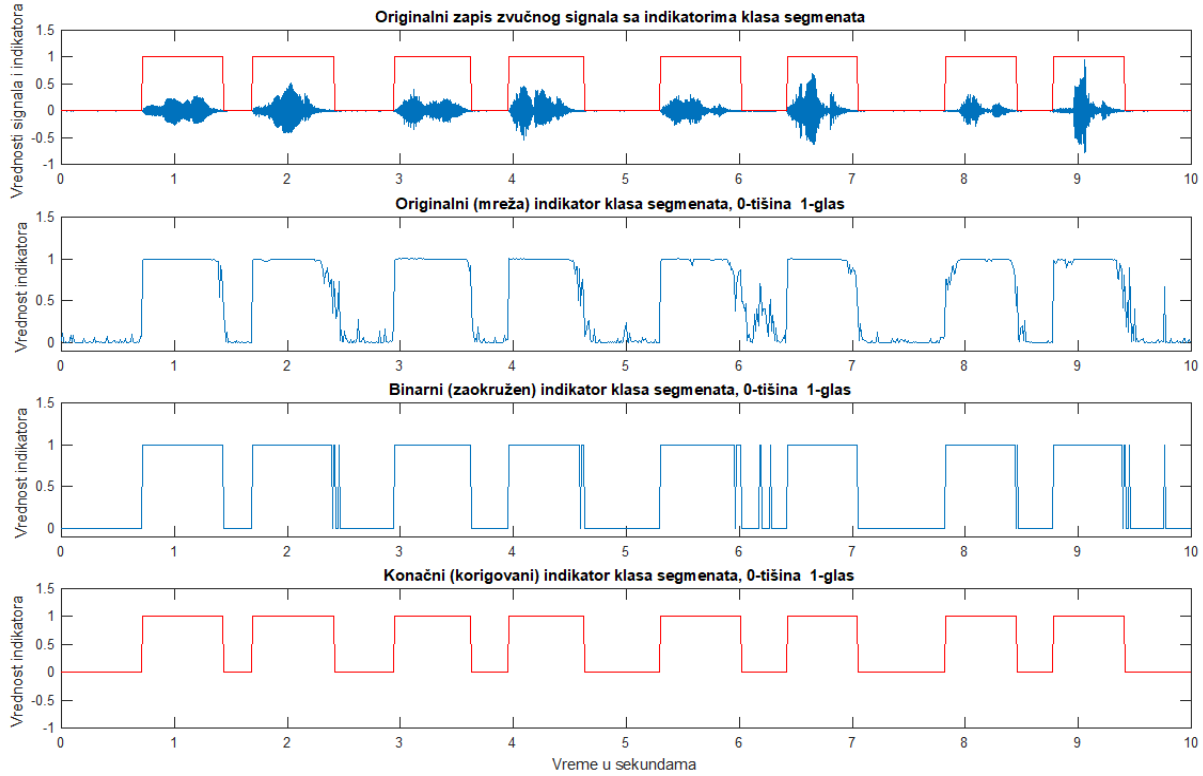
4.5.1.1 Grafički primeri ekstrakcije reči iz govornog signala primenom VAD algoritma

Grafički prikaz performandi rada VAD detektora dat je na Slikama 4.16, 4.17 i 4.18.

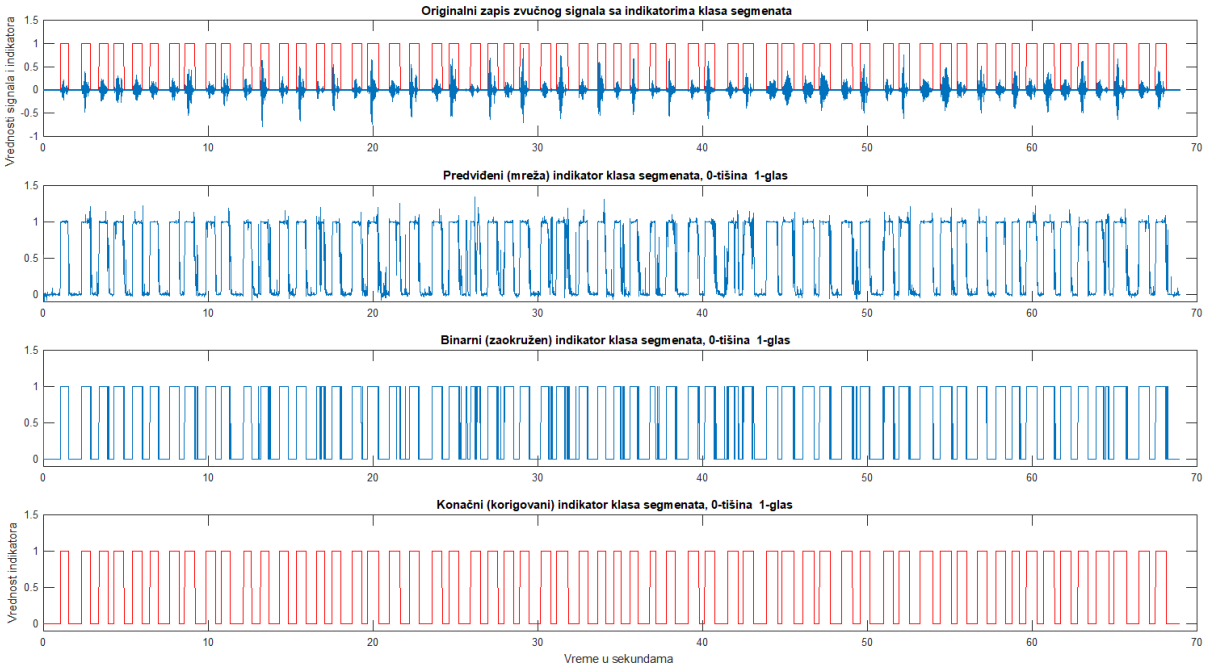
Na slici 4.16 je prikazan kontinuirani govorni signal koji sadrži niz od osam sukcesivno izgovorenih reči i devet sekvenci „tišine“ proizvoljno odabranih iz zvučnog zapisa GAT testa sa trideset reči. Ovaj signal, koji je produkovao ispitanik (A) najpre je korišćen za obuku klasifikatora (MLP) a zatim je isti signal uzet kao test uzorak za obučenu neuronsku mrežu (modul odluke) kod VAD agoritma. Frekvancija odabiranja analiziranih govornih signala bila je takođe 11.025 kHz dok je dužina frejmova bila 10 mm, što je zaokruženo na 110 odbiraka

po frejmu. Tačnu poziciju aktuelnih granicai ovde su odredili trenirani logopedi. I ovde je prihvaćen kriterijum tačnost predikcije po kom je dopuštena greška pri određivanju granice fonema bila u dužini od 30 ms, odnosno 3 frejma ili 330 odbiraka signala. Obučavajući, ručno kreirani, uzorak govornog signala i sadržao je 8 sekvenci sa rečima i 9 sekvenci tišine. Iz uzorka su računati vektori iz prve grupe VAD obeležja za svaki frejm a njima su pridružene vrednosti indikatora govra (1) i tišine (0). Tako je formiran obučavajući skup podataka za VAD. Kao test uzorak koristimo identični integralni kontinuirani sigal iz kog smo izdvojili obučavajući uzorak. Manja slika na Slici 4.16 prva po redu prikazuje originalni zvučni signal i finalni indikator klase (crvena linija). Manja slika druga po redu, prikazuje originalnu procenu klase dobijenu MLP klasifikatorom, dok su na trećoj i četvrtoj slici prikazane zaokružene i korigovane (ispeglane) vrednosti indikatora klasa. Korekcija zahteva heuristički pristup što ukazuje na nesavršenost klasifikatora. Po ovom kriterijumu svaki od izdvojenih segmenata je pravilno klasifikovan što je za očekivati jer je zadržan isti pozadinski signal i isti signal glasa (ispitanik A). Ovaj već obučeni VAD detektor, doduše sa malo znanja o glasu ispitanika (A), izložen je novom test uzorku, kontinuiranom govornom signalu cele dužine koji sadrži 30 izgovornih reči iz GAT testa a takođe pripada istom ispitaniku (A) koji je poslužio za obučavajući uzorak pa se može očekivati da on nema velikih varijacija signala glasa i tišine u odnosu na obučavajući signal. Rezultati testiranja obučenog VAD detektora dovođenjem velikog test signala prikazani su na Slici 4.17. Treba skrenuti pažnju da signal prikazan na slikama 4.18 i 4.19 sadrži izgovorne reči koje logoped izgovara pre ispitanika, tako da na Slikama 4.18 i 4.19 ima ukupno 60 reči.

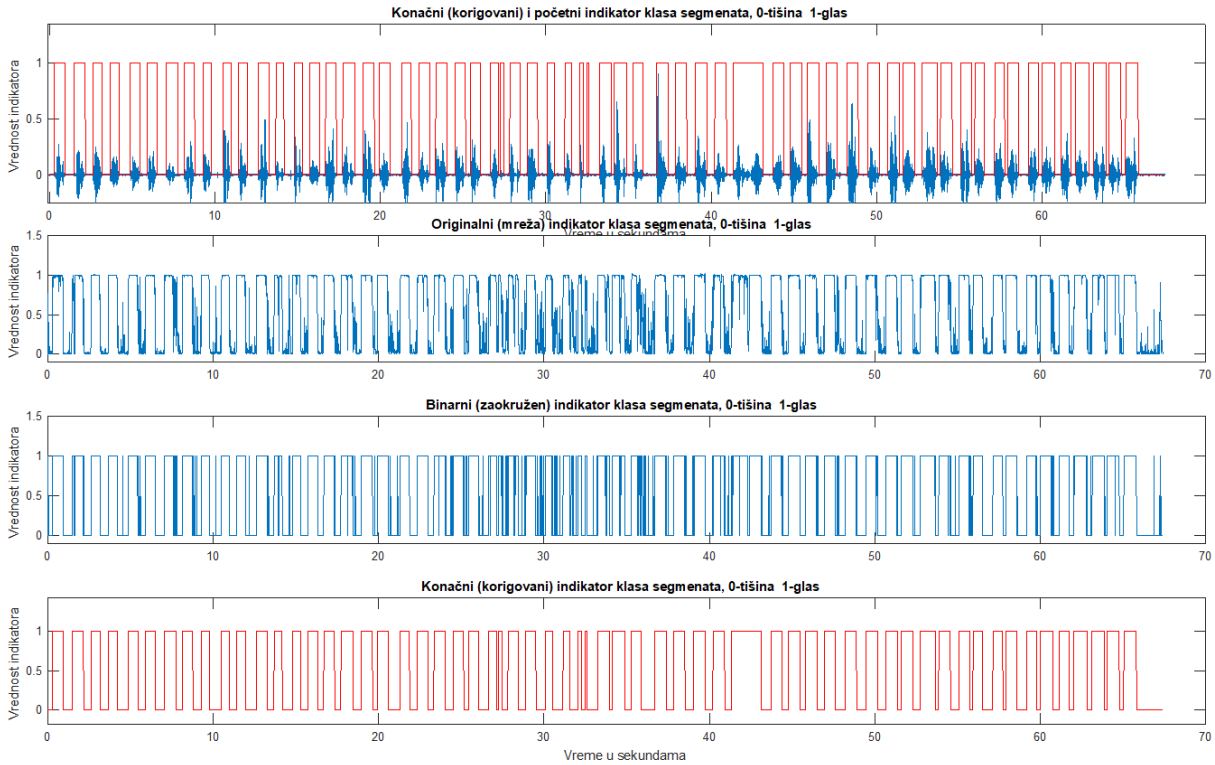
Sledeći test signal se odnosi na drugog ispitanika (B) i kao takav sadrži izvesne razlike u smislu akustičkih karakteristika kako govornog signala tako i signala tišine. Prethodno stečeno znanje VAD detektora stečerno na signalu (A) ograničene dužine testira se na novom nepoznatom signalu (B). Rezultati testiranja VAD drugim velikim test signalom prikazani su na Slici 4.18. Ovi primeri imaju za cilj proceenu mogućnosti dizajniranja i primena VAD algoritma za utvrđene uslove akvizicije govornog signala. Na osnovu prikazanih slika zaključujemo da ima smisla uvoditi ovako dizajniran VAD detektor u sitem za ocenu kvaliteta govora, pod uslovom povećanja njegove baze znanja. Na ovaj način se ubrzava ekstrakcija korisnog signala i dodaju u bazu novi zvučnih stimulusi reči, što predstavlja veliku uštedu vremena za logopede i takođe uvećava baza stimulusa reči čime se povratno povećava i efikasnost primenjenog VAD algoritma i otvara mogućnost potpune automatizacije aktuelnog procesa. Zahtevajući visok nivo tačnosti pri VAD ekstrakciji reči stvara se bolja osnova za efikasnu segmentaciju fonema iz tih reči uz nešto više utrošenog vremena.



Slika 4.16 VAD Ekstrakcija reči iz poznatog zvučnog signala, ispitanik (A).



Slika 4.17 VAD Ekstrakcija reči iz delimično poznatog zvučnog signala, ispitanik (A).



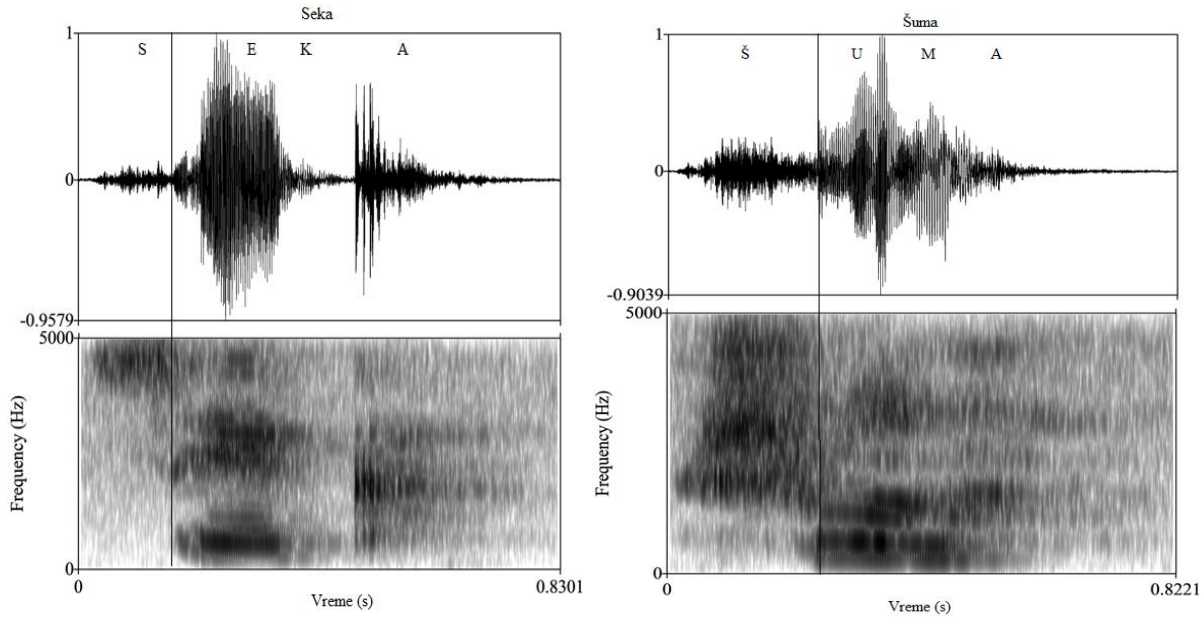
Slika 4.18 VAD Ekstrakcija reči iz nepoznatog zvučnog signala, ispitanik (B).

4.5.2. Rezultati segmentacije reči izdvojenih putem VAD

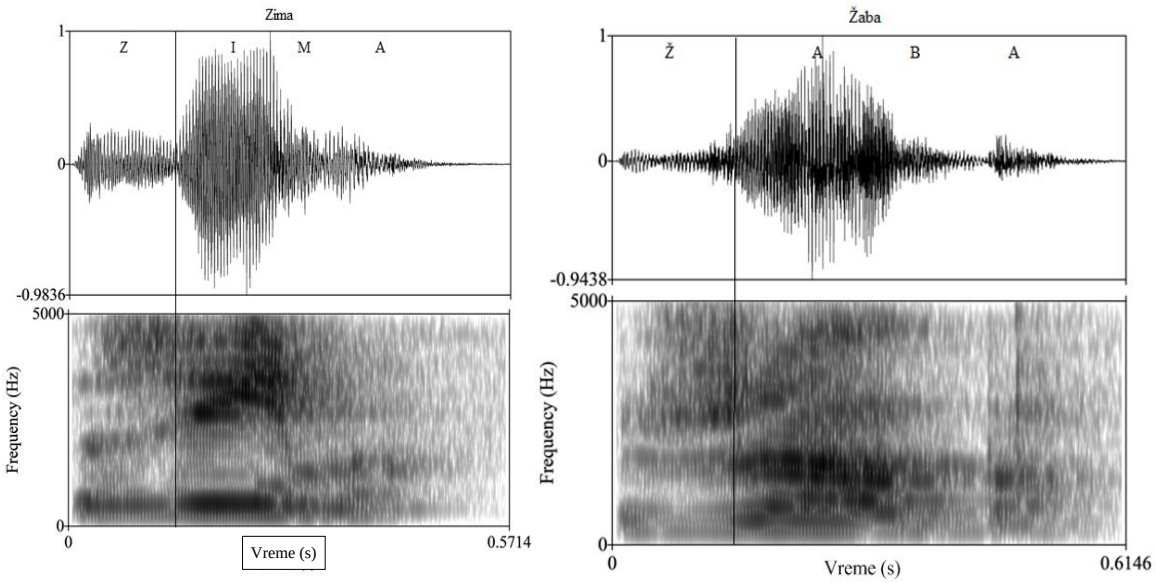
Ocena valjanosti procesa segmentacije reči primenom četiri različita metode izvedena je uporednim prikazom rezultata segmentacije. Testiranje procedure segmentacije je sprovedeno na govornim stimulusima reči sadržanih u GAT testu. Frekvancija odabiranja analiziranih govornih signala bila je 11.025 kHz dok je dužina frejmova bila 10 mm, što je iz praktičnih razloga zaokruženo na 110 odbiraka po frejmu. Tačnu poziciju granica su odredili trenirani logopedi sa velikim iskustvom u oblasti analize govora u vremenskom i spektralnom domenu kao i auditivne percepcije karakterističnih obeležja. Uveden je kruterijum tačnost predikcije po kom je dopuštena greška pri određivanju granice fonema bila u dužini od 30 ms, odnosno 3 frejma ili 330 odbiraka signala. Ovo je urađeno iz tog razloga što razlike u ručnoj definiciji granica između logopeda variraju u tom opsegu dužina. Pošto se algoritam za segmantaciju takođe zasniva na metodi prepoznavanja oblika, za obuku prediktora korišćeni su ručno segementirane sekvence fonema uzetih od 13 ispitanika za obuku svih preiktora. Teastiranje performansi obučenih prediktora za određivenje granica segmenata fonema u rečima koje su prethodno ekstrahovane pomoću VAD detektora, izvedeno je na test uzorku produkovanom od strane 137 ispitanika.

Ovaj uzorak služi kao dobar indikator svrsishodnosti celokupne primenjene procedure. Za prikaz rezultata procedure segmentacije su korišćene akustičke karakteristike sledećih četiri izdvojene reči: Seka, Šuma (Slika 4.19) i Zima i Žaba (Slika 4.20).

Sve uključene reči u inicijalnoj poziciji sadrže frikative koji spadaju u glasove sa najvećom frekvencijom odstupanjapri izgovoru po tipu distorzije, koja je ovde predmet analize u kontekstu kvaliteta artikulacije. Iz tog razloga rezultati dobijeni u analizi ovih fonema imaju najveću težinu. Prve dve reči u inicijalnoj poziciji sadrže bezvučne frikative S i Š a sledeće reči sadrže zvučne frikative Z i Ž. Segmeni sva četiri frikativa su odvojeni od ostataka reči vertikalnom linijom na slikama 4.19 i 4.20. Na slici 4.21 su prikazani talasni oblik i spektrogram segmenata signala „Tišine“ u svrhu komparacije distinktivnih akustičkih karakteristika frikativa, vokala i signala „Tišine“. Iza svih frikativa u svim rečima sa slika sledi signal vokala koji se odlikuje zvučnošću, regularnošću i harmoničnošću što implicira visok nivo energije, mali broj nultih prelaza, veliki korelacioni koeficijent i malu grešku predikcije i determiniše vrednosti svih 17 parametara iz druge grupe. Ukoliko su karakteristike inicijalnih fonema slične karakteristikama susednih vokala tada je stopa njihove diskriminacije manja i tačnost segmentacije opada. Zvučni konsonantni fonemi imaju više sličnosti sa vokalima u smislu pomenutih kvaliteta pa je za očekivati manju tačnost njihove segmentacije. Bezvučni fonemi pokazuju veću razliku inherentnih akustičkih kvaliteta u odnosu na vokale (nedostatak jesnih formantnih struktura u pojasu spektru sa nižim frekvencijama) pa je očekivana greška segmenatacije teoretski manja u odnosu na zvučne foneme.

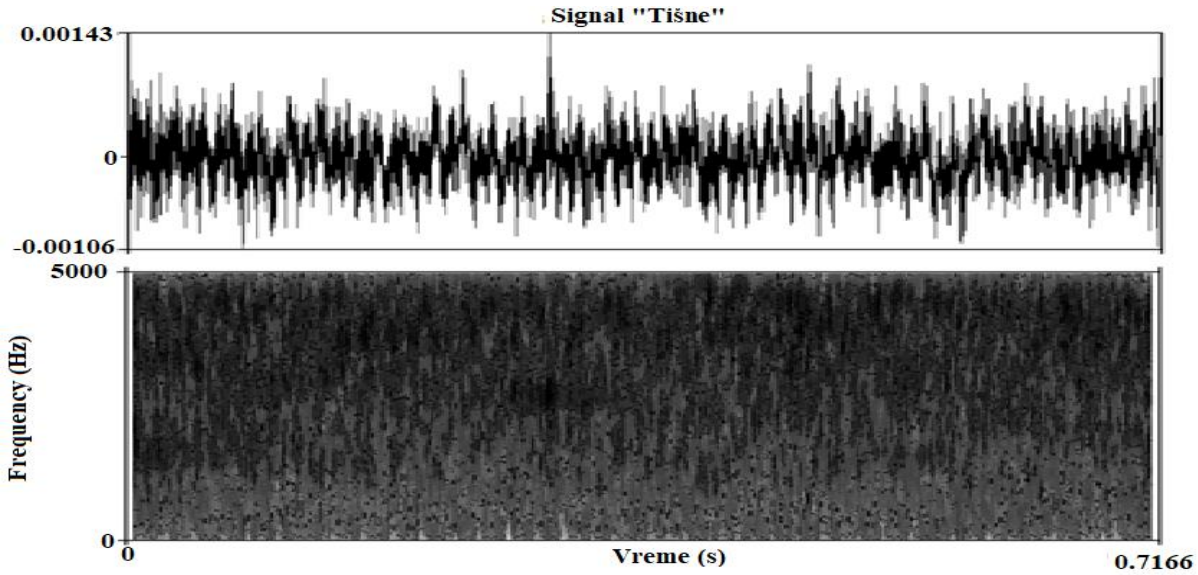


Slika 4.19 Talasni oblik i spektrogram govornih segmenata reči Seka (levo) i Šuma (desno), sa bezvučnim frikativima s i š u inicijalnoj poziciji.



Slika 4.20 Talasni oblik i spektrogram govornih segmenata reči Zima (levo) i Žaba (desno), sa zvučnim frikativima z i ž u inicijalnoj poziciji.

Zvučni fonemi imaju izražene formantne strukture u oblasti nižih frekvancija spektra, tako da se ove strukture zbog koartikulacije prelivaju u formante predstojećih vokala. Spektrogram signala tišine prikazan na slici 4. 21 ukazuje na stohastičku raspodelu energije na celom opsegu frekvencija pokazujući nedostatak regularnosti karakterističnu delimično i za bezvučne frikative što može ometati određivanje granice između tišine i ovih fonema



Slika 4.21 Talasni oblik i spektrogram segmenata signala „Tišine“.

Tabela 4.2 Greška segmentacije prediktora.

Klasifikatori	Izgovorne reči			
	Seka	Šuma	Zima	Žaba
	Greška klasifikacije [%]			
KNN Ansambl	5.8394	6.5693	10.2190	8.7591
Naive Bayes	10.2190	12.4088	16.0584	14.5985
SOM	19.7080	21.8978	27.0073	21.1679
MLP Ansambl	5.1095	4.3796	7.2993	7.2993

Ovi faktori nisu jedini uticajni faktori na tačnost segmentacije jer pored njih mogu delovati raspodele trening skupova i tipovi primenjenih prediktora. Tačnost segmentacije inicijalnoh fonema kao nosilaca vektora obeležja tipičnosti izgovora je bitan korak predobrade.

Prikaz rezultata segmentacije inicijalnih fonema iz navedenih reči dat je u Tabeli 4.2. Očigledna je razlika u greškama predikcije granica segmenata različitih prediktora, ali postoji i razlikia među greškama predicije segmenata među pojedinim rečima. Razlog prednosti Ansambla MLP u odnosu na ostale prediktore leži u većoj fleksibilnosti i robustnosti u odnosu na njih. Primećena razlika u tačnosti predikcije granica segmenata na nivou različitih reči posledica je izražene razlike raspodela važnih akustičkih kvaliteta po fonemima datih reči o kojima smo govorli u prethodnom pasusu. Ova procedura služi za formiranje baze zvučnih stimulusa fonema u tipičnoj i atipičnoj realizaciji. Iz ove baze biće formirani vektori distinktivnih obeležja za ocenu kvaliteta artikulacije tih fonema i zato je njihova kvalitetna segmentacija bitna za bolju procenu kvaliteta njihovih akustičkih karakteristika u smislu kvaliteta artikulacije.

5. PROBLEM NEIZBALANSIRANOG UČENJA

U ovom poglavlju je prkazan novi opšti pristup neizbalansirtanom učenju kao jednom od glavnih izazova u oblasti klasifikacije uzoraka. Preciznije, ovo poglavlje objedinjuje teoretski aspekt i eksperimentalnu potvrdu prednosti novog algoritma za balansiranje neizbalansiranih klasa, zasnovanog na principu maksimizacije entropije uzoraka. Metod podrazumeva detekciju distribucionih karakteristika idealno izbalansiranog uzorka datog u formi pravilne rešetke i prihvatljivi transfer ovih karakteristika na proizvoljni neizbalansirani uzorak (klasu), koristeći tehniku uzorkovanja kojom se menja struktura neizbalansiranog uzorka u pravcu porasta njegove reprezentativnosti i balansa. Predložena procedura podrazumeva uklanjanje postojećih instanci iz oblasti velike gustine raspodele verovatnoće (undersampling) u kombinaciji sa sintetičkom generacijom novih instanci u oblastima male gustine (oversampling). Ova procedura se primenjuje na svaku klasu pojedinačno u cilju redukcije unutrašnjeg imbalansa svake klase koje rezultuje značajnim poboljšanjem performamnsi involviranih klasifikatora.

Glavni funkcija ovog algoritma je povećanje entropije uzoraka (klasa) koja implicira redukciju tendencije induktivnih klasifikatora da favorizuju većinsku klasu ili klastere tokom treninga. Pored problema balansiranja klasa, ovaj algoritam može biti vrlo koristan kod problema prepoznavanja uzoraka i redukcije dimenzija uzoraka praćenog povećanjem njihove reprezentativnosti. Visok stepen opštosti metoda garantuje njegovu primenjivost u radu sa podacima velikih dimenzija i velikog stepena kompleksnosti strukture. Prikazana teoretska osnova metoda je verifikovana na adekvatnom sintetičkom skupu uzoraka a njegova praktična primenjivost je potvrđena na komparativnoj klasifikaciji velikog skupa raspoloživih empirijkih podataka koristeći nekoliko različitih poznatih klasifikatora. Krajnji cilj prikazane metode je poboljšanje performansi klasifikatora pri klasifikaciji i evaluaciji kvaliteta artikulacije fonema srpskog jezika.

Ovde je predstavljen novi generalni pristup problemu neizbalansiranog učenja, na osnovu prenosa određenih distributivnih svojstava uzorka u formi regularne rešetke na neizbalansirane klase uzorka proizvoljne strukture. Na osnovu karakteristika raspodele, usvojili smo uzorak u obliku regularne celobrojne rešetke kao idealanu paradigmu uniformnosti, reprezentativnosti i unutrašnjg balansa instanci.

Prenesene distributivne karakteristike su zasnovane na odnosu koji je uspostavljen između zapremine koja pripada svakoj instanci uzorka i njene lokalne srednje udaljenosti od unapred definisanog skupaa najbližih suseda. S obzirom na činjenicu da je ravnomerno raspoređeni trening uzorak ima dobre karakteristike u pogledu reprezentativnosti u odnosu na ciljnu

populaciju, prenos ovih karakteristika na proizvoljno distribuirane trening uzorke treba da pokaže pozitivan efekat na reprezentativnost tih uzorka i poboljšanje performansi involviranog klasifikatora.

U ovom poglavlju je prikazan novi opšti pristup neizbalansiranom učenju kao jednom od glavnih izazova u oblasti klasifikacije uzoraka (He i Garcia, 2009; Garcia i sar., 2008; Japkowicz, 2003; Weiss, 2004; Furundžić i sar., 2013a; Furundžić i sar., 2015.). Preciznije, ovde je data teoretska osnova i eksperimentalna potvrda prednosti novog algoritma za balansiranje neizbalansiranih klasa, zasnovanog na principu maksimizacije entropije uzoraka (Furundžić i sar. 2017b). Metod se zasniva na detekciji distribucionih karakteristika idealno izbalansiranog uzorka datog u formi pravilne rešetke i transferu ovih karakteristika na proizvoljni neizbalansirani uzorak, koristeći tehniku uzorkovanja kojom se menja struktura neizbalansiranog uzorka u pravcu porasta njegove reprezentativnosti i balansa. Predložena procedura podrazumeva uklanjanje postojećih instanci iz oblasti velike gustine raspodele verovatnoće (undersampling) u kombinaciji sa sintetičkom generacijom novih instanci u oblastima male gustine (oversampling). Ova procedura se primenjuje na sve klase pojedinačno u cilju redukcije unutrašnjeg imbalansa svih klasa što rezultuje značajnim poboljšanjem performamnsi involviranih klasifikatora. Observabilna manifestacija ovog algoritma je povećanje entropije uzoraka (klasa) koja implicira redukciju tendencije induktivnih klasifikatora da favorizuju većinsku klasu ili klaster tokom treninga. Prikazana teoretska osnova metoda je verifikovana na adekvatnom sintetičkom skupu uzoraka a njegova praktična primenjivost je potvrđena na komparativnoj klasifikaciji velikog skupa od dvadeset (20) raspoloživih empirijkih baza podataka. Unutrašnji imbalans klasa utiče na performanse klasifikatora (Furundžić i sar. 2014a), dok neravnoteža među klasama predstavlja mali problem kada postoji prihvatljivi balans unutar klasa, u smislu reprezentativnosti podataka (Japkovicz, 2003). Krajnji cilj primene ove metode je poboljšanje performansi klasifikatora pri klasifikaciji i oceni kvaliteta artikulacije fonema srpskog jezika (Furundžić i sar. 2017c, 2015).

U kontekstu neizbalansiranog učenja postoje ključne činjenice i koncepti čije je poznavanje neophodno za razumevanje i rešavanje ovog problema, a u nastavku teksta skrećemo pažnju na neke od njih, dok je, za detaljnu i iscrpnu analizu problema, preporučujemo čitalaca radovi izloženi u Japkovicz (2003), Veiss (2004), Chawla et al. (2004), He et al. (2008) i He i Garcia (2009). U radovima Vajs (2004) i He i Garsija (2009) autori daju detaljan pregled najnovijeg razumevanja pojmova neizbalansiranog učenja i aktuelnih pristupa rešenju ovog problema. Koncept klase pretpostavlja grupu primera koje karakteriše vektor diskriminatorskih karakteristika koji ih odvajaju od drugih, više ili manje sličnih grupa. Ovi koncepti, većinska

(negativna) klasa i manjinska (pozitivni) klasa su ranije opšte prihvaćena u oblasti, kao termini koji ukazuju da je jedna klasa znatno brojnija od druge klase (disbalans između klasa), (He i Garsija, 2009), koji se često izražava kroz velike vrednosti odnosa neravnoteže „Imbalance Rate“ (IR). U savremenom pristupu problemu neizbalansiranog učenje, zbog presudnog značaja interne neravnoteže involviranih klasa, pojmovi većinske i manjinske klase su zamenjeni pojmovima nedovoljno zastupljene klase i previše zastupljene klase, kao deskriptivnijim pojmovima (Japkovicz, 2001). Svaka instanca (primerak) može se predstaviti kao jedinstvena tačka u multidimenzionalnom prostoru. Uzorak može da obuhvati sve instance ciljne (target) populacije, ili samo jedan deo cele populacije (uzorak), što je najčešći slučaj u praksi. Izvorni skup instanci „sampling frame“ predstavlja sve raspoložive primere target populacije i kao takav je izvor iz kog se bira reprezentativni uzorak cele populacije. Izvorni uzorci mogu više ili manje ravnomerno pokrivati manji ili veći dio prostora ciljne populacije i ta činjenica ima odlučujući uticaj na reprezentativnost izvornog uzorka i samim tim na trening uzorak. Različite tehnike uzorkovanja, kao na primer, sistematsko i stratifikovano uzorkovanje se primenjuje za prevazilaženje problema reprezentativnosti. Raspodela verovatnoća raspoloživih izvornih podataka determiniše stepen reprezentativnosti izabranog uzorka. Veći stepen uniformnosti raspode primera u prostoru obeležja instanci, kao i veći prostor zauzet primerima, odgovaraju većem stepenu reprezentativnosti uzorka. Ovaj važan iskaz je predmet dokazivanja i diskusije u daljoj prezentaciji. Pojam kompleksnosti uzoraka ili “kompleksnosti klasa“ je termin koji se odnosi na varijabilnosti gustine verovatnoće slučajeva u prostoru obeležja poznat kao mešavina distribucija koja se karakteriše pojavom tzv. sub koncepata, sub klase ili klastera. Veća varijabilnost gustine instanci u prostoru obeležja odgovara višem stepenu kompleksnosti. Pojava sub koncepata ili sub klase u osnovnim klasama izaziva pojavu kod klasifikatora poznat kao “mali disjunkti” koji predstavlja skup pravila koji je podržan od strane malog broja primera za obuku koji obično imaju slabu prediktivnu tačnost (Holte et al., 1989). Prilikom izbora uzorka obuke, primarni cilj je da se postigne visok stepen njene reprezentativnosti. Prema ranijim činjenicama, ovaj stav znači da izabrani uzorak treba da što je moguće ravnomernije pokriva što više raspoloživog prostora instanci. Ovaj zahtev je jedan od glavnih zadataka neizbalansiranog učenja, i za tu svrhu smo dizajnirali novu resampling metodu za balansiranje klasa. To je metod balansiranja zasnovanog na međusobnim lokalnim rastojanjima instanci neuravnotežnih klasa, a naziva se „Distance Based Balancing“ (DBB), što se može smatrati glavnim doprinosom ovog istraživanja, kako u smislu teorije tako i u praktičnom smislu. Teorijska osnova stava o važnosti uniformnosti raspodele obučavajućih uzoraka u prostoru obeležja, zasniva se na maksimizaciju entropije uzoraka ili drugim rečima,

na minimiziranju tendencije klasifikatora ka favorizovanju zastupljenije klase tokom procesa obuke (Jaynes, 1957). Ostatak ovog poglavlja je organizovan na sledeći način. U odeljku 5.1, se bavimo važnim aspektima i implikacijama problema neizbalansiranih klasa, kao i standardnih metoda balansiranja, ističući svoje prednosti i ograničenja kao osnova za stvaranje novih metoda. Odeljak 5.2 je posvećen pojašnjenju "dobro izbalansiranih klasa" koji se sreće u literaturi, ali nije jasno definisan u smislu odnosa između neravnoteža u klasi i neravnoteže između klasa, gde ukazujemo na primarni značaj neravnoteže u samim klasama kao generatora problema naizbalansiranog učenja uopšte. Ovde objašnjavamo korespondenciju između neravnoteže unutar uzoraka i njihove reprezentativnosti kao osnove potrebne za razvoj novog algoritma balansiranja. U ovom delu takođe je potrebno pažnja je posvećena formalizaciji koncepta reprezentativnosti u svetlu teorije informacija. Sekcija 5.3 predstavlja glavni teoretski deo istraživanja gde smo uveli novi način balansiranja klasa, zasnovan na balansiranju funkcije gustine verovatnoće raspoloživih uzoraka za obuku posredstvom određivanja lokalnih distanci aktuelnih primera. Posebna pažnja je posvećena teorijskim odnosno matematičkim osnovama ovog modela.

5.1. Problem disbalansa klasa i aktuelni pristupi tom problem

Pristupi rešenju problema neizbalansiranog učenja mogu se podeliti u dve osnovne kategorije: rešenja na algoritamskom nivou ili na nivou unutrašnjih rešenja, koje se zasnivaju na stvaranju novog ili modifikaciju postojećih algoritama za rešavanje problema disbalansa klasa, Pazzani i dr. Pazzani i dr. (1994); Japkovicz i dr. (1995); Kubat i dr. (1998), i rešenja na nivou podataka ili spoljašnja rešenja, gde se koriste poznati algoritmi u izvornom obliku, ali se menja prvobitna distribucija podataka koji se koriste kako bi se smanjio negativan efekat inherentne neravnoteže klase (Levis i Gale, 1994, Levis i Gale, 1994a), Kubat i Matvin (1997); Ling i Li (1998); Dramond i Holte (2003); Maloof (2003); Japkovicz (2000); Weiss i Provost (2003); Chavla et al. (2004); Han et al. (2005); Monard et al. (2002). Algoritamski pristupi imaju određene prednosti, ali često imaju nadostatke kao algoritamske specifičnosti Estabrooks i dr. (2004). To je ozbiljan problem, jer isti skupovi podataka klasifikovani različitim algoritmima daju različitu tačnost klasifikacije, i to može predstavljati prepreku za prenos zadatka sa jednog klasifikatora na drug što se često zahteva. Algoritamski pristupi klasifikaciji naglašavaju primarni značaj različitih iduktivnih algoritama učenja za klasifikaciju kao što su: metode osetljive na cenu pogreške, support vector mašine (SVM) (Vuand i Chang, 2003) i Liu i Huang (2006), neuronske mreže (NN) (Lavrence i dr., 1998), metoda k najbližih susede

(KNN) (Vilson, 1972), genetski algoritmi (GA) (Garsija i dr., 2009), drvo odlučivanja, aktivno učenje, metode klasifikacije zasnovane na prepoznavanju i druge metode zasnovane diskriminaciji. Za razliku od algoritamskih internih pristupa, eksterni pristupi su nezavisni od algoritma za klasifikaciju koja se koristi, pokazujući, samim tim, veći stepen opštosti i prilagodljivosti. Na osnovu rezultata (Chavla i sar., 2004) metode uzorkovanja postaju de facto standard za suzbijanje disbalansa klasa. Na osnovu prethodnih činjenica i većeg stepena opštosti i primenljivosti eksternih metoda, izabrali smo eksterni pristup za pronalaženje opštijih i efikasnijih rešenja za problem klasifikacije neizbalansiranih klasa

5.1.1. Resampling podataka

Pošto je veliki deo istraživanja predstavljenog u ovom radu odnosi na novi resampling postupak potrebno je prikazati postojeće važne metode u ovoj kategoriji kroz kritički osvrt na njihov učinak, prednosti i nedostataka, kao i implikacije na nove pristupke. Takođe treba, u postojećoj konstelaciji da se odredi pozicija nove predložene metode u pogledu performansi i naglase njene moguće prednosti, nedostatci i perspektive. Eksterne metode odabiranja su standardni pristupi za rešavanje problema internog disbalansa u klasama. Ključna ideja je da se kroz preprocesing trening uzorka smanji disbalans između klasa. Drugim rečima, tehnika resamplinga treba da izmeni apriorne verovatnoće raspodele dominantnih i inferiornih klasa u obučavajućem uzorku u cilju dobijanja izbalansiranog odnosa broja instanci u svakoj klasi. Ova kategorija obuhvata različite tehnike koji se mogu svrstati u dve forme, oversampling i undersampling. Randomizirani oversampling i undersampling predstavljaju paradigmu resampling tehnike iz koje su se razvile druge heurističke ili vođene resampling tehnika kao što su dirigovani oversampling, dirigovani undersampling, oversampling sa slučajnom ili heurističkom generacijom novih uzoraka, i različiti oblici kombinacija pomenutih tehnika (Chavla i dr., 2002, Han i dr., 2005).

5.1.2. Osnovni resampling algoritmi

Opšti smisao i efekat metoda oversamplinga i undersamplinga je ekvivalentan i uključuje promene u veličini početnog seta podataka radi balansiranja inherentnog imbalansa aktuelne klase. Dakle, potreban balans klasa može se postići dodavanjem primeraka inferiorne ili slabije zastupljene klase (oversampling), redukcijom instanci superiorne ili bolje zastupljene klase (undersampling) ili kombinacijom obe metode. Iako obe metode imaju zajednički cilj, ipak,

pomenuti pristupi imaju različite efekte koji utiču na proces učenja (Estabrooks, 2000, Mease i dr., 2007, Drummond i Holte, 2003).

5.1.3. Oversampling (odabiranje sa dodavanjem primeraka)

Dobro poznat postupak za povećanje veličine inferiorne - manjinske ili pozitivne klase je slučajni oversampling, to je, jednostavna heuristička metoda za balansiranje distribucije slučajnom replikacijom primeraka manjinske klase. Ovaj problem je rano prepoznao u istraživačkoj oblasti neuronske mreže (Derouin i dr., 1991). Autori su ukazali rešenja kroz jednostavne replikacija primera, kreiranje novih primera, i porast parametra učenja za primere koji pripadaju manje zastupljenim inferiornim ili pozitivnim klasama. Ovaj metod povećava ravnotežu distribucije klasa bez dodavanja novih informacija podacima, Japkovicz (2001). Jednostavna replikacija postojećih pozitivnih primjera može dovesti do „overfittinga“ primeraka iz manjinske klase. Za klasifikator h se kaže da vrši overfitting trening uzorka podataka ako postoje neki alternativni klasifikator h’ iz iste kategorije, takav da h ima bolje performanse nad trening skupom primera, ali h’ ima bolje performance nego h nad skupom ciljane populacije (Mitchell., 1997, Estabrooks, 2000). Chavla i dr. (2002) predlažu tehniku oversempling na osnovu generacije novih sintetičkih primera manjinske klase koristeći efikasan interpolacioni algoritam. Ovaj metod se zove SMOTE (Synthetic Minority Over-sampling Technique). Taj algoritam pomaže klasifikatoru da poveća region odlučivanja koji sadrži primere iz manjinske klase. Od SMOTE algoritma, nastale su mnoge modifikacije koje su prikazane u literaturi. SMOTE Boost je algoritam, koji je dao Chavla i dr. (2003), kombinuje SMOTE sa busting procedurama. Han i dr. (2005) je kreirao granični SMOTE algoritam, koji generiše nove sintetičke primere u blizini postojećih slučajeva koji su blizu regije razgraničenja. U SMOTE algoritmu, problem over generalizacije je povezan sa procesom stvaranja sintetičkih uzoraka. Ustvari, SMOTE algoritam generiše isti broj sintetičkih uzoraka podataka za svaki originalni primer manjinske kategorije, bez obzira na broj njegovih susednih primera, što povećava preklapanje između klasa (He i Garcia, 2009). Prilagodljivi način uzorkovanja takođe je predložio He i dr. (2008) pokušava da prevaziđe ovaj problem. Ovaj metod je poznat kao Adaptivno Sinthetic Sampling algoritam (ADASIN), He i dr. (2008). Hongya i Herna (2004) su uveli Data Boost-IM metodu, zasnovanu na kombinaciji bustinga i generisanja podataka. Garsija i dr., 2008 su razvili metod zasnovan na konceptu susednih primera, uzimajući istovremeno u obzir blizinu i distribuciju primera.

5.1.4. Undersampling (odabiranje sa uklanjanjem primeraka)

Undersampling tehnika se zasniva na izboru manjeg skupa primera iz većinske klase uz zadržavanje svih instanci iz manjinske klase. Ova matoda je primenljiva u slučaju kada je količina primeraka i redundansa većinskih klasa veoma velika, pa smanjenje uzorka obuke smanjuje vreme i skladištenje podataka bez velikih gubitaka informacija. U literaturi važi stav da je random undersampling, uprkos poznatim nedostatacima, jedna od najefikasnijih resampling metoda (Garcia et al., 2008). Kod slučajnog undersamplinga, slučaj koji je prenaglašen u literature, (He i Garcia 2009, Garsija i dr., 2008 i Džo i Japkovicz, 2004), tvrde da uklanjanje primera iz većinske klase može dovesti do toga da klasifikator izgubi važne činjenice koji se odnose na većinsku klase i tako degradira svoje performanse. Ovaj stav je ispravan, ali problem se može lako prevazići različitim metodama heurističkog oversamplinga ili pomoću bagging prediktora (Breiman, 1994). Jedan od pionirskuh radova na poboljšanju slučajnih resampling metoda je rad koji su objavili Kubat i Matvin (1997). Oni su predložili tehniku jednostrane selekcije (OSS). OSS tehnika pokušava da heuristički ukloni instance koje pripadaju većinskoj klasi a koje su redundantne i / ili pripadaju graničnoj (zašumljenoj) oblasti. Oni zadržavaju sve pozitivne primere odnosno primere iz manjinske klase. Granični primeri se otkrivaju primenom odnosa koji su definisani kao Tomek link (Tomek, 1976), dok su redundantni primeri eliminisani putem kondenzovanog Hartovog algoritma (Hart, 1968). Ovaj algoritam efikasno rešava problem lokalnih razgraničenja između klasa, ali odbacivanje velikog broja slučajeva iz većinske klase može dovesti do značajnog gubitka informacija o ciljnoj populaciji. Vilson's editing (Vilson, 1972), je takođe dobro poznata metoda zasnovana na identifikaciji i uklanjanju zašumljenih primera većinske klase.

5.1.5. Problem kompleksnosti koncepata

Grupa metoda koje pristupaju resampling tehnikama na osnovu termina "kompleksnost koncepta" koriste se sve više zato što se bave suštinom neizbalansiranog učenja. Ovaj termin podrazumeva postojanje različitih struktura u okviru prostora obeležja a koje su posledica neujednačenog prostornog rasporeda instanci unutar jedne klase, poznatih kao sub koncepti, sub klase, klastera itd (Nickerson i dr., 2001). Ova grupa uključuje oversampling i undersampling zasnovane na klasterima. Pomenuti rad predstavlja kritički osvrt na resampling metode koje izbegavaju razmatranja slučaja gde su, u okviru jedne klase, podaci distribuirani

u skladu sa mešavinom gustina čije komponente imaju relativne gustine koje se razlikuju u velikoj meri. U ovoj situaciji, rešavanje jednog problema može da se kompromitovati stvaranjem drugog. Ovaj heuristički resampling postupak koristi tehniku klasterovanja bez nadzora, Principal Direction Divisive Partitioning, da odredi unutrašnje karakteristike svake klase pronalaženjem broja klastera u okviru klase. Pronađeni klasteri se uzorkuju tako da svaki klaster iz obe klase sadrži isti broj primera. Izjednačavanjem broja članova u svakoj podkomponenti izbegava se povećanje razlike u relativnim gustinama podkomponenti u svakoj klasi. Jedan od radova sa najdubljim uvidom u kompleksnost problema imbalansa klasa (CI) , (Japkovicz i Stephen, 2002.), predstavlja složenu i detaljnu studiju sa ciljem da se odgovori na nekoliko važnih pitanja. Prvo je namera da se razume priroda problema imbalansa klasa kroz uspostavljanje odnosa između složenosti koncepta, veličina trening skupa i stepena neravnoteže između klasa. Drugo pitanje se odnosi na poređenje i diskusiju nekoliko standardnih resampling tehnika za rešavanje CI problema. Treće pitanje odgovara na pretpostavku da CI problemi utiču ne samo na klasifikatore zasnovane na drvetu odlučivanja već i na neuronske mreže i SVM. Eksperimenti su pokazali da je imbalans klasa relativan problem determinisan sa 1) stepenom disbalansa klasa, 2) složenošću koncepta koja se manifestuje preko podataka, 3) ukupnom veličinom trening skupa, i 4) tipom uključenih klasifikatora. Drugim rečima, zaključeno je da je što je veći stepen neravnoteže klasa, veći stepen složenosti koncepta i manji obučavajući skup, tada neravnoteža klasa ima veći uticaj na performanse klasifikatora. Ovaj fenomen se objašnjava činjenicom da je visok stepen složenosti i neravnoteže, u kombinaciji sa malim skupovima obuke, uzrokuje pojavu veoma malih subklastera koji izazivaju pojavu malih disjunkta (Jo i Japkovicz, 2004). Mali disjunkti predstavljaju skup induktivnih pravila klasifikacije sklonih greškama zbog zasnovanosti na malom broju primera za obuku. Oni su takođe zaključili da metode zasnovane na random resamplingu mogu biti veoma korisne za poboljšanje performansi klasifikatorima osetljivih na imbalans. Od ovih metoda slučajni oversampling metod je efikasniji od slučajnog undersamplinga. Japkovicz takođe predlaže oversampling metodu zasnovanu na klasterima klastera (CBO) algoritam u Japkovicz (2003) i Japkovicz (2001), koja se uspešno nosi sa disbalansom unutar klasa i disbalsnsom među klasama istovremeno. Resampling strategija predstavljena u Japkovicz (2001) pokazuje da je neravnoteža unutar klasa i neravnoteža između klasa doprinosi povećanju stope netačnosti klasifikacije višeslojnog perceptron. Dalje, ova studija pravi razliku između dve vrste navedenih disbalansa i posmatra njihov uticaj na tačnost klasifikacije u vezi sa savršeno izbalansiranim situacijama ili rebalansiranim na veštačkim domenima podataka. Na sličan način operiše i resampling strategija data u Yen et al.

(2006) gde je predstavljen undersampling algoritam zasnovan na klasterovanju koji najpre klasteruje sve originalne primere većinske klase u određeni broj istaknutih pod klastera, a zatim bira odgovarajući broj uzoraka iz svakog klastera većinske klase u cilju balansiranja. Dakle zadržavaju se svi primeri iz manjinske klase. Pionirski rad na problemu disbalansa unutar klasa, Holte i dr. (1989) analizira odnos između malih pod klastera, malih disjunkta i greške klasifikacije. Weiss, u Weiss (2003) je uveo koncept koncentracije greške koje uspostavlja transparentan empirijski odnos između veličine malih disjunkta i vrednosti greške klasifikacije. Quinlan, u (Quinlan, 1991) takođe uspostavlja empirijski odnos između veličine malih disjunkta i stope greške klasifikacije za većinske i manjinske klase posebno. Obe random resampling metode imaju istu manu: ni jedna od njih ne menja prirodu apriorne raspodele verovatnoće unutar klasa tretirajući na isti način sve instance u prostoru obeležja što rezultira održanjem interne priorne distribucije verovatnoće instanci. Tokom izvođenja jednostavnog slučajnog oversamplinga, verovatnoća ponavljanja slučajeva iz deficitarnih područja (retki slučajevi) je proporcionalna njihovoj maloj priornoj raspodeli verovatnoće, što prouzrokuje ponavljanje malog broja slučajeva iz ovih oblasti, tako da, ova područja ostaju relativno retka i posle procedure slučajnog oversamplinga. Sa druge strane, primeri iz oblasti visoke gustine (centri sub klastera) imaju proporcionalno većoj gustini veću verovatnoću replikacije što se manifestuje zadržavanjem postojećeg neravnoteže instanci u okviru klase posle procedure slučajnog oversamplinga. Sličan efekat važi i za slučajni undersampling tehnika. Dakle, random resampling metode ne mogu imati bitnog uticaja na neravnotežu instanci unutar klase, zato balansiranje disbalansa između klasa primenom random resampling tehnika generalno ne može imati željeni efekat. Ova činjenica ukazuje na opšti nedostatak navedenih metoda slučajnog resampling i predstavlja značajan problem u ovoj oblasti. Balansiranje neravnoteže između klasa je najefikasnije kada je raspodela instanci u okviru pojedinačnih klasa uniforma. Pošto ravnomerna raspodela slučajeva u okviru pojedinačnih klasa značajno smanjuje efekat neravnoteže između klasa može se zaključiti da je primarni cilj resampling strategija eliminacija disbalansa unutar klasa, odnosno uspostavljanje najvećeg mogueg balansa u svakoj od klasa pojedinačno. Prema tome, glavni cilj metoda resampling bi trebalo da bude maksimalno moguće povećanje stepena uniformnosti raspodele verovatnoće unutar svake od klasa. Ovaj cilj se ne može postići jednostavnim postupcima random resampling ili navedenih resampling metoda zasnovanih na klasterima. Ovaj problem zahteva upotrebu kombinovanog resampling strategije koja sa jedne strane smanjuje broj slučajeva u oblasti prostora visoke gustine a na drugoj strani istovremeno, povećava broj primera u

oblastima niske gustine što se izvodi adekvatnim generisanjem novih sintetičkih instanci. Naša resampling strategija se zasniva na ovim upravo navedenim predlozima.

5.1.6. Kombinacija uzorkovanja i boosting tehnike

Ansambl tehnike učenja su već postale standard u rešavanju problema klasifikacije i aproksimacije funkcije, tako da je njihovo integrisanje sa resampling tehnikama takođe ispitivano u oblasti neizbalansiranog učenja. Na primer algoritma koji integriše SMOTE algoritam sa AdaBoost, SMOTEBoost, Chawla i dr. (2003), uvodi sintetičko generisanje uzoraka u svakoj od boosting iteracija tako da se svaki klasifikator iz ansambla fokusira na drugačiji način na manjinske klase. Budući da je svaki klasifikator iz ansambla konstruisan na drugom uzorku podataka, očekuje se da finalni klasifikator ima opštiji i bolje definisane granične oblasti za manjinske klasu (He i Garsija, 2009). U našem istraživanju smo koristili ansambl višeslojnih perceptron (MLP) koko bismo dobili pouzdaniji klasifikator.

5.2. Pojam izbalansirane klase

Mnogi autori, govoreći o problemima klasifikacije, pominju pretpostavku idealno balansirane klase, ali nijedan od njih ne definiše sam koncept idealno balansirane klase, zato je dublje razjašnjenje ovog pitanja jedan od važnih ciljeva ovog poglavlja. U većini radova koji se bave neizbalansiranim učenjem, neizbalansirani podaci još uvek se definišu kao skup podataka sa značajno poremećenom ravnotežom u broju predstavnika većinske i manjinske klase. Međutim, u nekim studijama (Japkovicz, 2003) je prikazan argument da kada ne postoji značajna interna neravnoteža instanci unutar klasa tada neravnoteža između klasa nije ozbiljan problem za klasifikatore, osim u slučaju veoma velikih vrednosti stope neravnoteže. Takođe treba napomenuti da u slučaju veoma naglašene neravnoteže unutar klasa , insistiranje na neravnoteži između klase nema nikakvog smisla i zato je neophodno da se fokusira pažnju na neravnoteže unutar klasa kao glavni generator problema učenja u uslovima disbalansa.

