УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Стефан М. Костић

ПРЕДИКЦИЈА ГУБИТКА КОРИСНИКА У МОБИЛНИМ ТЕЛЕКОМУНИКАЦИОНИМ МРЕЖАМА ПРИМЕНОМ НЕНАДГЛЕДАНОГ МАШИНСКОГ УЧЕЊА

докторска дисертација

Београд, 2021

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

Stefan M. Kostić

CUSTOMER CHURN PREDICTION IN MOBILE TELECOMMUNICATION NETWORKS APPLYING UNSUPERVISED MACHINE LEARNING

Doctoral Dissertation

Belgrade, 2021

ПОДАЦИ О МЕНТОРУ И ЧЛАНОВИМА КОМИСИЈЕ

Ментор:

др Мирјана Симић-Пејовић, ванредни професор

Универзитет у Београду – Електротехнички факултет

Чланови комисије:

др Милан Бјелица, редовни професор

Универзитет у Београду – Електротехнички факултет

др Ненад С. Митић, редовни професор

Универзитет у Београду – Математички факултет

др Бошко Николић, редовни професор

Универзитет у Београду – Електротехнички факултет

др Предраг Пејовић, редовни професор

Универзитет у Београду – Електротехнички факултет

Датум одбране:

Захвалница

Најпре, желим да се захвалим менторки, проф. др Мирјани Симић-Пејовић, на пруженој подршци, стручним саветима и великој посвећености током целокупног процеса мојих докторских студија.

Захваљујем се и члановима Комисије за преглед и оцену докторске дисертације, проф. др Милану Бјелици, проф. др Ненаду Митићу, проф. др Бошку Николићу и проф. др Предрагу Пејовићу на корисним сугестијама и саветима, који су допринели квалитету финалне верзије докторске дисертације.

Хвала и свим пријатељима, члановима шире породице, као и девојци на подршци и охрабривању у току докторских студија.

Посебну захвалност на великом труду и одрицању без којег не бих био у прилици да достигнем овај академски успех дугујем благопочившим баки Ђурени, баки Мирослави, деки Владимиру и деки Томиславу.

Највећу захвалност дугујем мојој породици, оцу Мирољубу, мајци Ирени и сестри Мини, на неизмерном разумевању, стрпљењу, љубави и подршци у току докторских студија. Њима посвећујем овај рад.

РЕЗИМЕ

Наслов докторске дисертације: Предикција губитка корисника у мобилним телекомуникационим мрежама применом ненадгледаног машинског учења

Резиме: Услед засићења и велике конкуренције на тржишту мобилних телекомуникација, постало је веома значајно да телекомуникациони оператори увек имају актуелан увид у динамику својих клијената. Управо из тог разлога анализа друштвених мрежа и њена примена са теоријом графова могу бити врло значајне. У овој докторској дисертацији биће анализирана друштвена мрежа која је представљена великим графом позива мобилне телекомуникационе мреже. Конкретно, биће приказано кластерисање њених чворова на основу широког скупа метрика, као што су улазни и излазни степен чвора, утицај чвора првог и другог реда, сопствени вектор чвора, као и вредност ауторитета и *hub* вредност чвора. У резултатима представљеним у докторској дисертацији биће показано да је могуће идентификовати неке важне чворове у анализираној друштвеној мрежи (графу) који су витални у погледу предвиђања губитка корисника. Биће доказано да ће, ако такав чвор напусти мрежу анализираног мобилног телекомуникационог оператора, клијенти који често комуницирају са тим одређеним чвором такође бити склонији напуштању мреже анализираног телекомуникационог оператора. Дакле, анализом претходно изгубљених корисника и анализом њихових претходних позива, проактивно се предвиђају нове кориснике који ће са високом вероватноћом такође одлучити на сличну акцију. Резултати модела за предвиђање губитка корисника квантификовани су коришћењем *Lift* метрике за најугроженијих десет процената анализиране популације. Предложени метод је довољно општи да се може лако усвојити у било ком пољу где се хомофилне или пријатељске везе могу претпоставити као потенцијални покретач губитка корисника.

Кључне речи: Ненадгледано машинско учење, кластеризација, теорија графова, предикција губитка корисника, анализа друштвених мрежа, наука о подацима

Научна област: Електротехника

Ужа научна област: Телекомуникације

УДК број:

ABSTRACT

Dissertation title: Customer churn prediction in mobile telecommunication networks applying unsupervised machine learning

Abstract: Due to telecommunications market saturation and competition, it is very important for telco operators to always have fresh insights into their customer's dynamics. In that regard, social network analytics and its application with graph theory can be very useful. In this doctoral dissertation a social network that is represented by a large telco network graph is analysed and clustering of its nodes is performed by studying a broad set of metrics. More specifically, the clustering process will be performed by using node in/out degree, first and second order influence, eigenvector, authority and hub values. This doctoral dissertation demonstrates that it is possible to identify some important nodes in analysed social network (graph) that are vital regarding churn prediction. It will be shown that if such a node leaves a monitored telco operator, customers that frequently interact with that specific node will be more prone to leave the monitored telco operator network as well; thus, by analysing existing churn and previous call patterns, we proactively predict new customers that will probably churn. The churn prediction results are quantified by using top decile lift metrics. The proposed method is general enough to be readily adopted in any field where homophilic or friendship connections can be assumed as a potential churn driver.

Keywords: Unsupervised machine learning, clustering, graph theory, churn prediction, social network analysis, data science

Scientic area: Electrical Engineering

Scientic subarea: Telecommunications

UDK code:

САДРЖАЈ

СПИСАК СЛИКА

СПИСАК ТАБЕЛА

СПИСАК СКРАЋЕНИЦА

ADASYN	Adaptive Synthetic Sampling approach
CART	Classification and Regression Trees
CCC	Cubic Clustering Criterion
CDC	Center for Disease Control and Prevention
CDR	Call Data Record
CHAID	Chi-square Automatic Interaction Detector
EH	Equal Height
EQ	Equal Width and Height
EV	Equal Volumes
EW	Equal Widths
ICOTE	Immune Centroids Oversampling Technique
KBS	Knowledge Based System
LP	Linear Perceptron
MLP	Multilayer Perceptron
MSISDN	Mobile Station International Subscriber Directory Number
MTDF	Mega-trend Diffusion Function
MWMOTE	Majority Weighted Minority Oversampling Technique
NRBF	Normalized Radial Basis Function
ORBF	Ordinary Radial Basis Function
PCA	Principal Component Analysis
PSF	Псеудо-F статистика

PST2	Псеудо-T2 статистика
RBF	Radial Basis Function
RDR	Ripple Down Rule
SE	Simulated Expert
SMOTE	Synthetic Minority Oversampling Technique
SVD	Singular Value Decomposition
TRkNN	Couples Top-N Reverse k-Nearest Neighbor
UN	Unequal Width and Height
UW	Unequal Widths

1. УВОД

Информационе технологије су промениле начин на који живимо и радимо и имају потенцијал да и у будућности наставе да утичу на начин на који људи живе и раде [1]. Уколико се посматрају карактеристике рачунара, може се видети да су у последњих пар деценија перформансе процесора повећане више од 10 000 пута, док је истовремено њихова цена постајала све приступачнија већини корисника [1]. Овај развој је био могућ константним напретком при пројектовању процесорских чипова (конкретно, смањивањем величине транзистора), као и увећавањем меморије рачунара и повезивањем појединачних рачунара у сложене рачунарске системе. Ипак, битна ставка при развоју била је компатибилност са претходним моделима; штавише, старији програми су могли да се извршавају брже на новим конфигурацијама, а уједно су новији рачунари омогућавали развој нових (сложенијих) апликација. Почетком 21. века напредак у повећању перформанси процесора (конкретно, броју операција које процесор може извршити у секунди) посустаје, те се прешло на алтернативне методе убрзавања рада рачунара попут коришћења више процесорских јединица унутар рачунара и паралелног процесирања. Додатно, у савременом рачунарству уведен је принцип „рачунарства у облаку“ (*cloud computing*) где се рачунарски ресурси посматрају као услуга која се пружа крајњем кориснику и којима корисник приступа путем интернета. На такав начин крајњем кориснику је омогућен флексибилан и бржи рад, чак и у случају да корисник не поседује квалитетан и напредни хардвер.

Са развојем информационих технологија, постојали су различити изазови који су ограничавали примену информационих технологија у науци и индустрији. Развој електронских рачунара је започео појавом револуционарних машина попут ENIAC-а (*Electronic Numerical Integrator Analyzer and Computer*) које су (поред бројних нових могућности које су пружале) биле изузетно сложени за управљање, а такође су имале врло ограничене нумеричке могућности као и изузетно малу меморију у поређењу са данашњим рачунарима. Са даљим развојем рачунарства, као и информационих технологија у целини, рачунари су постајали све моћнији како у смислу брзине извршавања операција, тако и у смислу количине података које могу процесирати. Развој рачунара и информационих технологија од 1945. године па до данас, може се посматрати као серију константних промена и усавршавања која су у свакој итерацији мењале суштину самог рачунара. Рачунари су прешли пут од најобичнијег научног рачунског помагала, преко машина за обраду свих врста података, до уређаја који је у својим различитим облицима постао неизоставан део живота свих нас [2].

Упоредо са развојем рачунара, мењао се и тип, механизми чувања и обим података које су обрађивали. Од бушених картица које су се користиле за похрањивање информација у најстаријим верзијама електронских рачунара, до савремених метода за смештање података „у облаку“ (*cloud storage*). У данашње доба, услед описаног развоја рачунара и информационих система у целини, количина података која се може забележити постаје изузетна. Дакле, главни изазов више није само где сместити

постојеће податке које је нека организација прикупила, већ је још већи изазов постала изузетно велика брзина генерисања нових података које организације могу прикупљати. Такође, поставља се велико питање на који начин проверавати квалитет таквих података (како валидирати велику количину података у реалном времену), као и на који начин их даље обрађивати како би се на основу те велике количине података могло доћи до неких информација које могу послужити за доношење информисаних одлука. Дакле, већина савремених научних и индустријских институција су почеле да губе контролу над великом количином података које прикупљају а које не могу да преточе у нове информације, тј. показатеље који, на пример, могу помоћи пословању фирме, или неком истраживању научног центра.

За превазилажење ових изазова коришћене су нове методе попут науке о подацима (*data science*) и анализе великих података (*big data analysis*). У литератури постоји више дефиниција термина велики подаци (*big data*). Аутори рада [3] под великим подацима подразумевају способност друштва да искористи податке на нов начин у циљу доласка до нових закључака који могу донети значајну додатну вредност. Са друге стране, [4] тврди да је немогуће дати поједностављену дефиницију таквог појма, већ је могуће дати само неке основне карактеристике података који се могу сматрати великим подацима. Те особине великих података, између осталих, подразумевају податке са великог броја различитих извора, податке великог броја различитих типова, податке које у целини није могуће сместити у традиционалне релационе базе, као и податке чија је брзина генерисања изузетно велика. Примери великих података би били подаци сакупљени путем великих сензорских мрежа (попут рутера код кабловског оператора), подаци о претраживању интернета, подаци о кретању цена на берзи, сви видео записи на платформи *YouTube* и слично. Велика подаци могу бити структуирани или неструктуирани, то јест потенцијално се могу уклопити у „стандардне“ парадигме релационих база. Процена дата у [3] оцењује да је само око 5% свих дигиталних података структуирано, а да је 95% података (укључујући странице на интернету и видео записе) неструктуирано. Ипак, кључна парадигма која је подвукла важност података је можда најбоље описана у цитату америчког статистичара Деминга Едвардса (Edwards Deming) *„In God we trust; all others must bring data.“* („У Бога верујемо, сви остали морају донети податке“). Ова парадигма описује колико је нова оријентација у пословању заснована на подацима добила предност у односу на доношење одлука на основу инстинкта, осећаја или чистог доменског знања експерата [5].

Са друге стране, концепт науке о подацима је често изједначаван са термином статистике и посматран је као њена пука примена на подацима. Заправо, наука о подацима се треба посматрати као механизам екстракције знања из велике количине сирових података којима нека институција располаже. Дакле, наука о подацима би подразумевала коришћење математике, статистике, машинског учења (*machine learning*), вештачке интелигенције (*artificial intelligence*), база података, али и дубоко разумевање начина формулисања проблема као и теоријско познавање материје којом се нека институција бави [5]. Један од главних елемената науке о подацима је и откривање информација и нових закључака из података (*knowledge discovery*). Постоји разлика између парадигме откривања закључака науке о подацима и класичне анализе података. Са једне стране, класична анализа података се базира на анализи који подаци задовољавају неки феномен (односно образац понашања) који је примећен у природи. Насупрот томе, наука о подацима се бави проналажењем образаца у великој количини података. Дакле, циљ науке о подацима је проналажење образаца које дефинишу подаци,

али узимајући у обзир да ти обрасци морају бити такви да је њихово појављивање у будућности врло вероватно и очекивано. Машинско учење је управо задужено за лоцирање ових образаца. При машинском учењу, рачунар крајњем кориснику даје могућност да, уколико је рачунару постављено право питање (тј. уколико му је омогућен приступ великој количини одговарајућих улазних података), пружи неке интересантне одговоре које би човек – аналитичар база података – тешко приметио управо услед преобимних улазних података. Наравно, треба имати у виду да једино подаци који имају добру предиктивну моћ могу бити коришћени за генерисање предиктивних модела, дакле пука способност података да опишу резултате из прошлости није довољна. Додатно, нису увек сви обрасци које би рачунар методом машинског учења пронашао употребљиви. Познат је пример илустрован у [6] где је наведено да је у неком хипотетичком граду примећено да са порастом продаје сладоледа долази до повећања броја кривичних дела. Иако би на први поглед ова два феномена деловала потпуно неповезани, лоциран је одговарајући образац понашања у доступним подацима. Наравно очигледно је да мора постојати неки додатни разлог за појаву оваквог обрасца, на пример доба године. Прецизније, у данима када је лепо време више људи ће се налазити на улици, па самим тим може доћи до више уличних крађа. Наравно, логично је да се лети продаје више сладоледа па је самим тим образац понашања који је пронађен машинским учењем неупотребљив за неку будућу предикцију. Овакви примери додатно илуструју колико је доменско знање, тј. дубинско познавање података и њихових релација, битно у науци о подацима. Ипак, и поред оваквих недостатака, машинско учење и рачунари сами по себи имају многе предности при раду са (великим) подацима, у смислу цене, прецизности и скалабилности. Прелаз на одлуке које доносе машине је прво примењен управо у најосетљивијим областима које су захтевале брзе одлуке на основу доступних информација и раних упозорења. Пример би свакако било берзанско пословање где рачунари на основу података доносе одлуке о продаји или куповини у делићу секунде, чим стигне нова битна информација.

Један интересантан пример комбиновања метода науке о подацима и анализе великих података је дат у [3]. Описан је пример развоја математичког модела заснованог на подацима које прикупља Гугл (*Google*) који је за циљ имао предикцију појаве грипа код корисника. Наиме, Амерички центар за контролу и превенцију болести (*Center for Disease Control and Prevention* – CDC), је имао проблем да благовремено упозори јавност на појаву избијања сезонске епидемије грипа. Наиме, узимајући у обзир да је CDC имао приступ искључиво подацима из америчког здравственог система које није имао могућности да обрађује у реалном времену, CDC је давао упозорења о епидемији сезонског грипа тек недељу или две након стварног избијања епидемије што је изузетно велики период код болести које се тако брзо шире попут грипа. С друге стране, инжењери из Гугла су имали приступ подацима о историјским интернет претрагама људи, као и о ранијим епидемијама које су се јављале на територији Америке и успели су да пронађу корелацију између ових променљивих. Конкретно, Гугл је анализирао 50 милиона најфреквентнијих претрага на њиховом претраживачу и упоредио их је са званичним подацима о епидемијама грипа у периоду од 2003. до 2008. године. Идеја је била заснована на томе да се провери да ли људи заражени грипом са већом вероватноћом претражују тачно одређене појмове (као што су на пример: „лек за кашаљ“, „адреса медицинске установе X“ и слично). Формирано је 450 милиона различитих статистичких и математичких модела од којих је један дао резултат који је био изнад очекивања. Наиме, један модел је пронашао јаку корелацију између 45 конкретних термина претраге и избијања епидемије грипа. На тај начин, добили су

могућност да раније алармирају здравствене власти у случају избијања епидемије грипа, с обзиром да је њихов алгоритам могао да у реалном времену процени појаву епидемије.

Постоје различити механизми машинског учења који се могу користити у науци о подацима. Основна подела би била базирана на надгледане (*supervised*) и ненадгледане (*unsupervised*) алгоритме [7]. Избор типа алгоритма умногоме зависи од врсте података који су доступни. У основи, постоји две врсте података који се користе при машинском учењу: означени (*labelled*) и неозначени (*unlabelled*). Надгледани алгоритми машинског учења су базирани на означеним подацима, док су ненадгледани алгоритми базирани на неозначеним подацима. Основна разлика између означених и неозначених података се најбоље може показати конкретним примером. Нека је циљ који је постављен дефинисање модела машинског учења који треба да предвиди да ли ће неки клијент банке успешно отплатити кредит. Подаци који би били интересантни за овакав тип анализе би свакако били финансијско стање клијента, статус запослења и слично. Посматрајући кретање тих података код клијената који су до сада успешно, односно неуспешно, вратили кредит банци могло би послужити као означени скуп података у коме би лабела била управо исход – клијент (ни)је вратио кредит. Дакле, овај скуп података би се понашао попут учитеља који би надгледао рад рачунара (то јест ученика). Уколико би алгоритам који би рачунар поставио неисправно предвидео исход (успешно враћање кредита банци), подаци би сигнализирали погрешан прорачунати исход и захтевали би промену математичког модела све до тренутка до када би исход био исправно прорачунат. Са друге стране, неозначени подаци представљају скуп података који немају унапред очекивани исход. На пример подаци који се односе на цео скуп артикала које ће купити купци у радњи нису означени – не постоји унапред дефинисан исход односно очекивана група артикала. Анализом података коришћењем неког алгоритма ненадгледаног машинског учења може се закључити да унутар свих анализираних куповина постоји већа група клијената који купују само основне потрепштине (хлеб, млеко и слично), друга група која је фокусирана на луксузнију робу, трећа група клијената која купује велику количину робе одједном (месечна набавка), и слично.

Битно је дефинисати још једну поделу алгоритама машинског учења: алгоритми могу бити предиктивни или дескриптивни [7]. Предиктивни алгоритам машинског учења представља модел који покушава да на основу познатих претходних исхода предвиди појаву истих исхода у будућности. Насупрот њима, дескриптивни алгоритми машинског учења за циљ имају да дефинишу групе опсервација које се понашају на исти начин у смислу кретања вредности података који их описују. Дакле, они се баве само груписањем, а не предвиђањем самим по себи. У највећем делу случајева, надгледано машинско учење се користи за генерисање предиктивних модела, али то не мора увек бити случај. На пример, посматраћемо горенаведени пример о враћању кредита банци као пример надгледаног учења са модификацијом циља алгоритма. Модификовани циљ надгледаног алгоритма машинског учења би био да се дефинишу три групе клијената које ће описивати да клијенти имају исподпросечну, просечну или надпросечну вероватноћу да ће успешно вратити кредит банци. У том случају би основни принцип био идентичан, алгоритам би предвиђао вероватноћу да ће конкретни клијент банке вратити кредит, а потом би се одрадило груписање резултата. Пример груписања би био следећи:

1. клијенти који имају до 33% вероватноће да ће успешно отплатити кредит би били маркирани као исподпросечни,

2. клијенти који имају између 33% и 66% вероватноће да ће успешно отплатити кредит би били маркирани као просечни,

3. клијенти који имају преко 66% вероватноће да ће успешно отплатити кредит би били маркирани као надпросечни.

Са друге стране, ненадгледани алгоритми машинског учења су махом дескриптивни (попут описаног примера анализе потрошачких корпи клијената). Ипак, и ненадгледани алгоритми могу бити предиктивни. Један пример ненадгледаног алгоритма машинског учења биће коришћен и детаљно представљен у наставку рада. Тај алгоритам је, поред машинског учења, базиран на концептима теорије графова и анализе друштвених мрежа у мобилним телекомуникацијама. Циљ предложеног алгоритма је генерисање модела за предикцију вероватноће да ће клијент напустити мобилног телекомуникационог оператора.

Наиме, услед засићења телекомуникационог тржишта, за мобилне телекомуникационе операторе широм света постало је витално да увек имају актуелан и свеж увид у динамику њихових клијената. Управо из тог разлога анализа друштвених мрежа и њене примене са теоријом графова могу бити веома корисне. Предмет истраживања које ће бити представљено у докторској дисертацији управо представља анализу друштвене мреже које је приказана помоћу графа који је формиран на основу свих позива у мрежи мобилног телекомуникационог оператора у периоду од месец дана. Чворови формираног графа биће предмет анализе кластера која ће бити извршена на основу структуралних метрика чворова унутар графа. Конкретно, у питању ће бити метрике: степен чвора, улазни степен чвора, излазни степен чвора, значајност чвора првог реда, значајност чвора другог реда, сопствени вектор чвора, вредност ауторитета чвора, *hub* вредност чвора и мера артикулације чвора. Истраживање ће показати да је могуће идентификовати неке битне чворове унутар графа (то јест друштвене мреже), који ће бити кључни за предикцију вероватноће да ће клијент напустити посматраног мобилног телекомуникационог оператора. Штавише, биће показано да уколико клијент (који представља чвор графа) напусти посматраног мобилног телекомуникационог оператора, други клијенти који често комуницирају са њим такође могу бити подложни да напусте посматраног мобилног телекомуникационог оператора. Дакле, модел који ће бити представљен у истраживању ће, на основу постојећих корисника који напуштају посматраног мобилног телекомуникационог оператора и њихових образаца комуникације, проактивно предвиђати нове клијенте који имају високу вероватноћу напуштања посматраног мобилног телекомуникационог оператора. *Lift* метрика ће бити коришћена за квантификовање резултата истраживања. Резултати предложеног истраживања су довољно општи да могу одмах бити употребљени у било којем пољу где се пријатељске или хомофилне везе могу посматрати као потенцијални узрочник осипања и умањења броја клијената.

Научни доприноси који ће проистећи из докторске дисертације су следећи:

1. Проучавање и опис широког спектра метрика чворова у графу на основу којих се могу открити структуралне карактеристике графа сачињеног од свих позива унутар мреже посматраног мобилног телекомуникационог оператора.

2. Дефинисање три групе метрика (мере за неусмерене графове, мере за усмерене графове и мера артикулације) које ће прецизно описати сваки елемент графа сачињеног од свих позива унутар мреже посматраног мобилног

телекомуникационог оператора; биће показано да, иако је граф мобилне телекомуникационе мреже по својој природи усмерен (услед природног усмерења позива од једног броја ка другом), мере за неусмерене графове такође могу пружити значајне додатне увиде који могу резултовати предвиђањем вероватноће губитка клијента.

3. Развој новог метода базираног на примени теорије графова за откривање неважећих чворова (чворова који не одговарају људима, већ телемаркетинг центрима, или корисничким сервисима) графа сачињеног од свих позива унутар мреже посматраног мобилног телекомуникационог оператора.

4. Развој модела базираног на реалним подацима који може пружити важан увид у сферу мобилног телекомуникационог пословања и може послужити као водич за осмишљавање нових и бољих стратегија за борбу против губитка корисника мобилних телекомуникационих оператора.

5. Дефинисање алгоритма за спајање резултата две засебне методе кластеризације (које су базиране на метрикама за неусмерене графове, односно метрикама за усмерене графове) у унифицирану скалу која ће представљати финални модел за предвиђање губитка корисника у мобилним телекомуникационим мрежама.

6. Због ограниченог обима основних променљивих које су анализиране у докторској дисертацији (само број и трајање позива), могуће је комбиновати предложени метод са другим савременим и квалитетним решењима базираним на другим променљивама (на пример, са моделима базираним на задовољству корисника и / или на преосталом времену трајања уговора корисника) и на тај начин добити још боље финалне резултате.

Наставак рада је конципиран на следећи начин: у поглављу 2 је дат преглед повезане литературе; поглавље 3 садржи детаљнији опис основних карактеристика науке о подацима и њених примена у науци и индустрији; поглавље 4 пружа детаљнији увид у технологије ненадгледаног машинског учења и конкретно на метод кластеризације; у поглављу 5 је дат опис примена теорије графова у анализи друштвених мрежа; поглавље 6 даје приказ истраживања, развоја као и резултате пројекта који је био базиран на примени науке у подацима у мобилним телекомуникацијама ради предикције вероватноће да ће клијент напустити мобилног телекомуникационог оператора; на крају поглавље 7 садржи закључак рада.

2. ПРЕГЛЕД ПОВЕЗАНЕ ЛИТЕРАТУРЕ

Широка распрострањеност и доступност мобилних телекомуникација је бројним научним радницима пружила могућност њихове исцрпне и детаљне анализе. Предикција губитка корисника је један од најбитнијих изазова за сваког мобилног телекомуникационог оператора. Један приступ за решавање овог проблема подразумева искоришћавање информација које се могу добити моделовањем интеракција између клијената која је представљена у [8]. Систем представљен у раду [8] има могућност да превазиђе ограничења метода базираних на анализи друштвених мрежа који, ради предвиђања будућег губитка корисника, захтевају информацију који клијенти су иницијално напустили мобилног телекомуникационог оператора. Остварена вредност *Lift* метрике за најугроженијих десет процената популације је једнака 2.5.

У истраживању представљеном у раду [9], аутори су анализирали неколико изазова који се могу појавити при комбинацији ненадгледаних метода кластеризације и *boosting* стабла одлучивања. Анализирано је укупно пет техника анализе кластера и максимална остварена вредност *Lift* метрике за најугроженијих десет процената популације је приближно једнака 2.6.

Једноставан али ефикасан приступ базиран на дифузији који је заснован на анализи друштвених веза ради идентификације значајног броја изгубљених корисника је представљен у раду [10]. У том раду аутори одређују утицај корисника највише на основу броја и трајања позива између претплатника. Остварена вредност *Lift* метрике за најугроженијих десет процената популације је једнака 5.

У раду [11] аутори предлажу технику за предвиђање губитка корисника базирану на позивима клијената (конкретно, обрасцима позива и њиховим променама у времену) и њиховим информацијама о уговорним обавезама. Предиктивни модел дефинисан на основу тих података је омогућавао остваривање вредности *Lift* метрике за најугроженијих десет процената популације од 5.4 (унутар периода од недељу дана између дефинисања модела и предвиђања напуштања оператора).

Систем представљен у раду [12] остварује вредност *Lift* метрике за најугроженијих десет процената популације који је једнак чак 10, коришћењем демографије клијената, информацијама о наплати потраживања, статусу уговора, записима о оствареним позивима у претходном периоду, као и контактима са корисничким центром. Ове додатне информације пружиле су вредан додатни увид у понашање клијената, што је и осликано у изузетно високој вредности *Lift* метрике за најугроженијих десет процената популације. Са друге стране, битно је нагласити да је анализиран само јако мали број клијената (укупно око 160 000 клијената), који су имали изузетно ниску стопу напуштања телекомуникационог оператора (око 0.71%). Аутори су у раду напоменули да улазни скуп података није довољно велики за генерисање

доброг продукционог предиктивног модела унутар сваког сегмента корисника. Ипак, овај рад пружа важне увиде у будућем развоју предиктивних модела за предвиђање губитка корисника мобилних телекомуникационих оператора.

Модел за предикцију губитка корисника у три корака је представљен у истраживању [13]. Прва фаза подразумева надгледану методу избора променљивих, а друга дефинисање и имплементацију система базираног на знању (*Knowledge Based System* – KBS) коришћењем *Ripple Down Rule* (RDR). У последњем кораку, техника симулираног експерта (*Simulated Expert* – SE) је предложена за процену и евалуацију аквизиције знања у систему базираном на учењу. Иако у резултатима рада нису представљени прецизни резултати вредности *Lift* метрике за најугроженијих десет процената популације, показано је да RDR систем има прецизност до 95%, а да је стопа предикције истински позитивних изгубљених корисника једнака 73%. На крају, симулирани експерт је успешно рекласификовао све погрешно класификоване опсервације. Битно је нагласити да је истраживање представљено у [13] извршено на изузетно малом скупу улазних опсервација који је садржао само 3333 корисника.

Различита истраживања која су рађена на подацима о изгубљеним корисницима мобилних телекомуникационих оператора показују да перформансе предикције различитих модела осетно варирају у различитим зонама улазних скупова података. У таквим ситуацијама, може се приметити корелација између нивоа тачности класификатора и извесност његовог предвиђања. Истраживање приказано у раду [14] предлаже механизам који може проценити извесност предвиђања класификатора у различитим зонама улазног скупа података. Аутори истраживања представљају алтернативан приступ проблему предвиђања губитка корисника мобилних телекомуникационих оператора базиран на концепту извесности предикције коришћењем фактора растојања. Наиме, улазни скуп података је груписан у различите зоне на основу дефинисаног фактора растојања. Све добијене зоне су накнадно подељене у две категорије: подаци са великом извесности напуштања мобилног телекомуникационог оператора, и подаци са ниском вероватноћом напуштања мобилног телекомуникационог оператора.

Проблем кластеризације података са пуно променљивих, попут проблема представљеног у истраживању које је окосница ове докторске дисертације, може бити решен и на алтернативни начин као што је и показано у раду [15]. Предложено решење у истраживању [15] је базирано на алгоритму *K*-средина на два нивоа. У првом кораку, аутори предлажу примену алгоритма *K*-средина са анализом силуета за сваку променљиву ради проналажења оптималног број кластера у подацима као и за саму кластеризацију података. Битно је нагласити да је свака променљива понаособ анализирана и кластерисана. У другом кораку, аутори врше трансформацију сваке опсервације на начин да свака променљива неке опсервације добија вредност центроида свог најближег кластера конкретне променљиве. Финални корак подразумева поновну примену $K-means$ кластеризације са анализом силуета за све трансформисане опсервације ради генерисања финалних кластера. Треба имати у виду да је ова кластеризација у два нивоа уведена да би се елиминисали изузеци у подацима, као и да би се уклониле екстремне вредности неких променљивих. Са друге стране, у истраживању представљеном у дисертацији постоји засебан механизам за елиминисање неважећих записа који не представљају реалне кориснике мобилне телекомуникационе мреже. Такође, истраживање [15] је базирано на предикцији губитка корисника мобилних друштвених апликација. Корисници таквих апликација имају драстично

другачије понашање у односу на кориснике мобилних телекомуникационих оператора, те се самим тим закључци представљени у раду [15] не могу директно применити на истраживање приказано у овој докторској дисертацији. Наиме, корисници мобилних друштвених апликација у највећем броју напуштају апликацију након првих неколико дана коришћења, што ни приближно није случај са претплатницима мобилних телекомуникационих оператора.

У истраживању које је представљено у раду [16], аутори предлажу сегментацију базирану на анализи друштвених мрежа ради проналажења значајних корисника, то јест клијената који имају висок утицај на своју друштвену мрежу контаката. Аутори су користили две метрике (степен чвора, као и ниво повезаности – изведену променљиву чија је основа број позива клијената и степен чвора) и дефинисали четири сегмента. Персонализовањем посебних понуда које су упућене значајним клијентима, мобилни телекомуникациони оператор је желео да подстакне утицајне кориснике да проширују свој утицај и величину њихових друштвених група. Резултати истраживања су показали да је оваквим акцијама стопа губитка корисника смањена (конкретне бројке нису представљене), док се потрошња претплатника осетно повећала.

У раду [17], аутори су описали дифузиони модел предиктивне шеме који је примењив на једног корисника, или на малу групу корисника. Модел уводи елементе друштвених наука у дифузиони алгоритам за предвиђање ширења енергије. Унапређења дифузионог модела укључују модификацију иницијалне енергије корисника коришћењем теорије друштвених наука, и модификацију самог алгоритма дистрибуције енергије укључивањем теорије друштвених статуса. Резултујући модел је остварио вредност *Lift* метрике за најугроженијих десет процената популације од 2.5.

Рад [18] пружа преглед различитих техника за *oversampling* података, ради превазилажења проблема небалансираних скупова података. Скупови података нису балансирани када је број појављивања једног исхода драстично мањи од броја појављивања других исхода. Пример оваквих небалансираних скупова података може представљати скуп свих корисника неког мобилног телекомуникационог оператора од којих је само мали број одлучио да промени оператора. Технике које су анализиране у раду [18] су *Mega-trend Diffusion Function* (MTDF), *Synthetic Minority Oversampling Technique* (SMOTE), *Adaptive Synthetic Sampling approach* (ADASYN), *Couples Top-N Reverse k-Nearest Neighbor* (TRkNN), *Majority Weighted Minority Oversampling Technique* (MWMOTE), као и *Immune centroids oversampling technique* (ICOTE). Укупно четири скупа података корисника неког мобилног телекомуникационог оператора је представљено и анализирано у овом раду, а емпиријски резултати су показали да је укупна предиктивна моћ највећа при коришћењу MTDF технике за *oversampling* улазних података.

Истраживања приказана у радовима [19], [20], [21] представљају покушај решавања проблема предикције губитка корисника мобилних телекомуникационих оператора коришћењем *rough set* теорије. *Rough set* теорија представља технику за доношење одлука и њена основна примена представља анализу података и класификацију непрецизних или несигурних информација. Радови [19], [20], [21] представљају сличне резултате у којима је показано да је генерички *rough set* алгоритам најоптималнији за предикцију губитка корисника, на основу анализираних података. Штавише, применом предложеног алгоритма, аутори остварују максималну прецизност при предвиђању корисника који ће напустити посматраног мобилног телекомуникационог оператора. Ипак, потребно је имати у виду да су реалне вредности

стопе губитка корисника значајно ниже од оних које су представљене у овим радовима, а такође и анализиране групе података нису претерано велике (у радовима је коришћен јавно доступан скуп података од 3333 корисника са стопом губитка корисника од приближно 15%).

Ефекат утицаја пријатељстава и хомофилије на стопе губитка корисника у мрежама мобилних телекомуникационих оператора је анализиран у раду [22]. Случајни узорак од 10 хиљада претплатника је анализиран у периоду од августа 2008. до маја 2009. године. Аутори тврде да се код неког корисника мобилних телекомуникационих мрежа вероватноћа напуштања оператора повећава у случају када пријатељи анализираног корисника промене телекомуникационог оператора. При томе, као пријатељи неког корисника су издвојени претплатници са којима је посматрани корисник комуницирао барем један, три или пет пута у току једног месеца. Закључак рада [22] тврди да губитак једног пријатеља из мреже мобилног телекомуникационог оператора доводи до повећања вероватноће губитка корисника за чак 24%.

Рад [23] представља комбинацију традиционалног приступа проблему губитка корисника мобилних телекомуникационих оператора и анализу друштвених мрежа. У истраживању аутори користе алгоритам базиран на анализи друштвених мрежа као наставак класичног модела ради увођења мултидимензионе анализе, као и да би се поспешила стопа прецизности целокупног система за предикцију губитка корисника мобилног телекомуникационог оператора. Конкретно, употреба анализе друштвених мрежа у случајевима када традиционални модел није давао задовољавајуће резултате (код клијената који имају изразито ниску или изузетно високу стопу позива) је довела до повећања укупне прецизности система за 10% у случају припејд корисника, односно 8% у случају постпејд корисника.

У истраживању које је представљено у раду [24], аутори предлажу свеобухватни приступ који за циљ има дефинисање најприкладније технике машинског учења за анализу графова позива ради предвиђања губитка корисника. Предложени модел је базиран на скалабилном *node2vec* алгоритму [25] који омогућава извлачење значајних информација о структури веома великих графова. Такође, коришћена је и метода за истовремену анализу интеракција клијената, као и њихових структуралних карактеристика у времену анализом карактеристика мрежног графа. Узимајући у обзир да анализирани подаци нису садржали информацију о томе који су клијенти напустили посматраног оператора, а који не, губитак корисника је дефинисан као неактивност корисника у унапред дефинисаном периоду. На основу ове дефиниције, предложени модел базиран на логистичкој регресији је остварио вредности *Lift* метрике за најугроженијих 0.5% популације који варира између 2 и 4 (у зависности од конкретне конфигурације, односно типа модела – то јест од тога да ли је модел намењен за припејд или постпејд кориснике).

Експеримент у коме су анализиране две технике: логистичка регресија и *Logit Boost* је представљен у раду [26]. Истраживање је користило скуп података који је садржао 3333 опсервације са стопом губитка корисника од приближно 15%, и остварени су високи резултати из угла прецизности: логистичка регресија је остварила укупну прецизност од 85.23%, док је *Logit Boost* модел имао укупну прецизност од 85.17%.

Укратко сумирано, већина радова који су представљени у повезаној литератури остварују вредности вредност *Lift* метрике за најугроженијих десет процената популације у опсегу између 2 и 5. У поглављу 6.3.3. биће представљени резултати

истраживања оствареног методом предложеном у овој докторској дисертацији. Приказани резултати ће потврдити да предложени систем представља добру и обећавајућу полазну основу за даљи развој модела за предикцију губитка корисника у мрежама мобилних телекомуникационих оператора.

3. НАУКА О ПОДАЦИМА: ПРЕГЛЕД И ПРИМЕРИ

У савременом свету, подаци се константно генеришу, прикупљају и преузимају. Свака акција коју човек може извршити ће неизоставно довести до генерисања неког записа података. Почевши од куповине платном картицом, коментара на фејсбуку, „лајка“ на инстаграму, коментара на твитеру, гледања садржаја на јутјубу... На све ово треба додати и податке које генеришу паметни уређаји, сензорске мреже, паметни аутомобили, али и податке генерисане од стране организација попут банака, авио-компанија, телекомуникационих оператора, универзитета, научних центара... Поред обима података, брзина којом се генеришу је постала велики проблем у последњој деценији. За институције је постало кључно да покушају да добију највећу додатну вредност из података које сакупљају како би могле да донесу одлуке базиране на информацијама а не само на основу полуинформација или осећаја. Да би се дошло до тих информација, неопходна је наука о подацима.

Као што је већ напоменуто, наука о подацима се може дефинисати као систематско проучавање особина неке организације и података које она прикупља, анализа тих података, доношења закључака и примена тих закључака као подршка при доношењу информисане одлуке [5]. Вредност науке о подацима у савременом свету у последњој деценији најбоље илуструје појава експлозије података (*data explosion*) [27]. Индикативни су подаци да је 2012. године процењено да ће се сваке године количина података који се прикупљају у свету бити дупло већа из године у годину [28], или да ће се седмоструко повећати пренос мобилних података у периоду од 2017. до 2022. године [29]. Иако би се могло претпоставити да ће се институције изгубити у оволикој количини података, најспособније компаније из потпуно различитих индустријских грана су врло брзо увиделе додатну вредност коју ови подаци могу донети [30]. Пословање и доношење одлука које је усмерено подацима (*data-driven business*) [31] је постало циљ разних компанија из области едукације, трговине, финансија, индустрије, медицине, транспорта али и државних управа широм света. Посебно у последњих неколико година је примећен брз раст интересовања за екстракцију корисних информација из сирових података. Ипак, да би се дошло до информације која се може посматрати попут „грумена злата у руднику“, потребно је прво одрадити „рударски“ посао, то јест издвојити информацију. Ти кораци при раду са подацима се могу груписати у следеће основне групе: прикупљање података, дефинисање модела података, трансформација података и њихово смештање у модел, препроцесирање података, моделовање и интерпретација резултата.

У оквиру докторске дисертације представљене у овом раду, наука о подацима заузима централну позицију. Процес реализовања пројекта коришћењем науке о подацима описан у овом поглављу је суштински идентичан поступку који ће бити примењиван у истраживању чији су резултати представљени у шестом поглављу.

3.1. Историјат развоја науке о подацима

Давно пре увођења термина науке о подацима и машинског учења (које представља један од стубова науке о подацима), у филозофији су уведени теоријски концепти индукције и дедукције [7]. Под индукцијом се подразумевала генерализација знања добијеног на основу анализе једног сета примера, и примена тих знања на свим сличним примерима. Са друге стране дедукција је представљала примену општих концепата на конкретним примерима. Најједноставнији пример разлике између дедукције и индукције би се могао описати коришћењем аналогије ученика и учитеља. Наиме, ако учитељ представља нове опште концепте ученику и након тога ученик може самостално решавати проблеме користећи упутство које му је задао учитељ, тада је у питању дедуктивни метод учења. Насупрот првом примеру, уколико учитељ демонстрира ученику серију примера који ученику увек приказују исти исход тада је у питању пример индуктивног учења. Тада би се од ученика очекивало да он сам уочи образац, односно принцип по коме се долази до исхода. Ипак, индуктивно учење има одређене мањкавости, пре свега у смислу уочавања образаца. Конкретно, образац који је примећен не мора бити довољно општи да се може применити на сваком примеру. Постоји пуно верзија такозваног проблема индукције. Једну од верзија је дефинисао шкотски филозоф Дејвид Хјум (*David Hume*). Он је тврдио да је једино оправдање за коришћење метода индукције, индуктиван сам по себи: наиме, пошто индукција даје резултате за неке индуктивне проблеме, онда ми претпостављамо да она важи за све индуктивне проблеме [32]. На овај начин он тврди да се индукција не може дедуктивно доказати, већ да се она доказује самом индукцијом, што у неку руку баца још већу сумњу на њену исправност.

