UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

DIPLOMSKO DELO

PODATKOVNO RUDARJENJE NA PRIMERU

Ljubljana, maj 2010 ANŽE MARN

IZJAVA

Študent **Anže Marn** izjavljam, da sem avtor/ica tega diplomskega dela, ki sem ga napisal pod mentorstvom **dr. Jurija Jakliča**, in da dovolim njegovo objavo na fakultetnih spletnih straneh.

V Ljubljani, dne_____ Podpis: _____

KAZALO

KAZALO SLIK

KAZALO TABEL

Uvod

V današnjem času se zmožnosti zajemanja in generiranja podatkov hitro povečujejo. K temu prispeva razmah uporabe računalnikov, digitalnih medijev in informatizacija znanstvenega, privatnega in poslovnega sveta. Poplava informacij, ki so na voljo sodobnemu človeku je že zdavnaj presegla njegove zmožnosti analize in pregleda. Zato so se pojavile potrebe po avtomatiziranem odkrivanju znanja in temu izzivu skuša odgovoriti znanost podatkovnega rudarjenja.

Rast količine podatkov pa verjetno ni bila nikjer tako hitra kot v poslovnem svetu. Podjetja so imela na razpolago vedno bolj zmogljive računalnike, podatkovne baze in informacijske sisteme, ki so beležili vse njihove dejavnosti, kar je sčasoma pripeljalo do ogromnih podatkovnih skladišč. Ker so podjetja podvržena tržnim mehanizmom, vedno skušajo pridobiti dodatno konkurenčno prednost in vedno bolj spoznavajo, da je največja konkurenčna prednost znanje. Največ še neodkritega znanja pa leži implicitno skritega v njihovih podatkovnih skladiščih in podatkovno rudarjenje predstavlja orodje, kako do tega znanja lahko pridejo.

Ker je podatkovno rudarjenje razmeroma mlada veda, še ni trdno ustaljenih načinov in pravil, ki bi veljala za uspešno implementacijo podatkovnega rudarjenja na konkretnih primerih. Fayyad, Piatetsky-Shapiro in Smyth predlagajo, da se projektov podatkovnega rudarjenja lotimo procesno. Pri tem poudarjajo nekaj ključnih korakov, ki vplivajo na uspešnost, kot so razumevaje področja uporabe aplikacije, čiščenje in priprava podatkov (vključuje osnovne operacije kot so odstranjevanje šuma in izjem, izbira strategije za manjkajoče podatke, upoštevanje časovne komponente podatkov in obvladovanje težav s podatkovnimi bazami, kot na primer različni podatkovni tipi in mapiranje praznih vrednosti). Prav tako reduciranje podatkov (iskanje najbolj primernih dimenzij za zastopanje podatkov, ki morajo biti vezane na cilj in namen projekta in preoblikovanje podatkov na način, da zadržijo želene odvisnosti, a se njihova količina zmanjša), izbira funkcije podatkovnega rudarjenja in izbira algoritma za podatkovno rudarjenje. Ključna koraka sta tudi interpretacija in uporaba odkritega znanja (vključitev znanja v učinkovitost sistema, reagirati na podlagi dobljenega znanja ali pa samo zabeležiti ugotovitve in jih posredovati zainteresiranim skupinam) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, str 37-54).

Nekateri drugi avtorji navajajo druge dejavnike za uspeh. Brachman in drugi (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996, str. 42-48) pravijo, da je za uspeh projekta podatkovnega rudarjenja pomembno šolanje uporabnikov, razpoložljivost podatkov, razpoložljivost vzorcev, spremenljivi in časovno usmerjeni podatki, prostorski podatki, kompleksnost podatkov in razširljivost modela.

Spekter dejavnikov, ki vplivajo na uspešnost projekta podatkovnega rudarjenja, je zelo širok. Ker je analiza vseh preobsežna, se bom v diplomski nalogi omejil na področje izvedbe podatkovnega projekta v najožjem smislu.

Namen diplomske naloge je identifikacija problemov, s katerimi se srečujemo pri izvajanju tipičnega projekta podatkovnega rudarjenja. Prav tako želim določiti katere faze izvajanja projekta podatkovnega rudarjenja so ključne za kvalitetne rezultate.

V diplomskem delu bom skušal predstaviti posamezne tehnike podatkovnega rudarjenja in izpostaviti njihove dobre in slabe lastnosti ter določiti primerna področja njihove uporabe. V drugem delu bom skušal skozi proces praktične izvedbe procesa podatkovnega rudarjenja identificirati probleme, s katerimi se srečujemo v praksi in prikazati končni rezultat.

Diplomska naloga je razdeljena na dva dela. V prvem delu bom na splošno opisal, kaj je podatkovno rudarjenje, od kod izvira in kam gre njegov razvoj. Poskušal bom tudi razložiti njegovo mesto in uporabnost v poslovnem svetu. Nato sledi predstavitev najbolj razširjenih in popularnih tehnik, ki se uporabljajo v podatkovnem rudarjenju. Temu delu sledi podrobna prestavitev algoritma za razvrščanje v skupine, ki bo uporabljen v praktičnem delu diplomske naloge. Drugi del pa bo namenjen analizi praktičnega primera podatkovnega rudarjenja. Tu bom predstavil izbrano programsko orodje za podatkovno rudarjenje in proces priprave podatkov. Sklepni del je analiza dobljenih rezultatov in njihova analiza.

1 Podatkovno rudarjenje in odkrivanje znanja v podatkovnih bazah ter poslovna inteligenca

Digitalna revolucija je omogočila preprosto zajemanje, procesiranje, hranjenje, razmnoževanje in prenašanje informacij (Fayyad et al., 1996, str. 37-40). Razmah računalništva in z njim povezanih tehnologij ter njihova vedno pogostejša uporaba v vsakdanjem življenju sta privedla do ogromnih količin raznovrstnih podatkov, shranjenih v podatkovnih bazah. Stopnja, s katero se povečuje količina shranjenih podatkov, je izredna. Lahko naredimo analogijo med Moorovim zakonom o rasti procesorske moči in rastjo količine informacij. Napredek na področju obdelave podatkov in pojav novejših aplikacij na tem področju sta bili deloma mogoča zaradi razvoja polprevodnikov in posledično računalniške industrije. Moorov zakon pravi, da se število tranzistorjev podvoji vsakih 18 mesecev, in do sedaj je to držalo tudi v realnosti. To lahko povežemo z opazovanjem rasti podatkov in informacij. Če se količina podatkov na svetu podvoji vsakih 20 mesecev (Auerbach, 2004), se velikost in število podatkovnih baz dvigata z enako hitrostjo. Odkrivanje znanja ali koristnih informacij v tej ogromni količini podatkov pa je velik izziv. Podatkovno rudarjenje je poizkus obvladovanja te eksponentne rasti informacij v ogromnih količinah podatkov.

Nekatera področja, ki se srečujejo s eksponentno rastjo zajetih podatkov, so: proizvodna industrija, trženje, bančništvo, telekomunikacijska podjetja in njihova omrežja, digitalne knjižnice, arhiviranje slik, spletni foto albumi, medicina z diagnostiko, znanstveni poizkusi in astronomija, raziskovanje vesolja, biologija, fizika, internet in še mnogo drugih.