5.2.1 Reprezentativnost uzorka

Reprezentativni uzorak se može intuitivno definisana kao nepristrasan indikator stanja ciljne populacije. Prema Pan i dr. (2005) i Ma i dr. (2011), reprezentativni uzorak je pažljivo osmišljen podskup originalnog skupa podataka (populacije), sa tri glavne osobine: taj podskup

je značajno redukovan u smislu veličine u odnosu na originalni izvorni skup, podskup bolje pokriva glavne karakteristike iz originalnog izvora u odnosu na druge podskupove iste veličine, ima što je moguće manju redundansu (R) među reprezentativnim primerima koje sadrži. S obzirom na savremeni ubrzani rast u količini raspoloživih informacija, potreba za efikasnim otkrivanjem i smanjenjem redundanse postojećih baza podataka postaje sve više i više realna (Pan i dr., 2005). Reprezentativnost kao informaciono teorijska kategorije u ovom kontekstu ima dualnu prirodu koja se manifestuje kroz: visok stepen pokrivenosti prostora instanci i nisku redundansu instanci uzorka. Relevantni autori u oblasti kombinuju informaciono teorijske (zasnovane na entropiji) tehnike merenja reprezentativnosti, kao što su uzajamna informacija (Pan i dr., 2005, Ma i dr., 2011), Kullback-Leibler divergencija (relativna entropija) (Pan i dr., 2005, Ma i dr., 2011, Paek i Hsu, 2011), indeks različitosti (Bertino, 2006) i klastering metode, da definišu i izmere reprezentativnost uzorka (Liu, 2007, Hanand Kamber, 2006). Glavni problemi u određivanju reprezentativnog uzorka se odnose na procenu njegove stope informacionog pokrivanja originalne populacije (Pan i dr., 2005, Zhai i dr., 2003), i procene njegove interne informacije redundanse (Pan i dr., 2005, Zhang i dr., 2002, Carbonell i Goldstein, 1998, Bertino, 2006, Kubat i Matvin, 1997). Reprezentativan uzorak koji pokriva najviše informacija ciljne populacije treba da ima veliku vrednost uzajamne informacije u odnosu na ciljnu populaciju i malu unutrašnu redundansu. U ovom istraživanju smo koristili stopu pokrivenosti, izraženu kroz Kullback-Leibler divergenciju uzoraka, i redundansu, izraženu kroz entropiju uzoraka, kao mere reprezentativnosti.

5.2.2. Mera stepena presecanja uzorka i populacije

U svrhu boljeg razumavenja pojma stepena presecanja prostora instanci, predstavimo osnovno značenje Kullback-Leibler divergencije poznate još kao relativna entropija. Za dve raspodele verovatnoća $P = P(X) = \{p(x_i) = p_i, \forall x_i \in X\}$ i $Q = P(Y) = \{p(y_i) = q_i, \forall y_i \in Y\}$ definisane nad diskretnim promenljivim X i Y, Kullback-Leibler Divergencija P od Q je definisana kao: $D_{KL}(P||Q) = -\sum_i p_i \log[p_i / q_i]$. Ova divergencija ime sledeće osobine: $D_{KL}(P||Q) \geq 0; D_{KL}(P||Q) = 0 \ \ iff \ \ P(x) = Q(x), \forall x \in X.$ Postoje i sledeće prateće pretpostavke od praktičnog značaja da, kada god je $P(x_i) = 0$ tada efekat $i-$tog člana takođe ima nultu vrednost pošto važi sledeće: $\lim_{x \to 0} x log(x) = 0.$

Razmotrimo dalje diskretne slučajne promenljive X i Y koje imaju zajedničku funkciju gustine verovatnoće odnosno mase verovatnoće $p(x, y)$ i marginalne verovatnoće $p(x)$ i $p(y)$. Možemo definisati uzajamnu informaciju $I(X; Y)$ kao relativnu entropiju između zajedničke

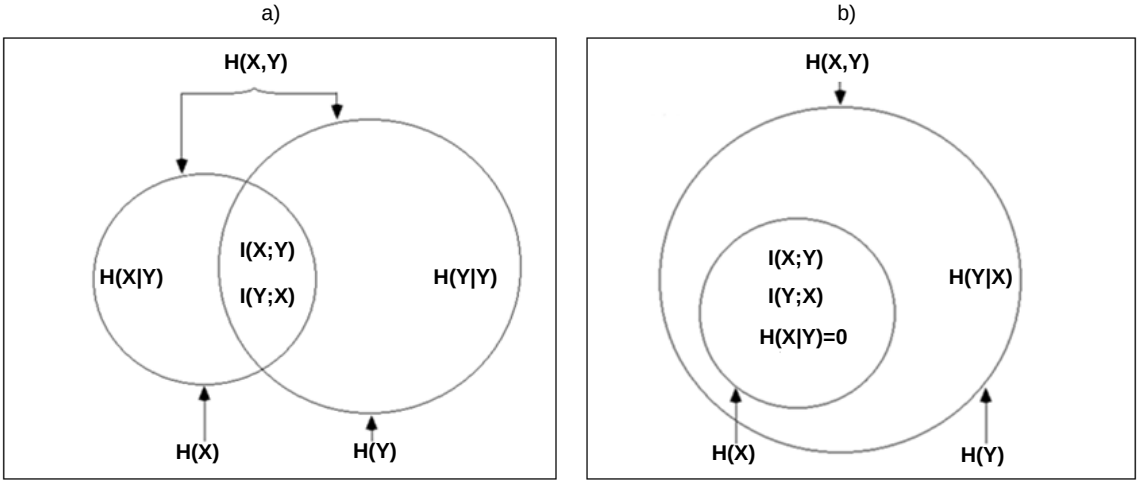
funkcije gustine verovatnoće $p(x,y)$ i proizvoda marginalnih raspodela verovatnoća $p(x)p(y)$. Dakle, $I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$.

Uzajamna informacija se može jednostavno definisati kao mera informacije koju međusobno dele promenljive X i Y odnosno, koliko znanje o jednoj od involviranih variabli redukuje neizvesnost o drugoj variabli. Očigledno je, da međusobno nezavisne varijable ne dele nikakvu zajedničku informaciju ili formalno: $p(x,y) = p(x)p(y) \Rightarrow log\left(\frac{p(x,y)}{p(x)p(y)}\right) = 0 \Rightarrow I(X;Y) = 0$.

Daljim razvijanjem relevantnih jednačina dobijamo sledeće važne relacije:

$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)$, gde $H(X,Y)$ predstavlja zajedničku entropiju za X i Y, $H(X)$ i $H(Y)$ su marginalne entropije a $H(X|Y)$ i $H(Y|X)$ su uslovne entropije.

Posmatrajući prethodne vrednosti i relacije možemo zapaziti njihovu očiglednu analogiju sa unijom, razlikom i presekom dva skupa, kako je to i predstavljeno na Venovom dijagramu sa Slike 5.1 (b). Prema Cover i dr., (1991) i Hastie i dr. (2001), uzajamna informacija $I(X;Y)$ korespondira preseku informacije $H(X)$, sadržanoj u slučajnoj diskretnoj varijabli X, sa informacijom $H(Y)$ sadržanoj u slučajnoj diskretnoj varijabli Y, $I(X;Y) = H(X) \cap H(Y)$ gde $H(\cdot)$ predstavlja Shannonovu entropiju (Shannon, 1976). Zamislimo situaciju gde X ($|X| = N_X$) predstavlja reprezentativni uzorak originalne target populacije Y ($|Y| = N_Y$).



Slika 5.1 Relaciona analogija - skupovi naspram entropija.

Znajući da je reprezentativni uzorak originalne populacije definisan nad prostorom obeležja kao značajno redukovan podskup target populacije Y $(X \subset Y, N_X \ll N_Y)$ lako možemo zaključiti da je vrednost stope prekrivanja (odnosno uzajamne informacije) *I(X;Y)* identična sa vrednošću informacije $H(X)$ sadržane u X, odnosno: $I(X;Y) = H(X) \cap H(Y) = H(X)$ (Fig. 1(b)), čije su granice definisane na sledeći način: $[0 \leq I(X;Y) \leq 1]$, gde je vrednost stope

prekrivanja direktno uslovljena kardinalnom vrednošću $|X|$. Sa ovog aspekta, primarni cilj pri određenju reprezentativnog trening uzorka podrazumeva izbor predefinisanog ograničenog broja instanci X koje na najbolji način pokrivaju prostor instanci odnosno prostor obeležja originalne Y populacije što implicira maksimalnu vrednost uzajamne informacije *I(X;Y)*.

5.2.3. Mera redundanse

U ovom poglavlju (Tabela 5.4., podpoglavlje 5.3.8) koristimo vrednosti Kullback-Leibler divergencije (D_{KL}) između target populacije i odbranih reprezentativnih uzoraka kao meru međusobne stope prekrivanja prostora obeležja. Što je veća stopa prekrivanja to je manja korespondentna vrednost divergencije D_{KL}. Uzmimo da slučajna variabla Y predstavlja originalnu populaciju sa relativno velikim brojem primera $N_Y = |Y|$ i želimo da izolujemo iz Y skupa reprezentativni uzorak predefinisane kardialnosti $N_X = |X|$ koja je po definiciji znatno manja od kardinalnosti originala ($|X| \ll |Y|$), (Slika 5.1(b)). Takođe želimo da izabrani skup X ima manju redundansu (R) od bilo kog skupa X', $X' \subset Y$ iste kardinalnosti $|X'| = |X|$ definisanog na domenu Y. Prema Shannonu (1976), validnost različitih reprezentativnih uzoraka može se predastaviti preko vrednosti redundanse $R(X) = 1 - H(X)/H_{max}(X)$, gde su $H(X)$ i $H_{max}(X)$ entropije aktuelnog konačnog uzorka i maksimalna moguća entropija uzorka respektivno. Apriorna pretpostavka je da uzorak ima maksimalnu vrednost entropije samo u slučaju identične verovatnoće za sve instance koje uzorak sadrži u sebi [$p(x) = 1/N_X, i = 1,2,\dots N_X$)], što implicira uniformnu raspodelu verovatnoće na prostoru obeležja, u kom slučaju entropija ima maksimalnu vrednost: $H(X) = H_{max}(X) = log_2(N_X)$ a redundansa ima minimalnu vrednost $R_{min}(X) = 1 - 1 = 0$. U slučaju potpune izvesnosti o instancma, odnosno nulte entropije uzorka $H(X) = 0$, redundansa postiže maksimalnu vrednost: $R_{max} = 1 - 0 = 1$. Ove činjenice ukazuju na to da svaka devijacija od uniformne raspodele verovatnoće u reprezentativnom uzorku prouzrokuje pad entropije i porast redundanse. Ovi zaključci nam ukazuju na to da naš primarni cilj pri izboru reprezentativnog uzorka predefinisane kardinalnosti treba da bude uzorak sa maksimalnom mogućom entropijom (Jaynes, 1957). Mi težimo ovom cilju preko resamplinga koji podiže nivo uniformnosti raspodele uzorka. Pošto je u realnim uslovima prostor obeležja n –dimenzionalni proctor sa limitiranim volumenom V_X, tada se uniformna raspodela instanci ograničenog skupa X sa N_X tačaka u ovoj oblasti na prihvatljiv način može aproksimirati pomoću n –dimenzionalne pravilne rešetka sa istm brojem N_X tačaka x_i koje predstavljaju centare ćelija u formi kocke konstantnog volumena $v_i = V_X/N_X$ za svaku od $i = 1,2, \dots, N_X$ instanci.

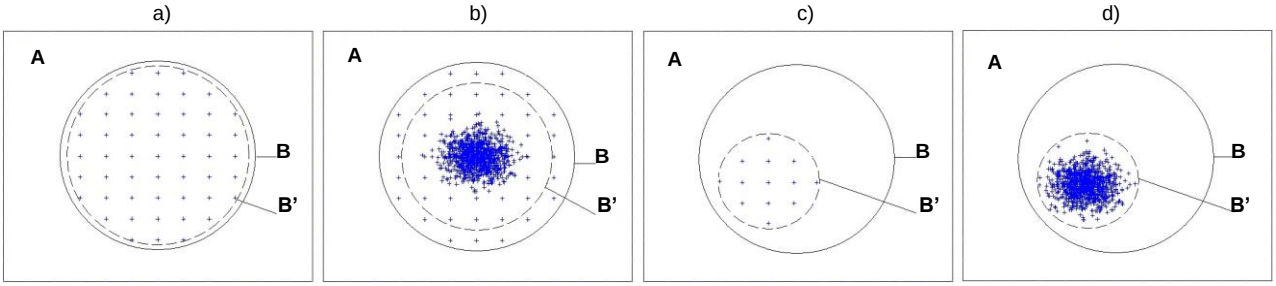
5.2.4. Reprezentativni trening uzorak

Mi ćemo koristiti prethodno izložene činjenice kao osnovu za novu metodu balansiranja neravnoteže unutar klasa. Stoga, pristup reprezentativnosti kao informacionoj kategoriji, na način koji kombinuje stopu pokrivenosti prostora obeležja i redundansu uzorka, izgleda sasvim logičan osnov za naš model balansiranja neravnoteže instanci unutar klasa. Koncept relevantnosti, koji je sinonim pojma reprezentativnosti, se takođe koristi u izboru trening skupa podata, Blum i Langlei (1997). Jedan od glavnih i opštih principa za izbor relevantnih trening skupova podrazumeva izbor uzoraka za koje učeći prediktori imaju veći stepen neizvesnosti, jer su informativniji za primenjene modele (Levis i Gejl, 1994). Ovaj stav podržava našu osnovnu ideju o generaciji izbalansiranog trening uzorka.

Da li su većinske klase su zaista većinske, i da li su manjinske klase zaista manjinske u smislu "dobro reprezentovane klase"? Reprezentativnost uzorka je u direktnoj srazmeri sa njegovom stopom pokrivenosti prostora originalne populacije i u obrnutoj srazmeri sa svojom redundansom. S obzirom na to da se visoka stopa pokrivenosti manifestuje kroz velike vrednosti relativne entropije između reprezentativnog uzorka i originalne populacije, dok se nizak nivo redundantnosti manifestuje kroz visok nivo entropije (neizvesnosti) reprezentativnog uzorka, idealno rešenje za izbor uzorka sa najboljim distributivnim osobina će biti onaj uzorak koji u potpunosti zadovoljava oba ova kriterijuma. Reprezentativni uzorak koji sadrži unapred definisani broj primera N_X (tačaka), sa uniformnom raspodelom verovatnoće u prostoru variabli celog sampling domena, bi zadovoljio oba neophodna kriterijuma pa se zato moramo fokusirati na efikasnu metodu za izbor takvog uzorka. Drugim rečima, veći stepen uniformnosti izabranog uzorka na celom prostoru obeležja, podrazumeva veću stopu pokrivenosti ciljane populacije i veći stepen neizvesnosti izabranih slučajeva za obuku, odnosno veća entropija i smanjena redundansa obučavajućeg uzorka, garantuju smanjenje tendencije klasifikatora da favorizuje dominantnu klasu ili određeni dominantni klaster raspoloživog uzorka.

Relativno visoka gustina instanci u prostoru obeležja pojedinih klasa/oblasti definiše se kao jako zastupljena klasa/oblast. Za ovakve klase/oblasti takođe se pretpostavlja da imaju relativno visoku redundansu. Za relativno niske gustine instanci u pojedinim klasama/oblastima koriste se termini retke klase i retki slučajevi kao ekvivalent koncepta slabo zastupljenih klasa/oblasti. Za ovakve klase/oblasti se obično pretpostavlja da imaju relativno nisku redundansu. Razmotrimo primere manjinske **B** (+) i većinske **A** (-) klase, sa različitim

distribucijama, gde se pokazuje uticaj uniformnosti raspodele na reprezentativnost klasa datih na Slici. 5.2 (a) - (b).



Slika 5.2. Uticaj raspodele instanci na performanse klasifikatora.

Primer na Slici. 5.2 (a) pokazuje raspoloživi uzorak sa relativno malim brojem pozitivnih (**B**) instanci, ravnomerno raspoređenih na celom prostora instanci klase. Predefinisana većinska klasa (**A**), koja je predstavljena dovoljno velikim i dovoljno gustim skupom ravnomerno raspoređenih slučajeva po celom prostoru obeležja klase, a priori zadržava istu distribuciju u svim slučajevima predstavljenim na Slici 5.2 (a) - (d). Krugovi predstavljeni punom linijom predstavljaju stvarnu unapred definisanu graničnu liniju ciljne populacije **B**, koja razdvaja klase **A** i **B**, dok isprekidana linija predstavlja graničnu liniju, procenjenu od strane induktivnog klasifikatora, koji izdvaja hipotetičku klasu **B (B')** iz njenog okruženje. U prvom primeru (Slika 5.2 (a)), klasa **B'** predstavlja procenjenu target klasu **B** i gotovo u potpunosti se poklapa sa target populacijom. To znači da je uzorak **B**, mada sastavljen od malog broja slučajeva dobar reprezent kompletne klase kojoj pripada i omogućava klasifikatoru da pravilno predvidi stvarnu granica između dve populacije **A** i **B**. Ovo je primer trening uzorka sa visokom reprezentativnošću i internim balansom instanci, koji ima malu redundansu i veliki stepen pokrivenosti aktuelnog prostora obeležja koji su posledica ravnomerne distribucije verovatnoće uzorka. Primer prikazan na Slici 5.2 (b) pokazuje nova klasa **B** sa mnogo većim brojem slučajeva, ali neravnomerno raspoređenih u istom prostoru obeležja, gde negativna klasa **A** ima nepromenjen broj slučajeva u odnosu na slučaj predstavljen na Slici 5.2 (a). Ovo je primer obučavajućeg uzorka niskog stepena reprezentativnosti koji ima visoku redundansu i nižu stopu pokrivenosti koje su posledica nerevnomerne respodela verovatnoća odnosno masa. Takođe je očigledno da klasifikator favorizuje negativnu klasu **A** u odnosu na klasu **B**, dajući joj veći prostor u prostoru obeležja, što se manifestuje povlačenjem procenjene granične linije prema gustom centralnom delu **B** klase. Ovi primeri pokazuju primarnu važnost prirode distribucije u odnosu na broj slučajeva iz uzorka i pobije opšti stav o primatu broja slučajeva u neizbalansiranom učenju. Primeri distribucije sa klasama datim na Slikama 2 (c) - (d) pokazuju manjinsku klasu **B** u identičnom okružuju kao pod (a), ograničenu punom linijom (stvarna

granica) u redukovanom prostoru obeležja detrmininsanom raspoloživim sampling okvirom. Treba napomenuti, kao što je očigledno na Slici 5.2, da uzorak klase **B** dat na Slici 5.2 (c) ima mnogo veći broj instanci u odnosu na uzorak dat na Slici 5.2 (g), ali je pomenuti veći broj slučajeva neravnomerno respoređen i koncentrisan na malom delu prostora instanci i ne daje klasifikatoru gotovo nikakve informacije o perifernim delovima prostora obeležja klase kojoj pripadaju. Stoga, klasifikatori izloženi neravnomerno distribuiranim anticipiraju neprihvatljive granice za ciljnu populaciju. Ovi primeri pokazuju rezultate uticaja obučavajućih uzoraka male reprezentativnosti i velikih vrednosti redundaanse na performanse klasifikatora. S druge strane, redukovan skup ravnomerno raspoređenih instanci funkcionalnom smislu potpunosti zamenjuju mnogo veći skup. Ovi primeri pokazuju rezultate uticaja uzoraka za obuku sa visokim stepenom reprezentativnosti i malom vrednošću redundanse na ponašanje klasifikatora. Međutim, u realnim uslovima, skoro bez izuzetka, imamo posla sa uzorcima klase, omeđenih nepoznatim granicama u formi hiper površi u multidimenzionalnom prostoru instanci. Performanse klasifikatora se degradiraju srazmerno povećanju interne neravnoteže u klasama i neravnoteže između klasa gde veći značaj ima interni disbalans. Ovaj stav je relativno lako dokazati prostom analizom distribucije primera klase u prostoru obeležja. Konkretno, visoka koncentracija svih instanci uzorka na malom području prostora obeležja datog na Slici 5.2(b) rezultuje veoma malom reprezentativnošću proporcionalnoj relativnoj veličini zapremine koju zauzima uzorak. Tako, kada volumen V_X zauzet od strane trening uzorka X kardinalnosti $|X| = N_X$, predstavlja mali deo ukupne zapremine prostora V_Y okupiranog od strane ciljne populacije $(V_X/V_Y \to 0)$ onda njegova entropija $H_x \to 0$, rezultujući minimalnom pokrivenosti i maksimalnom redundansom, odnosno minimalnim nivoom reprezentativnosti. S druge strane, ako je obučavajući uzorak X ravnomerno raspoređen po celoj zapremini prostora V_Y ciljne populacije $(V_X/V_Y \to 1)$, tada njegova entropija ima maksimalnu vrednost $H_X = H_{Xmax} = log_2(N_X)$ što rezultuje maksimalnom stopom pokrivenosti prostora obeležja i minimalnom redundansom, odnosno maksimalnim nivoom reprezentativnosti. U narednom poglavlju predstavljena je naša nova metoda za uravnoteženje klasa u svetlu povećanja reprezentativnost obučavajućeg uzorka.

5.3. Metod balansiranja baziran na lokalnim rastojanjima instanci (DBB)

U realnim situacijama, neuniformna distribucija instanci unutar klase (uzorka) se očekuje sa velikom verovatnoćom, i samo je pitanje stepena naglašenosti te verovatnoće. Krajnji cilj naše metode balansiranja je transformacija datog neuniformno distribuiranog, neizbalansiranog

uzorka u smeru idealnog, uniformno distribuiranog uzorka. Na osnovu idealnih distributivnih karakteristika, usvojili smo uzorak u formi pravilne celobrojne rešetke, kao idealanu paradigmu uniformnosti, reprezentativnosti i internog balansa instanci. Nastojaćemu da prenesemo pomenute distributivmne karakteristika pravilne rešetke na realne neizbalansirane uzorke sa proizvoljnom neravnomernom distribucijom instanci u cilju njihovog balansiranja. Dimenzije rešetke čije karakteristike prenosimo na proizvoljni realni uzorak determinisane su, u smislu broja instanci i broja obeležja, dimenzijama originala koji balansiramo. Posebno smo zainteresovani za odnos volumena koji pripada svakom pojedinačnom primeru i njegove sredinje lokalne udaljenosti od prethodno definisanog broja suseda (odnos zapremina/udaljenost). Ovaj odnos predstavlja kondenzovanu formu distributivnih karakteristika rešetkastog uzorka i zato ga treba najpre definisati a zatim naći način za njegov prenos na proizvoljni neizbalansirani uzorak. Pošto je pravilnost reaspodele instanci u rešetkastom uzorku očigledna, jasno je da sve tačke rešetke zauzimaju identičan prostorni volumen, pa je ovaj volumen lako izračunati, dok smo za izračunavanje srednjih lokalnih distance za proizvoljne instance rešetke razvili originalnu metodologiju. Procedura za izračunavanje i prenos odnosa zapremina/distanca je prikazana detaljno u podpoglavljima (5.3.1 do 5.3.4). Ova transformacija je izvedena korišćenjem dve originalne resampling tehnike: modifikovani stratifikovani undersampling uzoraka u oblastima prostora obeležjaa velike gustine verovatnoće, i modifikovani SMOTE sintetički oversampling u oblastima prostora obeležja male gustine. Veliki broj postojećih, specifičnih i parcijalnih rešenja ovog problema, predstavljaju odličnu osnovu u smislu spektra ideja za rešenje sa većim stepenom opštosti i pouzdanosti. Zadržaćemo se na balansiraanju na nivou podataka odnosno na resampling tehnikama pri kreiranju našeg resampling metoda balansiranja (DBB) zasnovanog na distancama koji omogućava značajno povećanje stepena uniformnosti raspodele instanci unutar originalnih neuravnoteženih uzoraka. Osnovna i jednostavna zamisao ovog metoda balansiranja je eliminacija određenog broja instanci iz oblasti domena visoke gustine verovatnoće podržane sintetičkom generacijom potrebnog broja novih instanci u oblastima domena sa niskom gustinom verovatnoće.

Ako imamo uzorak sa unapred definisanim brojem slučajeva (N) ravnomerno raspoređenih u n-dimenzionalnom prostora obeležja, jasno je da će gustina instanci u tom prostoru brzo opada sa porastom broja n a Euklidsko rastojanje između instanci će rapidno rasti. Da bismo uklonili instance iz gustih oblasti domena, bilo bi logično koristiti stratifikovani random resampling u prostoru obeležja. Ova standardna statistička procedura podrazumeva podelu originalne populacije u određeni broj manjih homogenih podgrupa (podprostora), a zatim

slučajni izbor malih uzorka iz svake podgrupe. Metod ima smisla u slučaju malog broja n relevantnih obeležja, ali u realnim situacijama n može imati velike vrednosti (n > 10000) ili više. Ova činjenica predstavlja nepremostivu prepreku operativnim algoritmima zbog ogromnog broja (h^n) mogućih podprostora dobijenih segmentacijom svih obeležja (varijabli) na h segmenata. Svaki od (h^n) subprostora treba da bude pretražen a samo relatrivno mali broj njih sadrži određeni broj instanci koje treba prebrojati, izbaciti ili sintetizovati nove, što zahteva procedure sa računarskom kompleksnošću $O(h^n)$ koja je ogromna o najvećem broju realnih slučajeva. Eto zašto tražimo indirektno rešenje zasnovano na korespondenciji zapremina/distanca. Predloženo rešenje, predstavljeno u sledećim podpoglavljima, operiše u jednodimenzionalnom prostoru rastojanja sa računarskom kompleksnošću $O(N)$ omogućavajući neuporedinvo kraće operativno vreme. Formula za izračunavanje Euklidovog rastojanja između instanci $x_i = \{f_{i1}, f_{i2}, \dots, f_{in}\}$ i $x_k = \{f_{k1}, f_{k2}, \dots, f_{kn}\}$ koje su definisane skupom vektora obeležja dužine n je predstavljena jednačinom 1.

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{n} (f_{i,j} - f_{k,j})^2} \qquad (1)$$

Predloženi DBB metod se koristi za selekciju uzorka instanci iz raspoloživog okvira na način koji garantuje da će broj instanci u svim subprostorima domena biti proporcionalan zapreminama tih subprostora čime se postiže raspodela verovatnoće slična uniformnoj raspodeli (kvaziuniformna) po kompletnom raspoloživom domenu. Ova procedura se sprovodi odvojeno za svaku klasu.

5.3.2. Detekcija relacije volumen-distanca u celobrojnoj pravilnoj rešetki

Kako smo već rtekli, usvojili smo konačni uzorak instanci raspoređenih u pravilnu celobrojnu rešetku kao idealnu paradigmu uniformnosti, reprezentativnossti i balansa. U cilju definicije korespondencije volumen-distanca, najpre ćemo uspostaviti korespondenciju između skupa eksperimenata izvedenih na tročlanom skupu $E = \{-1,0,1\}$ i n -dimenzionalne celobrojna pravilne rešetka sa ograničenim koordinatama $(\mu, \nu, \dots)$, gde $\mu, \nu, \dots \in [-1,0,1]$ predstavljaju tročlane celobrojne skupove. Razmotrimo eksperimente na tročlanom skupu $E = \{-1,0,1\}$ celobrojnih vrednosti odnosno $E = \{E_1, E_2, E_3\}$ gde su elementarni događaji $E_1 = \{-1\}, E_2 = \{0\}, E_3 = \{1\}$ međusobno disjunktivni i imaju identične verovatnoće.

Kada se izvodi samo jedan eksperiment ($n = 1$) nad skupom E, tada je reprezent prostora elementarnih događaja definisan kao $\{E_i, i = 1, 2, 3\}$. Prostor Ω_1 elementarnih

događaja $\{E_i,\ i=1,2,3\}$, čiji članovi imaju identične verovatnoće realizacije $\{P(E_i)=1/3,\ i=1,2,3\}$ je definisan u formi sledećeg jednodimanzionalnog niza

$$\Omega_1=\{E_1,E_2,\ E_3\}=\{-1,0,1\}. \qquad (2)$$

Kardinalana vrednost $|\Omega_1|$ predstavlja broj elemenata skupa E, $(|\Omega_1|=|E|=3)$.

U slučaju dva nazavisna eksperimenta $(n=2)$ izvedenih nad skupom E, reprezent prostora elementarnih događaja Ω_2 je definisan u formi uređenog para $(E_i,E_j), i,j=1,2,3$, iz Dekartovog proizvoda $E_1\times E_2=\{(E_i,\ E_j)|E_i\in E\wedge E_j\in E,\ i,\ j=1,2,3\}$, gde je proctor Ω_2 dat u formi Jed. (3), čiji članovi takođe imaju identične vrednossti verovatnoće realizacije $\{P(E_i,E_j)=1/3^2=1/9, i,j=1,2,3\}$.

$$\Omega_2=\begin{Bmatrix}(E_1,E_1),(E_2,E_1),(E_3,E_1)\\(E_1,E_2),(E_2,E_2),(E_3,E_2)\\(E_1,E_3),(E_2,E_3),(E_3,E_3)\end{Bmatrix}=\begin{Bmatrix}(-1,-1),(0,-1),(1,-1)\\(-1,\ \ 0),(0,\ \ 0),(1,\ \ 0)\\(-1,\ \ 1),(0,\ \ 1),(1,\ \ 1)\end{Bmatrix}. \qquad (3)$$

Kardinalna vrednost ovog skupa je $|\Omega_2|=|E\times E|=|3\times 3|=3^n=3^2$. Ovaj proces odgovara 2- dimenzionalnoj celobrojnoj kvadratnoj rešetki (Slika 5.3).

U slučaju tri nazavisna eksperimenta $(n=3)$ izvedenih nad skupom E, represent prostora elementarnih događaja Ω_3 je definisan u formi uređene trojke $(E_i,E_j,E_k), i,j,k=1,2,3.$, iz Dekartovog proizvoda $E\times E\times E=\{(E_i,E_j,E_k)|E_i\in E,E_j\in E,E_k\in E,i,j=1,2,3\}$. Prostor Ω_3 je dat u formi Jed. (4), čiji članovi takođe imaju identične vrednosti verovatnoće $\{P(E_i,E_j,E_k)=1/3^3=1/27,i,j,k=1,2,3\}$.

$$\Omega_3=\begin{Bmatrix}(E_1,E_1,E_1),(E_2,E_1,E_1),(E_3,E_1,E_1)\\(E_1,E_1,E_2),(E_2,E_1,E_2),(E_3,E_1,E_2)\\(E_1,E_1,E_3),(E_2,E_1,E_3),(E_3,E_1,E_3)\\(E_1,E_2,E_1),(E_2,E_2,E_1),(E_3,E_2,E_1)\\(E_1,E_2,E_2),(E_2,E_2,E_2),(E_3,E_2,E_2)\\(E_1,E_2,E_3),(E_2,E_2,E_3),(E_3,E_2,E_3)\\(E_1,E_3,E_1),(E_2,E_3,E_1),(E_3,E_3,E_1)\\(E_1,E_3,E_2),(E_2,E_3,E_2),(E_3,E_3,E_2)\\(E_1,E_3,E_3),(E_2,E_3,E_3),(E_3,E_3,E_3)\end{Bmatrix}\begin{Bmatrix}(-1,-1,-1),(0,-1,-1),(1,-1,-1)\\(-1,-1,\ \ 0),(0,-1,\ \ 0),(1,-1,\ \ 0)\\(-1,-1,\ \ 1),(0,-1,\ \ 1),(1,-1,\ \ 1)\\(-1,\ \ 0,-1),(0,\ \ 0,-1),(1,\ \ 0,-1)\\(-1,\ \ 0,\ \ 0),(0,\ \ 0,\ \ 0),(1,\ \ 0,\ \ 0)\\(-1,\ \ 0,\ \ 1),(0,\ \ 0,\ \ 1),(1,\ \ 0,\ \ 1)\\(-1,\ \ 1,-1),(0,\ \ 1,-1),(1,-1,-1)\\(-1,\ \ 1,\ \ 0),(0,\ \ 1,\ \ 0),(1,\ \ 0,\ \ 0)\\(-1,\ \ 1,\ \ 1),(0,\ \ 1,\ \ 1),(1,\ \ 1,\ \ 1)\end{Bmatrix} \qquad (4)$$

Kardinalna vrednost ovog skupa je $|\Omega_3|=|E\times E\times E|=|3\times 3\times 3|=3^n=3^3$ Ovaj proces odgovara 3- dimenzionalnoj celobrojnoj kubnoj rešetki.

Shodno prethodnim činjenicama, za opšti slučaj od n eksperimenata nad E, opšti član iz skupa elemtarnih događaja je definisan kao uređena n-torka $\underbrace{(E_i\times E_j\times E_k\ldots\times E_w)}_{n}$, Dekartovog proizvoda $\underbrace{E\times E\times E\ldots\times E}_{n}=\{(E_i,E_j,E_k,\ldots,E_w)|E_i\in E,E_j\in E,E_k\in E,\ldots,E_w\in E,i,j,k,\ldots,w=1,2,3.\}$ a sam prostor događaja Ω_n kao skup čiji članovi imaju jednaku

verovatnoću realizacije $P(\underbrace{E_i, E_j, E_k, \dots, E_w}_{n}) = 1/3^n, i, j, k, \dots, w = 1, 2, 3.$ Kardinalna vrednost ovog skupa data jednačinom (5) je:

$$|\Omega_n| = \left|\underbrace{E \times E \times E \dots \times E}_{n}\right| = \left|\underbrace{3 \times 3 \times 3 \dots \times 3}_{n}\right| = 3^n \quad (5)$$

Ovakav proces reprezentuje nD uniformnu pravilnu rešetku čiji je 2D slučaj prikazan na Slici. 5.3. Analizom ovog 2D slučaja dolazimo do bitnih opštih zaključaka o odnosu srednjih vrednosti rastojanja tačaka od tačaka iz okruženja i njima pripadajuće lokalne zapremine u prostoru instanci odnosno prostoru obeležja generisanih proizvoljnim tipovima raspodela.

Vratimo se eksperimentima nad tročlanim celobrojnim skupom tačaka $E = \{-1,0,1\}$ odnosno $E = \{E_1, E_2, E_3\}$ gde su elementarni događaji $E_1 = \{-1\}, E_2 = \{0\}, E_3 = \{1\}$ disjunktni i sa identičnim verovatnoćama $\{P(E_i) = 1/3, i = 1, 2, 3\}$. Definišimo karakteristični događaj A nad skupom E i njemu komplemetarni događaj B, tokom n nezavisnih eksperimentima na sledeći način:

$A = \{E_1, E_3\} \Leftrightarrow A = \{E_1\} \vee \{E_3\}, B = \{E_2\}\ ili\ A = \{-1,\ 1\} \Leftrightarrow A = \{-1\} \vee \{1\}, B = \{0\}$. U svakom od n nezavisnih eksperimenata, verovatnoća uspešne realizacije događaja A je konstantna, $p = P(A) = 2/3$ a verovatnoća da se događaj A ne realizuje je $q = 1 - P(A) = 1/3$. Svaki od n ovakvih eksperimenata predstavlja Bernulujev eksperiment i može se posmatrati kao bacanje kocke nepravilene verovatnoće. Shodno prethodnim činjenicama, za komplementarni događaj B je $p_b = P(B) = 1/3$ i $q_b = 1 - P(B) = 2/3$. Na osnovu prikazanih karakteristika događaja A i činjenice da se on u n eksperimenata može pojaviti r puta, $r = \{0,1,2, \dots, n\}$, očigledno je raspodela verovatnoća ovih događaja definisana zakonom binomne raspodele verovatnoća prikazane sledećom formulom:

$$P(X = r) = \binom{n}{r} p^r q^{n-r}. \quad (6)$$

Gde je $X = \{0,1,2, \dots, n\}$ diskretna slučajna promenljiva čije komponente predstavljaju broj uspešnih realizacija događaja A u n nezavisnih eksperimenata. Vrednosti verovatnoća $p = 2/3$ i $q = 1 - p = 1/3$ ostaju konstantne tokom svakog od n eksperimenata, a sleedeći izraz, $\binom{n}{r} = \frac{n!}{[r!(n-r)!]}$, predstavlja kondenzovanu formu binomnih koeficijenata. Uspostavimo idealnu korespondenciju između domena verovarnoća (P) realizacija događaja X nad nekim prostorom događaja $\Omega_n, \{P(X = r), X \epsilon \Omega_n\}$ i domena samih realizacija (R_l) događaja X na sledeći način:

$$R_l(X = r) = P(X = r)|\Omega_n|, \quad (7)$$

Gde je $|\Omega_n| = \underbrace{|E \times E \times E \ldots \times E|}_{n} = E^n$, prostor događaja nezavisnih eksperimenata na tročlanom skupu $E = \{E_1, E_2, E_3\}$, odnosno n –tostruki Dekartov proizvod nad skupom E koji predstavlja hiperkocku u formi celobrojne rešetke. Veličina $|\Omega_n|$ je kardinalna vrednost od Ω_n. Kompozicijom jednačine za binarnu raspodelu (6) i relacija (7), obzirom da je $|\Omega_n| = 3^n$ videti (5), $p = 2/3$ i $q = 1/3$, dobijamo elegantne jednačine (8) i (9) za traženu korespondenciju:

$$P(X = r) = C_n^r p^r q^{n-r} = \binom{n}{r}\left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{n-r} = \binom{n}{r}\frac{2^r}{3^n}. \quad (8)$$

$$R_l(X = r) = P(X = r)|\Omega_n| = \binom{n}{r}\frac{2^r}{3^n}3^n = 2^r\binom{n}{r}. \quad (9)$$

U Tabeli 5.1 su date raspodele verovatnoća događaja A, $P(X = r)$ i samih realizacija $R_l(X = r)$ događaja A za n eksperimenata na skupu E, gde je $n = 1,2,3,4,5.$ i $r = 0,1,2,3,4,5.$ Na Slici 5.3 dat je prikaz dvodimenzionalne kvadratne celobrojne rešetke koja se može opisati kao sledeća uniformno distribuirana ograničena matrica tačaka: $C = c_i(x_i, y_i)|x_i \in E, y_i \in E, i = 1,2,\ldots,9\}$, odnosno kao skup uređenih parova ($C \in Z^2$) u dvodimenzionalnom celobrojnom prostoru Z^2. Skup C se može smatrati Dekartovim proizvodom $x \times y$, gde je $x = E = \{-1,0,1\}, y = E = \{-1,0,1\}$. Kardinalne vrednosti skupova x i y su $|x| = |y| = 3$. Kardinalnost skupa C, odnosno proivoda $|x \times y|$ je $|C| = |x \times y| = 3 \times 3 = 3^2$. Ovakva kompozicija tačaka predstavlja pravilnu kvadratnu ($2D$) celobrojnu rešetku tačaka. Svaka tačka $C_i, i = 1,2,\ldots,9.$, (Slika 5.3) je pozicionirana u centar gravitacije kvadrata s_i^2, jedinične stranice $a_i = 1$, i jediničnog volumena $v_i = |s_i^2| = a_i^2 = 1, i = 1,2,\ldots,9$. U slučaju $i = 5$, taška c_5 predstavlja centar gravitacije kvadrata s_5^2 definisanog koordinatama temena $\{(-0.5,-0.5), (-0.5,0.5), (0.5,0.5), (0.5,-0.5)\}$.

Tabela 5.1
Verovatnoće $P(X = r)$ i odgovarajuće realizacije $R_l(X = r), r = 0,1,\ldots,5.$

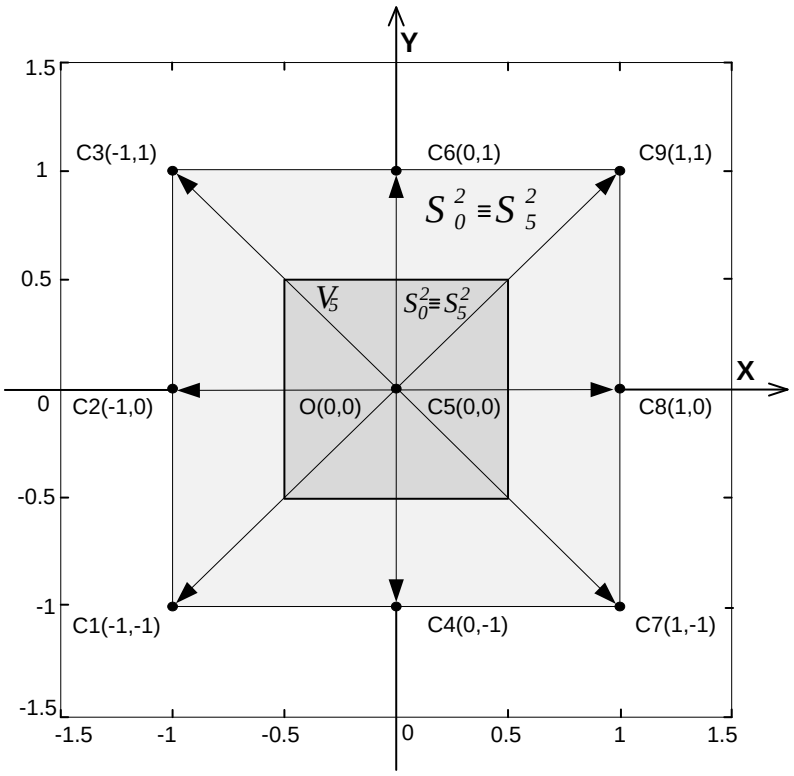
n \ r		0	1	2	3	4	5	$\|\Omega_n\| = 3^n$
$n = 1$	$P(X = r)$	1/3	2/3	0	0	0	0	3
	$R_l(X = r)$	1	2	0	0	0	0	
$n = 2$	$P(X = r)$	1/9	4/9	4/9	0	0	0	9
	$R_l(X = r)$	1	4	4	0	0	0	
$n = 3$	$P(X = r)$	1/27	6/27	12/27	8/27	0	0	27
	$R_l(X = r)$	1	6	12	8	0	0	
$n = 4$	$P(X = r)$	1/81	8/81	24/81	32/81	16/81	0	81
	$R_l(X = r)$	1	8	24	32	16	0	
$n = 5$	$P(X = r)$	1	10/243	40/243	80/243	80/243	32/243	243
	$R_l(X = r)$	1	10	40	80	80	32	

Zapremina kvadrata s_5^2 je $v_5 = s_5^2 = a^2 = 1$. U slučaju dvodimenzionalnoe kvadratne rešetke (Slika 5.3) možemo smatrati da svaka tačka zauzima identičan jedinični volumen odnosno

površinu v_i. Koordinatni početak $O(0,0)$ je referentna tačka podudarna sa tačkom $c_5(0,0)$, u prikazanoj shemi, za koju treba da definšemo srednje Euklidsko rastojanje od svih njenih datih suseda uključujući i samu tačku $c_5(0,0)$. Sasvim je jasno sa Slike 5.3 da tačke $c_i, i = 1,2, \dots ,9$. Predstavljaju temena četiri susedna kvadrata čije baze i volumeni imaju jedinične vrednosti a tačka $O(0,0)$ predstavlja njihovo zajedničko teme (zasenčena oblast). Temena c_1, c_3, c_9, c_7 zasenčene oblasti (S_0^2) na Slici 5.3 formiraju dvodimenzionalnu hiperkocku čija je baza $a_{s_0} = 2$, a volumen $v_{s_0} = a_{s_0}^2 = 2^2$ sa težištem u koordinatnom početku O. Prostor ograničen sa S_0^2 definišemo kao „S_0^2 okruženje" tačke O, dok sve tačka iz tog prosora $c \in S_0^2, i = 1,2, \dots, 3^2$ definišemo kao susede tačke O. Analogno prethodnoj definiciji, za opšti slučaj, n-dimenzionalne celobrojne rešetke ograničene hiperkockomS_i^n definišemo „S_i^n okruženje" i –te tačke rešetke $c_i \underbrace{(x_i, y_i, \dots, w_i)}_{n}, i = 1,2, \dots, n.$ Tačke $c_i \in S_i^n, i = 1,2, .., 3^n$ iz S_i^n okruženja c_i označavamo susedima i-te tačke a ove susedne tačke sačinjavaju celobrojnu rešetku $C = \{c_i \underbrace{(x_i, y_i, , \dots, w_i)}_{n} | x_i \in E,\ y_i \in E, z_i \in E, \dots, w_i \in E,\ c_i \in S_i^n, i = 1, 2, \dots, 3^n\}$. Sku C je Dekartov proizvod $\underbrace{x \times y \times z \times \dots \times w}_{n}, x = y = z = w = E, E = \{-1,0,1\}$.

Kardinalne vrednosti skupova $x, y, z, \dots, w$ su očigledno$|x| = |y| = |z| =, \dots, |w| = 3$.

Kardinalna vrednost od C jednaka je kardinalnoj vrednosti pomenutog Dekartovog proizvoda, $|C| = \left|\underbrace{x \times y \times \dots \times w}_{n}\right| = \underbrace{3 \times 3 \times 3 \times \dots \times 3}_{n} = 3^n$. Osnovni cilj u ovom delu istraživanja je defnicija odnosa između datog volumena v_i zauzetog od strane proizvoljne instance c_i iz n - dimenzionalne celobrojne pravilne rešetke, i njene srednje Euklidske distance d_{ain}, indeks *"$_a$"* znači srednji, *"$_i$"* se odnosi na i - tu tačku i *"$_n$"* se odnosi na dimenziju rešetke, od svih K_i^n susednih tačaka iz njenog S_i^n okruženja. Da bismo ispunili ovaj zahtev, treba odrediti tačan broj K_i^n tačaka u S_i^n okruženju i sve vrednosti njihovih koordinata. Za ovu svrhu, koristićemo dvodimenzionalnu celobrojnu rešetku $C \subset Z^2$ prikazanu na Slici 5.3. Izračunate vrednosti će biti predstavljene u opštoj formi koja je primenljiva na $n-$ dimenzionalni slučaj celobrojne rešetke $C \subset Z^n$.



Slika 5.3. $2D$ celobrojna rešetka (3×3).

Prikazani dvodimenzionalni slučaj rešetke omogućava dobru preglednost uspostavljenih relacija kao i pomenutu mogućnost generalizacije. Opšta jednačina Euklidskog rastojanja (1) za sve tačke iz S_0^2 okruženja od koordinatnog početka $O(0,0)$ sa Sl 3 je: $d(O,c_i) = \sqrt{(x_i-0)^2+(y_i-0)^2} = \sqrt{x_i^2+y_i^2}, i = 1,2,..,9$. Treba obratiti pažnju na podudarnost tačaka O i c_5, $d(c_5,c_i) = d(O,c_i), i = 1,2,..,9$. Očigledno je da postoje tri grupe susednih tačaka, koje su ekvidistantne u odnosu na tačku O u okviru svake od grupa u smislu Euklidskog rastojanja. Definišimo kao nultu grupu $Grupa_0$ skup $k_0 = 1$, ekvidistantnih tačaka c_i sa nultim Euklidskim rastojanjem od tačke $O, d_0 = d(O,c_i) = \sqrt{0}, i = 5$. Dakle ovaj skup čini samo jedna tačka koja je podudarna sa kordinatnim početkom $c_5 \equiv O$. Rastojanje d_0 smatramo rastojanjem *nultog reda* a k_0 brojem *suseda nultog reda*. $Grupa_1$ predstavlja skup od $k_1 = 4$ skvidistantnih tačaka c_i sa jediničnim Euklidskim rastojanjem od tačke O, $d_1 = d(O,c_i) = \sqrt{1}, i = 2,4,6,8$. Dakle ovu grupu čine tečke c_2, c_4, c_6, c_8. Rastojanje d_1 smatramo *rastojanjem prvog reda* a k_1 smatramo brojem *suseda prvog reda*. $Grupa_2$ predstavlja skup od $k_2 = 4$ ekvidistantnih tačaka c_i sa Euklid. rastojanjem u odnosu na koordinatni početak datim kao: $d_2 = d(O,c_i) = \sqrt{2}, i = 1,3,7,9$. Rastojanje d_2 je rastojanje *drugog reda* a k_2 je broj *suseda drugog reda*. Ukupan broj susednih tačaka iz S_0^2 okruženja je $K_0^2 = \sum_{r=1}^{2} k_r = 4+4 = 3^2-1 = 8$ (Slika 5.3). Pošto je broj relevantnih susednih tačaka poznat a takođe i njihove

koordinate, moguće je odrediti njihova rastojanja od referentne tačke kao i srednje rastojanje tačaka iz S_0^2 okruženja od referentne tačke na sledeći način: $d_{a02} = \sum_{r=1}^{2} k_r d_r / \sum_{r=1}^{2} k_r = (4\sqrt{1} + 4\sqrt{2})/8 = (1 + \sqrt{2})/2$. $Grupa_0$ $(k_0 = 1, d_0 = 0)$ je isključena zbog nulte distance jer ne može biti računata kao sused samoj sebi. Mala vrednost dimenzija prostora celobrojne rešetke $(n = 2)$ daje mogućnost direktnog uvida u pozicije svih tačaka (Slika 5.3) i omogućavaju nam jednostavno određivanje srednjeg rastojanja d_{a02}, dok uočene zakonitosti raspodele tačaka unutar rešetke treba da nam omoguće određivanje srednjeg rastojanja za opšti slučaj $(n > 3)$. Jednostavno određivanje rastojanja d_{a01} u slučaju jednog eksperimenta nad skupom E, $(n = 1)$, odnosno u slučaju jednodimenzionalne celobrojne rešetke, gde imamo jednodimenzionalnu referentnu tačku $O(0)$ i dve najbliže susedne tačke $c_1 = (+1)$ and $c_2 = (-1)$ na x osi. U ovom slučaju je broj svih suseda iz S_0^1okruženja dat kao $K_0^1 = \sum_{r=1}^{1} k_r = 1 + 1 = 3^1 - 1 = 2$, tako da je $d_{a01} = \sum_{r=1}^{1} k_r d_r / \sum_{r=1}^{1} k_r = (\sqrt{1} + \sqrt{1})/2 = 1$. Slučaj tri eksperimenta nad $E, (n = 3)$ ima svoj geometrijski ekvivalent u formi trodimenzionalne rešetke, gde takođe imamo mogućnost direktnog uvida u pozicije svih tačaka jednostavno izračunavanje broja i rastojanja svih tačaka iz S_0^3 okruženja. Broj svih suseda iz S_0^3 okruženja je $K_0^3 = \sum_{r=1}^{3} k_r = 6 + 12 + 8 = 3^3 - 1 = 26$ a srednje rastojanje d_{a03} je: $d_{a03} = \sum_{r=1}^{3} k_r d_r / \sum_{r=1}^{3} k_r = (6\sqrt{1} + 12\sqrt{2} + 8\sqrt{3})/26 = (3 + 6\sqrt{2} + 4\sqrt{3})/13$. Kada je broj eksperimenata nad skupom E, $n > 3$, ekvivalentni Dekartov proizvod E^n generiše celobrojnu rešetku koja ne dozvoljava jednostavan vizuelni pristup svim tačkama iz S_0^n okruženja tačke O. Zato koristimo observabilni slučaj E^2 sa Slike 5.3 kako bismo uspostavili pouzdan metod za odrđivanje individualnih rastojanja i aktuelnoh srednjeg rastojanja d_{a0n} u višedimenzionalnom prostoru. Broj svih suseda iz n-dimenzionalnog S_0^n okruženja i rastojanje d_{a0n} su dati u opštoj formi Jed. (10) čiji elementi zahtevaju dalju analizu i objašnjenje.

$$K_0^n = \sum_{r=1}^{n} k_r = 3^n - 1 \; ; \qquad (10)$$

$$d_{a0n} = \sum_{r=1}^{n} k_r d_r / \sum_{r=1}^{n} k_r . \qquad (11)$$

Rastojanja d_r iz Jed. (10) označavamo rastojanjima $r - tog\ reda$ a k_r su brojevi *suseda* $r - tog\ reda$. Napomenimo da zaključci izvedeni za referentnu tačku O iz neograničenog celobrojnog prostora Z^n imaju opšti značaj za svaku proizvoljno odabranu i-tu tačku tog prostora. Ovaj stav se lako dokazuje jednostavnom translacijom proizvoljne tačke c_i u poziciju koordinatnog početka. Očigledno je da pomenuta substitucija odnosno translacija ne menja uslove u okruženju tačaka celobrojne rešetke.

Takođe treba napomenuti da svi ovde izvedeni zaključci, za celobrojnu pravilnu kubnu rešetku čija jedinična ćelija ima stranicu $a = 1$, mogu biti u svakom slučaju kubne rešetke čije ćelije

imaju stranice proizvoljne dužine $a \in]0,1[$. U cilju određivanja opšteg izraza za predloženi odnos $volumen/rastojanje$ u n-dimenzionalnom prostoru, uspostavimo najpre formalnu korespondenciju između uspešnih realizacija ekskluzivnih događaja $A = \{-1, 1\}$ i $\bar{A} = \{0\}$ i vrednosti koordinata $\{x_i, y_i\}$ svih tačaka $C = c_i(x_i, y_i), c_i \in E^2, i = 1,2,..,9$ rešetke (Slika 5.3).

Tabela 5.2
Korespondencija distributivnih karakteristika 2D rešetk I raspodele događaja $A = \{-1, 1\}$ i $\bar{A} = \{0\}$.

$c_i(x_i,\ y_i)$	Ω^2	$A_{xi} = \lvert A \cap x_i$	$A_{yi} = \lvert A \cap y_i \rvert$	$A_{xi} + A_{yi}$	$x_i^2 + y_i^2$	r	$d_r = \sqrt{r}$	$Grupa_r$
$c_1(-1,-1)$	$\{-1,-1\}$	1	1	2	2	2	$\sqrt{2}$	$Grupa_2$
$c_2(-1,\ 0)$	$\{-1,\ 0\}$	1	0	1	1	1	1	$Grupa_1$
$c_3(-1,\ 1)$	$\{-1,\ 1\}$	1	1	2	2	2	$\sqrt{2}$	$Grupa_2$
$c_4(\ 0,-1)$	$\{\ 0,-1\}$	0	1	1	1	1	1	$Grupa_1$
$c_5(\ 0,\ 0)$	$\{\ 0,\ 0\}$	0	0	0	0	0	0	$Grupa_0$
$c_6(\ 0,\ 1)$	$\{\ 0,\ 1\}$	0	1	1	1	1	1	$Grupa_1$
$c_7(\ 1,-1)$	$\{\ 1,-1\}$	1	1	2	2	2	$\sqrt{2}$	$Grupa_2$
$c_8(\ 1,\ 0)$	$\{\ 1,\ 0\}$	1	0	1	1	1	1	$Grupa_1$
$c_9(\ 1,\ 1)$	$\{\ 1,\ 1\}$	1	1	2	2	2	$\sqrt{2}$	$Grupa_2$

Posmatrajući Tabelu 5.2 i Sliku 5.3 zapažamo da prva kolona u Tabeli 5.2 prikazuje coordinate tačaka pravilne rešetke $c_i(x_i, y_i)$ u (x, y) ravni (Slika 5.3). Druga kolona predstavlja prostor događaja Ω^2 prikazan u formi matrice u Jednačini (3), tako da je $\Omega^2 \equiv C^2$. Treća I četvrta kolona predstavljaju broj realizacija događaja A u vrednostima koordinata x_i i y_i i-te tačke c_i rešetke $(A_{xi} = |A \cap x_i|, A_{yi} = |A \cap y_i|)$. Peta kolona predstavlja kompilaciju druge I treće kolone. Šesta kolona predstavlja sumu kvadrata koordinata datih u prvoj koloni. Sedma kolona predstavlja broj realizacija (r) događaja A u svakoj tački $c_i(x_i, y_i)$ iz prve kolone. Osma kolona sadrži vrednosti Euklidskih rastojanja između tačke $O(0,0)$ i tačaka $c_i(x_i, y_i)$ iz prve kolone. Deveta kolona prikazuje pripadnost tačaka $c_i(x_i, y_i)$ odgovarajućim grupama. Grupe su indeksirane indeksom r ($r = \{0,1,2\}$). Jednačina (12) uspostavlja relaciju između pete, šeste i sedme kolone. Očigledno je da broj r, $r = \{0,1,2\}$ uspešnih realizacija događaja A u prostoru koordinata $(x_i,\ y_i)$ tačke $c_i, c_i \in Grupa_r$ determiniše distance d_r tačaka te grupe od koordinatnog početka O (Tabela 5.2) i jed. (13).

$$r = \sqrt{(x_i - 0)^2 + (y_i - 0)^2} = A_{xi} + A_{yi} = x_i^2 + y_i^2, c_i \in Grupa_r, \quad (12)$$

$$d_r = \sqrt{x_i^2 + y_i^2} = \sqrt{r}. \quad (13)$$

Kolona devet iz Tabele 5.2 prikazuje skup grupa $S_G = \{Grupa_r, r = 0,1,2.\}$, gde je: $Grupa_0 = \{c_i, i = 5\}$, $Grupa_1 = \{c_i, i = 2,4,6,8.\}$, $Grupa_2 = \{c_i, i = 1,3,7,9.\}$. Svaka $Grupa_r$ sadrži tačke koje su ekvidistantne u odnosu na referentnu tačku O i u sebi sadrže isti broj realizacija (r) događaja A. Rastojanja svih tačaka iz $Grupe_r$ su $d_r = \sqrt{r}, r = \{0,1,2\}$. U slučaju nD rešetke događaj A može imati bilo koju od n vrednosti, odnosno, $d_r = \sqrt{r}, r = \{0,1,2,\dots,n\}$.