Сличан проблем представља и теорема да „нема бесплатног ручка“ (*no free lunch theorem*) у којој се тврди да ниједан алгоритам не може бити бољи од неког другог уколико би се при провери њихових перформанси узеле у обзир све могуће варијације улазних параметара [7], [33]. Самим тим, тврди се да прецизност било ког алгоритма није ништа боља од пуког насумичног погађања. Као пример се може посматрати класичан пример наставка низова, на пример питање који број би требало да следи након бројева 1, 2, 4, 8, ...? Уколико је било која секвенца бројева подједнако вероватна онда нема могућности да се може дати бољи одговор од простог случајног покушаја. Наравно, неке секвенце су вероватније од других. Слично, дистрибуција проблема машинског учења у стварном свету је крајње неуједначена. Дакле, да би се избегао проблем теореме „бесплатног ручка“, потребно је да постоји добро познавање дистрибуције могућих резултата. Такође, потребно је и познавање карактеристика те дистрибуције при избору адекватног алгоритма машинског учења.

Са развојем рачунара, могућност употребе рачунара за решавање проблема методом индукције је постајала све већа и већа. Моћније конфигурације рачунара, су пружиле и инспирацију за дефинисање нових метода које би ефикасније могле да дођу до решења. Ипак, иако су рачунарски ресурси који су данас доступни неупоредиво већи него пре само десетак година, а управо услед изузетног обима података који се обрађују, потреба за ефикасним коришћењем ресурса је постала никад већа. Више није довољно само доћи до одговора на основу података, већ доћи до одговора на што ефикаснији начин. У паралели са развојем рачунара, постојао је тренд да се рачунарима обезбеди механизам самосталног размишљања (попут људских бића). На пример, изучавање начина рада људског мозга коришћењем неуралних мрежа је предмет разних студија још од четрдесетих година претходног века. Сам термин машинског учења је настао у овом

контексту и то као област изучавања која рачунару омогућава да учи, без потребе да се то учење експлицитно испрограмира [34]. Деведесетих година двадесетог века дошло је до популаризације новог термина „рударења података“ (*data mining*) управо из потребе да се подаци (које су компаније почеле масовније да прикупљају услед појефтињења система за њихово похрањивање) преточе у информације. Главна идеја рударења података је била да се била каква корисна информација извуче из (за то време) велике количине података. Са почетком двадесет првог века уведен је и термин велики подаци (*big data*) и то преко помоћу концепта три В (волумен, варијабилност и брзина података –*three V: volume, variety, velocity*), а након неког времена и термин наука о подацима као механизам за извлачење знања из великих података.

3.2. Процес реализовања пројекта коришћењем науке о подацима

Када се уопштено посматра неки пројекат науке о подацима, подразумевано је да постоји одређени циљ истраживања, као и скуп променљивих (података) који ће послужити као улазне информације за остваривање циља. На пример, циљ може бити предвиђање кретања параметара пословања неког предузећа, а променљиве помоћу којих би се ово кретање предвидело у будућности могу бити историјска кретања истих показатеља, затим предвиђени раст продаје, оперативни трошкови, трошкови развоја и слично. Уопштено посматрано, основни елементи при процесу реализације пројекта базираног на науци о подацима су [35]:

1. Пажљива дефиниција циља пројекта – дефинисање основних параметара пројекта, као и карактера пројекта;

2. Прикупљање података – проналажење и приступ подацима (интерним или екстерним) потребним за реализацију пројекта;

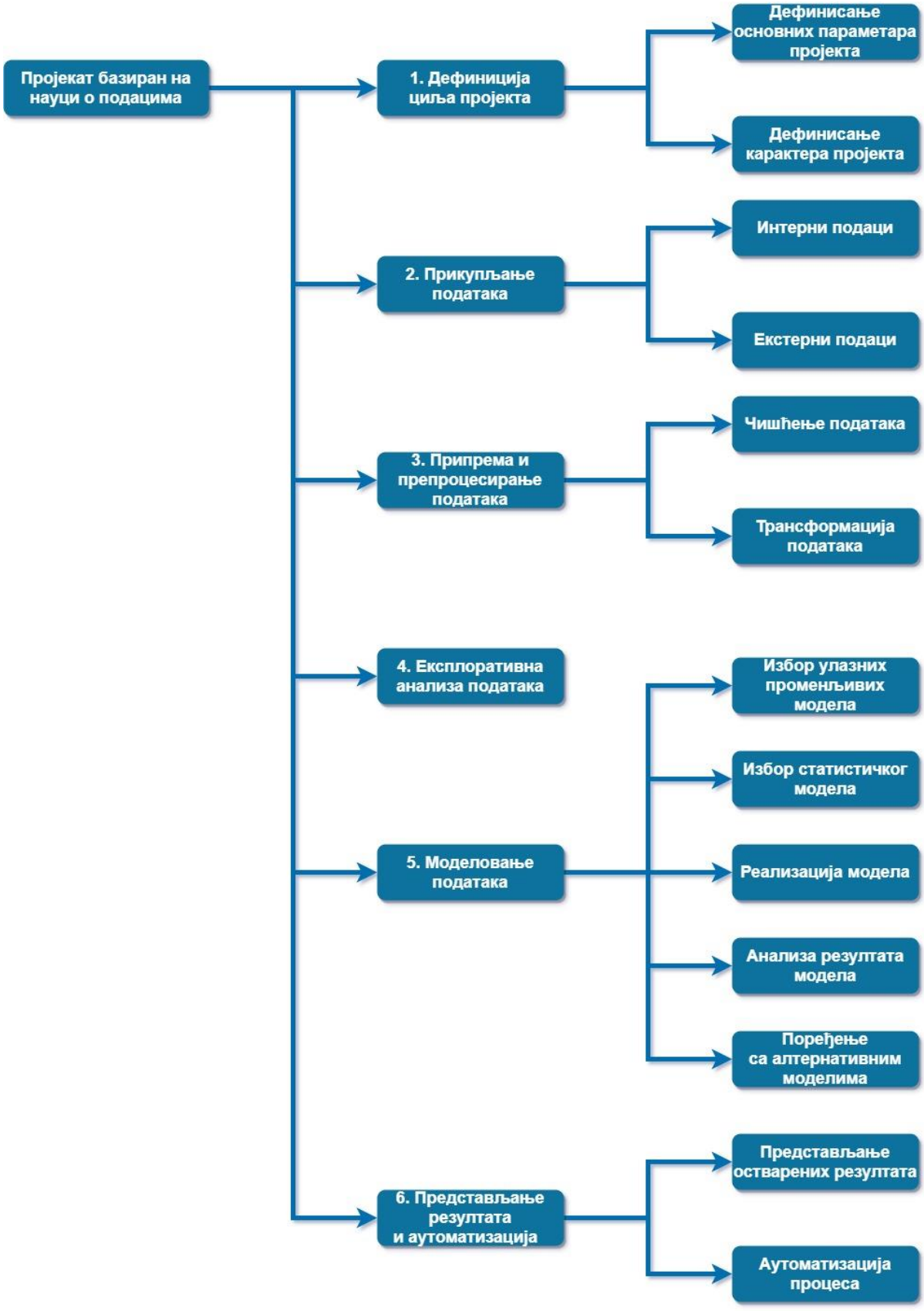
3. Припрема и препроцесирање података – провера и исправљање грешака у подацима ради припреме за моделовање;

4. Истраживачка (експлоративна) анализа података – прибављање детаљнијих информација о подацима коришћењем дескриптивне статистике и визуелних метода;

5. Моделовање података – достизање циљева пројекта коришћењем метода машинског учења и статистике;

6. Приказ резултата и аутоматизација процеса – представљање резултата заинтересованим странама и индустријализација процеса анализе за поновну употребу и интеграцију са другим алатима.

Шематски приказ дат на слици 3.1. сумира основне кораке при реализацији пројекта базираног на науци о подацима.



Слика 3.1: Основних шест корака при реализацији пројекта базираног на науци о подацима

Први корак, дефиниција циља пројекта, је кључан зато што би било какав проблем у овом првом кораку могао довести до краха целог пројекта. Најбитнија питања која се морају поставити пре почетка пројекта базираног на науци о подацима су: шта је захтевани резултат пројекта, зашто је пројекат започет и како се пројекат уклапа и „ширу слику“, односно у рад организације која је покренула пројекат. Ради успешне реализације пројекта, потребно је дефинисати следеће елементе: јасан циљ пројекта, контекст пројекта, начин анализе података, доступне ресурсе расположиве за рад, доказ могуће реализације пројекта или доказ концепта, временски оквир реализације пројекта, као и метрике које ће потврдити (не)успех пројекта.

Прикупљање и консолидација података из различитих извора представља други корак при реализација пројекта базираног на науци о подацима. Као што је већ описано, велика количина података која је постала доступна је постала и предност и мана при реализацији реалних пројеката. Подаци који се користе на пројекту могу се грубо поделити на интерне (податке које прикупља и структуира сама организација која је иницијатор пројекта) и екстерне (податке које прикупља трећа страна која своје податке уступа организацији зарад реализације пројекта). Иако би се дало претпоставити да су интерни подаци једноставнији за рад, то не мора бити случај. Наиме, бројне су организације које имају више различитих силоса (складишта) података, тј. засебних система за похрањивање података за сваки одвојени структурни део организације. Тада би поступак консолидације свих тих интерних података похрањених у више одвојених целина могао бити врло мукотрпан задатак. Са друге стране, екстерни подаци имају велику вредност пошто могу омогућити драстично другачији поглед на исти ентитет у односу опис истог ентитета помоћу искључиво интерних података. Ипак, при раду са екстерним подацима се могу јавити проблеми квалитета података (недоступних информација, некомплетних записа и слично). Иако је детаљна анализа квалитета података неопходан елемент трећег корака при имплементацији пројеката базираних на науци о подацима (тј. припреми података), пожељно је да се већ у фази прикупљања података одради базична анализа квалитета података. Та базична анализа квалитета података би могла показати да ли су подаци из неког конкретног извора изразито проблематични и неодговарајући за анализу, те их потенцијално треба у потпуности искључити из даљег рада. Други проблем са екстерним подацима би могао бити механизам њиховог обједињавања са интерним подацима. Наиме, да би се успешно користили подаци из интерних и екстерних извора потребно је да се ентитети из једног система успешно пресликају на други. Неретко се може догодити да је немогуће извршити ово пресликавање те се самим тим неки вредан извор информација мора изузети. На пример уколико бисмо посматрали две корпорације, од којих је једна оријентисана ка малопродаји, а друга ка банкарству, могло би се очекивати да би постојао заједнички интерес размене података. На пример, циљ би могао бити да се прикупе додатне информације о финансијским навикама клијената банке који посећују конкретни малопродајни објекат. Ипак, уколико не би постојала нека јединствена идентификација клијената у једном односно другом систему које би се могле користити за мапирање (тј. спајање) информација, никакво обједињавање података не би било могуће. Рачуни у малопродајним радњама нису персонализовани (изузимајући плаћања платним картицама), те се никако не би могли повезати са клијентима банке па би овај потенцијално вредан извор екстерних информација био изгубљен.

Припрема података, трећи корак у процесу реализације пројекта базираног на науци о подацима, је најчешће и временски најзахтевнији сегмент пројекта. Он се састоји од чишћења података као и трансформације и комбиновања података у облик

који је подесан за конкретну анализу. Под чишћењем података се подразумева уклањање података који су, између осталог, настали услед:

1. погрешног уноса података (тј. грешака у куцању и слично),
2. уноса немогућих вредности (на пример старост особе од -20 година),
3. недостајућих вредности,
4. уноса изузетака (*outlier*) – података који својим вредностима драстично одударају од осталих података; такви подаци могу правити велике проблеме при моделовању зато што мали број изузетака може упутити модел на потпуно погрешан правац решења,
5. уноса података који нису дозвољени (на пример пол може бити мушки или женски, друге вредности се не прихватају као исправне за тај податак)...

Постоје различити механизми за унапређивање квалитета података. Уколико је количина неодговарајућих података мала, често се прибегава њиховом уклањању. Међутим, уколико је неопходно искористити све податке и њихово избацивање није дозвољено, постоје и алтернативне методе за чишћење података попут њихове замене. Нетачни подаци могу заменити на више начина. На пример, уместо недостајућих информација могуће је уметнути средњу вредност свих осталих доступних података, медијану унутар одређене групе... Конкретно, уместо недостајућег податка за тежину неког ђака, може се користити просек свих ђака у школи, или само просек свих ђака који иду у исто одељење са ђаком за кога нису доступни подаци.

Трансформација података подразумева групу акција које омогућавају генерисање скупова података који су погодни за моделовање. Под трансформацијом података се може сматрати агрегирање података, спајање података, генерисање изведених променљивих, смањивање димензионалности (смањивање броја променљивих у финалном скупу података) и слично. Такође, често је за неке статистичке моделе пожељно да су улазни подаци што приближнији нормалној (Гаусовој) расподели, па је пожељно извршити трансформацију променљивих (најчешће путем логаритамске функције или кореновањем).

Истраживачка анализа података подразумева преглед и проверу свих добијених променљивих и њихових вредности. Овај тип провера се најчешће ради визуелно, тј. генеришу се одређене визуелизације података које помажу при провери веродостојности променљивих које ће представљати улазне параметре математичког модела. Заправо, анализом и прегледом графика се много једноставније могу уочити неке неправилности које су потенцијално и даље присутне у подацима. Уколико такве неправилности постоје, неопходно је да се поново исправе на начин описан у претходној тачки.

Пети корак при реализацији пројекта коришћењем науке о подацима представља конкретно моделовање података. Овај корак се може посматрати као унија следећа три елемента:

1. избор улазних параметара модела и избор статистичког модела,
2. реализација модела,

3. анализа резултата модела и поређење са алтернативним конфигурацијама модела.

Избор променљивих које ће се користити за моделовање је битан корак при моделовању. Постоји два главна проблема који се могу јавити при погрешном избору променљивих које се укључују у модел.

Први проблем подразумева да променљиве које се укључују у модел могу бити ирелевантне за дефинисање модела. Овај тип анализе се може спровести на више начина. Први начин је свакако да се добрим експертским познавањем конкретне области којом се пројекат науке о подацима бави, донесе експертска одлука да се неке променљиве (не)требају укључивати у статистички модел. Други начин подразумева нумеричку анализу везе између улазних променљивих и очекиваних резултата и елиминисање променљивих које нису у корелацији са очекиваним резултатима. На ове начине се може уклонити „шум“ који би непотребне променљиве могле да формирају унутар модела.

Други проблем који се може јавити при погрешном избору променљивих је укључивање више улазних променљивих које су снажно повезане и стога могу имати несразмеран утицај на вредност очекиваних излазних величина. На пример, нека постоје четири променљиве које би требале да опишу једну излазну вредност, а три од четири променљиве су јако корелисане зато што их покреће иста основна концептуална појава. Тада би се (код већине модела) догодило да се ефекат те основне појаве драстично више наглашава, те би потенцијално та појава могла да диспропорцијално утиче на резултате у поређењу са преосталом четвртом (независном) улазном променљивом.

Избор статистичког модела такође има велики утицај на употребну вредност целог пројекта базираног на науци о подацима. Неке од битнијих ставки при избору статистичког модела за реализацију пројекта (поред прецизности и тачности) морају бити и: сложеност модела и његова имплементација на продукционом и тестном окружењу; сложеност одржавања модела; једноставност модела (велика је предност када се статистички модел може мање или више једноставно појаснити крајњим корисницима који могу, али и не морају имати довољно познавање статистике и математике).

Реализација модела представља следећи битан корак. Постоје бројни софтверски пакети који омогућавају имплементацију статистичких модела, попут: *MATLAB, Python, R, SAS, KNIME* и други. У зависности од дефинисаног циља пројекта, могу се користити надгледане или ненадгледане методе машинског учења. Неки од алгоритама који се користе при реализацији надгледаних метода машинског учења су: регресија (линеарна или логистичка), стабла одлучивања, неуралне мреже и слично. Са друге стране, алгоритми које се користе при реализацији ненадгледаних метода машинског учења су кластеризација, неки типови неуралних мрежа (попут оних базираних на Кохоненовим самоорганизујућим мапама (*Kohonen Self-organizing Map*)) и слично. Анализа главних компоненти (*Principal component analysis* - PCA) је још један пример технике базиране на ненадгледаном машинском учењу који се користи за редукцију димензија скупова података са великим бројем димензија. Више детаља о конкретним алгоритмима ненадгледаног машинског учења који се користе за реализацију модела биће дато у каснијим поглављима. Битно је само подвући да је, узимајући у обзир доступна хардверска и софтверска ограничења, потребно да реализовани модел има разумно време извршавања (разумно време извршавања може варирати у зависности од конкретне примене модела). Брзина извршавања модела која је захтевана и потребна зависи од начина примене модела. На пример, модели који се користе за обраду података

у реалном времену морају бити јако брзи и са одговарајућом хардверском подршком. Са друге стране, модели који су конципирани тако да се извршавају једном месечно, а да се њихови резултати примењују до следећег извршавања не морају бити нити брзи нити једноставни.

Анализа резултата модела је неопходан елемент при развоју сваког типа модела. Узимајући у обзир да је сврха модела генерализација (тј. примењивање историјских правила на новим подацима), провера перформанси модела користећи исте податке помоћу којих је сам модел формиран није могућа ни исправна. Дакле, да би се потврдили резултати модела, најчешће је потребно потврдити његове перформансе на неким подацима са којима се није сусретао у процесу „учења“, односно генерисања модела. Ово се најчешће постиже поделом улазног скупа података на три дела:

1. део за тренирање модела – део података на основу којих се одређују параметри модела;

2. део за валидирање модела – део података на основу којих се оптимизацију параметри;

3. део за тестирање перформанси – део података који ће се користити за непристрасну оцену коректности модела.

При реализовању пројекта базираног на машинском учењу најчешће није довољно само реализовати један тип модела, већ је потребно формирати више различитих модела и поредити њихове резултате како би се изабрао најбољи. Постоји већи број метрика за поређење резултата модела, попут прецизности, просечне (квадратне) грешке модела, *Lift* метрике и слично. Више детаља о метрикама које се користе за поређење карактеристика модела биће дато у следећим поглављима. Управо рачунање метрика прецизности модела на претходно наведеном скупу података намењеном за тестирање перформанси би пружило непристрасну методу поређења њихових карактеристика. Ипак, уколико је почетни скуп података који се користе за развој модел недовољно велики за развој модела, може се одступити од ове праксе. Тада се користе механизми унакрсне провере, тј. дељења података на партиције које би се наизменично користиле за тренирање и валидацију модела.

Битно је подвући да ниједан модел није непромењив и увек прецизан, посебно узимајући у обзир да се са временом подаци (а свакако и понашања која се описују математичким моделом) мењају. Финални избор модела је увек примарно базиран на статистичким мерама које их нумерички пореде. Међутим, пошто су сви модели базирани на подацима који су доступни, а који се прикупљају у зависности од конкретног захтева који је дефинисан на почетку процеса дефинисања модела, закључује се да ће свака накнадна промена почетних претпоставки које су први корак при дефинисању модела довести до деградације перформанси финалног модела. Управо из овог разлога се види колико је први корак (пажљива дефиниција циља пројекта) битан и колико од њега заправо зависи успех целог пројекта.

Да би пројекат доживео успех, неопходно је и успешно представљање његових резултата свим заинтересованим странама. Овај корак који представља круну пројекта се често занемарује, иако јако често може довести до неких промена основне парадигме пројекта што може довести до промена основних постулата пројекта који су дефинисани

на старту, те самим тим могу оборити потпуно исправан пројекат базиран на науци о подацима.

Аутоматизација процеса представља завршни корак припреме модела за употребу и интеграцију са другим алатима који постоје у организацији. Уколико се при развоју модела не води рачуна о његовим могућностима аутоматизације може се доћи у ситуацију да се неки модел са фантастичним резултатима не може користити у редовним процесима.

4. НЕНАДГЛЕДАНО МАШИНСКО УЧЕЊЕ

Машинско учење представља подобласт вештачке интелигенције у којој рачунари уче из података у циљу побољшања перформанси при решавању неког уско дефинисаног задатка, без експлицитног програмирања [36]. Термин машинско учење увео је Артур Самјуел (Arthur Samuel) 1959. године [34], али и поред великог оптимизма у научној заједници шездесетих година прошлог века, машинско учење није довело до брзог напретка развоја вештачке интелигенције која би могла да парира људској интелигенцији. Ипак, вештачка интелигенција, а самим тим и машинско учење као њена окосница, је постала актуелна тема почетком двадесет првог века са напретком у развоју алгоритама, са повећаним обимом доступних података, као и са технолошким напретком у развоју хардвера.

Као што је описано у уводном поглављу, машинско учење се може поделити на надгледано и ненадгледано машинско учење. У надгледаном учењу, рачунар има приступ означеним подацима које може користити ради унапређивања перформанси при обављању неког задатка. Са друге стране, у ненадгледаном учењу рачунари немају приступ означеним подацима, те је из тог разлога задатак алгоритма машинског учења умногоме тежи зато што захтев није прецизно дефинисан а и перформансе се не могу прецизно измерити. Међутим, ненадгледано учење може пружити додатне информације које надгледано учење не може. На пример, уколико је дефинисан задатак раздвајања нежељених или злонамерних порука од регуларних порука електронске поште, метода надгледаног учења би захтевала одређене примере спам електронске поште који су раније слати и на основу њих и препознавала будућу спам електронску пошту. Са друге стране, методе ненадгледаног учења би захтевале само историјску електронску пошту и потенцијално би могле да раздвоји електронску пошту у више целина од којих би једна била спам електронска пошта, друга битна електронска пошта, трећа вести, четврта промоције, и слично. Дакле, ненадгледани алгоритми машинског учења могу пронаћи и обрасце понашања који нису изворно затражени, то јест ненадгледано машинско учење може дати одговор и на питање које му није постављено. Цитат Јана Л'Куна (*Yann LeCun*) који описује значај ненадгледаног машинског учења је дат у [36]: „Већи део учења у природи је ненадгледано учење. Ако би интелигенција била торта, ненадгледано учење би била сама торта, надгледано учење би било шлаг на торти... ...ми знамо како да направимо шлаг, али не знамо да направимо торту. Морамо решити проблем ненадгледаног машинског учења пре него што уопште можемо кренути да се бавимо правом и пуном вештачком интелигенцијом“.

Алгоритам кластеризације представља један од најзаступљенијих алгоритама ненадгледаног машинског учења. Управо ће та техника бити коришћена за груписање клијената мобилног телекомуникационог оператора у истраживању које је окосница ове

докторске дисертације. На основу тог груписања ће бити процењена вероватноћа да ће клијенти напустити посматраног мобилног телекомуникационог оператора.

4.1. Предности и мане надгледаних и ненадгледаних метода машинског учења

Надгледано учење ће имати несумњиво боље перформансе на уском скупу задатака за које се могу дефинисати прецизни обрасци који се не мењају много са проласком времена, као и када је доступан довољно велики означени скуп података за тренирање модела. На пример, када је доступан веома велики скуп слика од којих је свака означена, одговарајући алгоритам надгледаног машинског учења би могао да пружи одличне перформансе при класификацији нових слика истих објеката. Такође, надгледано машинско учење може израчунати своју прецизност на основу поређења својих предвиђања и стварних ознака на сликама. Алгоритам би покушао да смањи грешке на познатим подацима и самим тим побољша предвиђања на неком скупу података са којим се до тада није сусретао. На овај начин се види основна предност надгледаног машинског учења, да се уз помоћ означених података се могу радити фина подешавања прецизности модела. Ипак, врло је захтевно формирати ознаке великог скупа података. Наиме, човек – експерт мора ручно дефинисати сваку ознаку, а уколико је мало означених података на којима се алгоритам учи финални модел ће бити лошији. Такође, колико год моћни алгоритми надгледаног учења били, њихова прецизност је ограничена на прецизност дефинисања ознака података (лабела). Уколико експерт који означава податке у неком временском периоду није конзистентан, може се десити да модел не буде прецизан. Додатно, треба имати у виду да је велика већина података у свету око нас неозначена, па ће могућност вештачке интелигенције базиране на надгледаним алгоритмима машинског учења да унапређује перформансе користећи нове податке са којима до сада није имала додира, бити врло ограничена.

Као што је описано у претходном параграфу, надгледано учење је одлично за решавање уско дефинисаних проблема, али не тако добро у случајевима када су обрасци понашања непознати унапред, уколико се брзо мењају у времену, или уколико не постоји доступан довољно велики скуп означених података. У таквим случајевима предност има ненадгледано машинско учење. Ненадгледано учење се, насупрот надгледаном учењу, не води означеним подацима, већ искључиво обрасцима који се налазе у самим подацима помоћу којих формира модел. Уколико се посматра претходно описани пример класификације слика, модел базиран на ненадгледаном учењу би могао да групише слике на основу тога колико међусобно личе. На пример, слике које подсећају на аутомобил би биле груписане у једну, а слике које подсећају на дрво би биле груписане у другу групу. Наравно, алгоритам сам по себи не би могао да назове сваку групу, тј. да једну групу назове аутомобили а другу дрвеће – то би био задатак за експерта који би анализирао резултате модела. Такође, ненадгледани алгоритам би (након почетне фазе тренирања) након сусрета са новим подацима могао да формира групу недефинисаних слика, што би индицирало да се на сликама појавио неки нови објекат за који би поново требало истренирати модел.

4.2. Алгоритми ненадгледаног машинског учења

Основни типови алгоритма ненадгледаног машинског учења су [36]:

1. Редукција димензија података (*Dimensionality Reduction*),

2. Издвајање (екстракција) карактеристика (*Feature Extraction*),

3. Ненадгледано дубоко учење (*Unsupervised Deep Learning*),

4. Кластеризација (*Clustering*).

Алгоритми редукције димензије података подразумевају алгоритме који врше трансформацију високо димензионалних података на мањи број изразито описних димензија. Редукција димензија омогућава другим механизмима ненадгледаног машинског учења да ефикасније уоче обрасце у подацима тако што ће елиминисати промењљиве од маргиналног значаја. На тај начин се једноставније могу извршити сложени алгоритми који раде са великим количинама података. Постоје два основна типа алгоритама за редукцију димензија података:

1. Линеарна редукција димензија – линеарна трансформација промењљивих из оригиналног високо-димензионог скупа података у ниже-димензиони, примери: анализа главних компоненти (*Principal component analysis* – PCA); декомпозиција сингуларних вредности (*Singular value decomposition* – SVD)...

2. Нелинеарна редукција димензија – нелинеарна трансформација промењљивих из оригиналног високо-димензионог скупа података у ниже-димензиони, примери: изомапе (*Isomap*); учење речника (*Dictionary learning*)...

Анализа главних компоненти је базирана на провери карактеристика свих промењљивих у неком скупу података. Очекивано је да неће вредности свих промењљивих унутар неког скупа података подједнако варирати – вредности неких промењљивих ће варирати више од других и оне ће имати већу вредност при одређивању карактеристика тог скупа података. Применом анализе главних компоненти, модел ће дефинисати ниско-димензиону представу података, истовремено покушавајући да сачува што већу количину варијабилитета коју су изворне промењљиве носиле. Број димензија које генерише метод анализе главних компоненти је драстично мањи од броја изворних промењљивих у оригиналном скупу података. Део варијабилитета података ће дефинитивно бити изгубљен, али ће нова структура података бити много једноставнија за обраду и визуелизацију.

Декомпозиција сингуларних вредности представља приступ за смањивање димензија података коришћењем поступка смањивања ранга оригиналне матрице промењљивих на мањи ранг. Битно је нагласити да се оригинална матрица може рекреирати коришћењем линеарних комбинација неких вектора матрица мањег ранга. Да би се генерисала матрица мањег ранга алгоритам чува векторе оригиналне матрице који носе највише информација (то јест, оне који имају највеће сингуларне вредности). На тај начин матрица мањег ранга чува најбитније елементе из оригиналног скупа промењљивих.

Алгоритам изомапа је базиран на учењу унутрашње геометрије података рачунањем растојања између сваке опсервације и њених суседа али на нееуклидски начин (коришћењем геодетског или закривљеног растојања). Коришћењем ових прорачуна, алгоритам изомапа у следећем кораку трансформише високодимензиони скуп података у ниже-димензиони.

Алгоритам базиран на учењу речника подразумева увођење нових елемената – бинарних вектора где се свака оригинална опсервација може посматрати као пондерисана сума одговарајућих вектора. Матрица (односно речник) коју формирају ови вектори је слабо попуњена, то јест махом је попуњена вредностима 0 са само пар тежинских коефицијената који су различити од нуле. Након генерисања речника, алгоритам може једноставно издвојити најистакнутије представнике елемената изворног простора променљивих – тако што ће изабрати оне који имају највише вредности које су различите од нуле.

Издвајање карактеристика се користи ради генерисања алтернативног приказа оригиналних података. Екстракција карактеристика се може користити и за редукцију димензија тако што би се мањи број нових (изведених) димензија користио за описивање целог скупа података. За екстракцију карактеристика се могу користити нерекурентне неуралне мреже, где број елемената у излазном и улазном слоју неуралне мреже мора бити једнак. Овакве неуралне мреже су познатије као аутоенкодери и служе да ефективно реконструишу оригиналне карактеристике података, истовремено учећи нове представе података у скривеним слојевима неуралне мреже. Сваки скривени слој неуралне мреже учи нову репрезентацију оригиналних променљивих, а сваки следећи слој учи на основу знања претходног слоја. Дакле, из слоја у слој модел учи све сложеније представе оригиналних променљивих. Излазни слој садржи финалну трансформацију улазних променљивих. Ова представа почетних променљивих се може користити као улаз за неки предикциони модел који ће потенцијално бити прецизнији.

Дубоко учење представља грану машинског учења која је базирана искључиво на решавање проблема моделовањем коришћењем неуралних мрежа [37]. Праћење поступка тренирања неуралних мрежа представља велики изазов при дефинисању неуралне мреже. При процесу тренирања, скривени слојеви неуралне мреже се баве учењем представе улазних података како би решили постављени проблем. Да би се унапредио рад комплетне неуралне мреже анализира се градијент функције грешке ради корекције коефицијената сваког чвора неуралне мреже. Свака корекција коефицијената чворова мреже је врло рачунски захтевна, а додатно се могу јавити два проблема. Први проблем које се може јавити је врло ниска вредност градијента функције грешке што онемогућава тренирање неуралне мреже (*vanishing gradient problem*), а други је изузето висока вредност градијента што тренирање мреже чини врло нестабилним (*exploding gradient problem*). Да би се решили ови проблеми при тренирању сложених, вишеслојних неуралних мрежа, тренирање неуралних мрежа се врши у више сукцесивних фаза (корака), где свака фаза подразумева тренирање једне плитке неуралне мреже. Тада је излаз једне мреже, истовремено улаз следеће неуралне мреже. Типично је прва неуронска мрежа у овој каскади ненадгледана неуронска мрежа, док су касније мреже најчешће надгледане. Дакле, ненадгледано дубоко учење подразумева претренирање сложених неуралних мрежа ради оптимизације процеса тренирања неуралне мреже.

4.3. Кластеризација

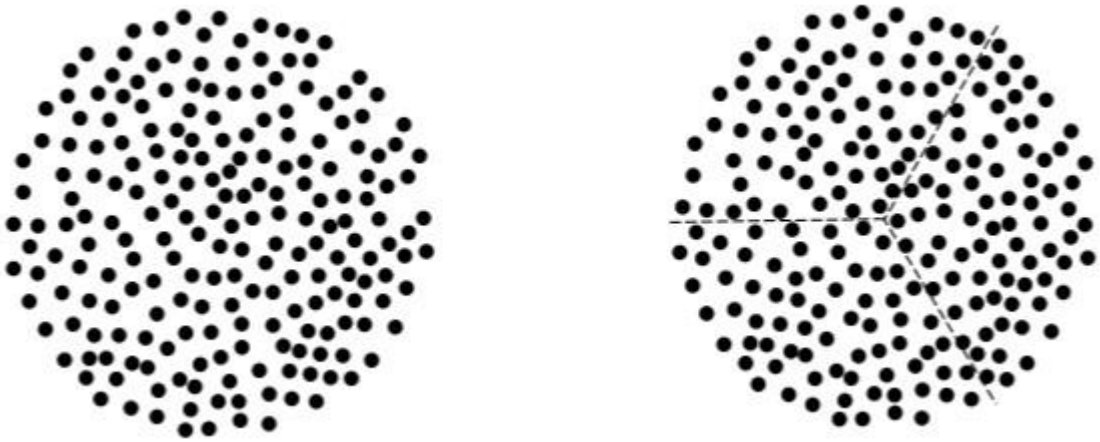
Кластер анализа (кластеризација) подразумева велику групу техника које имају за циљ партиционисање података у групе на начин да елементи унутар једне групе буду међусобно што сличнији, док су истовремено што различитији од осталих елемената у иницијалном скупу података. Кластер анализа представља скуп метода за генерисање интерпретабилне и информативне класификације иницијално некласификованих података, коришћењем података (променљивих) који су прикупљени на нивоу сваке

појединачне индивидуе (опсервације) [38]. Један од битнијих елемената претходне дефиниције представља тврдња да класификација мора бити интерпретабилна, што подразумева да се финалне групе морају описати на начин који може донети неку додатну вредност. Иако се обично подразумева да се финални резултати анализе кластера морају успешно интерпретирати, то некада не мора бити случај. Ипак, и у том случају кластер анализа може бити успешна у смислу да је указала на одређену групу података који се понашају на слични начин.

Иако се термини кластер, група и класа најчешће интуитивно користе за описивање истих објеката, прецизну дефиницију термина кластер је тешко дати [39]. У [40] и [41] термин кластер је дефинисан у смислу интерне кохезије (хомогености) и екстерне изолације (сепарације). На слици 4.1. су приказан примери кластера који задовољавају неке од ових услова. Кластери са слика су очигледни чак и без формалне дефиниције. Суштина је да примери са слике 4.1. приказују да је врло компликовано поставити дефиницију која би могла бити употребљива у свим екстремним случајевима.

Слика 4.1: Илустрација концепта хомогености и сепарације: лево – кластери су хомогени али нису изоловани; средина – кластери су изоловани али нису хомогени; десно – кластери су изоловани и хомогени [39]

Пример са слике 4.2. лево не показује никакве знаке „природне“ структуре кластера, већ је у питању мање-више хомогена структура. Иако би се очекивало да методе кластеризације не могу одредити никакве групе унутар оваквог скупа података, на слици 4.2. десно је приказана једно потенцијално груписање елемената у три кластера. Често се процес дељења хомогеног скупа података у више кластера назива дисекција и може имати примене у неким специфичним случајевима [39]. Дакле, потенцијални проблем у кластеризацији података може бити то што није унапред позната геометријска структура података па се сам тим може јавити ситуација да се сви примећени кластери интерпретирају као „природни“. На тај начин се занемарује могућност да је класификација која је продукт кластер анализе наметнула структуру подацима (уместо да је метода само пружила увид у стварну „природну“ структуру података). Ово је један од већих потенцијалних проблема у кластер анализи.



Слика 4.2: Лево – скуп података који не садржи „природне“ кластере; десно – пример дисекције података са слике лево на три кластера [39]

Једна од најчешћих примена метода кластеризације је и дефинисање образаца у комплексном скупу података. Ти обрасци се накнадно могу користити за унапређење процеса, система или остваривање циљева организација. Постоје разни примери коришћења метода анализе кластера, попут откривања превара у финансијама, дефинисања типова купаца у малопродајним радњама, груписања клијената по оствареном профиту, класификација клијената радњи за продају одеће по њиховим модним преференцама и слично. Откривање превара у финансијама је јако сложен и компликован процес који не може зависи само од једног фактора. На пример, уколико се посматра особа ХУ, може се претпоставити да ништа неуобичајено не представља куповина преко интернета у износу од 10.000 динара. Такође, ни куповина одеће особе ХУ у износу од 10.000 динара сама по себи не мора представљати ништа алармантно. Куповина у Пироту, такође, не мора бити сигнал за аларм сама по себи иако особа ХУ према доступним подацима живи у Београду. Међутим, куповина одеће преко интернета у износу од 10.000 динара при чему је захтев послат из Београда док се особа ХУ налази у Пироту, може представљати потенцијални аларм за неке малверзације.

Профилирање кластера се може дефинисати као процес додељивања назива кластерима који су добијени претходном кластер анализом. У овом процесу би циљ био идентификовати неке особине (или неку њихову комбинацију) које могу једнозначно описати сваки кластер.

4.3.1. Типови кластеризације

Две основне категорије метода кластеризације су [39]:

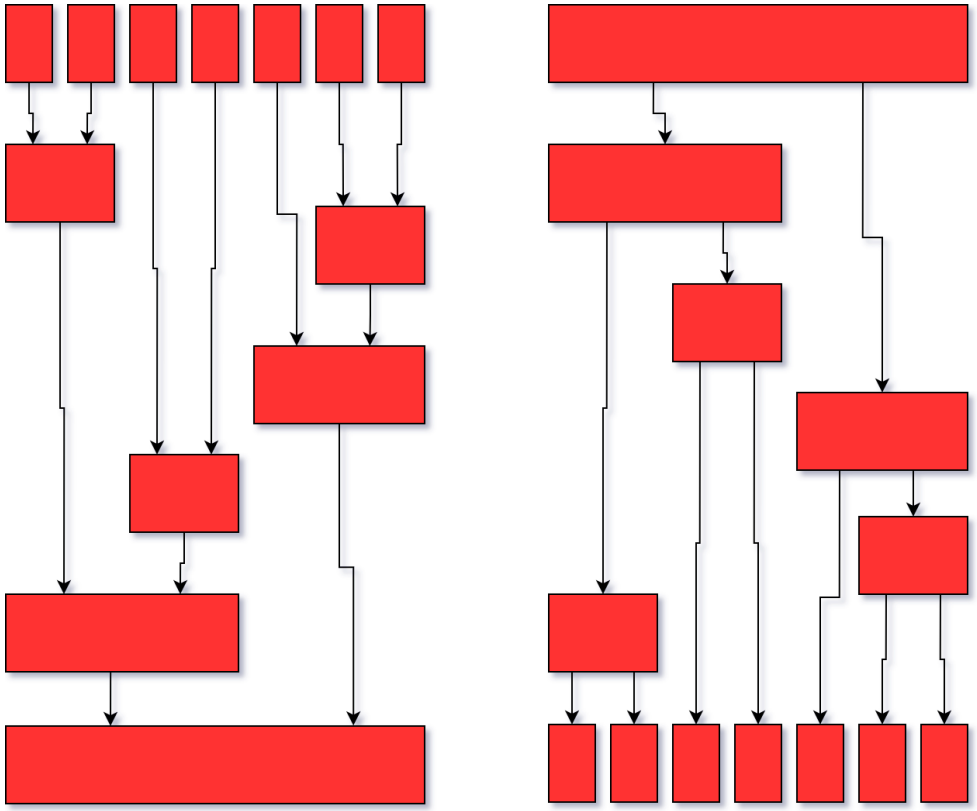
1. Хијерархијска кластеризација,
2. Партитивна (оптимизацијска) кластеризација.

Хијерархијска кластеризација је базирана на принципу генерисања кластера који су хијерархијски угњеждени унутар кластера формираних у ранијим итерацијама. У [39] се тврди да хијерархијска кластеризација представља окосницу свих кластер анализа у пракси зато што су широко распрострањене и имплементиране у различитим

софтверским пакетима, као и зато што су једноставне за употребу. Хијерархијска кластеризација се може јавити у два облика:

1. Сакупљајућа (*Agglomerative*) кластеризација,
2. Раздвајајућа (*Divisive*) кластеризација.

Сакупљајућа кластеризација подразумева процес базиран у три корака: у првом кораку се свака опсервација смести у сопствени кластер, у другом се спајају два најсличнија кластера, а трећи корак подразумева понављање другог корака до тренутка када преостане само један кластер који ће садржати комплетан скуп података. Са друге стране, раздвајајућа кластеризација подразумева процес инверзан сакупљајућој кластеризацији: у првом кораку се комплетан скуп података смешта у један кластер, у другом кораку се опсервације које су најмање сличне деле у два раздвојена кластера, а трећи корак подразумева понављање другог корака до тренутка када свака опсервација не постане засебан кластер. Графички приказ поступка сакупљајуће и раздвајајуће кластеризације је дат на слици 4.3.