Dva praktična primera, kjer se generira ogromna količina zajetih podatkov sta:

- Evropski VLBI (angl. *Very Long Baseline Interferometry*) (Goddard Space Flight Cente, 2009) ima 16 teleskopov, od katerih vsak generira 1 Gb/s (gigabit na sekundo) astronomskih podatkov. Opazovano obdobje pa je dolgo 25 dni. Zbrana količina podatkov je zares »astronomska«.
- Podatkovna baza organizacije Internet Archive (Internet Archive, 2009), ki arhivira internetne strani, obsega že več kot 2 petabajta podatkov (2.000 TB). Vsak mesec pa zraste za nadaljnjih 20 TB!

Profesorja Peter Lyman in Hal R. Varian (2003) z univerze Berkeley v svoji raziskavi ocenjujeta, da je bilo v letu 2002 ustvarjenih 5 exabajtov (5 milijonov TB) podatkov. V raziskavi ocenjujeta tudi, da se stopnja rasti povečuje za 30 odstotkov na leto. Drugi ocenjujejo, da je rast še hitrejša.

1.1 Podatkovno rudarjenje

Surovi podatki so redko koristni sami po sebi oziroma običajno ne predstavljajo dodane vrednosti. Prava vrednost zbranih podatkov je odvisna od naše sposobnosti, da iz njih izluščimo informacije, ki nam služijo za *podporo pri odločanju*, in da prepoznamo *trende in vzorce*, skrite v teh podatkih. V večini primerov se te podatke obdeluje ročno s pomočjo strokovnjakov, ki uporabljajo statistične metode za izdelavo povzetkov in poročil. Vendar pa je ta pristop vedno bolj neprimeren zaradi eksponentne rasti podatkovnih baz, ki presegajo 200 GB (Winter Corporation, 2005). Zahtevnost raziskovanja tako velikih podatkovnih baz presega človeške zmogljivosti in zato rabimo računalniško podprte procese, ki ta proces avtomatizirajo.

Vse to je spodbudilo razvoj inteligentnih podatkovno analitičnih metodologij, katerih cilj je odkrivanje znanja v podatkovnih bazah. Podatkovno rudarjenje je ne trivialen proces iskanja veljavnih, neobičajnih, potencialno uporabnih in razumljivih vzorcev v podatkih (Fayyad et al., 1996, str. 41). Obstajajo še druge definicije (Berry & Linoff, 2000, str. 5, 7):

- Podatkovno rudarjenje je proces raziskovanja velike količine podatkov z namenom odkriti uporabne vzorce in pravila.
- Podatkovno rudarjenje je proces odkrivanja zanimivega znanja iz velike količine podatkov, shranjenih bodisi v podatkovnih bazah, podatkovnih skladiščih (angl. data warehouse) ali v drugih hranilcih informacij. Preprosto povedano je podatkovno rudarjenje pridobivanje znanja iz velike količine podatkov.

Razlog za razmah podatkovnega rudarjenja v informacijski tehnologiji in tudi v širši javnosti je v dejstvu, da si želimo čim hitrejšega pridobivanja znanja in koristnih informacij iz vedno večje količine podatkov v našem okolju (Han & Kamber, 2006, str. 2-4).

1.2 Odkrivanje znanja v podatkovnih bazah

Podatkovno rudarjenje je jedro procesa odkrivanja znanja v podatkovnih bazah (angl. *Knowledge Discovery from Data*, v nadaljevanju KDD). KDD se nanaša na celotni proces odkrivanja znanja v podatkih.

KDD in je netrivialen proces odkrivanja implicitnega, doslej neznanega in potencialno uporabnega znanja iz podatkov. Rudarjenje podatkov je ena od faz odkrivanja znanja v podatkovnih bazah, v kateri dejansko pride do odkrivanja znanja (Han & Kamber, 2006, str. 6-7). KDD in podatkovno rudarjenje (angl. *Data mining*, v nadaljevanju DM) sta interdisciplinarna. Presegata omejitve tradicionalnih statističnih metod, vendar le-te vključujeta v svoj okvir. Cilj vsega tega je pridobivanje novega znanja, ki ga je mogoče uporabljati za boljše odločanje, klasifikacijo ali predvidevanje (Han & Kamber, 2006, str. 2).

Shapiro je leta 1989 začel uporabljati termin KDD. Rekel je, da je termin postal priljubljen v krogih umetne inteligence (angl. *artificial intelligence*) in strojnega učenja (angl. *machine learning*). Vendar so bili strokovnjaki s področja podatkovnih baz v boljših odnosih s poslovno in novinarsko skupnostjo in zato se je uveljavil termin podatkovno rudarjenje. Izraz podatkovno rudarjenje je starejši kot izraz KDD. Skovali so ga v združenju statističnih podatkovnih analitikov (angl. *Statistics Data Analysis Association*) (Shapiro, 2000, str. 60).

Podatkovno rudarjenje je prevod angleškega izraza »data mining« in je pravzaprav napačno ime. Ko rudarimo za zlatom v skali ali pesku, to v angleščini poimenujemo »gold mining« ali po naše »rudarjenje za zlatom« in ne »peskovno rudarjenje« ali pa »skalno rudarjenje«. Pravilen izraz bi bil torej »rudarjenje znanja v podatkih«, vendar je dolg. Krajši izraz bi bil lahko »rudarjenje znanja« (angl. *knowlegde mining*), vendar v tem izrazu ni zadosti poudarjeno, da se to rudarjenje aplicira na velike količine podatkov. Verjetno se je zato to napačno poimenovanje »podatkovno rudarjenje« obdržalo (Han & Kamber, 2006, str. 5).

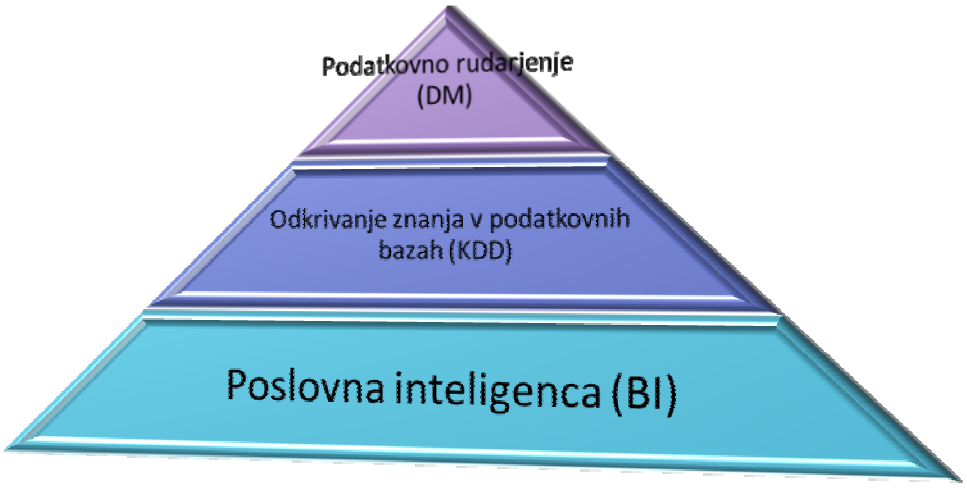
1.3 Poslovna inteligenca (BI)

Poslovno inteligenco (angl. *Business Intelligence*, v nadaljevanju BI) je organizacija Gartner (http://www.gartner.com) definirala kot »širok nabor aplikacij in tehnologij za zbiranje, hranjenje, analiziranje, izmenjevanje in omogočanje dostopa do podatkov z namenom pomagati poslovnim uporabnikom priti do boljše poslovne odločitve«.