U nastavku moramo definisati kompleksniju korespondenciju između raspodele diskretne promenljive X nad prostorom događaja Ω_2, u smislu frekvencije pojavljivanja komponenti promenljive X, i rasporeda grupa ekvidistantnih tačaka u odnosu na tačku O u našoj celobrojnoj kvadratnoj rešetki (Slika 5.3). Podsetimo se da $X = \{0,1,2,\dots,n\}$ predstavlja diskretnu slučajnu promenljivu čije komponente predstavljaju broj r mogućih uspešnih realizacija događaja A u n nezavisnih eksperimanata nad skupom E. Vrednosti definisane veličine $R_l(X = r)$, u Jednačini (9) i Tabeli 5.1, predstavljaju frekvencije uspešnih realizacija svih komponenti diskretne promenljive $X = \{0,1,2\}$ tokom dva nezavisna eksperimenta nad skupom E , odnosno (E^2). Pomenute frekvencije uspešnih realiozacija komponenti promenljive X predstavljaju broj mogućih načina realizacije aktuelnih komponenti. Frekvancija prve X komponente $R_l(X = 0)$ predstavlja broj (k_0) načina nulte realizacije događaja A u prostoru Ω_2, i ovaj broj je: $R_l(X = 0) = 1$ (Tabela 5.1). Povežimo ovaj događaj sa prostornim kordinatama tačke $c_5(0,0)$, Tabela 5.2 i Slika 5.3. Podsetimo da događaj $A = \{1, -1\}$ samo u tački $c_5(0,0)$, gde je $k_0 = 1$ (*susedi nultog reda*), ima nultu realizaciju $r = 0$. Takođe je očigledno da je rastojanje između tačaka c_5 i $O, d(O, c_5) = d_0 = 0$, (*rastojenje nultog reda*). Frekvencija druge komponente slučajne variable $R_l(X = 1)$ predstavlja broj k_1 svih mogućih načina jednostruke realizacije događaja A iz Ω_2 prostora i ovaj broj je $R_l(X = 1) = 4$, videti (Tabelu 5.1). Pogledom na Tabelu 5.2 i Sliku 5.3 možemo videti da, u koordinatama sledećih četiri $(k_1 = 4)$ tačke c_2, c_4, c_6, c_8, (*susedi prvog reda*), događaj A se odigrava jedan put ($r = 1$). Takopđe napominjemo da su ove tačke ekvidistantne u odnosu na $O, d(O, c_2) = d(O, c_4) = d(O, c_6) = d(O, c_8) = d_1 = \sqrt{1} = 1$ (*rastojenje prvog reda*). Frekvancija treće komponente X promenljive $R_l(X = 2)$ predstavlja broj (k_2) svih mogućih načina dvostruke realizacije događaja A u prostoru Ω_2 gde je $R_l(X = 1) = 4$ (Tabela 5.1). Iz Tabele 5.2 i Slike 5.3 možemo videti da se, u koordinatama sledećih četiri $(k_2 = 4)$ tačke c_1, c_3, c_7, c_9, događaj A odigrava dva puta $(r = 2)$ (*susedi drugog reda*). Takođe napominjemo da su ove tačke ekvidistantne u odnosu na tačku O , $d(O, c_1) = d(O, c_3) = d(O, c_7) = d(O, c_9) = d_2 = \sqrt{2}$ (*rastojanja drugog reda*). Prethodno iznesene činjenice sadrže sledeći generalni zaključak:

$k_r = R_l(X = r), r = 0,1,2,3,\dots,n.,$ (14)

Broj k_r predstavlja sa jedne strane frekvenciju $r - te$ komponente diskretne promenljive X $(X = r)$ u prostoru Ω_n (Tabela 5.1), i, sa druge strane, broj ekvidistantnih *suseda* $r - tog$ *reda* iz S_0^n okruženja referentne tačke $O\ \underbrace{(0,0,0,\dots,0)}_{n}$. Aktuelna identična Euklidska rastojanja za svaku grupu $Grupa_r$ tačaka predstavljaju (*rastojanja r-tog reda*) prikazana u Jed. (15):

$d_r = \sqrt{r}, r = 0, 1, 2, \dots, n.$ (15)

Na osnovu jednačina (9) i (14) izvodimo vrlo važan zaključak o broju k_r suseda *r-tog reda* u formi Jed. (16):

$k_r = 2^r \binom{n}{r}, r = 0,1, 2, \dots, n.$ (16)

Ukupan broj suseda iz S_i^n okruženja za proizvoljnu $i - tu$ tačku od N tačaka n -dimenzionalne rešetke može se prikazati u obliku sledeće sume:

$K_i^n = \sum_{r=1}^{n} k_r = \sum_{r=1}^{n} 2^r \binom{n}{r}, i = 1, 2, \dots, N.$ (17)

Ove zaključne relacije ispunjavaju sve potrebne uslove za daterminaciju aktuelnog odnosa *volumen/rastojanje*, ili indirektno, lokalne gustine instanci u ndimenzionalnoj pravilnoj rešetci sa uniformno raspodeljenim tačkama. Preciznije, odredili smo broj i poziciju svih tačaka iz S_i^n okruženja referentne tačke *i*, pa samim tim možemo odrediti srednje Euklidsko rastojanje d_{ain} za bilo koju tačku $c_i \underbrace{(x_i, y_i, z_i, \dots, w_i)}_{n}, c_i \in Z^n, i = 1,2, \dots, 3^n$., od svih susednih tačaka iz njenog S_i^n okruženja, i to na sledeći način:

$d_{ain} = \sum_{r=1}^{n} k_r d_r \,/\, \sum_{r=1}^{n} k_r = \sum_{r=1}^{n} 2^r \binom{n}{r} \sqrt{r} / \sum_{r=1}^{n} 2^r \binom{n}{r}$. (18)

Ako imamo praktičnih ili drugih razloga da uključimo manji broj najbližih suseda iz S_i^n okruženja proizvoljne tačke $c_i \in Z^n$, tada u Jed. (18) parametar r ima vrednosti manje od n, $r = \{1,2,3, \dots, t., t < n\}$, vidi Jed. (19).

$$\left.\begin{aligned} d_{ai1} = \sum_{r=1}^{1} 2^r \binom{n}{r} \sqrt{r} / \sum_{r=1}^{1} 2^r \binom{n}{r} = 2\sqrt{1}/2 = 1, n = 1, \\ d_{ai2} = \sum_{r=1}^{2} 2^r \binom{n}{r} \sqrt{r} / \sum_{r=1}^{2} 2^r \binom{n}{r} = (1 + \sqrt{2})/2, n = 2, \\ d_{ait} = \sum_{r=1}^{t} 2^r \binom{n}{r} \sqrt{r} / \sum_{r=1}^{t} 2^r \binom{n}{r}, t < n, \\ d_{ain} = \sum_{r=1}^{n} 2^r \binom{n}{r} \sqrt{r} / \sum_{r=1}^{n} 2^r \binom{n}{r}. \end{aligned}\right\} \quad (19)$$

U n – dimenzionalnoj celobrojnoj rešetki, svaka tačka $c_i \in Z^n, i = 1,2, \dots, N.$, ima tačno $k_1 = 2^1 \binom{n}{1} = 2n$ ekvidistantnih najbližih susednih tačaka c_j u svom S_i^n okruženju (*susedi prvog reda*), dakle, po dve tačke za svaku od n dimenzija sa minimalnim rastojanjem d_1 (*rastojanje prvog reda*), $(d_1 = d(c_i, c_j) = \sqrt{1} = 1, i = 1,2, \dots, N, j = 1, 2, 3, \dots, 2n)$. Ovo su instance iz $Grupe_1$. Iz prve od jednačina (19) i prve kolone Tabele 5.3 može se videti da srednje rastojanje d_{ai1} između tačaka $Grupe_1$ na zavise od broja dimanzija n pošto imaju konstantnu vrednost $d_{ai1} = 1$. Za sva ostala rastojanja na osnovu (18) i (19) važi sledeća relacija: $d_{ait} > 1,\ n \geq t > 1$. Tabela 5.3 predstavlja izračunate srednje vrednosti Euklidskih rastojanja tačaka, na osnovu Jed. (19), za celobrojne rešetke koje sadrže N primera u – dimenzionalnom prostoru za slučajeve $n = 1,2, \dots,10$.

Tabela 5.3

Vrednosti srednjih rastojanja pravilnih celobrojnih rešetki za $n = 1,2,3, \dots ,10$.

$n\downarrow$ \ $r\rightarrow$		1	2	3	4	5	6	7	8	9	10
1		1	0	0	0	0	0	0	0	0	0
2		1	1.2071	0	0	0	0	0	0	0	0
3	d_{air}	1	1.2761	1.4164	0	0	0	0	0	0	0
4		1	1.3107	1.5214	1.6171	0	0	0	0	0	0
5	$r \leq n$	1	1.3314	1.5779	1.7387	1.8045	0	0	0	0	0
6	$i = 1,..,N$	1	1.3452	1.6120	1.8093	1.9327	1.9781	0	0	0	0
7		1	1.3550	1.6343	1.8526	2.0127	2.1078	2.1393	0	0	0
8		1	1.3624	1.6499	1.8811	2.0635	2.1945	2.2678	2.2897	0	0
9		1	1.3682	1.6614	1.9009	2.0973	2.2517	2.3594	2.4158	2.4310	0
10		1	1.3728	1.6701	1.9153	2.1209	2.2906	2.4224	2.5108	2.5540	2.5645

5.3.3. Važne distributivne karakteristike pravilne celobrojne rešetke

Za ovu svrhu možemo definisati rešetku kao pravilno popunjavanje prostora pomoću osnovnih ćelija, koje se mogu smatrati kvadratima u dvodimenzionalnom slučaju ili kockama, odnosno hiperkockama u $3D$, odnosno nD slučaju. Razmotrimo pravilnu nD celobrojnu rešetku $C_L \subset Z^n, C_L = \{c_{1L}, c_{2L}, \dots, c_{N_L}\}$, sačinjenu od N_L osnovnih jediničnih ćelija $c_{iL}, i = 1, 2, \dots, N_L$. Svaka jedinična ćelija je u stvari hiperkocka ivice $a = 1$ i volumena $v_{iL} = a^n = 1, i = 1, 2, \dots, N_L.$, sa jednom tačkom u težištu kao na Slici 5.3 za $2D$ slučaj. Rešetku C_L smatramo kao prihvatljiv slučaj aproksimacije ravnomerne raspodele uzorka ograničenog broja N_L instanci u n-dimenzionalnom prostoru, sa konstantnom funkcijom gustine (mase) verovatnoće ($p_i = 1/N_L, \forall i = 1,2, \dots, N_L$) i konstantnom entropijom[$H_i = -(1/N_L)\log_2(1/N_L),\ \forall i = 1,2, \dots, N_L$] identičnom za svaku tačku c_i, gde $H(C_L) = \sum_{i=1}^{N} H_i = log_2 N_L$ predstavlja Shannon-ovu entropiju ovog uzorka i maksimalnu moguću entropiju za svaki uzorak od N_L tačaka u istom prostoru. Dakle, $H(C_L) > H(C), \forall C \neq C_L \wedge N_L = N$ gde N predstavlja broj instanci proizvoljnog uzorka C. Aktuelni uzorak C_L je okarakterisan sa dve očigledne distributivne karakteristike: srednje lokalno rastojenja d_{ain} izvesne konstantne vrednosti za određene grupe kao u Jed. (19), i identičnim volumenom ($v_{iL} = a^n = 1$) prostora instanci pridruženog svakoj tački c_{iL}. Svaka od ovih distributivnih karakteristika uzorka C_L pojedinačno reprezentuju indirektnu verifikaciju njegove uniformne raspodele verovatnoća. Definišimo korespondentni jednodimenzionalni diskretni uzorak, V_L, $V_L = \{v_{1L},\ v_{2L}, \dots,\ v_{N_L}\} = \underbrace{\{1, 1, \dots, 1\}}_{N_L}$, kao niz identičnih jediničnih volumena. Ovde ćemo uzeti novu jednostavniju konvencionalnu notacju d_{iL} kao zamenu za dosadašnju oznaku srednje lokalne distance d_{ain}, dakle $d_{iL} = d_{ain}$, i definišimo uzorak D_L, $D_L = \{d_{iL}, i = 1, 2, .., N_L\}$, sastavljen od identičnih konstantnih komponenti u skladu sa Jed. (19) i Tabelom 5.3. Pošto ova

dva uzorka (V_L i D_L) pojedinačno imaju konstantne vrednosti članova ($d_{iL} = d_{jL}, \forall\, i,j \in \{1,2,..,N_L\} \wedge v_{iL} = v_{jL} = 1, \forall\, i,j \in \{1,2,..,N_L\}$) oba se povinuju zakonu proste degenerativne (deterministične) raspodele verovatnoće sa funkcijom gustine raspodele datom u formi Dirakove delta funkcije sa korespondentnim izvornim vrednostima $O_{rd} = d_{iL}$ i $O_{rv} = v_{jL} = 1$. Koristeći nizove V_L i D_L možemo indirektno prezentovati distributivne karakteristike originalnih uniformnih nD uzoraka u formi rešetke C_L. Uspostavimo važnu korespondenciju između vrednosti v_{iL} i d_{iL} na sledeći način:

$$\left.\begin{matrix} v_{iL}/(d_{iL})^n \equiv a^n/(d_{iL})^n \\ v_{iL}/(d_{iL})^n \equiv 1/(d_{iL})^n, a = 1 \end{matrix}\right\}, \tag{20}$$

U Jed. (20) indeks „$_L$“ znači *lattice* (rešetka) a „^{n}“ predstavlja eksponent i dimenziju vektora obeležja instanci. Ovaj identitet predstavlje indirektno određivenje funkcije gustine verovatnoće aktuelne uniformne rtešetke. Ukoliko nađemo način da uspostavimo slično stanje između ovih atributa u proizvoljnom empirijskom uzorku sa proizvoljnom raspodelom verovatnoće, tada bi njegova reprezentativnost i njegova entropija težile svojoj maksimalno mogućoj vrednosti a empirijski uzorak bi aproksimirao celobrojnu regularnu rešetku u smislu raspodele verovatnoće. Svi prethodni relevantni rezultati i zaključci su izvedeni iz slučaja celobrojne kvadratne rešetke (Slika 5.3) u svrhu lakšeg razumevanja problema i lakše manipulacije, ali treba istaći da ovi rezultati i zaključcim imaju neograničenu univerzalnu primenljivost i u slučaju proizvoljnih rastojanja ($a_i \neq 1$) između najbližih susednih tačaka u realnom prostoru $C_L \in R^n$, tako da odgovarajući generalni represent srednjih rastojanja prikazan u Jed. (19) može biti redefinisan prostom inkluzijom faktora a u Jed. (19) na koji način ćemo dobiti sledeći skup jednačina (21).

$$\left.\begin{matrix} d'_{i1} = a\sum_{r=1}^{1} 2^r \binom{n}{r}\sqrt{r}/\sum_{r=1}^{1} 2^r \binom{n}{r} = 2an\sqrt{1}/2 = a, n = 1, \\ d'_{i2} = a\sum_{r=1}^{2} 2^r \binom{n}{r}\sqrt{r}/\sum_{r=1}^{2} 2^r \binom{n}{r} = a(1+\sqrt{2})/2, n = 2, \\ d'_{it} = a\sum_{r=1}^{t} 2^r \binom{n}{r}\sqrt{r}/\sum_{r=1}^{t} 2^r \binom{n}{r}, t < n, \\ d'_{in} = a\sum_{r=1}^{n} 2^r \binom{n}{r}\sqrt{r}/\sum_{r=1}^{n} 2^r \binom{n}{r}. \end{matrix}\right\} \tag{21}$$

U sledećim podpoglavljima prikazujemo resampling metod za povećanje reprezentativnosti eksperimentalnih neizbalansiranih trening uzoraka.

5.3.4. Transfer distributivnih karakteristika sa ragularne rešetke na uzorke proizvoljne raspodele

Zamislimo pogodnu transformaciju $\mathcal{T}$ koja preslikava kompleksnu eksperimentalnu klasu (uzorak) $C_e = \{c_{ie}, i = 1,2,\dots,N_e,\ C_e \in R^n\}$ proizvoljne raspodele u izbalansirani uzorak $C_b =$

$\{c_{ib}, i = 1, 2, \dots, N_b,\ C_b \in R^n\}$, sa raspodelom vrlo sličnom ravnomernoj, tako da je $\mathcal{T}(C_e) = C_b, C_b \approx C_L$, gde C_b predstavlja prihvatljivu aproksimaciju C_L, a C_L predstavlja klasu odnosno uzorak u formi n-dimenzionalne pravilne rešetke $C_L = \{c_{iL}, i = 1, 2, \dots, N_L, C_L \in R^n\}$. Glavni cilj ovog podpoglavlja je upravo pronalaženje metoda odnosno algoritma kojim ćemo uspostaviti ovu transformaciju kojom se postiže porast reprezentativnosti originalne eksperimentalne klase (uzorka) pre njegovog uključenja u process obule klasifikatora odnosno pre procesa klasifikacije instanci ciljne populacije kojoj uzorak pripada. Uzmimo eksperimentalnu klasu $C_e = \{c_{ie}, i = 1, 2, \dots, N_e,\ C_e \in R^n\}$ i izračunajmo srednje Euklidska rastojanja $d_{ie}, i = 1, 2, \dots, N_e,$ za svaku originalnu instancu c_{ie} u saglasnosti sa Jed. (19) pa zatim definišimo jednodimenzionalni niz $D_e = \{d_{ie}, i = 1, 2, .., N_e.\}$. Naglašavamo da niz D_e, izračunat iz eksperimentalnog uzorka, ima članove sa međusobno različitim vrednostima ($d_{ie} \neq d_{je}, i \neq j$) što ukazuije na neravnomernost raspodele originalnog uzorka, za razliku od niza $D_L = \{d_{iL}, i = 1, 2, .., N_L.\}$, izvedenog iz pravilne uniformno distribuirane rešetke (uzorka) a čiji su članovi identični, odnosno imaju konstantnu jedinstvenu vrednost ($d_{iL} = d_{jL}, \forall i, j$).

Pođimo od pretpostavke da svaka tačka $c_{ie} \in C_e$ zauzima određeni fiktivni volume v_{ie}, Ekvivalentan eksponencijalnoj vrednosti svojeg sredneg Euklidskog rastojanja odnosno $(d_{ie})^n$, po ugledu na tu relaciju kod pravilne rešetke. Prenesimo ovaj karakteristični odnos sa uzorka C_L, dato u Jed. (20), ne eksperimentalni uzorak C_e i po tom principu izračunajmo fiktivnu ekvivalentnu zapreminu v_{ie} hipotetički pridruženu svakoj instanci

$c_{ie}, i = 1,2, \dots, N_e.$

$$\left.\begin{array}{l} v_{ie}/(d_{ie})^n = v_{iL}/(d_{iL})^n \Rightarrow v_{ie}/(d_{ie})^n = a_{iL}^n/(d_{iL})^n \Rightarrow v_{ie} = a_{iL}^n (d_{ie}/d_{iL})^n, a \neq 1 \\ v_{ie}/(d_{ie})^n = v_{iL}/(d_{iL})^n \Rightarrow v_{ie}/(d_{ie})^n = 1/(d_{iL})^n \Rightarrow\ v_{ie} = (d_{ie}/d_{iL})^n, a = 1 \end{array}\right\} \quad (22)$$

Sada definišimo jednodimenzionalni uzorak volumena $V_e = \{v_{1e}, v_{2e}, \dots, v_{Ne}\}$. koji odgovaraju svakoj instanci c_{ie} eksperimantalnog uzorka C_e. Očigledno je da je uzorak V_e niz vrednosti v_{ie} koje su takođe me]usobno različite ($v_{ie} \neq v_{je}, i \neq j$), što je opet suprotno stanju niza $V_L = \{v_{1L}, v_{2L}, \dots, v_{N_L}\}$ dobijenog iz uzorka regularne rešetke gde važi sledeća relacija: $v_{iL} = v_{jL} = a_{iL}^n, \forall\ i, j$. Uporedimo neke distributivne karakteristike dva, za nas interesantna nD uzorka: a) prostu pravilnu rešetku C_L čija je *i-ta* jedinična ćelija definisana kao hiperkocka dužine a_{iL} i zapremine $v_{iL} = a_{iL}^n$, koja je indirektno predstavljena preko $1D$ uzorka D_L sa identičnim konstantnim članovima d_{iL}, i b) eksperimentalni uzorak proizvoljne neravnomerne raspodele instanci u realnom prostoru C_e, takođe indirektno predstavljenom preko $1D$ uzorka D_e sa međusobno rezličitim članovima d_{ie}. U podpoglavlju 4.2 smo istaklli da jednostanva kubične rašetka C_L predstavlja uniformnu raspodelu ograničenog broja instanci

N_L u prostoru obeležja koja sa maksimalnom entropijom $H_{C_e} = H(C_e) = \log_2 N_L$, u odnosu na sve ostale moguće raspodele uzoraka od N_L instanci u istom prostoru.

U skladu sa ovom činjenicom, jednostavno zaključujemo da je $H_{C_L} > H_{C_e}$. Podsetimo da su uzorci (nizovi) D_e i V_e sastavljeni od različitih članova što implicira pozitivne vrednosti njihovih standardnih devijacija ($\sigma_{D_e} = \sigma(D_e) > 0$ i $\sigma_{V_e} = \sigma(V_e) > 0$), dok su odgovarajuće devijacije za pravilnu rešetku $\sigma_{D_L} = \sigma(D_L) = 0$ i $\sigma_{V_L} = \sigma(V_L) = 0$, što rezultuje jednostavnim zaključkom u formi sledećih nejednačina: $\sigma_{D_e} > \sigma_{D_L}$ i $\sigma_{V_e} > \sigma_{V_L}$. Na osnovu ovih činjenica jednostavno zaključujemo da i entropije involviranih nizova stoje u sličnim relacijama kao i njihove standardne devijacije: $H_{D_e} = H(D_e) > 0$ i $H_{V_e} = H(V_e) > 0$, dok su $H_{D_L} = H(D_L) = 0$ i $H_{V_L} = H(V_L) = 0$ što implicira: $H_{D_e} > H_{D_L}$ and $H_{V_e} > H_{V_L}$.

Sumiranjem prethodnih činjenica dolazimo do generalnog zaključka da vrednosti entropija originalnog uzorka (H_{Ce}) i entriopija niza lokalnih srednjih rastojanja njegovih instanci (H_{De}) stoje međusobno u obrnutoj proporciji. Drugim rečima, što je vaća entropija originalnog uzorka (H_{Ce}) to je manja entropija korespondentnog niza lokalnih srednjih rastojanja (H_{De}). Iskoristićemo niz V_e za detekciju oblasti prostora obeležja sa različitim lokalnim gustinama verovatnoće. Reletivno male vrednosti v_{ie} odgovaraju instancama iz oblasti visoke gustine verovatnoće dok velike vrednosti volumena odgovaraju oblastima male gustine verovatnoće. Razmotrimo još jednom celobrojnu pravilnu rešetku C_L sa N_L instanci koje zauzimaju proctor ukupne zapremine $\mathbb{V}_L = \sum_{i=1}^{N_L} v_{iL}$. Pošto ceo volumen $\mathbb{V}_L$ sadrži N_L tačaka i svaka tačka (instanca) po pretpostavci zauzima jedinični volumen $v_{iL} = 1, i = 1,2, \dots, N_L.$, tada je volumen celog uzorka ekvivalentan broju N_L, odnosno $\mathbb{V}_L = \sum_{i=1}^{N_L} v_{iL} = N_L v_{iL} = N_L$. Definišimo pod-uzorak (stratum) C_{LS} od N_{LS} instanci volumena$\mathbb{V}_{LS}$, $\mathbb{V}_{LS} \subset \mathbb{V}_L$i $v_{iLS} = 1, i = 1, \dots, N_{LS}.$ tada je $\mathbb{V}_{LS} = \sum_{i=1}^{N_S} v_{iLS} = N_{Ls} v_{iLS} = N_{LS}$ što vodi ka jednačini. (23).

$$\mathbb{V}_{LS}/N_{LS} = \mathbb{V}_L/N_L \Longrightarrow N_{LS} = N_L(\mathbb{V}_{LS}/\mathbb{V}_L). \quad (23)$$

Relacije date u Jed. (22) i (23), definisane za nD celobrojnu regularnu rešetku, prilagođavamo exsperimentalnom uzorku C_e proizvoljne neravnomerne raspodele za koji važi $d_{ie} \neq d_{je} \forall i \neq j$. Transfer gornje relacije je od presudnog značaja za balansiranje proizvoljnog empirijskog uzorka na način koji garantuje dobijanje konačnog uzorka sa raspodelom verovatnoće koja je slična ravnomernoj. Ovaj cilj postižemo uz pomoć DBB resampling algoritma zasnovanog na undersamplingu u oblastima prostora obeležja visoke gustine verovatnoće podržanog sintetičkim oversamplingom u oblastima prostora male gustine.

Na osnovu činjenica Krasnopolsky (2013), trening uzorak treba da ima finiju rezoluciju u oblastima gde target populčacija nije adekvatno semplovana i grublju rezoluciju u oblastima sa gustim semplovanjem. Povećanje balansa instanci unutar uzoraka (klasa) povećavamo njegovu reprezentativnost i samim tim, pouzdanost involviranih induktivnih klasifikatora, što je u saglasnosti sa glavnim ciljem ovog poglavlja.

5.3.5. Praktčne prednosti pristupa odabiranju preko indirektnog odnosa volumen/distanca

Analizirajmo jednostavan pristup segmentaciji višedimanzionalnog prostora obeležja u cilju analize lokalne gustine. Uzmino sledeći empirijski uzorak:

$C_e\ (f),\ C_e = \{c_{ie} | c_{ie} = f_{i,j}, i = 1,2, \dots, N_e., j = 1,2, \dots, n.\}$, sastavljen od N_e članova determinisanih sa n obeležja. Zamislimo segmentaciju ovih n obeležja (osa) na $h = 100$ segmenata. Ovom operacijom mi dekomponujemo originalni skup C_e na h^n novih podskupova (podprostora) koje treba analizirati. Zamislimo realnu situaciju sa uzorcima gde je $n = 100$ i $h = 100$, što znači da je $h^n = 1.0000\text{e} + 2001$, što uzrokuje ogromnu računarsku komplekssnost $O(h^n)$. Realne mogućnosti za pretraživanje, analizu i resemplovanje ovolikog broja subprostora su vrlo male sa obzirom na operativna ograničenja kompjutera i algoritama.

Druga prepreka ovom direktnom pristupu reodabiranja uzorka je nedostatak mogućnosti vizuelne prezentacije uzoraka u višedimanzionalnom prostoru obeležja. Razmotrimo indirektni pristup istom problemu, analize lokalne gustine verovatnoće istog uzorka C_e koristeći Jed. (22). rednja Euklidska rastojanja d_e su determinisana na sledeći način:

$d_e = \{d_{ik} | d_{ik} = d(c_{ie}, c_{ke}), i = k = 1, 2, \dots, N_e\}$ .Za svaku od N_e instanci originalnog uzorka C_e, kardinalne vrednosti su $|d| = |N_e \times N_e| = N_e^2$, gde je $d(c_{ie}, c_{ke})$ rastojanje dato u Jed. (1).

Ova operacija, čak i u slučaju velikih vrednosti N_e je problem sa kompleksnošću $O(N_e^2)$ koja je više redova veličina niža od inicijalne kompleksnosti $O(h^n)$. Ove činjenice prikazuju praktični značaj primene predložene supstitucije. U nastavku ova procedura se odvija uglavnom u prostoru skupa odabranih srednjioh rastojanja kardinalne vrednosti $|N_e|$, Što znači da algoritam uglavnom radi u jednodimenzionalnom prostoru dužine N_e , omogućavajući neuporedivo kraće operactivno vreme $O(N_e)$.

5.3.6. Transformacija empirijskog uzorka u kvaziuniformni uzorak pomoću DBB stratifikovanog odabiranja.

Polazimo od prethodne pretpostavke da uzorak u formi pravilne rešetke predstavlja uzorak sa uniformnom raspodelom instanci sa maksimalnom vrednošću entropije i najvećim stepenom reprezentativnosti, jednostavno rečeno idealno balansirani uzorak.

Ova činjenica jasno definiše tok procesa balansiranja realnih empirijskih uzoraka, koje najčešće karakteriše naglašena neravnomernost raspodele instanci, značajno niža vrednost entropije i nizak nivo reprezentativnosti, odnosno, izražen interni disbalans. Adekvatna procedura balansiranja podrazumeva dopustivu i prihvatljivu transformaciju $\mathcal{T}$ koja transformiše originalni neuravnoteženi uzorak C_e u finalni izbalansirani uzorak ($C_b: C_b = \mathcal{T}(C_e)$) koji će imati strukturu sličnu strukturi regularne rešetke iste dužine (broja instanci) i istovremeno, ostati u domenu ciljne populacije koju reprezentuje. Transformacija $\mathcal{T}$ je kompozicija tri odvojene transformacije: a) *Normalizacija* $\mathcal{N}(\cdot)$ volumena uzorka, b) *Sintetički Selektivni Stratifikovani Oversampling* $\mathcal{O}(\cdot)$ i za produkciju novih sintetičkih instanci u oblastima prostora male gustine verovatnoće i c) *Selektivni Stratifikovani Undersampling* $\mathcal{U}(\cdot)$, procedura za eliminaciju redundantnih instanci u oblastima velike gustine instanci. Sledeći izraz predstavlja formalnu reprezentaciju pomenute kompozicije: $\mathcal{T} = \{\mathcal{O}, \mathcal{U}, \mathcal{N}\}: \mathcal{T}(\cdot) = \mathcal{O}(\mathcal{U}(\mathcal{N}(\cdot)))$. Podvrgavanjem originalnog neizbalansiranog uzorka C_e operatoru $\mathcal{T}$ dobijamo sledeću razvijenu formu:

$$C_b = \mathcal{T}(C_e)) = \mathcal{O}(\mathcal{U}(\mathcal{N}(C_e))) \quad (24)$$

Tekst koji sledi detaljno opisuje sve tri nabrojane elementarne transformacije:

a) *Normalizacija* $\mathcal{N}(\cdot)$

Odredimo ukupne zapremine koje zauzimaju uzorci iste dužine C_L i C_e za koje je $N_L = N_e = N$, a odgovarajuće zapremine su $V_L = \{v_{iL}, i = 1,2, \dots, N\}$ and $V_e = \{v_{ie}, i = 1,2, \dots, N\}$, gde je:

$$\mathbb{V}_L = \sum_{i=1}^{N} v_{iL} = N, \mathbb{V}_e = \sum_{i=1}^{N} v_{ie} \quad (25)$$

Pošto je prethodno definisana jednakost $\mathbb{V}_L = N$ definišimo procedure normalizacije $\mathcal{N}(\mathbb{V}_e) = \mathbb{V}_{en} = \mathbb{V}_L = N$, gde $\mathbb{V}_{en}$ predstavlja normalizovani ukupni volumen zauzet od strane uzorka C_e. Definišimo normalizovani volumen koji zauzima proizvoljna instanca empirijskog uzorka na sledeći način: $v_{ie\mathcal{N}} = N v_{ie} / \sum_{i=1}^{N} v_{ie}$ i saberimo zapremine svih instanci tog uzorka $V_{e\mathcal{N}} =$

$\{v_{ie\mathcal{N}}, i = 1,2, \dots, N\}$, na sledeći način: $\mathbb{V}_{e\mathcal{N}} = \sum_{i=1}^{N} v_{ie\mathcal{N}} = \sum_{i=1}^{N} N v_{ie} / \sum_{i=1}^{N} v_{ie} = N \sum_{i=1}^{N} v_{ie} / \sum_{i=1}^{N} v_{ie} = N = \mathbb{V}_L.$ (26)

Dakle, polazimo od izjednačavanja zapremina eksperimentalnog uzorka $V_{e\mathcal{N}}$ i pravilne rešetke V_L tako da njihovi ukupni volumeni $\mathbb{V}_L$ i $\mathbb{V}_{e\mathcal{N}}$ imaju identične vrednosti ($\mathbb{V}_L = \mathbb{V}_{e\mathcal{N}} = N$), gde volumeni instanci nisu identično raspodeljeni ($v_{iL} = 1, i = 1,2, \dots, N$) i ($v_{ie\mathcal{N}} \neq v_{je\mathcal{N}}, i \neq j, \forall\, i, j \in \{1, 2, .., N_L\}$). Ujednačavanje volumena instanci empirijskog uzorka $V_{e\mathcal{N}}$ je krajnji cilj procesa balansiranja koji ćemo ostvariti indirektnim resamplingom instanci u prostoru obeležja uzoraka. Ova procedura počinje

Stratifikacijom uzorka $V_{e\mathcal{N}}$ za kojim sledi sintetički oversampling stratuma koji sadrže mali broj instanci i undersamplingom stratuma koji sadrže veliki broj instanci.

Sratifikacija je process podele uzorka populacije na određeni broj homogenih subpopulacija pre uzorkovanja. Na ovaj način odabrani uzorak se može smatrati kao parcijalno aproksimativno uniformni uzorak. Proporcionalno stratifikovano uzorkovanje predstavlja slučajno odabiranje instanci iz svake podgrupe (stratuma) u direktnoj proporciji sa njenim učešćem u celoj populaciji. U slučaju standardnog proporcionalnog stratifikovanja broj instanci odabran iz svakog stratuma je direktno proporcionalan standardnoj devijaciji aktuelne variable u stratum tako da veća variabilnost implicira veću dužinu uzorka. Svrha ovakvog odabiranja je da se dobije izbalansirana reprezentacija različitih podgrupa u svrhu bolje klasifikacije ili klasterovanja. Uzmimo normalizovani empirijski uzorak $V_{e\mathcal{N}} = \{v_{ie\mathcal{N}}, i = 1,2, \dots, N\}$, gde $\mathbb{V}_{e\mathcal{N}} = \sum_{i=1}^{N} v_{ie\mathcal{N}} = N$ predstavlja ukupni volume eksperimentalnog skupa (Jed. (25)). Izvršimo stratifikaciju vektora $V_{e\mathcal{N}}$ na M $(M < N)$ homogenih grupa (stratuma) G_h, gde svaki od njih sadrži različit broj N_h različitih volumena koji odgovaraju instancama $\{v_{ie\mathcal{N}} : v_{ie\mathcal{N}} \neq v_{je\mathcal{N}}, i \neq j,\ i, j = 1,2,3, .., N\}$ gde $(M, N, N_h \in N+),\ N_h = |G_h|,\ h = 1,2,3, \dots, M.,\ N_h \neq N_k,\ h \neq k, \forall\, h, k \in \{1, 2, .., M\}$. Podsetimo da $N+$ ima značenje pozitivnog celog broja. Reprezentativni stratum definisan u Jed. (27) predstavlja posebnu oblast prostora obeležja ukupnog volumena $\mathbb{V}_h$, koji sadrži disjunktivne skupove instanci uzorka $G_h \cap G_k = 0,\ \forall\, h, k \in \{1, 2, .., M\}$, gde je: $\mathbb{V}_h = \mathbb{V}_{e\mathcal{N}}/M,\ V_h = V_k, \forall\, h, k \in \{1, 2, .., M\}$.

$G_h = \{v_{ie\mathcal{N}},\ i = (h-1)N_h + 1,\ (h-1)N_h + 2,\ (h-1)N_h + 3, \dots,\ (h-1)N_h + N_h,\ h = 1,2, \dots, M\}$. (27)

Pošto u ovoj situaciji imamo posla sa neravnomernom raspodelom instanci, stratifikacijom dobijamo *M* stratuma od kojih je svaki reletivno homogeni skup međusobno sličnih članova $v_{ie\mathcal{N}}$ (parcijalno uniforman), dok se istovremeno ove vrednosti znatno razlikuju između stratuma. Neki stratumi mogu sadržati ekstremno male $v_{ie\mathcal{N}}$ vrednosti, dok drugi mogu imati

relativno velike vrednosti. Pošto po definiciji svaka od h groupa (G_h), zauzima jednaku zapreminu prostora $\mathbb{V}_h$, ukoliko grupa sadrži instance malih $v_{ie\mathcal{N}}$ vrednosti znači da je broj instanci N_h u toj grupi relativno veliki. Analogno, velike vrednosti $v_{ie\mathcal{N}}$ unutar stratum implicira reletivno mali broj instanci N_h. Drugim rečima, neki stratum imaju veliku gustinu tačaka dok druge pripadaju retkim grupama. Očigledno, stratum male gustine mogu su slabo predstavljeni, dok grupe velike gustine smatramo previše zastupljenim, što u slučaju klasifikacije predstavlja disbalans unutar klase. Pošto je ovde naš glavni cilj balansirana distribuicija instanci po celom volumenu $\mathbb{V}_{e\mathcal{N}}$, krucijalni zahtev je uspostavljanje kvaziuniformnosti raspodele unutar svake od h grupa G_h. Imajući na umu početni uslov neravnomerne raspodele instanci unutar svakog stratum i znajući da svaki stratum zauzima identičan volumen $\mathbb{V}_h$, naša procedura odabiranja treba da obezbedi da finalni stratum imaju jednak broj instanci N_{hb}, kako bi se obezbedila ista gustina unutar celog prostora što znači izbalansiran uzorak. Za ispunjenje potrebnog zahteva, nad stratumima velike gustine treba sprovesti undersampling a nad stratumima male gustine oversampling kako bi se postigao prihvatljivi predloženi balans instanci po svim grupama $G_{hb}: N_{hb} = N_{kb}, h \neq k, \forall\, h, k \in \{1, 2, .., M\}$. Dakle, stratum sa malim $v_{ie\mathcal{N}}$ vrednostima imaju ulogu klastera velike gustine smeštenih u klasi i oni treba da budu podvrgnuti undersamplingu. Redukcija instanci se može postići slučajnim odbacivanjem instanci iz svake grupe pojedinačno. Posle undersampling procedure svaka od h obrađenih (izbalansiranih) grupa G_{hb} sadrži $N_{hb} = N/M$ instanci. Problem rešavamo primenom poznatog SMOTE metoda balansiranja. Ovaj metod smo nazvali *Syntetički Selectivni Stratifikovani Oversampling* odnosno restriktivni SMOTE. Posle završene procedure odabiranja, svaka od h izbalansiranih grupa G_{hb} sadrži $N_{hb} = N/M$ instanci.

U oba slučaja, uvažavamo kriterijume za određivanje pogodnog broja instanci u projektovanim izbalansiranim uzorcima. Postoji nekoloko načina za određivanje ove vrednosti u zavisnosti od prirode problema (distribucije) i broja N raspoloživih instanci.

Sa aspekta odnosa između brojeva instanci N u originalnom skupu $V_{e\mathcal{N}}$ i broja instanci N_b finalnog izbalansiranog skupa $V_{e\mathcal{N}b}$ prirodno se pojavljuje kao logičan sledeći kriterijum: $N/\,N_b = 1$. Na ovaj način želimo da zadržimo postojeći broj instanci, iako to nije uvek potrebno a nekada je i nepoželjno, jer, kako smo već rekli, dobro distribuirana mala grupa instanci u prostoru obeležja nosi mnogo više informacija o ciljnoj populaciji nego mnogo veća grupa inszanci koje su neravnomerno raspoređene u tom istom prostoru. Dakle, iz praktičnih razloga ova vrednost može imati i drugu vrednost $N/\,N_b \neq 1$. U slučaju kada usvojimo $N/\,N_b > 1$, tada kažemo da postoji kontrolisani disbalans koji se može predstaviti preko odnosa imbalansa (IR), koji predstavlja odnos veličina originala i finalnog uzroka ($IR =$

$N/\ N_b$). U slučaju velike redundanse ovaj odnos će imati veliku vrednost i ima praktični značaj pri redukciji veličine inicijalnog trening uzorka. Kada odredimo željeni broj u finalnom izbalansiranom uzorku N_b, time određujemo broj instanci N_{hb} u svakom od novih balansiranih stratuma ili grupa G_{hb} koji se može predstaviti na sledeći način: $N_{hb} = N_b/M$. Balansiranje pomenutih grupa se izvršava u saglasnosti sa definisanim pragom odabiranja, koji je funkcija broja stratuma M i broja ionstanci N_b. U našem slučaju je: $\theta = N_b/M$.

b) *Sintetički Selektivni Stratifikovani Oversampling* $\mathcal{O}(\cdot)$

Ova procedura ima epitet selektivna zato što se primenjuje samo u oblastima male gustine. Svaka grupa $G_h, |G_h| < \theta$, gde $|G_h|$ predstavlja kardinalnost grupe, se podvrgava sintetičkom oversamplingu, što podrazumeva interpolaciju novih sintetičkih tačaka na linijama koje spajaju odabranu instancu sa određenim skupom njenih najbližih suseda. Ovom interpolacijom instanci menjamo kardinalne vrednosti odabranih stratum:

$$\mathcal{O}(G_h) = G_{hb}, |G_{hb}| > |G_h|, |G_{hb}| = N_{hb} = \theta. \quad (28)$$

Standardna SMOTE procedura je detaljno objašnjena u Chawla et al. (2002). Ovde predstavljeni selektivni stratifikovani oversampling ima veliku prednost u odnosu na običnu SMOTE procedure jer ima tendenciju optimizacije raspodele instanci koja maksimizira entropiju uzorka (klase) menjajući početnu neravnomernu raspodelu u smeru uniformne raspodele izbalansiranog uzorka. Na suprot našem metodu, običan SMOTE izvršava interpolaciju instanci po celom prostoru obeležja i na taj način održava postojeći disbalans unutar klase što je suboptimalno rešenje problema balansiranja.

c) *Selectivni Stratifikovani Undersampling* $\mathcal{U}(\cdot)$

Grupe $G_h, |G_h| < \theta$ su podvrgnute sledećoj undersampling proceduri:

$$\mathcal{U}(G_h) = G_{hb}, |G_{hb}| < |G_h|, |G_{hb}| = N_{hb} = \theta. \quad (29)$$

Ova procedura podrazumave selektivno random uklanjanje redundantnih instanci iz grupa sa reletivno velikim brojem primera, što korespondira sa uklanjanjem instanci iz oblasti velike gustine. Ova procedura ima velike prednosti u poređenju sa jednostavnim uklanjanjem slučajno odabranih instanci, zato što ima tendenciju prema optimalnoj raspodeli instanci koja maksimizira entropiju uzorka promenom neravnomerne raspodele originala u smeru uniformne raspodele finalnog izbalansiranog uzorka, postižući na ovaj način povećanje reprezentativnosti uzorka i redukujući broj redundsdantnih instanci u gustim oblastima prostora obeležja. Nasuprot ovom metodu, metod nasumičnog uklanjanja instanci iz celog prostora obeležja, rezultuje uklanjanjem važnih instanci u retkim oblastima proctor, zadržavajući na taj način

inicijalni interni disbalans. Finalni rezultat gornjih procedura balansiranja je novi izbalansirani skup stratum (grupa) $G_{hb}, |G_h| = \theta$, koji determiniše finalni izbalansirani skup instanci $C_{ib}, i = 1, 2, 3, \dots, N_b$.

5.3.7. DBB Algoritam Balansiranja Zanovan na Distancama

Input

1) Inicialn training uzorak D koji ima N vector instanci x_i dužine n, $D = \{x_i | x_i = f_{ij}, i = 1, \dots, N., j = 1,2, \dots, n.\}$ gde je n dimenzija actuelnog uzorka.

Procedura

Do **Loop** od 1 do N:

(1) Izračunaj Euclidske distance d_{ik} za svaku vektor instancu $x_i = \{f_{i1}, f_{i2}, \dots, f_{in}\}$ od svih njenih $x_k = \{f_{k1}, f_{k2}, \dots, f_{kn}\}, i, k = 1,2,3, \dots N.$, kako sledi: $d_{ik} = d(x_i, x_k) = \sqrt{\sum_{j=1}^{n}(f_{i,j} - f_{k,j})^2}$, $i, k = 1,2, \dots, N., j = 1,2, \dots, n.$

(2) Sortiraj vrednosti d_{ik} svake instanci x_i u rastućem poretku.

(3) Uzmi prvih K_i predefinisan broj $K_i = \sum_{r=1}^{t} 2^r \binom{n}{r}$ sortiranih distanci d_{ik}, and izračunaj srenjeu vrednost d_{ie} za sbvaku instancu x_i.

(4) Izračunaj sredbnje Euclidsko rastojanje d_{iL} za svaku od N instanci u n-dimenzionalnoj uniformnoj celobrojnoj rešetki u skladu sa sistemom Jed. (19).

(5) Izračunaj empirijske vrednosti volumena v_{ie} za svaku od N tačaka u skladu sa Jed. (21), $v_{ie} = (d_{ie}/d_{iL})^n$ i definiši niz $V_e = \{v_{1e}, v_{2e}, \dots, v_{Ne}\}$.

(6) Normalizuj niz empirijskih volumena u skladu sa sledećim izrazom: $v_{ien} = v_{ie} / \sum_{i=1}^{N} v_{ie}$ so that $\sum_{i=1}^{N} v_{ien} = 1$.

End **Loop**

(1) Definiši pogodno izabran broj M stratuma $G_h, h = 1,2, \dots, M, 1 < M < N.$ i izvedi stratifikaciju V_e.

(2) Definiši ukupan volumen zauzrt od strane actuelog trening sempla na sledeći način: $V = N \sum_{i=1}^{N} v_{ien} = N$, zatim izračunaj vrednosti volumena za svaki stratum: $V_h = V/M = N/M, h = 1,2, \dots, M.$

(3) Izračunaj prag odabiranja $\theta = N/M$.

Do **Loop** od 1 do M:

(4) Za svaki od M stratuma $G_h, h = 1,2, \dots, M.$, nađi broj uključenih instanci $|G_h|$

(5) Za svaki stratum $|G_h| > \theta$, koji inicijalno ima broj od N_h instanci, nasumično smanji taj broj da se dobije novi balanirani stratum G_{hb} koji sadrži finalni broj $N_{hb} = |G_{hb}| = \theta$ slučajno izabranih instanci.

(6) Za svaki stratum sa $|G_h| < \theta$, upotrebi SMOTE algoritam za syntetički over-sampling instanci, proporcionalno volumenu V_h , kako bi se dobili novi balansirani stratumi $|G_h| > \theta$ koji sadrže broj instzanci $N_{hb} = |G_{hb}| = \theta$.

End **Loop**

(7) Sakupi sve balansirane stratum G_{hb} u novi balansuirani skup D_b: $D_b = \{x_{ib} | x_{ib} = f_{ij}, i = 1,2, \dots, N_b., j = 1,2, \dots, n.\}$

(8) Ponovi celu proceduru za druge klase i uzmi kompletan skup koji sada predstavlja dobro izbalansiran trening uzorak za klasifikaciju.

End **Process**

5.3.8. Opšti primer balansiranja baziranog na distancama

Na opštem primeru balansiranja internog disbalansa komplesnih klasa testiramo sve prednosti našeg (DBB) algoritma balansiranja, koje su teoretski dobro fundirane i dokazane u dosadašnjoj prezentaciji ovog poglavlja. Kako je rečeno, praktična ograničenja su posledica velike dimenzije uzorka ($n \gg 3$), kako u smislu analize tako i u smislu vizualizacije. Zato ovde razmatramo slučaj u dvodimenzionalnom prostoru instanci, kako bismo omogućili vizualizaciju pozitivnog uticaja DBB algoritma na distributivne karaklteristika uzorka u prostoru obeležja sa jedne strane, i korespondentnu indirektnu prezentaciju u jednodimenzionalnom prostoru srednjih (lokalnih) Euklidskih rastojanja između instanci i njihovih suseda sa druge strane. U tom smislu, ovo potpoglavlje se sastoji iz dva dela: prvi se odnosi na pomenutu direktnu prezentaciju efekata našeg algoritma a drugi se odnosi na indirektnu prezentaciju koja ima primarni praktični značaj.

5.3.9. Direktna prezentacija prednosti DBB algoritma

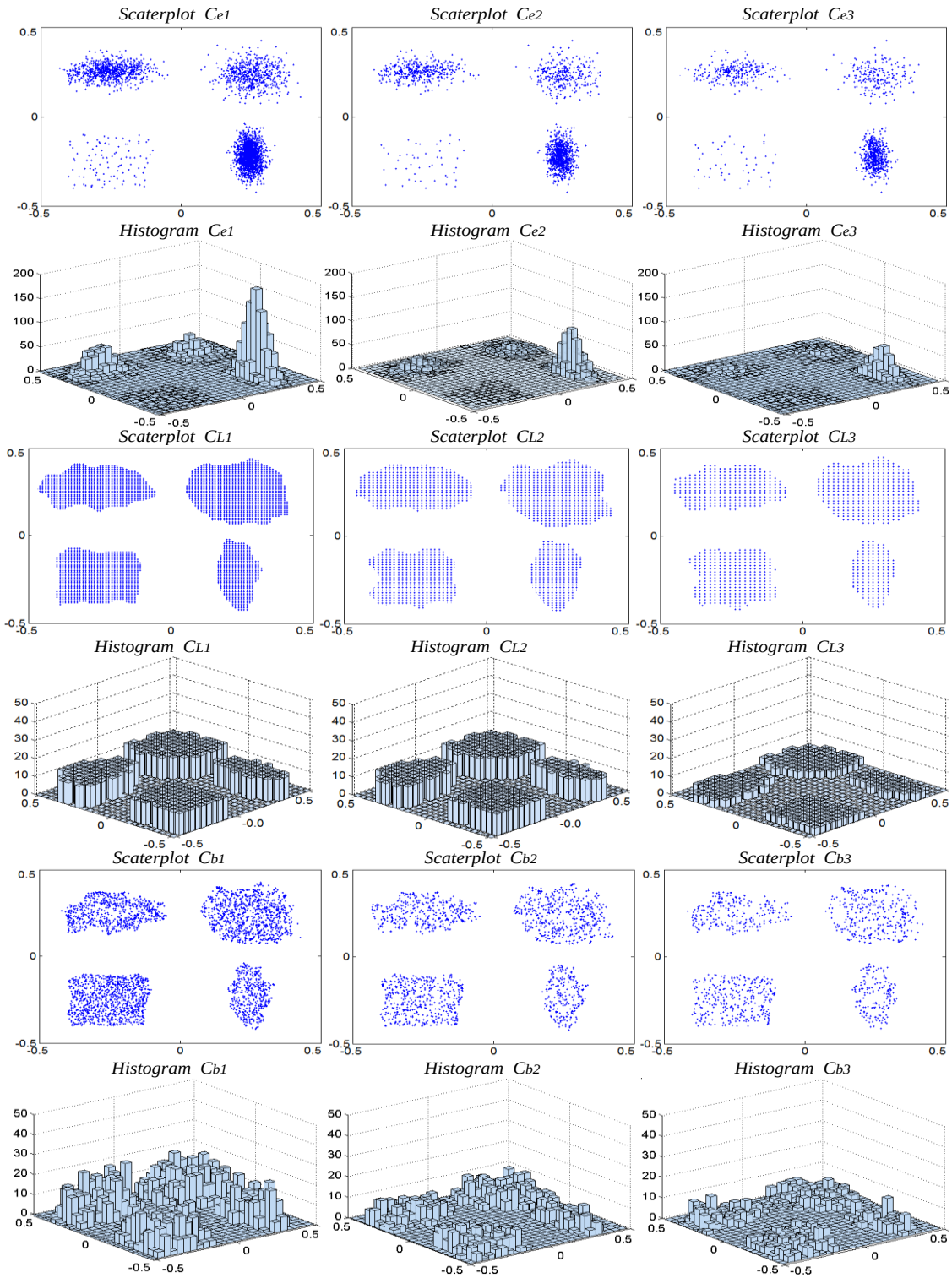
Slika 5.4 koja je data u formi matrice od (6×3) manjih slika, predstavlja rezultate DBB algoritma za odabiranje eksperimentalnih heterogenih klasa (uzoraka). Eksperimentalna, neuravnotežena klasa, prikazana je u prvom redu matrice Slika 5.4, koja se sastoji od subkoncepata različitih oblika, dimenzija i raspodela. Ovde skrećemo pažnju na strukturu

klastera u originalnom neizbalansiranim eksperimentalnim klasama (C_{e1}). Klaster originalnog uzorka koji se nalazi u prvom (I) kvadrantu ima 600 primera, klaster u drugom kvadrantu ima 800 primera, klaster u trećem kvadrantu sadrži 100 i klaster u kvadrantu IV ima 1.600 primera, što iznosi ukupno od 3100 primera. Klaster smešten u kvadrantu III ima distribuciju sličnu ravnomernoj, dok svi ostali klasteri imaju sličnu normalnoj, više ili manje iskrivljenu raspodelu instanci u 2D prostoru. Prema trećem poglavlju, reprezentativnost takvog uzorka je slaba i naš cilj je da se poveća stepen reprezentativnosti kroz balansiranje evidentne neravnoteže unutar klasa pomoću našeg novog metoda. Za potrebe detaljnog objašnjenja prenosa distributivnih karakteristika iz idealne rešetke (paradigme balansa) na neuravnoteženi empirijski uzorak pomoću DBB algoritam, moramo opisati sadržaj prikazan na Slici 5.4. Slika 5.4 je organizovana u obliku matrice od (6 × 3) malih slika. Prvi red sadrži a tačkast prokaz originalnog sintetičkog uzorka klase $C_{e1}, |C_{e1}| = 3100$ i sledećih clasa $\{ C_{e2,} |C_{e2}| = 3100{:}\,2 = 1550, C_{e3,} \approx 3100{:}\,3 = 1033\}$ izvedenih iz C_{e1} pomoću jednostavnog slučajnog odabiraanja instanci. Opšta definicija dužine ovih uzoraka prikazana je na sledeći način: $|C_{ei}| = round(3100/i), i = 1{,}2, \dots{,}3.$ Ovi uzorci se smatraju eksperimentalnim ili tzv empirijskim uzorcima u poređenju sa idealnim uzorkom pravilne rešetke i tako su uzorci tretirani u predsojećem tekstu. Treba napomenuti očigledne i istaknute neravnomerne raspodele slučajeva u skupu eksperimentalnih klasa, kao i heterogenost i raznolikost njihovih potklasa (klastera) kako u obliku tako i dimenzijama. Treba reći da drugi i treći primer klase iz prvog reda sa Slike 5.4, inače nastali slučajnom redukcijom instanci iz prvog primera, zadržavaju prirodu distribucije po celom prostoru i jedina bitna promena u odnosu na original je manji broj instanci što je lako uočljivo. Aktuelna raspodela uzoraka je uzrok njenog malog nivoa reprezentativnosti i zato su ove klase podvrgnute procesu balansiranja. Drugi red Slike 5.4 predstavlja 3D histograme sa bazom dimanzija (25×25) koji korespondiraju slikama iz gornjeg reda. Svaki histogram jasno odražava prirodu nejednake distribucije gustine verovatnoće odgovarajućeg uzorka, koju karakterišu relativno male vrednosti entropije, zbog tendencije akumuliranja velikog broja slučajeva u pojedinim manjim područjima prostora instanci. Mi koristimo normalizovane vrednosti histogram neravnotežnih uzoraka (C_e) za utvrđivanje entropija $\left(H_{C_e}\right)$ kao njihovih neposrednih komparativnih mera reprezentativnosti i neravnoteže. Treći red na Slici 5.4 predstavlja aproksimaciju pravilnih kvadratnih uzoraka rešetke $\{C_{Li}, |C_{Li}| = round(3100/i), i = 1{,}2, \dots{,}3.\}$ tako da je: $|C_{L1}| = 3100, \dots, |C_{L3}| = 1033$ koje su ravnomerno raspoređene po domenu koji zauzimaju odgovarajući empirijski uzorci iz prvog reda Fig. 4. Ovaj treći red sa Slike 5.4 apriori predstavlja ravnomerno raspoređene uzorke u smislu maksimalne reprezentativnosti i

ravnoteže praćenih maksimalnom entropijom. Krajnji očekivani benefit našeg DBB algoritma je prihvatljiva transformacija originalnih neravnotežnih uzoraka prema uniformnim uzorcima strukture slične odgovarajućim uzorcima rešetke.