Слика 4.3: Лево – Сакупљајући метод кластеризације; десно – раздвајајући метод кластеризације

Ипак, хијерархијске методе имају значајне недостатке. Као прво, перформансе хијерархијске кластеризације драстично опадају са порастом броја опсервација у скупу података који се анализира. Додатно, свака грешка која се догодила на неком ранијем нивоу (у некој претходној подели или спајању кластера) се не може кориговати. Управо због тога, у [42] се тврди да хијерархијске методе имају дефект да никада не могу исправити грешку коју су начиниле у неком претходном кораку. Трећи озбиљни недостатак се огледа у томе да постоји пуно различитих начина на који се може рачунати

сличност (код сакупљајуће кластеризације), односно различитост (код раздвајајуће кластеризације) између опсервација и ниједна метода се не може издвојити као најбоља; самим тим јавља се проблем коју методу изабрати.

Алтернативу хијерархијској кластеризацији представља партитивна (оптимизацијска) кластеризација. Партитивна кластеризација функционише по принципу смештања опсервација у кластере помоћу минимизације неке претходно дефинисане функције грешке. Основна идеја иза ове методе је да се поделом n опсервација на g захтеваних кластера израчунава индекс $c(n,g)$ чија вредност мери неки аспект успешности конкретне поделе. За неке индексе су пожељније високе вредности, док се за неке друге тражи што нижа вредност. Главна предност ових индекса је то што се могу међусобно поредити, па се самим тим могу тестирати више различитих предлога броја кластера у подацима и изабрати решење које је што приближније „природним“ кластерима. Постоји велики број индекса који је дефинисан у литератури [39], што само подвлачи колико је сложен проблем одређивања оптималног броја кластера у непознатом скупу података. Партитивне методе кластеризације су практично једини избор при кластер анализи великих скупова података, узимајући у обзир да су хијерархијске методе врло рачунски захтевне.

Ипак, постоји значајан број проблема који се јављају при раду са партитивним методама кластеризације. Већина партитивних метода кластеризације захтева број кластера који је присутан у подацима као параметар који се задаје пре започињања процеса (самим тим, неопходно је прорачунати га на неки начин). Такође, најчешће се унапред претпоставља облик кластера (у питању је у највећем броју случајева (хипер)сферичан облик). Партитивне методе кластеризације су веома осетљиве на појаву изузетака у подацима, а на коначни резултат прорачуна може утицати чак и редослед опсервација у иницијалном скупу података. Један од проблема је и то да је практично немогуће пронаћи једно оптимално решење груписања података из разлога што постоји велики број потенцијалних решења. Теоретски је могуће да се индекс „успешности“ кластеризације прерачуна за сваку могућу поделу опсервација, али је то практично немогуће због њиховог превеликог броја. Наиме, постоји формула која описује колико ће се различитих подела генерисати уколико се n опсервација кластерише у g група [43]:

$$N(n,g) = \frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^n \tag{4.1}$$

Уколико се, на пример, 100 опсервација из улазног скупа података дели на четири кластера, укупно би постојало $6.7\cdot10^{58}$ могућих комбинација. Управо услед експлозивног скока броја комбинација кластеризација у случају партитивне методе кластеризације уведен је принцип хеуристичке претраге (*heuristic search*) оптималних подела на кластере. Њеним коришћењем се смањује број комбинација које је потребно проверити. Принцип рада хеуристичке претраге је следећи: у првом кораку се изврши иницијална кластеризација n опсервација у g група (најчешће помоћу предефинисаних референтних вектора иницијалних кластера – *seeds*); рачуна се промена индекса „успешности“ при пребацивању било које опсервације из тренутног у неки други кластер; изврши се промена која ће довести до највећег побољшања индекса „успешности“; уколико је побољшање индекса изнад одређеног предефинисаног прага, поновити процедуру. На овај начин, хеуристичка претрага значајно смањује број подела које је потребно анализирати. Ипак, прецизност хеуристичке претраге умногоме зависи од почетног

избора вектора иницијалних кластера, то јест хеуристичка претрага не може гарантовати глобално најбоље решење, већ само решење које је најбоље за дефинисани избор вектора иницијалних кластера.

4.3.2. Рачунање сличности

У претходном поглављу је уведен термин сличности, односно различитости, између опсервација унутар неког скупа података. Иако тај термин може на први поглед деловати довољно интуитиван, те да не мора захтевати прецизну дефиницију, једноставан пример може показати колико је сложено дефинисати сличност и између два тривијална појма. На пример, уколико би се поставило питање: „Која птица је сличнија патки, врана или пингвин?“ одговор би у многоме зависио од избора типа сличности. Може се претпоставити да је прва асоцијација на термин сличности то да су два објекта слична уколико деле највише заједничких особина. Ипак, тада проблем постаје дефиниција особина [44]. Уколико се пореде птице из примера по њиховој могућности летења, може се рећи да су патка и врана сличне. Ако се са друге стране посматрају њихове могућности пливања, закључак је да су патка и пингвин сличнији. Дакле, одговор на питање „Која птица је сличнија патки, врана или пингвин?“ зависи од начина на који је изабрано да се пореде. Сличан принцип важи и у кластеризацији, коришћење различитих дефиниција поређења сличности опсервација даће различите финалне кластере.

Независно од конкретног избора метрике за одређивање сличности, потребна је дефиниција основних постулата (то јест принципа) који представљају добру основу за исправну метрику сличности d између опсервација x и y [39]:

1. Симетрија: $d(x, y) = d(y, x)$
2. Неидентична препознатљивост: $d(x, y) \neq 0\ x \neq y$
3. Идентична непрепознатљивост: $d(x, y) = 0\ x = y$
4. Неједнакост троугла: $d(x, y) \leq d(x, z) + d(y, z)$.

У литератури је дефинисан велики број различитих нумеричких метода за одређивање сличности. Избор конкретне нумеричке методе зависи од типа саме карактеристике која се пореди. Између осталих, могу се користити следеће методе:

1. Еуклидско растојање,
2. Блоковско (*city block*) растојање,
3. Минковски (*Minkowski*) растојање,
4. Пирсонов (*Pearson*) коефицијент корелације,
5. Говерово (*Gower*) растојање.

Еуклидско растојање је једна од најраспрострањенијих метрика које се користе за рачунање сличности. Може се користити и за интервалне и за континуалне променљиве. Оно представља својеврсну екстензију Питагорине теореме. Проширење Питагорине теореме је неопходно за рачунање растојања између било које тачке у n-

димензионом простору. Проширење подразумева замену имплицитног поређења са координатним почетком (тачком (0,0)), са произвољном *y* тачком у Еуклидском простору. Формула Еуклидског растојања је дата са [39]:

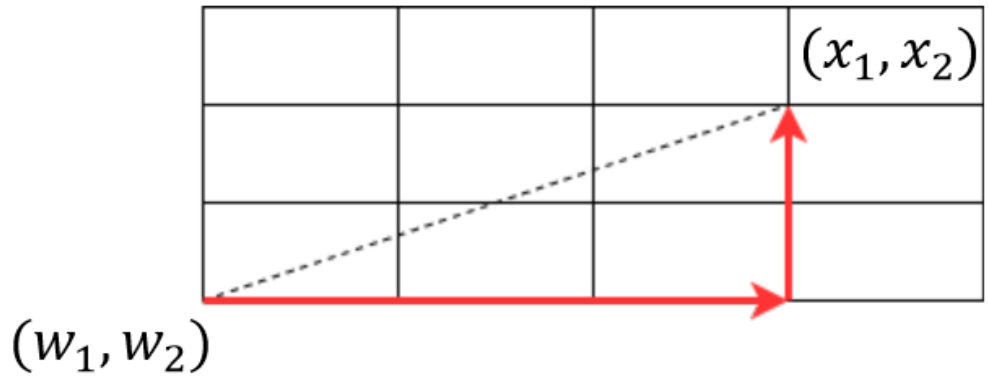
$$D_E = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.2}$$

Да би се формула искориговала за евентуалне недостајуће вредности у улазним подацима, користи се следећа модификација у формули (где је *n* укупан број променљивих а *v* број ненедостајућих вредности):

$$D_{E\,mod} = \|x - y\| = \sqrt{\left(\frac{n}{v}\right)\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.3}$$

Блоковско растојање је мање осетљиво на изузетке у подацима за разлику од Еуклидског растојања. Оно је базирано на мерењу растојања дуж катета троугла који спаја две посматране тачке. Блоковско растојање се може интерпретирати и као кретање по улицама града, где су све улице међусобно пода правим углом (слика 4.4.). Из овог разлога је ово растојање познато и као Менхетн растојање. Блоковско растојање се може користити за мерење растојања између и интервалних и континуалних променљивих. Формула за израчунавање блоковског растојања је:

$$D_B = \sum_{i=1}^{k}|w_i - x_i| \tag{4.4}$$



Слика 4.4: Приказ израчунавања блоковског растојања

Интересантно је да су еуклидско и блоковско растојање само специјални случајеви растојања Минковског [39]:

$$D_{M\lambda} = \left(\sum_{i=1}^{k}|w_i - x_i|^{\lambda}\right)^{1/\lambda} \tag{4.5}$$

За вредности параметра λ=1 добија се формула за блоковско растојање, док се за λ=2 добија формула за Еуклидско растојање.

У статистици, корелација се често користи за поређење сличности променљивих, то јест коефицијент корелације се најчешће користи као мера која указује на потребу уклањања редундантних променљивих из скупа података. Једна од најзаступљенијих корелационих статистика је и Пирсонов коефицијент корелације [39]. Формула за

израчунавање различитости два кластера *x* и *y* коришћењем Пирсоновог коефицијента корелације је:

$$k_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{4.6}$$

где је x_i *i*-та опсервација кластера *x*, y_i *i*-та опсервација кластера *y*, а $\bar{x}$ и $\bar{y}$ су, редом, центри кластера *x* и *y*. Пирсонов коефицијент корелације може имати вредности од -1 до 1, где вредности 1 означава највећу могућу позитивну везу, а -1 означава најјачу могућу негативну везу између анализираних кластера. Да би се коефицијент корелације превео у меру растојања може се користити формула:

$$D_k = \left(1 - k_{xy}\right)/2 \tag{4.7}$$

где D_k представља меру различитости између *x* и *y*, и узима вредности у интервалу [0,1].

Битно је запазити да Пирсонов коефицијент корелације не испуњава све принципе сличности који су дефинисани на почетку овог поглавља. Заправо, у [39] је наведено да је коришћење корелационих коефицијената у кластеризацији далеко спорније од његовог коришћења у оцењивању линеарног односа између променљивих. Додатно, једини ефективан начин при којем се коефицијент корелације може користити је да су све променљиве које се користе за кластеризацију дефинисане на истој скали. Ако су различите скале коришћене за анализу променљивих и њихових средњих вредности и варијанси, коефицијенти корелације постају безначајни [39]. На пример, нека су анализиране две серије бројева: 5, 4, 3, 2, 1 и 510, 420, 330, 240, 150. Ове две серије имају корелацију које је једнака тачно 1, те би на основу ове анализе коефицијената корелације ове две серије биле нераспознатљиве једна од друге (а очигледно нису). Дакле корелациони коефицијенти показују мањак способности да анализирају неидентичну препознатљивост.

Говерово растојање је метрика која се може користити за рачунање сличности свих типова променљивих. Дефинисана је као [45]:

$$D_G = \frac{\sum_{k=1}^{v} \delta_{ijk} s_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}} \tag{4.8}$$

где је s_{ijk} мера сличности између опсервација *i* и *j* уколико се посматра променљива *k*, а δ_{ijk} је типично један или нула у зависности од тога да ли се поређења сматрају важећим или неважећим. На пример, вредност δ_{ijk} је једнака нули ако је променљива *k* недостајућа за једну или обе анализиране опсервације *i* и *j*. Такође, вредност δ_{ijk} може бити једнака нули ако се сматра да је одговарајуће да се код бинарне променљиве *k* искључе сва негативна поклапања. Са друге стране, за бинарне и категоричке променљиве са више од две категорије, компонента сличности s_{ijk} може имати вредности један када две опсервације имају исте вредности и нула у супротном. За континуалне променљиве, предложено је коришћење следеће мере сличности:

$$s_{ij} = 1 - \left|x_{ik} - x_{jk}\right|/R_k \tag{4.9}$$

где је R_k опсег опсервација за k-ту променљиву (другим речима, након скалирања дозвољених вредности k-те променљиве на јединични опсег, примењује се класична формула за блоковско растојање).

4.3.3. Припрема података за процес кластеризације

Подаци који се користе за кластер анализу могу бити систематски или опортунистички. Иако су подаци који су систематског типа погоднији за моделовање, у реалном окружењу се може доћи у сусрет са драстично већим бројем опортунистичких података који нису „чисти“ попут систематских података. При анализи масивних опортунистичких података, постаје значајно теже наћи било какав образац коришћењем анализе кластера. Ипак, добра припрема података је кључ за успешнију анализу. Пет основних корака при припреми кластера су:

1. Избор скупа података за анализу и узорка података,
2. Избор променљивих,
3. Графичка анализа података,
4. Стандардизација променљивих и
5. Трансформација променљивих.

Први корак при припреми података за кластеризацију подразумева упознавање са подацима. Заправо прво питање које се јавља при анализи кластера је: „кога кластерисати?“. Најчешће није неопходно кластерисати целу популацију података већ само неки подскуп који може адекватно описати понашање целе популације – узорак. Постоји више метода за дефинисање узорка попут случајног, стратификованог и слично. На овај начин би се од почетног скупа података који има пуно опсервација и пуно променљивих, дошло до скупа са мањим бројем опсервација и истим бројем променљивих у односу на изворни скуп података.

Следећи корак би представљала редукција променљивих, то јест поступак смањивања димензија података. Ово је битно из три разлога: као прво, мањи број променљивих подразумева брже процесирање податка; друго, неке од променљивих које су прикупљене могу бити ирелевантне за анализу; и као треће, променљиве које су међусобно јако корелисане могу имати диспропорцијални утицај на резултујуће кластере. У моделима који су надгледаног типа ирелевантне променљиве могу бити елиминисане на основу њихове везе (на пример корелацијом) са циљном (означеном) променљивом. Са друге стране, пошто је кластеризација модел базиран на ненадгледаном машинском учењу, никаква зависна променљива није доступна па се самим тим процес избацивања ирелевантних променљивих компликује. Ирелевантне променљиве се, на пример, могу елиминисати на основу експертског мишљења.

Са друге стране, елиминисање редундантних (колинеарних) променљивих је битна ставка која у многоме може поспешити развој било надгледаних било ненадгледаних модела. На пример, претпоставимо да се за потребе анализе кластера анализира скуп података са четири променљиве, од којих су три међусобно јако корелисане. Модели за анализу кластера који би радили на основу оваквих података би заправо триплирали ефекат променљивих које деле концептуално значење у поређењу

са ефектом који генерише последња променљива, те би самим тим резултати били неодговарајући. Елиминисање редундантних променљивих се може извршити кластеризацијом самих променљивих. На тај начин би се изабрао тачно по један представник из сваког кластера променљивих и само би се тај подскуп променљивих користио за финалну анализу, док би све остале променљиве биле одбачене. Суштински, кластеризацијом променљивих би се формирале групе (кластери) променљивих на начин да су унутар једне групе променљиве које су максимално корелисане једна са другом, а истовремено максимално некорелисане са променљивама из других кластера. Распоред променљивих у кластере се одвија у две фазе [46]:

1. Прва фаза подразумева сортирање најближих компоненти – принцип сличан алгоритму сортирања најближем центроиду представљеном у [47]. У свакој итерацији, компоненте кластера променљивих се прерачунавају на начин да се променљиве смештају у кластер са којим имају највећи квадрат корелације.

2. Друга фаза подразумева алгоритам претраге који проверава да ли ће евентуално премештање неке променљиве у неки други кластер повећати укупну варијацију модела. Уколико се променљива пребаци из кластера у кластер, потребно је да се рекалкулишу центроиди изворног и одредишног кластера, након чега се поступак наставља са следећом променљивом.

Поставља се и питање на који начин из кластера променљивих изабрати тачно једног представника. Постоји више начина да се реши овај проблем, при чему је први чисто експертски (експерти могу проценити променљиве које се налазе унутар неког кластера и изабрати једну за коју сматрају да би највише погодовала даљој анализи). У случају алгоритама надгледаног учења, други начин може подразумевати избор променљиве која је најбоље корелисана са означеном циљаном променљивом. У случају ненадгледаног учења (када означена променљива није доступна), може се користити и $1 - R^2$ однос [48]:

$$\langle 1 - R^2 \rangle_{ratio} = \frac{1 - R^2_{own\ cluster}}{1 - R^2_{next\ closest\ cluster}} \tag{4.10}$$

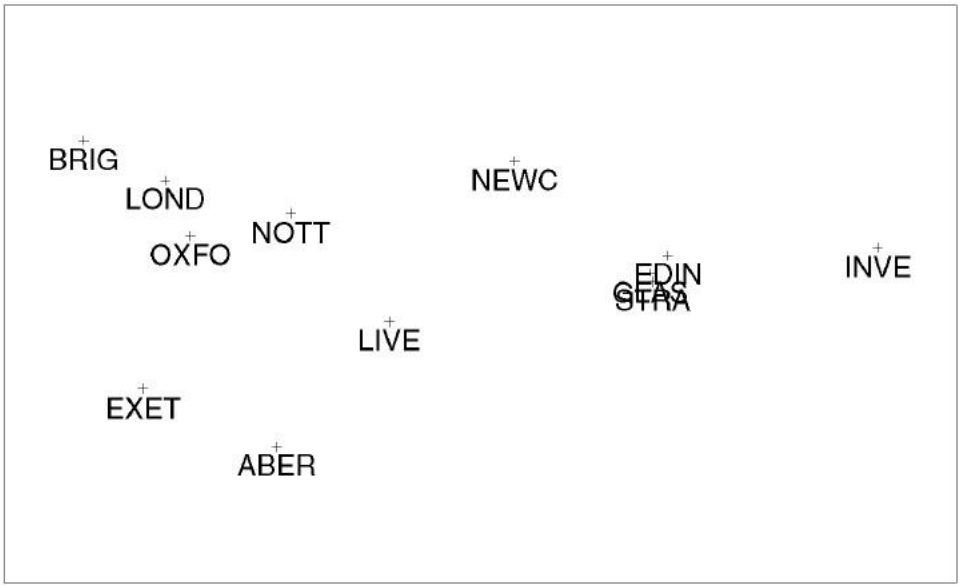
где је R^2 коефицијент детерминације, то јест статистичка мера која описује припадност анализиране променљиве свом кластеру променљивих. Очигледно је да је циљ да променљива која ће бити изабрана из свог кластера као карактеристични представник буде што сличнија свом кластеру (то јест $R^2_{own\ cluster} \to 1$) и да буде што различитији од следећег најближег кластера (то јест $R^2_{next\ closest\ cluster} \to 0$). На тај начин се закључује да је оптимално изабрати ону променљиву која има најмању вредност односа $\langle 1 - R^2 \rangle_{ratio}$ за карактеристичног представника свог кластера.

Графичка анализа података може бити јако значајна помоћ при кластер анализи. Наиме, график који садржи приказ дистрибуције података може пружити основне информације о томе колико приближно кластера постоји у подацима, колика је варијација података, као и какав је облик кластера. Узимајући у обзир да неке методе кластеризације имају тенденцију да дефинишу кластере одређеног типа и облика, коришћење графичке анализе података може омогућити увид у то да ли постоје неке специфичности у подацима које би захтевале искључиво одређене алгоритме кластеризације. Један од основних проблема при графичкој анализи података може представљати велики број димензија. Тада је потребно користити неке од метода

редукције димензија попут анализе главних компоненти (*principal component analysis*) [49], или мултидимензионо скалирање (*multidimensional scaling*) [50].

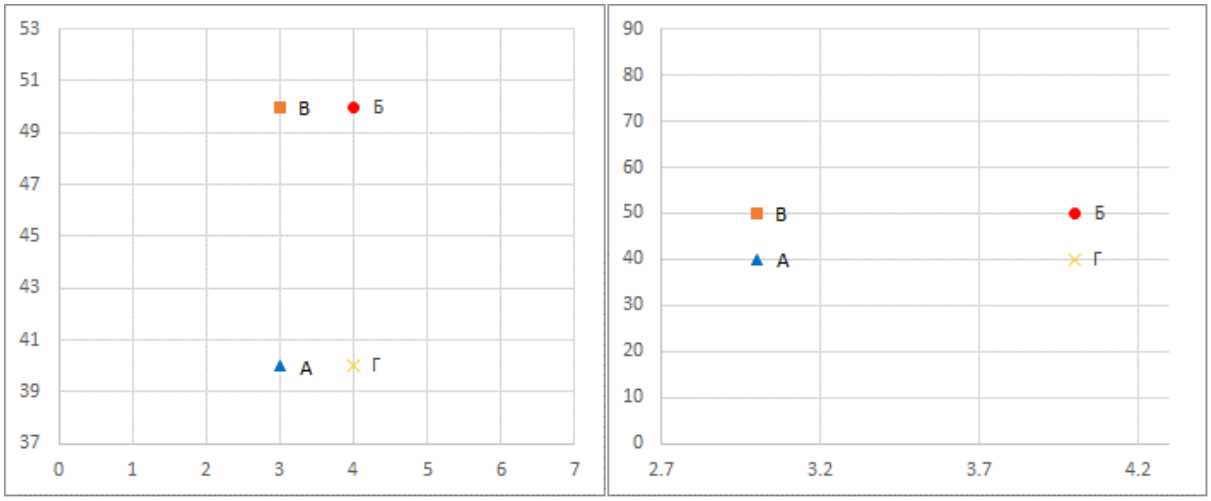
Анализа главних компоненти подразумева трансформацију основног скупа променљивих у нови скуп тако да нове променљиве буду међусобно некорелисане, али и да свака следећа компонента описује све мањи део изворне варијације података. Свака нова променљива може представљати неку линеарну комбинацију изворних променљивих. У случају када изворне променљиве нису корелисане, један део варијације података се може посматрати дуж неке линије која није паралелна ни са једном од оса (то јест изворних променљивих). Линија правца највеће варијације представља први еигенвектор који заправо минимизује (у смислу квадратних растојања) пројекције података на осе. Друга основна компонента се формира на начин да минимизује пројектована растојања на правац који је нормалан првој основној компоненти. На тај начин се постиже да основне компоненте никада не могу бити корелисане. Свака следећа основна компонента се формира на сличан начин, при чему ће свака следећа моћи да објашњава све мањи и мањи део изворне варијације података. Иако изворних променљивих може бити изузетно велики број (а може се формирати исти број главних компоненти колико је и изворних променљивих), често ће само пар главних компоненти описивати велики проценат варијабилитета свих изворних променљивих уколико су изворне променљиве корелисане. Обрнуто, уколико изворне променљиве уопште нису корелисане, анализа главних компоненти неће успети да постигне тако значајне резултате. Ипак, преостаје питање колико главних компоненти је довољно за исправан опис улазних променљивих. У [51] је приказан механизам који је популаризован у раду [52] те је данас познат као Гутман-Каизер метод који говори да је потребно анализирати све основне компоненте које имају сопствене вредности веће од један [53].

Мултидимензионо скалирање за циљ има лоцирање опсервација у простору редукованих димензија (обично је у питању Еуклидски простор) на начин да разлике између појединачних тачака што приближније приказују реалне сличности и разлике између опсервација. На пример, пример на слици 4.5. представља график добијен мултидимензионим скалирањем података о растојању између градова у Великој Британији [50]. Постоји велика сличност између реалног распореда градова у Британији са распоредом на слици (узимајући у обзир да је на слици дат ротиран приказ са сликом рефлектованом у огледалу).



Слика 4.5: Мапа градова у Великој Британији реконструисана на основу процењеног времена путовања коришћењем мултидимензионог скалирања; ABER - Aberystwyth, BRIG - Brighton, EDIN - Edinburgh, EXET - Exeter, GLAS - Glasgow, INVE - Inverness, LIVE - Liverpool, LOND - London, NEWC - Newcastle, NOTT - Nottingham, OXFO - Oxford, STRA -Strathc1yde [50]

Стандардизација променљивих подразумева скалирање зарад бољег представљања података у циљу што коректније анализе. Конкретно, уколико се исти подаци визуелно анализирају на два различита начина, то јест коришћењем две различите скале при приказивању података, груписање тих података може бити драстично другачије. Пример на слици 4.6. приказује један пример ове тврдње. И на слици лево и на слици десно су приказане идентичне опсервације, са том разликом да су скале графика на слици лево и на слици десно различите. Очигледно се намеће да је природно груписање опсервација на слици лево такво да тачке А и Г припадају једном, а тачке В и Б припадају другом кластеру. Са друге стране, слика десно индицира да је природно груписање истих тачака мало другачије, груписане би биле тачке А и В у један, а тачке Б и Г у други кластер.



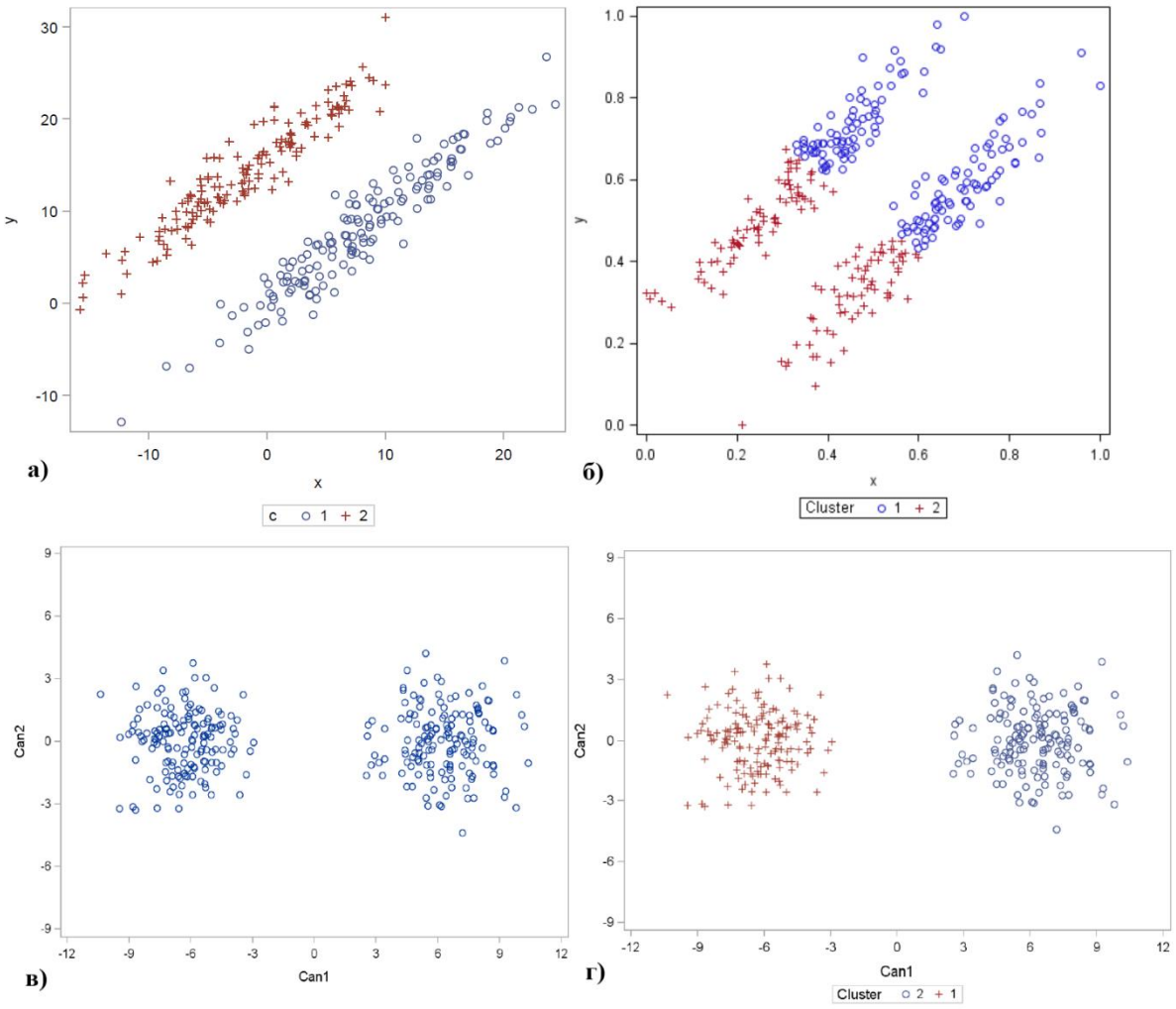
Слика 4.6: Приказ значајности стандардизације променљивих у кластеризацији

На сличан начин би се и рачунар понашао у оваквој ситуацији. Дакле, попут људског ока, алгоритам за одређивање кластера се може заварати. Генерално, промењљиве које имају велику варијансу могу имати већи утицај на формирање кластера у односу на промењљиве мање варијабилности. Да би се овај ефекат контролисао, потребно је извршити стандардизацију промењљивих. Постоји велики број алгоритама за стандардизацију промењљивих [54], при чему се општи принцип метода може описати следећом формулом:

$$S = add + multiply * \frac{X - location}{scale} \tag{4.11}$$

где је S стандардизовани резултат, X оригинална промењљива, $scale$ представља „скалу“, то јест меру опсега (која може бити једнака стандардној девијацији, разлици максималне и минималне вредности...), $location$ представља меру локације стандардизоване вредности (која може бити једнака средњој вредности, минимуму промењљиве...), add константа која се може додати на стандардизовану вредност, $multiply$ константа којом се може помножити стандардизована вредност. Стандардизација промењљивих се може посматрати као посебни случај додавања тежинских коефицијената на промењљиве [39], где су тежински коефицијенти реципрочне вредности варијабилности појединачних промењљивих. Ипак, овакав начин избора коефицијената стандардизације може у неким случајевима смањити диференцијабилност појединих кластера [55]. Управо из овог разлога је у [54] дата препорука да се као одговарајући стандарадизациони коефицијент при процесу анализе кластера користи опсег промењљиве то јест разлика максималне и минималне вредности промењљиве.

Последњи корак при припреми за анализу кластера представља трансформација промењљивих. Она је неопходна из разлога што се често може јавити проблем да изворни подаци могу генерисати кластере који нису сферичног облика. Наиме, кластери могу бити издужени (или на други начин неправилног сферичног облика) и у том случају кластер анализа стандардним методама које одређују сферичне кластере у подацима неће дати смислене резултате. Тада је неопходно трансформисати изворне промењљиве на начин да се корелација промењљивих у сваком кластеру смањи. Ипак да би се знала корелација појединачних промењљивих унутар кластера, потребно је знати саму расподелу опсервација по кластерима (што је изворно и циљ саме кластеризације), па се јавља проблем. Ипак, постоје механизми за апроксимативну процену матрице коваријанси унутар кластера [39] на основу које би се могле формирати трансформације промењљивих на начин да се омогући успешна кластеризација и података са природним несферичним кластерима. Пример резултата кластеризације без и са покушајем трансформације неправилних кластера је дат на слици 4.7.



Слика 4.7: а) Очекивани изглед два кластера који нису правилни; б) Предложено решење без почетне трансформације променљивих; в) Изглед оригиналних променљивих када су трансформисане на начин да се смањи корелација променљивих унутар кластера; г) Примена класичних метода кластеризације на трансформисаним подацима [56]

4.3.4. Партитивна кластеризација

Партитивна (оптимизацијска) кластеризација врши поделу неког скупа опсервација на начин да изврши оптимизацију неког критеријума грешке поделе, као што су на пример максимизација раздвојености кластера, или максимизација хомогености кластера. Узимајући у обзир да су методе партитивне кластеризације погодне за рад се великим бројем опсервација, могу се користити у случајевима када хијерархијска кластеризација једноставно није могућа. Међутим, партитивне методе захтевају *a priori* познавање броја кластера који се налазе у подацима. Одређивање броја кластера може представљати велики изазов. Теоретски је могуће генерисати сваку могућу поделу унутар неког скупа података и оценити је на основу претходно дефинисаних критеријума (и на тај начин обезбедити глобално најбоље решење за конкретан скуп података), али је у пракси ово најчешће немогуће из разлога превелике количине података и група које је потребно проверити. Из тог разлога дошло је до развоја хеуристичких метода базираних на следећим корацима:

1. Дефинисати неку поделу n опсервација у g група, то јест кластера

2. Прерачунати промену у критеријумима одређивања кластера до које би довело померање било које опсервације из једног кластера у други; извршити промену која би довела до највећег побољшања

3. Понављати корак два докле год доводи до значајног побољшања почетно задатог критеријума (до промене критеријума која је већа од неке унапред задате минималне смислене вредности)

Проблем код хеуристичких метода се огледа у томе да више не постоје гаранције да ће партитивна кластеризација довести до глобалног минимума функције грешке (најбољег могућег решења), већ само до неког локалног минимума функције грешке. Практично то значи да ће хеуристичка метода на различитим узорцима истог скупа података (а некад чак и на истом скупу података који је другачије сортиран) израчунавати различите вредности очекиваног броја кластера у подацима.

Постоје различити критеријуми за оцену грешке поделе опсервација на кластере, али два најзаступљенија су максимизација сума квадрата растојања између различитих кластера и минимизација сума квадрата растојања опсервација унутар једног кластера. Велика вредност сума квадрата растојања између различитих кластера показује да је подела генерисала кластере који су добро раздвојени, а са друге стране минимизација сума квадрата растојања опсервација унутар једног кластера показује да су кластери који су добијени хомогени.

Адекватност предложеног решења кластеризације се може нумерички показати раздвајањем тоталне варијације T на варијацију унутар кластера W и варијацију између кластера B, на начин да је $T = B + W$. Прецизније, важи да је [39]:

$$T = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x})\,(x_{ml} - \bar{x})' \tag{4.12}$$

где је g број кластера, n_m број опсервација унутар кластера m (такав да је $\sum_{m=1}^{g} n_m = n$, где је n укупан број опсервација унутар скупа података), x_{ml} вектор опсервације l која се налази у кластеру m, а $\bar{x}$ је вектор средњих вредности свих променљивих у скупу података. Сваки вектор ће имати тачно онолико димензија колико је променљивих у моделу. Варијација унутар кластера је тада једнака:

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \overline{x_m})\,(x_{ml} - \overline{x_m})' \tag{4.13}$$

где је $\overline{x_m}$ вектор средњих вредности свих променљивих унутар кластера m. Варијација између кластера је једнака:

$$B = \sum_{m=1}^{g} n_m (\overline{x_m} - \bar{x})(\overline{x_m} - \bar{x})' \tag{4.14}$$

па је одатле заиста:

$$T = W + B \tag{4.15}$$

У случајевима када је у улазном скупу података присутно више од једне променљиве, једначина 4.15. постаје драстично компликованија у поређењу са случајем када је у улазном скупу присутна само једна променљива. У том случају, предложено је проширење у смислу минимизације суме квадрата растојања унутар кластера; то јест минимализовати $trace(W)$ (што је уједно еквивалентно максимизовању $trace(B)$). Показује се [39] да је минимизација $trace(W)$ еквивалентна минимизовању суме квадратних Еуклидских растојања између појединачних опсервација и средишта њихових кластера, то јест да је:

$$trace(W) = E = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \overline{x_m})\,(x_{ml} - \overline{x_m})' = \sum_{m=1}^{g} \sum_{l=1}^{n_m} d_{ml,m}^2 \tag{4.16}$$

где је $d_{ml,m}$ Еуклидско растојање између l-те опсервације која се налази у кластеру m и средишта кластера m. Узимајући у обзир да се дијагонални елементи матрице једноставно сумирају, сви елементи који нису на дијагонали се игноришу (то јест, игноришу се сви коваријантни елементи) [57]. Ово подразумева да се све променљиве третирају као да су независне, то јест некорелисане. Дакле, вредност израза $trace(W)$ се мења са променом скале променљивих (из овог разлога различите методе стандардизације променљивих могу дати различите резултате, што је појашњено у претходним параграфима). Узимајући у обзир да $trace$ функција узима у обзир само елементе на дијагонали матрице W, долази до генерисања махом сферних кластера. Ово се може избећи трансформацијом променљивих (што је појашњено у претходним параграфима). Још један проблем до кога може доћи коришћењем минимизације вредности $trace(W)$ је и то да ова техника најчешће доводи до генерисања кластера приближно једнаких величина по броју опсервација. На овај начин ће природни велики кластери бити умањени, а природни мањи кластери бити повећани.

Постоје алтернативне методе за оптимизацију кластера, као што је метода израчуна детерминанте матрице квадратних растојања унутар кластера $det(W)$ [58], [59]. Овом методом се могу лоцирати не-сферични кластери у подацима, али се и даље могу јавити проблеми сличне величине кластера. Ипак треба имати у виду да је рачунање детерминанте матрице доста сложеније од рачунања вредности $trace$, па ће самим тим време потребно за оптимизацију кластера бити осетно веће.

Један од најпопуларнијих алгоритама за партитивну кластеризацију је алгоритам *K*-средина [60]. Један од основних разлога његове популарности је то што је време потребно за достизање конвергенције решења пропорцијално броју опсервација у улазном скупу података, што самим тим омогућава његово коришћење у великим скуповима података. Штавише, алгоритам *K*-средина је неодговарајући за изразито мале скупове података (са мање од 100 опсервација) из разлога што у таквим случајевима решење изразито зависи од редоследа опсервација у изворном скупу података. Промена редоследа опсервација може довести до драстично различитих предложених решења. Конкретно, може доћи до различитог избора почетних центроида. Овај ефект је познат као ефекат редоследа, а биће показано да и већи скупови података могу имати проблема услед истог ефекта при раду са алгоритмом *K*-средина. Три основна корака при извршавању алгоритма *K*-средина су:

1. Насумично изабрати почетне позиције *K* центроида – средишта кластера, то јест референтних вектора. Почетни центроиди могу бити и неке случајно изабране опсервације.

2. Свака опсервација се додељује најближем центроиду, дефинишући на тај начин привремене кластере. Постојећи центроиди се замењују средиштем одговарајућих привремених кластера. Процес се понавља докле год се не достигне конвергенција, то јест докле се позиције центроида значајно мењају у свакој итерацији. Алтернативно, процес се може зауставити и након одређеног броја итерација. На тај начин ће се добити локације финалних центроида.

3. Последњи пут анализирати све опсервације, и сваку опсервацију доделити најближем финалном центроиду.

Иако је рачунски најзахтевнији корак 2, основа алгоритма је у кораку 1, то јест у избору почетних референтних вектора. Такође, кључан елемент алгоритма *K*-средина представља избор вредности *K*, односно избор броја кластера у скупу података. Неке од метода могу бити:

1. Експертско познавање података и група које се могу јавити у њима (на пример, ако се анализирају деца која иду у основну школу и потребно је извршити њихову кластеризацију по разреду који похађају, логично је претпоставити да ће постојати осам група, по једна за сваки разред),

2. Једноставност даље обраде података које ће пружити метода анализе кластера (на пример, компанија која захтева кластер анализу својих клијената жели да их групише у тачно пет група, врло високи, високи, средњи, ниски и врло ниски потрошачи; у том случају број кластера треба бити постављен на пет)

3. Ограничења даље обраде података које ће пружити метода анализе кластера (на пример, компанија која захтева кластер анализу својих клијената има капацитет да клијентима пружи максимално три типа персонализованих понуда; тада би се кластеризација вршила само на три кластера),

4. Произвољан избор броја кластера,

5. Избор броја кластера на основу структуре самих података, нумеричком анализом.

4.3.5. Нумеричке методе за израчунавање броја кластера

Постоји велики број различитих нумеричких метода за израчунавање броја кластера које се могу користити у различитим применама анализе кластера. Већина метода је релативно неформална и суштински подразумева анализу графичког приказа неког критеријума кластеризације у зависности од различитог изабраног броја кластера [39]. Значајно велике и нагле промене које се могу приметити на графицима могу указивати на добар одабир броја кластера у подацима. Наравно, овакав метод процене је врло субјективан. Ипак, треба имати у виду да не постоји једна универзална метода за одређивање броја кластера у неком скупу података, најчешће се препоручује проналазак консензуса између више различитих метода (наравно, узимајући у обзир експертску

процену) [39]. У раду [61] је побројано чак 30 различитих метода за одређивање броја кластера што само наглашава значај овог процеса за успешну анализу кластера. Основна подела метода за одређивање броја кластера у подацима је на глобалне и локалне методе [41]. Глобалне мере су специфичне по томе да су оне базиране на прорачуну мере $G(c)$ која описује успешност поделе изворног скупа података на c кластера. Те мере су најчешће базиране на прорачуну варијације унутар кластера и варијације између кластера и идентификовању вредности $G(c)$ која је оптимална. Проблем оваквих дефиниција је да најчешће немају дефинисану вредност $G(1)$, то јест не могу да дају одговор на питање да ли је најоптималније уопште не делити улазни скуп података на било какве кластере. Локалне методе су, насупрот глобалним методама, базиране на анализи да ли се два конкретна кластера требају спојити (или аналогно, да ли се један постојећи кластер треба поделити на два мања). За разлику од глобалних метода, локалне методе су базиране на делу података који се анализира (изузев у првом кораку када се и анализира да ли се укупни скуп података уопште треба делити на два подскупа – кластера). Ипак, мана локалних метода је у томе што оне захтевају дефиницију неког прага (или ниво значајности), који би означавао да се подаци не требају даље делити. Најчешће је јако компликовано дефинисати овај праг узимајући у обзир да он може зависити од структуре самих података који су предмет анализе кластера. Неке од најзаступљенијих метода за процену броја кластера су:

1. кубни критеријум кластеризације (*cubic clustering criterion* - CCC) [62],
2. псеудо-F статистика (PSF) [63],
3. псеудо-T2 статистика (PST2) [64],
4. Билова F статистика [65].