Če se osredotočimo na zadnji del te definicije, pridemo do bistva, kaj je pravzaprav BI: tehnologija, ki pomaga poslovnemu uporabniku sprejemati boljše oziroma bolj informirane poslovne določitve.

Poslovna inteligenca ima za svoj cilj, da ugotovi, kako pravo informacijo ob pravem času ponuditi pravi osebi. Poslovna inteligenca zaobjema vse zmožnosti organizacije za pretvorbo surovih podatkov ter koristne in zanesljive informacije, ki so na voljo celotni organizaciji (Miller, Brautigam, & Gerlach, 2006, str. 3).

Slika 1: Razmerje med BI, KDD in DM s tehnološkega vidika



Vir: Lukawiecki, Technet Spotlight On Demand Video: Introduction to Data Mining, 2008.

Na sliki 1 vidimo, kakšno je razmerje med BI, KDD in podatkovnim rudarjenjem z vidika tehnologije. BI uporablja število tehnik, ki so del KDD, še posebej pa DM kot najnaprednejšega in najzanimivejšega področja KDD.

2 Podatkovno rudarjenje in področje napovedne analitike

Analize v podjetjih se močno razlikujejo po kompleksnosti. Najbolj preprosta in tudi najbolj razširjena oblika poročanja so že vnaprej določena in strukturirana poročila. So preprosta za izdelavo ali pa se generirajo celo avtomatsko in najbolj spominjajo na predstavitve. Od uporabnika ne zahtevajo veliko znanja s področja informatike. Njihova največja pomanjkljivost je, da niso prilagodljiva.

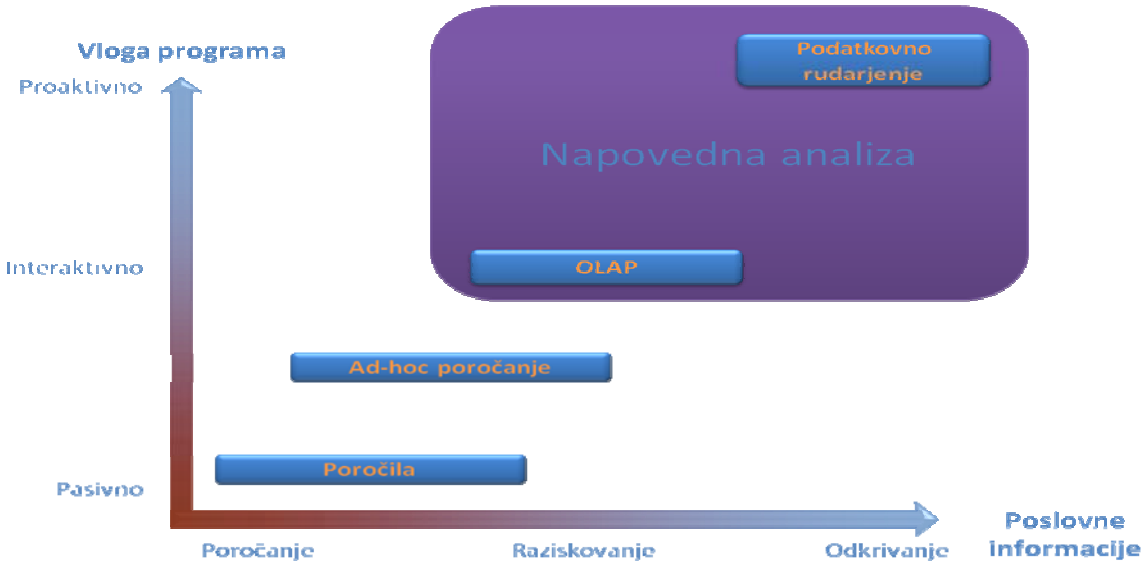
Problem prilagodljivosti delno odpravljajo »ad hoc« poročila, ki predstavljajo kompleksnejšo in bolj interaktivno poizvedovanje. Ta poročila zahtevajo bolj podkovanega uporabnika.

Naslednji nivo analize je analiza z orodji OLAP (angl. *OnLine Analytical Processing*). Ta tehnika je najbolj interaktivna in raziskovalna od treh naštetih tehnik. Zahteva veščega uporabnika in veliko specifičnega znanja.

Podatkovno rudarjenje je z vidika analize najbolj proaktivna in raziskovalna tehnika analize, kar jih poznamo. Zahteva visoko usposobljene uporabnike.

OLAP in DM definirata meje napovedne analize (angl. *Predictive Analysis*), ki je trenutno eno najbolj »vročih« programskih razvojnih področij. Podjetja se trudijo izdelati orodja, ki bi se umestila med OLAP in DM (Lukawiecki, 2008a).

Slika 2: Tehnike analize glede na kompleksnost



Vir: Lukawiecki, Technet Spotlight On Demand Video: Introduction to Data Mining, 2008

2.1 Uporaba podatkovnega rudarjenja

Podatkovno rudarjenje se je v kratkem obdobju razvilo v samostojno vedo, ki jo je možno uporabljati na najrazličnejših področjih. Digitalizacija in informatizacija vseh področij našega življenja pa pomenita, da se nabor le-teh samo še širi. V razdelku 2.1 bom na hitro predstavil različna področja in praktično uporabo podatkovnega rudarjenja. Posebno pozornost bom namenil uporabi podatkovnega rudarjenja v gospodarstvu.

Na področju proizvodnega procesa se pogosto srečujemo s poplavo podatkov iz mnogih merilnih naprav, ki so vgrajene v proizvodni proces. Za odkrivanje povezave med parametri v proizvodnem procesu in želeno lastnostjo produkta kot na primer kakovost jekla se uporablja podatkovno rudarjenje.

V medicini se podatkovno rudarjenje uporablja za napovedovanje razvoja bolezni in diagnoze.

Razcvet digitalnega zajema slike je povzročil poplavo fotografij, ki so shranjene na računalnikih in svetovnemu spletu. Za potrebe iskanja, klasificiranja, prepoznavanja in razvrščanja slik v skupine se uporabljajo tehnike podatkovnega rudarjenja.

Izboljševanje in odkrivanje novih učinkovin je glavna raziskovalna dejavnost farmacevtske industrije. Podatkovno rudarjenje se uporablja za predvidevanje lastnosti učinkovin glede na njihovo kemično strukturo, kar pospeši in poceni raziskovalni proces.