Na osnovu njenih idealnih distributivnih karakteristika, usvojili smo pravilnu rešetku C_{iL} kao idealnu paradigmu ravnomernosti, reprezentativnosti i ravnoteže, tako da se procena odgovarajućih svojstva uzoraka C_{ie} i C_{ib} svodi na poređenje ovih uzoraka sa odgovarajućim uzorcima rešetke C_{iL}. Četvrti red na Slici 5.4 prikazuje 3D histograme dimenzija osnove (25 × 25) koji odgovaraju uzorcima prikazanim na slikama koje su date u prethodnom redu i odražavaju njihovu prirodu uniformne raspodele gustine verovatnoće, koju karakterišu maksimalne vrednosti entropije u aktuelnim uzorcima. Mi koristimo vrednosti histograma rešetki uzoraka (C_{Li}) da se utvrde entropija ($H_{C_{Li}}$)

kao komparativne mere idealne ravnoteže i reprezentativnosti. Peti red Slike 5.4 prikazuje balansirane uzorke $\{C_{bi}, |C_{bi}| = round(3100/i), i = 1,2, \dots ,3.\}$, Tako da su uzorci $|C_{b1}| = 3100, \dots , |C_{b3}| = 1033$ dobijeni balansiranjem odgovarajućih originalnih uzoraka iz prvog reda Fig. 4 pomoću DBB algoritam. Ovi uzorci imaju visok stepen ravnomernosti distribucije u odnosu na odgovarajuće neravnotežne originale iz prvog reda Fig. 4, i imaju sličnu strukturu sa odgovarajućim rešetkama, što je evidentno, i, na direktan način potvrđuje očekivane rezultate primene DBB algoritma. Konačno, šesti red na Slici 5.4 prikazuje histograme dimenzije baze (25 × 25) koji odgovaraju balansiranim uzorcima petog reda Slike 5. 4. Histogrami takođe pokazuju visok stepen uniformnosti distribucije gustine verovatnoće po celom prostoru instanci koji potvrđuje sličnost balansiranih uzoraka sa odgovarajućim paradigmama (rešetkama) i očiglednu razliku od originala neravnotežnih pandana. Na osnovu posmatranja dijagrama rasturanja i histograma datih na Slici 5. 4, sasvim je jasno da a) uzorci paradigme (rešetke) (C_{Li}) imaju ravnomernu distribuciju gustine verovatnoće u prostoru instanci i maksimalnu reprezentativnost koju prati maksimalni vrednost entropije, b) originalni neuravnoteženi uzorci (C_{ei}) imaju distribuciju sa najnižim stepenom uniformnosti i minimalne reprezentativnosti kojui prati najniža vrednost entropije u odnosu na ostale uzorke, i, c) finalni balansirani (C_{bi}) imaju distributivne osobine koje su veoma slične odgovarajućim vrednostima uzoraka u formi rešetke, tako da reprezentativnost ovog uzorka teži maksimuma, odnosno znatno je veća u odnosu na original a neznatno manja u odnosu na idealni uzorak.



Slika 5.4. Prezentacija direktnih distributivnih karakteristika originalnog, idealnog i balansiranog uzorka.

Prateći numerički oblik neposrednog predstavljanja prednosti algoritma je dat u levom delu Tabele 5.4 (kolone 1 do 5), koje predstavljaju distributivne karakteristike stvarnih *2D* uzoraka izloženih na Slici 5.4. Prva kolona u Tabeli 5.4 sadrži imena uzoraka gde se prvih šest redova odnosi na neravnotežne originale, narednih šest redova predstavljaju idealno distribuirane uzorake rešetkaka i posljednjih šest redova predstavljaju uravnotežene uzoraka dobijenih od

strane DBB algoritma. Druga kolona predstavlja odgovarajući broj instanci (dužina) svih uzoraka. Treća kolona predstavlja odgovarajuće vrednosti entropija određene na osnovu normalizovanih $3D$ histograma sa osnovom od (25 × 25) kvadratnih binova, tako da je $\sum_i f_i b_i^2 = 1$, gde je $i = 625, f_i$ predstavlja normalizovan učestalost svakog sloja i b_i predstavlja širinu bina koja je u ovom slučaju: $b_i = 1/625 = 0.04$, jer je širina cele oblasti ima jediničnu vrednost, vidi Sliku 5.4.

Entropija uzorka je prvi pokazatelj efikasnosti algoritma i moramo razjasniti njenu ulogu. Kao što vidimo idealno distribuirane $2D$ rešetke $C_{Li}, i = 1,2,..,6.$, imaju maksimalne moguće entropija $H_{Li}, i = 1,2,\dots,6,$ u odgovarajućem broju slučajeva $H_{L1} = H_{L2} =,..,= H_{L6} \approx 8.0444$ dok eksperimentalni neizbalansirani (original) uzorci $C_{ei}, i = 1,2,..,6,$ imaju min. entropij $H_{ei}, i = 1,2,\dots,6., H_{e1} = H_{min} = 6.4507, H_{e2=}6.4582, \dots, H_{e6=}6.2732$. Odgovarajući balansirani uzorci $C_{bi}, i = 1,2,..,6,$ imaju veće vrednosti entropije i vrlo blizu maksimalnim vrednostima odgovarajućih uzoraka rešetki, $H_{b1} = 7.8523, H_{b2} = 7.8174, \dots, H_{b6} = 7.7320,$ kao što smo pretpostavljali i očekivali. Dakle, opšti odnosi entropija odgovarajućih uzorka su $H_{C_{Li}} > H_{C_{bi}} > H_{C_{ei}} \wedge H_{C_{Li}} \approx H_{C_{bi}}$ Druga, ali najosetljiviji direktni pokazatelj efikasnosti algoritma je dgovarajući Kulback-Leibler Divergencija (D_{KL}). Vrednosti D_{KL}, date u koloni 4. Tabele 5.4 predstavljaju relativnu entropiju datih uzoraka u odnosu na odgovarajuće idealnu rešetku C_{Li}. Relativno veća vrednost D_{KL} ukazuju na veću razliku između entropije ispitivanih uzoraka, tako da je razlika između dva identična uzorka jednaka nuli. Stoga, $D_{KL}(C_L||C_L) = 0$, pogledati redove 7 do 12 u koloni 4, dok je na drugoj strani $D_{KL}(C_L||C_e) > 0$ i $D_{KL}(C_L||C_b) > 0$, videti redove 1 do 6 i redove 13 do 18 u istoj koloni Tabele 5.4 respektivno. U pomenutoj koloni 4 se jasno vidi da su vrednosti odstupanja neravnotežnih eksperimentalnih uzoraka od odgovarajućih idealnih uzoraka rešetki značajno veći od vrednosti odstupanja izbalansiranih uzoraka od svojih idealnih pandana $D_{KL}(C_L||C_e) > D_{KL}(C_L||C_b)$. U koloni 5 Tabele 5.4 su date vrednosti redundanse kao drugačijom merom reprezentativnosti izračunatih za sve uzorke u skladu sa stavom 3.1.2. U ovoj koloni vrednosti entropije rešetke $H_L = 8.0444$ su ekvivalentne maksimalnoj vrednosti H_{max} $(H_{max} = H_L)$, jer uzorci rešetki predstavljaju ravnomernu raspodelu slučajeva. Kao što možemo videti redundanse reghularnih uzoraka rešetki C_L imaju nulte vrednosti (redovi 7 do 12 u koloni 5), dok redundanse neravnotežnih uzoraka C_e (redovi 1 do 6 u koloni 5), i balansiranih uzoraka C_b (redovi 13 do 18 u koloni 5) imaju vrednosti veće od nule. Suštinska informacija ove kolone je sadržana u činjenici da su vrednosti redundanse izbalansiranih uzoraka C_{ib} za red veličine manje od odgovarajućih vrednosti neuravnoteženih uzoraka C_{ie}, što jasno ukazuje na veću reprezentativnost izbalansiranih uzoraka u odnosu na

nebalansirane originale. Na osnovu rezultata navedenih u Tabeli 5.4 i na Slici 5.4 dokazano da je sličnost između finalnih balansiranih uzoraka C_{ib} i odgovarajućih uzoraka rešetki C_{iL} je značajno veća nego sličnost između korespondentnih parova C_{ie} i C_{iL}. Ova sličnost implicira povećanu reprezentativnost balansiranih uzoraka u odnosu na neuravnotežne originale. Ovaj dokaz je direktna manifestacija potencijala DBB resampling algoritma da izbalansira proizvoljni složeni neuravnoteženi uzorak, menjajući njgovu strukturu u smeru strukture pravilne rešetke, kao idealne paradigme uniformnosti, reprezentativnosti i ravnoteže.

5.3.10. Indirektna prezentacija prednosti DBB algoritma

Podsetimo opet da je direktna prezentacija i analiza distributivnih svojstava uzoraka moguća samo za uzorke malog broja dimenzija ($n \leq 3$) kao u slučaju prikazanom na Slici 5.4, pa u nastavku teksta, dajemo detaljan pregled indirektnog prezentacije i analize distributivnih svojstava uzoraka u jednodimenzionalnom prostoru lokalnih srednjih rastojanja. Takođe treba podsetiti da je stratifikovani resampling, kao sredstvo balansiranja uzoraka u cilju povećanja njihove reprezentativnosti, metoda koja je takođe ograničena na uzorke sa malim brojem dimenzija, pre svega zbog visoke složenosti izračunavanja. Zbog ovih ograničenja, uvodimo novu indirektnu analizu i prezentaciju distributivnih osobina multidimenzionalnih uzoraka, koristeći vrednosti srednjih lokalnih rastojanja instanci od fiksnog broja (KNN) njihovih suseda za svaku instancu. Koristeći pomenutu indirektnu korespondenciju, zamenjujemmo proces balansiranja u multidimenzionalnom prostoru instanci balansiranjem u jednodimenzionalnom prostoru lokalnih rastojanja, dodajući na taj način novi (DBB) algoritam visokog stepena opštosti kategoriji algoritama za uravnoteženje. Detaljna grafička indirektna prezentacija rezultata DBB postupka primenjenog na neravnotežnim uzorcima sa Slike 5.4 je data na Slici 5.5, čije razumevanje zahteva neke dodatne teoretske činjenic u vezi sa indirektnim distributivnim svojstvima pravilne rešetke. Razmotrimo rastojanja između tačaka pravilne n-dimenzionalne kubne rešetke koja se sastoji od N_L tačaka ravnomerno distribuiranih po celom prostoru instanci. Izračunajmo Euklidova rastojanja d_{ik}, za svaku tačku rešetke $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, od svih svojih susedima $x_k = \{x_{k1}, x_{k2}, \dots, x_{kn}\}, i, k = 1,2, \dots N_L$. kako sledi: $d_{ik} = d(x_i, x_k) = \sqrt{\sum_{j=1}^{n}(x_{i,j} - x_{k,j})^2}$, $i, k = 1,2, \dots, N_L., j = 1,2, \dots, n.$ Po definiciji, ivica osnovne ćelije kubne rešetke ima konstantnu vrednost, koja može biti jedinična vrednost ($a_L = 1$) u slučaju celobrojne rešetke ili neke druge vrednosti $a_L \ll 1$ u slučaju guste pravine rešetke iz realnog domena. Volumeni koji pripadaju svakoj instanci imaju vrednost $v_{iL} = a_L^n$.

Svaka tačka $x_i, i = 1,2, \dots, N_L,$, 2-dimenzionalne kvadratne rešetke date na Slici. 3 ima četiri($k_1 = 4$) suseda prvog reda ($r = 1$) čija je udaljenost $d_{1L} = a_L$ i četiri($k_2 = 4$) suseda drugog reda ($r = 2$) udaljenosti $d_{2L} = a_L\sqrt{2}$, dajući na taj način srednje euklidsko rastojanje $d_L = (4d_{1L} + 4d_{2L})/8 = (4a_L + 4a_L\sqrt{2})/8 = a_L(1 + 2\sqrt{2}1)/2 = 1.2071a_L$, prema jednačinama. 15, 16 i 18. Takođe, sva ostala prosečna rastojanja definisana za isti fiksni broj kNN tačaka su identična sa odgovarajućim realnim vrednostima $d_{iL} = d_L, i = 1,2, \dots, N_L$ u skladu sa jednačinom. (19) i Tabelom 5.3. Razmotrimo slučaj pravilne 2-D rešetke sa predefinisanim brojem tačaka N_L i predefinisanom ukupnom zapreminom $\mathbb{V}_L$. Pošto je ukupni volumen 2-D rešetke definisan na sledeći način: $\mathbb{V}_L = \sum_{i=1}^{N_L} v_{iL} = \sum_{i=1}^{N_L} a_{iL}^2 = \sum_{i=1}^{N_L} a_L^2 = N_L a_L^2$, imaćemo: $a_{iL} = a_L = \sqrt[2]{V_L/N_L}$. U slučaju ndimenzionalne pravilne rešetke odgovarajući volume je: $\mathbb{V}_L = \sum_{i=1}^{N_L} v_{iL} = \sum_{i=1}^{N_L} a_{iL}^n = \sum_{i=1}^{N_L} a_L^n = N_L a_L^n$ pa imamo: $a_{iL} = a_L = \sqrt[n]{V_L/N_L}$.

Tabela 5.4

Direktne i indirektne distributvne karakteristike aktuelnih $2D$ uzoraka.

C	l	H	$D_{KL}(C_L\|\|C$	R =1-H/H_L		D	l	min	max	μ	σ	H	$D_{KL}(D_L\|\|D$
C_{e1}	3100	6.4807	0.5703	0.1944		D_{e1}	3100	**0.0014**	0.0728	**0.0087**	**0.0088**	**3.5729**	**4.9224**
C_{e2}	1550	6.4582	0.4918	0.1972		D_{e2}	1550	0.0021	0.0860	0.0119	0.0115	4.0026	5.1043
C_{e3}	1033	6.4448	0.3885	0.1989		D_{e3}	1033	0.0027	0.0933	0.0144	0.0134	4.2143	4.9224
C_{e4}	775	6.3970	0.2738	0.2048		D_{e4}	775	0.0037	0.1141	0.0170	0.0169	4.4623	5.5069
C_{e5}	620	6.3420	0.2315	0.2116		D_{e5}	620	0.0032	0.1232	0.0187	0.0185	4.5838	5.9495
C_{e6}	516	6.2732	0.1285	0.2200		D_{e6}	516	**0.0042**	**0.1458**	**0.0199**	**0.0191**	**4.6396**	**6.6837**
C_{L1}	3100	8.0444	0.0000	0.0000		D_{L1}	3100	0.0108	0.0108	0.0108	0.0000	0.0000	0.0000
C_{L2}	1550	8.0444	0.0000	0.0000		D_{L2}	1550	0.0153	0.0153	0.0153	0.0000	0.0000	0.0000
C_{L3}	1033	8.0444	0.0000	0.0000		D_{L3}	1033	0.0188	0.0188	0.0188	0.0000	0.0000	0.0000
C_{L4}	775	8.0444	0.0000	0.0000		D_{L4}	775	0.0217	0.0217	0.0217	0.0000	0.0000	0.0000
C_{L5}	620	8.0444	0.0000	0.0000		D_{L5}	620	0.0242	0.0242	0.0242	0.0000	0.0000	0.0000
C_{L6}	516	8.0444	0.0000	0.0000		D_{L6}	516	0.0266	0.0266	0.0266	0.0000	0.0000	0.0000
C_{b1}	3100	7.8523	0.3450	0.0239		D_{b1}	3100	**0.0038**	**0.0324**	**0.0108**	**0.0027**	**2.8476**	**2.1604**
C_{b2}	1550	7.8174	0.2978	0.0282		D_{b2}	1550	0.0059	0.0452	0.0205	0.0048	3.5537	3.5518
C_{b3}	1033	7.7919	0.3190	0.0314		D_{b3}	1033	0.0084	0.0437	0.0192	0.0048	3.6447	2.7524
C_{b4}	775	7.7863	0.1718	0.0321		D_{b4}	775	0.0106	0.0572	0.0224	0.0059	3.8898	2.9361
C_{b5}	620	7.7505	0.1008	0.0365		D_{b5}	620	0.0134	0.0716	0.0256	0.0063	3.9190	3.0620
C_{b6}	516	7.7320	0.0558	0.0388		D_{b6}	516	**0.0153**	**0.0708**	**0.0277**	**0.0068**	**3.9618**	**2.8791**

Na ovaj način izračunavamo dužinu ivice proizvoljne jedinice pravilne rešetke čiji centar predstavlja poziciju odgovarajuće instance, i dalje lako izračunavamo srednje rastojanje ove instance od predefinisanog broja najbližih susednih tačaka a u skladu sa jednačinom (19). Pošto je ova srednja vrednost identična za sve instace rešetke dobićemo niz D_L sastavljen od identičnih realnih vrednosti srednjih lokalnih rastojanja za sve instance: $D_L = \{d_{iL}, d_{iL} =$

$d_{jL} = d_L, \forall i,j = 1,2,..,N_L\}$, gde d_L zavisi od dimenzije n uzorka (rešetke) i predefinisanog broja suseda kNN. U prikazanom eksperimentalnom uzorku imamo $n = 2$ i $kNN = 8$, dakle $d_L = d_{iL} = a_{iL}(1+\sqrt{2})/2 = (\sqrt[2]{V_L/N_L}\quad)\,(1+\sqrt{2})/2$. Nizovi D_L predstavljaju jednodimenzionalnu variablu sa degenerativnom (determinističkom) raspodelom verovatnoće ogrtaničenom skupom osobina formalno datih u Jed. (30).

$$\left.\begin{array}{c} P(X=c)=1 \\ F(x,c)=\begin{cases}1, if\ x \geq c \\ 0, if\ x < c\end{cases} \\ f(x)=\lim_{a\to 0}\delta_a(x), \delta_a(x)=\frac{1}{a\sqrt{\pi}}e^{-x^2/a^2}\end{array}\right\}, \qquad (30)$$

gde X, x, c predstavljaju D_L, d_{iL}, d_L respektivno, $P(X)$ predstavlja raspodelu verovatnoće , $F(x)$ je kumulativna funkcija raspodele verovatnoće, $f(x)$ je funkcija gustine raspodele (pdf), i $\delta_a(x)$ predstavlja the Dirakovu delta funkciju. U cilju objašnjenja indirektne komparacije distributivnih karakteristika neizbalansiranih I izbalansiranih uzoraka prikazanih u Tabeli 5.4 specificiramo broj primera eksperimentalnih (empirijskih) neizbalansiranih uzoraka dati u 2D prostoru obeležja na sledeći način: $\{N_{ei} = |C_{ei}| = 3100/i, i = 1,\ldots,6.,\}$. Komparativni uzorci-rešetke imaju identične dužine (broj instanci) sa svojim eksperimantalnim pandanima$\{N_{Li} = N_{ei}, i = 1,\ldots,6.\}$. Izbalansirani finalni uzorci takođe imaju iste dimenzije sa svojim originalnim parovima iz kojih su nastali primenom DBB algoritma, $\{N_{bi} = N_{ei}, i = 1,\ldots,6.\}$. Treba uočiti da uzorci iz istih kolona sa Slike 5.4 pripadaju istoj kategoriji u smislu broja elemenata (dužina).

U narednom tekstu ćemo koristiti termine *eksperimentalni* ili *empirijski* uzorci (e) za originalne uzorcima iz prvog reda Slike 5.4, *rešetke* ili idealni uniformni uzorci (L) se koriste za uzorke prikazanih u trećem redu Slike 5.4, dok termin *izbalansirani* uzorci (b) rezervisane su za uzorke trećeg reda Slike 5.4. U sledećoj analizi ćemo koristiti vrednost ukupne korisne zapremine $\mathbb{V}$ zauzete od strane uzoraka, koji mogu biti empirijski izračunati sa prihvatljivom tačnošću, i ukupnog a priori određenog broja N slučajeva distribuiranih u zapremini $\mathbb{V}$ kako bismo definisali opšte uslove koji su od suštinskog značaja za indirektno poređenja uzoraka i evaluaciju efikasnosti našeg DBB algoritma za uravnoteženje uzoraka.

Uvodimo sledeeće pojmove u cilju daljeg izlaganja: fiktivni ekvivalentni volumeni kocki v_i definisani u formi funkcijke: $v_i = f(\mathbb{V}, N)$; eqvivalentna ivica a_i pomenutih fiktivnih kocki određena na sledeći način: $a_i = f(v_i) = \sqrt[n]{v_i}$, gde n predstavlja broj obeležja instani, u našem konkretnom slučaju (Slika 5.4) je $n = 2$; i pojam srednjeg rastojanja d_i definisanog kako sledi: $d_i = f(a_{i,}kNN)$, u skladu sa Jed. (21) gde je kNN predefinisani broj aktuelnih suseda zavisan od dimenzije uzorka. U skladu sa ovim pretpostavkama, koristimo termin empirijski

ekvivalentni volumen v_{ie} za volume fiktivne kocke koje sadrže c_{ie}, dužina ivica ekvivalentnih fiktivnih kocki (a_{ie}), i prosečno rastojanje (d_{ie}) od unapred definisanih suseda koje se može lako izračunava na osnovu datog broja dimenzija i izabranog adekvatnog broja suseda. Treba napomenuti da se definicija ovih pojmova zasniva na vrednosti v_{ie} datoj jednačinom. (22). Takođe uvodimo koncept volumena jedinice rešetke v_{iL} , ivicu jedinične kocke rešetke (a_{iL}) i pojam prosečnog rastojanja (d_{iL}) od unapred definisanih susednih instanci rešetke. Ove vrednosti su povezani sa svakom instancom rešetke. U osnovi definicije ovih termina stoji identitet $v_{iL} = a_{iL}^2$, koji predstavlja osnovnu karakteristiku jednostavne pravilne kubne rešetke. Odgovarajući pojmovi koji se odnose na izbalansirane uzoraka su ekvivalent fiktivnih zapremina kocki sa oznakom v_{ib} , odgovarajuća ekvivalentna ivicu (a_{ib}) i srednja rastojanja (d_{ib}) od unapred definisanog skupa susednih instanci. Ovi pojmovi su takođe pridruženi svakoj instanci izbalansiranih uzoraka i definicija ovih pojmova u osnovi stoji na sledećoj relaciji aproksimativne jednakosti$(v_{ib} \approx a_{ib}^2)$, koja je takođe potvrđena u Jed. (22). Ova približna jednakost je zasnovana na očiglednoj sličnosti uzoraka datih u trećem i petom redu na Slici 5.4. Definišimo nizove vrednosti lokalnih srednjih rastojanja za empirijske uzorka na sledći način: $D_{e1} = \{d_{ie1},\ i = 1, 2, .., N_{e1}\},\ D_{e2} = \{d_{ie2},\ i = 1, 2, .., N_{e2}\},\ D_{e3} = \{d_{ie3},\ i = 1, 2, .., N_{e3}\}$, nizove vrednosti srednjih lokalnih rastojanja rešetkastih uzoraka kao: $D_{L1} = \{d_{iL1},\ i = 1, 2, .., N_{L1}\},\ D_{L2} = \{d_{iL2},\ i = 1, 2, .., N_{L2}\},\ D_{L3} = \{d_{iL3},\ i = 1, 2, .., N_{L3}\}$ i nizove vrednosti srednjih lokalnih rastojanja izbalansiranih uzoraka kao: $D_{b1} = \{d_{ib1},\ i = 1, 2, .., N_{b1}\},\ D_{b2} = \{d_{ib2},\ i = 1, 2, .., N_{b2}\},\ D_{b3} = \{d_{ib3},\ i = 1, 2, .., N_{b3}\}$. Svaki od gore definisanih pojmova su osnovni pojmovi neophodni za objašnjenja distributivnih karakteristika svih analiziranih uzoraka. Na osnovu prethodnih pojmova definišemo strukturu totalnih zapremina aktuelnih uzoraka koristeći sledeće izraze:

$\mathbb{V}_{e1} = \sum_{i=1}^{N_{e1}} v_{ie1} = \sum_{i=1}^{N_{e1}} a_{ie1}^2$, $\mathbb{V}_{e2} = \sum_{i=1}^{N_{e2}} v_{ie2} = \sum_{i=1}^{2} a_{ie2}^2$, $\mathbb{V}_{e3} = \sum_{i=1}^{N_{e3}} v_{ie3} = \sum_{i=1}^{N_{e3}} a_{ie3}^2$,

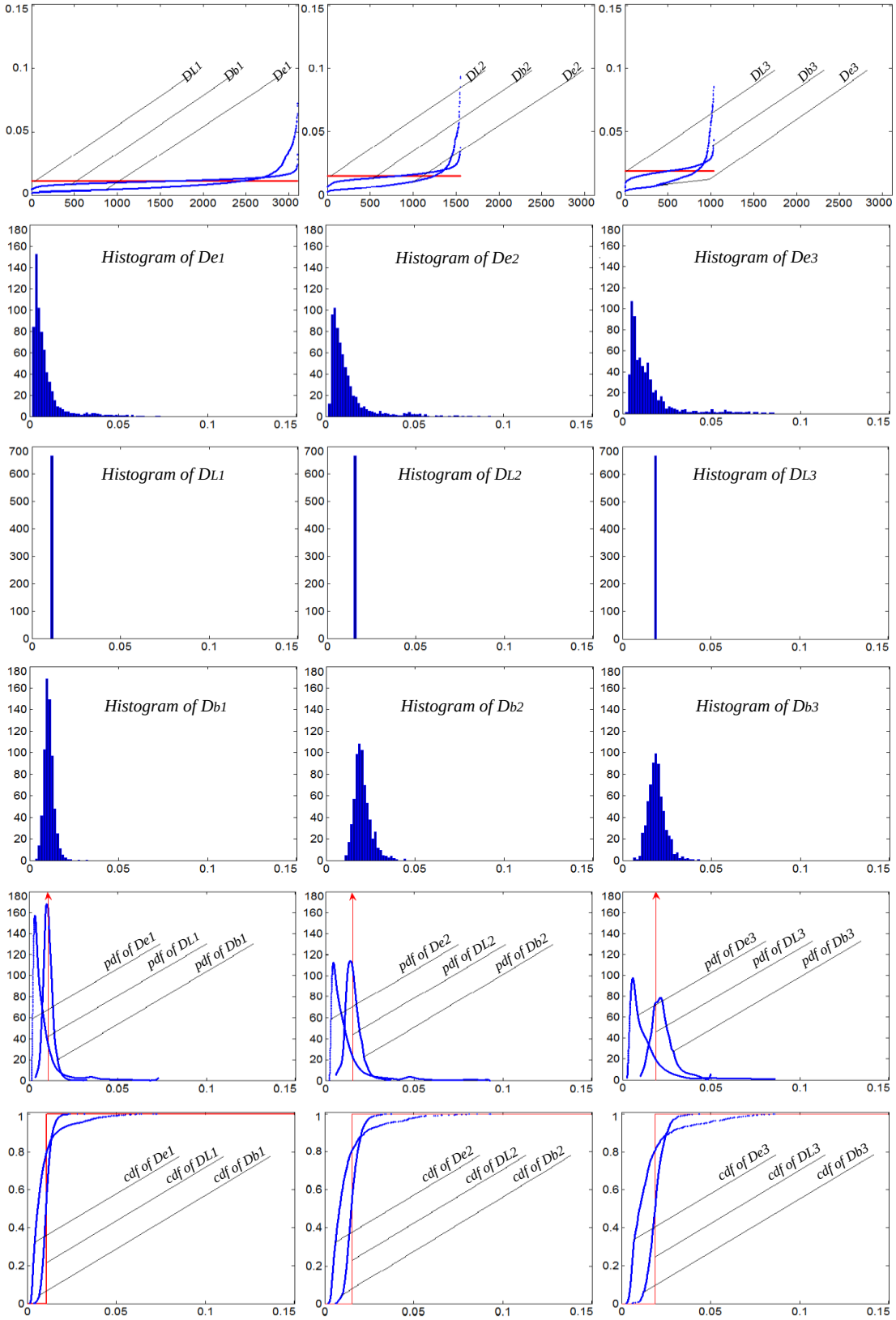
$\mathbb{V}_{L1} = \sum_{i=1}^{N_{L1}} v_{iL1} = \sum_{i=1}^{N_{L1}} a_{iL1}^2$, $\mathbb{V}_{L2} = \sum_{i=1}^{N_{L2}} v_{iL2} = \sum_{i=1}^{N_{L2}} a_{iL2}^2$, $\mathbb{V}_{L3} = \sum_{i=1}^{N_{L3}} v_{iL3} = \sum_{i=1}^{N_{L3}} a_{iL3}^2$, i

$\mathbb{V}_{b1} = \sum_{i=1}^{N_{b1}} v_{ib1} = \sum_{i=1}^{N_{b1}} a_{ib1}^2$, $\mathbb{V}_{b2} = \sum_{i=1}^{N_{b2}} v_{ib2} = \sum_{i=1}^{2} a_{ib2}^2$, $\mathbb{V}_{b3} = \sum_{i=1}^{N_{b3}} v_{ib3} = \sum_{i=1}^{N_{b3}} a_{ib3}^2$, gde indeksi 1, 2, i 3 odgovaraju pridruženim kolonama sa Slike 5.4.

Pošto DBB procedura ne izlazi van domena $X = \{([-0.5, 0.5], [-0.5, 0.5]), \mathbb{V}_X = 1\}$ originalnih uzoraka, ukupni korisni proctor svih individualnih uzoraka sa Slike 5.4 imaju sledeće izračunate vrednosti: $\mathbb{V}_{e1} \approx \mathbb{V}_{e2} \approx \mathbb{V}_{e3} \approx 0.25$, $\mathbb{V}_{L1} \approx \mathbb{V}_{L2} \approx \mathbb{V}_{L3} \approx 0.25$, $\mathbb{V}_{b1} \approx \mathbb{V}_{b2} \approx \mathbb{V}_{b3} \approx 0.25$, gde bez bitnih posledica možemo zadržati stav da zapremine svih uzoraka imaju približno iste vrednosti: $\mathbb{V}_e = \mathbb{V}_L = \mathbb{V}_b = 0.25$.

Očigledno je da ivice ekvivalentnih fiktivnih kocki empirijskih uzoraka nisu identične međusobno: $a_{ie1} = a_{je1}, a_{ie2} \neq a_{je2}, a_{ie3} \neq a_{je3}, \forall i \neq j$ i da nije moguće direktno i tačno odrediti pojedinačne vrednosti ivica i vrednosti korespondentnih vaolumena v_{ie} i zato u ovu svrhu koristimo Jed. (22). Gornje nejednakosti ivica a_{ie} fiktivnih hiperkocki
Su posledica neravnomerne raspodele instanci koje se dalje prenose na emprijski izračunata lokalna srednja rastojanja $d_{ie,}d_{ie1} \neq d_{je1}, d_{ie2} \neq d_{je2}, d_{ie3} \neq d_{je3}, \forall i \neq j$.

Unutar rešetkastih uzoraka je potpuno drugačija situacija u smislu pomenutih veličina i relacija: $\{a_{iL1} = a_{jL1} = a_{L1}, a_{iL2} = a_{jL2} = a_{L2}, a_{iL3} = a_{jL3} = a_{L3}, \forall i, j\}$. Na osnovu prethodnog dobijamo sledeće relacije: $\{d_{iL1} = d_{jL1} = d_{L1}, d_{iL2} = d_{jL2} = d_{L2}, d_{iL3} = d_{jL3} = d_{L3}, \forall i, j\}$. Pomenuti pojmovi i relacije izbalansiranih uzoraka se određuju na osnovu korespondentnih pojmova i relacija empirijskih uzoraka i rešetkastih uzoraka. Pošto su karakteristike izbalansiranih uzoraka vlo slične osobinama rešetkastih uzoraka (Slika 5.4), možemo usvojiti sledeću releciju: $\{a_{ib1} \approx a_{jb1}, \dots, a_{ib3} \approx a_{jb3}, \forall i, j\}$ i konsekventno: $\{d_{ib1} \approx d_{jb1}, \dots, d_{ib3} \approx d_{jb3}, \forall i, j\}$. Kako smo već rekli, komponente d_{ie}, d_{ib} respektivnih nizova $D_{e,}D_b$, izračunavamo u saglasnosti sa Jed. (21), dok vrednosti komponenti d_{iL} niza D_L možemo odrediti pomoću jednog od sledećih načina: a) Pomoću izračunatih procena ukupnih zauzetih prostora od strane rešetkastih uzoraka ($\{\mathbb{V}_{L1}, \dots, \mathbb{V}_{L3}\}$) i a priori datih vrednosti dužina uzoraka ($\{N_{L1} = 3100,\ N_{L2} = 1550,\ N_{L3} = 1033,\}$), i b) Koristeći Jed. (21). Na osnovu inherentznog identiteta jediničnih volumena rešetkastog uzorka ($v_{iL} = v_{jL} \forall i, j$) izračunavamo sledeće vrednosti zajedničkh veličina $v_{iL} = \mathbb{V}_L / N_L$, zatim izračunavamo ivice kocki: $a_{iL} = \sqrt[2]{v_i}$ i na kraju, na osnovu Jed. . (19) dobijamo : $d_{iL} = a_{iL}(1 + \sqrt{2})/2$. Primenom ovih proračuna na uzorke iz prve vrste Slike 5.2 dobijamo:
$\{v_{iL1} = 0.25/3100 = 0.000080, a_{iL1} = \sqrt[2]{0.000080} = 0.008980, d_{iL1} = 0.008980 * (1 + \sqrt{2})/2 = 0.010840, i = 1{,}2, \dots, 3100\}$. Dalje, imamo: $\{v_{iL2} = 0.000161, a_{iL2} = 0.012700, d_{iL2} = 0.015330, i = 1{,}2, \dots, 1550, v_{iL3} = 0.000242, a_{iL3} = 0.015556, d_{iL3} = 0.018778, i = 1{,}2, \dots, 1033, v_{iL4} = 0.000322, a_{iL4} = 0.020080, d_{iL4} = 0.021680, i = 1{,}2, \dots, 775\}$. Izračunavanjem individualnih vrednosti srednjih rastojanja d_{iL} za sve uzorke, formiramo nizove D_L, $D_{L1} = \{d_{iL1}, d_{iL1} = 0.010840, \forall i = 1, 2, .., 3100\}, D_{L2} = \{d_{iL2}, d_{iL2} = 0.015330,\ \forall i = 1, 2, .., 1550\}, D_{L3} = \{d_{iL3}, d_{iL3} = 0.018778, \forall i = 1, 2, .., 1033\}, D_{L4} = d_{iL4},\ d_{iL4} = 0.021680, \forall i = 1, 2, .., 775\}$.



Slika 5.5. Nizovi aktuelnih srednjih lokalnih rastojanja *D* i njihove distributivne karakteristike: histogrami, *pdf* i *cdf* funkcije.

Distributivne karakteristike dobijenih *1D* nizova srednjih lokalnih rastojanja $D_{e1}, \dots, D_{e6}, D_{L1}, \dots, D_{L6}, D_{b1}, \dots, D_{b6}$, koji se odnose na odgovarajuće *2D* uzorke

$C_{e1}, \dots, C_{e6}, C_{L1}, \dots, C_{L6}, C_{b1}, \dots, C_{b6}$, date u numeričkoj formi u desnom delu Tabele 5.4 (kolone 6 do 13), a njihova odgovarajuća grafička prezentacija je data na Slici 5.5. Prvi indirektni indicator potencijala DBB algoritma za balansiranje neizbalansiranih uzoraka je dat u formi (6×3) matrice manjih sa Slike 5.5 Pri direktnoj prezentaciji distributivnih karakteristika uzoraka, datoj u prostoru obeležja (Slika 5.4), usvojili smo rešetkasti uzorak C_{Li} kao idealnu paradigm uniformnosti, reprezentativnosti i balansa kako bismo pojednostavili evaluaciju odgovarajućih karakteristika za uzorke C_{ei} i C_{bi}.

Pri indirektnoj prezentaciji distributivnih karakteristika uzoraka, za svaki uzorak C generišemo jednodimenzionalni niz vrednosti srednjih lokalnih rastojanja D koje se odnose na svaku instancu uzoraka, a u skladu sa Jed. (19) i Jed. (21). Niz D u ovom slučaju predstavlja kondenzovana obeležja koja su indirektni reprezenti uzorka C, iz kog su ekstrahovana. Pošto smo već definisali paradigmu idealnog uzorka C_{Li}, iz nje ćemo ekstrahovati odgovarajući niz obeležja u obliku srednjih lokalnih rastojanja D_{Li} koji ćemo usvojiti kao idealni indirektni reprezent uzorka C_{Li} iz kog je izvedena. Zahvaljujući ovoj činjenici ocena karakteristika ostalih uzoraka C_{ei} i C_{bi} se može redukovati na jednostavnu komparaciju njihovih nizova srednjih rastojanja D_{ei} i D_{bi} sa nizom D_{Li}. (Slika 5.5) predstavlja izvedene distributivne karakteristike skupa uzoraka $\{C_{ei}, C_{Li}, C_{bi}, N_{ei} = N_{Li} = N_{bi} = 3100/i, i = 1, \dots, 3.\}$ u indirektnoj formi preko skupa odgovarajućih indirektnih reprezenata $\{D_{ei}, D_{Li,} D_{bi}, l_{D_{ei}} = N_{ei}, l_{D_{Li}} = N_{Li}, l_{D_{bi}} = N_{bi}\}$, respektivno, gde N predstavlja broj instanci odgovarajućeg uzorka dok l predstavlja dužinu odgovarajućih nizova. Svaka i-ta kolona sa Slike 5.5 prikazuje karakteristike i-tog uzorka dužine $l_i = 3100/i$, što znači da prva kolona prikazuje niz dužine $l_1 = 3100$ dok zadnja kolona predstavlja niz dužine $l_3 = 1033$. Prva vrsta na Slike 5.5 predstavlja kompoziciju originalnih vrednosti srednjih lokalnih rastojanja $D_{ei}, D_{Li}, D_{bi}, i = 1, \dots, 3$. Na svim slikama prve vrste sa Slike 5.5, vrednosti idealnih nizova D_{Li} su prikazani kao horizontalne prave linije, dok su odgovarajuće vrednosti nizova rastojanja D_{ei} i D_{bi} prikazane u sortiranom modu kako bismo jasnije uočili njihova odstupanja od paradigmi D_{Li}. Očigledno je da odstupanje od niza D_{ei} od niza D_{Li} mnogo veće od odstupanja D_{bi} od D_{Li}, što nam indirektno govori da balansirani uzorci C_{bi} imaju mnogo veći stepen uniformnosti, reprezentativnosti i ravnoteže nego original neuravnotežena uzorci C_{ei}. Ova činjenica je potvrda pretpostavljene efikasnosti DBB algoritma u procesu balansiranja uzoraka. Na osnovu drugih distributivnih karakteristika ovih nizova prikazanih na istoj slici, možemo dobiti precizniju predstavu o uzoracima iz kojih su nizovi izvedeni. Druga, treća i četvrta vrsta na Slici. 5.5 predstavljaju normalizovane vrednosti histograma lokalnih srednjih distanci D_{ei}, D_{Li} i

D_{bi} rerspektivno, sa dužinama i-tog iuzorka $l_{ei} = l_{bi} = l_{Li} = 3100/i, i = 1,2,\dots,3.$ Svaki normalizovan histogram zadovoljava sledeće jednakosti: $\sum_i f_i b_i = 1, i = 1,2,\dots,100$, gde i predstavlja broj slojeva (binova) na histogramu, f_i je normalizovana frekvencija i -tog stratuma a $b_i = 0.15/100 = 0.0015$ je širina sam i -tog bina. U trećoj vrsti Slike 5.5 su dati histogrami nizova D_{Li} kao indirektni represenati idealne rešetke C_{Li}. Posmatrajući histograme nizova D_{ei} (gornja vrsta) i D_{bi} (donja vrsta) u poređenju sa D_{Li}, jasno je da histogrami nizova D_{Li} potpuno odgovaraju maksimalnim vrednostima histograma D_{bi} nizova dok vrlo slabo korespondiraju sa maksimalnim vrednostima histograma D_{ei} nizova, što znači da postoji mnogo veća sličnost između uzoraka C_{bi} i idealne paradigme C_{Li} nego između C_{ei} i C_{Li}. Ova činjenica je indikator većeg stepena ravnomernosti raspodele instanci izbalansiranih uzoraka u poređenju sa originalnim neizbalansiranim uzorcima koji obiluiju oblastima kako vrlo velike tako i vrlo male koncentracije instanci u prostoru obeležja instanci. Peta vrsta Slike 5.5 prikazuje komparaciju estimata funkcija gustine verovatnoće nizova lokalnih srednjih rastojanja D_{ei}, D_{Li} i D_{bi}. Treba da obratimo pažnju na funkcije gustine verovatnoće nizova D_{Li} kao indirektnog reprezenta idealnog rešetkastog uzorka pošto ćemo sa njima porediti nizove D_{ei} i D_{bi}. Pošto svi članovi nizova D_{Li} imaju identične vrednosti, ovi nizovi su u skladu sa deterministićkom ili degenerativnom distribucijom, i zato njihove funkcije gustina raspodele (pdf_{Li}) predstavljaju Dirakove delta funkcije u specifičnim tačkama prikazane kao vertikalne linija sa strelicom usmerenom nagore, ukazujući na nultu variansu i beskonačnu vrednost amplitude. Treba istaći da su prikazane estimirane vrednosti funkcija gustine verovatnoća izračunate ekstrapolacijom vrednosti normalizovanih histograma nizova srednjih lokalnih rastojanja D, koristeći generalizaciju MLP ansambla, tako da je: $\int_{-\infty}^{\infty} \mathrm{f(x)dx} = 1$. Svaka od tri slike u petoj vrsti Slike 5.5 prikazuje funkcije gustine verovatnoća ($pdf_{D_{ei}}, pdf_{D_{Li}}, pdf_{D_{bi}}, i = 1,\dots,3.$) gde je evidentno da Dirakova funkcija označena kao $pdf_{D_{Li}}$ prolazi u blizini Mode $pdf_{D_{bi}}$ funkcije, odnosno u blizini maksimalne vrednosti niza D_{bi}. Ovo znači da većina od instanci balansiranog sempla C_{bi} ima vrednosti srednjih lokalnih rastojanja približno jednake srednjim vrednostima rastojanja instanci idealnih rešetkastih uzoraka C_{Li}, ukazujući na očiglednu uniformnost raspodele i interni balans balansiranih uzoraka C_{bi}, što se jednostavno može prikazati preko sledećih relacija: $D_{bi} \cong D_{Li}$ and $C_{bi} \cong C_{Li}$. Na ovim slikama, takođe su očigledne blizina i simetrija krivih $pdf_{D_{bi}}$ sa idealnim srednjim rastojanjima ($pdf_{D_{Li}}$). Ova osobina predstavlja indirektnu manifestaciju uniformnosti, reprezentativnosti i balansa korespondentnih uzoraka C_{bi}. Na istim slikama se može videti nepodudarnost između krivih $pdf_{D_{ei}}$ i $pdf_{D_{Li}}$, pošto su njihove mode međusobno udaljene u prostoru nizova D_{ei} i

D_{Li}. Očigledne su inkopatibilnosti širina i iskošenost krivih $pdf_{D_{ei}}$ u odnosu na vrednosti idealnih srednjih rastojanja ($pdf_{D_{Li}}$). Ove osobine predstavljaju indirektnu manifestaciju neravnomernosti raspodele, niskog stepena reprezentativnosti i disbalansa odgovarajućih C_{ei} uzoraka. Ove činjenice znače da izrazita većina instanci C_{ei} uzoraka ima srednja lokalna rastojanja različita od lokalnih rastojanja instanci koje pripadaju odgovarajućim idealnim rešetkastim uzorcima C_{Li} što ukazuje na izraženu neravnomernost raspodele, nizak nivo reprezentativnosti i interni imbalans C_{ei} uzoraka. Na ovim slikama su očigledne visoke vrednosti varianse neizbalansiranih uzoraka u odnosu na varianse balansiranih uzoraka što se manifestuje velikom širinom njihovih funkcija gustine raspodele. Šesta vrsta matrice Slike 5.5 prikazuje kumulativne raspodele srednjih lokalnoih rastojanja ($cdf_{D_{ei}}, cdf_{D_{Li}}, cdf_{D_{bi}}, i = 1, \dots, 3.$), observiranih uzoraka, koje potvrđuju viši stepen kongruencije krivih $cdf_{D_{bi}}$ i $cdf_{D_{Li}}$ u odnosu na krive $cdf_{D_{ei}}$ i $cdf_{D_{Li}}$, što je manifestacija sledećih relacija: $D_{bi} \cong D_{Li}$ i $C_{bi} \cong C_{Li}$. Prethodne činjenice ističu početnu različitost distributivnih karakteristika uzoraka C_{ei} i C_{Li} i veliku sličnost finalnih uzoraka C_{bi} sa pandanima C_{Li}, dokazujući predpostavljenu efikasnost DBB algoritma pri balansiranju uzoraka. Druga indirektna manifestacija efikasnosti DBB algoritma za balansiranja uzoraka je predstavljena u numeričkoj formi u Tabeli 5.4. Date distributivne karakteristike date u numeričkoj formi su suština indirektne prezentacje efikasnosti pomenutog algoritma. Objasnimo najpre sadržaj desnog dela Tabele 5.4. Kolone 6 i 7 predstavljaju oznake (D) i broj instanci (l) u aktuelnim nizovima respektivno. Kolone 8 do 11 predstavljaju statističke karakteristike odgovarajućih nizova (min, max, μ, σ) dok kolone 12 i 13 prikazuju vrednosti odgovarajućih entropija (H) i vrednosti relativnih entropija preko Kulbak-Leibler Divergencije (D_{KL}) respektivno. Treba istaći da su aktuelne vrednosti entropije izračunate na osnovu normalizovanih histograma odgovarajućih nizova D, tako da je $\sum_i f_i b_i = 1$, gde je i = 100 presdstavlja broj stratum (binova) histograma, f_i je normalizovana frekvencija i-tog stratuma, a $b_i = 0.15/100 = 0.0015$ je širina i-tog bina. Prvih šest vrsta u Tabeli 5.4 (1-6) sadrže vrednosti koje se odnose na originalne eksperimentalne uzorke C_{ei} i odgovarajuće nizove D_{ei}, vrste 7-12 sadrže vrednosti idealnih rešetkastih uzoraka C_{Li} i nizova D_{Li}, dok vrste 13-18 sadrže vrednosti koje se odnose na balansirane uzorke C_{bi} and D_{bi}. Potsetimo da smo usvojili uzorak u obliku pravilne rešetke C_{iL} kao idealnu komparativnu paradigmu uniformnosti, reprezentativnosti i balansa, tako da se ocena pomenutih karakteristika za uzorke C_{ie} i C_{ib} praktično svodi na direktnu komparaciju ovih uzoraka sa odgovarajućim rešetkastim uzorcima C_{iL} . Podsetimo se da su nizovi $D_{ei}, D_{Li}, D_{bi}, i = 1,2,..,6.$ Sa dužinama $l_{ei}, l_{Li}, l_{bi}, l_{ei} = l_{Li} = l_{bi} = 3100/i, i = 1,2, \dots, 6.$ indirektni pokazatelji

distributivnih karakteristika datih uzoraka $C_{ei}, C_{Li}, C_{bi}, i = 1,2 \dots ,6.$, kao njihovi derivati. Dakle, kao idealnu indirektnu komparativnu meru (paradigm) uniformnosti, reprezentativnosti i balansa za sve uzorke uzimamo distributivne karakteristike nizova D_{Li}. Ovi nizovi su u isto vreme indirktni indikatori efikasnosti DBB algoritma pri balansiranju uzoraka. Komparacijom vrednosti i karakteristika nizova D_{ei} i D_{bi} sa nizom D_{Li} dobijao posredno nivo reprezentativnosti nizova D_{ei} i D_{bi}. Podsetimo se glavne hipoteze ovog poglavlja koja predviđa da će balansirani uzorak C_{bi} dobijen pomoću DBB algoritma pokazati znatno veći stepen reprezentativnosti u odnosu na originalni neizbalansirani uzorak C_{ei}. Na nivou srednjih lokalnih rastojanja ovo znači da je mera podudarnosti nizova D_{bi} sa nizvima D_{Li} je znatno visa nego stepen podudarnosti između nizova D_{ei} i D_{Li}. Analitirajmo najpre karakteristike idealnih nizova D_{Li}. Iz Tabele 5.4 vidimo da je: min $min(D_{Li}) = max(D_{Li}) = \mu(D_{Li}), i = 1,2, \dots ,6.$ Ovae jednakosti ukazuju na to da se stvarni nizovi se sastoje od skupova identičnih članova i u skladu su sa degenerativnom ili determinističkom distribucijom, koja se karakteriše nultom standardnom devijacijom $\sigma_{D_{Li}} = \sigma(D_{Li}) = 0, i = 1, \dots ,6.$, i nultom entropija $H_{D_{Li}} = H(D_{Li}) = 0, i = 1,2, \dots ,6.$ Treba napomenuti da maksimalne vrednosti entalpije uzoraka u našem 2D primeru $H_{C_{Li}} = 8.0444, i = 1, \dots ,6$., odgovaraju nultoj vrednosti entropije odgovarajućih lokalnih srednjih rastojanja instanci od njihovih suseda $H_{D_{Li}} = 0, i = 1, \dots ,6.$ Na osnovu vrednosti entropije $H_{C_{ei}}, H_{D_{ei}}, H_{C_{bi}}, H_{D_{bi}}$ prikazanih u kolonama 4 i 12 Tabele 5.4, zaključujemo da je stepen neuređenosti unutar uzorka stoji u obrnutoj srazmeri sa stepenom neuređenosti koji odgovara nizu lokalnih srednjih rastojanja. Upoređujući vrednosti entropije i statističke karakteristike ostalih nizova sa odgovarajućim nizovima D_{Li}, primećujemo da je sličnost između sekvenci i D_{bi} i D_{Li} znatno veća u odnosu na sličnosti sekvenci D_{ei} i D_{Li}. Najvažniji pokazatelji sličnosti su vrednosti Kulbacki-Leibler divergencije (D_{KL}), što je veća vrednost D_{KL} to je manja sličnost poređenih uzoraka. Kolona 13 iz Tabele 5.4 jasno pokazuju sledeći odnos između podudarnosti: $D_{KL}(D_{Li}||D_{ei}) : D_{KL}(D_{Li}||D_{bi}) \approx 2:1$, koji indirektno potvrđuje našu osnovnu pretpostavku značajnog povećanja reprezentativnost balansiranih uzoraka C_{bi} u odnosu na originalne neuravnotežene uzorke C_{ei}. Ovaj dokaz je indirektna manifestacija potencijala DBB resampling algoritma da uspostavi interni balans prozvoljnih neizbalansiranih uzoraka menjajući njegovu strukturu u smeru strukture izbalansirane pravilne rešetke, kao idealne paradigme ravnomernosti, reprezentativnosti i ravnoteže. Ovaj dokaz je bio glavni cilj stava koji je u osnovi kompletnog istraživanja prikazanog u ovom poglavlju.

6. KLASIFIKATORI I KLASIFIKACIJA

Klasifikacija je generalno proces grupisanja entiteta u različite kategorije prema diskriminatornim atributima. Sa aspekta matematike, klasifikacija je proces preslikavanja vektora $\boldsymbol{X}$ iz domena atributa, u diskretni domen klasa $\boldsymbol{C}$. U kontekstu ovog istraživanja, klasifikacija je finalna procedura za kategorizaciju odnosno evaluaciju kvaliteta artikulacije odabranih fonema srpskog jezika, reprezentovanih adekvatno odabranim vektorima obeležja. U svrhu pronalaženja pouzdanog modela klasifikacije u predstavljenom istraživanju su korišćene četiri različite vrste klasifikatora koji su u potrebnoj meri predstavljeni u ovom poglavlju, dok su rezultati klasifikacije predstavljeni u osmom poglavlju o eksperimentalnim . U tu svrhu su takođe primenjene postojeće i razvijene nove metode preprocesinga u cilju povećanja reprezentativnosti obučavajućih uzoraka što je prikazano u poglavlju o učenju u uslovima neizbalansiranih podataka. Klasifikatori su kategorički induktivni učeći prediktori odnosno estimatori koji uspostavljaju fleksibilnu funkcionalnu korespondenciju između vektora obeležja konkretnih instanci i kategorija (klasa) kojima one pripadaju. U pitanju je nepotpuna indukcije koja na osnovu ograničenog reprezentativnog uzorka generiše zaključak o instancama cele populacije koji nije uvek istinit sud, za razliku od potpune indukcije koja je sa druge strane u praksi neprimenjiva zbog ograničenosti raspoloživih informacionih resursa. Algoritam podrazumeva određivanje skupa pravila, odnosno ograničenja ili parametara u opštem smislu, tokom procesa obuke koji svakom elementu skupa od n vektor instanci $\boldsymbol{x_i}$ reprezentativnog trening sempla X $(\boldsymbol{x_i} \in X, i = 1,2, \dots, n)$ pridružuje odgovarajući elemenat c_j skupa $\boldsymbol{C}$ od m klasa $(c_j \in, j = 1,2 \dots, m)$. Adaptivni modeli za koje su putem obuke determinisani parametri, pokazuju sposobnost generalizacije procesa klasifikacije izvan domena trening uzorka na kompletan domen takozvane ciljne ili target populacije koju reprezentuje ograničeni obučavajući skup. To znači da dobar klasifikator može sa visokim stepenom pouzdanosti ekstrapolirati znanje izvan granica domena trening uzorka pri klasifikaciji proizvoljne instance iz target domena. Ovakav klasifikator sa još većom pouzdanošću može interpolirati znanje pri klasifikaciji instanci koje su unutar granica domena trening uzorka, pošto se pri ekstrapolaciji klasifikator izlaže većoj neizvesnosti. Stepen pouzdanosti odabranog klasifikatora je determninisan stepenom reprezentativnosti trening skupa pa je jedan od glavnih preduslova određivanja pouzdanog klasifikatora dobar izbor reprezentativnog trening uzorka iz raspoloživog skupa instanci. Instance ili primerci su definisani vektorom diskriminatornih obeležja i u suštini, sa aspekta klasifikatora, predstavljaju tačku u Euklidskom prostoru obeležja pa je reprezentativnost posmatranog trening uzorka uslovljena prirodom raspodele

ovih tačaka u prostoru obeležja. Treba pomenuti da vektori obeležja mogu da se jave u više formi: binarni, kategorički, ordinalni, celobrojni, realni i kombinovani.