Као што је наведено у претходним параграфима, широко распрострањени критеријум у процесу кластеризације је минимизација неке мере варијације унутар кластера (W), као на пример $trace(W)$. Максимизација дела варијације који се може објаснити (RSQ (R^2) вредности) се такође често користи. Имајући у виду да је $trace(T)$ константа за сваки изабрани кластер, RSQ вредност се може израчунати као:

$$RSQ = 1 - \frac{trace(W)}{trace(T)} \tag{4.17}$$

Кубни критеријум кластеризације [62] је глобална метрика која апроксимира дистрибуцију RSQ вредности, узимајући у обзир претпоставку да су кластери хипер-димензионе коцке у хипер-димензионој кутији. Иако је ова претпоставка најчешће погрешна, овај критеријум може бити коришћен за процену нулте хипотезе да су подаци униформно расподељени. Треба имати у виду да *CCC* метрика није одговарајућа за случајеве кластера неправилних облика, или уколико су улазне променљиве корелисане једне са другима. Да би се прорачунала *CCC* метрика, потребно је прво израчунати очекивану RSQ вредност [62]:

$$\mathrm{E}(RSQ) \cong 1 - \left[\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j} + \sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right] \tag{4.18}$$

где је n укупан број опсервација унутар скупа података, p број променљивих у скупу података, q број кластера, и:

$$u_j = \frac{s_j}{c} \tag{4.19}$$

где је s_j дужина хипер-димензионе коцке по j-тој димензији и

$$c = \left(\frac{v}{q}\right)^{\frac{1}{p}} \text{ и } v = \prod_{i=1}^{p} s_i \tag{4.20}$$

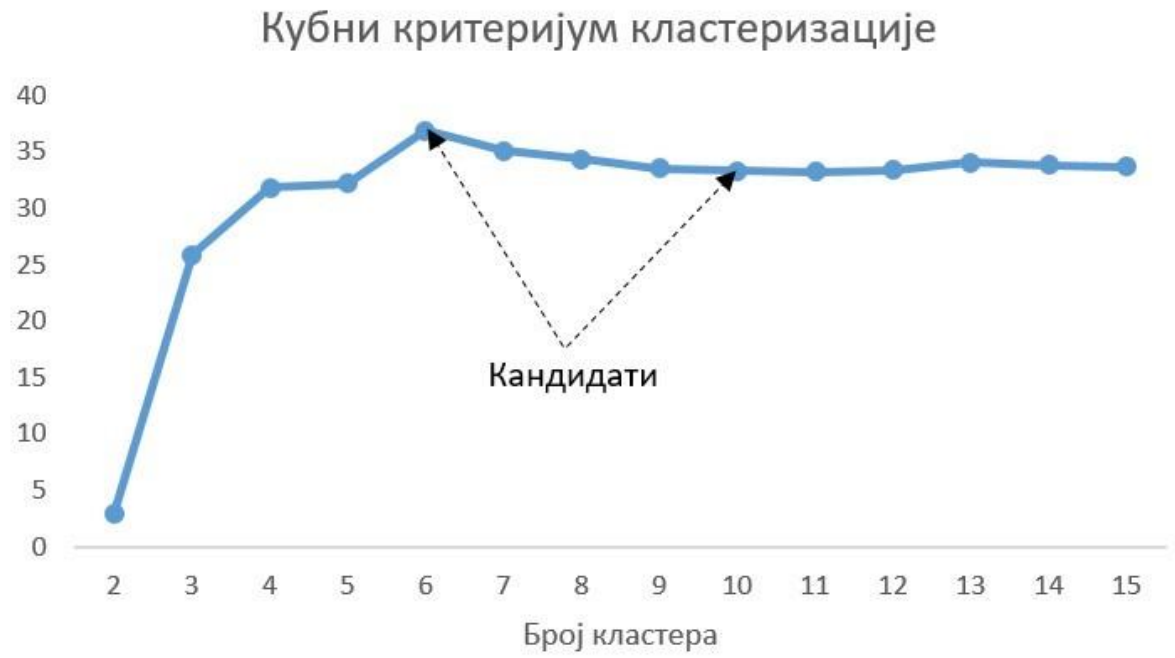
Други израз који је неопходан за израчунавање *CCC* метрике је *RSQ* вредност [62]:

$$RSQ = 1 - \frac{p^* + \sum_{j=p^*+1}^{p} u_j^2}{\sum_{j=1}^{p} u_j^2} \tag{4.21}$$

где је p^* параметар који описује димензионалност варијације између кластера (p^* мора бити мањи од укупног броја кластера q). На крају се *CCC* метрика израчунава помоћу емпиријске формуле која је дефинисана у покушају да се стабилизује варијанса за различити број опсервација, променљивих и кластера [62]:

$$CCC = ln\left[\frac{1-E(RSQ)}{1-RSQ}\right]\frac{\sqrt{\frac{np^*}{2}}}{(0.001+E(RSQ))^{1.2}} \tag{4.22}$$

Најбољи начин интерпретације *CCC* метрика је анализа графика зависности њених вредности од броја кластера у подацима. Генерално, локални максимуми на графику који имају вредност *CCC* метрике преко два би представљали потенцијални добар избор броја кластера. Локални максимуми који узимају вредности између нула и два означавају потенцијалне изборе, али их треба пажљиво тумачити. Уколико анализирани подаци имају хијерархијску структуру могуће је примећивање већег броја локалних максимума на графику. Битно је имати у виду да веома негативне вредности *CCC* метрике могу упућивати на проблеме са изузецима у подацима. Пример графичке интерпретације *CCC* методе је дат на слици 4.8.



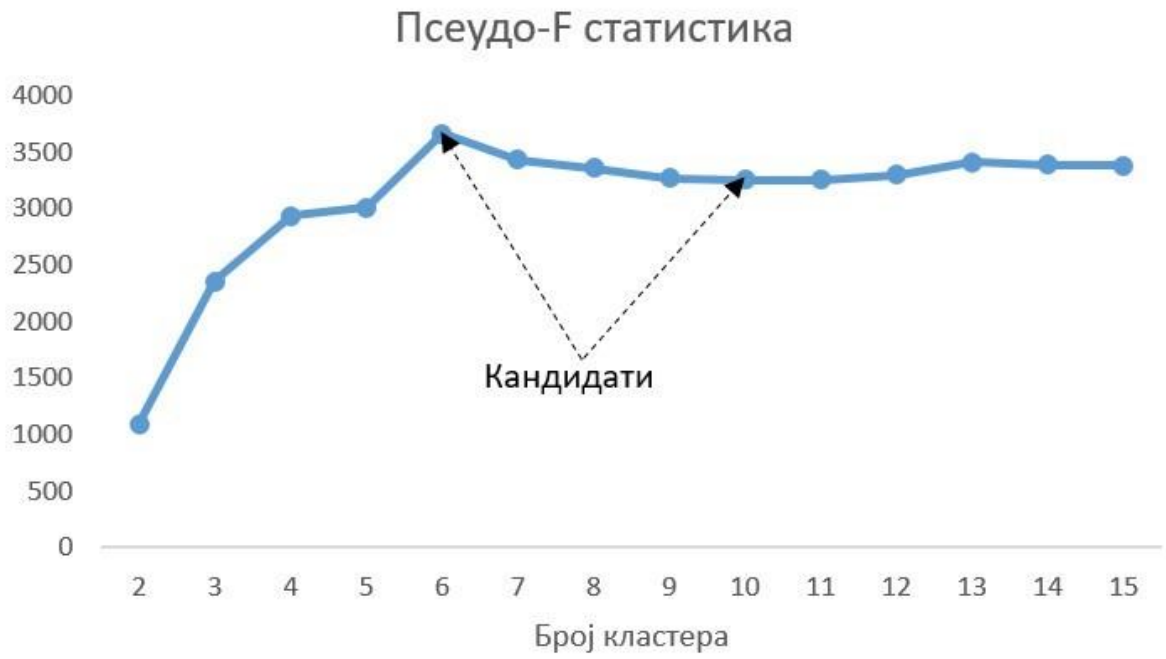
Слика 4.8: Интерпретација CCC методе одређивања броја кластера

Псеудо-F статистика (PSF) је глобална метода за процену броја кластера која покушава да квантификује меру сепарације између кластера. Ако *n* опсервација формира *g* кластера, и *B* представља суму квадрата растојања између кластера, а *W* је укупна сума квадрата растојања сваке опсервације унутар кластера са одговарајућим центроидом, вредност PSF метрике је дата са [63]:

$$PSF(g) = \frac{B/(g-1)}{W/(n-g)} = \frac{(T-W)/(g-1)}{W/(n-g)} \tag{4.23}$$

$$PSF = \frac{\left(\sum_{i=1}^{n}\|x_i - \bar{x}\|^2 - \sum_{i\epsilon C_j}\left\|x_i - \overline{x_j}\right\|^2\right)/(g-1)}{\left(\sum_{i\epsilon C_j}\left\|x_i - \overline{x_j}\right\|^2\right)/(n-g)}, \tag{4.24}$$

где је x_i вектор опсервације i, $\bar{x}$ је вектор средњих вредности свих променљивих у скупу података, а $\overline{x_j}$ је вектор средњих вредности свих променљивих у кластеру j. PSF метода се интерпретира графички, посматрањем локалних максимума на графику зависности PSF вредности од броја кластера; сваки локални максимум на графику који има задовољавајући број кластера може бити потенцијално решење. Пример графичке интерпретације PSF методе је дат на слици 4.9.



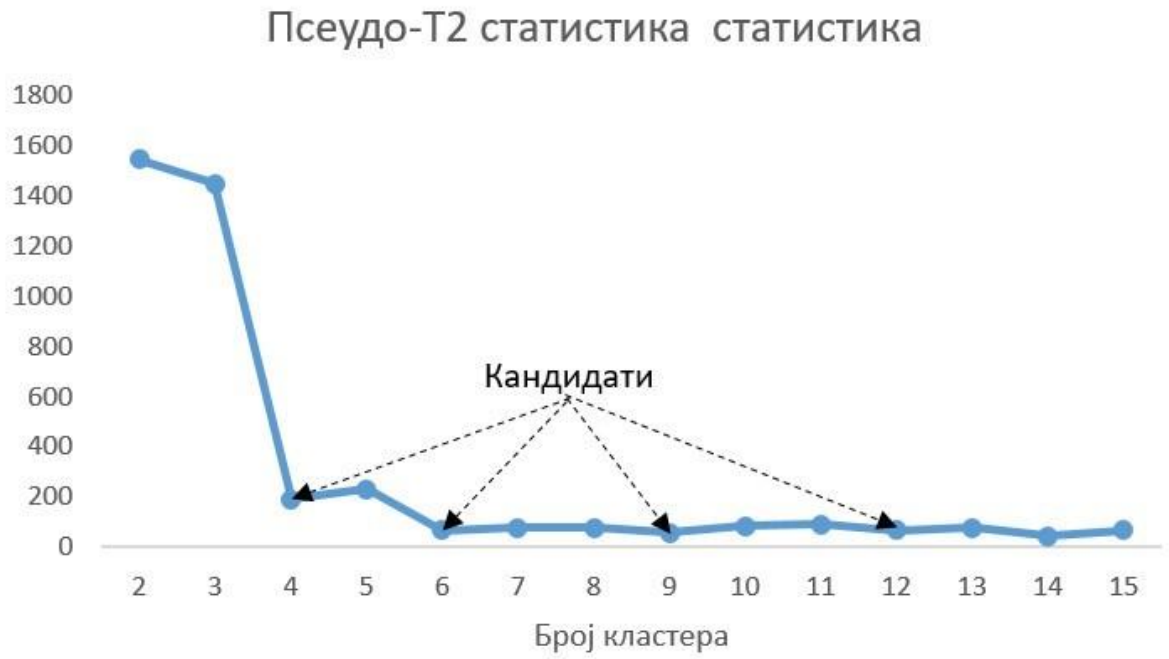
Слика 4.9: Интерпретација PSF методе одређивања броја кластера

Псеудо-T2 статистика (PST2) је локална метода, базирана на $J_e(2)/J_e(1)$ статистици [64]. Циљ PST2 алгоритма је провера да ли се два различита кластера требају спојити. Веће вредности PST2 статистике указују на то да су центроиди два посматрана кластера значајно различита и да се не требају спојити. Насупрот томе, ниска вредност PST2 метрике указује да се два анализирана кластера могу слободно спојити. Дакле, уколико би се PST2 метода графички анализирала, број кластера који одговара тачки на графику непосредно пре наглог скока би представљао потенцијално решење. Вредност PST2 метрике која се добија при спајању кластера *k* и *l* у кластер *m* је дата са:

$$PST2 = \frac{W_m - W_k - W_l}{(W_k + W_l)/(n_k + n_l - 2)}, \tag{4.25}$$

$$PST2 = \frac{\sum_{i\epsilon C_m}\|x_i - \overline{x_m}\|^2 - \sum_{i\epsilon C_k}\|x_i - \overline{x_k}\|^2 - \sum_{i\epsilon C_l}\|x_i - \overline{x_l}\|^2}{\left(\sum_{i\epsilon C_k}\|x_i - \overline{x_k}\|^2 + \sum_{i\epsilon C_l}\|x_i - \overline{x_l}\|^2\right)/(n_k + n_l - 2)}, \tag{4.26}$$

где је W_h укупна сума квадрата растојања сваке опсервације унутар кластера h од одговарајућег центроида, n_h представља број опсервација унутар кластера h, и $\overline{x_h}$ представља вектор центроида кластера h. Пример графичке интерпретације PST2 методе је дат на слици 4.10.



Слика 4.10: Интерпретација PST2 методе одређивања броја кластера

Још једна локална метода одређивања броја кластера која се добро показала на упоредном тесту предложеном у раду [61] је и Билова F статистика [65]. Она је базирана на поређењу некоригованих сума квадрата растојања између опсервација које припадају неком кластеру и њихових центроида (w_i за i-ти кластер). Уколико је некоригована сума квадрата w_1, за варијанту са c_1 кластера, дата са:

$$w_1 = \sum_{m=1}^{c_1} \sum_{l=1}^{n_m} (x_{ml} - \overline{x_m})\,(x_{ml} - \overline{x_m})' \tag{4.27}$$

много мања од некориговане суме квадрата за друго решење, w_2 (за варијанту са c_2 кластера), онда је прво решење боље од другог. Обрнуто, уколико је некоригована сума квадрата првог решења много већа од суме квадрата другог решења ($w_1 \gg w_2$), тада је друго решење оптималније. Билова F статистика одређује када је разлика између значајна:

$$F(w_1, w_2) = \frac{(w_1 - w_2)}{w_2} \frac{(n - c_2)k_2}{(n - c_1)k_1 - (n - c_2)k_2} \tag{4.28}$$

где је n укупан број опсервација унутар скупа података, c_1 број кластера у првој опцији (опцији са више кластера), c_2 број кластера у другој опцији (опцији са мање кластера), а w_1 и w_2 представљају некориговане сума квадрата растојања опсервација од центроида у првој и другој опцији. Коефицијенти k_1 и k_2 су редом једнаки:

$$k_1 = c_1^{(-2/p)} \tag{4.29}$$

$$k_2 = c_2^{(-2/p)} \tag{4.30}$$

где је p број променљивих у скупу података. Дакле, подела n опсервација у c_2 кластера је значајно боља од поделе на c_1 кластера (где је $c_2 > c_1$) уколико је тест статистика:

$$F(w_1, w_2) = \frac{(w_1 - w_2)}{w_2} \frac{(n - c_2)c_2^{(-2/p)}}{(n - c_1)c_1^{(-2/p)} - (n - c_2)c_2^{(-2/p)}} \tag{4.31}$$

већа од критичне вредности F-расподеле са $p(c_2 - c_1)$ и $p(n - c_2)$ степена слободе. Обично се за критичну вредност узима вредност 0.05. Уколико Билова F статистика није значајна, не постоји значајна разлика између предложених решења, те се по принципу Окамове бритве, треба изабрати једноставније решење које садржи мање кластера.

4.3.6. Хијерархијска кластеризација

Хијерархијске методе представљају окосницу кластеризације [39]. Велика популарност хијерархијских метода произилази из тога што оне, за разлику од партитивних метода не трпе последице различитог сортирања улазних података, то јест исти скуп улазних података ће бити смештен у исте кластере независно од начина на који су улазни подаци сортирани. Неке хијерархијске методе могу одредити кластере неправилног облика, понекад омогућавајући изостављање препроцесирања података. Наравно, једна од већих предности хијерархијске кластеризације је и то да код њихове употребе нема потребе за коришћењем метода за одређивање броја кластера у подацима.

Ипак, главна мана хијерархијских метода је њихово дуго време извршавања, што ограничава њихову употребу на кластеризацију података мале или средње величине. Такође, једно од питања при раду са хијерархијским методама кластеризације је избор праве мере за рачунање растојања између кластера. Постоји велики број различитих метода за прорачун растојања између кластера при коришћењу хијерархијске кластеризације и не може се рећи да је нека најбоља или да су све еквивалентне. Детаљна компарација различитих техника хијерархијске кластеризације које користе карактеристичне методе за прорачун растојања између кластера је дата у [66]. Методологија поређења перформанси описаних технологија је била базирана на Монте Карло симулацијама. Неке од различитих техника хијерархијске кластеризације су:

1. техника повезивања на основу просека (*average linkage*),
2. техника повезивања на основу центроида (*centroid linkage*),
3. техника потпуног повезивања (*complete linkage*),
4. техника једноструког (простог) повезивања (*single linkage*),
5. МекКитијева анализа сличности (*McQuitty's similarity analysis*),
6. техника повезивања на основу медијане (*median linkage*),
7. техника флексибилне бете (*flexible beta*),
8. техника повезивања на основу густине (*density linkage*),
9. Вардова техника минималне варијансе (*Ward's minimum variance*),

10. техника базирана на процени максималне вероватноће једнаке варијансе (*equal variance maximum likelihood*).

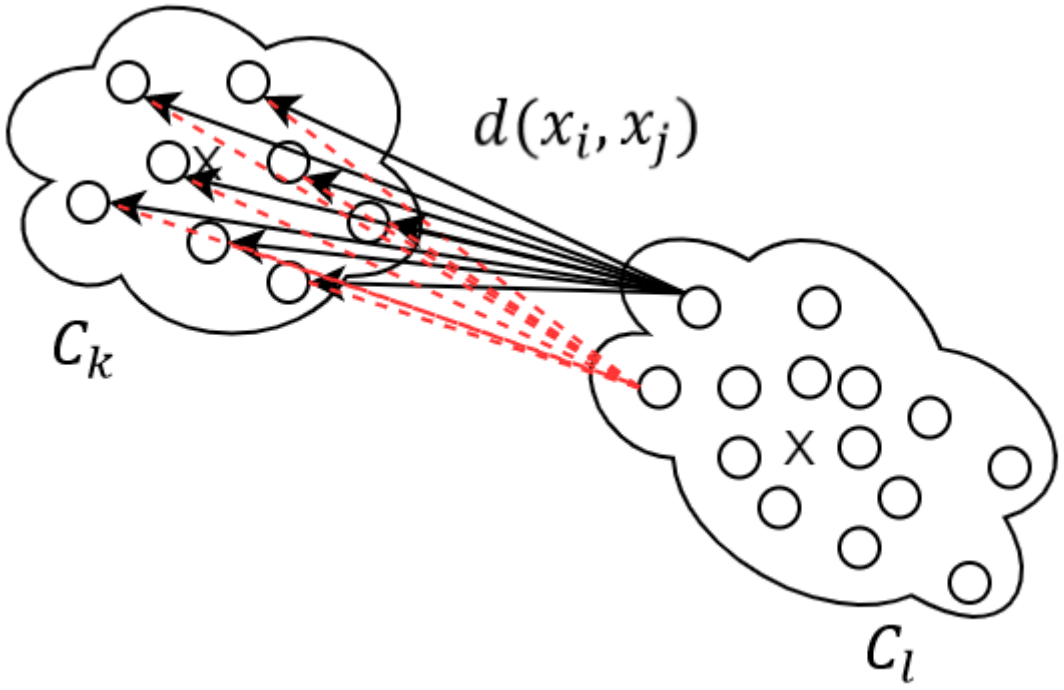
Различите технике суштински описују различите механизме проналажења сличности између индивидуе и групе, или између две групе. У мноштву техника које су доступне већина прихвата матрицу растојања као улазни податак. Из тог разлога, један од механизама за избор одговарајуће технике може бити и тип улазних података који је доступан, то јест евентуална доступност матрице растојања између опсервација. Уколико се користе „чисти“ изворни подаци (улазне променљиве, а не матрица растојања), већина техника ће у уводном кораку прерачунати матрицу растојања што може драстично увећати време процесирања.

Сакупљајуће технике хијерархијске кластеризације раде на принципу спајања кластера који су најсличнији. Све технике које су побројане изнад представљају само различите начине рачунања сличности, или прецизније растојање између кластера. Треба имати у виду да већина горепобројаних техника претпоставља да је структура кластера сферична, те би за неке технике било корисно трансформисати податке по претходно описаним процедурама за превазилажење проблема не-сферичних кластера. Ипак, три технике: техника повезивања на основу густине, техника повезивања на основу густине из два корака и техника једноструког (простог) повезивања, могу директно радити обраду неправилних кластера и не захтевају препроцесирање.

Техника повезивања на основу просека [67] дефинише сличност између два кластера као просечно растојање између сваког пара опсервација које припадају тим одговарајућим кластерима. Дакле, растојање између кластера C_k и кластера C_l дато је једначином:

$$D_{kl} = \frac{1}{n_k n_l} \sum_{i \in C_k} \sum_{j \in C_l} d(x_i, x_j) \tag{4.32}$$

где је n_k број опсервација у кластеру C_k, n_l број опсервација у кластеру C_l, $d(x_i, x_j)$ растојање између опсервација i и j. Техника повезивања на основу просека узима у обзир сва растојања између свих парова опсервација у два различита кластера, па је самим тим под мањим утицајем изузетака у подацима у односу на друге хијерархијске технике. Такође, ова техника може баратати директно са изворним подацима и не захтева комплетну матрицу растојања. Техника повезивања на основу просека је често бржа у односу на остале хијерархијске технике и добро се показала у компаративној анализи приказаној у [66]. Графички приказ рада технике повезивања на основу просека је дат на слици 4.11.

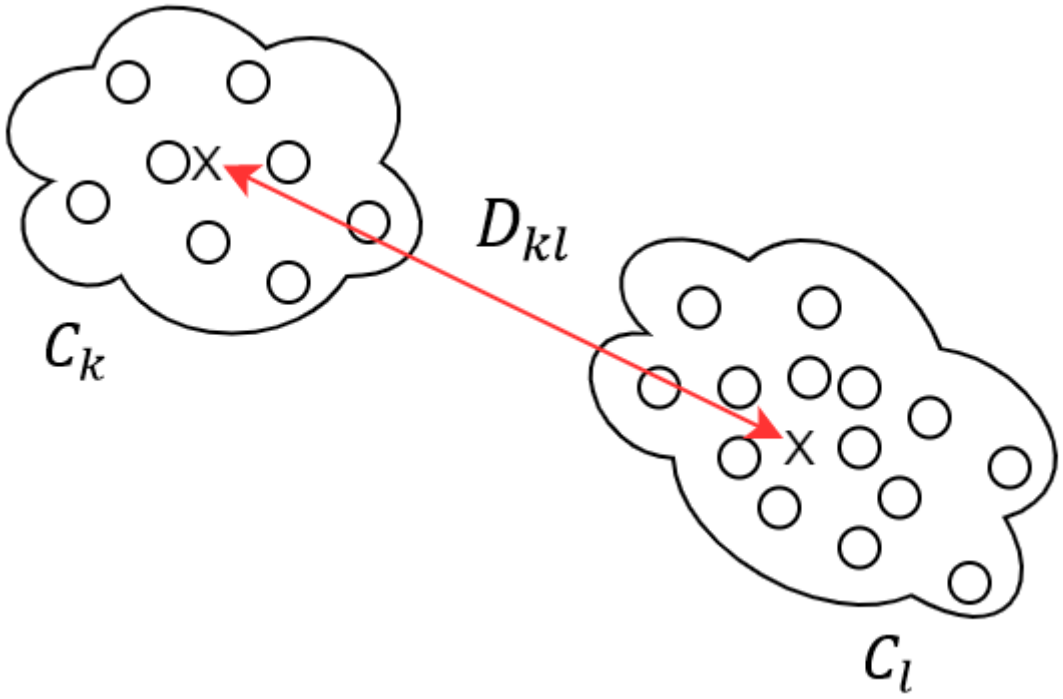


Слика 4.11: Приказ рада технике повезивања на основу просека

Најинтуитивнију технику хијерархијског повезивања кластера представља техника повезивања на основу центроида [67]. Она подразумева да је растојање између два кластера једнако растојању између њихова два карактеристична вектора ($\overline{x_k}$ и $\overline{x_l}$), то јест центроида:

$$D_{kl} = \|\overline{x_k} - \overline{x_l}\|^2 = \sum (\overline{x_k} - \overline{x_l})^2. \tag{4.33}$$

Узимајући у обзир да ова техника пореди центроиде два кластера, при чему се центроид генерише на основу свих опсервација унутар посматраног кластера, неће бити претерано осетљива на изузетке у подацима. Такође, ова техника је релативно брза и може радити директно са изворним подацима, али се Монте Карло симулацијама није показала подједнако успешном као техника повезивања на основу просека [66]. Још једна мана ове технике је да већи од два кластера који се спајају има тенденцију да „доминира“ над мањим кластером [68]. Заправо, ако две групе које се спајају имају јако различит број елемената (опсервација), онда ће карактеристични вектор нове заједничке групе бити драстично сличнији карактеристичном вектору изворно веће групе. Графички приказ рада технике повезивања на основу центроида је дат на слици 4.12.

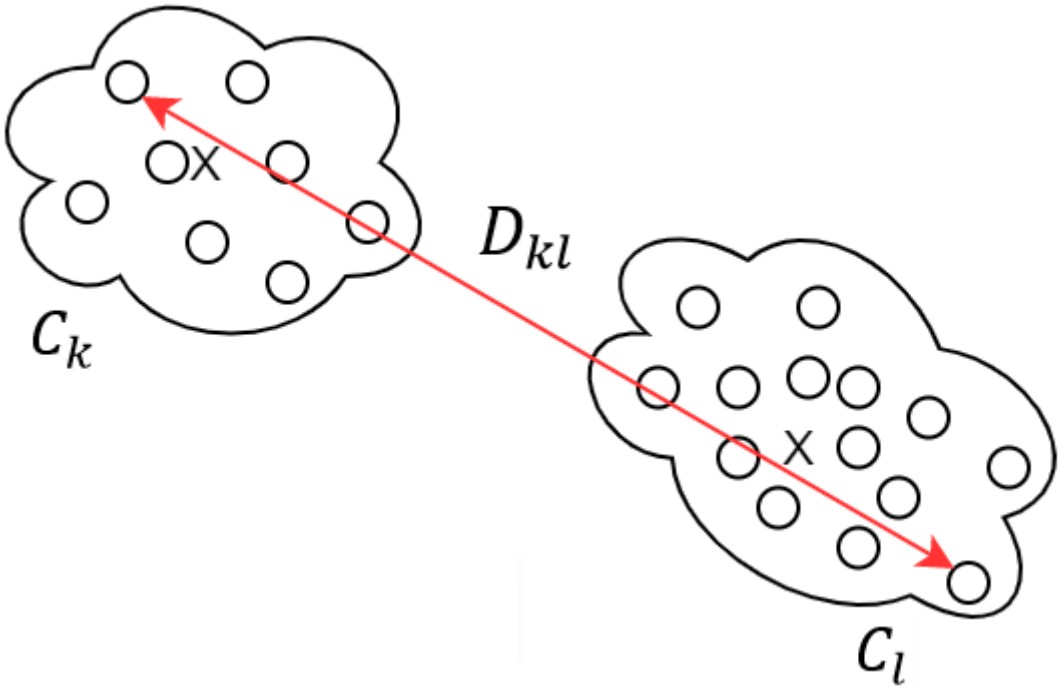


Слика 4.12: Приказ рада технике повезивања на основу центроида

Техника потпуног повезивања дефинише растојање између кластера као максимално растојање између опсервација у једном и другом посматраном кластеру [69]:

$$D_{kl} = i \in C_k \; j \in C_l \max d(x_i, x_j). \tag{4.34}$$

Техника потпуног повезивања има тенденцију да формира кластере са приближно истим дијаметрима, а такође је и врло осетљива на присуство изузетака у подацима. Графички приказ рада технике потпуног повезивања је дат на слици 4.13.

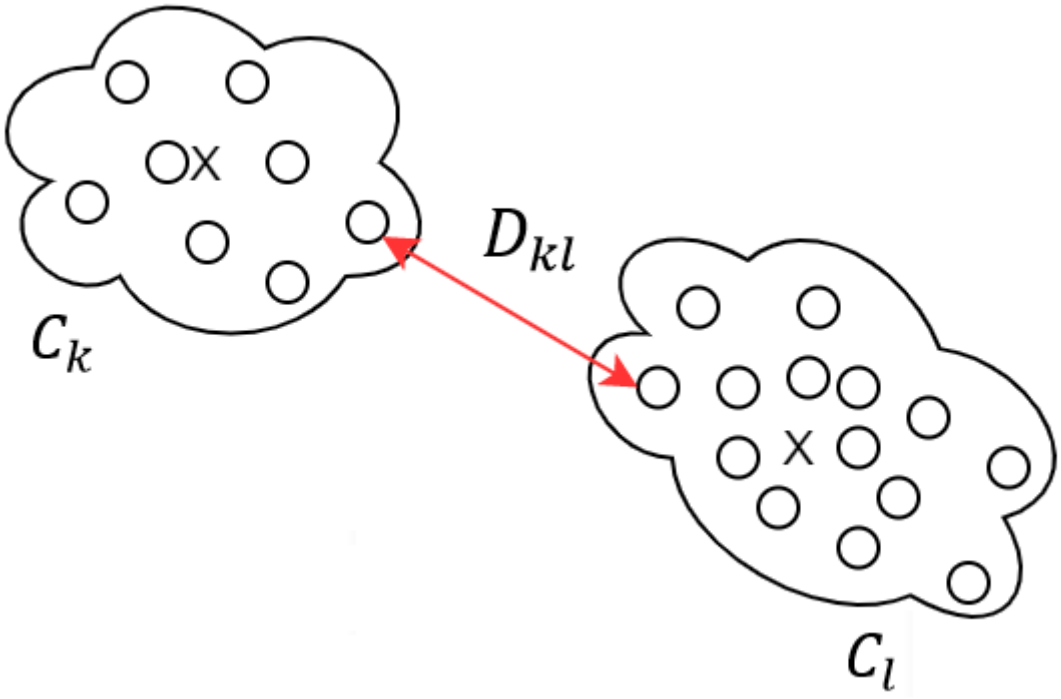


Слика 4.13: Приказ рада технике потпуног повезивања

Техника једноструког (простог) повезивања [70] дефинише растојање између два кластера као минимално растојање између опсервација у једном и другом посматраном кластеру:

$$D_{kl} = i \in C_k \; j \in C_l \min d(x_i, x_j). \tag{4.35}$$

Узимајући у обзир да техника простог повезивања не уводи никакве претпоставке ни ограничења по питању облика кластера, овом техником се могу директно идентификовати кластери не-сферичног облика. Иако техника простог повезивања има многе жељене теоријске особине, показала се релативно лоше у симулацијама и студијама. Чак штавише, јако често се дешава да најлошије перформансе има управо ова техника [66]. Главни разлог за ово је феномен „уланчавања“, то јест тенденција да се опсервације укључе у постојеће кластере уместо да дође до дефинисања нових (засебних) кластера [71]. Графички приказ рада технике простог повезивања је дат на слици 4.14.



Слика 4.14: Приказ рада технике простог повезивања

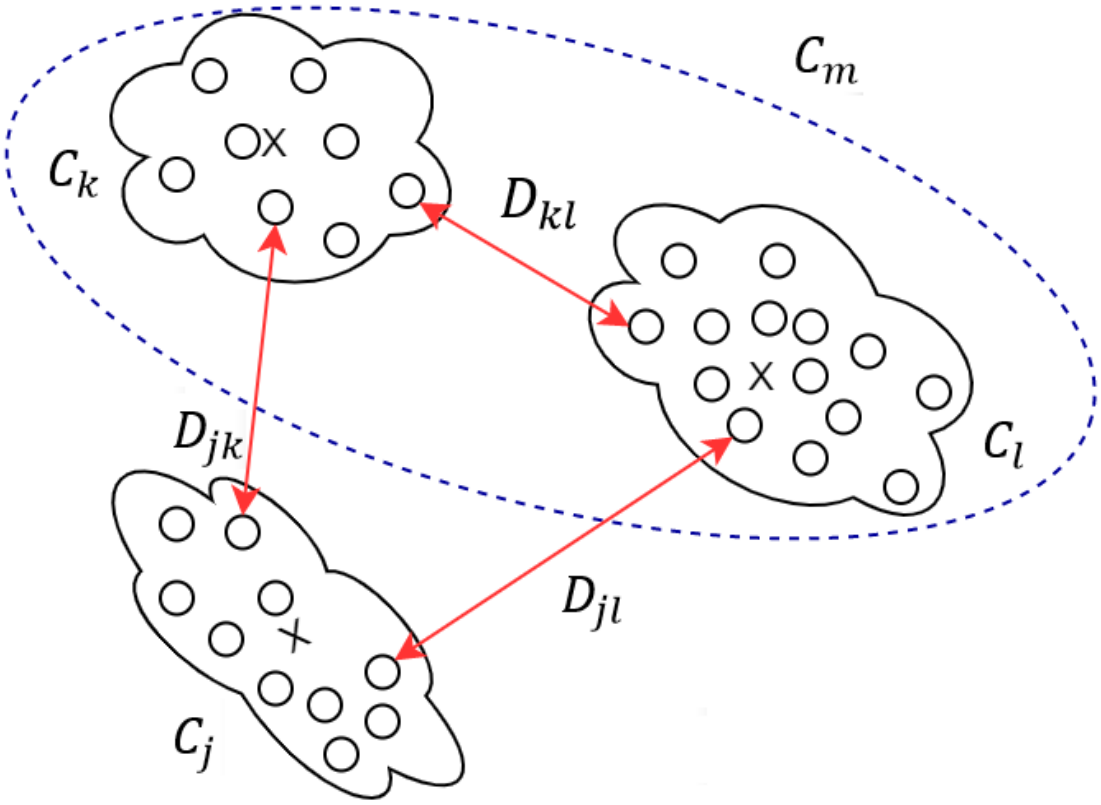
МекКитијева анализа сличности [72], техника повезивања на основу медијане [73] и техника флексибилне бете [74] су базиране на сличним принципима. Суштински, врши се рачунање растојања од новоформираног кластера M (који настаје спајањем кластера K и L) и неког другог кластера J.

$$D_{kl,McQuitty} = \frac{D_{jk} + D_{jl}}{2} \tag{4.36}$$

$$D_{kl,Median} = \frac{D_{jk} + D_{jl}}{2} - \frac{D_{kl}}{4} \tag{4.37}$$

$$D_{kl,Beta} = \left(D_{jk} + D_{jl}\right)\frac{(1-b)}{2} + D_{kl}b \tag{4.38}$$

где је b бета вредност за коју важи $0 < b < -1$, а компоненте D_{jk}, D_{jl} и D_{kl} се прерачунавају на основу улазне матрице растојања. Подразумевана вредност параметра b је -0.25 [74], иако [75] препоручује вредност -0.5. Битно је приметити да је МекКитијева техника посебан случај технике флексибилне бете за $b = 0$. За разлику од МекКитијеве анализе сличности и технике повезивања на основу медијане који нису били успешни у тесту, техника флексибилне бете је постигла обећавајуће резултате у симулационој студији [66]. Графички приказ рада технике простог повезивања је дат на слици 4.15.



Слика 4.15: Приказ рада МекКитијеве анализе сличности, технике повезивања на основу медијане и технике флексибилне бете

Техника повезивања на основу густине је процес из два корака. У првом кораку, нека непараметарска процена густине опсервација се користи за прорачунавање нове мере растојања, $d^*(x_i, x_j)$, и након тога се у другом кораку генерише финални прорачун кластера применом технике једноструког (простог) повезивања помоћу новоизрачунатих мера растојања $d^*(x_i, x_j)$. Постоји више техника за процену густине које се могу користити, попут оних које су представљене у радовима [76] и [77]. Техника *K*-најближих суседа [76] користи одређени број „суседа“ (блиских опсервација) за процену густине. Прецизније, процена густине опсервација коришћењем технике *k*-најближих суседа је дефинисана као:

$$d^*(x_i, x_j) = \begin{cases} \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) & \text{ако је } d(x_i, x_j) \leq \max(r_k(x_i), r_k(x_j)) \\ \infty & \text{иначе} \end{cases} \tag{4.39}$$

где је $r_k(x)$ растојање од неке опсервације x до њеног k-тог најближег суседа, $d(x_i, x_j)$ представља Еуклидско растојање између тачака x_i и x_j, а $f(x_i)$ је однос између броја опсервација (k, које је дефинисано као улазни параметар) унутар затворене сфере чији се центар налази у тачки x_i са радијусом $r_k(x_i)$ и запремине сфере (v_i):

$$f(x_i) = \frac{k}{nv_i} \tag{4.40}$$

где је n укупан број опсервација.

Вонгова хибридна техника [77] користи прелиминарне резултате које израчунава техника k-најближих суседа. На основу прелиминарних резултата су израчунати C_k (састав k прелиминарних кластера), n_k (број опсервација унутар k прелиминарних кластера) и W_k (сума квадрата растојања унутар k кластера), као и средишта прелиминарних кластера. Прелиминарни кластери C_i и C_j се сматрају суседни уколико је средиште између њихових центроида ($\overline{x_{lj}}$ је средиште између $\overline{x_l}$ и $\overline{x_j}$) ближе једном од центроида $\overline{x_l}$ или $\overline{x_j}$ него било ком другом прелиминарном центроиду. Еквивалентан услов је:

$$d^2\left(\overline{x_l}, \overline{x_j}\right) < d^2\left(\overline{x_l}, \overline{x_v}\right) + d^2\left(\overline{x_j}, \overline{x_v}\right) \tag{4.41}$$

за све друге прелиминарне кластере C_v такве да је $C_i \neq C_v$ и $C_j \neq C_v$. Тада је растојање између два прелиминарна кластера C_i и C_j једнако:

$$d^*\left(\overline{x_l}, \overline{x_j}\right) = \begin{cases} \dfrac{\left(W_i + W_j + \frac{1}{4}\left(n_i + n_j\right)d^2\left(\overline{x_l}, \overline{x_j}\right)\right)^{\frac{p}{2}}}{\left(n_i + n_j\right)^{1+\frac{p}{2}}} & \text{ако су } C_i \text{ и } C_j \text{ суседни} \\ \infty \quad \text{иначе} \end{cases} \tag{4.42}$$

где је p број променљивих. Пошто не захтева никакве претпоставке облика кластера, техника повезивања на основу густине може директно генерисати кластере неправилног облика. Ипак, ова техника није претерано ефикасна при анализи компактних кластера у поређењу са претходно наведеним техникама. Такође, ова техника захтева дефинисање броја кластера који ће бити дефинисани у прелиминарној анализи. Предложен број прелиминарних кластера је $n^{0.3}$ [77], где је n укупан број опсервација у подацима.