Izdelovalci netrivialnih računalniških iger kot so šah, go in podobne se zanašajo na tehnike podatkovnega rudarjenja, da izenačijo ali včasih celo presežejo sposobnosti človeškega igralca. Leta 1997 je tak računalnik imenovan Deep Blue v šahu premagal šahovskega velemojstra Kasparova (Kononenko & Kukar, 2007, str. 24 - 29).

Sledi še sedem najbolj tipičnih rab podatkovnega v podjetjih:

Iskanje dobičkonosnih strank – DM omogoča podjetjem, da identificirajo stranko, ki je zanje donosna, in tudi odkrivajo razloge, zakaj je tako.

Razumevanje potreb strank/zaposlenih – DM upodabljamo za razumevanje vsake entitete, ki izraža kakršno koli vedenje. To lahko pomeni preučevanje spletnega obiskovalca in tega, kako se »sprehaja« po spletni strani, pa tudi ugotavljanje, zakaj nikoli ne odpre določenega dela strani, ki je sicer zanimiva.

Management prehajanja strank – To je zelo specifična uporaba DM. Gre za poskus identifikacije uporabnika, ki je ravno pred tem, da bo zamenjal ponudnika storitve, in pozneje seveda tudi za preprečitev te zamenjave. Odkrivanje takih uporabnikov je pomembno na današnjih nasičenih razvitih trgih, na katerih praktično ni novih mobilnih uporabnikov. So samo tisti, ki menjajo operaterja.

Napovedovanje prodaje in inventarja – Predvidevanje prodaje in inventarja je ena najstarejših aplikacij napovedne analize. Tu se trudimo predvidevati, koliko in kaj se bo prodalo, ali bo v skladišču zadosti prostora, kolikšen bo moj prihodek ali pa dolg.

Izdelava učinkovitih trženjskih akcij – Verjetno nobena organizacija nima dovolj virov, da bi s svojimi trženjskimi akcijami ciljala kar na vse ljudi. Končni cilj uporabe napovedne analize je, da bi se na akcijo odzvalo ker se da veliko ljudi.

Zaznava in preprečevanje prevar – je eno najzahtevnejših področij podatkovnega rudarjenja. Omogoča nam odkrivanje nezakonitih transakcij ali transakcij, ki bodo vodile v kriminalna dejanja. To področje je še na svojem razvojnem začetku in še ni izpolnilo vseh možnosti. Končni cilj pa je, da bi bilo možno v živo gledati ustreznost transakcije.

Popravljanje podatkov med ETL – ETL je okrajšava ukazov Extract, Transform in Load, ki se uporabljajo v podatkovnih skladiščih. Ko pretakamo podatke iz različnih sistemov in z njimi polnimo podatkovno skladišče, se pogosto pojavijo zapisi, ki jim manjka kak atribut, npr. spol. Tehnike DM nam omogočajo, da v realnem času z veliko verjetnostjo predvidimo manjkajoči atribut in ga zapišemo (Lukawiecki, 2008a).

Slika 3: Sedem tipičnih nalog, za katere lahko uporabljamo podatkovno rudarjenje

Vir: Lukawiecki, Technet Spotlight On Demand Video: Introduction to Data Mining, 2008

3 Naloge podatkovnega rudarjenja

Tan, Steinbach in Kumar (2006, str. 7–11) delijo podatkovno rudarjenje na dve glavni kategoriji:

- **Napovedovalne naloge.** Tukaj je cilj napovedati vrednost izbranega atributa na osnovi drugih atributov, katerih vrednosti so nam znane. Slednje atribute običajno imenujemo neodvisne spremenljivke. Odvisna spremenljivka pa je atribut, ki ga iščemo.
- **Opisovalne naloge.** Ta kategorija podatkovnega rudarjenja skuša predvsem opisati oz. najti vzorce v množicah podatkov. Opisovalne tehnike po navadi uporabljamo za raziskovanje podatkov in dostikrat jih kombiniramo z drugimi tehnikami, da dobljene rezultate sploh lahko razložimo. Dober primer za to je razvrščanje v skupine (angl. *clustering*), ki ga pogosto uporabimo kot prvi korak pri raziskovanju podatkov. Če imamo nalogo, da moramo raziskati veliko bazo podatkov, jo lahko z uporabo razvrščanja v skupine razdelimo na homogene skupine, ki jih je nato lažje analizirati.

Znotraj teh dveh glavnih skupin pa so štiri glavne naloge podatkovnega rudarjenja, ki so nanizana v nadaljevanju (Tan, Steinbach, & Kumar, 2006, str. 7–11).

3.1 Napovedno modeliranje

Napovedno modeliranje (angl. *predictive modeling*) se nanaša na izdelavo modela, ki napove vrednost napovedne spremenljivke kot funkcijo neodvisnih spremenljivk. Ločimo med dvema tipoma napovednega modeliranja.

Klasifikacija (angl. *classification*) je proces iskanja modela oz. funkcije, ki lahko razločuje med razredi podatkov (angl. *data classes*) z namenom, da vanje razvrsti objekte brez razreda. Dobljeni model je rezultat analize množice podatkov za urjenje modela (angl. *training data set*), v katerem so objekti z znanim razredom.

Dobljeni model je lahko predstavljen v različnih oblikah, kot so (Han & Kamber, 2006, str. 24):

- pravila za klasificiranje (angl. *classification rules*) v obliki ČE – POTEM (angl. *IF-THEN*),
- odločitvena drevesa (angl. *decision tree*),
- matematične formule,
- naivno Bayesian klasifikacijo (angl. *naive Bayesian classification*),
- metoda podpornih vektorjev (angl. *support vector machines*, v nadaljevanju SVM),
- najbližji sosed (angl. *k-nearest neighbor*).

Za klasifikacijo je značilno, da jo uporabljamo za napovedovanje diskretnih spremenljivk. Primer je napovedovanje, ali bo spletni uporabnik opravil spletni nakup ali ne. Napovedna spremenljivka je v tem primeru lahko samo v dveh stanjih (0 ali 1) in zato je to klasifikacija.

Napovedovanje (angl. *prediction*) pa lahko prikažemo na primeru napovedi vrednosti delnic, pri kateri je napovedna spremenljivka zvezna. Tehnike uporabljene za napovedovanje, so:

- linearna regresija (angl. *linear regression*),
- nelinearna regresija (angl. *nonlinear regression*).
- nevronske mreže (angl. *neural network*)

Cilj obeh vrst napovednega modeliranja je izdelati model, ki ima najmanjšo napovedno napako, se pravi najmanjšo razliko med napovedanimi in dejanskimi vrednostmi. Napovedno modeliranje lahko med drugim uporabljamo za predvidevanje, ali se bo stranka odzvala na trženjsko akcijo, za predvidevanje motenj v Zemljinem ekosistemu ali pa za presojanje, katero bolezen ima pacient, na podlagi izvidov.