Klasifikacija je ključni zadatak za ekstrakciju znanja iz baza podataka (Knowledge Data Discovering) i manipulaciju podacima (Data Mining). Tokom obuke i konstrukcije modela klasifikacije, algoritam učenja detektuje relacije između skupa atributa i indikatora klasa, i definiše model koji najbolje odgovara obučavajućem uzorku podataka i najbolje opisuje modelirani proces. Zadatak ovako dobijenog klasifikatora je predviđanje pripadnosti konkretnoj klasi za bilo koje nepoznate instance iz target populacije. Dakle, cilj obuke je puzdan klasifikacioni model sa dobrom generalizacijom tj. model sa tačnom i pouzdanom kategorizacijom slučajeva nepoznate pripadnosti. Imajući za cilj najbolju generalizaciju, prediktivni model treba pravilno prilagoditi obučavajućem uzorku podataka. Kao rezultat procesa obuke mogući su sledeći ishodi kvaliteta klasifikatora: a) Loša identifikaciona faza modela koja se manifestuje lošom klasifikacijom obučavajućeg skupa što najčešće implicira lošu klasifikaciju test uzorka iz target populacije, odnosno lošu generalizaciju modela. U ovom slučaju javljaju se velike vrednosti greške pri klasifikaciji kako trening tako i test uzorka, pa se model smatra nepouzdanim za praktičnu upotrebu; b) Dobra identifikaciona faza modela koja se manifestuje dobrom klasifikacijom obučavajućeg skupa i dobrom generalizacijom potvrđenom kroz dobru klasifikaciju test uzorka. U ovom slučaju vrednosti greške klasifikatora na celom obučavajućem uzorku su male, pa se model smatra pouzdanim za klasifikaciju nepoznatih instance odnosnu za praktičnu upotrebu; c) Prividno dobra identifikaciona faza modela se manifestuje odličnom klasifikacijom obučavajućeg skupa koja rezultuje lošom generalizacijom, odnosno lošom klasifikacijom test populacije. U ovom slučaju vrednosti greške klasifikatora na obučavajućem uzorku su male dok su vrednosti greške pri klasifikaciji test uzorka velike, pa se model smatra nepouzdanim za klasifikaciju nepoznatih instanci odnosnu za praktičnu upotrebu. Ova pojava je poznata kao *overfiting* modela što znači pretreniranost ili nekontrolisana specijalizacija za prepoznavanje trening instanci. Za klasifikator c se kaže da overfituje obučavajući uzortak ako postoji neki alternativni klasifikator c' iz iste kategorije, takav da c bolje klasifikuje instance obučavajućek skupa od c', ali c' bolje klasifikuje ukupne podatke iz ciljne populacije od klasifikatora c (Mitchell, 1997) Jedan od poznatih uzroka overfitinga manjinske klase je prekomerne prosta replikacija postojećih primera manjinske klase, koja se uvodi iz razloga balansiranja klasa u trening uzorku, što izaziva suprotan efekat od željenog, odnosno redukciju prostora u kom se mogu pojaviti instance manjinske klase t.j. favorizaciju pojavljivanja instanci većinske klase. Ovaj fenomen je važan kod mašinskog učenja pa su uvedene nove tehnike u cilju prevencije

ovakvog iscrpnog učenje na obučavajućem uzorku. Pomenute tehnike sa najčešće zasnivaju na traženju inherentnih regularnosti u raspoloživim podacima za poboljšanje performansi generalizacije. Jedna od njih je princip minimalne dužine opisa (MDL) (Rissanen, 1978). koji se zasniva na sledećem stavu: S obzirom na ograničeni skup raspoloživih podataka (najčešći slučaj u relnim uslovima) najbolje objašnjenje daju modeli koji dopuštaju najveću kompresiju odnosno redukciju dimenzija podataka obučavajućeg uzorka. To jest, što smo više sposobni za kompresiju podataka, to znači da dobjamo više informacija o osnovnim pravilnostima pokretačkog mehanizma koji generiše te podatke. Ovakav proces neizbežno kod modela razvija sklonost ka maksimalnoj opštosti odnosno favorizuje otkrivanje najopštijih pravila (odn., Većih disjunkta) (Holte, 1989; Ting, 1994). Međutim, takva induktivna sklonost maksimalnoj opštosti predstavljala ozbiljan nedostatak pri klasifikacijom neuravnoteženih podataka. Inače je problem neizbalansiranih podataka i korespodentni problem neizbalansiranog učenja prepoznat kao jedan od najozbiljnijih problema u radu sa induktivnim prediktorima. Kao jednom od najaktuelnijih problema mašinskog učenja neizbalansiranom učenju je u ovom radu posvećena posebna pažnja i detaljno je objašnjen u šestom poglavlju.

6.1. Stabilni i nestabilni prediktori

Razmatrali smo način kako mere greške igraju ulogu u problemu nadgledanog učenja i ocenjujući prediktore sa kojima radimo. Postoje dve kategorije prediktora (klasifikatora) u smislu njihove stabilnosti i to: nestabilni i stabilni. Nestabilan prediktor je onaj koji ima stohastičku prirodu odnosno naglašenu zavisnost od slučajno izabranog obučavajućeg skupa, zato hipoteza koju on formira tokom testa zavisi u velikoj meri od izabranog domena kom pripada trening skup. Primeri nestabilnih prediktora su stabla odlučivanja (Duda i sar., 2001) i veštačke neuronske mreže. Stabilan prediktor je onaj koji nema takvih jakih zavisnost od podataka o obuci; Primeri stabilnih prediktora su klasifikatori zasnovani na k-najbližih suseda i Fischerov linearni diskriminator (Duda i sar., 2001). Nestabilni klasifikatori su poznati po karakteristično velikoj varijansi signala greške pri generalizaciji, dok stabilni klasifikatori imaju nisku varijansu (Breiman, 1996). To je osobina koja ukazuje na smisao kreacije ansambla klasifikatora kao jedne od aktuelnih tema u tezi. Sa ovog aspekta potpuno je shvatljiva ideja za formiranje ansambla ovakvih klasifikatora u cilju uravnoteženja kompromisnog donošenja odluke prediktora. Varijansa je samo jedna komponenta dobro poznate dekompozicije greške generalizacije na bias-varijansa komponente.

6.2. Bias i variansa

Jedan od najvažnijih teoretskih alata u istraživanju mašinskog učenja je dekompozicija Bias – Variansa (Geman i sar.,1992). Prvobitna dekompozicija od strane Geman i sar., (1992) odnosi se na gubitke prikazane u formi kvadratne greške, i navodi da se greška generalizacije može podeliti u dve zasebne komponente koje se mogu posebno tumačiti, bias i variansa. Ove dve vrednosti obično su u koliziji jedna s drugom: pokušaj smanjivanja bias komponente uzrokuje povećanje varijanse i obrnuto. Postojeće tehnike u literaturi mašinskog učenja često se ocenjuju po kriterijumu stepena optimizacije odnosa između ove dve komponente (Wahba i sar., 1999, Valentini i sar., 2002). Bias se može okarakterisati kao mera razlike predviđenih i target vrednosti, na osnovu proseka odziva prediktora obučenih sa nekoliko različitih trening uzoraka. Varijansa je merilo stabilnosti rešenja. Blaga razlika podatkaka u obuci kod estimatora sa visokom varijansom će imati tendenciju da proizvede ogromne razlike u performansama. Dugortajna obuka prediktora ima tendenciju da smanjuje bias, ali postepeno povećava varijansu tako da se u nekom trenutku ostvaruje optimalni odnos baias-variansa koji minimizira grešku generalizacije. To je takozvana baias-variansa dilema.

Bootstrap agregacija, poznat u skraćenoj formi kao ***bagging*** ili perturb and combine (P&C), Breiman (1998), predstavlja meta-algoritam za obuku prediktora, namenjenih poboljšanju stabilnosti i tačnosti algoritama za mašinsko učenja koji se koriste u statističkoj klasifikaciji i regresiji. Ovaj algoritam je poznat. Takođe smanjuje varijansu signala greške i pomaže u prevenciji overfitinga, odnosno, sklonosti ka specijalizaciji za bolje prepoznavanje instanci trening uzorka. Iako se izvorno primenjuje na stabla odlučivanja, može se bez razlike koristiti sa bilo kojim tipom prediktora. Bagging je jedan od modela nastalih na principu usrednjavanja performansi ansambla modela obučenih na različitim podskupovima trening uzorka. Skorašnji rezultati u mašinskom učenju pokazuju da performanse finalnog modela treba poboljšavati ne kroz biranje strukture najboljeg očekivanog prediktivnog modela, već stvaranjem modela na osnovu kompozicije rezultata (odziva) modela koji imaju različite strukture. Razlog je u tome što, u stvari, svaka hipoteza je samo estimacija stvarne ciljne vrijednosti i, kao i svaka ocena, i ona je pod uticajem biasa i varijance. Teorijski rezultati (Breiman, 1996) pokazuju da se smanjenje varijanse može dobiti jednostavnim kombinovanjem nekorelisanih hipoteza o stvarnim target vrednostima. Ova jednostavna ideja je osnova jedne od najefikasnijih skorašnjih tehnika u mašinskom učenju. Bagging dovodi do poboljšanja stabilnosti modela klasifikacije (Breiman, 1996), kao što su, na primer, veštačku neuronske mreže, klasifikatori u

formi stabla odluučivanja, kao i izbor podskupa kod linearne regresije (Breiman, 1994). Sledeći tekst sadrži prikaz neophodnih informacija o vrstama klasifikatora koji su korišćeni u ovom istraživanju.

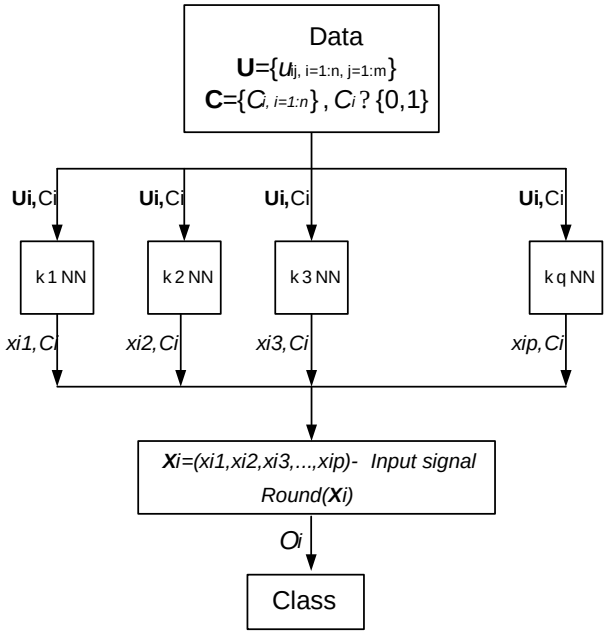
6.3. Klasifikator zasnovan na proceni preko najbližih suseda - kNN

Kombinovanje više klasifikatora iste kategorije, poznato kao ansambli klasifikatora pokazale su se kao pouzdane metode za poboljšanje performansi klasifikacije nepouzdanih individualnih klasifikatora i u zadnjih par decenija i time izazvale stekao veliki interes u oblasti (Bauer i Kohavi 1999; Melville et al. 2004; Barandela i sar. 2013). Ove metode vode bitno smanjuju greške pri klasifikaciji u mnogim aplikacijama u realnoj upotrebi i, što je bitno, otpornije su na neinformativne prediktorske variable (obeležja) nego pojedinčni klasifikatori (Melville i sar., 2004; Khoshgoftaar i sar., 2011). Jedna od najjednostavijih i najstarijih metoda za klasifikaciju je k najbližih suseda (kNN) klasifikator. Ona klasifikuje nepoznate realizacije (instance) u klasu kojoj pripada većina od k njenih njima najbližih suseda, definisanih merom rastojanja raspoloživog trening skupa instanci iz kog se uzimaju susedne instance, (Cover i Hart 1967, Guvenir i Akkus 1997). Uprkos imanentnoj jednostavnosti, kNN metoda daje rezultate uporedive sa mnogo složenijim metodama a u nekim slučajevima čak i nadmašuje neke složene algoritme klasifikacije. Međutim, kNN često pogađaju ne-informativne variable u podacima, slučaj sa visokim dimenzionalnim podacima. U literaturi postoji više primera poboljšanja performanse k NN klasifikatora primenom ansambl tehnika. Evo nekih primera iz ove kategorije: Grabovski (2002), Domeniconi i Ian (2004), Zhou i Yu (2005), Hall i Samworth (2005) i Samworth (2012).

Kombinovanje više klasifikatora iste kategorije, poznato kao metoda ansambla, može dati značajne poboljšanje performansi predikcije u odnosu na pojedinačne klasifikatore iz ansambla (Gul i sar. 2015). Ansambli generalno povećavaju stabilnost i robustnost u odnosu na pojedinačne klasifikatore.

KNN ansambl kao drugi klasifikator primijenjen u ovom istraživanju, sastoji se od sledećih pet pojedinačnih k NN klasifikatora: 1NN, 3NN, 5NN, 7NN i 9NN klasifikatora. Klasifikatori k NN-a jednostavno klasifikuju test uzorke koristeći Euclidsko rastojanje određenog broja najbližih susjeda iz uzorka obuke, koji praktično služi kao referentni skup primera. Klasifikator k NN ansambl takođe formuliše svoj odgovor na osnovu kriterijuma većinskog glasanja, zato smo koristili neparan broj (5) klasifikatora u ansamblu. Treba napomenuti da su k NN klasifikatori determinističke strukture koje uvek daju nedvosmislena

jednoznačna rešenja za isti skup uzoraka. Jedini stohastički uticaj na njihov odgovor ima slučajna selekcija trening i test uzoraka ali ovaj uticaj je mali. Takođe treba reći da k NN klasifikatori imaju brži odgovor u odnosu na MLP strukture. Budući da DBB algoritam omogućava značajno povećanje reprezentativnosti podataka, njegova primena treba da poboljša performanse klasifikatora koji se primenjuju. Kao praktičan dokaz ove tvrdnje, trebao bi služiti apriori pretpostavljena korelacija između mera reprezentativnosti uzoraka i mera performanse klasifikatora obučenih na ovim uzorcima. Preciznije, pretpostavlja se da je visina mere performanse klasifikatora direktno korelisana sa reprezentativnošću (balansom) uzoraka za obuku. Stoga, kao indirektna mera efikasnosti balansnih algoritama, uzimamo vrednost merenja performansi svih naših klasifikatora.

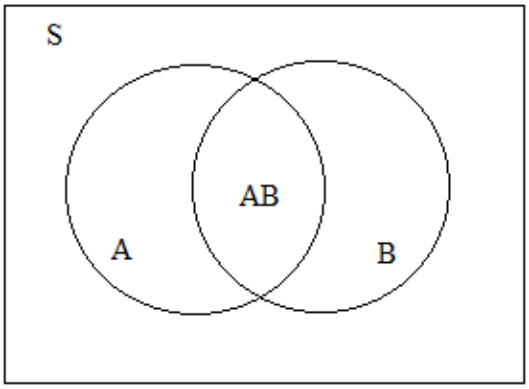


Slika 6.1 Ansambl k Najbližih Suseda kao klasifikator.

6.4. Jednostavni Bajesov klasifikator

Kao i svi drugi klasifikatori Jednostavni Bajesov klasifikator podrazumeva postojanje trening skupa *S* instanci definisanih u formi vektora obeležja $\boldsymbol{x} = (x_1, .., x_j, .., x_n)$ dužine *n* i njima korespondentnih (pridruženih) kategoričkih ili numeričkih indikatora klasa c_i kao elemenata skupa $\boldsymbol{c} = \{c_1, .., c_i, .., c_L\}$ kardinalnosti $|\boldsymbol{c}| = L$. U fazi obuke formiraju se parametri modela koji kasnije svakoj test instanci $\boldsymbol{x}$ pridružuju ocenjene vrednosti verovatnoće pripadnosti datim klasama $\hat{P}(c_i|\boldsymbol{x})$, odnosno $\left(\hat{P}(\boldsymbol{x} \in c_i), i = 1, .., L\right)$ i na osnovu maksimalne vrednosti verovatnoće vrši se konačna klasifikacija test instanci. Ovaj klasifikator je u stvari

primenjena Bayesova teorema (Thomas Bayes, 1701 - 1761), koja se neposredno zasniva na računanju uslovnih verovatnoća. Posmatrajmo moguće događaje *A* i *B* koji stoje u izvesnoj probabilističkoj vezi Slika 6.2, i definišimo osnovne probabilističke kategorije relevantne za definiciju ovog klasifikatora.



Slika 6.2. Presek događaja *A* i *B*.

Prethodna (Aprior) verovatnoća: *P(A), Uslovna verovatnoća:* Verovatnoća ostvarenja događaja *A* pod uslovom da je ostvaren događaj *B*: *P(A|B), P(B|A), Zajednička verovatnoća dva događaja: C=(A,B), P(C)=P(A,B).*

Presek dva slučajna događaja *A* i *B* je određen ishodima zajedničkim za oba događaja i označava se sa *A∩B* odnosno *AB*.

Verovatnoća preseka dva događaja: Definiše se u funkciji uslovne verovatnoće:

$$P(AB) = P(A)P(B|A) \text{ ili } P(AB) = P(B)P(A|B). \quad (31)$$

Iz prethodnih stavova sledi: $P(B)P(A|B) = P(A)P(B|A)$, odnosno sledi Bajesova teorema za dva događaja:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (32)$$

Za konačni broj disjunktnih slučajeva A_i, $i+1,\ldots,N$ Bajesova teorema izgleda ovako:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}. \quad (33)$$

U slučaju *nezavisnih događaja A* i *B* važi:

$P(A|B) = P(A)$ i $P(B|A) = P(B)$, odnosno $P(AB) = P(A)P(B)$.

Gornje relacije su važne za Jednostavni Bajes (Naive Bayes) Klasifikator. U sledećem tekstu ćemo prikazati algoritam obuke klasifikatora na obučavajućem skupu parova $(\boldsymbol{x}, c)$ u situacijama kada su vektori obeležja instanci date a) u formi diskretnih ili kategoričkih vrednosti i b) u formi kontinualnih numeričkih vrednosti. U tom smislu ćemo putem analogije gornju formulu prilagoditi aktuelnim veličinama $\boldsymbol{x}$ i c.

$$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c_i)P(c_i)}{P(\boldsymbol{x})}. \quad (34)$$

U ovoj jednačini $P(c_i|\boldsymbol{x})$ predstavlja parametar aposteriorne verovatnoće dobijen na trening uzorku, i služi za ocenu verovatnoće da neka test instanca $\boldsymbol{x}$, pripada klasi c_i. $P(c_i)$ je parametar prethodne verovatnoće klase c_i u odnosu na ostale klase dobijen na trening uzorku. $P(\boldsymbol{x})$ je evidentna prediktorska prethodna verovatnoća. $P(\boldsymbol{x}|c_i)$ je *likelihood* parametar odnosno verovatnoća pojave prediktora $\boldsymbol{x}$ pod uslovom realizacije date klase c_i. Prethodna prediktorska verovatnoća $P(\boldsymbol{x})$ ima konstantnu vrednost i ne utiče na vrednosti aposteriornih verovatnoća pa se ona zanemaruje pa jednačina za proračun aposteriorne verovatnoće test instance dobija sledeću konačnu formu:

$$P(c_i|\boldsymbol{x}) \propto P(\boldsymbol{x}|c_i) \times P(c_i) = P(x_1|c_i) \times P(x_2|c_i) \times \dots \times P(x_n|c_i) \times P(c_i) \quad (35)$$

Konačna odluka o klasi $c_i, i + 1,2, \dots, L$ se donosi u skladu sa maksimumom od svih L vrednosti verovatnioća. $\underset{i}{argmax}(P(c_i|\boldsymbol{x}))$.

Gornji algoritaam se može direktno koristiti u slučaju diskretnih vrednosti obeležja odnosno komponenti vetora $\boldsymbol{x}$. Kada vrednosti obeležja imaju kontinualnu prirodu tada postoje dva pristupa za proračun parametara klasifikatora.

Prvi pristup je diskretizacija kontinualne variable (binning, eng.) na pogodan broj segmenata ili stratusa. Tada jednostavno nastavljamo obuku klasifikatora u skladu sa prethodnom procedurom. Kod drugog pristupa uzorku sa obeležjima datim u formi kontinualnih numeričkih vrednosti, estimirane aposteriorne uslovne verovatnoće načešće se modeliraju u formi normalne raspodele:

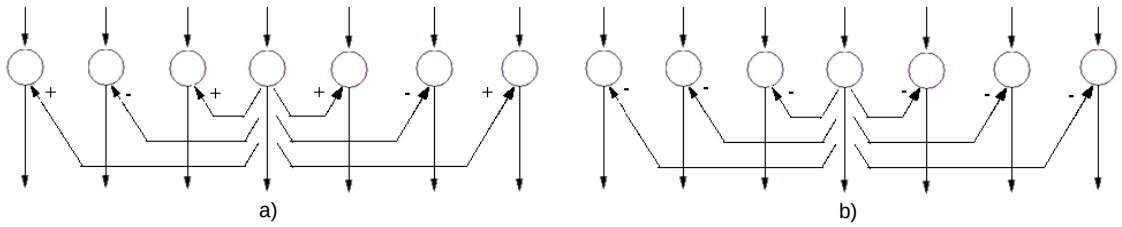
$$\hat{P}(x_j|c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right). \quad (36)$$

Gde je μ_{ji} parametar dobijen na trening skupu i predstavlja srednju vrednost onih vrednosti j-te kolone obeležja x_j za koje je klasa $\boldsymbol{c} = c_i$. Standardna devijacija σ_{ji} se takođe odnosi na iste vrednosti x_j za koje je $\boldsymbol{c} = c_i$. I u ovom slučaju konačna odluka o klasi $c_i, i + 1,2, \dots, L$ kojoj pripada vektor $\boldsymbol{x}$ se donosi u skladu sa maksimumom od svih L vrednosti verovatnioća $\underset{i}{argmax}(\hat{P}(c_i|\boldsymbol{x}))$.

6.5. Samoorganizujuće Mape (SOM) kao klasifikator

Kod modela koji se obučavaju pod nadzorom na osnovu primera kao što su Perceptron, Adalina i slični ne koriste se osnovna adaptivna svojstva i funkcije neurona Hebb, 1949, jer

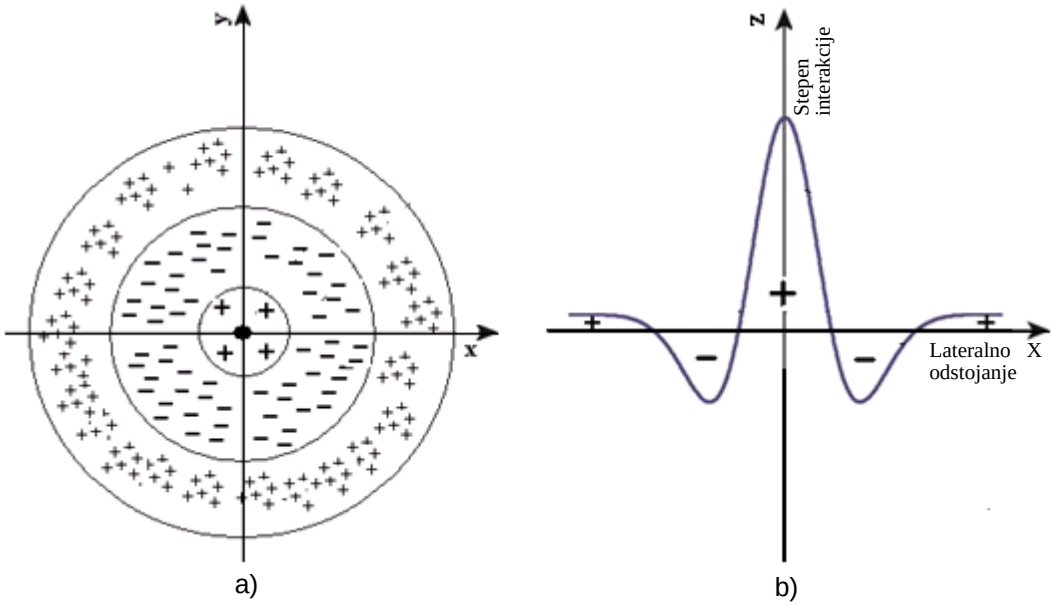
uvek uključuju komparaciju sopstvenog i željenog odziova na svim jedinicama i na svim nivoima. Obučavanje uz pomoć učitelja karakteristično je za modifikaciju generalnog ponašanja učećeg sistema na najvišem nivou, ali to ne znači da se i na nižim nivoima organizacije odvija sličan proces bar kada je ljudski mozak u pitanju. Ovaj bitan nedostatak fiziološke fundiranosti karakterističan za perceptrone s jedne strane i njihova praktična primenljivost sa druge, doveli su do zapostavljanja u istraživanju i razvoju modela sa univerzalnim svojstvima kakva ima mozak s jedne strane, i istovremeno do pojave ogromnog broja međusobno sličnih algoritama za obuku perceptrona i sličnih modela. Pomenuti modeli su dakle bihevioristički modeli naučenih relacija stimulus – odgovor na višem nivou organizacije. Najveći nedostatak ovih modela je to što se podrazumeva da svaka neuronska ćelija operiše uglavnom samostalno, iako paralelno sa drugim ćelijama jer se apstrahuje njihova lateralna interakcija. Na suprot dokazanim kolektivnim i kooperativnim osobinama biološkog pandana ovi modeli su zadržali samo jednu kooperativnu karakteristiku i to u samom neuronu koji na neki način vrši prostorno vremensku integraciju ulaznih signala. Očigledno da modeli neuronskih mreža ne mogu dosledno odslikavati originalni fenomen kolektivnog proesa većeg broja neurona ukoliko se na uzmu u obzir povratne grane formirane na mnogo raznih načina. Na osnovu znanja o morfološko funkcionalnoj organizaciji korteksa velikog mozga može se reći da je izuzetno važan tip organizacije struktura sa povratnim granama takozvani *lateralni povratni model* ili *laminarni model mreža* (Kohonen, 1984).



Slika 6.3 Ekcitatorno-inhibitorna a) i inhibitorna lateralna interakcija neurona unutar sloja b).

Ova struktura je predstavljena u jednostavnoj formi na Slici 6.4 a, dok su stepen i raspodela električnog polja u funkciji interakcije prikazani na Slici 6.4 b. Ova funkcija je poznata kao „mexican hat“. Model ima epitet laminarni po anlogiji sa organizacijom kore velikog mozga gde su neuroni organizovani u slojeve (lamine) debljine nekoliko milimetara. Lateralna povratna interkcija podrazumeva međusobni uticaj grupe neuronskih jedinica iste lamine (sloja) koja se manifestuje kao promena odziva aktuelnih jedinica odnosno kao ekscitacija (povećanje odgovora“+“) ili kao inhibicija (umanjenje odgovora“-“) u odnosu na primarni signal, tokom niza vremenski diskretnih koraka. Ova promena je posledica promene

ulaznih signala nastale pod dejstvom signala iz povratnih grana susednih jedinica tokom niza koraka.



Slika 6.4 Lateralna interakcija neurona u dve ravni (Mexican hat).

6.5.1. Računarska simplifikacija procesa samoorganizacije

Evidentno je da se fenomen *klasterovanja* može ostvariti primenom različitih formi lateralne povratne funkcije, tako da se funkcija iz prethodnog primera može smatrati samo jednom od njih.U osnovi svake topološke reprezentacije u kori mozga leži upravo pojava lokalizovanih *klastera aktivacije* piramidalnih ćelija (Charles i sar., 2005), pa je matematička formalizacija ove pojave osnova računarkog modeliranja samoorganizacije neurona. Obzirom na složenost originalne pojave neophodna je simplifikacija kompjuterskog algoritma u kojoj će se sačuvati funkcionalni principi originalnog procesa s jedne strane i omogućiti praktična efikasnost s druge strane.

6.5.2. Osnove algortam učenja – adaptacije SOM

Algoritam učenja Samo-Organizujućih Mapa (SOM) čine četiri glavna koraka:

1. Iz skupa svih mogućih ulaznih vektora slučajno se odabira jedan ulazni vektor $x = (\xi_1, \xi_2, \dots, \xi_n),\ x \in \Re^n$.

2. Koristeći furmulu funkcije aktivacije izračunava se stanje aktivacije svake pojedinačne izlazne procesorske jedinice *y* nastalo kao posledica slučajno odabranog ulaznog vektora *x.*

3. Procesorska jedinica čiji sinaptički težinski vektori pokažu najveći stepen podudarnosti sa aktuelnim vektorsikim ulazom *x*, u skladu sa već usvojenim kriterijumom podudarnosti, biva deklarisana kao pobednička jedinica odnosno kao pobednik sa indeksom *c*.

4. Vektor težinskih parametara pobedničke jedinice *c*, *mc* kao i sve procesorske jedinice iz njene neposredne okoline, se adaptivno transformišu na takav način da se povećava njihova podudarnost sa aktuelnim ulaznim vektorom *x*. Ovaj proces se iterativno ponavlja i postepeno konvergira ka preslikavanju topološkog rasporeda skupa ulaznih vektora na skup procesorskih jedinica neuronske strukture. Detaljnija analiza svakog od navedenih koraka će razjasniti ceo proces u funkciji diskretne vrmenske variable *t* kojom se određuje korak svake iteracije i omogućava algoritamsku prezentaciju celog procesa učenja koja sledi.

Neka je dat skup ulaznih signala $\Xi = x_k, 1 < k < p$ koji potiču iz proizvoljno kompleksnog ulaznog vektorskog prostora $\Re^n$. Pri svakom iterativnom koraku algoritma iz skupa Ξ, na slučajan način koji garantuje ravnomernu prezentaciju svih ulaznih signala $x_k \epsilon \Xi$, bira se jedan ulazni signal $x(t) = (\xi_1, \xi_2, \dots, \xi_n)$. Ne predstavljajući nikakvo ograničenje ulazni vektori se normalizuju na standardnu vrednost dužine 1, čime se ulazni prostor transformiše u n-dimenzioalni prostor poluprečnika dužine 1. Definišući dužinu vektora *x* u obliku Euklidske norme:

$$dužina\_x = \|x\| = \sqrt{\sum_{k=1}^{n} \xi_k^2} \tag{37}$$

ulazni vektori se mogu normalizovati na vrednost 1 tako što se njegove komponente ξ_k transformišu na sledeći način:

$$\xi_k' = \xi_k / \|x\| \tag{38}$$

U sledećem koraku se izračunava stanje aktivacije svake jedinice uzrokovano prispećem ulaznog vektora *x(t)* u skladu sa adekvatnom funkcijom aktivacije. Postoje dva tipa funkcije aktivacije koje se koriste za određivanje pobedničke procesorske jedinice pri uvođenju ulaznog vektora. Najčešće korišćena funkcija aktivacije procesorske jedinice indeksa *i* je definisana kao Euklidsko odstojanje između ulaznog vektora *x* i vektora sinaptičih težina aktuelne jedinice m_i. Sledeća jednačina opisuje proračun razlike vektora težina m_i aktuelne jedinice u odnosu na ulazni vektor *x(t)* u vremenskom trenutku *t*, odnosno izlazne vrednosti aktivacije $\eta_i(t)$ aktuelne procesorske jedinice u tom trenutku.

$$\eta_i(t) = \|m_i(t) - x_i(t)\| = \sqrt{\sum_{k=1}^{n} (\mu_{ik}(t) - \xi_k(t))^2}. \tag{39}$$

Druga funkcija aktivacije zasniva se na unutrašnjem ili skalarnom proizvodu ulaznog vektora *x* i vektora težinskih faktora m_i aktivne procesorske jedinice. U ovom slučaju se evaluacija stanja aktivacije date jedinice izračunava u skladu sa sledećom formulom:

$\eta_i(t) = \|x_i(t)^T m_i\ (t)\| = \sum_{k=1}^{n} \xi_k(t)\mu_{ik}(t).$ (40)

U sledećem koraku, jedinica čiji vektor težina m_i ima najveći stepen podudarnosti sa ulaznim vektorom *x*, proglašava se za pobedničku jedinicu. Ako se koristi aktivaciona funkcija u formi jednačine 9, tada se za pobedničku jedinicu bira ona sa sa najnižim vektorom aktivacije η_i odnosno jedinica sa najmanjom Euklidskom distancom između vektora m_i i vektora *x*. Kriterijum izbora pobedničke jedinice za koju se usvaja indeks *c* se može definisati na sledeći način:

$c(t)$: $\eta_c(t) = \min\big(\eta_i(t)\big) = \min(\ \|m_i(t) - x_i(t)\|).$ (41)

U slučaju kada je aktivaciona funkcija definisana jednačinom 9, očigledno će kriterijum izbora jedinice izgledati drugačije od prethodnog i to u skladu sa jednačinom

$c(t)$: $\eta_c(t) = \max\big(\eta_i(t)\big) = \max(x_i(t)^T m_i(t)).$ (42)

U poslednjem koraku algoritma, vektori težinskih faktora pobedničke jedinice kao i njenih susednih jedninica se podvrgavaju procesu adaptacije koja zavisi od tri faktora, i to od:

1) vremenski promenljivog parametra učenja $\alpha(t)$,

2) vremenski promenljive funkcije indeksa procesorskih jedinica odnosno funkcije topološke distance aktuelne pobedničke jedinice indeksa *c* i susednih jedinica $\phi_{ci}(t)$, koja uključuje lateralnu povratnu interakciju između okolnih jedinica (*i*) i pobedničke jedinice (*c*), i

3) razlike između vektora težinskih faktora pridruženih pobedničkoj jedinici $m_i(t)$ i ulaznog vektora $x(t)$ odnosno od vrednosti vektora izlazne aktivacije pobedničke jedinice $\eta_c(t)$.

Proces adaptacije vektora težina $m_i(t)$ rezultira novim vektorom težina $m_i(t+1)$ u skladu sa sledećom jednačinom:

$m_i(t+1) = m_i(t) + \alpha(t_k)\phi_{ci}(t)[x_i(t) - m_i(t)]$ (43)

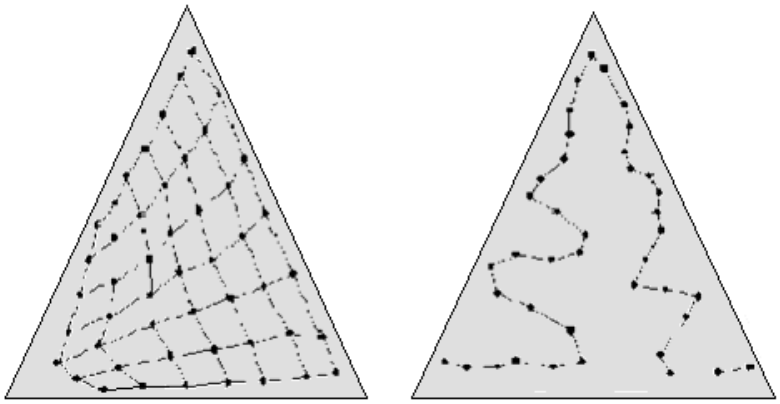
Detaljne informacije o SOM su date u priručniku Kohonen, 1984.

6.5.3. Primeri preslikavanja sa očuvanjem topologije uzorka

Primeri kompjuterske simulacije ilustruju kroz grafičku prezentaciju pojavu gde vektori sinaptičkih težina imaju tendenciju da aproksimiraju originalnu raspodelu odnosno topologiju uzorka ulaznih vektora. U primerima su odabrani dvodimenzionalni ulazni vektori, zbog lakše vizuelne prezentacije, a njihova raspodela gustine verovatnoća je ravnomerna na površini unutar graničanih linija.

Vektori $x(t_k)$ se tokom adaptacije slučajno biraju iz postojećeg skupa i determinišu promenu vrednosti sinaptičkih vektora m_i. Vektori m_i su prikazani kao pune crne tačke u istom

koordinatnom sistemu gde i vrlo gusto i ravnomerno distribuirani vektori $x(t_k)$ koji su prikazani kao siva površ na slici 6.5. U nameri da simbolično prikažu kojoj tački svaki vektor m_i pripada, krajnje tačke vektora m_i su povezane mrežom linija koje su podudarne sa topologijom strukture procesorskih jedinica. Dakle linija koja povezuje vektore m_i i m_j ima ulogu da prikaže da su dve korespondentne jedinice sa indeksima i i j susedi u topološkoj strukturi procesorskih jedinica. Na Slici 6.5 a je prikazan vrlo gust skup dvodimenzionalnih vektora uniformne raspodele na površini trougla (siva podloga) koji je reprezentovan topološkom strukturom redukovanog broja procesorskih jedinica (vektori sinaptičkih težina predstavljeni punim kružićima meosobno povezanih). Na ovoj slici svaka procesorska jedinica predstavlja klaster vektora koji pripadaju njenoj okolini i imaju najmanje rastojanje od nje u odnosu na druge jedinice. Na ovaj način se ostvaruje očuvanje topologije velikih originalnoh uzorka vektora proizvoljnih dimenzija posredstvom znatno manjeg broja reprezentativnih vektora redukovanih dimenzija (1D, 2D i 3D). Zato se ove strukture koriste za redukciju dimenzija podataka.



a) b)

Slika 6.5 Skup 2D vektora gusto i ravnomerno raspodeljenih po površima trougla (siva podloga) i skup procesorskih reprezentativnih jedinica (mreža crnih tačaka).

Slika 6. 5b prikazuje isti uzorak vektora koji je reprezentovan jednodimenzionalnom strukturom procesorskih jedinica. U oba slučaja je jasno uočljiva tendencija preslikavanja generalne prostorne topološke organizacije ulaznih vektora na redukovanu strukturu jedinica. Linearne arhitekture jedinica Slika 6. 5b imaju tendenciju da aproksimiraju dvodimenzionalnu raspodelu ulaznih vektora formirajući pritom takozvane Peanove Krive. Ovo ponašenje linearne strukture procesorskih jedinica u slučaju dvodimenzionalnih vektorskih ulaza se prenosi i na slučajeve kada ulazni vektori imaju tri ili mnogo više dimenzija a preslikavaju se u strukturu sa manjim brojem dimenzija. Karakteristična je analogija između ove pojave i kognitivne kategorije abstrakcije karakteristične za ljudski mozak. Bitno je uočiti da su obe vrste vektora predstavljene u istom koordinatnom sistemu pa se vidi da određeni prostor

vektora ulaza pokriva odgovarajući vektor sinaptičkih težina praktično reprezentujući određeni klaster vrednosti ulaznih vektora. Naime određena adaptivna jedinica je senzitivna na određenu grupu ulaznih vektora i to tako da postoji izvesna podudarnost vrednosti specifične grupa ulaznih vektora i težinskih faktora procesorske jedinice koja ih reprezentuje što rezultuje okisdanjem upravo te jedinice na pojavu ulaznih vektora samo iz njenog okruženja. Drugim rečima ulazni vektori sa određenim vrednostima komponenti preslikavaju se u jedinicu sa sličnim, vrednostima težinskih faktora. Takođe je uočljiva i podudarnost zakonomernosti rasporeda ulaznih vektora i korespodentnih procesorskih jedinica kao i podudarnost opštih prostornih formi skupa ulaznih vektora i skupa procesorskih jedinica. Ove karakteristike SOM struktura su razlog za njihovu upotrebu pri kategorizaciji vektora obeležja akustičkih signala koji odgovaraju različitim kvalitetima artikulacije fonema Srpskog jezika, odnosno za evaluaciju kvaliteta artikulacije istih.

6.6. MLP Ansambl kao klasifikator

Pored aproksimacije funkcija, jedna od najvažnijih primena neuronskih mreža je klasifikacija uzoraka. U ovom potpoglavlju su prikazane karekteristike MLP mreže u meri potrebnoj da se razume njena primena u klasifikaciji.

6.6.1. Veštačke neuronske mreže, opšte karakteristike

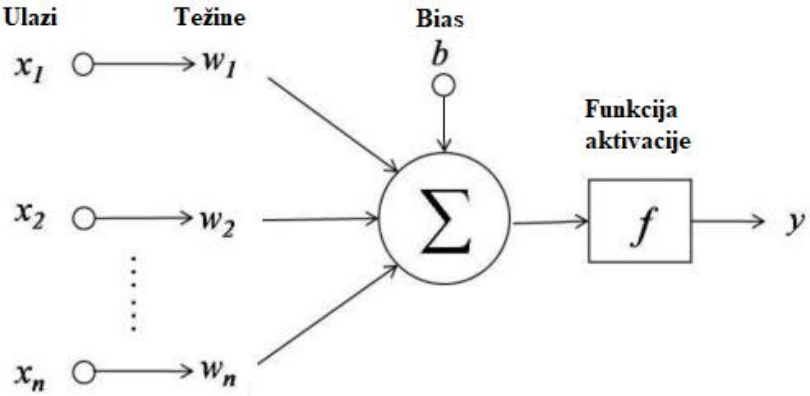
Veštačke neuronske mreže su računarske strukture za obradu informacija zasnovane na generalizovanim matematičkim modelima principa morfološko funkcionalne organizacije nervnog sistema. Najbitnije opšte karakteristike ovih struktura su sledeće:

1. Obrada informacija se odvija u jednoj ili više jednostavnih procesorskih elemanata nazvanih neuroni ili jedinice.
2. Signali se prenose među neuronima preko usmerenih veza.
3. Svaka veza ima pridruženu vrednost težinskih faktora, koji u tipičnoj neuronskoj mreži služi kao multiplikator prenošenog signala.
4. Svaki neuron ima sopstvenu aktivacionu funkciju koja preslikava sumirane ulazne signale u sopstveni izlazni signal.
5. Svaka neuronska mreža ima sopstvenu arhitekturu odnosno tačno definisan raspored i broj neurona i težinskih faktora.
6) Svaki tip neuronske mreže ima definisan odgovarajući algoritam obuke.

Procesorske jedinice sa usmerenim komunikacionim kanalima se nazivaju i čvorovima a u njima su locirane osnovne funkcije. Na Slici 6. 6 je prikazana struktura abstraktnog neurona sa *n* ulaza. Svaki ulazni kanal *i,i =1,2,...,n,* prenosi realne vrednosti signala *xi*.

Osnovna funkcija f, koja predstavlja funkciju transformacije pobudnog signala u odgovor abstraktnog neurona, može biti definisana posebno za svaki čvor ili zajedniška za sve čvorove što je najčešći slučaj. Tipična osnovna funkcija procesorskih elemenata mreže je sigmoidna funkcija *tangens hyperbolicus* prikazana je na Slici 6.8. Skup aktuelnih osnovnih funkcija je veliki i izbor zavisi od problema u kom se mreža primenjuje.

Ulazni kanali imaju pridruženu skalarnu realnu vrednost koja simbolizuje težinski faktor w_i, i skalira ulazni signal x_i. Ovako multiplicirani ulazni siganali se sumiraju u neuronu i zatim se nad ovim zbirom izračunava funkcija f . Ako svaki čvor u neuronskoj mreži shvatimo kao jedinicu sa osnovnom funkcijom koja svaki ulaz transformiše u precizno definisani izlaz, tada se veštačka neuronska mreža može smatrati mrežom ovakvih osnovnih funkcija. Razni modeli neuronskih mreža zasnivaju svoju originalnost na vrsti odabranih osnovnih funkcija, tipu interkonekcije među jedinicama, odnosno arhitekturi, i tajmingu prenosa informacija.

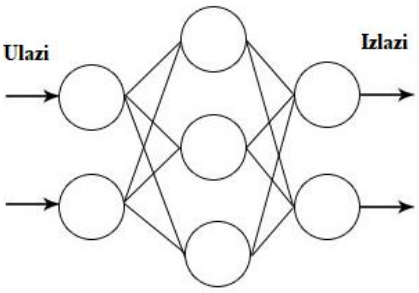


Slika 6.6 Shema abstraktnog neurona.

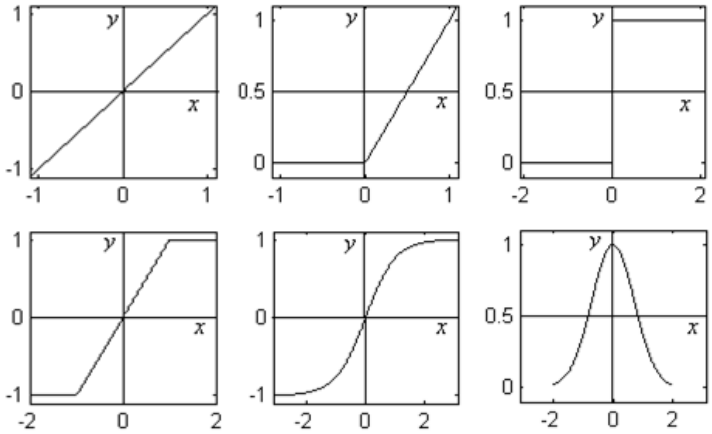
6.6.2. Višeslojni perceptron

Višeslojni perceptron (MLP) spada u neuronske mreže sa propagacijom signala unapred, čija se obuka odvija pod nadzorom, odnosno u prisustvu signala željenog odziva na predefinisani ulaz. Obučavajući skup je dat u formi konačnog skupa parova ulaz-izlaz. MLP koristi generalizovano delta pravilo učenja, odnosno pravilo povratne propagacije signala greške, (Rumelhart I sar., 1986). Perceptron je arhitektura organizovana u slojeve: ulazni, skriveni i izlazni sloj. Ako je sloj "skriven" tada nema direktne veze sa ulazom ili izlazom mreže, tako da mreža prikazana na Slici 6.7 ima jedan skriveni sloj koji sadrži tri čvora.

Arhitekture veza između čvorova mogu da variraju ali u ovoj tezi mi samo razmatramo potpuno povezane mreže odnosno MLP structure. Informacije u mreži se kreću s leva na desno na prikazanom dijagramu. Čvorovi u skrivenim i izlaznim slojevima izračunavaju njihovu vrednost aktivacije koja je ponderisana težinskim faktorima i transformisana preko aktivacione funkcije u izlaznu vrednost. Različite funkcije aktivacija su prikazane na Slici 6.8 od kojih su u tezi korišćene hiperbolički tangens za ulazne skrivene slojeve i funkcija identičnog preslikavanja za izlazne slojeve.



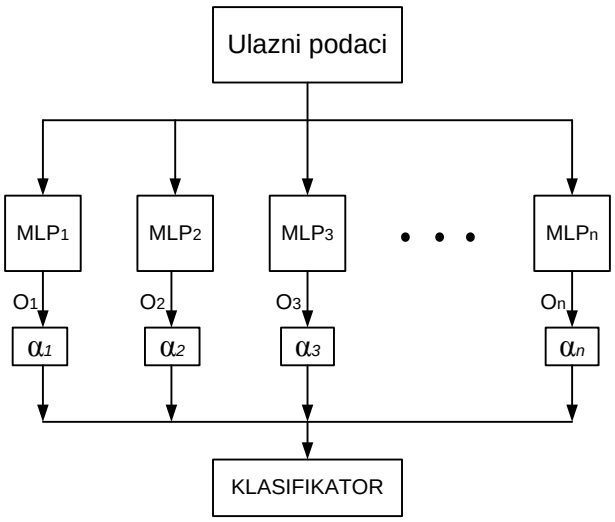
Slika 6.7 Višeslojni perceptron.



Slika 6.8 Različiti tipovi osnovnih funkcija aktivacije.

6.6.3. Ansambli MLP klasifikatora

Strukturna shema ansambla MLP kao klasifikatora data je na Slici 6.9 (Furundžić i sar., 2012a).



Slika 6.9 Ansambl MLP klasifikatora.

6.6.4. Algoritam optimalnog MLP Ansambla

Kombinovani usrednjeni odziv ansambla se dobija u skladu sa jednačinom 15. Težinski parametri a_i zadovoljavaju sledeća ograničenja

$$\sum_{i=1}^{n} a_i = 1, 0 \leq a_i \leq 1, \quad (44)$$

Izlaz ansambla je definisan na sledeći način:

$$\boldsymbol{O} = \sum_{i=1}^{N} a_i \, O_i. \quad (45)$$

U našem slučaju su faktori a_i iz jednačine 14 definisani na sledeći način:

$$a_i \leq 1/N, \quad (46)$$

Gde N predstavlja broj odabranih MLP klasifikatora. Dakle izlazna vrednost celog ansambla je jednaka srednjoj vrednosti izlaza svih eksperata ansambla. Funcija transformacije individualnih ulaza klasifikatora može biti linearna i nelinearna u zavisnosti od kompleksnosti modela.

6.6.5. Primenjeni algoritam

Pojednostavljena prezentacija algoritma za determnaciju optimalnog ansambla iz kompletne grupe obučenih induktivnih MLP klasifikatora sadrži sledeće korake:

1. Podeli ceo raspoloživi reprezentativni skup **S** u dva skupa, obučavajući S_L i test skup S_T jednakih kardinalnih vrednosti $|\mathbf{S_L}|=|\mathbf{S_T}|$ odnosno sa jednakim brojem instanci, vodeći računa o podjednakoj zastupljenosti svih klasa, što znači da sve klase imaju pouzdane reprezentativne uzorke.
2. Pošto u konkretnom slučaju, kao i generalno, negativna klasa je znatno bolje zastupljena od

pozitivne, koristili smo naš DBB algoritam za balansiranje broja instanci pozitivne i negativne klase. Negativna klasa ovde ima značenje negativnog nalaza u medicinskoj dijagnostici koji je prohvaćen kodproblema klasifikacije.

for i=1:I, (I=100)

3. Odabrati slučajno 80% primeraka iz od obučavajućeg skupa S_L za trening skup a 20% za validacioni skup za svaki *i*-ti MLP_i klasifikator, gde je *i*=1,2,3,…, I, dok I predstavlja predeterminisani ukupan broj klasifikatora koje treba obučiti.

4. Odredi strukturu *i*-tog induktivnog MLP_i klasifikatora koristeći metod postepenog povećanja kompleksnosti strukture.

5. Izvedi proces obuke MLP_i klasifikatora i sačuvaj prikupljeno znanje u formi vektora vrednosti težinskih faktora $\mathbf{W_i}$ i biasa $\mathbf{b_i}$. Na ovaj način svaki obučeni MLP_i klasifikator je predstavljen vektor modelom koji se jednostavno transformiše u odgovarajući obučeni klasifikator tokom testiranja ili ekspoloatacije. U vektor modelu svaka komponenta vektora predstavlja jednu vrednost težinskog faktora čiji je indeks definisan samom pozicijom komponente u vektoru.

6. Uzeti test skup S_T za izračunavanje estimiranih izlaznih vrednosti $\boldsymbol{o_i}$ aktuelnog MLP_i klasifikatora.

7. Odredi *i*-tu skalarnu komponentu e_i vektora performansi $\boldsymbol{E}$, dobijenu na izlazu *i*-tog MLP_i eksperta u skladu sa jednačinom 17.

$$e_i = mse(\boldsymbol{t} - \boldsymbol{o_i}), \quad (47)$$

gde je $i = 1,2,3, \dots I$, $\boldsymbol{t}$ predstavlja target vektor a $\boldsymbol{o_i}$ predstavlja izlazni vektor predviđen od strane *i*-tog eksperta. Član *mse* predstavlja srednju kvadratnu grešku.

end

8. Kompozicijom komponenti $e_i, i = 1,2,3, \dots, I$, dobijamo vektor performansi $\boldsymbol{E} = \{e_1, e_2, \dots, e_I\}$ svih *I* klasifikatora.

9. Sortirajmo vrednosti vektora $\boldsymbol{E}$ u rastućem nizu tako da njegov prvi član ima najmanju vrednost, odnosno $e_1 = min(\boldsymbol{E})$ a poslednji član maksimalnu, $e_{100} = max(\boldsymbol{E})$. Treba podsetiti da minimalna vrednost prvog člana predstavlja performansu najboljeg modela na koji se odnosi, drugi član se odnosi na sledeći najbolji model i.t.d.

10. Definiši broj $M, M \geq 5, M < I$, prvih najboljih obučenih modela. Broj M je uslovljen računarskim resursima ali treba da ima vrednost manju od polovine ukupnog broja obučenih uzoraka. U našem slučaju smo uzeli $M = 20$, što znači da smo odbacili 80 obučenih modela kao manje relevantne u smislu generalizacije zbog reletivno velikih odstupanja performansi u odnosu na očekivane.

11. Naći najpre odzive svih M odabranih modela $\boldsymbol{O}_j, j = 1,2,..,M$. Slučajno birati N obučenih modela iz odabranog skupa M, $N < M$ i naći iscrpne kombinacije odziva odabranih N modela pri svakoj iteraciji $\boldsymbol{O}_k, k = 1,2,..,N$. Iscrpan broj kombinacija od M modela N te klase je prikazan sledećom jednačinom:

$$\boldsymbol{C} = C_N^M = \binom{M}{N} = \frac{M!}{N!(M-N)!} \quad (48)$$

Na ovaj način dobili smo optimalni skup od $\boldsymbol{C}$ ansambla $\{A_i\}$, $A_i = \{MLP_1, MLP_2, .., MLP_N\}, i = 1,2,\ldots,\boldsymbol{C}$.

Pošto su odzivi svih pojedinačnih modela poznati, kombinovani usrednjeni odziv dobijenih ansambala će predstavljati klasifikaciju test uzorka pomoću aktuelnih ansambala, pa se naš optimizacioni postupak svodi na izbor ansambla sa najboljom generalizacijom na test uzorku, odnosno sa najmanjom greškom.

12. Za svaku od $\boldsymbol{C}$ kombinacija od N modela MLP treba izračunati objedinjeni izlaz definisan kao u Jed. (15 i 16) kako bi se izračunao vektor od $\boldsymbol{O}_c$ vrednosti, $c = 1,2,\ldots,\boldsymbol{C}$.

13. Izračunati objedinjene vrednosti performansi za sve ansamble na sledeći način: $E_c = mse(\boldsymbol{t} - \boldsymbol{O}_c), c = 1,2,\ldots,\boldsymbol{C}$. (49)

14. Na kraju, nalazimo najbolji ansambl u skladu sa sledećom jednačinom:

$$A = \underset{c}{argmin}(E_c) \quad (50)$$

15. Parametre ovako izabranog obučenog ansambla MLP memorisati kao konačni model optimalnog klasifikatora koji sadrži relevantno znanje.

6.5.6 Analiza Senzitivnosti Neuronske Mreže kao Indikatora Funkionalne Zavisnosti Odziva od Ulaznih Variabli.

Pojam senzitivnosti neuronske mreže (NM) se odnosi na metode analize uticaja ulaznih variabli i/ili perturbacija težinskih parametara na njene izlaze. Analiza osetljivosti počinje tokom šesdesetih godina 20-tog veka, sa radom gde Widrov istražuje verovatnoću netačne klasifikacije uzrokovanih perturbacijom sinaptičkih vrednosti , koji su bili uzrokovani nepreciznošću mašine i zašumljenim ulazom (Widrov i Hoff, 1960). U hardverskoj realizaciji NM, takvi poremećaji su morali biti biti analizirani pre svoje njenog dizajna, jer oni značajno utiču na obuku i generalizaciju mreže. Ova prva ideja o analizi senzitivnosti mreža proširena je kasnije na njihovu optimizaciju, kao što su problemi redukcije uzoraka, izbora optimalnog skupa obeležja i analiza prirode uticaja ulaznih variabli na izlazne vrednosti NM.

Praktična primena analize senzitivnosti se zasniva na merenju promena osmotrenih na izlaznoj variabli y_k, koje su posledica uticaja pogodno dizajniranih promena ulazne veličine x_i (Widrow and Hoff, 1960). U tom smislu važi: što je veći efekat osmotren na izlazu, to je veća osetljivost NM na relevantne ulazne promenljive. Jakobian matrica nastala iračunavanjem parcijalnih izvoda izlaza y_k po ulazima x_i odnosno $\partial y_k / \partial x_i$, predstavlja analitičku verziju ispitivanja osetljivosti. Veličina $S_{ik}= \partial y_k / \partial x_i$ predstavlja senzitivnost izlazne veličine y_k u funkciji ulazne veličine x_i i izračunavanje ovih vrednosti ima značaj za određivanje relevantnosti ulaznih veličina, jer predstavlja nagib funkcije relacije između aktuelnog para izlaza y_k i ulaza x_i. Vrednosti Jakobian matrice ne zavise samo od informacije naučene od strane NM, koja je sačuvana u distribuiranoj formi matrica težinskih faktora w_{ij}, već takođe zavise od aktivacija neurona skrivenog i izlaznog sloja koji dalje zavise od vrednosti ulaznih veličina. Analizom senzitivnosti mreže na perturbacije ulaznih variabli može se odrediti značajnost uticaja pojedinih variabli na odziv obučene NM. Analiza senzitivnosti se moze izvesti posmatranjem efekta determinisane funkcije greške izazivanjem promena ili poremećeja na ulaznim veličinama. Uobičajen način za ovaj metod je takozvana „*clamping tehnika*“ (Wang i sar., 2000), koja se zasniva na komparaciji greške nastale unošenjem originalnih vrednosti ulaznih uzoraka i greške nastale restrikcijom na fiksiranu vrednost (obično srednju vrednost) jedne analizirane ulazne veličine po svim realizacijama. Po ovom kriterijumu veća ukupna greška po svim realitzacijama na izlazu NM po fiksiranoj variabli pokazuje veći značaj same variable u konstelaciji izabranih variabli. Treba istaći da ista variabla u konstelaciji sa drugim skupom variabli može imati manji ili vaći uticaj na izlaz. Engelbrecht i sar. (1995) koriste analizu osetljivosti za procenu značaja ulaznih variabli, koristeći precizno ozračunavanje izvoda kako bi odredili značaj uticaja pojedinih dominantnih ulaznih variabli. U radu Furundic, 1988 autor analizira značaj uticaja ulaznih variabli u modeliranju procesaa padavine-oticaj, koristeći metodu merenja uticaja kontaminacije ulaznih variabli malim slučajnim vrednostima na promenu performansi izlaza. Pokazalo se da pojedine variable imaju presudan značaj za posmatrani izlaz dok su neke druge imale minorni uticaj na model pa su kao takve eliminsane iz dalje analize u cilju redukcije dimanzija ulaza i kompleksnosti modela. U radu Furundzic i sar., 1998,autori su analizirali uticaja ulaznih variabli (faktora rizika) na pojavu malignih bolesti, gde je ispitan uticaj perturbacije ulaznih variabli na promenu performansi obučene NM. Perturbacija je izvedena preko niza slučajnih binarnih (1,0) vrednosti sa obzirom na to da je veliki broj analiziranih faktora rizika definisan u formi binarnog vektora. U radu Furundžić i sar., 2009 autori detaljno prikazuju proces određivanja uticaja relevantnih parametara

Na kvalitet artikulacije fonema š, koristeći metod Numeričke Analize Senzitivnosti (NSA) (Montano i Palmer, 2003) i NM strukture. Isti autor, kao koautor u radu Naumović i sar., 2010 učestvuje u istraživanju gde koriste NM za estimaciju prediktivnih faktora i efekata terapije kod problema idiopatske membranozne nefropatije. Rezultati tog istraživanja su redukcija broja uključenih faktora i rangiranje različitih terapeutskih protokola po broju izlečenih pacijenata. Metod numeričke analize senzitivnosti (NSA), (Montano i Palmer, 2003) je zasnovan na izračunavanju nagiba između ulaza i izlaza, bez ograničenja i pretpostavki u vezi sa prirodom uključenih variabli. U ovom istraživanju je korišćena jednostavna i efikasna metoda za ocenu uticaja ulaznih variabli, pomenuta '*clamping tecnique'*, koja se pored ocene uticaja može prilagoditi i za problem detekcije prirode uticaja ovih variabli na izlaze, odnosno za aproksimaciju funkcionalne zavisnosti izlaza od ulaznih variabli. Informacija o prirodi zavisnosti izlaza od ulaznih variabli ima veliki značaj u inverznoj analizi uzročno posledičnih veza u procesu artikulacije.