Вардова техника минималне варијансе [78] представља једну од најпопуларнијих техника хијерархијске кластеризације. Вардова техника хијерархијски спаја кластере на начин да, у свакој итерацији, сума квадрата растојања унутар кластера буде најмања од свих могућих опција спајања нека два кластера из претходне итерације. Прецизније, растојање између два кластера C_k и кластера C_l дато је једначином:

$$D_{kl} = \frac{\|\overline{x_k} - \overline{x_l}\|^2}{\left(\frac{1}{n_k} + \frac{1}{n_l}\right)} = \frac{\sum_i (\overline{x_{ki}} - \overline{x_{li}})^2}{\left(\frac{1}{n_k} + \frac{1}{n_l}\right)} \tag{4.43}$$

где је $\overline{x_k}$ центроид кластера C_k, $\overline{x_l}$ центроид кластера C_l, а n_k и n_l представљају број опсервација у кластерима C_k и C_l. Вардова техника минималне варијансе може да обрађује чисте улазне податке (координатне податке), па може да ради брже од других хијерархијских техника, посебно на рачунарима који имају ограничену меморију. Основна мана ове технике је то што има тенденцију да спаја кластере мањих димензија како би формирала финалне кластере приближно истих димензија. Такође, Вардова техника је јако осетљива на изузетке у подацима [66].

Техника базирана на процени максималне вероватноће једнаке варијансе [79], је слична Вардовој техници са разликом да уклања негативну карактеристику да формира кластере истих димензија. Формула за растојање између кластера C_k и кластера C_l који се спајају у кластер C_m је дата једначином:

$$D_{kl} = np \ln\left(1 + \left(\frac{W_m - W_k - W_l}{P_g}\right)\right) - p_k(n_m \ln(n_m) - n_k \ln(n_k) - n_l \ln(n_l)) \quad (4.44)$$

$$P_g = \sum_{i=1}^{G} W_i \quad (4.45)$$

где је n_i број опсервација унутар кластера i, W_i сума квадрата растојања унутар кластера i, p број променљивих, n укупан број опсервација, G укупан број кластера на посматраном нивоу хијерархије. Као што је већ напоменуто, насупрот Вардовој техници, техника базирана на процени максималне вероватноће једнаке варијансе формира кластере различитих величина. Фактор p_k (који је најчешће постављен на вредност 2) пружа могућност да се смањи овај ефекат.

4.3.7. Непараметарска кластеризација

Већина метода кластеризације формирају кластере одређеног типа (одређене величине или облика). На пример алгоритам *K*-средина и друге сличне методе базиране на методи најмањег збира квадрата формирају кластере који имају приближно исти број опсервација. Са друге стране, методе које су најмање склоне формирању кластера одређеног облика су базиране на непараметарској процени густине опсервација [80], [81]. Уколико су матрице коваријанси неједнаке или драстично не-нормалне, непараметарска процена густине је често најбољи приступ. Један од предуслова за развој модела методом непараметарске кластеризације је разумно велики број опсервација у улазном скупу података (ова метода не може радити када нема довољно података у улазном скупу података). Још једна предност метода непараметарске кластеризације је то да су оне мање осетљиве на проблеме промена скале података (који су описани у поглављу 4.3.3. припрема података за кластеризацију). Такође, метод непараметарске кластеризације не захтева унапред број кластера као параметар. Захтева се само параметар изравнавања (*smoothing parameter*) ради генерисања кластера. Треба имати у виду да већи параметар изравнавања доводи до генерисања мање различитих кластера у истом скупу података. Не постоји универзални одговор на питање који параметар изравнавања користити при дефинисању модела, најчешће треба симулирати поступак са више различитих опција и изабрати ону која даје решење које има највише смисла.

Општи кораци при раду метода непараметарске кластеризације су:

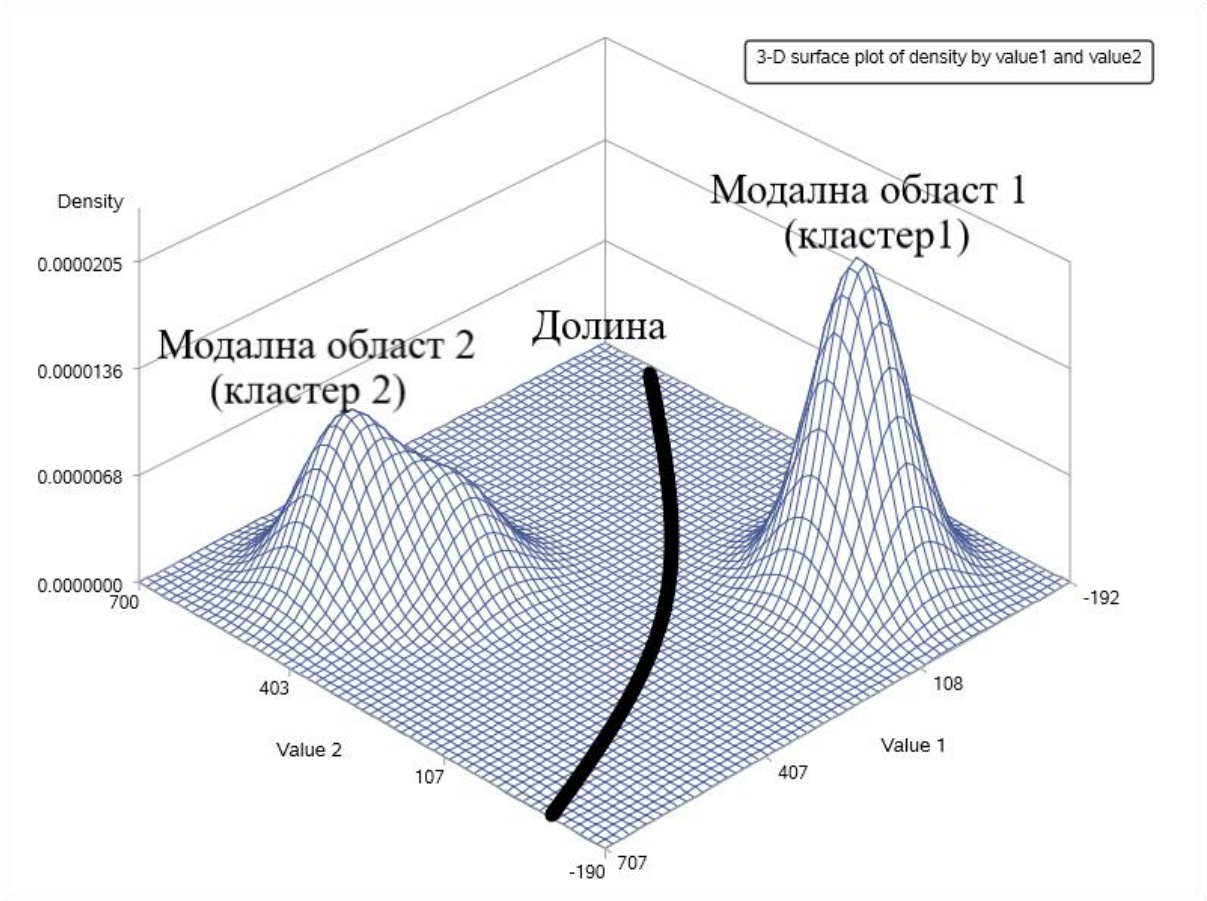
1. процена иницијалне густине коришћењем униформних кернела фиксног радијуса,

2. генерисање прелиминарних кластера коришћењем методе „тражења долине“

3. рачунање одговарајуће *p*-вредности за сваки кластер поређењем максималне густине у кластеру са максималном густином на граници кластера,

4. понављање корака три докле год *p*-вредности нису испод предефинисане вредности или докле не остане само један кластер (све опсервације у једном кластеру).

У непараметарској кластеризацији, кластер представља скуп опсервација које праве модалну област (*modal region*), где модална област подразумева вероватноћу да ће се случајно изабрана тачка наћи унутар хиперсфере радијуса *r* од тачке *x*. Дакле, кластер би представљао неку област око локалног максимума функције густине вероватноће. На пример, уколико се у скупу података налазе два кластера, анализом тог скупа података би се пронашла две модалне области и једна долина (локални минимум) који их раздваја. Дакле, управо та долина би представљала границу која дели два кластера, те би главни задатак алгоритма представљало управо дефиниција долина, то јест „тражење долине“. Графички приказ функције густине вероватноће скупа података са два кластера је дат на слици 4.16.



Слика 4.16: Приказ функције густине вероватноће скупа података са два кластера

4.3.8. Пост-процесирање анализе кластера

Након што је изведена успешна кластер анализа улазног скупа података коришћењем неке од техника описаних у претходним поглављима, поставља се питање интерпретације добијених резултата – профилирање кластера. Циљ профилирања кластера је додељивање јединствених описа сваком од добијених кластера. Ти описи могу бити једноставна ознака типа или врсте (на пример у кластер анализи животиња добијени кластери могу бити пси, мачке, птице и слично), или сложени називи (попут

кластер анализе купаца који би се делили на групе: „млади велики купци“, „породични купци робе широке потрошње“, „штедљиви купци“ и слично). Називи се не додају кластерима само из разлога додељивања описа кластера, већ се ти називи могу користити за поређење различитих метода кластеризације. На пример, уколико једна метода кластеризације издваја неку групу која нам је од интереса а друга не, очигледно је да као финалну опцију треба изабрати ону методу која нам пружа прецизније дефинисане кластере. Постоје више метода за профилирање кластера. Једна од најчешће коришћених техника је базирана на поређењу центроида свих добијених кластера. Наиме, на тај начин се може приметити да ли опсервације унутар једног кластера осетно одударају од просека популације и / или од других кластера.

Поређење центроида кластера међусобно може бити корисно када не постоји велики број кластера који је добијен неком методом. Ипак, када је добијени број кластера изразито велики, могу се јавити проблеми. Такође, уколико је жеља да се профилише само један део од великог броја добијених кластера, међусобно поређење не може бити опција. Тада се користи поређење центроида појединачних кластера и просека целе изворне популације. На овај начин ће бити наглашене разлике просечне опсервације и сваког појединог анализираног кластера.

Ова стратегија профилирања кластера је базирана на следећим корацима:

1. Свака доступна опсервација се класификује помоћу једне додатне ознаке (лабеле) на начин да ће ознака имати вредност један (односно нула) уколико посматрана опсервација припада (односно не припада) кластеру који се профилише,

2. Коришћењем метода логистичке регресије формирати помоћни модел који ће на основу улазних променљивих предвиђати вредност додатне ознаке уведене у кораку један,

3. Лоцирати које су улазне променљиве најзначајније у моделу логистичке регресије који је формиран у кораку два,

4. Формирати хистограмску расподелу најзначајнијих променљивих опсервација које припадају посматраном кластеру чије се профилирање врши, као и опсервација које не припадају посматраном кластеру.

Ипак, треба имати у виду да кластер анализа не мора увек резултовати интерпретабилним кластерима. Уколико се догоди да ниједна од улазних променљивих не генерише значајне разлике од просека популације и / или других кластера, потенцијално може бити потребно преиспитати избор конкретне методе и других параметара кластеризације.

Један од проблема који се такође може јавити након комплетиране кластер анализе може бити и додељивање нових опсервација неким од постојећих (претходно дефинисаних) кластера. За те потребе се користе центроиди, на начин да се свака нова опсервација додељује центроиду којем је најближа, при чему је блискост дефинисана на исти начин као и при формирању кластера. На пример, ако се посматра случај у коме је кластер анализа формирала два финална кластера, један који означава клијенте који су склони малверзацијама, док други кластер подразумева клијенте који нису склони преварама. Сваки нови клијент би се поредио са центроидима једног и другог кластера

и на тај начин би се процењивала вероватноћа да ће клијент извршити неку малверзацију у будућности. Дакле, иако кластеризација не садржи означене податке (то јест циљне податке), кластер анализа може служити и као предиктивни алат.

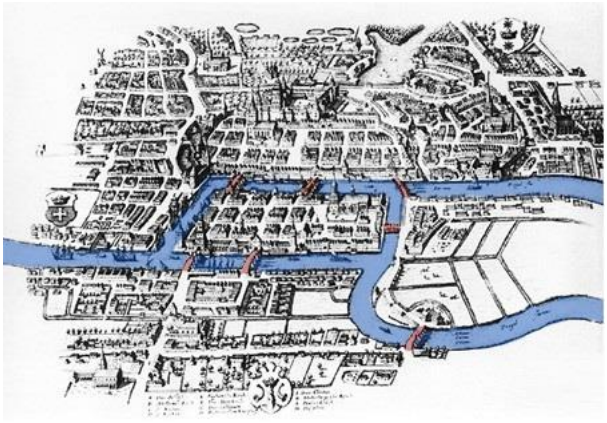
5. АНАЛИЗА ДРУШТВЕНИХ МРЕЖА И ТЕОРИЈА ГРАФОВА

Теорија графова представља алат којим се могу верно описати карактеристике чворова разних комплексних мрежа, попут оне која се анализира у овој докторској дисертацији. Метрикама теорије графова се може квантификовати повезаност и значајност чворова у некој комплексној мрежи, при чему свака појединачна метрика може описати један засебан аспект сваког чвора. У овом поглављу биће дат преглед основних карактеристика теорије графова, као и дефиниције метрика од интереса које ће се користити у истраживању које представља окосницу ове докторске дисертације.

5.1. Историјат теорије графова и проблем Кенигсбергшких мостова

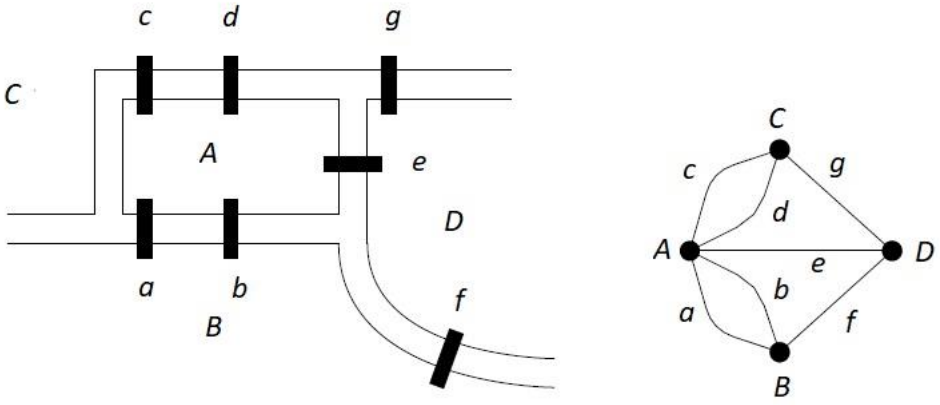
Почеци и историјат теорије графова су били врло скромни [82]. Док су друге гране математике биле мотивисане махом фундаменталним рачунским проблемима, проблеми који су довели до развоја теорије графова су били замишљени да развијају домишљатост. Али насупрот привидној тривијалности таквих проблема, теорија графова је остварила велики број теоретских резултата.

Један од најстаријих проблема који се појавио у развоју теорије графова је проблем Кенигсбергшких (Калињинградских) мостова [82]. На слици 5.1. је приказана мапа града Кенигсберга, то јест Калињинграда, и реке Прегел која протиче кроз Кенигсберг. Као што се види са слике, река Прегел формира једно острво након којег се дели на две гране. Да би се омогућило грађанима Кенигсберга да путују са једног краја града на други, река је премошћена са седам мостова. Једна од недоумица која се јавила посматрањем овакве архитектуре је била да су грађани Кенигсберга покушавали да открију руту којом би могли да се крећу градом користећи сваки од седам мостова. Ипак, ови покушаји нису успевали, што је многе навело на размишљање да то није могуће. Тек је 1736. проблем Кенигсбергшких мостова третиран из математичког угла, и тек тада је коначно доказано да механизам проналаска такве руте кретања по граду не постоји. Наиме, Леонард Ојлер је у раду [83] представио метод којим се тај конкретан проблем (али и други методи сличног типа) могу решити.



Слика 5.1: Приказ града Кенигсберга (црвеном бојом су означени мостови преко реке Прегел) [84]

Ојлерово решење проблема Кенигсбергшких мостова се састоји од два основна корака. У првом је формирана упрошћена шема града која је приказана на слици 5.2. Дефинисане су четири области означене симболима *A*, *B*, *C* и *D*, као и седам мостова означених са *a*, *b*, *c*, *d*, *e*, *f* и *g*. Ова упрошћена шема представља граф. Други корак представља проналажење специјалног пута унутар овог графа. Ипак, да би се Ојлерово решење проблема разјаснило, потребно је прво увести дефиниције основних појмова. По дефиницији, граф би представљао коначни скуп чворова и грана, као и скуп правила које нам објашњавају који је пар чворова повезани којом граном [85]. Обично, грана повезује различите чворове, али се може догодити и да једна грана повезује два иста чвора; у том случају говори се о петљи. Под путањом би се дефинисала секвенца чворова и грана $v_0, e_1, v_1, e_2, \ldots, v_{r-1}, e_r, v_r$ у којој свака грана e_i спаја чворове v_{i-1} и v_i (где је $1 \leq i \leq r$). У Ојлеровом примеру, постоје четири чвора и седам грана, док би правила говорила да грана *a* повезује чворове *A* и *B*, грана *c* чворове *A* и *C*, и слично. Ради илустрације проблема теорије графова, формирају се шематски прикази попут оног приказаног на слици 5.2. На таквим шематским приказима, сваки чвор би био представљен тачком, а грана неком дужи која спаја две тачке.



Слика 5.2: Ојлеров упрошћени шематски приказ проблема Кенигсбергшких мостова

Коришћењем терминологије која је представљена у претходном параграфу, проблем Кенигсбергшких мостова се може дефинисати као потрага за путањом у графу

који има особину да сваку грану садржи тачно једанпут. Путања тог типа је позната под називом Ојлерова путања [85].

Да би се појаснило Ојлерово решење, потребно је увести још неколико појмова. Граф је повезан уколико су било која два чвора из посматраног графа повезани неком путањом. Са друге стране граф је неповезан уколико се састоји од више повезаних графова, то јест компоненти повезаности [85]. Такође, под степеном неког чвора се подразумева број грана које садрже анализирани чвор. У примеру Кенигсбергшких мостова, степен чвора *A* је пет, док је степен свих других чворова једнак три. Може се приметити да уколико се саберу степени свих чворова унутар једног повезаног графа, добија се збир који је једнак двоструком броју чворова у графу. Ова тврдња је наведена у [83], а потом и проширена на начин да је речено да у сваком графу, број чворова који имају непарни степен мора бити паран. Последња тврдња се у литератури [86] може наћи и под називом „лема руковања“, пошто она тврди да би се на забави у којој се гости међусобно рукују морало бити паран број костију који су се руковали непаран број пута.

Финална тврдња Ојлеровог решења проблема Кенигсбергшких мостова је да уколико повезани граф има више од два чвора непарног степена, неће моћи да садржи Ојлерову путању. Самим тим, узимајући у обзир „лему руковања“, закључено је да повезани граф који садржи Ојлерову путању мора садржати или ниједан или два чвора непарног степена. С обзиром да граф из примера Кенигсбергшких мостова садржи чак четири чвора непарног степена, очигледно је да тај граф не садржи Ојлерову путању, те да се самим сви мостови града Кенигсберга не могу обићи на начин да се сви мостови пређу тачно једанпут. Интересантно је да Ојлер у свом раду није оставио доказ ове тврдње, те да је доказ први пут објављен тек 1873. године. Такође, може се запазити да би граф приказан у проблему Кенигсбергшких мостова садржао Ојлерову путању уколико би се изградио мост који би повезивао области *B* и *C*.

5.2. Теорија графова и комплексне мреже

Напад на Светски трговински центар 11. септембра 2001. године је, поред напада на Њујорк, представљао и својеврсан напад на интернет као целину. Наиме, три јако битна трансатлантска интернет кабла, као и станица за рутирање саобраћаја се налазе јако близу зграда које су се обрушиле те су нападом на Светски трговински центар претрпеле значајну штету. Истраживања која се баве перформансама на интернету су скренула пажњу да је доступност сервера на интернету пала за чак 9% непосредно након напада [85]. Ипак, након пола сата вредност се вратила на вредност која је била актуелна пре напада. Овај податак може показати да, иако је витални део инфраструктуре интернета био у прекиду, цела мрежа је успела да настави да функционише са високим капацитетима. Такође, види се и да је након свега пола сата мрежа била поново на својим оптималним вредностима перформанси, што потврђује робусност интернета као мреже и својеврсну могућност самоизлечења мреже у ванредним ситуацијама коришћењем савремених протокола рутирања података.

Интернет се може дефинисати и као комплексна мрежа [85] – велика група међусобно повезаних чворова. Јако је битно напоменути да су комплексне мреже генерално изразито велике, те предикција њиховог понашања у целини анализом понашања свих појединачних чворова практично није могућа. Комплексне мреже су свуда око нас, а као њихов главни пример се могу посматрати комуникационе мреже. Развој комуникационих мрежа је почео још од прадавних покушаја преноса информација голубовима писмоношама, преко телеграфске комуникације, до фиксне и

мобилне телефоније и интернета. Поред комуникационих мрежа, постоје комплексне мреже које се формирају на основу релација између људи – друштвене мреже. Иако термин друштвена мрежа последњих година може доминантно асоцирати на интернет заједнице попут Фејсбука или Инстаграма, друштвене мреже су постојале и давно пре него што је дошло до развоја интернета. Те традиционалне друштвене мреже су у литературу уведене још тридесетих година прошлог века у раду Јакоба Морена који је увео термин социограма [85]. Социограм је подразумевао графички приказ мреже сличан графу, људи би били представљени тачкама (односно чворовима), а релације између људи би биле приказане линијом (односно граном) која би повезивала две тачке. Веза између људи би могла бити позитивна или негативна (у зависности од односа између две индивидуе), док би одсуство везе представљало неутралност, односно одсуство било каквог односа. Ипак, за разлику од социограма, графови представљају математичке објекте који омогућавају анализу структуре мреже на основу теоретског математичког алата. Ипак, друштвене мреже су подстакле неке нове правце развоја теорије графова, као на пример увођењем нових метрика које могу описивати значај особе (или групе особа) у друштвеној мрежи. Са друге стране, теорија графова пружа алат за формално описивање значаја или утицаја чворова у графу.

Иако су комуникационе и друштвене мреже најзаступљеније и најпознатије комплексније мреже са којима се људи сусрећу, постоји и много других примера комплексних мрежа свуда око нас. Неки од примера комплексних мрежа су и:

1. Мреже путева, где би раскрснице представљале чворове, а делови пута који спајају раскрснице представљали гране (где би се на посебан начин могле анализирати једносмерне и двосмерне улице),

2. Авио транспорт, где би аеродроми представљали чворове, а летови између аеродрома представљали гране (сваки засебни авио-превозник би могао формирати засебну мрежу летова, која би се могла интерконектовати са мрежама других авио-превозника),

3. Мреже цитата, где би сваки рад представљао чвор, а сваки цитат ка представљао грану (анализом оваквих мрежа би се могли лоцирати значајни радови у некој области),

4. Мреже репутације на интернет платформама за продају, где би сви чланови представљали чворове (и продавци и купци), а оцене производа би представљала гране,

5. Трансакције у банкарству, где би бројеви рачуна представљали чворове, а трансакције и пренос новца представљали гране (анализом оваквих мрежа би се лоцирале финансијске малверзације као и прање новца),

6. Телефонски позиви, где би сваки број у мрежи представљао чвор, а позиви гране (овакав тип комплексних мрежа биће предмет истраживања које ће бити представљено у следећим поглављима)...

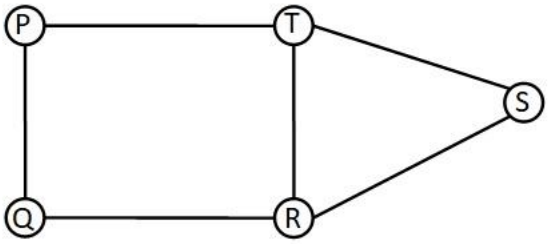
Очигледно је да се комплексне мреже могу јавити у различитим научним дисциплинама, попут економије, социологије, логистике, биологије, и слично. Ипак, за разумевање већине комплексних мрежа довољан је један алат – теорија графова.

5.3. Основни појмови и метрике теорије графова

Описно говорећи, графови су фигуре састављене од тачака и линија, при чему свака линија повезује по две (не обавезно различите) тачке. Граф се дефинише као апстрактни математички објекат, а фигура састављена од тачака и линија је заправо геометријска представа или цртеж графа. Формална дефиниција графа би се могла поставити на следећи начин [87]:

Дефиниција 5.1: Нека је *X* непразан скуп и *ζ* бинарна релација у *X*. Уређени пар *G=(X, ζ)* се назива граф. Елементи скупа *X* су чворови графа, а елементи скупа *ζ* гране графа.

Граф се може представити цртежом на следећи начин. Чворови графа $x_1, \dots, x_n$ ($\in X$) се представљају произвољним међусобно различитим тачкама у равни (или простору). Ако је $(x_i, x_j) \in \zeta$, тачка која представља чвор x_i се спаја непрекидном глатком линијом са тачком која представља x_j. Тада се може рећи да су чворови x_i и x_j суседни. Ако $(x_i, x_j) \notin \zeta$ чворови x_i и x_j нису директно повезани на цртежу. На слици 5.3. је приказан граф са чворовима P, Q, R, S и T. Дужи које спајају чворове се називају гране. Ако грана g садржи чворове x_i и x_j, може се рећи да је грана g инцидентна са чворовима x_i и x_j. Грана која спаја чвор са самим собом назива се петља. Пут представља низ грана које следе једна за другом, на пример на примеру са слике 5.3. пут *P-Q-R-S-T* је дужине 4, а пут *P-Q-R-T-P* је такође дужине 4. Пут *T-R-S-T* се из очигледног разлога назива контура. Прецизнија дефиниција појма пута биће дата у следећим параграфима. Дакле, граф представља скуп тачака и дужи које их повезују док су све његове метричке особине неважне.



Слика 5.3: Пример једноставног неусмереног графа

Са друге стране, граф се може представити и квадратном матрицом чији је ред једнак броју чворова графа. Тада би елемент матрице на пресеку *i*-те врсте и *j*-те колоне (a_{ij}) био једнак броју грана које полазе из чвора x_i а завршавају се у чвору x_j. Таква матрица се зове матрица суседства [87] и обележавала би се са *A*. Матрица суседства која би одговарала графу са слике 5.3. је:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Појмови простог графа и подграфа ће бити битни при дефинисању осталих елемената теорије графова који се користе у анализи друштвених мрежа [87]:

Дефиниција 5.2: Граф је прост ако не постоје два чвора која повезују више од једне гране и који не садржи петље.

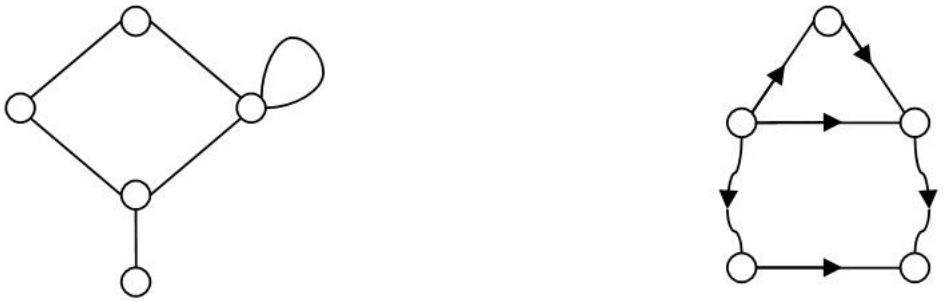
Дефиниција 5.3: Нека је дат граф *G=(X,U)*. Граф облика *H=(Y,T)*, при чему је $Y \subset X$ и $T = U \cap Y \times Y$ (T је подскуп скупа U који садржи све оне парове из U који су образовани само од елемената скупа Y) назива се подграф графа *G*, образован скупом чворова *Y*.

Дакле, подграф из датог графа добија се на тај начин што се уочи неки подскуп скупа чворова (*Y*) и удаље из почетног графа сви остали чворови заједно са гранама које су суседне удаљеним чворовима. У подграфу остају само гране које повезују чворове из редукованог скупа *Y*. Ако је $X \neq Y$, граф *H* из дефиниције 5.3. се назива прави подграф.

Један од битнијих елемената при анализи друштвених мрежа је и смер релације. Појам смера релације између неких ентитета се најлакше може објаснити примером. Ако се посматра скуп трансакција у банкарству, очигледно је да постоји велика разлика уколико је са рачуна *A* пребачен неки износ новца на рачун B, у поређењу са случајем када је исти износ пребачен са рачуна *B* на рачун *A*. Слично, ако се посматрају мреже позива неког мобилног телекомуникационог оператора, очигледно је да није исто уколико је број *X* звао број *Y*, или уколико је број *Y* иницирао позив према броју *X*. Из тог разлога је појам оријентисаног и неоријентисаног графа из теорије графова изразито битан [87]:

Дефиниција 5.4: Граф је оријентисан (антисиметричан) ако и само ако је ζ антисиметрична релација (односно ако важи да $(\forall a, b \in X) a \ \zeta \ b \wedge b \ \zeta \ a \Longrightarrow a = b$). Слично, граф је неоријентисан (симетричан) ако и само ако је ζ симетрична релација (односно ако важи да је $((\forall a, b \in X) a \ \zeta \ b \Longrightarrow b \ \zeta \ a$).

Пример неоријентисаног графа са петљом је дат на слици 5.4. лево, док је пример оријентисаног графа дат на слици 5.4. десно. Специфичност оријентисаног графа је у томе да се његове гране могу дефинисати као уређени парови његових темена.



Слика 5.4: Лево – пример неоријентисаног графа са петљом; десно – пример оријентисаног графа

Битно је приметити да је матрица суседства неоријентисаног графа симетрична матрица за коју важи $A = A^T$. Са друге стране, за оријентисане графове матрица суседства не мора бити симетрична. Додатно, уколико су елементи на главној дијагонали матрице суседства једнаки нули, тада граф неће имати петљи.

Поред основних појмова теорије графова, потребно је дефинисати и основне операције са графовима [88]:

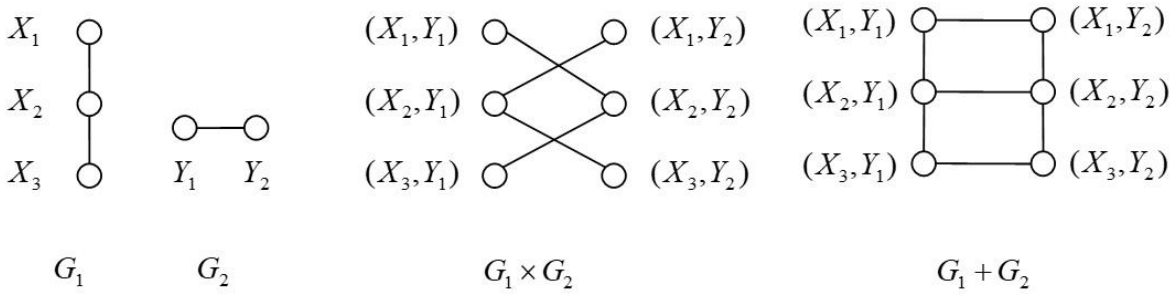
Дефиниција 5.5: Нека је G неоријентисан граф без петљи. Комплемент $\bar{G}$ графа G је неоријентисан граф без петљи који има исте чворове као G, при чему су два (међусобно различита) чвора суседна у ако и само ако ти чворови нису суседни у G, то јест ако је $G=(X,U)$, тада је $\bar{G} = (X, [X]^2 - U)$.

Дефиниција 5.6: Унија графова $G_1 = (X_1, U_1)$ и $G_2 = (X_2, U_2)$ је граф $G=(X,U)$, где је $X = X_1 \cup X_2$ и $U = U_1 \cup U_2$ и означава се као $G = G_1 \cup G_2$.

Дефиниција 5.7: Нека су G_1 и G_2 графови, а $X = \{x_1, \dots, x_n\}$ и $Y = \{y_1, \dots, y_m\}$, редом њихови скупови чворова. Производ $G_1 \times G_2$ ових графова је граф чији је скуп чворова једнак $X \times Y$, тј. чворови графа $G_1 \times G_2$ су сви уређени парови облика $(x_i \times y_j)$. Два чвора $(x_{i1} \times y_{j1})$ и $(x_{i2} \times y_{j2})$ су суседни у $G_1 \times G_2$ ако и само ако су x_{i1} и x_{i2} суседни чворови у G_1 и ако су y_{j1} и y_{j2} суседни у G_2.

Дефиниција 5.8: Сума $G_1 + G_2$ графова G_1 и G_2 је граф са истим скупом чворова као и $G_1 \times G_2$. Чворови $(x_{i1} \times y_{j1})$ и $(x_{i2} \times y_{j2})$ су суседни у $G_1 + G_2$ ако и само ако су $x_{i1} = x_{i2}$ и ако су y_{j1} и y_{j2} суседни у G_2 или су x_{i1} и x_{i2} суседни чворови у G_1 и $y_{j1} = y_{j2}$.

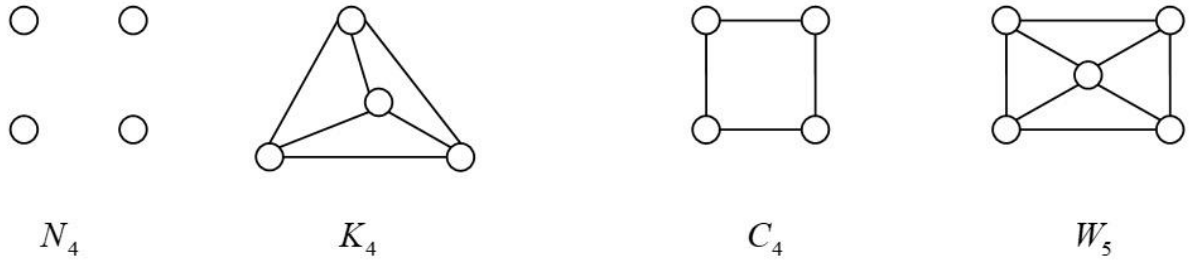
На слици 1.5 је приказан производ односно збир графова G_1 и G_2.



Слика 5.5: Лево – пример изгледа графова G_1 и G_2; средина – производ графова G_1 и G_2; десно – збир графова G_1 и G_2

Неке од најбитнијих специјалних класа графова су приказане на слици 5.6. а дефинисани су као [86]:

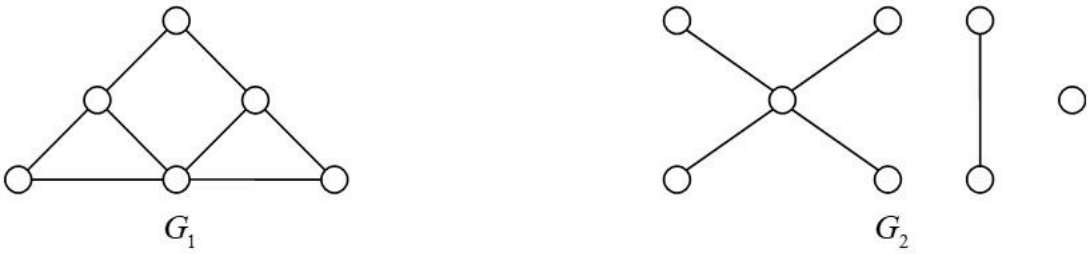
1. Нула граф (N_n): граф који садржи n чворова и не садржи ни једну грану,
2. Комплетан граф (K_n): прост граф чија су свака два темена спојена граном; Граф K_n има тачно $n(n-1)/2$ грана,
3. Циклични граф (C_n): повезан граф који садржи n чворова и чији је сваки чвор степена два,
4. Точак (W_n): граф са n чворова који је збир графова C_{n-1} и N_1.



Слика 5.6: Са лева на десно редом: нула граф, комплетни граф и циклични граф са по 4 чвора, и точак са 5 чворова

Путања, односно пут у графу је дефинисан у поглављу 5.1. Битно је запазити да произвољна путања унутар неког графа може више пута да пролази истом граном или кроз исти чвор. Елементарна путања би представљала путању која кроз сваки чвор графа пролази највише једанпут. Путања која се завршава у истом чвору у којем почиње, назива се кружни (затворени) пут или контура. Као грана у путу се може појавити и нека петља.

Неоријентисани граф је повезан уколико се његова два произвољна чвора могу повезати путањом [87]. Ако постоје чворови који се не могу повезати путањом, граф је неповезан. Неповезан граф се састоји од два или више одвојених делова. Ови одвојени делови графа се називају компоненте повезаности графа. Тачније, компонента повезаности графа којој припада неки чвор x_i је подграф образован скупом свих оних чворова који се могу спојити путањом са чвором x_i, укључујући ту и чвор x_i. На слици 5.7 је приказан граф G_1 који је повезан и граф G_2 који је неповезан и има три компоненте повезаности.



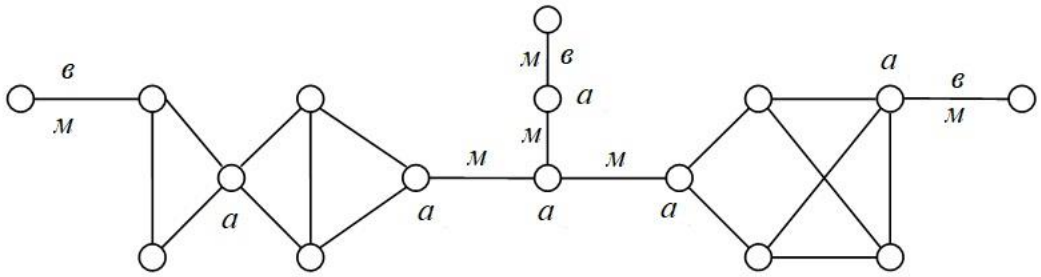
Слика 5.7: Лево – повезани граф G_1; десно – неповезани граф G_2 са три компоненте повезаности

Уско повезани са термином компоненти повезаности графа су концепти артикулационог чвора и моста [89]:

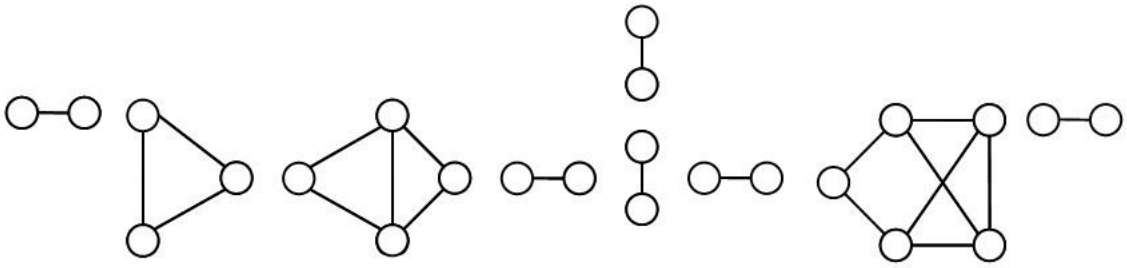
Дефиниција 5.9: Чвор *v* представља артикулациони чвор графа *G=(X,U)* уколико се скуп грана графа *U* може поделити на два подскупа U_1 и U_2 тако да графови G_1 и G_2 које формирају скупови грана U_1 и U_2 и чворови који им припадају имају само чвор *v* као заједнички.

Дакле, артикулациони чвор графа је чвор чијим се удаљавањем повећава број компонената повезаности графа. Мост графа је грана којом се постиже исти ефекат. Грана која је инцидентна са чвором степена један назива се висећа грана. Може се приметити да је свака висећа грана је уједно и мост графа. Са друге стране, блок графа је сваки његов максимални повезани подграф без артикулационих чворова [87]. На

слици 5.8 је приказан граф на коме су словима *а*, *м* и *в* означени редом артикулациони чворови, мостови и висеће гране. С друге стране, на слици 5.9 су приказани сви блокови графа приказаног на слици 5.8.

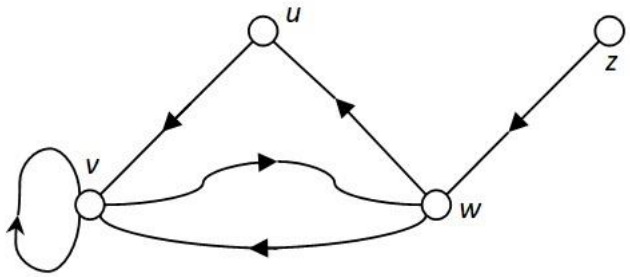


Слика 5.8: Пример артикулационих чворова (а), мостова (м) и висећих грана (в) у графу



Слика 5.9: Пример блокова у графу са слике 5.8.

Битно је нагласити да многе дефиниције које важе за неоријентисане важе и за оријентисане графове. Тако, на пример, за два чвора *u* и *v* се каже да су повезана ако у скупу грана оријентисаног графа постоји грана облика (*u*,*v*) или (*v*,*u*). Тада се за чворове *u* и *v* каже да припадају овим гранама. Аналогно могу се дефинисати и оријентисани путеви и оријентисани контуре. Може се приметити да оријентисана путања не може садржати грану (*u*,*v*) више од једанпут, али може да садржи гране (*u*,*v*) и (*v*,*u*). На пример, на слици 5.10. може се запазити оријентисана путања *z*-*w*-*v*-*w*-*u*. За оријентисан граф *D* каже се да је повезан ако за свака два чвора *u* и *v* које припадају оријентисаном графу *D* постоји оријентисана путања која води из *u* у *v*.