3.2 Analiza asociacij

Analiza asociacij (angl. *association analysis*) se uporablja za odkrivanje in opisovanje močnih asociacij ali povezav v podatkih. Odkrite povezave med podatki so tipično podane v odliki implicitnih pravil. Ker je povezav med obravnavanimi podatki lahko zelo veliko, nas zanimajo samo tiste, ki imajo največjo podporo. Primeri uporabe asociacij so odkrivanje genov s podobnimi učinki, odkrivanje povezanih spletnih strani ali pa za razumevanje povezav med različnimi elementi zemeljskega ekosistema (Tan et al., 2006, str. 9). Najbolj znan primer uporabe asociacij pa je analiza nakupne košarice (angl. *market basket analysis*), ki ima za cilj povezati različne izdelke, ki so bili kupljeni skupaj.

3.3 Analiza skupin

Analiza skupin (angl. *cluster analysis*) v primerjavi z napovednim modeliranjem razvršča objekte v skupine brez vnaprej poznanih razredov ali njihovega števila. Je primer neusmerjenega rudarjenja podatkov. Skupine se tvorijo na osnovi pravila, da naj bodo skupine elementov navznoter čim bolj homogene, navzven pa čim bolj heterogene (Tan et al., 2006, str. 24).

Tehnike analize skupin so bile uporabljene za širok spekter raziskovalnih problemov. Hartigan (Hartigan, 1975) nudi analizo objavljenih raziskav na področju analize skupin. Na področju medicine se uporablja za razvrščanje bolezni, zdravil in simptomov iz katerih nato naredijo zelo uporabne taksonomije. Na področju psihiatrije se tehnike uporabljajo za postavljanje diagnoze za shizofrenijo, paranojo in podobno. Na splošno je analiza skupin zelo uporabna tehnika, ko imamo »goro« podatkov in bi jih radi razčlenili v manjše smiselne enote.

3.4 Zaznavanje anomalij

Odkrivanje zapisov, ki imajo občutno drugačne lastnosti od večine, se imenuje zaznavanje anomalij (angl. *anomaly detection*). Ko iščemo anomalije, se moramo potruditi, da normalnih zapisov ne identificiramo kot anomalije. Dober sistem za iskanje anomalij mora imeti visoko stopnjo zaznave in nizko stopnjo napačne identifikacije zapisov kot anomalij. To je zlasti pomembno, ko se tak sistem uporablja za preprečevanje zlorab kreditnih kartic. Če sistem zlorabe ne zazna, nastane velika škoda, če pa kot prevaro zazna legitimno transakcijo, pa uporabniku povzroči veliko preglavic (Witten & Frank, 2005, str. 314–315).

4 Pregled tehnik podatkovnega rudarjenja

V četrtem poglavju sledi splošen pregled tehnik podatkovnega rudarjenja. Opisi posameznih tehnik je splošen in zato ne vključuje podrobnih opisov algoritmov. Izjema je razvrščanja v skupine, ki je uporabljena tehnika v praktičnem delu diplomskega dela in je podrobno predstavljen na koncu poglavja.

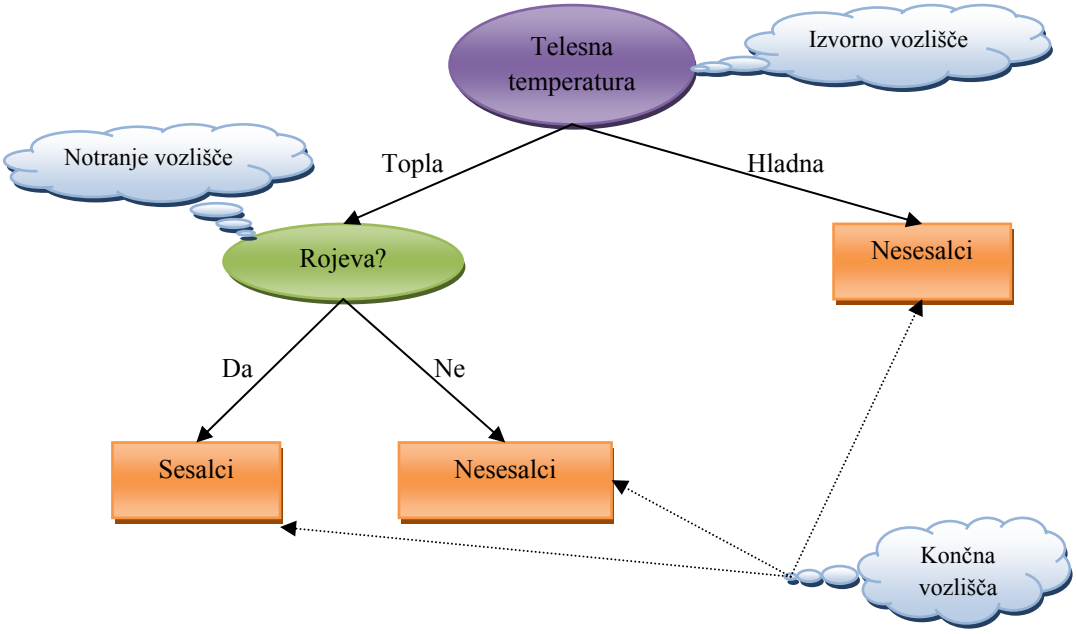
4.1 Odločitvena drevesa

Odločitvena drevesa so zelo široko uporabljana orodja, ki se jih uporablja tako za klasificiranje kot tudi za predvidevanje (Berry & Linoff, 1997, str. 243).

Odločitvena drevesa si lahko predstavljamo kot niz zaporednih vprašanj, ki imajo za cilj klasificirati neki zapis. Vsakemu vprašanju sledi novo vprašanje, ki je odvisno od predhodnega odgovora. Ta proces traja, dokler ne ugotovimo, v katero skupino spada konkretni zapis.

Niz teh vprašanj in odgovorov lahko grafično prikažemo kot narobe obrnjeno drevo – odločitveno drevo. Slika 4 prikazuje preprosto odločitveno drevo za klasifikacijo sesalcev.

Slika 4: Preprosto odločitveno drevo za klasifikacijo sesalcev



Vir: P. N. Tan, M. Steinbach & V. Kumar, Introduction to Data Mining (2nd ed.), 2006, str. 151, slika 4.4.

Sestavljajo ga (Tan et al., 2006, str. 150):

- izvorno vozlišče (angl. *root node*) – nima nobenega vhodnega signala in nima nobenega izhodnega signala ali ima več izhodnih signalov;
- notranje vozlišče (angl. *internal node*) – ima en vhodni signal in dva ali več izhodnih signalov;
- končno vozlišče (angl. *terminal node*) – ima točno en vhodni signal in nobenega izhodnega.

Zapis vstopi v odločitveno drevo pri izvornem vozlišču. Tu se naredi test, ki določi, v katero notranje vozlišče bo zapis vstopil. Za ta test se uporabljajo različni algoritmi, a princip je vedno enak: izbrati test, ki najbolje diskriminira med ciljnimi razredi. Ta proces se ponavlja, dokler zapis ne pride do končnega vozlišča. Vsaka pot od izvornega do končnega vozlišča je edinstvena in predstavlja pravilo, ki se nato uporablja za klasifikacijo. Vsi zapisi, ki končajo v istem končnem vozlišču, spadajo v isti razred (Berry & Linoff, 1997, str. 244).