7. REZULTATI PROCENE KVALITETA ARTIKULACIJE

Pouzdana klasifikacija je uslovljena reprezentativnošću obučavajućeg uzoraka. Reprezentativnost uzoraka direktno korelisana sa balansom suprotstavljenih klasa, odnosno sa stepenom uniformnosti raspodele primeraka klasa po prostoru njihovih obeležja. U realnim uslovima balans podataka u smislu njihove ravnomernosti raspodele često je ozbiljno narušen. Iz tog razloga su razvijene razne tehnike balansiranja uzoraka a problem je definisan kao imbalanced learning. Ovo je jedan od najvećih izazova u oblasti učećih induktivnih prediktora. Istraživanje prikazano u disertaciji u najvećoj meri je posvećeno ovom problemu i kao rezulat toga dizajniran je novi algoritam za optimalno određivanje obučavajućeg uzorka u, **D**istance **B**ased **B**alancing (DBB) algoritam (Furundžić i sar. 2017c). Krajnji korak u obradi podataka se svodi na donošenje odluke na osnovu raspoloživih podataka što često podrazumeva njihovu klasifikaciju. Zato bi doprinos na ovako bazičnom nivou mogao imati značaj.

Glavni cilj ovog potpoglavlja je prikaz rezultata poređenje nekoliko standardnih metoda balansiranja datih u literaturi sa novo predloženim DBB metodom kako bi se procenilo da li predloženi metod može ravnopravno da se nosi u praksi sa problemima vezanim za disbalans klasa. Ovde su ukratko opisani primeri oversamplinga [a), b), ..., g)] i metode undersamplinga [e), f), ..., m)] koje koristimo. Da bi izvršili potrebno poređenje, odabrali smo dvadeset skupova podataka različitih tipova i stepena disbalansa.

7.1. Primenjene metode balansiranja

U ovim eksperimentima korišćene su sledeće standardne tehnike u svrhu upoređivanja sa DBB algoritmom:

a) SMOTE oversampling,

je tehnika heurističkog oversamplinga (Chavla i sar., 2002), zasnovana je na generisanju novih sintetičkih primera manjinske klase koristeći efikasnu tehniku sintetičkog oversamplinga. Ovaj algoritam dodaje unapred definisani skup novih primera sintetičkih manjinskih klasa ekstrapolacijom između postojećih instanci manjinskih klasa, a ne prostom replikacijom postojećih primeraka. Sintetičko generisanje novih slučajeva manjinske klase dovodi do širenja granice odlučivanja u region većinske klase, smanjujući tendenciju klasifikatora da zanemaruje manjinske slučajeve. Pored prednosti, ova tehnika takođe ima svoje nedostatke što je dovela do mnogih modifikacija ovog algoritma predstavljenog u literaturi (Chavla i sar., 2003; Hongiu i Herna, 2004; He et al., 2008; Garcia et al., 2008; Garcia, 2009).

b) ADASYN algoritam (He et al., 2008) koristi sistematski metod za adaptivnu kreaciju različitih količina sintetičkih podataka prema njihovim raspodelama. Osnovna ideja je da se raspodela gustine koristi kao kriterijum za definisanje broja sintetičkih uzoraka koje treba generisati za svaki primer manjinske klase.

c) Borderline-SMOTE tehnika se koristi za identifikaciju uzoraka manjinske klase koji služe kao polazni uzorak za stvaranje sintetičkih primera za manjinske primjere u blizini granične oblasti.

d) Distance Based Balancing (DBB), је нови алгоритам подвргнут компаративној analizi. Овај алгоритам је техника заснована на подацима намењена повећању репрезентативности узорака, и стога га упоређујемо са одговарајућим најсавременијим техникама узорковања. *e) Random oversampling (ROS),* je ne-heuristički metod za balansiranje klase kroz slučajnu replikaciju primera manjinskih klasa (DeRouin et al., 1991). Ovaj metod povećava ravnotežu u klasama bez dodavanja novih informacija podacima što može dovesti do overfitinga manjinske klase Japkovicz (2001), Mitchell (1997) i Estabrooks (2000). Replikacija manjinske klase ne dovodi do širenja granice oblasti u region većinske klase (Chavla i sar., 2002).

f) SMOTE+Tomek algoritam (Batista i sar., 2004) je kreiran kako bi se stvorili bolje definisani klasteri unutar klase koji se primjenjuju na većinsku klasu u cilju uklanjanja nepotrebnih instanci.

g) SMOTE+ENN algoritam (Batista i sar., 2004) je jednostavna modifikacija algoritma Smote + Tomek. ENN ima tendenciju da ukloni više primera nego što Tomek link metoda, pa se očekuje da će pružiti detaljnije čišćenje redundantnih instanci.

h) Wilson's Edited Nearest Neighbor Rule (ENN) algoritam (Wilson, 1972) balansira uzorke uklanjanjem bilo kog primera čije se oznake klase razlikuje od klase najmanje dva od tri najbližih suseda.

i) Most Distance metod (Zhang and Mani, 2003), balansira uzorke odabirom primera većinske klase čija je prosečna razdaljina do tri najbližih primera manjinskih klasa najveća.

j) NearMiss-2 metod (Zhang and Mani, 2003), bira primere većinske klase čija je prosečna razdaljina od tri najdalje primera manjinskih klasa najmanja.

k) Neighborhood Cleaning Rule (NCL) (Laurikkala, 2001) koristi ENN algoritam za uklanjanje instance većinske klase.

l) Random undersampling (RUS), ne-heuristički metod koji smanjuje broj slučajeva većinske klase pri očuvanju manjinske klase. Najveći nedostatak slučajnog undersamplinga je to što ovaj metod može odbaciti potencijalno korisne podatke koji bi mogli biti važni za indukcioni proces

(Batista i sar., 2004). Empirijski je dokazano da slučajni undersampling može biti efektivan metod rezampinga (Garcia et al., 2008). Jedan od pionirskih radova na poboljšanju slučajnog ponovnog uzorkovanja, koji su izvodili Kubat i Matvin (1997), predlaže jednostranu tehniku selekcije (OSS) na osnovu smanjenja broja instanci većinske klase (Hart, 1968) i uklanjanja primjera iz zašumljenih područja (Tomek, 1976).

m) Tomek links (Tomek, 1976), algoritam se može koristiti kao metod balansiranja ili kao metod uklanjanja redundantnih podataka ili šuma. Kada se koristi za undersampling, ovaj metos eliminiše samo primere koji pripadaju većinskoj klasi. Kada se koristi kao metod za čišćenja podataka (Smote + Tomek), uklanjaju se primeri iz obe klase radi povećanja separabilnosti klasa.

7.2. Opis baza podataka

Dvadeset (20) skupova podataka koji su korišćeni u ovoj eksperimentalnoj studiji su navedeni u Tabeli 7.1. Uključene skupove podataka karakteriše raznolikost broja atributa koji varira od 2 do 32, raznolikost odnosa neravnoteže koji varira od 1 (uravnoteženo) do 72 (Visoko neuravnoteženo). Neki od uključenih skupova podataka sadrže kontinualne i kategorične atribute. Za svaki skup podataka u Tabeli 7.1 prikazan je broj primera (N), broj primera u negativnom (No) i pozitivnom (N +) klasi, IR vrednosti ($IR = N^+/N^o$) i broj atributa. Za skupove podataka sa više od dve klase definisali smo manjinsku klasu kao pozitivnu, a sve ostale tretirane su kao negativne klase. Podsećamo da u ovom radu razmatramo samo probleme binarne klasifikacije. Svi algoritmi su testirani na tri sintetička skupa podataka, trinaest benchmark skupova podataka i na kraju, metode su testirane na četiri uzorka laboratorijskih podataka dobijenih u toku projekata iz patologije govora.

a) Sintetički podaci

Generisali smo određeni broj dvoklasnih sintetičkih uzoraka na domenima 2D X-OR sa različitim neizbalansiranim raspodelama podataka. Uzorci za obuku su podeljeni u dve kategorije, originalne uzorke i izbalansirane uzorke dobijene pomoću opisanih metoda za resemplovanje. Svaki original (XOR2, XOR8 i XOR32), dat u Tabeli 5, sastoji se od dve klase N_i, instance gde je $i = 1, 2, 3$. Manjinska klasa se sastoji od dva sub-klastera identičnog broja ($N^+/2$) primera koji se nalaze na prvom (I) i trećem (III) kvadrantu koordinatnog sistema. Većinska klasa se sastoji od dva pod-klastera identičnog broja ($N^o/2$) primera lociranih u drugom (II) i četvrtom (IV) kvadrantu. Sub klasteri se ne preklapaju i svaki od njih predstavlja skup podataka E, gde je svaki primer $E_j \in E$ uređena dvojka $E_j = (x_j, y_j)$. Originalni sub-

klasteri sadrže primere koji su normalno distribuirani u 2D domenu. Normalna raspodela instanci se smatra unutrašnjim disbalansom klase. Pored ovoga, postoji i projektovana neuravnoteženost između klasa, koja je predstavljena kroz IR vrijednosti, gdje je $IR = \{2, 8, 32\}$. Svi izbalansirani obučavajući uzorci, koji su proizvedeni oversampling tehnikama (DBB, ROS, SMOTE), sastoji se od dve klaseod po N^o instanci, dok izjednačeni uzorci proizvedeni od strane RUS-a sadrže dve klase N^+ instanci. Svi izbalansirani uzorci imaju identičan broj pozitivnih i negativnih primera ($N^+ = N^o \Rightarrow IR = 1$).

b) Uci repository data

Ova grupa sadrži četiri skupa podataka izabranih iz benchmark podataka UCI repozitorijuma (Blake i Merz, 1998) koji su koristili (Chavla, et al., 2002; Vu and Chang, 2003; Batista i sar., 2004; Phung et al., 2009) . Tabela 7.3 (vrste 9 do 12) prikazuju podatke korišćene u ovoj studiji. Za bazu kvasca definisali smo CIT klasu kao pozitivnu a sve ostale su tretirane kao negativne klase. Za bazu Abalone smo koristili četvrtu klasu kao pozitivnu, a sve druge klase kao negativnu klasu. Za bazu Arrhythmia smo označili šestu klasu kao pozitivnu i smanjili broj atributa na 14. Za bazu Letter koristili smo slovo A kao pozitivnu, zadržavajući sve atribute.

c) Elena project data[1]

Ova grupa podataka koja se sastoji od sledećih tri skupa podataka: Iris, Phoneme i Satimege, se često koristi za procenu metoda za balansiranje podataka pri obuci induktivnih učećih algoritama (Chavla, et al., 2002; Batista, i sar., 2004).

d) Podaci o kvalitetu artikulacije

Naši laboratorijski uzorci sadrže skup akustičnih karakteristika govornog signala fonema koji pripadaju grupi frikativa srpskog jezika. Specifični fonemi su izgovoreni u početnoj poziciji reči gde se pojavljuju. Svaka instanca uzorka pripada jednoj od dve klase, od kojih većinska klasa sadrži primere normalne artikulacije dok manjinska klasa sadrži primere patološke artikulacije. Baza podataka čija su glavne karakteristike prikazane u Tabeli 7.2, uključuju podatke od sledećih četiri fonema: *sh* /ʃ/, *zh* /ʒ/, *s*, i *z*, kao tipične predstavnike grupe frikativa. Svaka instanca predstavljena je vektorom izabranih atributa dužine 19. Prva grupa od dvanaest karakteristika predstavlja Mel Frekventnih Kepstralnih Coefficienata (MFCC) uzima se iz svakog frejma. Broj okvira varira za svaku fonemu, tako da atributi MFCC-a svakog fonema predstavljaju matricu od M vrsta i 12 kolona, gdje M predstavlja broj frejmova u stvarnom fonemu, a za svaki okvir postoji 12 vrednosti koeficijenata. Tokom procesa obuke svakom okviru aktuelnog fonema pridružena je odgovarajuća labela klase. Tokom procesa testiranja,

[1] Elena project data, https://www.elen.ucl.ac.be/neural-nets/Research/Projects/ELENA/databases/REAL/

izlazi iz svakog ulaznog okvira svake foneme se sabiraju i usrednjavanjem se dobija procena pripadnosti nekoj od datih klasa fonema.

Druga grupa od tri atributa datih u formi realne vrednosti koja se odnose na dužinu talasnog signala, sastoji se od broja semplova (nw) talasnog signala koji se odnosi na reč koja u početnoj poziciji sadrži aktuelne foneme, broj uzoraka talasnog oblika signala koji se odnose na aktuelne foneme (nph), i odnos dužine ($nq = nph/nw$).

Na kraju, treća grupa se sastoji od četiri sledeća atributa koja se odnose na distribuciju energije frejmova: energija frejma foneme (Ef), energija frejma ostatka odgovarajuće reči (Er), ukupna energija fonema (Efu) i ukupna energija ostatka aktuelne reči (Eru).

Glavni problem analitičkog sistema za procenu kvaliteta artikulacije je adekvatan izbor akustičkih parametara koji su dovoljno pouzdani za finu diskriminaciju unutar iste kategorije fonema.

Tabela 7.1
Opšte karakteristike baza podataka.

	Data set	N	N^o	N^+	IR	#Atr.
1	Abalone IV class	4177	4120	57	72.2800	9
2	Arrhythmia VI class	452	427	25	17.0800	14
3	Br. Cancer	699	458	241	1.9004	9
4	EEG Eye State	14976	8255	6721	1.2282	14
5	E. Coli	336	301	35	9.6000	7
6	Heart Disease	294	188	106	1.7736	13
7	Iris	150	100	50	2.0000	4
8	Letter A	20000	19211	789	24.3480	16
9	Magic	19020	12332	6688	1.8439	10
10	Phoneme	5404	3818	1586	2.4073	6
11	Phon. sh/ʃ/	314	218	96	2.7080	19
12	Phoneme s	185	121	64	1.8906	19
13	Phon. zh/ʒ/	279	188	91	2.0659	19
14	Phoneme z	189	117	72	1.6250	19
15	Pima Diabetes	768	500	268	2.1194	8
16	Satimage	6435	5809	623	9.3242	36
17	XOR2	2700	1800	900	2.0000	2
19	XOR32	1856	1800	56	32.1420	2
19	XOR8	2025	1800	225	8.0000	2
20	Yeast	1484	1021	463	2.2052	8

Ovaj problem kategorizacije iste grupe fonema (frikativa) u dve kategorije (normalni i patološki) je znatno teži od klasifikacije fonema različitih grupa. Upoređivanjem rezultata klasifikacionih frikativa iz naše baze podataka sa benchmark bazom Phoneme (*Elena project data*) dobijamo priliku da procenimo relevantnost akustičkih karakteristika koje smo odabrali. Naime ukoliko tačnost klasifikacije naših podataka bude u rangu tačnosti klasifikacije baze Phoneme tada možemo tvrditi da su obeležja izabrana za našu bazu fonema pouzdani prediktori kvaliteta artikulacije. Naravno, ovde podrazumevamo komparaciju rezultata istih kategorija klasifikatora.

7.3. Primenjeni klasifikatori

Jedan od važnijih ciljeva ovog rada je pouzdana indirektna komparativna ocena uticaja DBB algoritma na performanse izabranih klasifikatora, u poređenju sa uticajima skupa od trinaest (13) postojećih relevantnih metoda balansiranja. Unapređenje performansi klasifikatora ovde nije jedini cilj pošto klasifikatori služe i kao indikator uticaja pomenutih algoritama. Pošto je glavni problem disertacije što pouzdanija ocena kvaliteta artikulacije fonema, odnosno kategorizacija na normalne i patološke, koristili smo različite klasifikatora kako bismo kroz njihovu komparaciju došli do optimalnog rešenja. U tom smislu izabrali smo četiri različita klasifikatora: ansambl MLP (višeslojni perceptron) i ansambl KNN (k najbližih suseda), Samoorganizujuće Mape (SOM) i Naive Bayes. Sledeća bitna napomena odnosi se na pojam objektivizacije logopedskog načina ocenjivanja jer su njihove ocene korišćene kao izlazni trening skup podataka. Korišćene su usrednjene vrednosti ocena grupe od pet logopeda kao objektivan uzorak, odnosno etalon, i prema njemu su izvršeni obuka i testiranje klasifikatora.

MLP ansambl je robustni klasifikator (Hansen i Salamon, 1990), koji i sam spada u kategoriju algoritama za balansiranje zasnovanih na modelima. Ansambl MLP uspostavlja uslove za konvergenciju ka optimalnom rešenju, za razliku od individualnih MLP struktura koje često upadaju u lokalne minime (suboptimalna rešenja). Ovaj algoritam teži stvaranju jednakih preduslova za indirektno upoređivanje efikasnosti algoritama za balansiranje. MLP ansambl kao i većina klasifikatora spada u kategoriju induktivnih prediktora koji su determinisani podapodacima (data-driven) a koji su uglavnom osetljivi na unutrašnji disbalans obučavajućih uzoraka. Stanje disbalansa uzrokuje favorizaciju većinske klase i klastera tokom treninga i takođe slabiju generalizaciju, odnosno ekstrapolaciju znanja izvan prostora obuke. Treba napomenuti da stohastička selekcija inicijalnih parametara pojedinačnih MLP struktura i način i izbora tzrening test uzoraka utiču na determinizam performanse pojedinačnih MLP klasifikatora. Zbog toga su korišćeni ansambli MLP kako bi se redukovala ova pojava. Ovaj klasifikator, dizajniran je originalnim algoritmom izbora optimalnog skupa obučenih MLP struktura (Furundžić i sar. 2012a), i detaljno je prikazan u poglavlju 6. Za svaku izvornu bazu smo napravili klasifikator u obliku 50 MLP struktura po ansamblu. Pojedinačne MLP strukture smo odredili postepenim povećanjem njihove složenosti povećavajući broj neurona u skrivenim slojevima, do određivanja prihvatljivih modela. Korišćen je Levenberg-Markuardt algoritam sa adaptivnim momentom kako bi se osigurala efikasna konvergencija i obuka. Overfiting modela je kontrolisan skupom validacionih podataka. Tokom procesa obuke svaki

MLP ansambl (klasifikator) formuliše svoj odgovor na osnovu kriterijuma većinskog glasanja (usvaja se odgovor većine od 50 individualnih MLP struktura).

KNN ansambl kao drugi primijenjeni klasifikator, sastoji se od sledećih pet pojedinačnih KNN klasifikatora: 1NN, 3NN, 5NN, 7NN i 9NN klasifikatora. Klasifikatori KNN-a jednostavno klasifikuju test uzorke koristeći Euklidsko rastojanje određenog broja najbližih suseda iz uzorka obuke, koji praktično služi kao referentni skup primera. Klasifikator KNN ansambl takođe formuliše svoj odgovor na osnovu kriterijuma većinskog glasanja, zato smo koristili neparan broj (5) klasifikatora u ansamblu. Treba napomenuti da su KNN klasifikatori determinističke strukture koje uvek daju nedvosmislena jednoznačna rešenja za isti skup uzoraka. Jedini stohastički uticaj na njihov odgovor ima slučajna selekciju trening i test uzoraka ali ovaj uticaj je mali. Takođe treba reći da KNN klasifikatori imaju brži odgovor u odnosu na MLP strukture. Budući da DBB algoritam omogućava značajno povećanje reprezentativnosti podataka, njegova primena treba da poboljša performanse klasifikatora koji se primjenjuju. Kao praktičan dokaz ove tvrdnje, trebao bi dokazati a priori pretpostavljenu direktnu korelaciju između mera reprezentativnosti uzoraka i mera performanse klasifikatora obučenih na ove uzorke. Preciznije, pretpostavlja se da su performanse klasifikatora indirektna posledica reprezentativnosti (ravnoteže) uzoraka za obuku. Stoga se kao indirektna mera efikasnosti balansnih algoritama uzima vrednost mere performansi naših klasifikatora.

SOM Klasifikator je korišćen kao komparativni metod a njegova primena se oslanja na klasterovanje uzoraka pri čemu se između klasa ne formira fleksibilna granična hiperpovrš kao kod perceptrona što je nedostatak ovog alata. Željeni izlaz ovog prediktora je definisaan samo kroz broj mogućih klastera trening uzorkaa i ovaj broj je iznossuo 2, tako da model na svaki ulazni test vektor generiše numeričku vrednost njegove pripadnosti klasi 1 ili klasi 0.

Naive Bayes Klasifikator je poznat kao vrlo efikasan probabilistički orijentisan klasifikator uporedivih performansi sa medernim klasifikatorima i zato je ovde uvršten kao komparativna metoda. Detalji o svim klasifikatorima su dati u poglavlju 6.

Treba istaći da smo u svim slučajevima obuke koristili sledeću raspodelu uzoraka: 50 % slučajno odabranih podataka korišćeno je za obuku klasifikatora, od čega je 10% korišćeno za proces validacije u slučaju MLP klasifikatora, ostalih 50% je korišćeno za testiranje performansi klasifikatora. Obučavajući i test uzoraci su se sastojali od slučajno odabranih 50% primera većinskih klasa i 50% primera manjinskih klasa za svaku bazu podataka, čuvajući originalnu stopu disbalansa. Uzorci za obuku su zatim podvrgavani procesu balansiranja primenom svih navedenih algoritama, dok su test uzorci uzeti u originalnoj formi korišćeni da bi se procenio pravi efekat balansiranja na performanse klasifikatora. Poredeći efikasnost

algoritama za balansiranje određujemo optimalni algoritam koji ćemo koristiti u preprocesingu obučavajućeg skupa za obuku navedenih klasifikatora.

7.4. Metrike za ocenu performansi klasifikatora

Definisali smo dve odvojene kvantitativne metrike procene za komparativno vrednovanje efikasnosti gore navedenih metoda za balansiranje odabranih baza podataka (direktne i indirektne metrike). Prve mere se odnose na kvantitativne vrednosti distributivnih svojstava originalnih baza podataka i njihovih izbalansiranih derivata. Ove distribucione karakteristike su predstavljene skupom vrednosti $\{\sigma(D),\ \mathrm{H(D)}\ \text{and}\ \sigma\mathrm{H(D)} = \sigma(D) * \mathrm{H(D)}\}$, gde je σ standardna devijacija, H je entropija jednodimenzionalnog niza srednjih lokalnih rastojanja (D) i σH je jednostavni proizvod ovih vrednosti koji pretpostavlja relevantnu kompoziciju oba faktora sa dokazanim uticajem. Kao što je navedeno u Odjeljku 3.6, maksimalnu reprezentativnost svih uzoraka od N instanci ima regularna rešetka čiji niz D_L ima nultu standardnu devijaciju $(\sigma_{D_L} = 0)$ i nultu entropiju ($H_{D_L} = 0$) Tako da je $\sigma H_{D_L} = 0$. Dakle, najveću reprezentativnost od realnih baza podataka imaće ona sa najmanijim vrednostima σ, H i σH (najbliže vrednostima σ_{D_L}, H_{D_L} i σH_{D_L} . Dakle kao najreprezentativniju bazu podataka, prihvatamo onu koja ima minimalnu vrednost $\min(\sigma H_D)$. Metoda balansiranja koja proizvodi najizbalansiranije derivate iz originalnog skupa podataka, u smislu minimuma ove mere, smatra se najefikasnijom metodom. Ukupna tačnost klasifikacija kao jedan od kriterijuma ocenjivanja nije pogodan za ocenu neuravnoteženih podataka (Guo i Viktor, 2004). Zato smo koristili pokazatelje za procenu odgovarajuće krive relativne operativne karakteristike (ROC) (Kubat et al., 1998, Favcett, 2003, Provost i Favcett, 1997 i Maloof, 2003) za procenu relevantnosti DBB metoda. Pretpostavimo da su p i n pozitivni i negativni test primeri a Y i N su predviđeni pozitivni i negativni rezultati klasifikacija dobijeni od strane obučenog klasifikatora. Vizuelni prikaz izvedene klasifikacije mogu se dati u obliku konfuzione matrice kao u Tabeli 7.2. Smatramo da je manjinska klasa pozitivna klasa a većinska klasa negativna..

Tabela 7.2
Konfuziona matrica za ocenu performansi.

Prava Klasa		Klasifikovan kao	
		Y	N
	n	$TP\ (True\ Positive)$	$FP\ (False\ Positive)$
	p	$FN\ (False\ Negative)$	$TN\ (True\ Negative)$

Koristili smo konfuzionu matricu da definišemo prikazane jednačine za metriku evaluacije obuke klasifikatora iz neuravnoteženih skupova podataka na sledeći način:

P: Broj pozitivnih realizacija,

N: Broj negativnih realizacija,

Sensitivity (True positive rate): $TPR = TP/P = TP/(TP + FN)$

Specificity (True negative rate): $SPC = TN/N = TN/(TN + FP)$

Specificity (Falce positive rate): $FPR = FP/N = FP/(FP + TN) = 1 - SPC$

ROC kriva prikazuje parametarsku kompoziciju vrednosti TPR(T) u funkciji FPR(T), gde je T promenljivi parametar koji predstavlja vrednost praga razgraničenja između funkcija gustina raspodele posmatranih klasa.

AUC: Predstavlja vrednost površine ispod ROC krive i performanse klasifikatora su direktno proporcionalne ovoj vrednosti, koja u slučaju tačne klasifikacije svih instanci iznosi 1.

Ostali, dole prikazani korisni derivati konfuzione matrice, se takođe koriste pri oceni performansi klasifikatora.

Accuracy: $A_{cc} = (TP + TN)/(TP + FP + TN + FN)$

Recall: $Rec = TP/(TP + FN)$

Precision: $Precis = TP/(TP + FP)$

G_mean: $G_{mean} = \sqrt{[TP/(TP + FN)] \times [TP/(TN + FP)]}$

F_Measure: $F_{Measure} = [(1 + \beta^2) * Recall * Precision]/[\beta^2 * Recall + Precision]$

Drugi, indirektni indikator efikasnosti metoda balansiranja predstavljen je merom performansi klasifikatora koji je podvrgnut obuci pomoću uravnoteženih derivata izvorne baze podataka, koji su dobijeni pomoću aktuelnih metoda balansiranja. Ovaj indikator je kondenzovani derivat konfuzione matrice (površina ispod ROC krive poznato pod nazivom AUC (Area Under the Curve). Ovaj indirektni indikator reprezentativnosti uzoraka treba da bude u negativnoj korelaciji sa direktnim pokazateljima reprezentativnosti ($\sigma(D)$, $\mathrm{H(D)}$ and $\sigma\mathrm{H(D)}$). Treba napomenuti da na AUC indikator takođe može uticati priroda ponašanja klasifikatora, koja nije uvek deterministička. Performanse idealnog klasifikatora izraženog ovim parametrom imaju sledeću vrednost: AUC = 1. Ovo je maksimalna vrednost koju parametar može imati. Prema tome, klasifikatori sa najvišim AUC vrijednostima će biti najbolje rangirani. Metoda balansiranja koja proizvodi balansirane derivate izvornih baza podataka koje povećavaju performanse (AUC) korišćenog klasifikatora do najveće vrednosti, u poređenju sa drugim metodama, smatraće se najefikasnijim čime pretenduje na upotrebu u realnim uslovima.

7.5. Eksperimentalna evaluacija efikasnosti DBB algoritma

U ovom poglavlju predstavljamo eksperimentalne rezultate komparativne evaluacije efektivnosti DBB algoritma u poređenju sa ostalih trinaest (13) relevantnih tehnika balansiranja primenjenih na prikazane originalne baze podataka. Za tu svrhu svaka od dvadeset (20) korišćenih baza podataka podleže svakoj od 13 metoda balansiranja, tako da za svaku bazu podataka dobijamo trinaest novih baza podataka, koje smo nazvali derivatima originala. Originali su takođe uključeni kao četrnaesti nebalansirani skup uzoraka. Opšta simbolička šema generisanja derivata je: Derivat_i_j = Metod_j (Original_i), gde se i = 1, 2, ..., 20., odnosi se na bazu i a j = 1, 2, ..., 14., se odnosi na metod j, tako da pored 20 originala dobijamo još 260 balansiranih derivata što je ukupno 280 baza podataka. U kontekstu uvedene šeme, Derivat_15_2 predstavlja balansiranu bazu podataka dobijenu od izvorne baze pomoću metoda ADASIN. Derivative_15_1 predstavlja bazu podataka Pima ORIGINAL. Svi derivati imaju svoje pozicije prikazane zaglavlju tabela (7.3, 7.4, 7.5 i 7.6). Prvo ćemo predstaviti rezultate dobijene prvim setom metrika za direktnu procenu efikasnosti. Oni predstavljaju distributivne karakteristike nizova lokalnih rastojanja D originalnih baza podataka D (ORIGINAL) i nizova D njihovih uravnoteženih derivata: D(ADASIN), D(BORDER SMT), D(DBB), ..., D(TOMEK), dobijenih korišćenjem Trinaest metoda koje smo testirali. Rezultati eksperimenta su prikazani u tabelama 7.3 i 7.4. Tabela 7.3 u prvoj koloni sadrži imena baza podataka, a druge kolone sadrže skup vrednosti $\sigma(\mathrm{D})$ za originalne baze podataka (druga kolona) i njihovih trinaest uravnoteženih derivata dobijenih pomoću dve grupe tehnika, oversampling (kolone 2: 8) i undersampling (kolone 9:14), respektivno. Bold-face podaci označavaju numeričke vrednosti koje prikazuju minimalne vrednosti $\sigma(\mathrm{D})$. Tabela 7.4 u prvoj koloni sadrži imena korišćenih baza podataka, a druge kolone sadrže skup vrednosti H(D) za originalne baze podataka (druga kolona) i njihovih trinaest uravnoteženih derivata identično raspoređenih kao u tabeli 6. Bold-face numeričke vrednosti se odnose na minimalne vrednosti H(D). Podsetimo na negativnu korelaciju između $\sigma(\mathrm{D})$ i H(D) s jedne strane i reprezentativnosti aktuelnog uzorka sa druge strane. Bold-face vrednosti u tablicama ukazuju na najefektivniju metodu koja se odnosi na određeni skup podataka. Vrednost proizvoda $\sigma(\mathrm{D})$* H(D), koja nije prikazana u tabeli, predstavlja pokazatelj reprezentativnosti uzoraka sa opštim značenjem i biće dalje računat i razmatran zajedno sa drugim pokazateljima. Ilustrativno predstavljanje eksperimentalne procene takođe je prikazano na slici .17, date u formi matrice 4 × 3 malih slika Sl. (i, j), i = 1: 4, j = 1: 3. Ova slika predstavlja vizuelni interni pogled na suštinu efekta balansiranja uzoraka korišćenjem predstavljenih tehnika balansiranja. Slika 7.1 predstavlja

distributivne karakteristike niza lokalnih distanci D preko funkcije gustine verovatnoće ilustrativnog primera originalne baze podataka Pima i njenih derivata. Svaka mala slika predstavlja *pdf(D)* originalnu bazu Pima i njene derivate dobijene jednim od prikazanih dvanaest tehnika balansiranja. Kod svih slika, puna plava linija se odnosi na originalne a tačkaste crvene linije se odnose na balansirane derivate, tako da je prikazano ukupno trinaest uzoraka. Prva mala slika Sl. (1,1) na Slici 7.1. predstavlja *pdf(D)* izvorne baze podataka i njen DBB balansirani derivat. DBB algoritam balansira gustine instanci u prostora instanci, što se manifestuje značajnim povećanjem amplitude *pdf* krive, što pokazuje značajno smanjenje σ(D) balansiranog uzorka u poređenju sa originalnom.

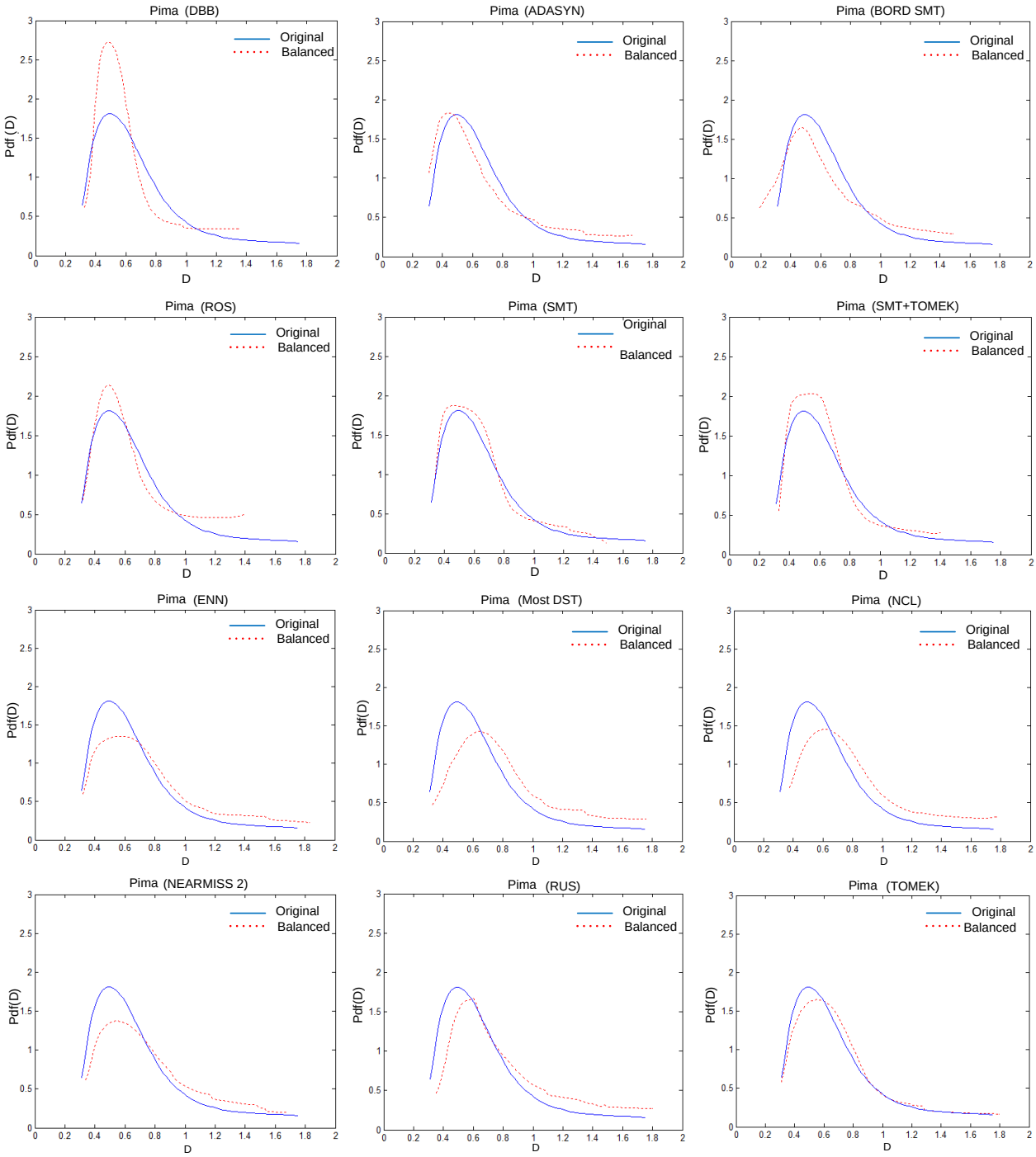
Tabela 7.3
Vrednosrti standardne devijacije $\sigma(D)$ za originalne i balansirane baze podataka.

		over-sampled data							under-sampled data					
Data Set	ORIG	ADAS	BOR SMT	DBB	ROS	SMT	SMT+ ENN	SMT+ TOM	ENN	MOST DIST	NCL	NEAR MISS2	RUS	TOM
Abalone	0.0836	0.0868	0.0882	**0.0821**	0.0914	0.1176	0.0884	0.0882	0.0830	0.3618	0.0826	0.5020	0.5493	0.0836
Arrhythmia	0.3781	0.3922	0.3792	**0.3485**	0.4288	0.3860	0.4038	0.3972	0.3788	0.4391	0.3931	0.3565	0.3419	0.3788
Br. Cancer	0.1497	0.1512	0.1787	0.1616	**0.1416**	0.1498	0.1524	0.1503	0.1551	0.2087	0.1567	0.1573	0.1725	0.1523
EEG Eye St.	0.0757	0.0750	0.0772	**0.0699**	0.0746	0.0736	0.0745	0.0742	0.0762	0.0743	0.0798	0.0882	0.0787	0.0761
E. Coli	0.3190	0.2854	0.3012	**0.2643**	0.2980	0.2922	0.2926	0.2916	0.3194	0.5855	0.3076	0.2880	0.3052	0.3128
Heart	0.5057	0.4720	0.4140	**0.2456**	0.4220	0.4538	0.4270	0.4674	0.6762	0.4444	0.4053	0.4935	0.4591	0.5170
Iris	0.1204	0.1372	0.1542	0.1204	0.1512	0.1388	0.1388	0.1388	**0.1200**	0.1382	0.1204	0.1255	0.1468	0.1204
Letter A	0.1329	0.2426	0.2688	**0.1303**	0.2424	0.2216	0.2216	0.2216	0.1329	0.2669	0.1329	0.2340	0.3124	0.1329
Magic	0.1390	0.1294	0.1324	0.1274	**0.1235**	0.1266	0.1314	0.1292	0.1522	0.1514	0.1511	0.1574	0.1519	0.1459
Phoneme	0.3327	0.2406	0.2488	0.2349	**0.2341**	0.2349	0.2457	0.2395	0.3447	0.4637	0.3482	0.3495	0.3833	0.3385
Phon. sh/ʃ/	0.3078	0.3236	0.3234	**0.2978**	0.3012	0.3144	0.3292	0.3218	0.3180	0.4251	0.3126	0.3237	0.3734	0.3117
Phoneme s	0.0706	**0.0618**	0.0686	0.0626	0.0640	0.0640	0.0648	0.0648	0.0742	0.0890	0.0739	0.0868	0.0823	0.0748
Phon. zh/ʒ/	0.2586	0.2450	0.2428	0.2350	**0.2010**	0.2336	0.2474	0.2428	0.2861	0.2804	0.2606	0.2679	0.2745	0.2743
Phoneme z	0.1620	0.3021	0.3570	**0.1529**	0.3215	0.2992	0.3045	0.3004	0.1683	0.2090	0.1627	0.1933	0.1958	0.1649
Pima	0.2125	0.1928	0.2056	**0.1622**	0.1926	0.1880	0.2048	0.1960	0.2426	0.2232	0.2170	0.2207	0.2158	0.2272
Satimage	0.3217	0.2986	0.3158	0.2912	**0.2853**	0.2990	0.3048	0.3028	0.3258	0.5860	0.3288	0.4120	0.6528	0.3242
XOR2	0.0271	0.0230	0.0228	0.0215	0.0226	0.0216	**0.0214**	0.0214	0.0268	0.0303	0.0272	0.0292	0.0321	0.0270
XOR32	0.0360	0.0218	0.0244	0.0192	**0.0166**	0.0178	0.0172	0.0178	0.0352	0.1008	0.0362	0.1022	0.0873	0.0363
XOR8	0.0328	0.0272	0.0264	0.0183	0.0202	0.0186	**0.0180**	0.0181	0.0328	0.0490	0.0330	0.0502	0.0464	0.0324
Yeast	0.2141	0.1950	**0.1841**	0.1847	0.2022	0.1996	0.2032	0.2120	0.2315	0.2510	0.2461	0.1845	0.2458	0.2318

Tabela 7.4
Vrednosti entropije $H(D)$ za originalne i balansirane baze podataka.

		over-sampled data							under-sampled data					
Data Set	ORIG	ADAS	BOR SMT	DBB	ROS	SMT	SMT+ ENN	SMT+ TOM	ENN	MOST DIST	NCL	NEAR MISS2	RUS	TOM
Abalone	3.6751	2.8598	2.7824	**2.6126**	2.6268	2.7731	2.7599	2.7669	3.2363	4.2023	3.2314	3.6470	4.2451	3.2306
Arrhythmia	4.7447	4.2576	4.3692	**3.1504**	4.3640	4.3945	4.4441	4.4458	4.2183	3.3360	4.2271	3.5425	3.6368	4.2183
Br. Cancer	3.6755	4.9279	4.5862	3.6406	4.8875	4.9509	4.9052	4.9112	3.7054	3.7686	3.5574	3.8451	3.7554	3.6922
EEG Eye St.	3.7284	3.9255	3.9653	3.5656	3.8047	3.8345	3.8361	3.8316	3.6054	3.5787	3.5587	**3.5253**	3.6314	3.6020
E. Coli	**3.2634**	3.5406	3.5861	3.3236	3.7243	3.5896	3.6332	3.6310	3.5986	3.8000	3.5117	3.5102	4.1973	3.4977
Heart	4.7454	4.7691	4.6783	**4.0779**	4.5358	4.6400	4.6082	4.5883	4.8283	4.4849	4.4275	4.5992	4.4967	4.7235
Iris	2.8561	3.5003	3.5598	2.8552	3.4363	3.4898	3.4898	3.4898	2.8552	3.0245	2.8552	2.9724	2.9265	**2.7024**
Letter A	4.5590	3.3005	2.4585	3.3852	**3.1456**	3.4182	3.4182	3.4182	3.9982	3.7796	3.9987	4.2412	4.5418	3.9988
Magic	3.2688	3.5491	3.2641	**3.1057**	3.4110	3.4727	3.2073	3.5164	3.3966	3.6179	3.4883	3.5116	3.5014	3.3904
Phoneme	4.2298	4.1161	4.2147	3.8784	4.1168	4.0887	4.1173	4.1093	4.0899	4.3689	4.0305	4.1849	**3.8672**	4.0500
Phon. sh/ʃ/	4.4621	4.9071	4.9293	4.3405	4.8215	4.8984	4.9467	4.9393	4.0352	4.3943	**4.0122**	4.1519	4.0185	4.0412
Phoneme s	4.4719	4.2179	4.3616	**3.8505**	4.2859	4.2439	4.2283	4.2251	4.3889	4.0880	4.1595	4.0980	4.2151	4.6330
Phon. zh/ʒ/	3.8532	3.7674	3.6723	**3.2331**	3.7532	3.6632	3.6883	3.6999	3.9068	3.9464	3.6698	3.6943	3.8456	3.8392
Phoneme z	3.8800	3.7998	4.1092	**3.5269**	3.9785	3.8847	3.8534	3.8317	3.8079	4.0321	3.6749	3.7744	4.1113	3.9813
Pima	4.3833	4.2948	4.5776	**4.0557**	4.4245	4.4633	4.4135	4.4962	4.4427	4.4465	4.3844	4.3768	4.2886	4.4420
Satimage	5.0698	4.3854	4.5628	4.2686	**4.1682**	4.3373	4.3520	4.3476	4.3417	4.4255	4.3313	4.1985	4.2127	4.3402
XOR2	3.6915	3.7297	3.5817	**3.4945**	4.0401	3.9104	3.9055	3.9055	3.6872	3.7629	3.7055	3.7177	3.9286	3.6899
XOR32	2.6913	3.5997	2.3470	2.6449	**2.2434**	3.2117	2.6243	3.2123	2.6811	4.4332	2.7059	4.5444	4.4215	2.6907
XOR8	3.2061	2.8826	2.6856	**2.4375**	2.8184	3.4052	3.3979	3.3979	3.2158	4.1729	3.2142	4.0786	4.0473	3.2081
Yeast	3.5360	3.5223	3.4804	3.5498	3.6791	3.6064	3.6015	3.6586	3.4852	3.8024	3.6320	**3.4028**	3.6284	3.8777

Dakle, ova kriva pokazuje sličnost sa krivom $pdf(D_{bi})$ dobijenom od balansiranog uzorka, datog u poglavlju 5. na Slici 5.5. Sve ostale slike pokazuju *pdf* krivu koje su približnog oblika kao *pdf* krive originala [Slika. (1,2), Sl. (2,2)], ili su mnogo bliže uniformnoj distribuciji, u odnosu na *pdf* krivu originalnog [Sl. (3,1), Sl. (3,2) , Sl. (3,3), Sl. (4,1)]. Ove činjenice ukazuju na povećanje vrednosti $\sigma(\mathrm{D})$ i $H(\mathrm{D})$, drugim rečima, smanjenje reprezentativnosti transformisanih uzoraka u odnosu na original.



Slika 7.1 Uticaj raznih tehnika balansiranja na karaktersistike raspodele (*pdf*) srednjih lokalnih rastojanja za originalni Pima skup podataka (puna plava linija) I korespondentne balansirane derivate (isprekidana crvena linija).

Da podsetimo da DBB algoritam balansira sve klase trening uzorka, tako da većinska klasa zadržava isti broj primera, dok se broj instanci manjinske klase povećava do broja primera većinske klase. Treba istaći da prezentacija *pdf* krivih koje se odnose na srednje lokalne udaljenosti predstavlja ilustrativni indirektni unutrašnji uvid u raspodelu instanci n-dimenzionalnog uzorka. Tabele 7.5 i 7.6 prikazuju indirektne mere efikasnosti metoda balansiranja, dobijenih merenjem performansi dva različita klasifikatora: ansambl MLP i KNN ansambl, koji prolaze kroz uravnotežene derivate originala, kao i originalne skupove podataka. Ove mere su rezultati interakcije algoritama za balansniranje, karakteristika klasifikatora i slučajnog izbora obučavajućeg uzorka iz baze podataka. Zbog stohastičke prirode MLP struktura, takođe smo uključili KNN ansambl kao deterministički algoritam kako bismo dobili komparativnu procjenu efekata balansiranja. Pored originalnih performansi MLP i KNN, u analizi rezultata izračunate su performanse vrednosti fiktivnog hibridnog klasifikatora (MLP + KNN) korišćenjem usrednjene vrednosti performansi MLP i KNN klasifikatora. Treba imati na umu da je u prvoj grupi direktnih metrika [$\sigma(\mathrm{D}), H(\mathrm{D}), \mathrm{i}\, \sigma H(\mathrm{D})$] rangiranje rezultata u pozitivnoj korelacija sa metrikom, bolji rang predstavljen je nižom vrednošću i on odgovara nižoj vrednosti metrike, dok je u slučaju indirektne grupe pokazatelja (AUC), pozicija rangiranja u negativnoj korelaciji sa metrikom, što znači da bolji rang (niža vrijednost) odgovara većoj vrednosti AUC.

Kompletni eksperimentalni rezultati performansi koji se odnose na sve skupove podataka, sve metode balansiranja i sve klasifikatore predstavljeni su u Tabelama 7.5 i 7.6. Koristeći rezultate prikazane u tabelama 7.3, 7.4, 7.5 i 7.6, prvo moramo dokazati pretpostavku o negativnoj korelaciji između direktnih i indirektnih pokazatelja balansa. U tu svrhu izračunavamo Pearsonove koeficijente korelacije. Tabela 7.7 pokazuje koeficijente korelacije R, *p*-vrednosti kao i vrednosti RL i RU. R koeficijenti mogu da se kreću od -1 do 1, gde -1 negativnu korelaciju, 0 predstavlja nepostojanje korelacije, a 1 predstavlja direktnu, pozitivnu korelaciju. Predstavljene *p*-vrednosti testiraju nultu hipotezu da ne postoji veza između posmatranih metrika. Prikazane *p*-vrednosti se kreću od 0 do 1, pri čemu vrednosti blizu 0 odgovaraju statistički značajnoj korelaciji u R. Ako je *p*-vrednost manja od usvojenog nivoa značajnosti (0,05), onda se odgovarajuća korelacija smatra značajnom. Prikazane vrednosti RL su 95% od intervala pouzdanosti donje granice za aktuelne koeficijente R, dok su vrednosti RU gornje granice za aktuelne koeficijente R. Tabela 7.7 pokazuje matricu 3x12 koja sadrži koeficijente korelacije između vrednosti direktne i indirektne metrike. Dimenzije podataka

direktnih i indirektnih metrika su 20x14. Ove matrice su prvo transformisane u vektor kolonu dimenzija 280x1, a zatim su izračunate vrednosti koeficijenta korelacije.

Tabela 7.5
Rezultati klasifikacije (AUC) za originalne i balansirane baze podataka za MLP Ansambl klasifikator.

		over-sampled data							under-sampled data					
Data Set	ORIG	ADAS	BOR SMT	DBB	ROS	SMT	SMT+ ENN	SMT+ TOM	ENN	MOST DIST	NCL	NEAR MISS2	RUS	TOM
Abalone	0.9831	0.9801	0.9799	**0.9989**	0.9856	0.9777	0.9825	0.9844	0.7799	0.9856	0.9880	0.7212	0.9712	0.9878
Arrhythmia	0.8310	0.8082	0.8046	**0.8678**	0.8014	0.7891	0.8397	0.8505	0.8639	0.7895	0.8588	0.7140	0.8454	0.8537
Br. Cancer	0.9928	0.9785	0.9789	0.9892	0.9787	0.9763	0.9912	0.9914	0.9907	0.9913	0.9921	0.9915	**0.9931**	0.9917
EEG Eye St	0.9225	0.9851	0.9893	0.9959	**0.9974**	0.9836	0.9963	0.9945	0.9417	0.9325	0.9175	0.9654	0.9122	0.9352
E. Coli	0.9074	0.9085	0.8956	0.9456	0.9044	0.9052	0.9481	**0.9589**	0.9185	0.9270	0.9078	0.9019	0.9263	0.9074
Heart	0.8623	**0.8868**	0.8557	0.8682	0.8641	0.8601	0.8639	0.8675	0.8264	0.8661	0.8647	0.8671	0.8739	0.8683
Iris	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Letter A	0.9824	0.9935	0.9891	0.9944	**0.9956**	0.9916	0.9941	0.9927	0.9861	0.9770	0.9848	0.9766	0.9935	0.9875
Magic	0.9013	0.9003	0.8957	**0.9152**	0.9014	0.9013	0.9003	0.9009	0.8847	0.8742	0.9018	0.8568	0.9013	0.9007
Phoneme	0.9709	0.9697	0.9729	0.9643	0.9704	0.9751	**0.9753**	0.9742	0.8655	0.8439	0.9546	0.8094	0.9063	0.9411
Phon. sh/ʃ/	0.9030	0.8750	0.9060	0.9068	0.9066	0.9087	0.8993	**0.9258**	0.8654	0.8221	0.9056	0.6880	0.8897	0.8973
Phoneme s	0.8660	0.8645	0.8653	0.8992	0.8690	0.8911	0.8725	0.8668	**0.9045**	0.8547	0.8722	0.7604	0.8674	0.8685
Phon. zh/ʒ/	0.9818	0.9761	0.9692	0.9939	0.9639	0.9535	**0.9951**	0.9836	0.9724	0.8077	0.9373	0.7172	0.9843	0.9819
Phoneme z	0.9607	0.9385	0.9105	0.9628	0.9065	0.9388	0.9291	0.9447	0.9434	0.9685	0.9582	**0.9776**	0.9695	0.9683
Pima	0.7024	0.7723	**0.8462**	0.8205	0.7736	0.8233	0.8004	0.8045	0.6993	0.6809	0.6948	0.6611	0.6964	0.7017
Satimage	0.8743	**0.8918**	0.8743	0.8890	0.8826	0.8731	0.8823	0.8791	0.8832	0.8712	0.6809	0.6522	0.8828	0.8647
XOR2	0.9995	0.9082	0.9983	0.9996	0.9989	0.9993	0.9997	**0.9998**	0.9996	0.9981	0.9996	0.9985	0.9995	0.9995
XOR32	0.9916	0.9681	**0.9990**	0.9988	0.9902	0.9925	0.9988	0.9678	0.9980	0.9734	0.9988	0.9928	0.9904	0.9974
XOR8	0.9904	0.9832	0.9851	0.9951	0.9944	0.9904	0.9848	0.9946	**0.9993**	0.9938	0.9944	0.9828	0.9963	0.9989
Yeast	0.7716	0.7787	0.7484	**0.8092**	0.7855	0.7746	0.7704	0.7792	0.7506	0.7912	0.7696	0.5635	0.7629	0.7723

U cilju lakšeg tumačenja rezultata, predstavljamo primer vrednosti koeficijenta korelacije između $\sigma(D)$ data u obliku matrice (Tabela 7.3) i AUC (KNN) matrice (Tabela 7.6). Koeficijent čija je pozicija u matrici (1, 5) ima vrednost: R($\sigma(D)$,AUC(KNN)) = -0,5990.
Svi koeficijenti korelacije R imaju negativne vrednosti, a za sve p vrednosti smatra se da je p <0,05.

Ova činjenica potvrđuje da postoje statistički značajne korelacije između direktne i indirektne mere balansa uzoraka, što je stav hipoteze P3 koja tvrdi da direktne mere balansa raspodele instanci različitih kvaliteta artikulacije u prostoru obeležja obučavajućeg uzorka (Entropija i Standardna devijacija) stoje u jakoj pozitivnoj korelaciji sa indirektnim merama njegove reprezentativnosti (stepen tačnosti predikcije involviranih prediktora).

Dakle, sa porastom standardne devijacije lokalnih rastojanja instanci $\sigma(D)$, koji je indikator smanjenja entropije uzorka, dolazi do smanjenja AUC vrednosti, što znači do smanjenja tačnosti klasifikacije. Potvrda hipoteze P3, koja je bitna hipoteza ovog istraživanja, zahtevala je dosta truda za dokazivanje. Rezultati pokazuju da je korelacija između direktnih i indirektnih mera balansa nešto veća u slučaju KNN klasifikatora nego u slučaju MLP klasifikatora. Ova razlika je verovatno posledica stohastičkog karaktera MLP klasifikatora odnosno većeg determinizma KNN klasifikatora, ali istovremeno i pokzatelj konkurentnosti KNN klasifikatora sa najboljim tipovima klasifikatora.

Tabela 7.6

Rezultati klasifikacije (AUC) za originalne i balansirane baze podataka za KNN klasifikator.