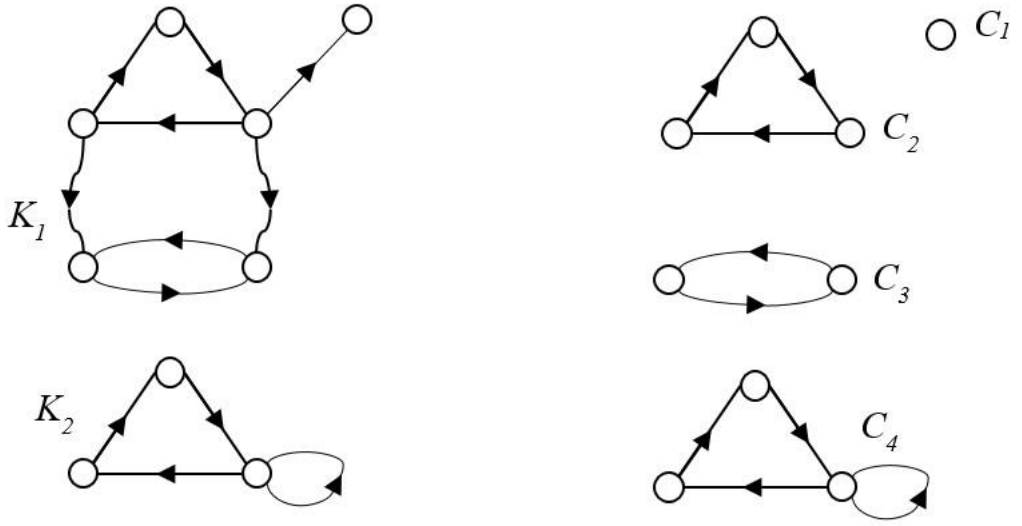
Слика 5.10: Пример оријентисаног графа са петљом

Треба напоменути да оријентација петље нема неког посебног значаја јер, ма како да се стрелица постави, петља води из, на пример, чвора *u* у исти чвор *u*. Стога се код петље стрелица на цртежу може изоставити.

Треба напоменути да концепт повезаности и компоненти повезаности графа код оријентисаних графова треба проширити дефиницијом компоненти јаке повезаности [87]:

Дефиниција 5.10: Нека је у скуп *X* чворова оријентисаног графа *G* уведена бинарна релација δ помоћу дефиниције: чворови *x* и *y* су у релацији δ ако и само ако је *x=y* или се *x* и *y* налазе на неком затвореном оријентисаном путу графа *G*. Подграфови графа *G* индуковани класама еквиваленције релације δ, представљају компоненте јаке повезаности оријентисаног графа *G*.

Пример компоненти јаке повезаности графа и поређења са компонентама повезаности је дат на слици 5.11. На слици 5.11. лево је приказан комплетан оријентисани граф са две компоненте повезаности K_1 и K_2. На слици 5.12. десно су са C_1, C_2, C_3 и C_4 дате компоненте јаке повезаности графа са слике лево. Може се запазити да компоненти јаке повезаности има више од компоненти повезаности.



Слика 5.11: Лево – пример оријентисаног графа са две компоненте повезаности; десно – компоненте јаке повезаности графа са слике лево

Различити проблеми анализе друштвених мрежа своде се на посматрање графова код којих је свакој грани придружен неки број. При овоме, је могуће да графови буду оријентисани или неоријентисани и са или без петљи у зависности од конкретног проблема. Такође, бројеви који се придружују гранама могу припадати различитим скуповима: скупу природних, реалних, ненегативних бројева, и слично. У свим оваквим случајевима се каже да је на скупу грана (укључујући и петље) дефинисана једна функција [87]. Ако се посматрају два чвора, *u* и *v*, и ако w_{uv} означава јачину везе гране која спаја чворове *u* и *v*, тада се матрица јачине веза може дефинисати као:

$$W = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{N1} \\ w_{12} & w_{22} & \cdots & w_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1N} & w_{2N} & \cdots & w_{NN} \end{bmatrix} \tag{5.1}$$

где N означава укупан број чворова у графу. Ако чворови u и v нису повезани, тада је тежина везе w_{uv} једнака 0. Важно је нагласити да је за неоријентисане графове $w_{uv} = w_{vu}$, док за оријентисане графове то не мора бити случај.

5.4. Метрике чворова у графовима

За потребе анализе чворова у графовима уведене су бројне метрике ради нумеричке потврде значаја чворова за граф. Неке од метрика су [85]:

1. Степен чвора (*node degree*),
2. Улазни степен чвора (*node in-degree*),
3. Излазни степен чвора (*node out-degree*),
4. Значајност чвора првог реда (*node first-order influence*),
5. Значајност чвора другог реда (*node second-order influence*),
6. Сопствени вектор чвора (*node eigenvector value*),
7. Вредност ауторитета чвора (*node authority value*),
8. *Hub* вредност чвора (*node hub value*).

Степен посматраног чвора представља укупан број чворова који су инцидентни са посматраним чвором. Ова мера се може рачунати и за неоријентисане и оријентисане графове. Код неоријентисаних графова степен чвора би био једнак броју чворова који су суседни са посматраним чвором. Са друге стране, код оријентисаних графова степен чвора би био једнак суми улазног и излазног степена посматраног чвора.

Улазни степен посматраног чвора код оријентисаних графова представља укупан број улазних грана инцидентних са посматраним чвором. Насупрот њему, излазни степен посматраног чвора код оријентисаног графа представља укупан број излазних грана које су инцидентне са посматраним чвором.

Значајност чвора првог реда представља генерализацију метрике степена чвора која узима у обзир и јачине грана које су инцидентне са посматраним чвором (везе ка суседима посматраног чвора). Општа формула за значајност чвора првог реда је:

$$I_1(u) = \frac{\sum_{v \epsilon N_u} w_{uv}}{N} \tag{5.2}$$

где је u посматрани чвор, N укупан број чворова у графу, N_u листа чворова који су суседни посматраном чвору, а w_{uv} представља јачину везе гране која спаја чворове u и v. Значајност чвора првог реда се може прорачунати и за неоријентисане и оријентисане графове. Главна разлика у прорачуну би представљала дефиниција суседног чвора, код

оријентисаних графова би се подразумевали искључиво суседи који су инцидентни са излазним гранама посматраног чвора.

Значајност чвора другог реда представља генерализацију метрике степена чвора која узима у обзир јачине грана које су инцидентне чворовима који су инцидентни са посматраним чвором (везе ка суседима суседа посматраног чвора). Општа формула за значајност чвора првог реда је:

$$I_2(u) = \sum_{v \epsilon N_u} I_1(v) \tag{5.3}$$

где је u посматрани чвор, N_u листа чворова који су суседни посматраном чвору, а $I_1(v)$ представља значајност првог реда чвора v. Значајност чвора другог реда се може прорачунати и за неоријентисане и оријентисане графове, са разликама које су идентичне као у случају рачунања значајност првог реда чвора.

Сопствени вектор чвора графа је метрика која је представља екстензију степена чвора, која подразумева додељивање вредности спрам мере „централности“ чвора у графу. Наиме, узимајући у обзир да нису сви чворови подједнако битни за структуру графа, грана која иде ка битнијем чвору би требала да значајније утиче на вредност „централности“ чвора графа спрам гране која иде ка мање битним чворовима. Општа формула за вредност сопственог вектора чвора графа је [86]:

$$E(u) = \frac{1}{\lambda} \sum_{v \epsilon N} w_{uv} E(v) \tag{5.4}$$

где је λ константа, N укупан број чворова у графу, w_{uv} представља јачину везе гране која спаја чворове u и v, а $E(v)$ је вредност сопственог вектора чвора v. У матричном облику се претходна формула може написати као:

$$WE = \lambda E \tag{5.5}$$

где је W матрица јачине веза. Сопствени вектор чвора графа се може рачунати само за неоријентисане графове.

Вредност ауторитета чвора и *hub* вредност чвора су мере значајности чворова. Ове метрике је увео Џон Клајнберг у својој анализи значаја и рангирања интернет страница [90]. Наиме, одређене странице се могу посматрати као „каталози“ страница и усмеравају кориснике ка значајним страницама на интернету. На овај начин су дефинисане две оцене за интернет странице:

1. Оцена ауторитета која оцењује квалитет саме странице

2. *Hub* оцена која оцењује вредност линкова као другим страницама које нека страница садржи

Дакле, „добра“ *hub* страница би подразумевала страницу која показује на велики број „добрих“ страница које поседују високи ауторитет. Такође, „добра“ ауторитативна страница се рекурзивно може дефинисати као страница на коју показује велики број „добрих“ *hub* страница. Опште формуле за вредност ауторитета чвора и *hub* вредност чвора у графу су [90]:

$$A(u) = \alpha \sum_{v \epsilon N} w_{uv} H(v) \tag{5.6}$$

$$H(u) = \beta \sum_{v \epsilon N} w_{vu} A(v) \tag{5.7}$$

где су α и β константе и $H(v)$ и $A(v)$ представљају, респективно, *hub* вредност чвора и вредност ауторитета чвора v. У матричном облику се претходна формуле могу написати као:

$$WW^T A = \lambda A \tag{5.8}$$

$$WW^T H = \lambda H \tag{5.9}$$

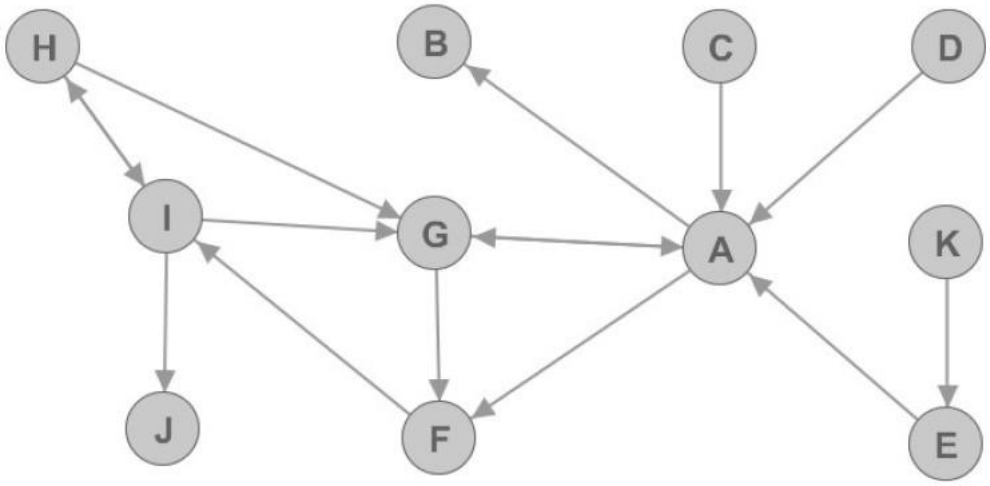
где је $\lambda = \alpha\beta$ константа, а W матрица јачине веза. Вредност ауторитета чвора и *hub* вредност чвора у графу се могу рачунати и за неоријентисане и за оријентисане графове.

Битно је напоменути да се као метрика чворова у графу може посматрати и артикулациони чвор. Наиме, узимајући у обзир да елиминисање артикулационих чворова из графа повећава број компоненти повезаности у графу, очигледно је да артикулациони чворови имају велику важност за структуру и повезаност графа. Може се усвојити следеће правило за одређивање вредности метрике артикулације неког чвора:

$$articulation(u) = \begin{cases} 0, & \text{чвор } u \text{ није артикулациони чвор} \\ 1, & \text{чвор } u \text{ је артикулациони чвор} \end{cases} \tag{5.10}$$

Приказане метрике чворова у графовима могу имати велику примену у анализи комплексних мрежа. Наиме, узимајући у обзир да комплексне мреже могу бити представљене графом на начин који је описан у поглављу 5.2, проналажење битних чворова по одређеним метрикама се може директно пресликати на претрагу битних чланова неке комплексне мреже. На пример, уколико би се анализирала мрежа путева (где би раскрснице представљале чворове, делови пута који спајају раскрснице представљали гране, а просечан број возила у делу пута представљао јачину везе гране) анализом чворова (раскрсница) би се могле лоцирати саобраћајно преоптерећени делови града. На тај начин би се могао оптимизовати транспорт, кориговати распоред семафора и времена трајања пропуштања саобраћаја у одређеним правцима, као и оптимизовати градски саобраћај (пребацивањем на алтернативне путање мањег саобраћајног оптерећења а веће пропусне моћи). Рад [91] даје један практичан пример примене метрика чворова графова у анализи друштвених мрежа које настају позивима у мрежи мобилне телефоније ради предвиђања вероватноће да ће клијент напустити мобилног телекомуникационог оператора.

Прорачун вредности метрика које су наведене у овом поглављу биће представљен на примеру графа који је приказан на слици 5.12.



Слика 5.12: Пример једноставног усмереног графа

На једноставном усмереном графу са слике 5.12. биће приказан поступак прорачуна метрика чворова у графовима. Додатно, интуитивно ће се показати да ће чворови са већим вредностима метрика бити значајнији за структуру графа у поређењу са чворовима који имају ниже вредности метрика. Са слике 5.12. се може видети да су неки чворови боље повезани са остатком графа, па би самим тим требали да буду битнији за саму структуру графа од других чворова. На пример, чворови *A* и *G* су значајнији од осталих чворова у графу управо зато што би њихово изостављање из графа довело до драстичног смањења укупног броја грана у графу. Штавише, и број компоненти повезаности графа са слике 5.12. би био повећан изостављањем тих тачака. Резултати прорачуна метрика чворова за граф са слике 5.12. су приказани у табели 5.1. Вредности из табеле 5.1. могу бити коришћене за процену важности чворова у структури графа. Битно је напоменути да су све константе које су неопходне за процену појединих метрика изабране тако да резултујуће метрике буду нормализоване у опсегу између нуле и јединице. Такође, зарад једноставности, вредности јачине веза свих грана у графу су постављене на један.

Табела 5.1: Резултујуће вредности метрика чворова једноставног усмереног графа

Чвор	Степен чвора	Улазни степен чвора	Излазни степен чвора
A	6	4	3
B	1	1	0
C	1	0	1
D	1	0	1
E	2	1	1
F	3	2	1
G	4	3	2
H	2	1	2
I	4	2	3

J	1	1	0
K	1	0	1

Чвор	Значајност чвора првог реда	Значајност чвора другог реда	Артикулација чвора
A	0.27	0.27	1
B	0.00	0.00	0
C	0.09	0.27	0
D	0.09	0.27	0
E	0.09	0.27	1
F	0.09	0.27	0
G	0.18	0.36	0
H	0.18	0.45	0
I	0.27	0.36	1
J	0.00	0.00	0
K	0.09	0.09	0

Чвор	Сопствени вектор чвора	*Hub* вредност чвора	Вредност ауторитета чвора
A	0.96	1.00	0.63
B	0.30	0.00	0.41
C	0.30	0.30	0.00
D	0.30	0.30	0.00
E	0.33	0.30	0.00
F	0.87	0.16	0.67
G	1.00	0.62	1.00
H	0.57	0.64	0.33
I	0.84	0.80	0.33
J	0.26	0.00	0.33
K	0.10	0.00	0.00

Ако се посматрају резултујуће вредности метрика чворова, може се приметити да је чвор *E* означен као артикулациони чвор. Ипак, са слике 5.12. се може приметити да чвор *E* није претерано битан или виталан за структуру комплетног графа. Такође, може се приметити да је чвор *G* веома добро повезан са остатком графа и очигледно јако битан за структуру графа, али истовремено није означен као артикулациони чвор. Дакле, артикулација чвора се може користити при анализи битности чворова у графу или

друштвеној мрежи која је апроксимирана графом, али се истовремено мора бити врло пажљив при њеној интерпретацији.

Резултујуће вредности метрика представљене у табели 5.1. показују да су чворови *A* и *G* врло значајни и то по свим прорачунатим метрикама. Међутим ако се посматра чвор *B* може се видети да он има релативно високу вредност ауторитета чвора, иако се визуелним прегледом конфигурације мреже такав закључак не може донети. Овај феномен се објашњава преносом ауторитета. Наиме, уколико чвор који није претерано битан из угла анализе комплексних мрежа (у овом случају чвор *B*) буде повезан са веома битним чвором у смислу *hub* оцене (у овом примеру чвор *A*), тада ће мање битан чвор добити веома високу оцену ауторитета. Ипак, насупрот овој оцени, на основу свих других оцена, очигледно је да чвор *B* није виталан за структуру и повезаност комплексне мреже приказане графом са слике 5.12. На сличан начин се могу приметити и феномени преноса *hub* оцене, као и пренос метрике сопственог вектора чвора. Дакле, при анализи чворова графа комплексне мреже треба узимати у обзир и потенцијални утицај горeописаних ефеката.

5.5. Друштвене мреже у мобилним телекомуникацијама

Велики телекомуникациони оператора губе кориснике сваког месеца и улажу велике напоре да тај број буде што мањи. У ранијем периоду, док није дошло до засићења тржишта мобилних телекомуникационих услуга, раст на тржишту телекомуникација је био експоненцијалан и губитак корисника није био посматран као битан изазов. Ипак у данашње доба, управо услед засићења и сазревања тржишта, растућа конкуренција је довела до тога да мобилни телекомуникациони оператори промене свој главни фокус са прибављања нових корисника, на очување постојеће корисничке базе. Посебан фактор који је утицао на ову промену у филозофији пословања мобилних телекомуникационих оператора представља могућност преноса бројева између оператора. На тај начин крајњи корисник може без икаквих непријатности прећи од једног ка другом мобилном телекомуникационом оператору, истовремено чувајући свој претплатнички број. Додатно, израчунато је да је за телекомуникационог оператора осетно јефтиније да, у условима какви тренутно владају на светском тржишту телекомуникацијама, своје ресурсе усмери ка очувању корисника у поређењу са ценом привлачења екстерног корисника [92]. Дакле, главни задатак сваког савременог мобилног телекомуникационог оператора је контролисање броја корисника које губе сваког месеца.

Кључан механизам за превенцију губитка корисника представља разумевање разлога зашто корисници одлазе, као и унапређивање услуга које су доступне корисницима. Упоредо са овим процесом, конкуренција на тржишту мобилних телекомуникација се труди да својим предностима преотме део клијената, па се самим тим сви провајдери на тржишту морају континуирано питати на који начин могу побољшати свој приступ овом битном проблему. Анализа друштвених мрежа може представљати метод којим се може доћи до предности на конкуренцијом.

Анализа друштвених мрежа може подразумевати анализу друштвених веза и друштвених комуникација између индивидуа коришћењем теорије графова [93]. Друштвене везе и комуникације могу подразумевати и позиве и поруке које људи размењују путем мобилних телефона. У том случају, у питању је телекомуникациони домен анализе друштвених мрежа. Телекомуникациони графови који се анализирају у телекомуникационом домену анализе друштвених мрежа се састоје од записа о позивима (*Call Data Record* – CDR). Ипак, нису сви чланови друштвених мрежа

подједнако битни. Постоји одређени број изразито значајних клијената, то јест клијената који могу пренети поруку о својим искуствима у погледу мобилних телекомуникационих оператора својим пријатељима или познаницима са којима су у редовном друштвеном контакту.

Пример који можда најбоље илуструје значај битних чланова друштвених мрежа је представљен у [94] и потиче још од доба Америчког рата за независност. Наиме, на самом почетку рата контингент британских војника је изненада послат да похапси вође побуне. Иако нису очекивали никакав отпор услед очекиваног фактора изненађења, британске војнике дочекала је добро организована група америчких побуњеника која их је потукла у директној борби. Овај догађај је довео до распламсавања Америчког рата за независност. Поставља се питање како је побуњеничка војска добила информацију о нападу Британаца и како су стигли да се организују. Испоставило се да су два човека Пол Ривиер (*Paul Revere*) и Вилијам Давс (*William Dawes*) упоредо са Британцима похитали да упозоре побуњенике. Пол Ривиер је ишао једним правцем којим се није очекивао велики одзив побуњеника, али је успео да прикупи велики број људи који су успели да одбију напад. Са друге стране, Вилијам Давс је прикупио знатно мање побуњеника који су дошли у битку са малим закашњењем. На овај начин је Пол Ривиер остао запамћен у америчкој историји. Испоставило се да је приступ прикупљању људи ова два побуњеника био драстично различит. Вилијам Давс је био млади обућар који је кроз места која је обилазио ужурбано пројахивао и лупао на врата насумично изабраних кућа вичући да Британци долазе. Људи у местима које је обилазио му нису веровали зато што га нису познавали, нису веровали да је опасност од Британаца непосредна. Насупрот њему, Пол Ривиер је био искусни трговац који је познавао доста људи. Његов приступ је био другачији. У местима које је обилазио Пол Ривиер је директно одлазио код локалних вођа побуњеника који су имали мрежу својих сарадника те су могли брзо да пренесу кредибилну поруку на великом растојању. На овај начин се може видети колико је био битан кредибилан извор информација са одговарајућим конекцијама још пре више од 200 година. У данашњем свету у коме се информације и дезинформације могу јако брзо пренети, лоцирање таквих значајних чланова друштвених мрежа може бити кључно у различитим сферама, од пословања до науке.

Као што је описано у претходним параграфима, индустрија мобилних телекомуникација је веома компетитивна те је зарад борбе за очување корисника битно користити утицај који одређени корисници имају на друге кориснике. Анализа друштвених мрежа се користи као алат којим се може продубити знање о интерним друштвеним групама корисника [95]. Групе корисника се природно могу приметити анализом образаца коришћења телекомуникационих производа и сервиса као што су примљени и упућени позиви, *SMS* поруке, мултимедијалне поруке, *e-mail* поруке и слично. Ове, некад мале некад велике заједнице, са формирају конекцијама између клијената независно од конкретног типа конекције. Ове заједнице – друштвене мреже – подразумевају више индивидуа које могу бити у некој релацији, било јакој или слабој, фреквентној или реткој, са осталим индивидуама у заједници. Као и у сваком типу заједнице, неки чланови могу имати већи или мањи утицај на своју околину. Ипак, не мора сваки члан заједнице имати подједнаки ефекат по сваком питању. На пример, одређена група корисника може имати утицаја на одлуку других корисника о напуштању мреже, док друга група може имати утицај на повећање продаје неког конкретног новог производа. Наравно, што је већи утицај који индивидуа има на своје окружење, то је тај корисник битнији за мобилног телекомуникационог оператора.

У веома компетитивним индустријама, попут индустрије мобилних телекомуникација, веома је важно раздвојити изразито вредне кориснике од оних који то нису. Постоји велики број различитих приступа за откривање вредних корисника попут оних базираних на коришћењу садржаја, профита који компанија остварује на њихов рачун, демографије, ризика, и слично. Анализа друштвених мрежа може бити још један од приступа за процену вредности клијената на основу разгранатости њихових конекција. Вредност клијента у овом контексту зависи и од количине, учесталости, дужине и снаге утицаја на друге клијенте.

Поред нивоа утицаја, битно је и дефинисати колико се брзо утицај преноси са лидера неке друштвене групе ка осталим члановима те групе. Да би се добили корисни резултати при анализи друштвене мреже мобилног телекомуникационог оператора, потребно је извршити дубљу анализу у дужем временском периоду. Потребно је узети у обзир околину утицајног корисника у дужем временском периоду и посматрати да ли ће се понашање лидера „пренети“ и на остале кориснике исте друштвене подгрупе. На пример, уколико лидер купи одређени нови уређај, да ли клијенти из његове групе такође купују исти уређај након неког времена? Овај тип анализе може помоћи при објашњавању еволуције друштвених група унутар телекомуникационе мреже, као и при планирању маркетиншких кампања. Задржавањем најбољих и најпрофитабилнијих клијената у својој бази корисника, телекомуникационе компаније могу унапредити продају нових производа, сервиса и промоција. Очигледно, једноставније је продати нови производ добром кориснику (задовољном и / или платежно способном) у поређењу са неким лошијим корисником.

Узимајући у обзир да телекомуникационе компаније могу чувати и обрађивати велику количину података, попут података о корисницима, позивима, порукама и слично, потребно је описати процес формирања друштвене мреже телекомуникационог оператора. Запис о позиву – CDR – чува информације о позиву који је упућен преко мреже неког мобилног телекомуникационог оператора. Узимајући у обзир огромну количину CDR записа, може се посматрати само део CDR записа унутар неког временског периода. Сваки пут када се изврши неки позив на мрежи мобилног телекомуникационог оператора генерише се запис CDR који у себи садржи неке дескриптивне информације о конкретном позиву. То би биле анонимизиране вредности MSISDN (*Mobile Station International Subscriber Directory Number* – број који једнозначно одређује претплатника односно корисника мобилног телекомуникационог оператора) позиваоца и позиваног корисника, време почетка позива, трајање позива, приход који је остварен за мобилног телекомуникационог оператора, квалитет позива, и слично.

При формирању друштвене мреже позива у мобилном телекомуникационом оператору, као чворови се могу посматрати сви MSISDN бројеви који су примећени, било као иницијатори, било као примаоци позива, унутар свих CDR-ова анализираних у току неког временског периода. Надаље, гране графа који би представљао друштвену мрежу мобилног телекомуникационог оператора би биле дефинисане на следећи начин [91]:

1. Грана $u \rightarrow v$ између два чвора u и v ће постојати уколико у скупу свих посматраних CDR-ова постоји барем један који има MSISDN број u као позивајући, а MSISDN број v као позивани број.

2. Уколико постоји више CDR-ова који имају MSISDN број *u* као позивајући, а MSISDN број *v* као позивани број, неће доћи до генерисања више грана између чворова *u* и *v*; сви ти CDR-ови ће бити посматрани као искључиво једна грана.

3. Уколико постоје два CDR-а таква да један има MSISDN број *u* као позивајући, а MSISDN број *v* као позивани број, а други MSISDN број *v* као позивајући, а MSISDN број *u* као позивани број, биће формиране две гране графа. Једна грана биће означена као $u \rightarrow v$, а друга као $v \rightarrow u$.

4. Први тежински коефицијент гране $u \rightarrow v$ између два чвора *u* и *v* графа друштвене мреже мобилног телекомуникационог оператора представља број позива (CDR-ова) у којима је MSISDN број *u* позивајући, а MSISDN број *v* позивани број.

5. Други тежински коефицијент гране $u \rightarrow v$ између два чвора *u* и *v* графа друштвене мреже мобилног телекомуникационог оператора представља укупно трајање позива (CDR-ова) у којима је MSISDN број *u* позивајући, а MSISDN број *v* позивани број.

На овај начин се на основу CDR записа може формирати граф. За чворове таквог графа се могу прорачунати метрике које су описане у поглављу 5.4. Уколико би се кластер анализа извршила на тим чворовима и њиховим атрибутима може се издвојити група значајних и утицајних корисника мреже мобилног телекомуникационог оператора.

6. ПРИМЕНА МЕТОДА КЛАСТЕРИЗАЦИЈЕ У МОБИЛНИМ ТЕЛЕКОМУНИКАЦИЈАМА РАДИ ПРЕДВИЂАЊА ГУБИТКА КЛИЈЕНТА

6.1. Методологија

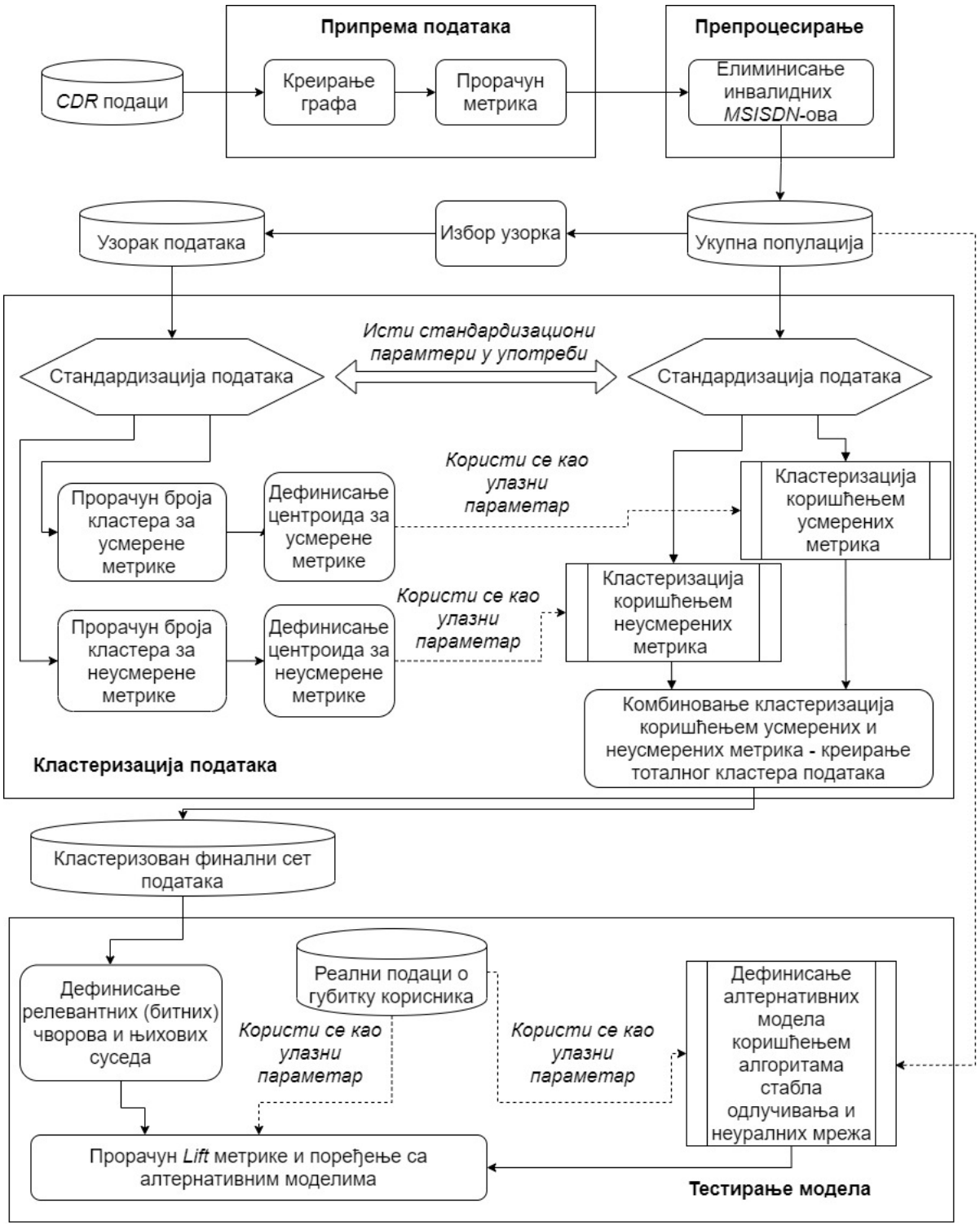
У овом поглављу биће представљен детаљан опис предложеног модела за предикцију губитка корисника у мрежи мобилних телекомуникационих оператора на основу анализе друштвених мрежа применом теорије графова [91].

Предвиђање вероватноће губитка корисника је проблем бинарне класификације у којем су сви корисници неког мобилног телекомуникационог оператора подељени у две групе. Једна група је означена као високо ризична група клијената (у смислу губитка корисника) која захтева неку реакцију (посебну понуду за наставак сарадње, попуст, бесплатне додатне услуге и слично), док друга група подразумева стандардне кориснике који нису у непосредној опасности да напусте мобилног телекомуникационог оператора. Додатно, губитак корисника може бити подељен у две основне категорије: вољни и невољни [14]. Невољни губитак корисника подразумева ситуацију у којој мобилни телекомуникациони оператор одлучује да раскине уговор са својим корисником. То се може десити из више различитих разлога, као на пример: малверзације корисника, неизмиривање обавеза корисника према мобилном телекомуникационом оператору, и слично. Са друге стране, вољни губитак корисника подразумева свесну одлуку корисника да напусти мобилног телекомуникационог оператора. Та одлука може бити проистећи из различитих узрока, при којима је за мобилног телекомуникационог оператора најпроблематичнији онај у којем се корисник одлучује за другог телекомуникационог оператора услед боље и / или приступачније понуде. Главни фокус истраживања представљеног у овој докторској дисертацији ће бити управо на оваквом вољном екстерном губитку корисника (у којем се корисник одлучује за услуге другог мобилног оператора), који су изразито битни јер могу оставити драстичне последице на главне финансијске индикаторе мобилних телекомуникационих оператора.

У даљем тексту биће представљен модел базиран на теорији графова и анализи друштвених мрежа који комбинује усмерене и неусмерене карактеристике образаца позива клијената мобилног телекомуникационог оператора. Модел анализе кластера ће дефинисати групе клијената и дефинисати значајне чланове друштвених мрежа – кориснике који могу мотивисати своје пријатеље и / или познанике да промене мобилног телекомуникационог оператора који им је пружалац услуга.

6.1.1. Опис модела

На слици 6.1. је дат шематски приказ свих корака који су неопходни за генерисање и тестирање предложеног модела.



Слика 6.1: Шематски дијаграм процеса реализације и тестирања предложеног модела за предикцију губитка корисника мобилног телекомуникационог оператора

У првом кораку се врши припрема података. Тај корак подразумева прикупљање свих CDR-ова из одређеног временског периода и генерисање графа на основу позива унутар мреже мобилног телекомуникационог оператора. Процес генерисања графа је

детаљно објашњен у поглављу 5.5. Након успешног генерисања графа, у склопу првог корака (то јест припреме података) потребно је извршити и прорачун свих метрика чворова графа (MSISDN бројева). Сваки чвор графа ће на основу дефиниција представљених у поглављу 5.4. имати укупно девет метрика. Битно је нагласити да је део метрика које ће се рачунати до сада био примењиван махом у анализи графова који настају повезивањем интернет страница. Метрике чворова ће представљати улазне променљиве за процес кластеризације о којем ће више података бити дато у следећим параграфима.

Следећи корак представља препроцесирање података. У овом кораку ће сви неважећи MSISDN бројеви бити елиминисани из мреже. Наиме, у друштвеној мрежи која је генерисана на основу свих CDR-ова постоје и неки чворови који не одговарају реалним корисницима (физичким лицима), већ специјалним службама (полиција, ватрогасци, хитна помоћ, тачно време...), позивним центрима, телемаркетинг центрима и слично. Детаљни поступак елиминисања неважећих MSISDN бројева из графа биће представљен у поглављу 6.2.1.

У трећем кораку се врши конкретна кластеризација података. Први задатак који је неопходан пре саме кластеризације је избор узорка података, зато што је комплетан скуп чворова изузетно велики и није погодан за кластеризацију. Након тога се врши стандардизација података. Битно је имати у виду да су и узорак и комплетна популација података стандардизовани истим стандардизационим методама и параметрима. На основу стандардизованог узорка података биће израчунати број кластера који се може издвојити из података и за усмерене и за неусмерене метрике (више детаља о подели метрика на групе биће дато у следећем поглављу). Вардова метода минималне варијансе (представљена у поглављу 4.3.6.) биће коришћена као изабрана метода за кластеризацију дефинисаног стандардизованог узорка. На основу кластеризованог узорка биће одређени центроиди свих добијених кластера како за усмерене, тако и за неусмерене метрике. Потом ће, на основу добијених центроида, цела популација података бити предмет анализе кластера. На крају ће се извршити комбинација добијених резултата за усмерене и неусмерене метрике и биће формиран тотални кластер података.

Четврти и последњи корак подразумева верификацију модела. На основу претходно дефинисаног тоталног кластера података биће прецизно дефинисани чворови унутар предложеног модела који су битни за предикцију губитка корисника. Ти чворови биће означени као Лидери и Битни чворови. Стога, њихови суседи биће изоловани у одвојени скуп података који ће бити означен као ризична група клијената из угла напуштања посматраног мобилног телекомуникационог оператора. Коришћењем реалних података о губитку корисника, биће показано да предложени модел пружа задовољавајуће резултате. Додатно, да би се потврдио избор алгоритма и модела који је представљен, биће извршено поређење предложеног решења са алтернативним методама моделирања као што су модели базирани на алгоритмима неуралних мрежа или стабла одлучивања. Битно је напоменути да да ће алтернативни модели бити дефинисани коришћењем истог изворног скупа података који је коришћен и за предложени модел базиран на анализи кластера. На крају, биће представљено поређење између предложеног модела базираног на анализи кластера и алтернативних метода за предикцију губитка корисника коришћењем *Lift* метрике.

6.1.2. Метрике чворова графа и њихово груписање

Као што је и описано у поглављу 5.4, девет метрика чворова графова је дефинисано. Ипак треба имати у виду да је део метрика био дефинисан само за усмерене, односно неусмерене графике. На основу горепоменутих дефиниција, биће извршена подела метрика на три групе:

1. Метрике за усмерене графове: улазни степен чвора, излазни степен чвора, значајност чвора првог реда, значајност чвора другог реда, вредност ауторитета чвора и *hub* вредност чвора. Ове метрике су најзначајније за графове мобилних телекомуникационих оператора управо из разлога што су друштвене мреже које настају позивима унутар мобилних телекомуникационих оператора по природи усмерене (постоји позивајући и позивани број; битно је да ли је неки корисник позиван или позива неког другог).

2. Метрике за неусмерене графове: степен чвора и сопствени вектор чвора. Ове мере нису подједнако битне за графове мобилног телекомуникационог оператора попут метрика за усмерене графове, али и даље могу носити неке битне и вредне информације које се односе на структуру графа.

3. Мера артикулације чвора: мера артикулације ће бити посматрана као засебна метрика зато што она представља меру утицаја чвора графа на његову структуру и повезаност, али не може директно утицати на то да је чвор (корисник у мрежи мобилног телекомуникационог оператора) битан у смислу утицаја на друге чворове (кориснике). Ово је већ илустровано у примеру са слике 5.12. из поглавља 5.4.

6.2. Процес кластеризације

6.2.1. Опис и препроцесирање података

У процесу развијања модела коришћени су реални подаци – CDR записи – који су прикупљени у периоду од једног месеца. Само CDR записи који су настали на основу стандардних гласовних телефонских позива су коришћени за генерисање мрежног графа мобилног телекомуникационог оператора. Битно је нагласити да су, због потребе за заштитом тако осетљивих података, сви подаци били детерминистички анонимизовани. То се постигло на начин да је сваки MSISDN број (било позивајући, било позиван) био кодиран засебном шифром. Дакле, за сваки појединачни MSISDN број, који би се једном или више пута појављивао у бази, постојала би шифра која ће бити једнозначно повезана на конкретан MSISDN број. Саме шифре ће се заправо користи за генерисање графа мреже телекомуникационог оператора (уместо конкретних MSISDN бројева).

Као што је описано у литератури [92]; [10], веома кратки позиви који су трајања до пет секунди и који су се појавили само једанпут унутар посматраног периода ће бити игнорисани. Наиме, овакви позиви могу означивати погрешно позване бројеве или активацију говорне поште при пропуштеним позивима, те свакако могу навести анализу на погрешне резултате. На основу анонимизованих CDR-ова формиран је граф мобилне телекомуникационе мреже позива који укупно садржи приближно 8.2 милиона чворова и 67.2 милиона грана. Просечно, сваки чвор у добијеном графу телекомуникационе мреже је инцидентан са 8.2 гране. Детаљнијом анализом структуралних карактеристика графа мобилне телекомуникационе мреже позива добијени су следећи резултати:

- Граф мобилне телекомуникационе мреже позива садржи 948 неусмерених компоненти повезаности,

- Граф мобилне телекомуникационе мреже позива садржи 2.8 милиона компоненти јаке повезаности,

- Највећи комплетан граф који се може лоцирати унутар графа мобилне телекомуникационе мреже позива је величине 22 чвора.

Узимајући у обзир да је циљ истраживања дефинисање модела за предикцију губитка корисника у мрежи мобилног телекомуникационог оператора, битно је елиминисати све неважеће MSISDN бројеве. Ти неважећи MSISDN бројеви могу генерисати велики саобраћај у мрежи мобилног телекомуникационог оператора и самим тим унети значајну грешку у предикциони модел. Примери неважећих MSISDN бројева су телефонски бројеви посебних служби (попут броја полиције, ватрогасаца, хитне помоћи...), позивним центрима, телемаркетинг центрима и слично. Део ових бројева се може лоцирати на основу интерних база података мобилног телекомуникационог оператора које се директно надовезују на јавно доступне телефонске именике, али један део се мора пронаћи анализом добијених података о чворовима [96].

Наиме, логично је претпоставити да ће позивни центри (на пример, позивни центри који служе за пријем позива корисника везаних за жалбе, приговоре и информације о неком предузећу) имати изразито велики улазни степен чвора, док ће са друге стране њихов излазни степен чвора у графу мобилне телекомуникационе мреже бити доста нижи. Насупрот позивним центрима, за телемаркетинг центре (попут центара за позивање корисника поводом телефонске продаје) је очекивано да имају изразито велики излазни степен чвора, док ће са друге стране улазни степен чвора у графу мобилне телекомуникационе мреже бити доста нижи. Ове претпоставке је требало потврдити кроз реалне податке.