4.2 Bayesov klasifikator

Bayesov klasifikator spada med statistične klasifikatorje (Han & Kamber, 2006, str. 310). Ti delujejo tako, da predvidevajo, s kolikšno verjetnostjo posamezni zapis pripada določenemu razredu. Bayesov klasifikator za svojo osnovo uporablja Bayesov teorem.

Bayesov teorem je dobil ime po svojem izumitelju, Thomasu Bayesu. Bayes je bil britanski matematik in prezbiterijanski menih, ki je živel v 18. stoletju v Londonu. Svoje ugotovitve o verjetnosti je zapisal v delu »Essay Towards Solving a Problem in the Doctrine of Chances«, ki so ga objavili dve leti po njegovi smrti (International Society for Bayesian Analysis, 2009).

V praksi se uporabljata dve metodi (Han & Kamber, 2006, str. 310):

- *Naivna verzija Bayesovega klasifikatorja* (angl. *naive Bayesian classification*) domneva, da je vpliv atributov na razred zapisa neodvisen od drugih vrednosti atributa istega zapisa. Ta domneva je postavljena zato, da poenostavi količino potrebnega preračunavanja. Zaradi te domneve tudi rečemo, da je ta klasifikator naiven.
- *Bayesovske mreže* (angl. *Bayesian belief networks*) so grafični modeli, ki v nasprotju s predhodno metodo upoštevajo tudi odvisnosti med spremenljivkami.

4.3 Pravila za klasificiranje

Pravila za klasificiranje (angl. *rule-based classification*) je tehnika, ki temelji na pravilih tipa ČE – POTEM. ČE del izraža pogoje, ki morajo biti izpolnjeni (npr. starost = 29, število otrok = 4), POTEM del pa predstavlja »posledice« tega pravila (npr. poročen = da). Primer tega pravila če – potem bi torej bil:

ČE starost = 29 IN število otrok = 4 POTEM poročen = da

Han in Kamber (2006, str. 321–333) opišeta dva načina pridobivanja pravil za klasificiranje:

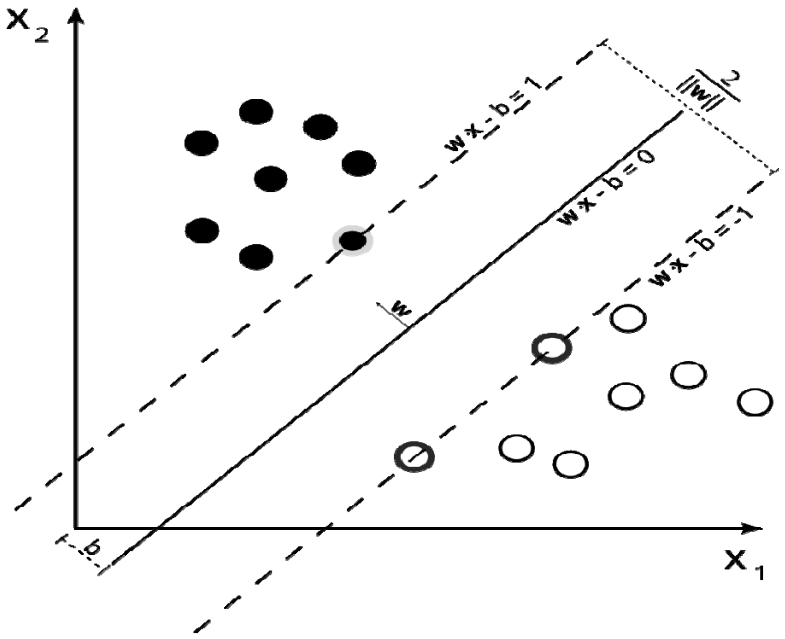
- *Pravilo, dobljeno iz odločitvenih dreves,* deluje tako, da najprej povzame pravila iz odločitvenih dreves, nato pa pravila, ki ne izboljšajo točnosti pravila, odstrani. To izvede z uporabo algoritma C4.5.
- *Pravila, dobljena z zaporednim pokrivajočim algoritmom* (angl. *sequential covering algorithm*). Algoritem deluje tako, da se ustvari pravilo (v smeri splošno -> specifično) za zapise, ki mu ustrezajo. Nato se te vrstice odstranijo iz podatkovne množice in proces se ponavlja, dokler ne zmanjka zapisov. Najpopularnejši algoritmi so CN2, AQ in RIPPER.

4.4 Metoda podpornih vektorjev

Prva raziskava na temo SVM je bila objavljena leta 1992, njene zasnove pa segajo v leto 1960. SVM je nova metoda, ki je v zadnjem času deležna zelo veliko pozornosti. Strokovna javnost se nad to metodo navdušuje predvsem zaradi njene sposobnosti formiranja zapletenih nelinearnih in večdimenzionalnih ravnin, ki odlično razmejujejo skupine objektov. Še ena njihova prednost je, da dajejo zgoščen opis dobljenega modela. Velika pomanjkljivost metode SVM pa je, da je za njihovo urjenje potrebnega zelo veliko časa oz. zahtevajo veliko procesorske moči.

SVM je algoritem nadzorovanega učenja, ki se uporablja tako za klasifikacijo kot tudi za predvidevanje. SVM deluje tako, da ustvari hiperravnina (angl. *hyperplane*) ali več hiperravnin v večdimenzionalnem prostoru, te pa razmejujejo posamezne skupine. Ravnina, ki razmejuje dve skupini, ju razmeji tako, da je med njima kar največja razdalja (W). Na ta način se minimizira preveliko prilagajanje (angl. *overfitting*). Koncept je prikazan na spodnji sliki v dvodimenzionalnem prostoru (Han & Kamber, 2006, str. 337-342).

Slika 5: Maksimalna razmejitev prostora s tehniko SVM



Vir: J. Han & M. Kamber, Data Mining Concepts and Techniques, 2006, str. 340.

4.5 Nevronske mreže

Nevronske mreže (angl. *neural networks*) (Berry & Linoff, 1997, str. 287) sta leta 1943 idejno zasnovala nevropsiholog Warren McCulloch in logik Walter Pitts, ki sta poskušala razložiti, kako deluje nevron v človeškem telesu. Čeprav je bil njun namen razložiti delovanje človekovih možganov, se je izkazalo, da ta model ponuja nov pristop k reševanju problemov zunaj nevrobiologije.

V petdesetih letih so razvili perceptrone, katerih zasnova je temeljila na delu McCullocha in Pittsa. Z njimi so imeli nekaj uspehov v laboratoriju, a na splošno niso izpolnili pričakovanj. Razlogi za to so bili zelo majhna zmogljivost tedanjih računalnikov in pa tudi teoretične pomanjkljivosti teh preprostih nevronskih mrež. Te pomanjkljivosti je leta 1982 odpravil John Hopfield z novim načinom učenja nevronskih mrež. Poimenoval ga je »backpropagation«, kar bi lahko po slovensko prevedli kot »vzvratno učenje« oz. »širjenje nazaj«.