Data Set	ORIG	over-sampled data							under-sampled data					
		ADAS	BOR SMT	DBB	ROS	SMT	SMT+ ENN	SMT+ TOM	ENN	MOST DIST	NCL	NEAR MISS2	RUS	TOM
Abalone	0.7259	**0.9734**	0.9478	0.9698	0.8974	0.9507	0.9719	0.9724	0.6932	0.6825	0.7435	0.7518	0.9660	0.7603
Arrhythmia	0.5832	0.7246	0.7283	**0.7375**	0.6399	0.7181	0.7235	0.7293	0.5731	0.5376	0.5829	0.4915	0.6194	0.5731
Br. Cancer	0.9688	0.9522	0.9653	0.9652	0.9726	0.9769	0.9758	0.9758	0.9735	0.9749	0.9728	0.9686	**0.9813**	0.9775
EEG Eye St	**0.9830**	0.9676	0.9623	0.9757	0.9788	0.9775	0.9657	0.9702	0.9703	0.9567	0.9704	0.9321	0.9807	0.9777
E. Coli	0.9159	0.8844	0.8752	0.8965	0.9043	0.9000	0.8837	0.8837	0.9161	0.8263	0.9172	0.8898	0.9000	**0.9185**
Heart	0.6531	0.6229	0.6400	0.6702	0.6460	0.6442	0.6702	**0.6823**	0.5807	0.5559	0.6414	0.5914	0.5880	0.6254
Iris	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Letter A	0.9906	0.9887	**0.9950**	0.9949	0.9918	0.9926	0.9905	0.9895	0.9924	0.7031	0.9949	0.9197	0.9894	0.9949
Magic	0.8081	0.7801	0.7834	0.7959	0.8131	0.8062	0.7983	0.8010	0.7752	0.7451	**0.8173**	0.7136	0.8037	0.8008
Phoneme	0.8361	0.8748	0.8576	**0.8776**	0.8739	0.8701	0.8508	0.8585	0.7240	0.6325	0.8489	0.6515	0.7900	0.8378
Phon. sh/ʃ/	0.9483	0.9641	0.9479	0.9677	**0.9679**	0.9468	0.9272	0.9569	0.8213	0.6569	0.9359	0.6918	0.8685	0.9380
Phoneme s	0.8904	0.8941	0.8526	0.8916	**0.9032**	0.8826	0.8732	0.8967	0.8660	0.7922	0.8971	0.7930	0.8915	0.8924
Phon. zh/ʒ/	0.7929	**0.8054**	0.6964	0.7964	0.7981	0.7939	0.7682	0.7828	0.7064	0.6965	0.7800	0.6191	0.7673	0.8005
Phoneme z	0.9901	0.9723	0.9144	0.9796	0.9937	0.8896	0.9673	**0.9958**	0.9462	0.7640	0.9724	0.6982	0.9154	0.9725
Pima	0.7150	0.6975	0.7448	0.7535	0.7241	0.7277	0.7373	0.7260	0.6986	0.6451	**0.7560**	0.6313	0.7394	0.7290
Satimage	0.8743	0.9166	0.8888	0.9039	**0.9265**	0.9220	0.9142	0.9177	0.8290	0.5876	0.8765	0.6407	0.9095	0.8951
XOR2	0.9984	**0.9997**	0.9918	0.9993	0.9993	0.9993	0.9993	0.9993	0.9993	0.9924	0.9993	0.9983	0.9993	0.9993
XOR32	0.9722	0.9717	0.9708	0.9738	0.9850	0.9717	0.9713	0.9717	0.9583	0.9363	0.9721	0.9717	**0.9977**	0.9722
XOR8	0.9917	0.9953	0.9983	0.9988	0.9983	0.9994	**0.9994**	0.9983	0.9917	0.9489	0.9917	0.9843	0.9971	0.9917
Yeast	0.7310	0.7277	0.7531	**0.7616**	0.7329	0.7515	0.7526	0.7541	0.7023	0.7256	0.7532	0.6170	0.7347	0.7448

Tabela 7.7

Pearsonovi korelacioni parametri između direktnih mera [$\sigma(D), H(D)$ i $\sigma H(D)$] i indirektnih mera disbalansa (AUC) koje su dobijene pomoću tri različita klasifikatora (MLP Ansaml, KNN Ansaml i hibridni MLP+KNN klasifikator).

Sve matrice (20x14) su convertovane u vektor kolone (280x1) pre računanja korelacije između njih.												
$corr(x,y), x \in \{\sigma, H, \sigma H\}$, $y \in \{\text{AUC(MLP)}, \text{AUC(KNN)}, \text{AUC(MLP + KNN)}\}$	AUC(MLP)				AUC(KNN)				AUC(MLP) + AUC(KNN)			
	R	*p-value*	R_L	R_U	*R*	*p-value*	R_L	R_U	*R*	*p-value*	R_L	R_U
$\sigma(D)$	-0.4334	< .0001	-0.5240	-0.3331	-0.5990	< .0001	-0.6693	-0.5182	-0.5737	< .0001	-0.6474	-0.4893
$H(D)$	-0.3623	< .0001	-0.4599	-0.2559	-0.3342	< .0001	-0.4344	-0.2259	-0.3720	< .0001	-0.4688	-0.2664
$\sigma H(D)$	-0.4361	< .0001	-0.5264	-0.3360	-0.5950	< .0001	-0.6658	-0.5136	-0.5722	< .0001	-0.6461	-0.4877

Eksperimenti su izvedeni preko identičnog skupa podataka obuke u slučaju oba klasifikatora, kako bi se izbegao stohastički uticaj razlike uzoraka za obuku na performanse klasifikatora. Takođe, postoji nešto manji stepen korelacije između H(D) i AUC, što je rezultat aproksimacije u izračunavanju entropije uzorka na osnovu histograma s jedne strane, i nesrazmernog odnosa gustina instanci i broja obeležja uzoraka sa druge strane. Za komparativni proračun korelacije između rangova direktnih i indirektnih metrika koristili smo Spearmanov korelacioni koeficijent rangova ρ (*ro*) kao meru statističke zavisnosti između rangiranja ove dve varijable. Tabela 7.8, koja je organizovana analogno Tabeli 10, prikazuje ρ i *p*-vrednosti za iste grupe podataka. I u ovom slučaju potvrđena je postojanje negativna korelacija između aktuelnih mera kao i statistička značajnost te korelacije.

Visok stepen podudarnost rezultata dve različite statističke metode dodatno učvršćuje stav o potvrdi tačnosti hipoteze P3.

Posebnu pažnju posvetićemo razultatima iz kolone 4 u Tabelama 7.5 i 7.6, koji se odnose na naš algoritam. Tabela performansi fiktivnog hibridnog klasifikatora MLP + KNN nije potrebno prikazivati. Izmerene vrednosti AUC dobijene ansamblom MLP imaju nešto veću srednju

vrednost od KNN ansambla, što ukazuje na prednosti neuronskih mreža usled veće fleksibilnosti.

Tabela 7.8
Spearmanovi korelacioni parametri ρ (rho) između direktnih mera $[\sigma(D), H(D)$ i $\sigma H(D)]$ i indirektnih mera disbalansa (AUC) koje su dobijene pomoću tri različita klasifikatora (MLP Ansaml, KNN Ansaml i hibridni MLP+KNN klasifikator).

Sve matrice (20x14) su convertovane u vektor kolone (280x1) pre računanja korelacije između njih.						
$corr(x, y), x \in \{\text{Rank}(\sigma, H, \sigma H)\}$, $y \in \{\text{Rank}[\text{AUC(MLP)}, \text{AUC(KNN)}, \text{AUC(MLP + KNN)}]\}$	Rank AUC(MLP)		Rank AUC(KNN)		Rank AUC(MLP)+AUC(KNN)	
	ρ	*p-value*	ρ	*p-value*	ρ	*p-value*
Rank $\sigma(D)$	-0.5538	< .0001	-0.6125	< .0001	-0.6020	< .0001
Rank $H(D)$	-0.4574	< .0001	-0.3812	< .0001	-0.4047	< .0001
Rank $\sigma H(D)$	-0.5766	< .0001	-0.6137	< .0001	-0.6098	< .0001

Na osnovu podataka iz tabela 7.6 i 7.7, neophodno je dokazati statistički značajno poboljšanje performansi klasifikatora pod uticajem DBB algoritma (kolona 4 u Tabelama 7.6 i 7.7) u odnosu na uticaj drugih algoritama (ostale kolone). Da bismo dokazali ovaj stav, potreban nam je statistički test. U tu svrhu koristimo test poznat pod nazivom two-tailed Wilcoxon test, za analizu statističke značajnosti razlika ostvarenih prosečnih vrednosti performansi. Rezultati ove analize prikazani su u tabelama 7.9 i 7.10. Prva kolona Tabele 7.9 pokazuje ocenjeni ukupni rang algoritama za balansiranje koji se odnosi na efekat balansiranja. Druga kolona sadrži nazive algoritama, a ostale kolone iz Tabele 7.9 predstavljaju srednje vrednosti (*mean*) i standardnu devijaciju (*sd*) performansi (AUC) svih klasifikatora koji su podvrgnuti obučavajućim uzorcima svih baza podataka generisanih primenom svih algoritama za balansiranje. Rangiranje ocenjenih uticaja algoritama balansranja na performanse klasifikatore (MLP, KNN i MLP + KNN) su vrlo konzistentne, jer se na prvih pet pozicija nalazi isti skup algoritama, uključujući DBB, sa neznatnom razlikom u rasporedu. Treba napomenuti da svi ovi algoritmi spadaju u kategoriju metoda oversamplinga, što je u skladu sa rezultatima prikazanim u radovima Japkovicz i Stephen, 2002 i Batista i sar., 2004. Z i *p*-vrednosti za Wilcoxon test su date u Tabeli 7.10. Prikazani rezultati potvrđuju našu pretpostavku, pokazujući statistički značajnu razliku ($p < 0.05$) između uticaja DBB i drugih algoritama za balansiranje. Detaljniji uvid u Tabelu 7.10 pokazuje da se veoma visok stepen značajnosti razlike odnosi na mere performanse MLP klasifikatora, kao i klasifikator MLP + KNN, dok je ovaj stepen značajnosti razlike nešto niži u KNN Klasifikator. Konkretno, u slučaju KNN klasifikatora, u četiri od trinaest slučajeva (SMT + TOM, ROS, SMT i TOM) balansnih algoritama, nema značajne razlike ($p > 0.05$) uticaja u odnosu na DBB algoritam (bold-face vrednosti u Tabeli 7.10). U ostalih 9 slučajeva postoji značajna razlika ($p < 0.05$) uticaja. Dobijeni rezultati u potpunosti potvrđuju našu pretpostavku statistički značajnog poboljšanja ($p < 0.05$) performansi MLP i MLP + KNN klasifikatora pod uticajem DBB

algoritma, u odnosu na uticaj drugih algoritama. U slučaju performanse klasifikatora KNN, u četiri od 13 slučajeva, ova razlika nije bila statistički značajna, te zato smatramo da MLP ansambl ima prednost nad KNN ansamblom.

Tabela 7.9
Srednja vrednost tačnosti klasifikacije AUC dobijena primenom različitih technika balansuranja i klasifikatora.

Total rank	Method of balancing	MLP		Method of balancing	KNN		Method of balancing	MLP + KNN	
		mean AUC	*sd* AUC		*mean* AUC	*sd* AUC		*mean* AUC	*sd* AUC
1^0	DBB	0.9407	0.0638	DBB	0.8955	0.1055	DBB	0.9181	0.0799
2^0	SMT+TOM	0.9330	0.0696	SMT+TOM	0.8931	0.1086	SMT+TOM	0.9131	0.0837
3^0	SMT+ENN	0.9312	0.0730	ROS	0.8873	0.1208	SMT+ENN	0.9091	0.0848
4^0	SMT	0.9253	0.0716	SMT+ENN	0.8870	0.1085	SMT	0.9057	0.0864
5^0	ROS	0.9235	0.0755	SMT	0.8860	0.1096	ROS	0.9054	0.0926
6^0	ADAS	0.9232	0.0738	BOR SMT	0.8857	0.1176	BOR SMT	0.9020	0.0876
7^0	TOM	0.9212	0.0825	ADAS	0.8757	0.1137	ADAS	0.8994	0.0870
8^0	ORIG	0.9197	0.0829	RUS	0.8719	0.1294	TOM	0.8956	0.0963
9^0	BOR SMT	0.9184	0.0715	NCL	0.8712	0.1260	RUS	0.8950	0.0977
10^0	RUS	0.9181	0.0832	TOM	0.8701	0.1303	ORIG	0.8941	0.0971
11^0	NCL	0.9091	0.0972	ORIG	0.8685	0.1304	NCL	0.8901	0.0969
12^0	ENN	0.9037	0.0890	ENN	0.8359	0.1448	ENN	0.8698	0.1102
13^0	MOST DIST	0.8974	0.0907	NEARMISS2	0.7778	0.1645	MOST DIST	0.8327	0.1124
14^0	NEARMISS2	0.8399	0.1441	MOST DIST	0.7680	0.1527	NEARMISS2	0.8088	0.1462

Tabela 7.10
Vilkoksonov test rangiranja za ocenu ranga za procenu razlike između uticaja DBB-a i drugih algoritama na postignute performanse različitih klasifikatora. (Interval povjerenja = 95%)

Total rank	Method of balancing	MLP		Method of balancing	KNN		Method of balancing	MLP + KNN	
		p-value	*Z-value*		*p-value*	*Z-value*		*p-value*	*Z-value*
1^0	DBB	-	-	DBB	-	-	DBB	-	-
2^0	SMT+ENN	0.0156	2.4170	SMT+TOM	**0.3719**	2.1934	SMT+TOM	0.0269	2.2135
3^0	SMT+TOM	0.0283	2.1934	ROS	**0.7605**	3.1795	SMT+ENN	0.0011	3.2596
4^0	ROS	0.0015	3.1795	SMT	**0.0582**	3.2405	ROS	0.0222	2.2867
5^0	TOM	0.0016	3.1593	SMT+ENN	0.0113	2.4170	SMT	0.0004	3.5413
6^0	SMT	0.0012	3.2405	BOR SMT	0.0149	3.2596	BOR SMT	0.0004	3.5413
7^0	RUS	0.0022	3.0584	RUS	**0.0707**	3.0584	ADAS	0.0004	3.5413
8^0	ORIG	0.0006	3.4206	TOM	0.0394	3.1593	RUS	0.0011	3.2596
9^0	ADAS	0.0029	2.9779	NCL	0.0442	3.4794	TOM	0.0005	3.5011
10^0	NCL	0.0005	3.4794	ORIG	0.0421	3.4206	ORIG	0.0004	3.5413
11^0	BOR SMT	0.0011	3.2596	ADAS	0.0002	2.9779	NCL	0.0007	3.3752
12^0	ENN	0.0012	3.2445	ENN	0.0010	3.2445	ENN	0.0003	3.5929
13^0	MOST DIST	0.0005	3.4813	NEARMISS2	0.0002	3.4612	MOST DIST	0.0002	3.7425
14^0	NEARMISS2	0.0005	3.4612	MOST DIST	0.0002	3.4813	NEARMISS2	0.0002	3.7425

Naši eksperimenti pokazuju da oversampling metode balansiranju uzoraka u kombinaciji sa DBB algoritmom, daju vrlo dobre rezultate na uzorcima različitih veličina, kompleksnosti i stepena disbalansa. Rezultati ukazuju na to da je opravdano istraživanje uticaja entropije obučavajućih uzoraka na performanse involviranog klasifikatora. U tom smislu, balansiranje zasnovano na indirektnom pristupu promeni entropije uzoraka, prikazana u ovom radu, može biti zanimljiv predmet daljih istraživanja.

7.5.1 Doprinos, prednosti i nedostaci DBB metode

Glavni teorijski rezultat predstavljen u ovom radu predstavlja uspostavljena formalnu korespodencija između srednjih lokalnih rastojanja između instanci regularne *nD* rešetke s

jedne strane i volumena koje zauzimaju ove instance s druge strane. Ovaj formalizam može imati značaj u karakterizaciji raspodele uzoraka velikih dimenzija.

Ova metoda nudi efikasan indirektan pristup balansiranju velikih baza uzoraka, što smanjuje vreme računanja i korišćenje računarskih resursa. U prilog ovoj tvrdnji, prikažimo stratifikaciju instanci u prostoru obeležja.

Direktna stratifikacija ima smisla u slučaju malog broja (n) relevantnih obeležja uzorka, ali u realnim situacijama n uzima velike vrijednosti i ova činjenica predstavlja prepreku operativnim algoritmima, zbog ogromnog broja (h^n) podprostora dobijenih segmentacijom svih karakterističnih varijabli u h segmenata. Svaki od ovih podprostora treba pretraživati i resemplovati, iako mali dio njih sadrži instance koje treba uzeti obraditi. Razmotrimo skup C_e, koji se sastoji N_e instanci određenih sa $n = 100$ obeležja. Zamislimo podelu ovih n koordinata na h = 100 segmenata. Ovom operacijom delimo originalni skup C_e na $h^n = 1.0\text{e}200$ novih podprostora koje treba analizirati, što podrazumeva ogromnu računsku složenost $O(h^n)$. Još jedan nedostatak stratifikacije u prostoru obeležja je nemogućnost vizuelne prezentacije u prostoru (n> 3). Razmotrimo indirektni DBB pristup problemu lokalne gustine. Za određivanje d_{ie} vrednosti koristimo $2D$ matricu $d_{2e} = \{d_{ik} | d_{ik} = d(c_{ie}, c_{ke})\}$, $i = k = 1, 2, \dots, N_e$., gde je N_e=$|C_e|$ i $|d_{2e}| = |N_e \times N_e| = N_e^2$.

Vrednost složenosti $O(N_e^2)$ indirektnog pristupa je za više redova of veličine manji veći od vrednosti direktnog pristupa $O(h^n)$. Predstavljeni selektivni stratifikovani oversampling, kao deo DBB procedure, menja početnu neravnomernu raspodelu uzorka ka uniformnoj distribuciji, povećavajući njenu reprezentativnost, dok SMOTE algoritam, na primjer, vrši interpolaciju instanci na čitavom prostoru instanci, održavjući na taj način iniocijalno stanje raspodele.

Metoda ROS povećava redundantnost i količinu podataka bez povećanja njihove informativnosti. Još jedan deo DBB algoritma, selektivni stratifikovanoi undersampling, ima očiglednu prednost u poređenju sa slučajnim undersamplingom. Ova procedura na celom prostoru obeležja uspostavlja kvazi uniformnu raspodelu uklanjajući redundantne instance iz gustih područja prostora, čime povećava reprezentativnost uzorka.

S druge strane, RUS metod vrši slučajno uklanjanje instanci po celom prostoru obeležja što rezultira zadržavanjem početnog stanja reprezentativnosti i uklanjanjem važnih instanci u oblastima male gustine. Zbog brzog rasta količine raspoloživih informacija, potreba za otkrivanjem i smanjenjem redundanse postojećih skupova podataka postaje sve više aktuelna što otvara prostor za primenu DBB algoritma. Uzimajući u obzir ove činjenice, možemo zaključiti da predloženi DBB algoritam ima prihvatljive karakteristike u poređenju sa aktuelnim algoritmima balansiranja podataka. Treba napomenuti da prezentacija *pdf* krivih

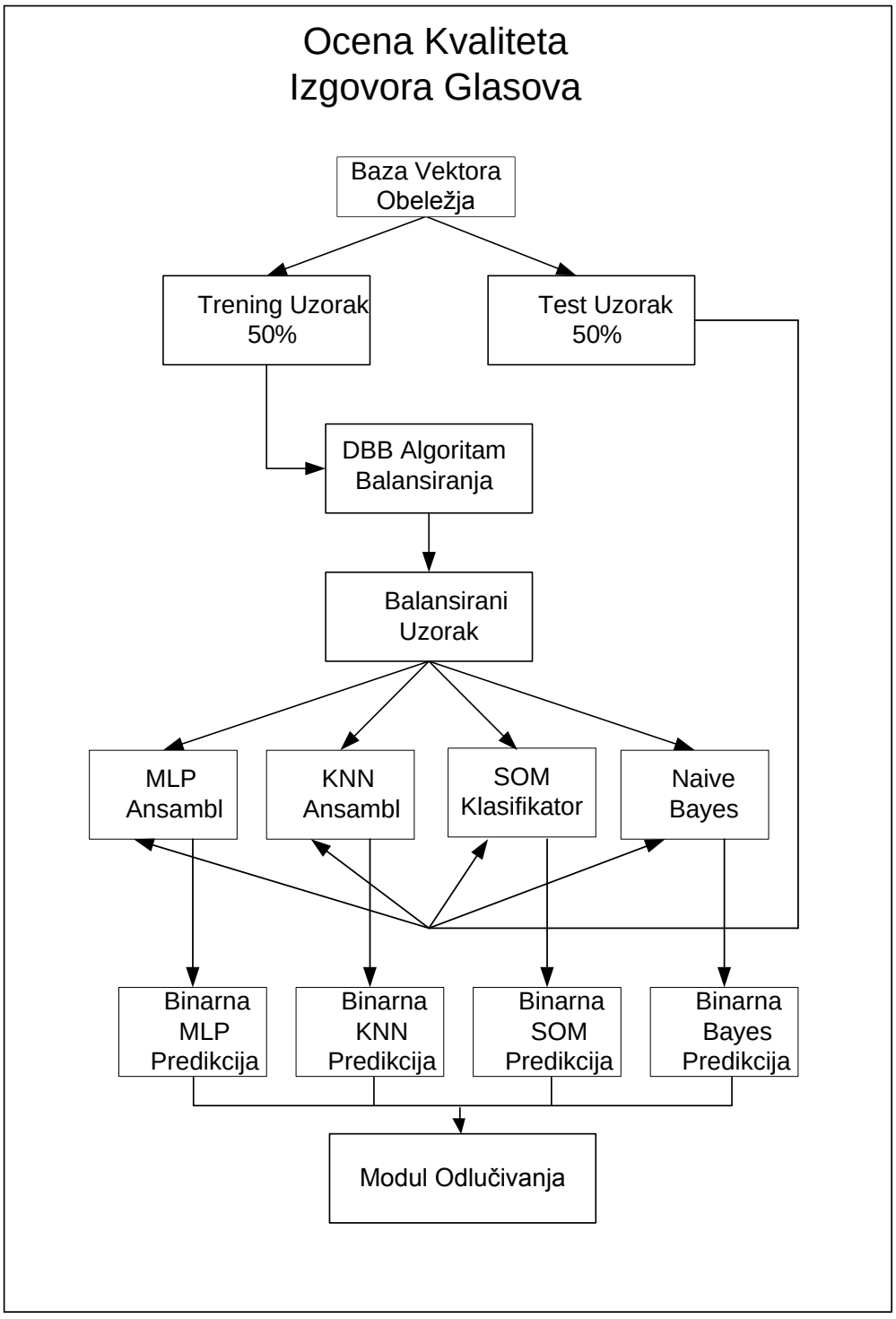
koje se odnose na srednje lokalne udaljenosti predstavlja ilustrativan metod za dobijanje indirektnog unutrašnjeg uvida u raspodelu instanci u *n*-dimenzionalnom prostoru uzorka. Ovaj metod daje vizuelnu prezentaciju uticaja različitih tehnika ponovnog uzorkovanja na raspodelu instanci u prostoru obeležja. Efikasnost DBB metode prevazilazi efikasnost standardnih metoda za balansiranje podataka jer zadržava svoje prednosti, dok istovremeno eliminiše njihove značajne nedostatke. Konkretno, ova metoda tokom obrade ne uklanja važne instance kao što je RUS metoda, ne generiše nove redundantne instance u gustim oblastima uzorka prostora kao što su ROS i SMOTE metode, uklanja postojeće redundantne instance gustih područja i generiše nove instance u retkim oblastima prostora obeležja. Ova studija jasno otkriva i prikazuje složenu korespondenciju između distribucionih karakteristika uzoraka i performansi klasifikatora obučenih na istim podacima. Problemi koji se mogu javiti ovim algoritmom tokom faze oversamplinga u prostoru niske gustine odnose se na količinu sintetičkih instanci koja može biti prevelika u slučaju previše male vrednosti θ (videti pododeljak 5.3.6 i 5.3.7). Ovu činjenicu treba imati na umu pri radu na PC mašinama sa ograničenom radnom memorijom.

7.5.1.1 Ostale mogućnosti za primenu algoritma

Učeći modeli podležu uticaju nepravilnosti u raspodeli prediktivnih varijabli, bilo u slučaju klasifikacije ili regresije, jer kao modeli upravljani podacima (*data driven models*) skloni su favorizovanju dominantnih struktura relativno visokih gustina (klastera) u prostoru obeležja prediktivnih varijabli. Osnovna namera DBB algoritma je eliminisanje takvih struktura koje predstavljaju unutrašnji disbalans. Ovaj algoritam se može primeniti i za regularizaciju ponašanja induktivnih prediktora u slučaju regresije, tj. aproksimacije funkcija, identifikacije i predikcije procesa. Nagle varijacije ili ekstremne amplitude vrednosti izlaznog signala često su u korelaciji sa visokim vrednostima signala greške predicije, zahvaljujući sklonosti prediktora da zanemaruju retke događaje kao irelevantne. DBB algoritam eliminiše takve pojave. Ovaj algoritam se takođe može uspešno primeniti za klasteriranje podataka podešavanjem praga θ na izabrane niže vrednosti, nakon čega se uklanjaju sve instance koje su veće od θ i izostavlja procedura selektivnog stratifikovanog undersamplinga prikazanog u pododeljku 5.3.6. Osetljivost algoritma na strukture različitih gustina, prikazane na Slici 7.1, garantuje svoju fleksibilnost u klasterovanju podataka. Algoritam može poslužiti za efikasno otkrivanje i uklanjanje inkompatibilnih instanci iz baza podataka. Prikazani algoritam može poslužiti za indirektnu detekciju i vizualizaciju klastera instanci u multidimenzionalnom prostoru, kao i smanjenje dimenzija uzoraka na osnovu mogućnosti jednostavnog otkrivanja i uklanjanja redundantnih instanci.

7.6. Rezultati ocena kvaliteta artikulacije primenom nekoliko modela

U prethodnim poglavljima i potpoglavljima prikazani su alati i rezultati procedura koje predstavljaju važnu pripremu za najvažniji deo istraživanja, odnosno, određivanja optimalnog računarskog modela logopedskog postupka za ocenu kvaliteta artikulacije glasova srpskog jezika. Od četiri različita modela, zasnovana na „pattern recognition“ pristupu, koji su podvrgnuti komparativnoj analizi u ovom potpoglavlju, biće odabran optimalni model. Opšti prikaz blok dijagrama algoritma za određivanje optimalnog modela za ocenu kvaliteta artikulacije dat je na Slici 7.2. Koristeći dva obučavajuća skupa različitih reprezentativnosti (50 % originalnog uzorka i njegov balansirani pandan) imamo priliku da iskoristimo prednosti DBB algoritma u cilju povećanja pouzdanosti predikcije korišćenih klasifikatora. Moduli za obuku su detaljno prikazani u šestom poglavlju. Modul odlučivanja generiše niz binarnih vrednosti ocene kvaliteta artikulacije fonema pri čemu je 0 rezervisana za tipične artikulacije a vrednost 1 predstavlja atipičan izgovor. Krajnji rezultat istraživanja je određivanje optimalnog računarskog modela procesa logopedske ocene kvaliteta artikulacije glasova srpskog jezika. Za optimalni model biće proglašen onaj model koji pokaže najveću pouzdanost predikcije kvaliteta izgovora glasova u skladu sa srednjom logopedskom ocenom (Tabela 7.11) i definisanim kriterijumima rangiranja prikazanim u Tabeli 7.12. Kao što smo rekli u poglavlju o kvalitetu artikulacije, naša objektivna ocena kvaliteta po GAT testu predstavlja usrednjenu ocenu grupe od pet (5) iskusnih logopeda koja se za svaki od analiziranih glasova predstavlja u obliku binarnog vektora kolone sa vrednostima 1 za svaki atipičan izgovor i 0 za tipične izgovore bez značajnijih odstupanja. Na ovaj način postavljen je polazni idealni etalon tačnih ocena kvaliteta artikulacije koji po pretpostavci ima maksimalnu AUC vrednost (1) za svaki fonem, maksimalnu srednju AUC vrednost (Mean AUC =1) i minimalnu vrednost standardne devijacije (Std AUC=0). Pouzdanost ovog vrlo važnog koraka pri formiranju računarskog ekspertskog modela ima veliki značaj za postizanje krajneg cilja - pouzdanog modela. Rezultati o tačnosti različitih prediktora za ocenu kvaliteta izgovora fonema s, š, z i ž u odnosu na etalon (logopedski konsenzus) prikazani su na Tabeli 7.11. u formi AUC vrednosti. Podaci o ukupnom broju pozitivnih (1) i negativnih (0) primera dati su u tebeli 7.1. Odnos trening i test skupa bio je rigorozan (50% : 50%) a rezultati prikazani u tabeli se odnose na test skup. Tabela 7.11 sadrži komparativni prikaz rezultata prediktora obučavanih na dva trening uzorka: 50% nebalansiranog originala i uzorak dobijen balansiranjem istog trening uzorka primenom DBB algoritma dizajniranog tokom istraživanja (Furundžić i sar. 2017 b).



Slika 7.2 Algoritam determinacije optimalnog modela za ocenu kvaliteta artikulacije.

Kao test skup za obe kategorije obučavanja prediktora uzet je ostatak od 50% originala. Treba napomenuti da se Optimalni MLP ansambl određuje na osnovu algoritma iz šestog poglavlja (Furundzic i sar. 2012a) dok je Običan usrednjeni MLP ansambl određen prostim usrednjavanjem odgovora svih članova ansambla. AUC vrednosti etalona su definisane samo za originalni uzorak. AUC vrednosti za balansirani skup su preslikane sa originala jer

balansirani skup predstavlja skup balansiranih obeležja koji nema mogućnost logopedske auditivne inspekcije ali implicitno odgovara originalu.

Tabela 7.11 Tačnost predikcije na test uzorku grupe prediktora (AUC) obučavanih na nebalansiranim (originali) i balansiranim (DBB) trening uzorcima.

	O B U K A											
Logopedi i Prediktori	ORIGINALNI TRENING UZORAK (50%)						(DBB) BALANSIRANI TRENING ORIGINAL					
	T E S T											
	ORIGINALNI TEST UZORAK (50%)						ORIGINALNI TEST UZORAK (50%)					
	S	Š	Z	Ž	Mean (AUC)	Std (AUC)	S	Š	Z	Ž	Mean (AUC)	Std (AUC)
Etalon-Logopedi Većinska odluka	1	1	1	1	1	0	1	1	1	1	1	0
kNN Ansambl	0.8904	0.9483	0.9901	0.7929	0.9054	0.0854	0.8916	0.9677	0.9796	0.7964	0.9071	0.0787
Naive **B**ayes	0.7983	0.8291	0.8453	0.7550	0.8069	**0.0397**	0.8831	0.9405	0.9391	0.9670	0.8697	0.0756
SOM	0.6939	0.5988	0.6524	0.6124	0.6394	0.0429	0.7275	0.7510	0.7704	0.7650	0.6964	0.0683
Optimalni odabrani MLP Ansambl	0.8660	0.9030	0.9607	**0.9818**	**0.9279**	0.0530	0.8992	0.9068	0.9628	**0.9939**	**0.9343**	0.0462
Običan usrednjeni MLP Ansambl	0.8560	0.9134	0.9507	0.8114	0.8829	0.0615	0.8739	0.8851	0.9481	0.9632	0.9176	**0.0446**

Svaki test primerak ima apriori pridruženu ocenu nastalu većinskom odlukom grupe logopeda, što za rezultat ima vektor binarnih vrednosti ocena koji se pridružuje celom test uzorku. Sa druge strane svaki prediktor generiše odgovarajući vektor binarnih vrednosti koje pridružuje istom test uzorku. Koristeći konfuzionu matricu (Tabela 7.2) i prateće jednačine metrike za procenu performansi klasifikatora dobili smo AUC vrednosti prikazane u Tabeli 7.11 za svaki fonem i svaki prediktor kao meru podudarnosti sa etalonskim vrednostima. Dakle, veća AUC vrednost odgovara većem stepenu tačnosti predikcije, odnosno boljim preformansama prediktora. U Tabeli 7.11 boldovane su vrednosti koje se odnose na maksimalne vrednosti predikcije (AUC). Naročito su važne srednje vrednosti Mean (AUC) kao opšta mera performansi prediktora, i po toj meri MLP Ansambl ima maksimalne vrednosti kako u domenu modela obučenih na nebalansiranom trening uzorku tako i u domenu modela obučanih na balansiranom uzorku.

Dodatne informacije za izbor optimalnog modela date su u Tabeli 7.12, gde su izvorne AUC vrednosti zamenjene njihovom pozicijom (rang) na skali postojećih vrednosti sortiranih u opadajućem nizu. Dakle, rang prediktora sa najvećom AUC vrednošću će biti 1 a rang prediktora sa najnižom vrednošću će biti 5. Rangiranje je obavljeno za obe grupe trening uzoraka po svim fonemima i po svim srednjim vredostima za foneme. Vrednosti standardne devijacije ukazuju na robustnost modela i prema Tabeli 7.11 MLP strukture sa balansiranim podatacima imaju najmanje vrednosti devijacije što odgovara najvećoj robustnosti modela.

Tabela 7.12 Rangiranje prediktora po efikasnosti u funkciji srtednjih AUC vrednosti.

Prediktori	RANG PREDIKTORA U FUNKCIJI AUC, ORIGINALNI TEST UZORCI FONEMA							RANG PREDIKTORA U FUNKCIJI AUC, (DBB) BALANSIRANI TEST UZORCI FONEMA						
	S	Š	Z	Ž	Mean (AUC)	SUM	RANG	S	Š	Z	Ž	Mean (AUC)	SUM	RANG
kNN Ansambl	**1**	**1**	**1**	3	2	8	**1**	**1**	**1**	**1**	4	3	10	2
Naive **B**ayes	4	4	4	4	4	20	4	3	2	4	2	4	13	3
SOM	5	5	5	5	5	25	5	5	5	5	5	5	25	5
Optimalni odabrani MLP Ansambl	2	3	2	**1**	**1**	9	2	2	3	2	**1**	**1**	9	**1**
Običan usrednjeni MLP Ansambl	3	2	3	2	3	13	3	4	4	3	3	2	16	4

Sumiranjem vrednosti rangova izračunavamo ukupni rang za sve prediktore. Shodno ovim vrednostima zaključujemo da MLP ansambl ostvaruje najveće srednje AUC vrednosti za oba slučaja trening uzoraka (Tabela 7.11). Takođe MLP ansambl u Tabeli rangova zauzima prvo mesto u slučaju balansiranog trening uzorka, dok KNN prediktor ima primat u slučaju nebalansiranog trening uzorka. Ovi rezultati prikazuju kako prednosti MLP ansambla, tako i visoku komparabilnost KNN prediktora u odnosu na proverene fleksibilne klasifikatore kakve su neuronske mreže. Ostali prediktori retko ostvaruju kompetitzivan rezultat što je bilo i očekivano, s obzirom na fleksibilnost KNN i MLP modela pri formiranju graničnih struktura u prostoru obeležja instanci.

Ovim rezultatima smo potzvrdili pretpostavku P4 koja predviđa najbolje performanse MLP ansambla obučenog na balansiranom skupu.

7.7. Rezultati komparacije ocena kvaliteta artikulacije logopeda i modela

Prethodno postavljeni idealni etalon tačnih ocena kvaliteta artikulacije (većinska odluka pet logopeda), biće korišćen u svrhu poređenja tačnosti prediktora i svakog od pet logopeda pojedinačno. Pošto pouzdanost ovog vrlo važnog koraka pri formiranju računarskog ekspertskog modela ima veliki značaj za postizanje krajneg cilja, treba istaći činjenicu da Inter Logopedska korelacija ocena kvaliteta izgovora fonema izražena u formi korelacije iznosi 0.80 do 0.90 kada su u pitanju logopedi sa preko 20 godina iskustva. Za mlađe logopede ova mera ide maksimalno do 0.70. U radu Schipor i sar., 2012, autori tvrde da interekspertska korelacija ocena artikulacije iznosi 0.76, što se dosta slaže sa podacima iskusnih logopeda uključenih u ovo istraživenje.

Rezultati utvrđenja nepostojanja značajnosti razlike u tačnosti različitih prediktora za ocenu kvaliteta izgovora fonema s, š, z i ž u odnosu na tačnost logopeda (Pretpostavka 2), prikazani su na Tabeli 7.12. Podsetimo da su podaci o ukupnom broju pozitivnih (1) i negativnih (0) primera dati su u tebeli 7.1. Svi podaci o trening i test skupovima ostaju isti kao i u prethodnom potpoglavlju. Pošto je utvrđen etalon tačnih ocena (vektor $O(L_m)$) kao konsenzus svih pet logopeda, jedini način za poređenje tačnosti vektora ocena prediktora {*O(KNN), O(NB), O(SOM) i O(MLP)*} i vektora ocena svih pet logopeda pojedinačno $\{O(L_1), O(L_2), O(L_3), O(L_4), O(L_5)\}$ je posredna komparacija sličnosti njihovih vektora ocena sa etalonom $O(L_m)$.

U tu svrhu izračunavamo Pearsonove parametre korelacije. Tabela 7.13 pokazuje koeficijente korelacije R, *p*-vrednosti kao i vrednosti R_L i R_U. R koeficijenti mogu da se kreću od -1 do 1, pri čemu -1 predstavlja direktnu, negativnu korelaciju, 0 predstavlja nepostojanje korelacije, a 1 predstavlja direktnu, pozitivnu korelaciju. Predstavljene *p*-vrednosti testiraju nultu hipotezu da ne postoji veza između posmatranih metrika. Prikazane *p*-vrednosti se kreću od 0 do 1, pri čemu vrednosti blizu 0 odgovaraju statistički značajnoj korelaciji u R. Ako je *p*-vrednost manja od usvojenog nivoa značajnosti (0,05), onda se odgovarajuća korelacija smatra značajnom. Prikazane vrednosti R_L su 95% od intervala pouzdanosti donje granice za aktuelne koeficijente R, dok su vrednosti R_U gornje granice za aktuelne koeficijente R.

Na osnovu rezultata Pirsonovih koeficijanata korelacije R i ostalih parametara iz tabele 7.13 zaključujemo da svi prediktori, izuzev SOM, generišu vektore ocena koji imaju veći stepen korelacije sa etalonom nego što imaju vektori ocena svih logopeda pojedinačno.

Ovi rezultati dokazuju stav pretpostavke P2 u kojoj se tvredi da ne postoji značajna razlika u tačnosti ocene kvaliteta artikulacije između logopeda i izabranog algoritma za automatsku ocenu kvaliteta artikulacije.

Ovi rezultati pored toga što potvrđuju najvažniju premisu ostvarivosti računarskog modela za ocenu artikulacije glasova, potvrđuju i potrebu i opravdanost njgovog dizajniranja.

Na osnovu rezultata prikazanih u Tabelama 7.11 – 7.13 posredno zaključujemo da je ispunjena prva premisa ostvarivosti aktuelnog modela P1 koja tvrdi da višedimenzionalni prostor artikulaciono-akustičkih atributa izgovornog glasa omogućava pouzdanu distinkciju između njegovih tipičnih i atipičnih realizacija.

Vektori obeležja, detaljno prikazanih u potpoglavljima 4.4.2 i 7.2, koji karakterišu kvalitet artikulacije glasova služe kao glavni izvor informacija za diskriminaciju njihove tipične i atipične produkcije. U interakciji sa vrednostima obeležja u prostoru obeležja, adaptivni

prediktori generišu fleksibilnu graničnu hiperpovrš koja razdvaja ove dve klase kvaliteta, što je prikazano u rezultatima.

Tabela 7.13
Pearsonovi parametri korelacije srednjih logopedskih ocena $\{O(L_m)\}$ sa pojedinačnim logopedskim ocenama $\{O(L_i), i = 1,2..,5\}$, i ocenama pet induktivnih prediktora, obučenih na balansiranim uzorcima i celom skupu analiziranih fonema.

Vektori srednjih $O(L_m)$ i pojedinačnih $\{O(L_i), i = 1,2..,5\}$ Logopedskih ocena za foneme s, š, z i ž su konvertovane u ukupni vektor kolonu dimenzije (484x1)* pre računanja korelacija. * (50% primera fonema iz tabele 7.1)									
$corr(x,y), x = \{O(L_m)\}, y \in \begin{Bmatrix} O(L_1), O(L_2), \\ O(L_3), O(L_4), O(L_5) \end{Bmatrix}$	LOGOPEDI				$corr(x,y), x \in \{L_m\}, y \in \begin{Bmatrix} \mathrm{O(KNN), O(NB), O(SOM)} \\ \mathrm{O(MLP_{opt.}), O(MLP_{sred.})} \end{Bmatrix}$	PREDIKTORI			
	R	*p-value*	R_L	R_U		*R*	*p-value*	R_L	R_U
$corr(O(L_m), O(L_1))$	0.7700	<0.0001	0.7068	0.8211	$corr(O(L_m), \mathrm{O(KNN)})$	0.9300	<0.0001	0.9085	0.9466
$corr(O(L_m), O(L_2))$	0.8100	<0.0001	0.7563	0.8529	$corr(O(L_m), \mathrm{O(NB)})$	0.8500	<0.0001	0.8064	0.8845
$corr(O(L_m), O(L_3))$	0.8300	<0.0001	0.7813	0.8687	$corr(O(L_m), \mathrm{O(SOM)})$	0.6805	<0.0001	0.5983	0.7486
$corr(O(L_m), O(L_4))$	0.8500	<0.0001	0.8064	0.8845	$corr(O(L_m), \mathrm{O(MLP_{opt.})})$	0.9400	<0.0001	0.9214	0.9543
$corr(O(L_m), O(L_5))$	0.8215	<0.0001	0.7706	0.8620	$corr(O(L_m), \mathrm{O(MLP_{sred.})})$	0.9200	<0.0001	0.8956	0.9389

Pošto primenjeni skup obeležja obezbeđuje pouzdanu ocenu kvaliteta artikulacije, uporedivu i čak superiornu u odnosu na pouzdanost ocene logopeda, zaključujemo da odabrani relevantni skup obeležja zadovoljava prethodno postavljeni zahtev iz pretpostavke P1. Ovaj stav ne znači da je odabran optimalni mogući skup obeležja sa obzirom na diverzitet i varijacije inherentnih akustičkih manifestacija artikulacije glasova srpskog jezika.

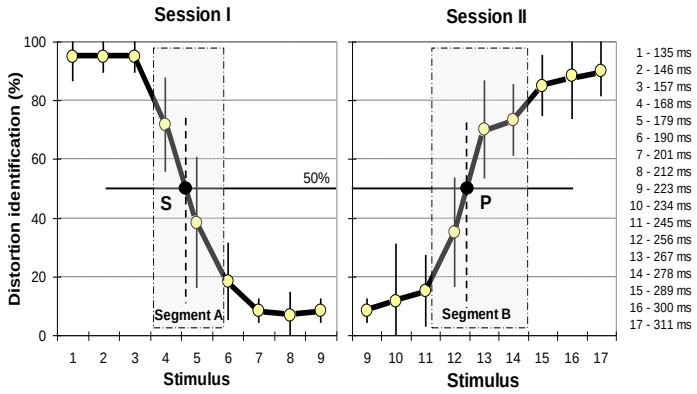
Ovim rezultatima smo potvrdili ostvarivost i opravdanost projektaa predloženog računarskog modela logopedskog procesa za ocenu kvaliteta artikulacije, zasnovanog na standardnoj logopedskoj proceduri datoj u formi GAT testa, prihvaćenog u logopedskoj praksi.

7.8. Primeri primene senzitivnosti neuronskih mreža za ocenu prirode uticaja karakterističnih ulaznih varijabli na izlaz.

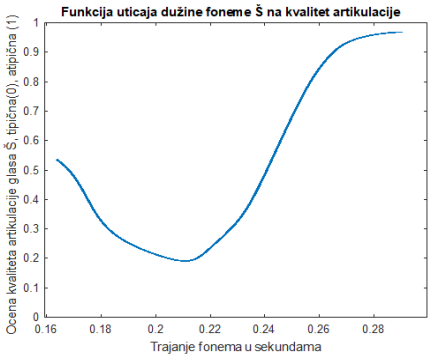
Eksperimentalni rezultati logopedske definicije identifikacione funkcije za prepoznavanje tipičnog/atipičnog trajanja frikativa š (Punišić i sar. 2017) prikazani su u formi grafa na Slici 7.3. Grupi treniranih eksperata je prezentovan signal fonema š koji je bio podvrgnut sintetičkoj rekonstrukciji trajanja u smislu smanjenja i produženja. Na slici 7.3 je prikazana usrednjena logopedska ocena odstupanja kvaliteta artikulacije fonema š. Na x osi je prikazano trajanje foneme u funkciji broja stimulus odsečaka identične dužine od 10 ms, a na *y* osi je prikazan stepen distorzije izgovora foneme u procentima, što znači da 100% predstavlja atipičan izgovor po tipu distorzije a 0% se odnosi na fonem izgovoren u tipičnoj realizaciji-bez distorzije. Tako je empirijski dobijena funkcionalna zavisnost stepena percipirane distorzije od Trajanja fonema.

Sa druge strane, u skladu sa *'clamping tecnique'* ispitivana je senzitivnost mreže obučena na realnim (nesintetičkim) vrednostima foneme š koje kao što se vidi na slici 7.4 nemaju ekstremne vrednosti trajanja karakteristične za graf 7.3. Na x osi su vrednosti trajanja fonema u

sekundama (s) tako da vrednost 0.16 na x osi predstavlja 160 *ms*, dok je na *y* osi presdstavljena vrednost ocene u rasponu od 0 do 1 gde 1 predstavlja atipičan izgovor po tipu distorzije. Ova vrednost (1) je potpuno ekvivalentna vrednosti 100% sa slike 7.3Obučena NM je podvrgnuta fiksiranju na srednju vrednost svih 19 ulaznih variabli za ocenu kvaliteta artikulacije izuzev variable trajanja fonem (*nf*) koja je sortirana u rastućem nizu od minimuma do maksimuma i tako prezentirana NM. Odziva obučene NM na tako transformisane ulaze prikazan je na Slici 7.4. Očigledna je podudarnost dva prikazana grafa od kojih je jedan empirijski a drugi je rezultat modeliranja. NM je obučavan na realnim vrednostima vektora obelažja kvaliteta izgovora glasa š i zato nisu korišteni sintetički ulazi iz eksperimenta 7.3. U domenu realnih dužina postoji vrlo dobro poklapanje prikazanih grafova što ukazuje na potencijal NM za analizu kompleksnih funkcionalnih veza između odabranih obeležja i odgovora NM. Ovaj graf predstavlja svojevrstan interni uvid u strukturu modeliranog procesa. Ovakve analize nisu rađene u oblasti analize kvaliteta artikulacije govora pa zato mogu imati značaj za tu oblast.

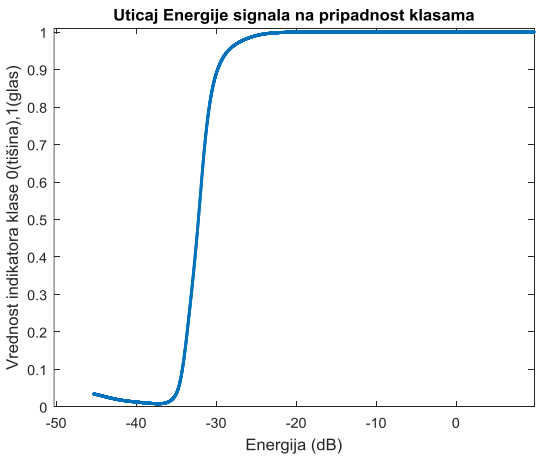


Slika 7.3. Identifikacione funkcije za prepoznavanje foneme na kvalitet tipičnog/atipičnog trajanja frikativa š.(Punišić i sar. 2017)

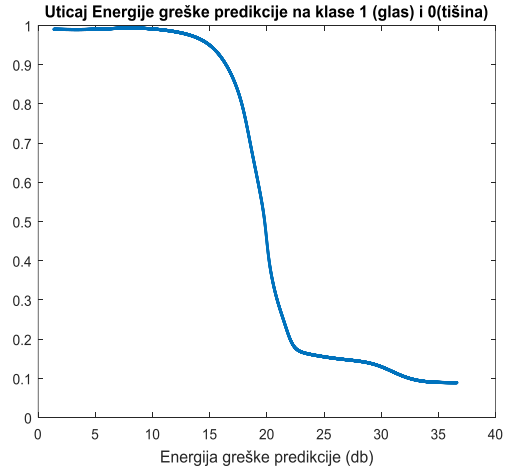


Slika 7.4. Funcija uticaja dužine artikulacije dobijena primenom MLP ansambla.

Pored ovog primera, na slikama 7.5, 7.6 i 7.7 su prikazane funkcije uticaja karajkterističnih obeležja na ponašanje VAD detektora pri ekstrakciji aktivnog govora iz kontinuiranog signala. U ovim slučajevima se radi o uticajima variabli čiji je efekat pri VAD ekstrakciji poznat (Energija signala, Energija greške predikcije i vrednosti prvog LPC koeficijenta). Nagla promena nagiba krive energije na slici 7.5 govori o njenom naglašenom značaju za distinkciju govornog signala i signala tišine što je potpuno u skladu sa praksom u analizi govora generalno.

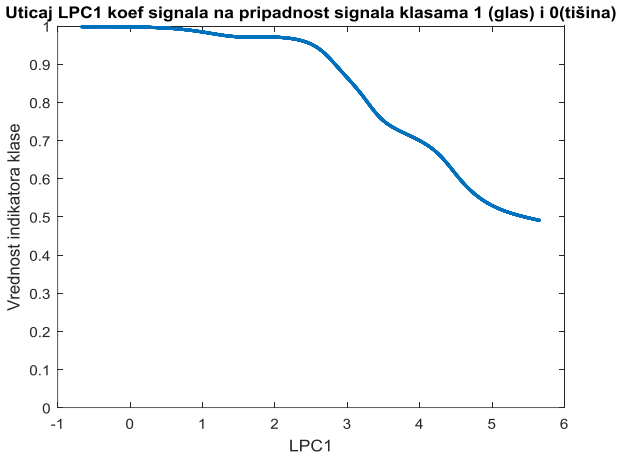


Slika 7.5 Funkcija uticaja energije na distinkciju na distinkciju aktivnog govora i signala tišine (VAD).



Slika 7.6 Funkcija greške linearne predikcije na distinkciju aktivnog govora i signala tišine (VAD)

Preostale dve krive su potpuno u skladu sa postojećim iskustvom u VAD detekciji i generalno u oblasti govornog signala. Ovi primeri ukazuju na mogućnost modeliranja vrlo važnih kauzalnih veza u oblasti logopedije i bolje razumevanje složenih procesa artikulacije a naročito procesa auditivne percepcije koja je manje istražena oblast.



Slika 7.7 Funkcija uticaja prvog LPC koeficijenta aktivnog govora i signala tišine (VAD).

8. ZAKLJUČAK

Na osnovu utvrđenih metodoloških koraka za ocenu kvaliteta artikulacije glasova srpskog jezika, u ovom istraživanju je detaljno predstavljen dizajnirani računarski model procesa logopedske evaluacije kvaliteta artikulacije glasova srpskog jezika, zasnovan na učećim modelima za prepoznavanja oblika.

Primetan pad kvaliteta artikulacije se ne retko dovodi u vezu sa naglim tehnološkim razvojem u oblasti telekomunikacija, interneta i mobilne telefonije. Favorizacija neverbalne i indirektne komunikacije na štetu direktne govorne komunikacije bitno utiče na pad kvaliteta artikulacije glasova i kvaliteta govora generalno. Pravilan razvoj govora predstavlja neophodan uslov zdravog razvoja i kvalita svih aspekata života jedinke. U logopediji već postoje standardne metode ocene kvaliteta artikulacije, zasnovane na komparativnoj auditivno-perceptivnoj analizi karakteristika govora u tipičnoj i atipičnoj realizaciji.

Naglašen subjektivistički karakter, nedovoljna efikasnost i pouzdanost kao i drugi nedostaci metoda za procenu kvaliteta artikulacije glasova srpskog jezika, zasnovanih na tradicionalnoj iskustvenoj evaluaciji logopeda, bili su motiv za simplifikaciju i objektivizaciju tog procesa prikazanu u okviru ovog istraživanja. Uočavanjem nedostataka postojećih rešenja, došlo se do ideje o povećanju dimenzija i diverziteta komponenti vektora obeležja kvaliteta artikulacije i primeni modernijih pouzdanih alata za njihovu klasifikaciju, u cilju kreiranja pouzdanog, objektivnog i efikasnog računarskog modela sa visokim stepenom autonomije kojim se unapređuje postojeći način ocene kvaliteta artikulacije. Generalna pretpostavka ostvarivosti i svrasishodnosti dizajniranog modela morala je da zadovolji nekoliko preduslova i prođe proveru kroz proces dokazivanja početnih premisa.

Ove pretpostavke impliciraju potebu pronalaženja skupa relevantnih atributa izgovornih glasova i fleksibilnih klasifikatora koji će na osnovu tih atributa oceniti kvalitet artikulacije produkovanih glasova sa tačnošću na nivou logopeda.

Ostvarivost predloženog objektivnog modela za ocenu kvaliteta izgovora glasova srpskog jezika uslovljena je potvrdom valjanosti još nekih premisa a realizacija modela zahtevala je istraživanje i korišćenje raspoloživih relevantnih znanja i alata sa jedne strane i kreiranje neophodnih originalnih rešenja sa druge strane. Od primarnog značaja za prikazano rešenje postavljenog problema bilo je razumevanje procesa artikulacije glasova kao složenog mehanizma zasnovanog na naučenim paradigmama koordinisane aktivacije organa govornog aparata i čula sluha.

Sledeća, vrlo značajna premisa rešenja definisanog zadatka, predstavlja razumevanje mehanizma naučene percepcije kvaliteta izgovorenog sadražaja, od strane logopeda, pre svega u cilju identifikacije relevantnih akustičkih manifestacija izgovorenih fonema, odnosno skupa obelažja na osnovu kojih logoped procenjuje kvalitet izgovora. Ova obeležja su u velikoj meri perceptabilna, tehnički detektabilna i merljiva, a mogu se predstaviti u numeričkoj ili simboličkoj formi, što nam je omogućilo dizajn računarskog modela procesa logopedske evaluacije kvaliteta artikuilacije glasova. Sa druge strane veliki broj ovih obeležja predstavlja praktičnu prepreku za uključenje svih obeležja u analizu pa je zato korišten odabrani skup relevantnih obeležja visoke informativnosti i separabilnosti.

Značajan momenat istraživanja se odnosio na vrlo aktuelni problem informativnosti i reprezentativnosti raspoloživog i obučavajućeg skupa instanci koji direktno utiče na performanse korišćenih učećih modela pa je zato pažljivo izveden postupak izbora ulaznih variabli i balansiranja uzoraka u cilju poboljšanja performansi klasifikatora.

Sledeći istraživački korak se odnosio na izbor tipa i strukture predloženog matematičkog modela prediktora koji treba na osnovu raspoložive ograničene baze podataka da uspostavi prihvatljiv algoritamski model korespondencije između vektora obeležja i analiziranog kvaliteta artikulacije fonema. Kroz komparativnu analizu performansi nekoliko predloženih tipova i struktura modela izbrali smo optimalni model- MLP ansambl. Pri definiciji optimalnog MLP ansambla korišćen je originalni algoritam za izbor manjeg podskupa dobro obučenih MLP struktura iz kompletnog velikog skupa.

Ostvarenje postavljenog cilja zahtevalo je sledeće, specifične metodološke korake:

- Priprema baze stimulusa kontinualnog govora ispitanika u skladu sa GAT testom;
- Određivanje skupa akustičkih indikatora za razgraničenje signala aktivnog govora i signala tišine za obuku VAD detektora;
- Ekstrakcija govornih stimulusa reči iz baze stimulusa kontinualnog govora primenom VAD detektora i formiranje baze stimulusa reči;
- Određivanje skupa akustičkih obelžja za diskriminaciju segmenata različitih fonema iz govornih stimulusa reči;
- Ekstrakcija govornih stimulusa fonema segmentacijom signala reči iz baze stimulusa reči;
- Određivanje skupa akustičkih obelžja za ocenu kvaliteta artikulacije glasova srpskog jezika;
- Formiranje baze vektora izabranih obeležja za foneme iz baze segmenata i prateće baze ekspertskih ocena kvaliteta artikulacije za svaki fonem.
- Analiza i povećanje nivoa reprezentativnosti izabranog obučavajućeg uzorka primenom nove DBB metode;

- Uspostavljanje formalnih modela algoritamske korespondencije (učeći prediktori) između vektora akustičkih mera i numeričkih indikatora, klasa različitog kvaliteta artikulacije, apriori ocenjenih od strane tima logopeda;
- Izbor optimalnog modela (MLP ansambl) kroz komparativnu analizu performansi nekoliko predloženih tipova i struktura modela.

Potvrđena je valjanost osnovnih pretpostavki (P1,...,P4) kroz sprovedene adekvatne statističke testove i putem analize rezultata.