Ради потврђивања изворних теза коришћен је принцип из три корака. Први корак је подразумевао прорачун метрика чворова у графу мобилне телекомуникационе мреже позива за све чворове, укључујући и телемаркетинг центре и позивне центре. Потом су сви добијени подаци укрштени са интерним база оператора о познатим специјалним бројевима. У трећем кораку, груписани су сви јавно доступни неважећи чворови и потврђена је појава два најчешћа обрасца понашања (позивни центри и телемаркетинг центри). Додатно, постављене су границе на начин да барем 90% свих јавно доступних неважећих бројева који потичу од позивних центара буду груписани заједно. Конкретно, на овај начин су груписани чворови који су имали степен чвора већи од 1000 и код којих је улазни степен чвора био барем 100 пута виши од излазног степена чвора. Након што је овај образац понашања MSISDN бројева у мобилној телекомуникационој мрежи позива дефинисан, сви остали чворови који су испуњавали наведени услов (а који нису били дефинисани као неважећи путем јавно доступних именика), су такође означени као неважећи позивни центри.

Сличан принцип је примењен и за телемаркетинг центре. Код њих је предложена граница била дефинисана као степен чвора од најмање 1000, као и излазни степен чвора који је барем 100 пута већи од улазног степена чвора. На основу ових критеријума је избачено укупно 800 чворова који су представљали инвалидне чворове у мрежи.

Битно је нагласити да се, иако конкретне границе за степен чвора и однос улазног и излазног чвора у графу не могу бити примењене као општа вредност која се може применити у мрежи неког другог мобилног оператора у свету (услед локалних специфичности, величине конкретног мобилног телекомуникационог оператора, и слично), овај принцип може једноставно прилагодити и применити у било ком другом сличном случају.

Додатни посебни проблем могу представљати позиви код којих је коришћен *GSM gateway*. У случају коришћења такве опреме CDR-ови нису исправно приказали позивајући или позивани број. Ово је био случај са приближно још 200 неважећих MSISDN бројева.

Препроцесирање података се завршава након што су сви неважећи MSISDN бројеви и CDR-ови у којима су они учествовали било као позивајући, било као позивани број елиминисани. Након што је комплетиран процес препроцесирања података и њихове припреме за генерисање модела, може се формирати финални граф мобилне телекомуникационе мреже позива. Финални граф мобилне телекомуникационе мреже позива садржи приближно 8.2 милиона чворова и 66.4 милиона грана, са просеком од 8.1 гране по чвору у графу. Анализа структуралних карактеристика финалног графа мобилне телекомуникационе мреже позива је показала да:

- Граф мобилне телекомуникационе мреже позива садржи 1003 неусмерених компоненти повезаности,
- Граф мобилне телекомуникационе мреже позива садржи 2.8 милиона компоненти јаке повезаности,
- Највећи комплетан граф који се може лоцирати унутар графа мобилне телекомуникационе мреже позива је величине 22 чвора.

Дакле, може се приметити да је елиминисање оквирно 1000 чворова довело до кидања приближно 800 000 грана, што је довело до повећања компоненти повезаности графа за 6%.

Уколико се пажљиво анализирају компоненте повезаности графа, битно је приметити да једна (убедљиво највећа) компонента повезаности садржи чак 99.96% од сви чворова целог графа мобилне телекомуникационе мреже позива. Све друге компоненте повезаности су доста мање, али не мање важне из угла предикције губитка корисника, што ће и бити показано у следећим поглављима.

Као што је наведено у претходном параграфу, све метрике чворова графа су израчунате за финални граф мобилне телекомуникационе мреже позива. Сам процес је одрађен коришћењем *SAS Enterprise Guide* софтвера за статистичку анализу и обраду великих података. Додатно, сваки анонимизовани MSISDN број је идентификован као корисник посматраног мобилног телекомуникационог оператора или као корисник неког другог мобилног телекомуникационог оператора. Узимајући у обзир да обрасци коришћења мобилних телефона припејд и постпејд корисника мреже посматраног мобилног оператора нису исти (услед великих разлика у цени пакета и услуга мобилне телефоније за припејд и постпејд кориснике), појавила се потреба за њиховим раздвајањем и засебном анализом. У следећим поглављима биће приказан детаљни процес анализе кластера само постпејд корисника коришћењем метрика теорије

графова, док ће припејд корисници бити изостављени. У тоталу, биће анализирано оквирно 1.7 милиона постпејд корисника.

6.2.2. Кластеризација MSISDN бројева

Пре него што се започне са процесом анализе кластера изабраних корисника, потребно је извршити припрему података која подразумева три основна корака: трансформацију података, избор узорка и стандардизацију података. Трансформација података је неопходна из разлога што је дистрибуција већине метрика изразито искривљена, те би се коришћењем логаритамске функције или коренивањем постигла максимизација нормалности расподеле појединачних променљивих. Променљиве степен чвора, улазни степен чвора и излазни степен чвора су трансформисане на следећи начин:

$$sqrt\_var = \sqrt{var} \tag{6.1}$$

где је var назив конкретне променљиве која се трансформише. Са друге стране, променљиве значајност чвора првог реда, значајност чвора другог реда, вредност ауторитета чвора, *hub* вредност чвора и сопствени вектор чвора су трансформисане на следећи начин:

$$log\_var = \ln(var + 0.1) \tag{6.2}$$

где је var назив конкретне променљиве која се трансформише. Битно је напоменути да је коефицијент 0.1 уведен у формулу узимајући у обзир да наведене променљиве могу имати вредност 0, те да без додатка коефицијента функција потенцијално не би била дефинисана.

Следећи корак би подразумевао избор узорка из целог скупа података на основу којег би се дефинисала правила кластер анализе. На основу тих правила би се накнадно кластерисала комплетна популација података. Узорак је изабран коришћењем једноставног случајног одабирања без понављања, те је одабрано 170 000 чворова (приближно 10% од укупне популације података која се анализира).

Скуп података који представља узорак је потом стандардизован коришћењем следеће формуле:

$$S = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6.3}$$

где је X променљива која се стандардизује, X_{min} минимум променљиве X у тестном скупу података, X_{max} максимум променљиве X у тестном скупу података, а S представља стандардизовану вредност. Може се приметити да је формула 6.3. аналогна раније представљеној формули 4.11. ($add = 0$, $multiply = 1$, $location = X_{min}$ и $scale$=$X_{max} - X_{min}$).

Као што је и објашњено у претходним поглављима, свих девет метрика чворова графа мобилне телекомуникационе мреже позива је подељено у три групе. То су метрике за усмерене графове, метрике за неусмерене графове и мера артикулације чвора. Ово је главни разлог за формирање два засебна процеса анализе кластера, један за метрике за усмерене графове, а други процес за метрике за неусмерене графове. Као што је и

описано у поглављу 4.3.5, основни корак при анализи кластера је одређивање броја кластера у подацима. Велики број различитих метода за одређивање броја кластера у подацима само додатно наглашава значај овог корака за успешност анализе кластера. У истраживању су коришћене три методе: кубни критеријум [62], псеудо-F статистика [63] и псеудо-T2 статистика [64]. Више детаља о овим методама за одређивање броја кластера у подацима је дато у поглављу 4.3.5. Битно је напоменути да не постоји један универзалан метод за одређивање броја кластера у подацима који би дао ултимативно исправно решење [97], [98]. Из тог разлога је потребно пронаћи консензус између више нумеричких метода за одређивање броја кластера ради проналажења оптималног решења.

Све три методе за одређивање броја кластера у подацима су имплементиране коришћењем *SAS Enterprise Guide* софтвера за статистичку анализу и обраду великих података, и то и за усмерене, и за метрике за неусмерене графове. Тестирано је педесет различитих варијанти броја кластера у подацима и то варијанте од 2 до 51 кластера. Анализиране су промене сваке од горенаведених статистика при промени броја кластера и то:

1. За кубни критеријум је као потенцијални кандидат за број кластера у подацима сматран сваки локални максимум који има вредност статистике већи од два,

2. За псеудо-F статистику је као потенцијални кандидат за број кластера у подацима сматран сваки локални максимум,

3. За псеудо-T2 статистику је као потенцијални кандидат за број кластера у подацима сматрана свака вредност која је одговарала локалном минимуму вредности статистике.

Сви критеријуми су подржавали решење са три кластера за метрике за неусмерене графове, односно са четири кластера за метрике за усмерене графове.

Након што је број кластера познат, у истраживању је коришћена Вардова метода минималне варијансе [78] као алгоритам који је изабран за саму кластер анализу. Као што је описано у поглављу 4.3.6, постоји велики број алтернативних алгоритама за кластер анализу. У истраживању је изабрана управо Вардова метода из разлога што је идеја аутора била да се минимализује варијанса унутар сваког добијеног кластера. Вардова метода хијерархијски спаја кластере на начин да, у свакој итерацији, сума квадрата растојања унутар кластера буде најмања од свих могућих опција спајања нека два кластера из претходне итерације. Рад [66] наводи да је Вардова метода је јако осетљива на изузетке у подацима. Управо из овог разлога препроцесирање података и уклањање неважећих чворова из графа телекомуникационе мреже позива додатно добијају на важности јер су се на тај начин елиминисали чворови који представљају изузетке, па нема формалних разлога који би оспоравали одлуку избора Вардове методе кластеризације. Још један од разлога за коришћење Вардове методе може представљати њена популарност, односно њено коришћење у повезаним радовима попут рада [99]. Додатно, Вардов алгоритам је нешто бржи од алтернативних метода што свакако није занемарљиво при обради овако великих скупова података (посебно хијерархијском методом кластеризације).

На основу кластера добијених Вардовом методом кластеризације, формирани су центроиди – то јест карактеристични представници сваког кластера. На основу добијених центроида се може применити партитивна кластеризација алгоритмом *K*-средина како би се сви преостали чворови (који се нису налазили у узорку), доделили себи најближем центроиду. Заправо, у питању је редуковани алгоритам *K*-средина у којем би се као почетне позиције *K* центроида користиле већ израчунате вредности центроида добијене Вардовом хијерархијском методом. Затим би се свака опсервација доделила најближем центроиду (коришћењем Еуклидског растојања), одређујући на тај начин финалне кластере. Уколико се анализира комплетан процес од три корака алгоритма *K*-средина представљен у поглављу 4.3.4, овај примењени редуковани алгоритам представља изворни алгоритам који има само једну итерацију (без померања почетних позиција *K* центроида). Резултати кластеризације на нивоу узорка и целе популације су представљени у табелама 6.1. и 6.2

Табела 6.1: Дистрибуција MSISDN бројева по кластерима коришћењем неусмерених метрика чворова графа мобилне телекомуникационе мреже позива

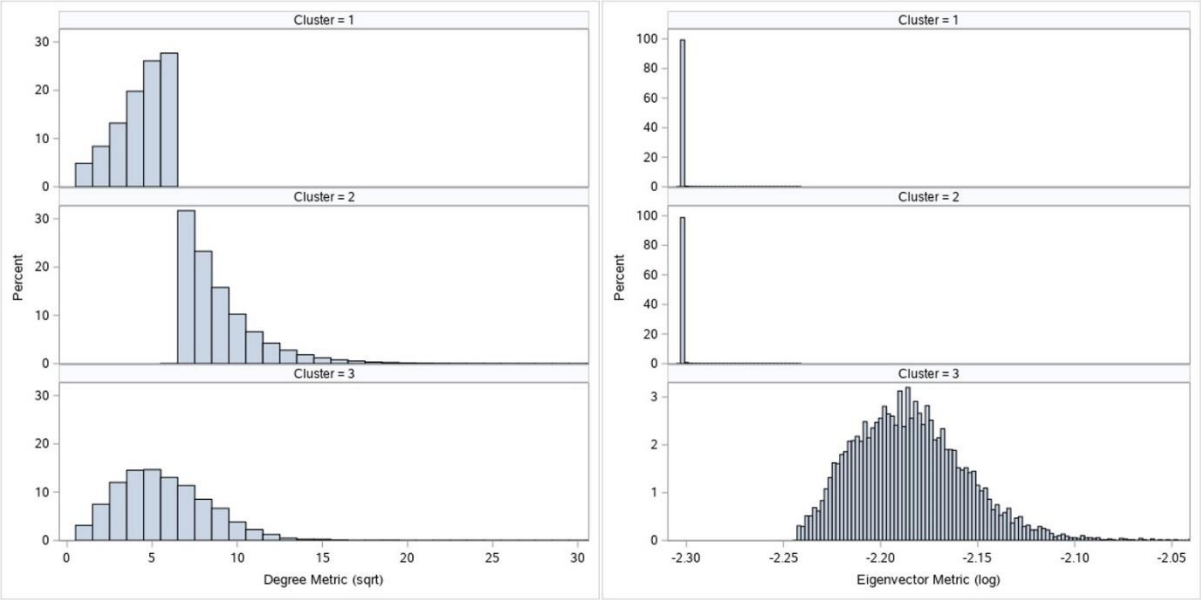
Кластер	Неусмерене метрике чворова графа	
	Проценат чланова у узорку	Проценат чланова у популацији
1	53.48	56.89
2	46.13	42.72
3	0.39	0.39

Табела 6.2: Дистрибуција MSISDN бројева по кластерима коришћењем усмерених метрика чворова графа мобилне телекомуникационе мреже позива

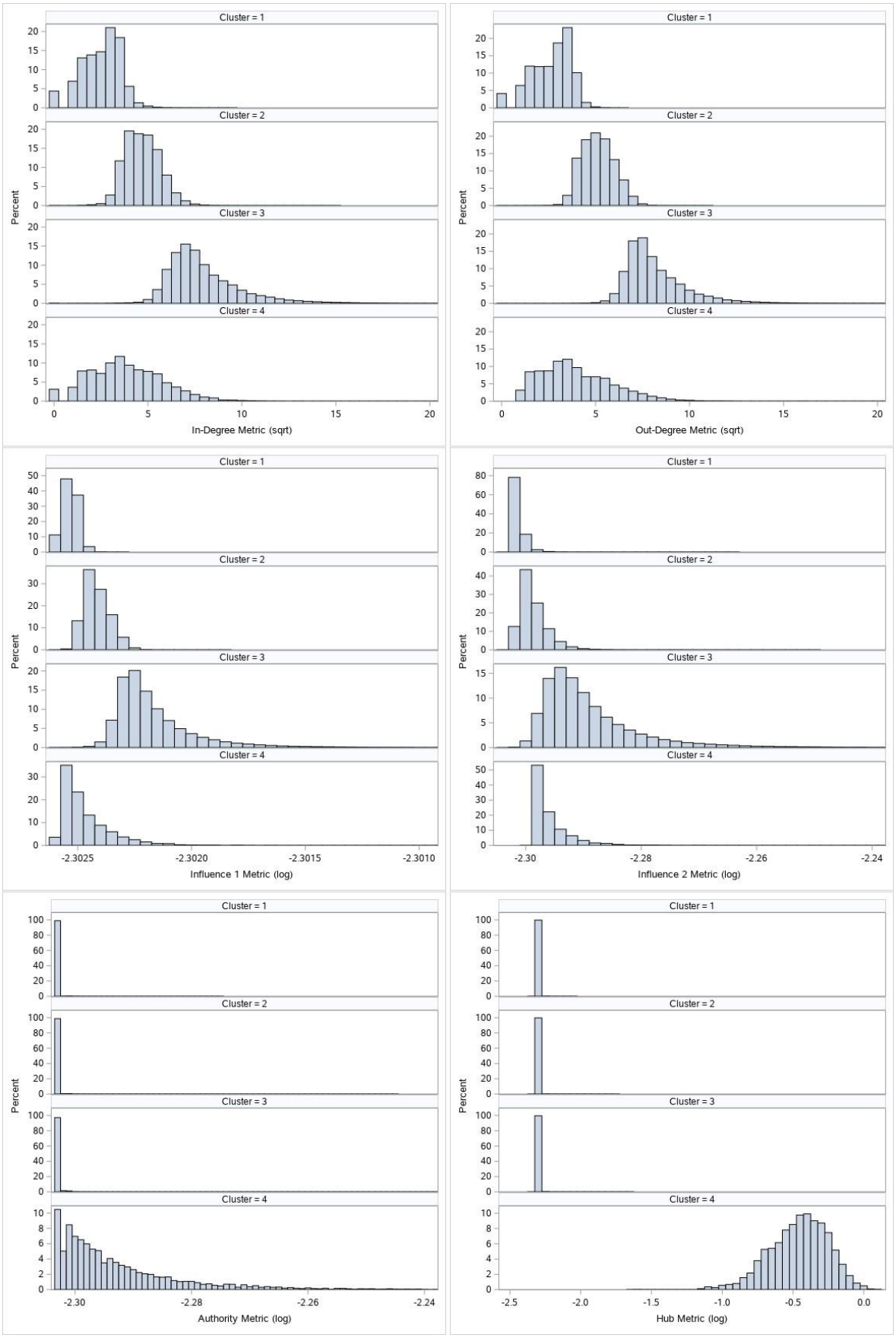
Кластер	Усмерене метрике чворова графа	
	Проценат чланова у узорку	Проценат чланова у популацији
1	42.53	39.9
2	41.04	45.25
3	16.07	14.47
4	0.36	0.38

Табеле 6.1. и 6.2 показују да су подаци унутар узорка исправно изабрани, узимајући у обзир да је дистрибуција чворова између различитих кластера унутар узорка и целе популације приближно иста.

Слике 6.2. приказује компаративне хистограмске графике расподеле броја чворова унутар сваког кластера (добијених кластеризацијом неусмерених метрика чворова графа) спрам неусмерених метрика чворова графа. Аналогно, слика 6.3. приказује компаративне хистограмске графике расподеле броја чворова унутар сваког кластера (добијених кластеризацијом усмерених метрика чворова графа) спрам усмерених метрика чворова графа.



Слика 6.2: Компаративне хистограмски графици расподеле броја чворова унутар сваког кластера добијених кластеризацијом неусмерених метрика



Слика 6.3: Компаративне хистограмски графици расподеле броја чворова унутар сваког кластера добијених кластеризацијом усмерених метрика

Сваки график је дефинисан својим нумеричким карактеристикама и дата му је нумеричка ознака. Ове нумеричке ознаке иду од један до три за неусмерене метрике, односно од један до четири за кластере добијене кластер анализом усмерених метрика чворова графа. Уколико се посматра слика 6.2. може се приметити да чворови који припадају кластеру два имају већи степен чвора у поређењу са чворовима који припадају кластеру један. Узимајући у обзир да чворови који припадају кластерима један и два добијеним кластеризацијом неусмерених метрика имају сличне вредности сопственог вектора, може се рећи да су чворови који припадају кластеру два значајнији за структуру графа од чворова који припадају кластеру један. Ако се детаљније анализирају чворови који припадају кластеру три добијеним кластеризацијом неусмерених метрика може се закључити да су они издвојени искључиво на основу високих вредности сопственог вектора. Високе вредности метрике сопственог вектора чвора могу указивати на то да су у питању јако добро повезани чворови који формирају окосницу, то јест језгро графа телекомуникационе мреже позива. Са друге стране, у примеру који је приказан у поглављу 5.4. приказан је феномен преноса метрике сопственог вектора чвора. На основу тог феномена се може тврдити да нису сви чворови са високим вредностима метрике сопственог вектора чвора подједнако битне за анализу којом се бави истраживање. Из тог разлога се уводи раздвајање чворова који су елементи кластера три добијеним кластеризацијом неусмерених метрика и то на следећи начин:

1. Чворови који се по својим другим карактеристикама (конкретно, по степену чвора) не разликују од чворова који припадају кластеру један, биће деградирани у кластер један добијен кластеризацијом неусмерених метрика,

2. Сви преостали чворови ће остати у кластеру три и представљаће право језгро и најбитније чворове графа мобилне телекомуникационе мреже позива, из угла анализе неусмерених метрика чворова.

Усмерене метрике чворова графа мобилне телекомуникационе мреже позива су формирале четири кластера. На сличан начин као и код кластеризације коришћењем неусмерених метрика, може се закључити да се прва три кластера разликују по значају. Очигледно је да је кластер један најмање значајан по питању усмерених метрика, следи кластер два, док би кластер три у себи садржао чворове који су најбитнији по свим метрикама изузев ауторитета и *hub* вредност чвора. Са друге стране, кластер четири (који би по логици ствари требао да представља најбитнији кластер), се издваја искључиво по вредностима ауторитета и *hub*-а, па се самим тим мора поново увести раздвајање чворова из најјачег кластера. Раздвајање чворова који су елементи кластера четири добијеним кластеризацијом усмерених метрика се изводи по следећем принципу:

1. Чворови који се по својим другим карактеристикама (конкретно, по улазном степену чвора, излазном степену чвора, значајности чвора првог реда и значајности чвора другог реда) не разликују од чворова који припадају кластерима један, односно два, биће редом деградирани у кластер један, односно два, добијен кластеризацијом усмерених метрика,

2. Сви преостали чворови ће остати у кластеру четири и представљаће право језгро и најбитније чворове графа мобилне телекомуникационе мреже позива, из угла анализе усмерених метрика чворова.

Може се приметити да су чворови који припадају кластерима означеним већим нумеричким ознакама битнији од оних који припадају кластерима са нижим нумеричким

ознакама (за кластеризације базиране и на усмереним и неусмереним метрикама). Треба приметити да мера артикулације чвора може све чворове поделити у две основне групе; једна група ће садржати све кластере који представљају артикулационе чворове, а друга група ће садржати све чворове који нису артикулациони чворови. Пошто је основни циљ истраживања дефинисање јединствене кластеризације свих чворова, биће уведена формула за унифицирање свих представљених кластеризација:

$$V = 2D + U + A/2 \tag{6.4}$$

где V означава нумеричку ознаку тоталног кластера чвора, D означава нумеричку ознаку кластера чвора базираног на усмереним метрикама, U означава нумеричку ознаку кластера чвора базираног на неусмереним метрикама, а A означава вредност метрике артикулације чвора. Коефицијенти у формули 6.4. су изабрани на начин да наглашавају значај усмерених метрика теорије графова у поређењу са неусмереним метрикама и метриком артикулације чвора у графовима мобилне телекомуникационе мреже позива. Наиме, узимајући у обзир да су мобилне телекомуникационе мреже позива усмерене по својој природи (услед својих специфичности, постојања позивајућег и позиваног броја и слично), за очекивати је да су усмерене метрике графова битније од неусмерених метрика при дефинисању значаја чворова. То се директно рефлектовало кроз коефицијенте формуле који су постављени на начин да је највећи коефицијент додељен усмереним метрикама, нешто нижи неусмереним, а најнижи мери артикулације чвора.

6.3. Резултати предикционог модела

6.3.1. Анализа података и предикционог модела

У претходним поглављима описан је процес генерисања јединствене кластеризације чворова (односно клијената) унутар мобилне телекомуникационе мреже позива комбиновањем усмерених метрика, неусмерених метрика, као и мере артикулације чвора. Када је описани процес примењен на целу популацију података, добијена је дистрибуција чворова по тоталним кластерима која је представљена на табели 6.3.

Табела 6.3: Дистрибуција чворова по тоталним кластерима

Кластер усмерене метрике	Кластер неусмерене метрике	Мера артикулације	Проценат корисника	Вредност тоталног кластера	Назив тоталног кластера
1	1	0	28.0125	3	Пратилац
1	1	1	12.0505	3.5	
1	2	0	0.0094	4	
1	2	1	0.0111	4.5	
2	1	0	7.9166	5	Стандардни
2	1	1	9.1662	5.5	
2	2	0	10.4609	6	
2	2	1	17.7645	6.5	
2	3	0	0.054	7	Лидер

2	3	1	0.023	7.5	
3	2	0	3.4634	8	
3	2	1	11.0023	8.5	
3	3	0	0.0023	9	Језгро
3	3	1	0.004	9.5	
4	2	0	0.0003	10	
4	2	1	0.0012	10.5	
4	3	0	0.0314	11	
4	3	1	0.0265	11.5	

Уколико се посматра дистрибуција чворова по свакој нумеричкој вредности тоталног кластера, могу се запазити четири целине – групе – које се издвајају у целој популацију чворова. Прва група је названа Пратиоцима, који подразумевају кориснике који нису витални у смислу анализе чворова метрикама теорије гафова. Ипак, чворови који припадају групи Пратилац показују неке интересантне карактеристике (а о којима ће бити више речи у следећим поглављима) које их раздвајају од просечних корисника. Друга група представља Стандардне кориснике. Стандардни корисници су боље повезани са остатком графа од Пратилаца, али ипак не представљају окосницу модела за предикцију губитка корисника. Група Стандардних корисника најбоље описује понашање просечног корисника мобилне телекомуникационе мреже позива, што ће и бити показано у следећим поглављима. Трећа и најбитнија група корисника представља Лидере – кориснике који представљају најзначајније чланове у анализираној друштвеној мрежи. То су корисници који имају разгранату мрежу комуницирања и који би требало да имају битан утицај на губитак корисника у мрежи. Последња група, Језгро, представља кориснике који су уско повезани у средишту мреже позива мобилног телекомуникационог оператора и који би (управо услед своје међусобне интерконекције) требали да буду у мањој опасности из угла напуштања мобилног телекомуникационог оператора.

Посматрањем дистрибуције чворова која је приказана у табели 6.3. се може приметити да не постоји један-на-један веза између кластера добијених коришћењем усмерених метрика (кластера са нумеричким ознакама од један до четири) и финалних кластера (Пратилац, Стандардни, Лидер и Језгро). На пример, чворови који припадају кластеру два добијеним коришћењем усмерених метрика могу бити финално сегментирани као Стандардни или као Лидери, у зависности од њиховог конкретног кластера добијеног коришћењем неусмерених метрика. Додатно, ако се посматрају чворови који припадају кластеру два добијеним коришћењем неусмерених метрика, може се приметити да:

1. Приближно 0.02% укупне популације чворова припада кластеру један који је добијен коришћењем усмерених метрика,

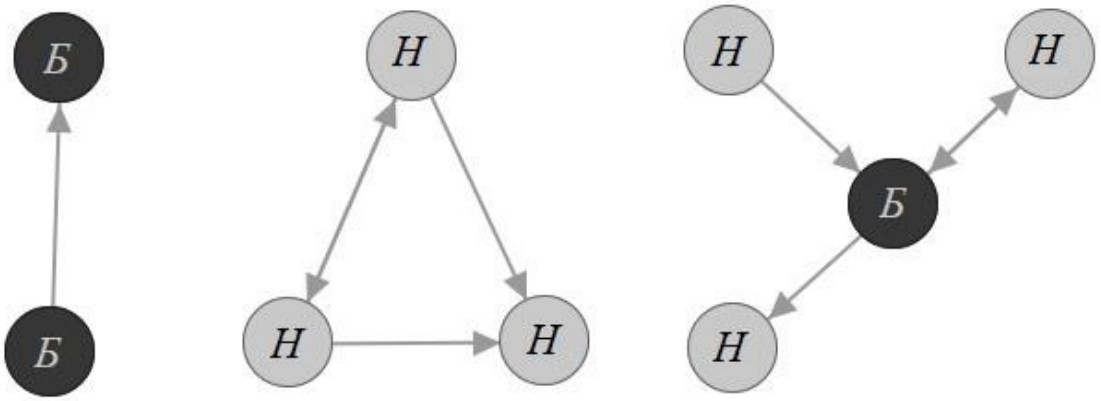
2. Приближно 28.22% укупне популације чворова припада кластеру два који је добијен коришћењем усмерених метрика,

3. Приближно 14.47% укупне популације чворова припада кластеру три који је добијен коришћењем усмерених метрика,

4. Приближно 0.002% укупне популације чворова припада кластеру четири који је добијен коришћењем усмерених метрика.

Дакле, види се да не постоји ни директна веза између кластеризације која је добијена коришћењем усмерених метрика и кластеризације која је добијена коришћењем неусмерених метрика.

У поглављу 6.2.1. је наведено да су све компоненте повезаности (независно од своје величине) битне за предложени предиктивни модел предвиђања вероватноће губитка корисника мобилног телекомуникационог оператора. Ипак, јавља се проблем из разлога што највећа компонента повезаности садржи 99.96% чворова графа, па ће све мање компоненте имати изузетно мали број чворова и самим тим неће имати могућности да имају високе вредности метрика. Штавише, скоро сви чворови који не припадају највећој компоненти повезаности ће припадати групи Пратилаца управо из горепоменутог разлога. Да би се овај проблем превазишао, потребна је додатна анализа. На слици 6.4. су приказани неки од најчешћих изгледа мањих компоненти повезаности.



Слика 6.4: Три примера мањих компоненти повезаности графа позива мобилног телекомуникационог оператора

На слици 6.4. чворови означени са *Б* представљају Битне чворове, а чворови означени са *Н* означавају Небитне чворове. Битни чворови ће представљати циљну групу модела, подједнако као и чворови означени као Лидери. Три универзална правила ће означавати чвор као Битан:

1. У мањим компонентама повезаности графа мобилне телекомуникационе мреже позива које садрже тачно два чвора, оба чвора ће бити означени као Битни,

2. У свим другим мањим компонентама повезаности графа мобилне телекомуникационе мреже позива, чвор ће бити маркиран као Битан уколико је артикулациони чвор,

3. Сви преостали чворови који припадају мањим компонентама повезаности графа мобилне телекомуникационе мреже позива ће бити означени као Небитни.

У анализи графа телекомуникационе мреже позива који је био предмет истраживања, постојало је 1970 чворова који су припадали мањим компонентама повезаности и сви су били припадници тоталног кластера Пратилац. Од тих чворова, њих 645 је накнадном анализом означено као Битни, док су преосталих 1325 означени као Небитни чворови.

Као што је описано у претходним поглављима, главна идеја истраживања је била да, уколико неки специфични и значајни корисник напусти мрежу посматраног мобилног телекомуникационог оператора, сви корисници који су комуницирали са тим значајним корисником представљају групу корисника која ће имати повишен ризик напуштања мобилног телекомуникационог оператора. На основу анализе која је представљена, дефинисана је група таквих значајних корисника. Значајни клијенти би били сви корисници који припадају кластерима Лидери и Битни корисници.

6.3.2. Верификација модела

Модел представљен у претходном поглављу је верификован коришћењем података о изгубљеним корисницима посматраног мобилног телекомуникационог оператора у периоду од четири месеца. Подаци о изгубљеним корисницима који су били доступни су садржали информације о разлогу напуштања оператора, датуму деактивације броја, као и (у случају пребацивања бројева између различитих мобилних оператора) о мобилном телекомуникационом оператору код кога је деактивирани корисник прешао. Процес верификације резултата модела предложеног у истраживању је извршен на следећи начин:

1. Посматрани су сви разлози деактивације постпејд бројева посматраног телекомуникационог оператора. Један део разлога за деактивацију бројева је игнорисан из разлога што нису представљали прави екстерни губитак корисника, На пример, уколико се корисник пребацује са постпејд на припејд тарифу посматраног мобилног телекомуникационог оператора, тада се из угла компаније није извршио никакав екстерни губитак корисника. Битно је нагласити да је један од главних разлога за дефинисање дистинкције између клијената који се пребацују са припејд на постпејд пакете унутар једног оператора и клијената који комплетно напуштају мрежу посматраног мобилног телекомуникационог оператора може бити и различита тарифа која се примењује за позиве унутар мреже, у поређењу са тарифом позива између бројева који припадају различитим операторима. Логична је претпоставка да ће клијенти који често међусобно комуницирају желети да се пребаце у мрежу истог мобилног телекомуникационог оператора да би смањили своје трошкове.

2. Сваки деактивирани чвор је анализиран и рачунат је број њему суседних чворова који су такође деактивирани у наредном периоду.

Резултати верификације модела су приказани у табелама 6.4. и 6.5. Конкретно, дистрибуција укупних и деактивираних корисника по тоталним кластерима је дата у табели 6.4. Са друге стране, у табели 6.5. је приказана мера деградације графа позива мобилног телекомуникационог оператора коју ће изазвати неки деактивирани корисник који је припадао одређеном тоталном кластеру. Због поверљивости података, део резултата је анонимизован (могу бити приказане само релације између вредности по кластерима, али не и конкретне вредности).

Табела 6.4: Дистрибуција укупних и деактивираних корисника по тоталним кластерима

Назив тоталног кластера	Укупан проценат корисника	Проценат деактивираних корисника
Пратилац	40.05	1.27 A
Стандардни	45.31	0.80 A
Лидер	14.54	0.88 A
Језгро	0.07	0.32 A
Битни	0.03	3.38 A
Тотална популација	100	A

Табела 6.5: Деградација графа позива мобилног телекомуникационог оператора у зависности од тоталног кластера деактивираног корисника

Назив тоталног кластера	Проценат деактивираних корисника који имају барем једног деактивираног суседа	Просечан проценат деактивираних суседа по једном деактивираном кориснику	Просечан број деактивираних суседа по једном деактивираном кориснику
Пратилац	31.3	25.7	A
Стандардни	48.9	30.4	4.09 A
Лидер	74.7	38.2	16.35 A
Језгро	57.1	2.6	1.01 A
Битни	88.2	100	-
Тотална популација	43.2	33	4.08 A

На основу података из табела 6.4. и 6.5. се могу донети следећи закључци:

1. Може се приметити да је вредност процента деактивираних корисника унутар тоталних сегмената Стандардни и Лидер, као и тоталне популације приближно једнак.

2. Проценат деактивираних корисника унутар тоталног кластера Језгро је осетно мањи у односу на преостале кластере. Ово се може објаснити тиме да је сваки чвор који се налази у кластеру Језгро представља окосницу графа позива мобилног телекомуникационог оператора, и самим тим га је много теже „преузети“ од стране конкурентских оператора. Са друге стране, чворови који припадају тоталном кластеру Пратилац су слабо повезани са остатком мреже

и самим тим су рањивији и подложнији за промену оператора у поређењу са остатком популације.

3. Проценат деактивираних корисника који имају барем једног деактивираног суседа је највећи за чворове који припадају кластеру Лидер. Ово конкретно значи да уколико један корисник из кластера Лидер напусти посматраног мобилног оператора, постоји приближно 75% вероватноће да ће барем један сусед посматраног чвора такође напустити мрежу посматраног мобилног телекомуникационог оператора. Овај проценат вероватноће је значајно нижи за преостале кластере, а посебно за кластер Пратилац где је вероватноћа више него дупло нижа и једнака скоро 31%.

4. Следећа битна статистика је просечан проценат деактивираних суседа по једном деактивираном кориснику. Ова вредност представља меру колико један деактивирани чвор реално деградира своју околину графа позива мобилног телекомуникационог оператора. Поново се може видети да су чворови који припадају кластеру Лидер највише склони деградацији своје околине у графу позива мобилног телекомуникационог оператора. Ипак, битно је приметити да чворови који припадају кластеру Језгро немају тенденцију деградирања структуре графа. Ово је случај и због тога што су чворови који припадају кластеру Језгро најчешће повезани са другим чворовима из истог кластера који нису склони напуштању посматраног мобилног телекомуникационог оператора.

5. Последња статистика представљана на табели 6.5. је просечан број деактивираних суседа по једном деактивираном кориснику. Као што је напоменуто, услед поверљивости података, резултати су анонимизовани. Ипак, и на основу анонимизованих података се може видети да чворови из кластера Лидер изазивају велику деградацију структуре графа позива, док са друге стране чворови из кластера Лидер и Пратилац изазивају најниже ремећење структуре графа позива мобилног телекомуникационог оператора.

6. Чворови који су означени као припадници кластера Битни се посматрају засебно, узимајући у обзир да обухватају само јако мали део укупне анализиране популације. Може се приметити да тај кластер има изузетно висок проценат деактивираних корисника, као и да ће сваки деактивирани корисник који припада кластеру Битни, након деактивирања, разорити у потпуности своју мању компоненту повезаности. Самим тим потврђује се почетна претпоставка да су мале компоненте повезаности графа позива мреже мобилног телекомуникационог оператора изузетно битне за предикцију губитка корисника.

Битно је подвући да се, уколико се посматра случај када је неки MSISDN број који представља чвор графа пребачен у мрежу неког конкурентског мобилног оператора, може приметити да се и његови суседи из графа позива такође пребацују у исту конкурентску мрежу у будућем периоду.

Сви модели за предикцију губитка корисника као свој финални резултат захтевају груписање тоталне популације клијената у две групе:

1. „ризични“ – висока вероватноћа губитка корисника

2. „сигурни“ – ниска вероватноћа губитка корисника

Из тог разлога ће улазни скуп података бити подељен на два дела. Први скуп корисника ће представљати све клијенте који су интераговали са неким клијентима из кластера Лидери или Битни, а који су у претходном периоду напустили посматраног мобилног телекомуникационог оператора. Тај скуп корисника ће бити означен као „ризични корисници“. Сви преостали корисници ће бити у другом скупу и биће означени као „сигурни корисници“.

6.3.3. Резултати кластеризације

Природа проблема предикције губитка корисника захтева нестандардну метрику за прорачун перформанси модела. Наиме, традиционалне метрике процене прецизности класификационих модела нису одговарајуће и не требају бити коришћене за случај када је један исход драстично мање вероватан од другог. Узимајући у обзир да се вероватноће губитка корисника (код иоле озбиљнијих оператора који нису у драматичним проблемима са пословањем) на нивоу целе популације клијената мере у неколико процената, употреба стандардне метрике попут прецизности није одговарајућа. На пример, уколико је стопа губитка корисника неког мобилног телекомуникационог оператора на нивоу целе популације једнака 1%, тада ће најтривијалнији модел који би сваког клијента означавао као „сигурног“ из угла вероватноће губитка корисника имати прецизност од чак 99%. Очигледно је да употребна вредност таквог предложеног модела не постоји, те се самим тим мора дефинисати нова метрика провере успешности предиктивног модела.

Један од главних пословних циљева при развоју модела за предикцију губитка корисника у телекомуникационој индустрији је генерисање мањег подскупа клијената који су врло ризични (из угла губитка корисника) и контактирати их у што краћем року ради превенције њиховог одласка од оператора. У литератури која се подробно бави предвиђањем губитка корисника у мрежама мобилних телекомуникационих оператора ([10], [9], [8], [11] и [12]) метода која је најчешће коришћена за проверу успешности модела је *Lift* метрика за најугроженијих десет процената популације. *Lift* метрика за најугроженијих десет процената популације корисника представља однос између стопе губитка корисника у десет процената најугроженијих клијената (у смислу губитка корисника) по процени неког модела, и просечне стопе губитка корисника на целој анализираној популацији. Формула за прорачун *Lift* метрике за најугроженијих *X* процената популације је дата са [100]:

$$Lift(X\%) = \frac{churn\ rate_{X\%}}{churn\ rate_{total\ population}} = \frac{\frac{n_{c,X\%}}{n_{t,X\%}}}{\frac{N_c}{N_t}} \tag{6.5}$$

где $churn\ rate_{X\%}$ представља стопу губитка корисника у *X%* процената најугроженијих корисника по неком предложеном моделу, $churn\ rate_{total\ population}$ представља просечну стопу губитка корисника на целој анализираној популацији, N_t укупан број корисника у целој популацији, N_c укупан број изгубљених корисника у целој популацији, $n_{t,X\%}$ укупан број корисника у изабраних *X%* процената популације и $n_{c,X\%}$ укупан број изгубљених корисника у изабраних *X%* процената популације. Управо прорачун *Lift* метрике на најугроженијих *X* процената популације може да омогући издвајање одговарајућег ризичног подскупа корисника које мобилни

телекомуникациони оператор може контактирати (где *X* означава проценат популације који оператор може да контактира у одговарајућем року).

Употреба *Lift* метрике биће описана следећим примером. Нека посматрани оператор садржи 1000 корисника и нека је примећена стопа губитка корисника 4%. Нека је предложени модел за откривање губитка корисника способан да издвоји десет процената популације и генерише вредност *Lift* метрике од 2. То би заправо значило да модел може, унутар 100 клијената које је дефинисао као најризичније, исправно предвидети 8 корисника који су напустили посматраног оператора. Са друге стране, уколико би се случајним избором изабрало 100 клијената (то јест 10% укупне популације), њих 4 би били клијенти који напуштају оператора. Дакле намеће се закључак да што је већа вредност *Lift* метрике, то је боља предикциони моћ предложеног модела за предвиђање губитка корисника мреже мобилног телекомуникационог оператора.