Tan s sodelavci (2006, str. 254) »učenje nazaj« razloži kot način posodabljanja uteži na višjih nivojih pred posodabljanjem uteži na nižjih nivojih. Ta postopek omogoča, da uporabimo napako nevrona na višjem nivoju za oceno napake nevrona na nižjem nivoju.

Odkritje »učenja nazaj« je sprožilo renesanso v raziskovanju nevronskih mrež. V osemdesetih letih se je njihova uporaba preselila iz laboratorijev na komercialno področje, na katerem so bile uporabljene domala povsod. Nekateri primeri uporabe so sprotna zaznava zlorab kreditnih kartic, prepoznavanje številk, napisanih na čekih, ocenjevanje vrednosti nepremičnin, odkrivanje bolezni pacientov in podobno.

Prednosti nevronskih mrež so (Han & Kamber, 2006, str. 327-337):

- Z njimi lahko rešujemo širok spekter problemov. So zelo prilagodljive.

- Dobro se obnesejo tudi na zapletenih področjih.
- Lahko uporabljajo zvezne in diskretne spremenljivke.
- So zelo razširjene in lahko dostopne. Najdemo jih v skoraj vseh komercialnih aplikacijah za podatkovno rudarjenje.

Nevronske mreže so zelo prilagodljive. Lahko jih uporabimo za napovedovanje zveznih spremenljivk – v tem primeru opravljajo napovedovanje, lahko pa jih uporabimo za napovedovanje diskretnih spremenljivk – v tem primeru opravljajo klasifikacijo. Če malo preuredimo nevrone v mreži, pa nevronske mreže lahko uporabimo za odkrivanje skupin v podatkih.

Nevronske mreže se dobro obnesejo z zveznimi in diskretnimi spremenljivkami kot vstopnimi podatki ali pa kot rezultatom. Veliko dela pa je treba vložiti v pripravo podatkov. Nevronske mreže namreč delujejo samo s števili od 0 do 1, zato moramo vse spremenljivke zreducirati na to obliko.

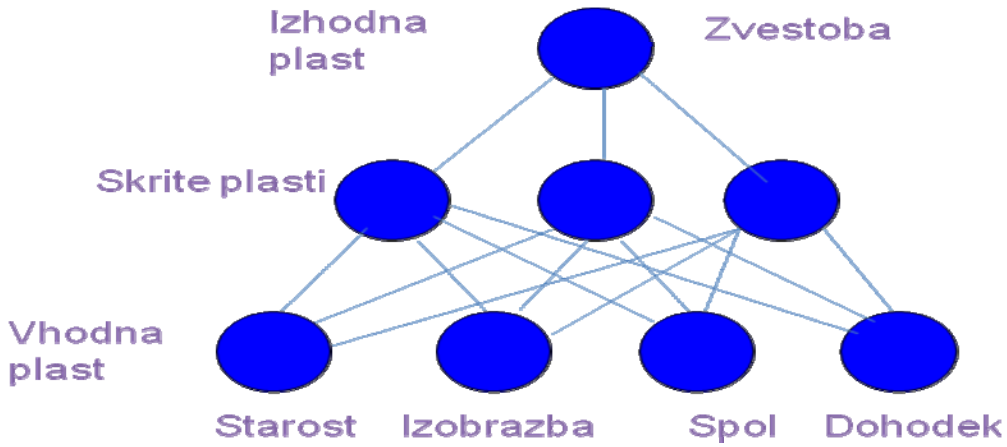
Pomanjkljivosti nevronskih mrež so (Han & Kamber, 2006, str. 327-337)

- Vhodni podatki morajo biti v obliki številke v razponu od 0 do 1.
- Ne moremo razložiti, zakaj so rezultati takšni, kot so.
- Obstaja možnost, da ne najdejo optimalne rešitve.

Nevronske mreže za svoje delovanje potrebujejo vhodne podatke v obliki številke v razponu od 0 do 1. To seveda zahteva veliko dela, potrebnega za transformacijo podatkov. Prav tako je ključnega pomena izbira tehnike za transformacijo podatkov in lahko znatno vpliva na rezultat rudarjenja.

Delovanje nevronskih mrež si lahko predstavljamo kot »črno skrinjo«, v katero damo vhodne podatke in dobimo rezultat. Ne moremo pa videti, kako smo prišli do tega rezultata, in tudi ne moremo izvleči previl, kot jih na premier lahko pri asociativnih pravilih ali odločitvenih drevesih. Na sliki 6 je ilustracija nevronske mreže z vhodnimi, skritimi in izhodnimi plastmi.

Slika 6: Nevronska mreža, ki napoveduje zvestobo



Vir:R. Lukawiecki, Technet Spotlight On Demand Video: Using Data Mining in Your IT Systems (Part 1), 2008.

Vidimo lahko, da poznamo vhodno plast, ki jo določimo sami, poznamo izhodno plast, ki je cilj celotne operacije, ne poznamo pa skrite plasti in zato iz nje ne moremo izluščiti nobenega pravila.

Nevronske mreže se dobro obnesejo za večino napovednih in klasifikacijskih nalog, pri katerih so rezultati pomembnejši od razumevanja, kako model deluje. Nevronske mreže dobro delujejo tudi pri razvrščanju v skupine. Za to lahko uporabimo tehniko SOM (angl. *Self-Organizing Map*). Pri uporabi te tehnike sicer dobimo posamezne skupine, vendar ne vemo, zakaj so si elementi znotraj skupin podobni, in jih moramo analizirati s pomočjo povprečnih vrednosti elementov v posameznih skupinah.

Edini primer, ko nevronske mreže ne dajo dobrega rezultata, je, ko imamo elemente z več sto ali tisoč atributi. Zaradi velikega števila atributov nevronske mreže težko najdejo vzorec, pa tudi čas, potreben za urjenje modela, se izrazito poveča. Eden od načinov, kako rešimo ta problem, je, da nevronske mreže kombiniramo z uporabo odločitvenih dreves, ki se obnesejo pri prepoznavanju pomembnih atributov.

4.6 Najbližji sosed

Tehnika najbližjega soseda (angl. *nearest neighbor*) ali krajše k-NN je metoda klasifikacije zapisov glede na relativno bližino zapisa v množici podatkov, uporabljenih za urjenje (angl. *training set*).