Tokom istraživanja je formirana govorna baza stimulusa u formi uzorka adekvatne veličine i reprezentativnosti u skladu sa definisanim kriterijumima izbora ispitanika. Baza je korišćena za ekstrakciju relevantnih obeležja fonema, proveru relevantnosti pojedinih parametara pri segmentaciji fonemskih i subfonemskih struktura analiziranih reči, trening, validaciju i verifikaciju performansi klasifikatora, kao i komparaciju rezličitih tipova i arhitektura klasifikatora. Istraživanje uključuje dve grupe ispitanika: kontrolna sa korektnim izgovorom i eksperimentalna sa različitim vrstama i nivoima odstupanja u izgovoru glasova. Pripadnici obe kategorije su podeljene na obučavajući uzorak i test uzorak približnih kardinalnih vrednosti. Pored dve grupe govornika, čiji izgovor će poslužiti za formiranje govorne baze, predviđene je i grupa od pet (5) treniranih slušalaca - eksperata, koja će poslužiti za objektivizaciju logopedskog procesa ocene kvaliteta artikulacije i, samim tim, objektivizaciju računarskog modela procesa, što implicira poboljšanje performansi našeg algoritma. Ova grupa je generisala etalon ocena kvaliteta artilulacije izabranih fonema na osnovu konsenzusa pri odlučivanju. U istraživanju su korišćeni adekvatni standardni testovi za ocenu kvaliteta artikulacije primenjeni u logopediji i relevantno ekspertsko znanje.

8.1. Pregled rezultata

Glavni rezultat, kao krajnji cilj ovog istraživanja, je računarski model procesa logopedske evaluacije kvaliteta artikulacije glasova srpskog jezika, zasnovan na učećim „data driven“ modelima za prepoznavanje oblika.

Ostali rezultati dobijeni tokom ovog istraživanja su prikazani u skladu sa redosledom sprovedenih metodoloških koraka.

Dizajn VAD modela za ekstrakciju signala aktivnog govora iz kontinuiranog govornog signala GAT testa, sproveden je kroz pristup prepoznavanja oblika primenom učećih prediktora od kojih KNN i MLP ansambl potpuno zadovoljavaju u smislu postignute tačnosti izraženoj u funkciji greške koja se kreće između 6% i 7 % za MLP i 7% i 10% za KNN prediktor.

Korištena akustička obeležja za ekstrakciju aktivnog govora, izabrana u skladu sa aktuelnim metodama, a čije su raspodele i karakteristike detaljno analizirane, potpuno ispunjavaju očekivane zahteve u smislu diskriminacionog potencijala.

Informativnost obeleža koja karakterišu različite klase kvaliteta artikulacije glasova uslovljena je separabilnošću njihovih histograma, odnosno funkcija gustine verovatnoće. Stepen preklapanja krivih gustine verovatnoće obeležja različitih klasa kvaliteta stoji u negativnoj korelaciji sa njihovom separabilnošću, odnosno informativnošću. Ovaj kriterijum je bio jedan od važnijih pri izboru vektora obeležja za različite probleme klasifikacije tokom istraživanja (VAD, Segmantacija, Ocena artikulacije) pa se njegova upotreba može proširiti bez ograničenja u domenu klasifikacije uzoraka.

Među izabranim obeležjima najveći uticaj na VAD predikciju imaju energija govornog signala i energija signala greške predikcije, a zatim vrednost broja promena znaka signala. Izvesno je da povećanje baze trening uzoraka prati povećenje tačnosti prediktora pa u tom smislu možemo prihvatiti prikazane modele.

Izbor modela za računarsku segmentaciju reči izdvojenih putem VAD-a sprovedena je takođe pomoću učećih prediktora za prepoznavanja oblika od kojih KNN i MLP ansambl prednjače i potpuno zadovoljavaju u smislu postignute tačnosti koja je ovde izražena u funkciji greške predikcije a koja se kreće između 5% i 7 % za MLP i 5% i 10% za KNN prediktor. Skup akustičkih obeležja za segmentaciju signala reči na foneme, po uzoru na savremene metode segmentacije uključuju i MFCC parametre tako da ovaj skup potpuno ispunjavaju očekivane zahteve u smislu diskriminacionog potencijala. Među ovim obeležjima najveći uticaj na predikciju granica segmenata fonema energija govornog sigmala, energija signala greške predikcije i prvih nekoliko MFCC koeficijenata.

Dizajniran je novi algoritam za povećenje reprezentativnosti trening uzorka zasnovan na balansiranju lokalnih rastojanja susednih instanci (DBB), uklanjanjem primeraka iz oblasti velike gustine raspodele i sintetičkom generacijom primeraka u oblastima sa malom gustinom raspodele. Novi DBB algoritam obezbeđuje povećanje entropije uzorka koje se manifestuje povećanjem njegove reprezentativnosti. U cilju dokaza valjanosti novog algoritma izvedena su opsežna testiranja na 20 standardnih baza podataka, uključujući i podatke iz oblasti artikulacije glasova, uz komparativnu analizu 13 različitih algoritama za balansiranje.

U svrhu rangiranja performansi ovog algoritma sproveden je Vilkoksonov statistički test koji ovaj algoritam pozicionira u klasu boljih algoritama. Ovaj algoritam se može smatrati doprinosom nastalim tokom aktuelnog istraživanja. Primenom ovog algoritma dokazana je uzročno posledična veza između direktnih mera reprezentativnosti uzoraka (entropija i

varijansa) i indirektnih mera koje se manifestuju tačnošću klasifikatora pri klasifikaciji tih uzoraka. Ovim rezultatom je potvrđena treća bitna pretpostavka ostvarivosti aktuelnog računarskog modela.

Definicija modela za ocenu kvaliteta artikulacije glasova sprovedena je takođe primenom učećih modela za prepoznavanja oblika gde je Optimizirani MLP ansambl rangiran kao najbojli dok su KNN anbsambl i NB prediktor rangirani na mesta 2 i 3. Na četvrtom mestu po rangu nalazi sa SOM prediktor. Tačnost predikcije MLP ansambla se kreće između 81% i 97% sa srednjom vrednišću od 93.43%. Tačnost predikcije KNN ansambla se kreće između 79% i 99% sa srednjom vrednišću od 90.71%. Srednja tačnost predikcije za NB i SOM prediktore iznosila je 86.97% i 69.64% respektivno. Skup akustičkih obeležja fonema za ocenu kvaliteta artikulacije uključuje vrednosti energije foneme, energije reči koja u inicijalnoj poziciji sadrži aktuelni fonem, 12 MFCC vrednosti, podatke o trajanju foneme i odgovarajuće reči i podatak o relativnom odnosu dužina foneme i date reči. Ovaj skup obeležja je odabran u skladu sa iskustvom logopeda ali se potpuno uklapa u aktuelni trend. Među ovim obeležjima najveći uticaj na predikciju kvaliteta artikulacije fonema energija govornog signala fonema, energija signala prateće reči i prvih 2 MFCC koeficijenata.

Prikazana je komparacija tačnosti računarskog modela za ocenu kvaliteta glasova u odnosu na tačnost ocene pet logopeda pojedinačno. Ova komparacije je izvedena posredno, preko etalona ocena koji je definisan konsenzusom pet iskusnih logopeda. Rezultat te komparacije govori o očiglednoj superiornosti računarskog modela u odnosu na logopede.

Ovim rezultatom je potvrđena druga (najvažnija pretpostavka istraživanja) a za njenu potvrdu korišten se Pirsonov statistički test korelacije. Na ovaj način posredno je potvrđena i prvaa polazna pretpostavka ovog istraživanja. Ovaj metod komparacije je razvijen tokom istraživanja i može se smatrati doprinosom.

Prikazani su primeri analize senzitivnosti NM u cilju modeliranja vrlo važnih kauzalnih veza u oblasti logopedije i bolje razumevanje složenih procesa artikulacije a naročito procesa auditivne percepcije koja je manje istražena oblast. Ovakvi modeli do sada nisu korišćeni u logopediji i to predstavlja novi pristup u domenu ocene kvaliteta artikulacije.

Tokom istraživanja ocenjana je valjanost polaznih pretpostavki P1, P2, P3 i P4 o ostvarivosti i opravdanosti projekcije sistema za ocenu kvaliteta izgovora glasova. Iz razloga preglednopsti ove pretpostavke su ponovo navedene uz prikaz rezultata njihove potvrde.

P_1 „Višedimenzionalni prostor artikulaciono-akustičkih atributa izgovornog glasa omogućava pouzdanu distinkciju između njegovih tipičnih i atipičnih realizacija“. Rezultati potvrđuju pretpostavku P_1.

P_2 „Ne postoji značajna razlika u tačnosti ocene kvaliteta artikulacije između logopeda i izabranog algoritma za automatsku ocenu kvaliteta artikulacije. Rezultati potvrđuju pretpostavku P_2.

P_3 „Direktne mere balansa raspodele instanci različitih kvaliteta artikulacije u prostoru obeležja obučavajućeg uzorka (Entropija i Standardna devijacija) stoje u jakoj pozitivnoj korelaciji sa indirektnim merama njegove reprezentativnosti (Tačnost predikcije involviranih prediktora)“. Rezultati potvrđuju pretpostavku P_3.

P_4 „Od planiranih modela za klasifikaciju višedimenzionih vektora akustičkih obelžja, najbolje performanse se očekuju od ansambla višeslojnih perceptrona sa balansiranom obukom.“ Rezultati potvrđuju pretpostavku P_4.

8.2. Doprinos disertacije

Glavni doprinos ovog istraživanja, kao krajnji cilj ovog istraživanja, je računarski model procesa logopedske evaluacije kvaliteta artikulacije glasova srpskog jezika, zasnovan na „data driven“ inteligentnim modelima za prepoznavanje oblika.

Za ostvarenje glavnog cilja, neophodno je bilo izabrati skup artikulaciono-akustičkih atributa izgovornih glasova srpskog jezika visoke informativnosti i definisati strukturu fleksibilnih klasifikatora koji na osnovu tih atributa generišu ocenu kvaliteta artikulacije glasova sa tačnošću na nivou logopeda. Dizajnirani klasifikatori predstavljaju uprošćeni računarski model usklađen sa logopedskom procedurom određivanja ocene kvaliteta artikulacije glasova. Na osnovu prikazanih rezultata i predstavljenih alata, otvara se mogućnost za korišćenje istih alata za modele opštije namene u logopediji.

Tokom istraživanje kreirane su nove metode i algoritmi koji se mogu koristiti kako u obradi govornog signala tako i u domenu klasifikacije i prepoznavanja uzoraka generalno.

Prvi doprinos je nova metoda, razvijena tokom istraživanja, je „Distance Based Balancing“ (Balansiranje instanci obučavajućeg uzorka zasnovano na lokalnim rastojanjima instanci). Ovo fleksibilna metoda za detekciju i eliminaciju neravnoteže u bazama podataka zasnovana na povećanju reprezentativnosti uzoraka kroz povećanu uniformnost distribucije njegovih instanci u prostora obeležja uzoraka, odnosno na maksimizaciji entropije uzorka. Algoritam kombinuje kontrolisani „oversampling“ i „undersampling“ u različitim oblastima prostora obeležja uzorka, zavisno od gustine instanci u ovim oblastima. Tačnije rečeno DBB algoritam za balansiranje obavlja sintetičko generasanje novih instanci u područjima niskih gustina instanci u prostoru obeležja, praćeno uklanjanjem instanci iz oblasti visoke gustine, što rezultuje povećanjem

uniformnosti distribucije uzorka po celom prostoru obeležja instanci. Ovim postupkom uklawaju se redundantni primeri, a primeri od najvećeg informativnog značaja su sačuvani i podržani od strane novih sličnih sintetičkih primera. Kao mera gustine instanci u oblasti poslužilo je srednje lokalno rastojanje svih instanci datog uzorka od predefinisanog skupa njihovih suseda. Ovaj algoritam predstavlja transfer distributivnih karakteristika pravilne rešetke na realne baze podataka i na taj način menja njihovu strukturu u pravcu maksimalne reprezentativnosti.

Drugi doprinos je novi primenjeni algoritam za izbor optimalnog ansambla MLP čije su prednosti primene detaljno prikazani u disertaciji. On se odlikuje originalnim načinom izbora članova optimalnog ansambla obučenih MLP struktura, koji mu obezbeđuje bolje performanse u odnosu na običan ansambl koji uključuje ceo skup obučenih MLP. Primena ovog algoritma nije ograničana samo na aktuelni problem već se bez ograničenja može koristiti u oblasti klasifikacije uzoraka. Ovaj algoritam takođe služi za balansiranje obučavajućih uzoraka.

Treći doprinos predstavlja algoritam za posrednu komparaciju tačnosti ocenjivanja kvaliteta artikulacije između obučenih prediktora i pojedinih logopeda, gde se kao posredni etalon koriste vrednosti ocena donesene većinskim odlučivanjem grupe iskusnih logopeda. Cilj ovakvog pristupa je da se na osnovu većunska odluka grupe logopeda o uzorku ograničene dužine, obuči fleksibilni klasifikator koji će kroz povećenje baze znanja povećavati tačnost, robustnost i sposobnost generalizacije do momenta njegove optimalne strukture, kada se može koristiti kao etalon za ocenui kvaaliteta artikulacije.

Sledeći doprinos predstavlja pristup problemu ocene kvaliteta artikulacije zasnovan na analizi osetljivosti obučenih NM na perturbacije ulaznih variabli, koji nije do sada korišćen u ovoj oblasti.

8.3. Mogućnosti za dalje istraživanje

Prikazani DBB algoritam se može primeniti i za regularizaciju ponašanja induktivnih prediktora u slučaju regresije, tj. aproksimacije funkcija, identifikacije i predikcije procesa kao i klasterovanja velikih grupa podtaka u cilju smanjenja postojeće redundanse. U situaciji enormnog porasta informacija u realnom okruženju, sve viša će biti cena filtriranja relevantnih informacija što je osnovna funkcija DBB algoritma. Govrni signal se odlikuje velikim dimenzijama, pa u njegovoj obradi ovakvi algoritmi mogu imati primenu. Prikazani novi pristup problemu ocene kvaliteta artikulacije zasnovan na analizi osetljivosti obučenih NM na perturbacije ulaznih variabli u buduće bi mogao biti interesantan.

LITERATURA

Armani, L.; Matassoni, M.; Omologo, M.; Svaizer, P. (2003). Use of a CSP-based voice activity detector for distant-talking ASR, *Proc. EUROSPEECH 2003*, Geneva, Switzerland, pp. 501–504.

Aslam, J., Ppopaa, R., and Rivest, R., 2007. On estimating the size of a statistical audit. In Proc. 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT'07).

Atal, B., Rabiner L., 1976. A pattern recognition approach to voiced unvoiced silence classification with applications to speech recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, 24 (3) (1976), pp. 201-212

Atal, B., and Hanuer, S., 1971. Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Amer., vol. 50, pp. 637-655, Aug. 1971.

Barandela, R, Valdovinos, RM, Sánchez, J.S., 2013. New applications of ensembles of classifiers. Pattern Anal Appl 6(3):245–256.

Bilibajkić, R., Subotić, M., Furundžić, D., 2014b. Primena neuralnih mreža u detekciji patološkog izgovora srpskih glasova. ZBORNIK RADOVA XXII TELEKOMUNIKACIONI FORUM TELFOR 2014, Izdavači: Društvo za telekomunikacije -Beograd, Akademska misao - Beograd, 25-27 Novembar, Beograd, Srbija. ISBN: 978-1-4799-6190-0, pp 873-876.

Bilibajkić, R., Šarić, Z., Jovičić, S., Punišić, S., Subotić, M., 2016. Automatic detection of stridence in speech using the auditory model. Computer Speech and Language, Volume 36, pp.122–135.

Breiman, L., 1996. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

Breiman, L., 1998. Arcing classifiers (with discussion). Ann. Statist. 26, 801–849.

Breiman, L., 1994. Heuristics of instability in model selection, Technical Report, Statistics Department, University of California at Berkeley (to appear, Annals of Statistics).

Breiman, L., 1996a. Bagging predictors. Mach Learn 24(2):123–140.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 36(1):105–139.

Basbug, F.; Swaminathan, K.; Nandkumar, S. (2004). Noise reduction and echo cancellation front-end for speech codecs, IEEE Trans. Speech Audio Processing, vol. 11, 1, pp. 1–13.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, vol. 6(1), pp.20–29.

Bertino, S., 2006. A measure of representativeness of a sample for inferential purposes. International Statistical Review, 74, pp. 149-159.

Blake, C., and Merz, C., UCI Repository, 1998. Irvine, CA: University of California, School of Information and Computer Science.

Bouquin-Jeannes, R.L.; Faucon, G. (1995). Study of a voice activity detector and its influence on a noise reduction system, Speech Communication, vol. 16, pp. 245–254.

Buhlmann, P., and Yu, Bin., 2000. Explaining bagging. Technical Report 92, ETH Zurich, Seminar Fur Statistik, May 2000Carbonell, J., and Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, pp. 335-6.

Cardie, C., and Howe, N.,1997. Improving minority class predication using case specific feature weights. In Proceedings of the fourteenth International Conference on Machine Learning, pages 57–65, Nashville, TN, July 1997.

Carney, J., and Cunningham, P., 1999. Tuning diversity in bagged neural network ensembles. Technical Report TCD-CS-1999-44, Trinity College Dublin, 1999.

Chang and Fallside, An adaptive training algorithm for bp networks," Computer Speech and Language, pp. 205-218, 1987.

Cesar, M. E., Hugo, R. L., 2000. Acoustic Analysis of Speech for Detection of Laryngeal Pathologies, Proc. 22nд Annual EMBS .lnt. Conf., pp. 2369-2372 July 2000.

Charles, R., Norman, L., Strominger,Robert J. Demarest, and David.Ruggiero, The Human Nervous System Structure and Function, 6th ed, Humana Press Inc, New Jersey, 2005.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling TEchnique, Journal of Artificial Intelligence Research 16, pp. 321–357.

Chawla, N.V., Japkowicz, N., and Kolcz, A., 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 1-6.

Chawla, N., Moore, Thomas., Bowyer, K., Hall, L., Springer, C., and Kegelmeyer, Philip., 2001. Bagging is a small-data-set phenomenon. In International Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

Cho, Y.D.; Kondoz, A. (2001). Analysis and improvement of a statistical model-based voice activity detector, *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278.

Cover, T.M., and Thomas, J.A., 1991. *Elements of Information, Theory*. Wiley, New York, 1991.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27.

DeRouin, E., Brown, J., Beck, H., Fausett, L., and Schneider, M., 1991. Neural Network Training on Unequally Represented Classes, In Dagli, C.H., Kumara S.R.T., and Shin, Y.C., Intelligent Engineering Systems Through Artifical Neural Networks, ASME Press, pp. 135-145.

Domeniconi, C., Yan, B., 2004. Nearest neighbor ensemble. In: IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol 1, pp 228–231.

Drummond, C., and Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling Beats Over-Sampling, Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II.

Duda, R., Hart, P., and Stork, David., 2001. Pattern Classification. John Wiley and Sons, 2001. 0-471-05669-3.

Engelbrecht, A.P., Cloete, I., and Zurada, J.M., 1995.Determining The Significance Of Input Parameters Using Sensitivity Analysis, international Workshop on Artificial Neural Networks (1995).

Estabrooks, A., Jo, T., and Japkowicz, N., 2004. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, 20, pp. 18–36.

Estabrooks, A., 2000. A combination scheme for inductive learning from imbalanced data sets, Master's thesis, Dalhousie University, Halifax, Nova Scotia.

Fawcett, T., 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Technical Report HPL-2003-4, HP Labs.

Fawcett, T., 2006. An Introduction to ROC Analysis, Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874.

Fant, G., 1981. The source filter concept in voice production, QPSRSpeech Transmission Laboratory, vol. 1, pp. 21-37, 1981.

Friedman, J,. and Hall, P., On bagging and nonlinear estimation Technical report, Stanford University, 1999.Furundzic, D., Application example of neural networks for time series analysis: rainfall-runoff modelling, Signal Processing, vol. 64 (3), 1998, pp 383–396.

Furundzic, D. Application example of neural networks for time series analysis: rainfall-runoff modelling, Signal Processing, vol. 64 (3), 1998, pp 383–396.

Furundzic D., Djordjevic, M., and Bekic A. J., 1998. Neural Networks approach to early breast cancer detection. Journal of Systems Architecture, vol. 44 (8), 1998, pp 617-633.

Furundžić, D., Subotić M.,i Pantelić, S.,2006. Ocena poremećaja govora na nivou fonema primenom neuronskih mreža, Zbornik radova Digitalna obrada govora i slike, DOGS 2006, Vršac, Srbija, oktobar 2006.

Furundžić, D., Subotić, M., Pantelić, S., 2007. Primena neuronskih mreža u klasifikaciji poremećaja izgovora frikativa, ETRAN 2007, AK 5.4, Herceg Novi – Igalo, juni 2007.

Furundžić, D., Subotić, M., Punišić, S., 2009. Determination of relevant parameters influence in articulation regularity rating for Serbian phoneme " š " using neural networks, Speech and Language 2009, Proceedings 3rd International Conference on Fundamental and Applied Aspects of Speech and Language, pp. 158-167, Belgrade, Serbia, November 2009.

Furundzic, D., Djurovic, Z., Celebic, V., and Salom, I., 2012a. Neural Network Ensemble for Power Transformers Fault Detection, Proceedings of Eleventh Symposium on Neural network Applications in Electrical Engineering, Neurel 2012, Beograd, Septembar 2012, ISBN: 978-1-4673-1571-5, IEEE Catalog Number: CFP12481-PRT, pp. 247-251

Furundžić, D., Jovičić, S., Subotić, M., Punišić, S., 2012b. Acoustic Features Determination for Regularity Articulation Quantification of Serbian Fricatives, Proceedings of Eleventh Symposium on Neural network Applications in Electrical Engineering, Neurel 2012, Beograd, Septembar 2012, ISBN: 978-1-4673-1571-5, IEEE Catalog Number: CFP12481-PRT, pp. 197-201.

Furundžić, D., Jovičić, S., Subotić, M., Punišić, S., 2013a. Imbalanced Learming Approach to the Categorization Articulation, In Proceedings SPEECH AND LANGUAGE, pp. 89-100, Belgrade, October 2013.

Furundzic, D., Jovicic, S., Subotic, M., and Grozdic, D.,2013b. Evaluation of Phonemes Quality Articulation Using Neural Network Ensembles In Proceedings SPEECH AND LANGUAGE, pp. 183 -190, Belgrade, October 2013.

Furundžić, D., Stanković, S., Dimić, G., 2014a. Error Signal Distribution as an Indicator of Imbalanced Data, Proceedings of Twelfth Symposium on Neural network Applications in Electrical Engineering, Neurel 2014, Beograd, November 2014, ISBN: 978-1-4799-5887-0, IEEE Catalog Number: CFP14481-PRT, pp. 189-194.

Furundžić, D., Subotić, M., Punišić, S., 2015. Optimal Resampling Of Imbalanced Data: Speech Pathology Detection. SPEECH AND LANGUAGE 2015, 5th International Conference on Fundamental and Applied Aspects of Speech and Language, Belgrade, 17-18 October, 2015. ISBN: 978-86-89431-07-0, Publisher: Life activities advancement center, The Institute for Experimental Phonetics and Speech Pathology, Editors: Mirjana Sovilj, Miško Subotić, pp.282-297.

Furundzic, D., Punisic, S., Bilibajkic, R., 2017a. Probabilistic approach to the k nearest neighbor classifiers in the characterization of the phonemes, 6 th International Conference on Fundamental and Applied Aspects of Speech and Language, October 27-29, 2017, 147-153, Belgrade, Serbia.

Furundzic, D., Stankovic, S., Jovicic, S., Punisic, S., Subotic, M., 2017b. Distance based resampling of imbalanced classes: With an application example of speech quality assessment, Engineering Applications of Artificial Intelligence,vol. 64 (2017c),pp 440–461.

Garcia, S., Fernandez, A., Luengo, J., and Herrera, F., 2009. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing, 13 (10), pp. 959–977.

Garcia, V., Sanchez, J.S., Mollineda, R.A., 2008. On the use of surrounding neighbors for synthetic over-sampling of the minority class. In: Proceedings of the 8th WSEAS International Conference on Simulation, Modelling and Optimization, Santander, Spain, pp. 389–394.

Gazor, S.; Zhang, W. (2003). A soft voice activity detector based on a Laplacian-Gaussian model, *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505.

Geman, S., Bienenstock, E., and Doursat, R., 1992.Neural networks and the bias/variance dilemma. Neural Computation, 4(1):1:58, 1992.

Advances in Neural Information Processing Systems, number 11, pages 620-626. MIT Press, 1999.

Godino-Llorente, J. I., Gómez-Vilda, P., 2004. Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors, IEEE Transactions on Biomedical Engineering,, vol. 51, pp. 380-384.

Gray, A., and Markel, J., 1976. Distance measures for speech processing, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 381-391, Oct. 1976.

Grabowski, S., 2002. Voting over multiple k-nn classifiers. In: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science IEEE, pp 223–225.

Grozdić, Đ., Marković, B., Galić, J., Jovicic, S., Furundzic, D., 2013c. Neural Network Based Recognition of Whispered Speeach, In Proceedings SPEECH AND LANGUAGE, pp. 223 - 230, Belgrade, October 2013.

Guo, H., and Viktor, H.L., 2004. Learning from Imbalanced Data Sets with Boosting and Data Generation: the DataBoost-IM Approach, in SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets, vol. 6, issue 1, pp. 30 – 39.

Gustafsson, S.; Martin, R.; Jax, P.; Vary, P. (2002). A psychoacoustic approach to combined acoustic echo cancellation and noise reduction, IEEE Trans. Speech and Audio Processing, vol. 10, no. 5, pp. 245–256.

Guvenir, H.A,, Akkus, A., 1997. Weighted k nearest neighbor classification on feature projections. http:// www.cs.bilkent.edu.tr/tech-reports/1997/BU-CEIS-9719.pdf. Accessed 3 October 2014.

Hadjitodorov, S., Boyanov, B., Teston, B., 2000. Laryngeal pathology detection by means of class-specific neural maps, IEEE Transactions on Information Technologies and Biomedicine., vol. 4, pp. 68–79.

Hadjitodorov, S., P. Mitev, P., 2002. A computer system for acoustic analysis of pathological voices and laryngel diseases screening. Journal ELSEVIER, Medical Engineering & Physics, 24(6), pp. 419-29.

Hall, P., Samworth, R., 2005. Properties of bagged nearest neighbour classifiers. J. R. Stat. Soc. Ser. B (Statistical Methodology) 67(3):363–379.

Hansen, L.K., Salamon, P. , 1990. Neural network ensembles, IEEE Trans Pattern Anal vol. 12, pp. 993-1001.

Hart, P.E., 1968. The condensed nearest neighbor rule, IEEE Transactions on Information Theory 14, pp. 515–516.

Hastie, T., Tibshirani, R., and Friedman, J., 2001. The Elements of Statistical Learning, Springer-Verlag, New York.

Haykin, C., 1998. Neural Networks: A Comprehensive Foundation 2nd Edition, Prentice Hall.

He, H., Bai, Y., Garcia, E.A., and Li, S., 2008. ADASYN, Adaptive Synthetic Sampling Approach for Imbalanced Learning, Proc. Int. J. Conf. Neural Networks, pp. 1322-1328.

He, H., and Garcia, E.A., 2009. Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9), pp. 1263–1284.

Hebb, D. O., 1949. The Organization of Behavior, Wiley, New York, 1949.

Holte, R.C., Acker, L., and Porter, B.W., 1989. Concept Learning and the Problem of Small Disjuncts, Proc. Int'l J. Conf. Artificial Intelligence, pp. 813-818.

Hongyu, G., Herna, V.L., 2004. Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, SIGKDD Explorations Newsletter 6 (1), pp. 30–39.

Hossein, K., Babak, S. A., Mansour, N. B., 2009. Optimal feature selection for the assessment of vocal fold disorders, Computers in Biology and Medicine, vol. 39, pp. 853-946.
Hothorn, T., Lausen, B., 2003a. Bagging tree classifiers for laser scanning images: a data-and simulation-based strategy. Artif Intell Med 27(1):65–79.
Hunt, B., Qi, Y., and Dekruger, D., 1992. Fuzzy classification using set membership functions in the back propagation algorithm, Heuristics, J. Knowledge Eng., vo. 5, no. 2, pp. 62-74, 1992.
Japkowicz, N., 2003. Class imbalance: Are we focusing on the right issue? In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets, Washington DC.
Japkowicz, N., 2001. Concept learning in the presence of between-class and within-class imbalances, In: Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence, pp 67-77, Springer-Verlag.
Japkowicz, N., Myers, C., and Gluck, M., 1995. A Novelty Detection Approach to Classification. In: Proceedings of the Fourteenth Joint Conference on Artificial Intelligence, pp. 518—523.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study, Intelligent Data Analysis 6 (5), pp. 429–449.
Jaynes, E.T., 1957. Information theory and statistical mechanics. Physical Review, Vol. 106(4), pp. 620–630.
Joshi, M. V., Kumar, , and Agarwal, R. C.,2001. Evalating boosting algorithms to classify rare classes: Comparison and improvements. In Proceeding of the First IEEE International Conference on Data Mining(ICDM'01), 2001.
Jovičić, S.T. (1999). Govorna komunikacija: fiziologija, psihoakustika i percepcija, Nauka, Beograd.
Jovičić, S. T., Punišić, S. (2007a). Perceptivno prepoznavanje akustičkih obeležja koja karakterišu odstupanja u izgovoru frikativa /š/, Zbornik radova LI Konferencija ETRAN, Herceg Novi, AK5.1.
Jovičić, S.T., Kašić, Z., Punišić, S. (2008). Trajanje frikativa /š/: analiza u izolovanim rečima, Zbornik radova, XVI TELFOR, Belgrade, 715-718.
Jovičić, S. Punišić, S., Šarić, Z. (2008). Time frequency detection of stridence in fricatives and affricates, Int. Conf. Acoustics 08, Paris, 5137-5141.
Jovičić S., Kašić Z., Punišić S. (2010). Production and perception of distortion in word-initial friction duration. Journal of Communication Disorders 43/5., DOI:10.1016/j. jcomdis.2010.04.007.
Karray, L.; Martin. A. (2003). Toward improving speech detection robustness for speech recognition in adverse environments, Speech Communication, no. 3, pp. 261–276.
Kašić, Z., Jokanović-Mihajlov, J., Tomić, T., Peter, S., Filipović, M. (1986). Neki međuodnosi percepcije i produkcije glasova sintezom. Psiholingvistički susreti, Beograd.
Kašić, Z., Peter, S., Urošević, Z., Filipović, M. (1987). Varijantnost trajanja glasova u reči. U: Zbornik radova ETAN, Bled, 229-233.
Kašić, Z., (1990). Sandhi i neutralizacija distinktivnih obeležja. Književnost i jezik, br.1, Beograd, 71-73.
Kašić, Z., (1997). Promene glasova uslovljene ritamsko-intonacionom organizacijom govora, Beogradska defektološka škola, br. 1, 77-82.
Kašić, Z., (1998). Slogovi i konsonantski skupovi u artikulacionoj bazi srpskog jezika. Beogradska defektološka škola, br. 1, 101-109.
Kašić, Z., (2000a). Funkcija suprasegmenata u govornom izrazu. Beogradska defektološka škola, br.2-3, 113-124.
Kašić, Z., (2000b). Segmentna i suprasegmentna organizovanost govora. U: S. Golubović i Z. Kašić Segmentna i suprasegmentna organizovanost govora i poremećaji fluentnosti. Beograd: Društvo defektologa Jugoslavije.

Kašić, Z., (2003a). Fonetika. Autorizovani rukopis, udžbenik za studente defektološkog fakulteta, Beograd.
Kašić. Z., (2003b). Percepcija distinktivnih obeležja u izolovanim jednosložnim rečima kod dece mlađeg školskog uzrasta. Istraživanja u defektologiji - Smetnje u razvoju. Beograd: Defektološki fakultet - CIDD, 217-240.
Kašić Z., Jovičić S.T., Ivanović M. (2004). Problemi segmentacije glasova u prirodnom kontinuiranom govoru. Zbornik radova DOGS 2004. Novi Sad: Fakultet tehničkih nauka: 37-40.
Kašić Z., Jovičić S., Đorđević M. (2006). O definisanju kraja izgovorene reči. Beograd: ETRAN 2006, Zbornik radova, 462-465.
Khoshgoftaar, T., Van Hulse, J., Napolitano, A., 2011. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern Part A Syst Hum 41(3):552–568.
Kohonen, T., 1984. Self Organization and Associative Memory, Third ed., Springer Verlag Berlin, 1984.
Kostić, Đ., Vladisavljević, S., Popović, M. (1983): Testovi za ispitivanje govora i jezika. Zavod za udžbenike i nastavna sredstva, Beograd.
Kostić, Đ., Nestorović, M., Kalić, D. (1964). Akustička fonetika srpskohrvtskog jezika, 2, Glasovno polje, Institut za eksperimentalnu fonetiku i patologiju govora, Beograd.
Kubat, M., Holte, R., and Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 30, pp. 195–215.
Kubat, M., and Matwin, S., 1997. Addressing the curse of imbalanced training sets: One sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186, Morgan Kaufmann.
Laurikkala, J., 2001. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere.
Lawrence, S., Burns, I., Back, A., Tsoi, A. Chung., and Giles, C. Lee., 1998. Neural network classification and prior class probabilities. Lecture Notes in Computer Science, 7700 LECTURE NO 299-314.
Lee, M., van Santen, J., Mobius, B., and Olive, J., 2005. Formant tracking using context-dependent phonemic information", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 5, Sept 2005, pp. 741-751.
Lewis, D., and Gale, W.A., 1994. A sequential algorithm for training text classifiers. In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 3-12, London, Springer-Verlag.
Li, Q.; Zheng, J.; Tsai, A.; Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition, IEEE Trans. Speech Audio Processing, vol. 10, no. 3, pp. 146–157.
Liebermann, P., 1961. Perturbations in vocal pitch, The Journal of the Acoustical Society of America Vol. 33, No. 5.
Ling, C. and Li, C., 1998. Data Mining for Direct Marketing: Problems and Solutions. In: Proc. of 4th International Conference on Knowledge Discovery and Data Mining, pp.73-79.
Lippman, R., 1987. An introduction to computing with neural nets, IEEE ASSP Mag., vol. 1, pp. 4-22, 1987.
Liu, B., 2007. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag, Berlin, Heidelberg, New York, NY.
Ma, B.J., Wei, Q., and Chen, G.Q., 2011. A combined measure for representative information retrieval in enterprise information systems. Journal of Enterprise Information Management, 24 (4), pp. 310–321.

Maguire, C., De Chazal, P., Reilly, R. B., Lacy, P., Automatic classification of voice paihology using speech analysis", Ward Congress on Biотедгcal Engineering and Medical Physics, Sydney, August 2003;

Maguire, C., De Chazal, P., Rei11y, R. B., Lacy, P. 2003. Identification of voice pathology using automated speech analysis", Proc. Of the 3rd International Workshop on Models and Analysis of Voice Emissian for Biomedieal Applications, Florence, December 2003a.

Maier, A., Hönig, F., Bocklet, T., Nöth, E., Stelzle, F., Nkenke, E., Schuster M., 2009. Automatic detection of articulation disorders in children with cleft lip and palate, Journal of the Acoustical Society of America, Vol. 126, No. 5, pp.2589–2602.

Maloof, M., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In: Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets.

Manfredi, C., 2001. Adaptive noise energy in pathological signals, IEEE Transactions on Biomedical Engineering, vol. 47, pp. 1538–1543.

Markaki, M., Stylianou, Y., 2011. Voice Pathology Detection and Discrimination Based on Modulation Spectral Features, Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.7, pp.1938,1948.

Melville, P., Shah N., Mihalkova, L., Mooney, R., (2004) Experiments on ensembles with missing and noisy data. In: Roli, F., Kittler, J., Windeatt, T., (eds) Lecture Notes in Computer Science: Proceedings of the Fifth International Workshop on Multi Classifier Systems (MCS-2004), Cagliari, Italy. Springer, Heidelberg, pp 293–302.

Miehaelis, D., Frohlich, M., Strube, H.W., 1998. Selection and combination of acoustic features for the description of pathologic voices, The Journal of the Acoustical Society of America, Vol. 103, No. 3, pp 1628-1639.

Mitchell, T.M., 1997. Machine Learning, McGraw-Hill Series in Computer Science, WCB McGraw-Hill, Boston, MA.

Montano J.J., Palmer A., (2003) .Numeric sensitivity analysis applied to feedforward neural networks. Neural Comput Appl 12:119–125.

Moran, R. Reilly, R. B., De Chazal P., Lacy, P., 2004. Telephone based voice pathalogy assessment using automated speech analysis and VoiceXML", ISSC 2004, Belfast, June 2004.

Naumovic, R., Furuncic, D., Jovanovic, D., Stosovic, M., Basta-Jovanovic, G., Lezaic, V., 2010. Application of artificial neural networks in estimating predictive factors and therapeutic efficacy in idiopathic membranous nephropathy, Biomedicine and pharmacotheraphy, Vol. 64. ,No. 9, November 2010, pp 633-638.

Nickerson, A., Japkowicz, N., and Millos, E., 2001. Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets. In: Proceedings of the 8th International Workshop on AI and Statistics, pp. 261-265.

Niles, L., Silverman, H., Tajchman, G., and Bush, ,1989. How limited training data can allow a neural network to outperform an optimal statistical classifier, in Proc. ICASSP89, vol. 1, pp. 17-20, 1989.

Niles, L., Silverman, , Tajchman, G., and Bush, M.,1989. The effects of training set size on relative performance of neural network and other pattem classifiers, Tech. Rep. LEMS-51, Brown University, Providence, RI, 1989.

Paek, T., Hsu, P., 2011. Sampling representative phrase sets for text entry experiments: A procedure and public resource. In: Proceedings of CHI, pp. 2477-2480. ACM Press.

Pan, F., Wang, W., Tung, A.K.H., and Yang J., 2005. Finding representative set from massive data. In: The 5th IEEE International Conference on Data Mining (ICDM), pp. 338–345. IEEE Computer Society, Washington, DC (2005).

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., and Brunk, C., 1994. Reducing Misclassification Costs. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 217-225.

Paulraj, M.P.; Yaacob, S.; Hariharan, M., 2009. Diagnosis of vocal fold pathology using time-domain features and systole activated neural network, CSPA 2009. 5th International Colloquium on Signal Processing & Its Applications (, vol., no., pp.29,32, 6-8 March 2009.

Perperoglou, Gul, A., Khan, A., et al. 2016. Ensemble of a subset of *k*NN classifiers. Advances in Data Analysis and Classiffication, pp. 1-14.

Poggio, T., Rifkin, R., Mukherjee, S., and Rakhlin, Alex., 2002. Bagging regularizes. Technical Report AI Memo 2002-003, CBCL Memo 214, MIT AI Lab, 2002.

Provost F., and Fawcett, T., 1997. Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions. In: Proceedings of the 3rd International Conference on Knowledge Representation and Data Mining, Cambridge, AAAI Press, pp. 43–48.

Punišić, S., Subotić, M., Furundzic, D., 2017. Identificational probability functions of the perceptual recognition of africate's and fricative's duration, 6th International Conference on Fundamental and Applied Aspects of Speech and Language, October 27-29, 2017, 154-167, Belgrade, Serbia

Rabiner, L.,and Sambur, M., 1977. Application of an LPC distance measure to the Voiced-Unvoiced-Silence detection problem, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 338-343, Aug. 1977.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, A.; Rubio, A. (2003). A new adaptive longterm spectral estimation voice activity detector, Proc. EUROSPEECH 2003, Geneva, Switzerland, pp. 3041–3044.

Rissanen, J., 1978. Modeling by the shortest data description. Automatica, 14:465– 471, 1978.

Ruck, D., Rogers, S., Kabrisky, M., Oxley, M., and Suter, B., 1990. The multilayer perceptron as an approximation to a Bayes optimal discriminant function, IEEE. Trans. Neural Networks, vol. pp. 296-268, Dec. 1990.

Rumelhart, D., Hinton, G., and Williams, R., 1986. Learning intemal representations by error propagation, in Parallel Distributed Processing: Explorations in the Microstructures of Cognition D. Rumelhart and J. McClelland, Eds., vol. l,\quad Cambridge, MA: MIT Press, 1986, pp. 318-362.

Samworth, R. J., 2012. Optimal weighted nearest neighbour classifiers. Ann Stat 40(5):2733–2763.

Sangwan, A.; Chiranth, M.C.; Jamadagni, H.S.; Sah, R.; Prasad, R.V.; Gaurav, V. (2002). VAD Techniques for Real-Time Speech Transmission on the Internet, IEEE International Conference on High-Speed Networks and Multimedia Communications, pp. 46-50.

Shannon, C.E., 1976. A mathematical theory of communication. Bell System Tech. J., 27, pp. 379-423.

Sharkey, A.J. , 1999. Combining Artificial Neural Nets, Springer, London.

Schipor, O.A., Pentiuc, S.G., Schipor, M.D. (2012). Automatic Assessment of Pronunciation Quality of Children within Assisted Speech Therapy. Electronics and Electrical Engineering., 2012. – No. 6(122). pp. 15–18.

Siegel, L., 1979. A procedure for using pattem classification techniques to obtain a Voicednlnvoiced classifier," IEEE Trans. Acousr., Speech, Signal Processing, vol. ASSP-27, pp. 83-88, Feb. 1979.

Siegel, L., and Bessey, A., 1982. Voiced/unvoiced/Mixed excitation classification of speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, pp. 451460, June 1982.

Stevanović, M. (1981). Savremeni srpskohrvatski jezik I, Naučna knjiga, Beograd.

Ishizaka, K., Flangan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell System Technical Journal, 51, 1233-1268.

Ting, K. M., 1994. The problem of small disjuncts: its remedy in decision trees. In Proceedings of the Tenth Canadian Conference on Artificial Intelligence, pages 91–97, Banff, Alberta, May 1994.

Titze, I. R., Talkin, D. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. Journal of the Acoustical Society of America, 66, 60-74.

Tomek, I., 1976. Two modifications of CNN, IEEE Transactions on Systems, Man and Cybernetics 6 (11), pp. 769-772.

Valentini, G., and Thomas, G., Dietterich, 2002. Bias-variance analysis and ensembles of SVM. In Proc. Int. Workshop on Multiple Classi¯er Systems (LNCS 2364), pages 222-231, Calgiari, Italy, June 2002. Springer.

Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. Journal of Speech and Hearing Research, 1, 227-244.

Van den Berg, J., Zantema, T., Doornenbal, P. Jr. (1957). On the air resistance and the Bernoulli effect of the human larynx. Journal of the Acoustical Society of America, 29, 626-631.

Vasilakis, M., Stylianou, Y., 2009. Voice pathology detection based on short-term jitter estimations in running speech, Folia Phoniatrica et Logopaedica, 61 (3), pp.153-170.

Vladisavljević, S. (1981). Poremećaj artikulacije. Privredni pregled. Beograd, 1981.

Wallen, E. J., Hansen, H.L., 1996. A Screening Test for Speech Pathology Assessment Using Objective Quality Measures, ICSLP, Proceedings of Fourth International Conference, PA, USA, vol.2, pp.776-779.

Weiss, G.M., 2004. Mining with Rarity: A Unifying Framework, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7-19.

Weiss, G.M., 2003. The effect of small disjuncts and class distribution on decision tree learning: PhD thesis, Rutgers University.

Weiss, G.M., and Provost, F., 2003. Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research, pp. 315–354.

Wahba, G., Lin, X., Gao, F., Xiang, D., Klein, R., and Klein, B.,1999. The bias-variance tradeoff and the randomized GACV. In M. Kearns, S. Solla, and D. Cohn, editors,

Widrow, B., and M. Hoff (1960), "Adaptive Switching Circuits", 1960 WESCON Convention Record, New York, in: [Anderson and Rosenfeld 1989].

Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man and Cybernetics, 2, pp. 408–421.

Wu, G., and Chang, E.Y., 2003. Class-Boundary Alignment for Imbalanced Dataset Learning. In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets, Washington DC.

Zhai, C.X., Cohen, W.W., and Lafferty, J., 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, pp. 10-17.

Zhang, J., and Mani, I., 2003. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: Proceedings of International Conference of Machine Learning (ICML '2003), Workshop Learning from Imbalanced Data Sets.

Zhang, Y., Callan, J., and Minka, T., 2002. Novelty and redundancy detection in adaptive filtering. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 2002, Vol. 02, ACM Press, New York, NY, pp. 81-8.

Zhou, Z. H., Tang,W. W. , 2002. Ensembling neural networks: many could be better than all, Artificial Intelligence vol. 137, pp. 239–263.

Zhou, ZH., Yu, Y., 2005. Adapt bagging to nearest neighbor classifiers. J Comput Sci Technol 20(1):48–54.

Zurada,J.M., Malinowski, A., and Cloete,I., 1993. Sensitivity Analysis for Minimization of Input Data Dimension for Feedforward Neural Network, IEEE International Symposium on Cirquits and Systems, London, June, 1994.

Zurada J. M., Malinowski A., Usui S., 1997. Perturbation method for deleting redundant inputs of perceptron networks, Neurocomput. vol. 14 pp. 177-193 1997.

Zurada JM, Malinowski A, Cloete I (1994) Sensitivity analysis for minimization of input data dimension for feedforward neural network. In: Proceedings IEEE International Symposium on Circuits and Systems, IEEE, New York, pp. 447–450 12.

Wang W, Jones P, Partridge D (2000) Assessing the impact of input features in a feedforward neural network. Neural Computing & Applications 9:101–112 11.

Biografski Podaci Autora

Kandidat, Draško (Vučić) Furundžžić je rođen 21. 03. 1958. godine u Bijelom Polju u Jugoslavenskoj republikci sr. C. Gora gde je završio osnovnu školu i gimnaziju sa odličnim uspehom. Završio je studije na Mašinskom fakultetu u Beogradu na smeru za medicinsku tehniku 1988. Posebno impresiju na kandidata ostavljajuju predavanja profesora Mašinskog fakulteta Ljubomira T. Grujića (Automatsko upravljanje) i prof. Medicinskog fakulteta Mihajla Mojović (Medicinska Fiziologija). Interesovanje za biomedicinske nauke bio je motiv da upiše postdiplomske studije na Medicinskom fakultetu u Beogradu, smer eksperimentalna fiziologija i patološka fiziologija, sa posebnim interesovanjem za fiziologiju nervnog sistema, kao osnove budućeg angažovanja u oblasti veštačke inteligencije. Od 1989 godine kandidat radi u Institutu Mihajlo Pupin u Beogradu, gde je i danas u stalnom radnom odnosu. U tadašnjoj grupi za inženjering znanje dobio je mogućnost da radi u domenu primene veštačke inteligencije u analizi sistema. Magistarski rad pod nazivom "Metodologija primene veštačkih neuronskih mreža u obradi i klasifikaciji signala" kandidat je odbranio 2009 godine na Elektrotehničkom fakultetu u Beogradu pod mentorskim nadzorom profesora Srđana Stankovića. Kandidat je radio na istraživanjima i projektima u oblastima hidrologije, meteorologije, klimatologije, geofizike, energetike, prepoznavanja teksta, prepoznavanja govora, patologije govora, onkologije, urologije i nefrologije, neurokardiologije, analize EEG i ECG signala i laserske tehnike. Objavio je radove u prestižnim međunarodnim časopisima, monografijama i zbornicima radova sa međunarodnih i domaćih konferencija. Učestvovao je u pregledanju radova za međunarodne časopise i simpozijume. Od 2000. godine kandidat održava seminar "Primena veštačke inteligencije u medicini" za studente smera za tehničku fiziku (ETF Beograd) kod profesora Dejana Rakovića. Posebno mesto u biografiji kandidata pripada katedri za Automatskog upravljanja (Signali i sistemmi) Elektrotehničkog fakulteta u Beogradu a naročito profesoru Srđanu Stankoviću.

PRILOZI

Prilog I – Globalni artikulacioni test - GAT

ИЕФПГ

Глобални артикулациони тест
(Костић Ђ., Владисављевић С.)

ИЕФПГ рег. бр.
Датум:
Лист: 203 од 211

Име и презиме _____ датум рођења_____

дужина трајања тестирања_____ испитивач_____

речи	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**примедба**
и - види								
е - беба								
а - мама								
о - вода								
у - буба								
п - пада								
б - баба								
т - тата								
д - деда								
к - кока								
г - гума								
ц - цица								
ћ - ћебе								
ђ - ђак								
ч - чело								
џ - џеп								
ф - фес								
в - воз								
с - сека								
з - зима								
ш - шума								
ж - жаба								
х - ходи								
ј - јаје								
р - риба								
м - мој								
н - нога								
њ - њива								
л - лице								
љ - људи								
Укупно								

PRILOG II - Test za analitičku ocenu artikulacije srpskog jezika - AT

ИЕФПГ

Аналитичка оцена артикулације српског језика

ИЕФПГ рег. бр.
Датум:
Лист: 204 од 211

Име и презиме _____датум рођења _____

дужина трајања тестирања_____ испитивач_____

В О К А Л И

Гласовна одступања	**И**	**Е**	**А**	**О**	**У**	**Укупно**
1. продужен						
2. скраћен						
3. обезвучен						
4. обезвучен на почетку						
5. обезвучен на крају						
6. озвучен на крају						
7. висок тон						
8. низак тон						
9. назализован						
10. јако наглашено						
11. отворенији						
12. затворенији						
13. вокал као						
14. заокружено И, Е						
15. развучено У						
16. нема гласа						
17. централни глас						
18. супституција						
СВЕГА						

И	игла, иње, сир, лист, очи	Иди и купи ми новине.
Е	ексер, ера, џеп, пећ, дете	Две девојчице седе поред пећи и веселе се.
А	ауто, ашов, пас, сат, маца	Данас пада киша.
О	одело, орао, топ, воз, око	Дошли смо возом око осам сати.
У	уво, уста, лук, зуб, једу	Буди код куће сутра ујутру.

П Л О З И В И

Гласовна одступања	**П**	**Б**	**Т**	**Д**	**К**	**Г**	**Укупно**
1. безвучно							
2. звучно							
3. продужена оклузија							
4. скраћена оклузија							
5. обезвучена оклузија							
6. обезвучена оклузија на почетку							
7. обезвучена оклузија на крају							
8. јака експлозија							
9. слаба експлозија							
10. дентолабијално							
11. лингволабијално							

12. интердентално							
13. алвеоларно							
14. посталвеоларно							
15. фрикативно							
16. палатализовано							
17. аспировано							
18. веларизовано							
19. назализовано							
20. нема гласа							
21. централни глас							
22. супституција							
СВЕГА							

П	пиле, капа, џеп	Пази кад се пењеш на ову опасну планину.
Б	баба, беба, зуб	Обећао си Бошку да ћемо у суботу ићи у клуб.
Т	топ, ауто, лист	У твојој ташни стоје апарат и карта.
Д	деда, одело, лед	Данас смо однели дете у обданиште, мада је далеко.
К	кућа, рука, лук	Како да купим кућу када толико кошта.
Г	голуб, вага, снег	Погледајте снег и гавранове на гранама.

А Ф Р И К А Т И

Гласовна одступања	**Ц**	**Ћ**	**Ђ**	**Ч**	**Џ**	**Укупно**
1. безвучно						
2. звучно						
3. продужена оклузија						
4. скраћена оклузија						
5. продужена африкација						
6. скраћена африкација						
7. обезвучена оклузуја						
8. обезвучен на почетку						
9. обезвучен на крају						
10. оштра африкција						
11. слаба африкција						
12. потпуна фрикција						
13. стриденс						
14. коронално						
15. интердентално I степен						
16. интердентално II степен						
17. интердентално III степен						
18. адентално						
19. десна латерална африкција						
20. лева латерална африкција						
21. напред померена африкција						
22. назад померена африкција						
23. сажет глас						
24. заокружене усне						
25. нема гласа						
26. централни глас						
27. супституција						
СВЕГА						

Ц	ципеле, маца, ловац	На цести се играју деца и не могу да процене опасност.
Ћ	ћурка, кућа, коњић	По овој мећави ћурке су се шћућуриле уз кућу.
Ђ	ђак, леђа, чађ	Ђаци желе да пођеш и да им будеш вођа.
Ч	чамац, очи, кључ	Свако вече пијемо чај и читамо.
Џ	џак, оџак, беџ	Џудиста скида џемпер и тренира џудо.

Ф Р И К А Т И В И

Гласовна одступања	Ф	В	С	З	Ш	Ж	Х	Ј	Р	Укупно
1. безвучно										
2. звучно										
3. продужено										
4. скраћено										
5. јака фрикција										
6. слаба фрикција										
7. високо										
8. ниско										
9. назализовано										
10. билабијално Ф										
11. билабијално В										
12. стриденс										
13. коронално										
14. интердентално I степен										
15. интердентално II степен										
16. интердентално III степен										
17. адентално										
18. заокружене усне за С										
19. десна латерална фрикција										
20. лева латерална фрикција										
21. алвеоларизовано										
22. палатализовано										
23. Ј слично И										
24. Ј слично Љ										
25. преоштро Х										
26. Х померено назад										
27. гутурално Х										
28. преградно Х										
29. енглеско										
30. веларно										
31. ресично										
32. грлено без вибрација										
33. унилатерално										
34. неодређено										
35. нема гласа										
36. централни глас										
37. супституција										
СВЕГА										

Ф	фењер, кафа, шраф	Имам фини доручак: кафу са кифлом.
В	воз, авион, лав	Дувао је ветар а војници су имали вежбе у рову.
С	сир, лист, пас	Помислио сам на његову досетку и слатко сам се смејао.
З	зуби, коза, воз	Зашто не пазиш на тај зумбул у зеленој вазни.
Ш	шума,кишобран, миш	Нашли смо у шупи кишобран и шиваћу машину.
Ж	жаба, ружа, нож	Жарко је пожелео жуту ружу.
Х	хармоника,јахач, шах	Хтео бих да вас победим у шаху.
Ј	јагоде, јаје, змај	Јутрос су јавили Дејану да му је мајка болесна.
Р	риба, буре, лептир	Има примерено владање и препоруке својих професора.

Н А З А Л И

Гласовна одступања	М	Н	Њ	Укупно
1. продужено				
2. скраћено				
3. обезвучено				

4. обезвучено на почетку					
5. обезвучено на крају					
6. слаба назализација					
7. високо					
8. ниско					
9. непотпуна преграда за М					
10. М личи на Б					
11. непотпуна преграда за Н					
12. интердентално					
13. јача експлозија за Н					
14. Н ближе гласу Њ					
15. непотпуна преграда за Њ					
16. Њ померено према Н					
17. веларизовано					
18. нема гласа					
19. централни глас					
20. супституција					
СВЕГА					
М	маца, лампа, сом	Мала Милица моли маму да јој да мало меда.			
Н	нос, динар, лимун	Није се никада ни у сну тако нечем надао.			
Њ	њушка, диња, коњ	Иње се окачило о грање које се њише.			

Л А Т Е Р А Л И

Гласовна одступања	**Л**	**Љ**	**Укупно**
1. продужено			
2. скраћено			
3. обезвучено			
4. обезвучено на почетку			
5. обезвучено на крају			
6. високо			
7. ниско			
8. назализовано			
9. интердентално			
10. слабо одизање језика			
11. десна латерализација			
12. лева латерализација			
13. језик се не диже			
14. померање уназад			
15. слично гласу Љ			
16. језик јако пресавијен			
17. Љ слично гласу Л			
18. Љ померено напред			
19. нема гласа			
20. централни глас			
21. супституција			
СВЕГА			
Л	лист, кола, шал	Под липом је велики и лепи хлад.	
Љ	љуљашка,уље, пасуљ	Узми кључ и закључај врата мало боље.	

Прилог 1.

Изјава о ауторству

Потписани-а Draško Furunžić

број уписа _____

Изјављујем

да је докторска дисертација под насловом

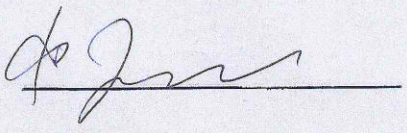
Ocena kvaliteta artikulacije glasova srpskog jezika primenom neuronskuh mreža

_____

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 4. 05. 2018.



Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Draško Furundžić

Број уписа

Студијски програм Signali i sistemi

Наслов рада Ocena kvaliteta artikulacije glasova srpskog jezika primenom neuronskih mreža

Ментор Profesor Dr. Srđan Stanković

Потписани Draško Furundžić

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 4. 05. 2018.

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Ocena kvaliteta artikulacije glasova srpskog jezika primenom neuronskuh mreža

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство
2. Ауторство - некомерцијално
3. Ауторство – некомерцијално – без прераде
4. Ауторство – некомерцијално – делити под истим условима
5. Ауторство – без прераде
6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 4. 05. 2018.

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.