Као што је и напоменуто у претходном поглављу, два одвојена скупа података су формирана, један скуп садржи „ризичне кориснике“, а други „сигурне кориснике“. Вредност *Lift* метрике за дефинисани скуп „ризичних корисника“ је приказан у табели 6.6. Из разлога што број корисника који су означени као Битни или Лидери, а који су напустили посматраног мобилног телекомуникационог оператора у претходном периоду, није велики, укупан број корисника који се налазе унутар скупа „ризичних корисника“ мањи је од десет процената укупне популације. Узимајући у обзир да је *Lift* метрика за најугроженијих десет процената популације „златни стандард“ у литератури, скуп „ризичних корисника“ биће проширен неким корисницима из скупа „сигурних корисника“. Конкретно, скуп „ризичних корисника“ биће проширен корисницима који су у претходном периоду интераговали са клијентима који су припадали кластеру Пратилац, односно Стандардни (а који су у међувремену напустили посматраног мобилног телекомуникационог оператора). Вредности *Lift* метрике за ове проширене скупове података су представљене у табели 6.6. Може се лако приметити да вредност *Lift* метрике полако опада са убацивањем додатних корисника у скуп „ризичних корисника“. Ова појава је очекивана управо из разлога јер се група корисника који су најризичнији из угла губитка корисника проширује корисницима који нису толико угрожени. На тај начин се свесно обара предиктивна моћ модела.

Табела 6.6: Lift статистика различитих добијених скупова података; вредност Lift метрике за најугроженијих десет процената популације

Скуп података	Укупан проценат корисника	*Lift*
Интерагују са Битним	0.002	45.80
Интерагују са Лидерима	4.543	3.88
"Ризични корисници"	4.545	3.89
"Ризични корисници" + интерагују са Пратиоцима	6.815	3.67

"Ризични корисници" + интерагују са Пратиоцима и Стандардним	10.000	2.80

Такође је битно запазити да је систем који је предложен у истраживању реактиван, то јест захтева да један иницијални клијент напусти мобилног телекомуникационог оператора да би предвидео потенцијалне будуће изгубљене клијенте. Управо из овог разлога, предиктивна моћ модела предложеног у истраживању је ограничена (не постоји механизам за предикцију иницијалног губитка корисника). Имајући у виду ово ограничење, може се закључити да су остварене перформансе модела више него задовољавајуће и обећавајуће. Заправо, уколико се посматрају вредности *Lift* метрике само за скуп података означен са "ризични корисници", види се да се остварују изузетно високе вредности од приближно 3.9. Ово практично значи да се посматрањем отприлике 4.5% укупне популације клијената може лоцирати приближно 20% од укупног броја свих клијената који су напустили посматраног телекомуникационог оператора у посматраном периоду. Као што је и очекивано, највеће вредности *Lift* метрике су остварене анализом клијената који су комуницирали са клијентима који су припадали кластеру Битни пре деактивације. Тако остварене вредности *Lift* метрике иду до 45.8. Ово практично значи да је лоциран један мањи подскуп корисника мобилног телекомуникационог оператора који је веома рањив и подложан промени оператора, те се самим тим брзо мора активирати одређени напредни механизам задржавања таквих корисника.

6.4. Дискусија – поређење са алтернативним предиктивним методама

Да би се ставио акценат на остварене резултате добијене моделом анализе кластера предложеним у истраживању, као и да би се оправдао избор анализе кластера као типа алгоритма на коме је базирано истраживање, формиран је већи број алтернативних модела који се могу користити као референте тачке за прецизност предикције губитка корисника. Прецизније, да би се потврдио избор анализе кластера као изабраног алгоритма за развој модела, биће формиран већи број компаративних модела коришћењем два алтернативна алгоритма. Избор додатних алгоритама је извршен коришћењем повезане литературе и статистичких метода које се у њима примењују. Изабрани су алгоритми неуралних мрежа и стабла одлучивања [37].

Битно је нагласити да су и неуралне мреже и стабла одлучивања надгледани алгоритми машинског учења. Као што је наведено у претходним поглављима, алгоритми надгледаног машинског учења, за разлику од ненадгледаних метода машинског учења (чији је представник и кластер анализа), захтевају податке који су означени. Та ознака (лабела), ће управо представљати једину модификацију у изворним подацима, и биће названа ознака деактивације. Ознака деактивације неког корисника ће бити постављена на нулу уколико анализирани корисник није деактивиран (то јест није напустио посматраног мобилног оператора) у дефинисаном временском периоду. Насупрот томе, ознака деактивације неког корисника ће бити постављена на један уколико је анализирани корисник напустио мобилног телекомуникационог оператора у дефинисаном временском периоду. Ознака деактивације ће, заправо, представљати циљну променљиву чију ће вредност изабрани алгоритми надгледаног машинског учења

покушати да предвиде. Додатно, велики број различитих конфигурација модела за сваки одговарајући алгоритам су тестиране и оне које генеришу најбоље резултате у смислу *Lift* метрике су коришћене. Треба имати у виду да је, ради равноправности и уједначавања услова, постављено правило да се сваки анализирани алтернативни алгоритам мора извршити у времену које је било потребно алгоритму анализе кластера за комплетан прорачун.

Ради једноставности, у следећем делу текста ће опсервације које имају ознаку деактивације једнаку један, односно нула, бити нотиране као опсервације означене са један, односно нула, респективно. И алгоритми неуралних мрежа и стабла одлучивања су имплементирани на сличан начин. У првом кораку је формиран узорак на начин да је унутар узорка драстично повећана вероватноћа појављивања опсервација означених са један, у поређењу са целом анализираном и доступном популацијом. Конкретно, узорак је формиран тако да је унутар њега 33% опсервација означених са један и 67% опсервација означених са нула. Узорак је изабран на овакав начин да би модел који се формира имао боље перформансе при раздвајању изгубљених и неизгубљених корисника. Овај специфични узорак ће се користити за генерисање и валидацију сваког од алгоритама, док ће, са друге стране, тестирање модела ради поређења перформанси бити извршено на целој популацији. У другом кораку ће подаци из узорка бити раздвојени на две целине:

1. 70% опсервација из узорка ће бити коришћено за тренирање нових модела,
2. 30% опсервација из узорка ће бити коришћено за валидацију нових модела.

Као што је и наведено у претходним поглављима, имајући у виду тип променљиве која се прорачунава (ознака деактивације представља бинарни циљ), велики број различитих конфигурација алгоритма стабла одлучивања може бити коришћен. На пример, постоји велики број критеријума за гранање у стаблима одлучивања, попут критеријума базираних на *Gini* индексу [101], *chi-squared* тесту [102], ентропији [103], и слично. Конкретно, у истраживању су коришћени критеријуми базирани на *Gini* индексу и *chi-squared* тесту. Дакле, на основу критеријума гранања стабла, коришћени су алгоритми *Classification and Regression Trees* [101] (CART) и *Chi-square Automatic Interaction Detector* [102] (CHAID) за генерисање стабала одлучивања. Додатно, анализиран је ефекат промене времена *Bonferroni* прилагођавања на моменат пре или након избора критеријума гранања [102]. На крају, тестиране су и различите конфигурације у смислу максималног броја грана стабла које могу настати из сваког чвора, као и максимална дубина стабла. Када су формирани финални модели, вредност *Lift* метрике за најугроженијих десет процената популације је рачуната за тестни скуп података (целу популацију) и та вредност је коришћена за поређење резултата модела.

Дакле, конкретне вредности конфигурација које су тестиране за алгоритам стабла одлучивања су:

1. Максимална дубина стабла: 6-10,
2. Максимални број грана који настаје из једног чвора: 2-6,
3. Време *Bonferroni* прилагођавања: пре или после избора критеријума гранања.

Различити модели су формирани и коришћењем алгоритма неуралних мрежа. Конкретно, посматрани су различити типови мрежне архитектуре, попут *Linear Perceptron* (LP), *Multilayer Perceptron* (MLP) и *Radial Basis Function* (RBF) мреже. У истраживању су анализиране и *Ordinary RBF* (ORBF) и *Normalized RBF* (NRBF) мреже [104]. Додатно, анализирани су ефекти коришћења различитих комбинационих функција које се користе за генерисање RBF мрежа. Такође, посматран је ефекат коришћења различитог броја скривених неурона. Као и у случају коришћења алгоритма стабла одлучивања, након што су формирани финални модели израчуната је вредност *Lift* метрике за најугроженијих десет процената популације за тестни скуп података (целу популацију).

Конкретне вредности конфигурација које су тестиране за алгоритам неуралних мрежа су:

1. Број скривених неурона: 3-20,
2. Тип мрежна архитектуре:
    - *Linear Perceptron* (LP),
    - *Multilayer Perceptron* (MLP),
    - *Radial Basis Function* (RBF) мреже које користе комбинационе функције:
        i. *Unequal Width and Height* (UN),
        ii. *Unequal Widths* (UW),
        iii. *Equal Width and Height* (EQ),
        iv. *Equal Widths* (EW) ,
        v. *Equal Height* (EH),
        vi. *Equal Volumes* (EV).

Модели базирани на стаблима одлучивања као и модели базирани на неуралним мрежама су имплементирани коришћењем *SAS Enterprise Miner* софтвера за статистичку анализу великих скупова података. Финалне остварене вредности *Lift* метрике за конфигурације које су се показале најбоље у својим класама су представљене у табелама 6.7. и 6.8.

Табела 6.7: Lift статистика код алтернативних модела базираних на алгоритму стабала одлучивања

Тип стабла одлучивања	Метод процене стабла одлучивања	Максимална дубина стабла	Максимални број грана који настаје из једног чвора	Број чворова у стаблу	*Lift*

SAS Default Tree	*SAS Default*	6	2	65	2.38
SAS Default Tree	Просечна квадратна грешка	6	2	65	2.42
CHART	Просечна квадратна грешка	10	2	317	2.46
CHART	Просечна квадратна грешка	6	2	101	2.48
CHAID	*SAS Default*	6	6	83	2.48
CHAID	Просечна квадратна грешка	6	4	73	2.49

Табела 6.8: Lift статистика код алтернативних модела базираних на алгоритму неуралне мреже

Тип неуралне мреже	Комбинационе функције	*Lift*
LP	*N/A*	2.47
MLP	*N/A*	2.56
ORBF	*EQ*	2.41
ORBF	*UN*	2.48
NRBF	*EQ*	2.57
NRBF	*EV*	2.51
NRBF	*EH*	2.65

Остварени резултати представљени у табели 6.7. показују да се највећа вредност *Lift* метрике за најугроженијих десет процената популације код модела базираних на стаблима одлучивања може остварити коришћењем CHAID алгоритма са 73 чворова, максимално четири гране из сваког чвора, дубином од шест чворова, са временом *Bonferroni* прилагођавања након избора гранања. Тако остварена вредност *Lift* метрике за најугроженијих десет процената популације је једнака 2.49. Са друге стране, табела 6.8. показује да модели базирани на алгоритмима неуралних мрежа имају нешто боље вредности *Lift* метрике за најугроженијих десет процената популације у односу на моделе базиране на алгоритму стабла одлучивања. Најбоља конфигурација код алгоритма неуралних мрежа користи *Normalized Radial Basis Function* и комбинационе функције *Equal Height* (NRBFEH) са шест неурона у скривеном слоју. Та конфигурација остварује вредност *Lift* метрике за најугроженијих десет процената популације од 2.65, што је и даље ниже од вредности коју може остварити метод анализе кластера предложен у истраживању.

Битно је напоменути да обе алтернативне методе имају предиктивни елемент који је другачијег типа од предложеног модела базираног на анализи кластера. Наиме, метод анализе кластера је базиран на дефинисању значајних корисника који могу утицати на друге клијенте, док алтернативне методе раде на проналажењу вредности ознаке деактивације једнаке један. Дакле, метод анализе кластера је реактиван (не може лоцирати изворног корисника који напушта мобилног телекомуникационог оператора), а алтернативне методе су директне (то јест директно рачунају вероватноћу да ће анализирани клијент напустити мобилног телекомуникационог оператора). Теоретски би директни модели требали да имају већу предиктивну моћ од реактивних, али у овом конкретном истраживању се показује да избор алгоритма може преокренути ову логичну претпоставку.

На основу остварених вредности *Lift* метрике за сваки од предложених алтернативних модела за предикцију губитка корисника у мобилним телекомуникационим мрежама, може се закључити да је изабрани метод анализе кластера заиста најоптималнији. Дакле, обе алтернативне методе (неуралне мреже и стабла одлучивања), и различите конфигурације сваког алгоритма нису успеле да остваре вредност *Lift* метрике која би била већа од вредности добијене коришћењем анализе кластера.

7. ЗАКЉУЧАК

Анализа друштвених мрежа коришћењем теорије графова има велике примене и може се користи у великом броју различитих дисциплина, као што је пословна интелигенција [105], маркетинг [106], здравствена заштита [99], откривање превара везаних за осигурања у аутомобилској индустрији [107], откривање превара у финансијским трансакцијама [108], анализи сурфовања интернетом [109], телекомуникацијама [110], предикцији губитка корисника [111], и слично. Узимајући у обзир да је индустрија мобилних телекомуникација веома компетитивна, витално је за све мобилне телекомуникационе операторе да стекну максимални могући увид у понашање и преференце својих клијената. Студија представљена у [112] је показала да скоро 75% клијената који напуштају мобилне телекомуникационе операторе има навику да говори своје негативно искуство са оператором барем једној особи из свог окружења. Дакле, информација о обрасцу комуницирања неког конкретног незадовољног клијента може бити витална за предикцију његове евентуалне одлуке за напуштањем свог тренутног мобилног телекомуникационог оператора. Управо анализа друштвених мрежа може пружити тај дубљи увид у понашање клијената.

Стандардни приступ при предикцији вероватноће губитка клијената мобилних телекомуникационих оператора подразумева анализу сваког клијента понаособ и генерисање десетина до стотина различитих индикатора (*Key Performance Indicators* – KPI). Ти индикатори могу описивати широки спектар понашања клијената унутар мобилног телекомуникационог оператора попут: основних демографских података, информација о наплати рачуна, тарифном пакету, истеку уговорених обавеза, и слично. Једни од индикатора који могу бити јако вредни су везани за активност корисника у мрежи мобилног телекомуникационог оператора у смислу броја позива, трајања позива и слично, као и трендови тих индикатора у посматраном периоду. Други системи за предикцију губитка корисника могу анализирати учесталост позива конкретног корисника ка корисничком центру, истовремено покушавајући да квантификују незадовољство корисника анализом њиховог тоналитета у току позва и на основу тога предвиђати вероватноћу губитка претплатника. На крају, постоје и радови који су базирани на анализи друштвених мрежа и дефинисању група значајних корисника који могу утицати на губитак корисника у мрежи (као на пример радови [10] и [8]).

У истраживању представљеном у овој докторској дисертацији представљен је модел за предикцију губитка корисника мобилних телекомуникационих оператора који је базиран на принципима анализе друштвених мрежа, ненадгледаних метода машинског учења и теорији графова. При развоју модела, прво је формиран граф мреже позива мобилног телекомуникационог оператора коришћењем записа о позивима унутар посматраног периода. Израчунате су вредности девет изабраних метрика свих чворова анализираног графа позива, а тих девет метрика је подељено у три групе: усмерене

метрике, неусмерене метрике и мера артикулације чвора графа. У следећем кораку, елиминисани су сви неважећи чворови који не одговарају реалним корисницима мобилног телекомуникационог оператора. Такође, извршена су два засебна поступка анализе кластера (коришћењем усмерених и неусмерених метрика) на преосталим подацима. Уведена је формула за спајање две засебне кластеризације и формирано је четири финална кластера: Пратилац, Стандардни, Лидер и Језгро, уз додатак једне мање групе коју представљају Битни корисници. Након тога, дефинисана је финална група значајних корисника (клијенти који припадају кластерима Лидер и Битни), чији су чланови у могућности да своје незадовољство (које је резултовало напуштањем анализираног оператора) пренесу на своју друштвену мрежу коју чине клијенти посматраног мобилног телекомуникационог оператора који комуницирају са њима. Коришћењем реалних података о изгубљеним корисницима посматраног мобилног телекомуникационог оператора, лоцирана је једна група значајних корисника која је напустила посматраног оператора у предефинисаном периоду. Њихова околина је анализирана у наредном периоду ради провере вероватноће напуштања мобилног телекомуникационог оператора. Тврдња да ће велики број клијената из околине значајних корисника (који су променили оператора) напустити посматраног мобилног телекомуникационог оператора је потврђена коришћењем реалних података о деактивираним претплатницима мобилног телекомуникационог оператора. На крају, у последњем кораку су перформансе предложеног решења базираног на ненадгледаним методама машинског учења поређене са перформансама алтернативних решења базираних на надгледаним методама машинског учења (коришћењем алгоритама стабала одлучивања и неуралних мрежа).

Резултати истраживања које је представљено у овој докторској дисертацији су квантификовани коришћењем *Lift* метрике. Показано је да постоји група ризичних корисника (приближно 4.5% од свих анализираних корисника) код којих је вредност *Lift* метрике приближно једнака 3.88. Са друге стране, укупна остварена вредност *Lift* метрике за најугроженијих десет процената анализиране популације је једнака 2.8. Узимајући у обзир да већина радова који су представљени у повезаној литератури остварују вредности *Lift* метрике за најугроженијих десет процената популације у опсегу између 2 и 5, може се закључити да су резултати истраживања представљеног у овој докторској дисертацији обећавајући и да представљају добру полазну основу за даљи развој модела за предикцију губитка корисника у мрежама мобилних телекомуникационих оператора.

Битно је нагласити да, иако је донекле очекивано да ће хомофилија и друштвене везе између клијената утицати на губитак претплатника посматраног мобилног телекомуникационог оператора, додатна вредност коју пружа истраживање представљено у овој докторској дисертацији представља могућност квантификације овог ефекта. Додатно, показано је да један посебни кластер клијената посматраног мобилног телекомуникационог оператора (клијенти који припадају кластеру Језгро), неће погубно утицати на своју околину. Дакле, показано је да уколико клијент који припада кластеру Језгро напусти посматраног мобилног телекомуникационог оператора, то неће довести до масовног губитка корисника из његове околине у мрежи мобилног телекомуникационог оператора. Овај ефекат се јавља из разлога што је околина клијената који припадају кластеру Језгро најчешће такође члан истог сегмента који је јако везан за посматраног мобилног телекомуникационог оператора и као такви неће бити у опасности да промене свог провајдера телекомуникационих услуга. Сумирано, очекивано хомофилно понашање клијената посматраног мобилног

телекомуникационог оператора није исто код клијената који припадају различитим финалним кластерима.

Додатно, још једна од предности модела за предикцију губитка корисника мобилних телекомуникационих оператора који је представљен у истраживању је то да може релативно једноставно бити комбинован са другим моделима који су базирани на алтернативним изворима информација. На пример, комбиновањем резултата модела базираног на анализи кластера приказаног у истраживању и модела за предикцију губитка корисника базираног на задовољству корисника (попут оног који је представљен у раду [112], у коме су анализирани подаци о цени, непријатностима које корисник доживљава, кваровима у мрежи, и слично), стопа прецизности комбинованог модела би била још већа.

Нажалост, услед специфичног типа података који је потребан за тип анализе који је представљен у дисертацији, јако је тешко обезбедити још неки скуп података истог типа (CDR податке неког другог мобилног телекомуникационог оператора и податке о деактивираним корисницима) који би се могао искористити за додатну потврду успешности предложеног модела. Такође, потребно је имати у виду да центроиди који су формирани на основу анализе метрика теорије графова добијених анализом графа позива једног мобилног телекомуникационог оператора не морају бити исти као и центроиди који би се добили анализом записа позива мобилног телекомуникационог оператора са неког другог тржишта, или друге величине (корисничке базе). Дакле, предложени модел би морао да буде рекалибрисан ради успешне и исправне примене код других мобилних телекомуникационих оператора. Ипак, принцип анализе података и њиховог моделовања представљен у истраживању је довољно општи да се може применити са било којим доступним скупом података.

Закључци истраживања представљеног у овој докторској дисертацији могу бити коришћени као смернице при развоју сличних модела базираних на анализи друштвених мрежа. Ипак, битно је подвући да је потребно имати у виду све специфичности које могу произаћи из коришћења анализе друштвених мрежа у другим индустријама. Конкретно, потребно је узети у обзир да свака различита примена анализе друштвених мрежа захтева детаљно познавање друштвених и структуралних карактеристика добијеног графа, ради дефинисања успешног предиктивног модела.

8. ЛИТЕРАТУРА

[1] S. H. Fuller and L. I. Millett, The Future of Computing Performance: Game Over or Next Level?, Washington, DC: National Academy Press, 2011.

[2] P. E. Ceruzzi, E. Paul и W. Aspray, A history of modern computing, Cambridge, MA: MIT press, 2nd ed, 2003.

[3] V. Mayer-Schönberger and K. Cukier, Big data: A revolution that will transform how we live, work, and think, New York, NY: Houghton Mifflin Harcourt, 2013.

[4] P. Simon, Too big to ignore: the business case for big data, Hoboken, NJ: John Wiley & Sons, 2013.

[5] V. Dhar, "Data science and prediction," *Communications of the ACM,* vol. 56, no. 12, pp. 64-73, 2013.

[6] N. J. Salkind and B. B. Frey, Statistics for people who (think they) hate statistics, Los Angeles, CA: Sage Publications, Incorporated, 2019.

[7] P. Flach, Machine learning: the art and science of algorithms that make sense of data, New York, NY: Cambridge University Press, 2012.

[8] Y. Richter, E. Yom-Tov and N. Slonim, "Predicting Customer Churn in Mobile Networks through Analysis of Social Groups," in *SIAM International Conference on Data Mining (SDM 2010)*, Columbus, Ohio, USA, April - May 2010.

[9] I. Bose and X. Chen, "Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn," *Journal of Organizational Computing and Electronic Commerce,* vol. 19, no. 2, p. 133–151, 2009.

[10] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. Nanavati and A. Joshi, "Social Ties and their Relevance to Churn in Mobile Telecom Networks," in *11th International Conf. Extending Database Technology: Advances in Database Technology (EDBT 08)*, Nantes, France, 25–29 March 2008.

[11] C. Wei and I. Chiu, "Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach," *Expert systems with applications,* vol. 23, p. 103–112, 2002.

[12] S. Hung, D. Yen and H. Wang, "Applying Data Mining to Telecom Churn Management," *Expert systems with applications,* vol. 31, p. 515–524, 2006.

[13] A. Amin, F. Rahim, M. Ramzan and S. Anwar, "A prudent based approach for customer churn prediction," in *International Conf. Beyond Databases, Architectures and Structures (BDAS 2015),* Ustron, Poland, 26–29 May 2015.

[14] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research,* vol. 94, p. 290–301, 2019.

[15] C. Yang, X. Shi, L. Jie and J. Han, "I Know You'll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application," in *24th ACMSIGKDD International Conference on Knowledge Discovery & Data Mining*, London, UK, 19–23 August 2018.

[16] E. Xevelonakis and P. Som, "The impact of social network-based segmentation on customer loyalty in the telecommunication industry," *Journal of Database Marketing & Customer Strategy Management,* vol. 19, p. 98–106, 2012.

[17] U. Droftina, M. Štular and A. Košir, "A diffusion model for churn prediction based on sociometric theory," *Advances in Data Analysis and Classification,* vol. 9, p. 341–365, 2015.

[18] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access,* vol. 4, pp. 7940-7957, 2016.

[19] A. Amin, C. Khan, I. Ali and S. Anwar, "Customer churn prediction in telecommunication industry: With and without counter-example," in *Mexican International Conference on Artificial Intelligence (MICAI 2014),* Tuxtla Gutiérrez, Mexico, 16–22 November, 2014.

[20] A. Amin, S. Shehzad, C. Khan, I. Ali and S. Anwar, "Churn prediction in telecommunication industry using rough set approach," *New Trends in Computational Collective Intelligence*, pp. 83-95, 2015.

[21] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain and K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing,* vol. 237, pp. 242-254, 2017.

[22] Q. Han and P. Ferreira, "The role of peer influence in churn in wireless networks," in *2014 International Conference on Social Computing*, Beijing, China, 4–7 August, 2014.

[23] A. Rehman and A. R. Ali, "Customer churn prediction, segmentation and fraud detection in telecommunication industry," in *4th ASE International Conference on Big Data*, Harvard University, Cambridge, MA, USA, 14-16 December, 2014.

[24] S. Mitrović, B. Baesens, W. Lemahieu and J. D. Weerdt, "tcc2vec: RFM-informed representation learning on call graphs for churn prediction," *Information Sciences,* 2019.

[25] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Francisco, CA, USA, 13-17 August, 2016.

[26] H. Jain, A. Khunteta and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," *Procedia Computer Science,* vol. 167, pp. 101-112, 2020.

[27] J. Moreira, A. Carvalho and T. Horváth, A general introduction to data analytics, Hoboken, NJ: John Wiley & Sons, 2019.

[28] J. Gantz and D. Reinsel, "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East Executive Summary: A Universe of Opportunities and Challenges," IDC, 2012.

[29] "Cisco visual networking index: global mobile data traffic forecast update, 2017–2022," [White paper], CISCO Forecast Global Mobile Data Traffic, 2017 : 2022, update 2019.

[30] S. Kostić, M. Đuričić, M. Simić and K. M., "Data Mining and Modeling use Case in Banking Industry," in *IEEE 26th Telecommunications Forum (TELFOR),* Belgrade, Serbia, November 2018.

[31] A. Sorescu, "Data-driven business model innovation," *Journal of Product Innovation Management,* vol. 34, no. 5, pp. 691-696, 2017.

[32] D. F. Norton and M. J. Norton, David Hume: A Treatise of Human Nature: Volume 1: Texts, London: Oxford University Press, 2007.

[33] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation,* vol. 1, no. 1, pp. 67-82, 1997.

[34] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development,* vol. 3, no. 3, pp. 210-229, 1959.

[35] D. Cielen, A. Meysman and M. Ali, Introducing data science: big data, machine learning, and more, using Python tools, Shelter Island, NY: Manning Publications Co., 2016.

[36] A. A. Patel, Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data, Sebastopol, CA: O'Reilly Media, 2019.

[37] P. Dangeti, Statistics for machine learning, Birmingham: Packt Publishing Ltd., 2017.

[38] B. Everitt and A. Skrondal, The Cambridge dictionary of statistics, New York, , :, NY: Cambridge University Press, 4th ed., 2010.

[39] B. S. Everitt, S. Landau, M. Leese and D. Stahl, Cluster analysis, Hoboken, NJ: John Wiley & Sons, 5th ed., 2011.

[40] R. M. Cormack, "A review of classification," *Journal of the Royal Statistical Society: Series A (General)*, vol. 134, no. 3, pp. 321-353, 1971.

[41] A. Gordon, Classification, London: Chapman & Hall, 2nd ed., 1999.

[42] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Hoboken, NJ: John Wiley & Sons, 2009.

[43] C. L. Liu, Introduction to combinatorial mathematics, New York, NY: McGraw-Hill, 1968.

[44] W. V. O. Quine, Ontological relativity and other essays, New York, NY: Columbia University Press, 1969.

[45] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857-871, 1971.

[46] B. D. Nelson, "Variable reduction for modeling using PROC VARCLUS," in *Twenty-Sixth Annual SAS Users Group International Conference*, Cary, NC, USA, 2001.

[47] M. R. Anderberg, Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks, New York, NY: Academic press, 1973.

[48] R. Sanche and K. Lonergan, "Variable reduction for predictive modeling with clustering," in *Casualty Actuarial Society Forum*, March 2006.

[49] B. S. Everitt and G. Dunn, Applied multivariate data analysis, Hoboken, NJ: John Wiley & Sons, 2nd ed, 2001.

[50] T. F. Cox and M. A. Cox, Multidimensional scaling, London: Chapman & Hall, 2nd ed, 2001.

[51] L. Guttman, "Some necessary conditions for common-factor analysis," *Psychometrika*, vol. 19, no. 2, pp. 149-161, 1954.

[52] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and psychological measurement,* vol. 20, no. 1, pp. 141-151, 1960.

[53] K. A. Yeomans and P. A. Golder, "The Guttman-Kaiser criterion as a predictor of the number of common factors," *Journal of the Royal Statistical Society. Series D (The Statistician),* vol. 31, no. 3, pp. 221-229, 1982.

[54] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *Journal of classification,* vol. 5, no. 2, pp. 181-204, 1988.

[55] J. L. Fleiss and J. Zubin, "On the methods and theory of clustering," *Multivariate Behavioral Research,* vol. 4, no. 2, pp. 235-250, 1969.

[56] SAS Institute. SAS® 9.4 and SAS® Viya® 3.4 Programming Documentation / SAS/STAT User's Guide, "Introduction to Clustering Procedures: Elongated Multinormal Clusters," [Online]. Available: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_introclus_sect008.htm&docsetVersion=15.1&locale=en. [Accessed 20. 01. 2021.].

[57] R. M. Elashoff, Perspectives in biometrics, New York, NY: Academic Press, 1975.

[58] F. H. C. Marriott, "Practical problems in a method of cluster analysis," *Biometrics,* vol. 27, no. 3, pp. 501-514, 1971.

[59] F. H. C. Marriott, "Optimization methods of cluster analysis," *Biometrika,* vol. 69, no. 2, pp. 417-421, 1982.

[60] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Fifth Berkeley symposium on mathematical statistics and probability*, June, 1967.

[61] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika,* vol. 50, no. 2, pp. 159-179, 1985.

[62] W. S. Sarle, "The Cubic Clustering Criterion; SAS Technical Report A-108," SAS Institute Inc., Cary, NC, USA, 1983.

[63] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics,* vol. 3, no. 1, pp. 1-27, 1974.

[64] R. Duda and P. Hart, Pattern Classification and Scene Analysis, New York, NY: John Wiley & Sons, 1973.

[65] E. M. L. Beale, Euclidean cluster analysis, 1969: Scientific Control Systems Limited.

[66] G. W. Milligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika,* vol. 45, no. 3, pp. 325-342, 1980.

[67] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin,* vol. 38, pp. 1409-1438, 1958.

[68] Q. He, A review of clustering algorithms as applied in IR, M.S. thesis, Graduate School of Library and İnformation Science University of İllinois at Urbana-Compaign, 1999.

[69] T. J. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Copenhagen, Denmark: I kommission hos E. Munksgaard, 1948.

[70] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki, "Sur la liaison et la division des points d'un ensemble fini," *Colloquium Mathematicum,* vol. 2, no. 3-4, pp. 282-285, 1951.

[71] G. Der and B. S. Everitt, A handbook of statistical analyses using SAS, London: Chapman & Hall, 2008.

[72] L. L. McQuitty, "Similarity analysis by reciprocal pairs for discrete and continuous data," *Educational and Psychological measurement,* vol. 26, no. 4, pp. 825-831, 1966.

[73] J. C. Gower, "A comparison of some methods of cluster analysis," *Biometrics,* vol. 23, no. 4, pp. 623-637, 1967.

[74] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies I Hierarchical systems," *The Computer Journal,* vol. 9, no. 4, pp. 378-380, 1967.

[75] G. W. Milligan, "A study of the beta-flexible clustering method," *Multivariate behavioral research,* vol. 24, no. 2, pp. 163-176, 1989.

[76] M. A. Wong and T. Lane, "A kth nearest neighbour clustering procedure," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 45, no. 3, pp. 362-368, 1983.

[77] M. A. Wong, "A hybrid clustering method for identifying high-density clusters," *Journal of the American Statistical Association,* vol. 77, no. 380, pp. 841-847, 1982.

[78] J. H. J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association,* vol. 58, no. 301, pp. 236-244, 1963.

[79] M. J. Symons, "Clustering criteria and multivariate normal mixtures," *Biometrics,* vol. 37, no. 1, pp. 35-43, 1981.

[80] B. W. Silverman, Density estimation for statistics and data analysis, London: Chapman & Hall, 1998.

[81] D. W. Scott, Multivariate density estimation: theory, practice, and visualization, Hoboken, NJ: John Wiley & Sons, 2015.

[82] N. Biggs, E. K. Lloyd and R. J. Wilson, Graph Theory, 1736-1936, London: Oxford University Press, 1986.

[83] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Commentarii academiae scientiarum Petropolitanae,* pp. 128-140, 1741.

[84] Merian-Erben, "Image of Koenigsberg, Map by Merian-Erben," [Online]. Available: https://commons.wikimedia.org/wiki/File:Image-Koenigsberg,_Map_by_Merian-Erben_1652.jpg. [Accessed 20. 01. 2021.].

[85] M. V. Steen, Graph theory and complex networks, An introduction, Enschede: Maarten Van Steen, 2010.

[86] J. L. Gross, J. Yellen and P. Zhang, Handbook of graph theory, London: Chapman & Hall, 2013.

[87] D. Cvetković and M. Milić, Teorija grafova i njene primene, Beograd: Beogradski izdavačko - grafički zavod, 1971.

[88] R. Diestel, Graph Theory, volume 173 of Graduate texts in mathematics, New York, NY: Springer-Verlag, 2005.

[89] J. A. Bondy and U. S. R. Murty, Graph theory with applications, New York, NY: Elsevier, 1976.

[90] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM),* vol. 46, no. 5, pp. 604-632, 1999.

[91] S. M. Kostić, M. I. Simić and M. V. Kostić, "Social Network Analysis and Churn Prediction in Telecommunications Using Graph Theory," *Entropy,* vol. 22, no. 7, p. article no. 753, 2020.

[92] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea and A. Joshi, "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications," in *15th ACM International Conf. Information and Knowledge Management (CIKM 06),* Arlington, VA, USA, 6–11 November 2006.

[93] H. Seetha, M. N. Murty and B. K. Tripathy, Modern technologies for big data classification and clustering, Hershey, PA: IGI Global, 2017.

[94] M. Gladwell, The tipping point: How little things can make a big difference, New York, NY: Little, Brown and Company, 2006.

[95] C. A. R. Pinheiro, Social network analysis in telecommunications, Hoboken, NJ: John Wiley & Sons, 2011.

[96] S. Kostić, J. Sretenović, M. Simić and M. Kostić, "Detekcija ekstremnih korisnika u telekomunikacionim mrežama pomoću analize socijalnih mreža," in *60. konferencija za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku (ETRAN)*, Zlatibor, Serbia, June 2016.

[97] B. S. Everitt, "Unresolved Problems in Cluster Analysis," *Biometrics,* vol. 35, p. 169–181, 1979.

[98] H. H. Bock, "On Some Significance tests in Cluster Analysis," *Journal of classification,* vol. 2, no. 1, p. 77–108, 1985.

[99] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider and P. G. Bagos, "Using Graph Theory to Analyze Biological Networks," *BioData mining,* vol. 4, no. 1, p. article no. 10, 2011.

[100] S. Tufféry, Data mining and statistics for decision making, Chichester, West Sussex: John Wiley & Sons, 2011.

[101] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, New York, NY: Chapman & Hall, 1984.

[102] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 29, no. 2, p. 119–127, 1980.

[103] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1993.

[104] D. S. Broomhead and D. Lowe, "Radial Basis Functions, Multivariable Functional Interpolation and Adaptive Networks," Royal Signals and Radar Establishment, Malvern, UK, 1988.

[105] X. Chai, O. Deshpande, N. Garera, A. Gattani, W. Lam, D. S. Lamba, L. Liu, M. Tiwari, M. Tourn, Z. Vacheri, S. T. S. Prasad, S. Subramaniam, V. Harinarayan, A. Rajaraman, A. Ardalan, S. Das, P. Suganthan and A. Doan, "Social Media Analytics: The Kosmix Story," *IEEE Data Engineering Bulletin,* vol. 36, no. 3, p. 4–12, 2013.

[106] D. Kempe, J. Kleinberg and É. Tardos, "Maximizing the Spread of Influence through a Social Network," in *9th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining (KDD 2003),* Washington, DC, USA, 24–27 August 2003.

[107] L. Šubelj, Š. Furlan and M. Bajec, "An Expert System for Detecting Automobile Insurance Fraud Using Social Network Analysis," *Expert Systems with Applications,* vol. 38, no. 1, p. 1039–1052, 2011.

[108] J. C. Wang and C. Q. Chiu, "Detecting Online Auction Inflated-reputation Behaviors Using Social Network Analysis," in *the North American Association for Computational Social and Organizational Science (NAACSOS 2005)*, Notre-Dame, IN, USA, 26–28 June 2005.

[109] H. W. Park, "Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web," *Connections,* vol. 25, no. 1, p. 49–61, 2003.

[110] Y. Dong, Y. Yang, J. Tang, Y. Yang and N. V. Chawla, "Inferring User Demographics and Social Strategies in Mobile Social Networks," in *the 20th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining (KDD 2014),* New York, NY, USA, 24–27 August 2014.

[111] W. Gruszczynski and P. Arabas, "Application of social network inferred data to churn modeling in telecoms," *Journal of Telecommunications and Information Technology,* vol. 2, p. 77–86, 2016.

[112] S. M. Keaveney, "Customer Switching Behavior in Service Industries: An Exploratory Study," *Journal of marketing,* vol. 59, no. 2, p. 71–82, 1995.

Биографија

Стефан Костић је рођен 29.09.1990. године у Београду. Основну школу „Вељко Дугошевић“ завршава са одличним успехом (просек оцена 5,00) и Вуковом наградом, те као ђак генерације. Носилац је више награда на савезном и републичком нивоу из физике и математике. Средњу школу „Математичка гимназија“ завршава са одличним успехом (просек оцена 5,00) и Вуковом наградом. Основне академске студије на Електротехничком факултету у Београду уписао је 2009. године. Дипломирао је на одсеку за Телекомуникације и информационе технологије, смер Системско инжењерство 2013. године, са просечном оценом 9,18. Дипломски рад на тему „Принципи *UWB* технологије“ одбранио је са оценом 10. Мастер академске студије уписао је 2013. године на Електротехничком факултету, модул Системско инжењерство и радио комуникације, и завршио их је 2014. године са просечном оценом 9,83. Мастер рад на тему „Принципи компресије аудио и видео сигнала у савременим телекомуникационим системима“ одбранио је са оценом 10.

Докторске академске студије на Електротехничком факултету у Београду, модул Телекомуникације, уписао је 2014. године, где је положио све испите са просечном оценом 10. Области истраживања током докторских студија обухватају припрему и анализу података, алгоритме надгледаног и ненадгледаног машинског учења, као и науку о подацима у целини.

Стефан Костић је од октобра 2013. до маја 2015. године био запослен у фирми „*P3 Communications*“ као инжењер за верификацију података, где му је свакодневна обавеза била анализа записа на релацији мобилни уређај - базна станица ради откривања потенцијалних проблема у конфигурацији мрежних елемената. Од јуна 2015. до маја 2016. године, Стефан Костић је био запослен у фирми „Телеком Србија“ као софтверски инжењер. На тој позицији, бавио се подршком пословној интелигенцији у бизнису, као и моделовањем и анализом података које телекомуникациони оператор прикупља у редовном раду. Од јуна 2016. до октобра 2016. године, Стефан Костић је био запослен у фирми „*Ibis-Instruments*“ као аналитичар великих података где се бавио анализом података из сензорских мрежа. Од новембра 2016. до априла 2019. године, Стефан Костић је био запослен у фирми „Ерсте банка“ као *Data Scientist*, односно као *Lead Data Scientist*. У том периоду између осталог бавио се и моделовањем података за регулаторне потребе, креирањем стрес тестова пословања банке, моделовањем података за подршку пословној интелигенцији и продаји, као и различитим пројектима везаним за пословање банке на локалном и регионалном нивоу. Од маја 2019. запослен је у фирми „*United Group*“ на позицији *Lead Data Scientist* где је наставио са радом на моделовању различитих врста података кабловских и мобилних оператора у земљи и региону југоисточне Европе.

Стефан Костић је аутор рада објављеног у научном часопису међународног значаја са SCI листе категорије M22. Аутор је и два рада објављена на конференцијама од међународног значаја, као и два рада објављена на конференцијама од националног значаја.

Изјава о ауторству

Име и презиме аутора Стефан Костић

Број индекса 5036 / 2014

Изјављујем

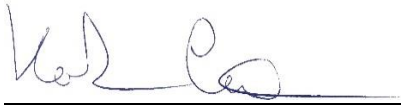
да је докторска дисертација под насловом

Предикција губитка корисника у мобилним телекомуникационим мрежама применом ненадгледаног машинског учења

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 01.02.2021.



Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Стефан Костић

Број индекса 5036 / 2014

Студијски програм Телекомуникације

Наслов рада Предикција губитка корисника у мобилним телекомуникационим мрежама применом ненадгледаног машинског учења

Ментор др Мирјана Симић-Пејовић, ванредни професор

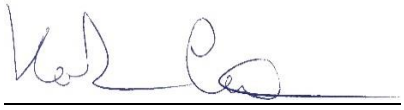
Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива 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 аутора

У Београду, 01.02.2021.



Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Предикција губитка корисника у мобилним телекомуникационим мрежама применом ненадгледаног машинског учења

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

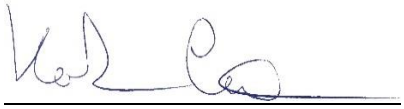
5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци. Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 01.02.2021.



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 прерада**. 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 прерада**. 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