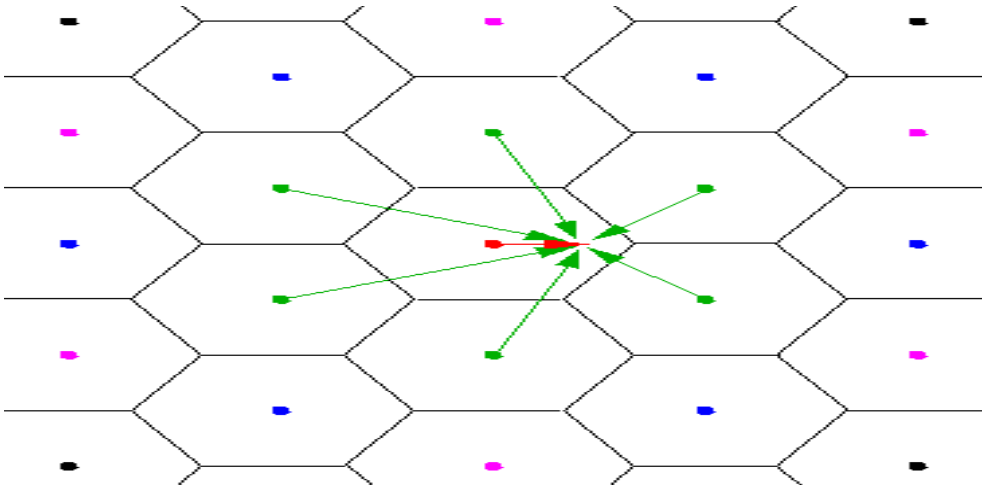
Če je tehnika odločitvenih dreves poznana kot »eager learners«, kar v dobesednem prevodu pomeni »vedoželjni učenci«, je metoda najbližjega soseda poznana kot »lazy learners« ali v prevodu »počasni učenci«. Princip delovanja »željnih učencev« je, da ustvarijo model, ki poveže vstopne atribute zapisa z ustreznim izhodnim razredom takoj, ko je to mogoče oz. je podatek za urjenje modela dostopen.

Primer tehnike, ki uporablja metodo »počasnih učencev«, je *rote-klasifikator* (angl. *rote classifier*). Deluje tako, da shrani celotno množico podatkov za urjenje (angl. *training set*). Nato samo v primeru, da so atributi vstopnega podatka identični z enim izmed podatkov, s katerimi smo urili model, izvede razvrščanje v skupine (Tan et al., 2006, str. 223).

Problem tehnike *rote-klasifikatorja* je, da nekateri zapisi ne bodo razvrščeni v skupine, ker ne ustrezajo nobenemu zapisu iz množice podatkov za urjenje (angl. *training set*). Tehnika k-NN ta problem reši tako, da »sprosti« kriterije in v skupine razvrsti tudi tiste zapise, ki so samo podobni zapisom, uporabljenim za izdelavo podatkovnega modela, ali z drugimi besedami, so *najbližji sosedje* teh zapisov.

Tehnika klasifikacije z najbližjim sosedom je eden najbolj preprostih algoritmov za strojno učenje (angl. *machine learning*). Klasifikacija zapisa ali objekta se izvede glede na zapise, ki so mu najbliže. Če ima zapis na primer za najbližje sosede dva zapisa, ki spadata v skupino A, in enega, ki sodi v B, potem ga bo algoritem razvrstil v skupino A. Pri tem sistemu »glasovanja« lahko uporabimo tudi sistem uteži, ki so odraz oddaljenosti sosedov od obravnavanega zapisa. Večja ko je oddaljenost, manjši pomen ima sosed pri glasovanju. To je prikazano na sliki 7.

Slika 7: Merjenje oddaljenosti zapisa od sosedov



Vir: College of Saint Benedict and Saint John's University, Tools for Science, 2009.

Tehnika najbližjega soseda ima to dobro lastnost, da lahko prostor razmeji bolj prosto kot pa, recimo, odločitvena drevesa, ki prostor razmejijo vzporedno z osmi prostora (Tan et al., 2006, str. 227).

Ena izmed največjih pomanjkljivosti tehnike k-NN je, da je procesorsko zahtevna. To izhaja iz dejstva, da mora algoritem za vsak zapis, ki ga želimo razvrstiti v skupine, izračunati oddaljenost od njegovih sosedov. Količina procesiranja se povečuje sorazmerno z velikostjo množice podatkov za urjenje. Da bi odpravili to pomanjkljivost, je bilo opravljenih že veliko raziskav. Navadno poskušajo zmanjšati število dejansko izvedenih izračunov za oceno razdalje od zapisa od sosedov.

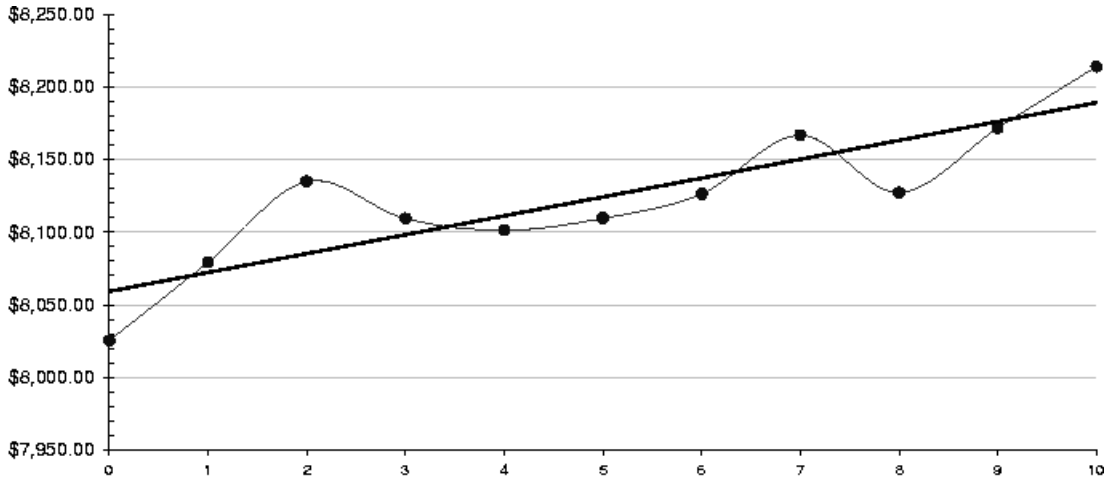
Tehniki, ki poskušata zmanjšati število potrebnih izračunov algoritma k-NN, sta *iskanje najbližjega soseda* (angl. *nearest neighbor search*), in *kD-tree* (Witten & Frank, 2005, str. 130).

4.7 Linearna regresija

V statistiki je sinonim za predvidevanje ena ali druga vrsta regresije. Obstajajo različne oblike regresije, vendar vse delujejo po enakem načelu, in sicer poskušajo napovedati vrednosti odvisne spremenljivke na podlagi neodvisnih spremenljivk ob minimalni napovedni napaki (Witten & Frank, 2005, str. 119–121).

Delovanje linearne regresije si lahko predstavimo v dvodimenzionalnem modelu, v katerem so vrednosti odvisne spremenljivke na osi Y, neodvisne spremenljivke pa na osi X. Preprost linearni regresijski model lahko prikažemo kot premico, ki ponazarja razmerje med znano vrednostjo zapisa (X) in med napovedno vrednostjo (Y). Od vseh možnih leg premice algoritem izbere tiso, ki minimizira razdaljo med premico in točkami. Na tak način minimizira napovedno napako (Berson, Smith, & Thearling, 2009).

Slika 8: Preprosta linearna regresija



Vir:M. Dumas, Aldred, Governatori, Hofstede, & Russell, 2002, slika 3.

4.8 Asociativna pravila

Asociativna pravile (angl. *association rules*): gre za ekstrakcijo uporabnih »če – potem« pravil na osnovi statistične pomembnosti (v kolikšni meri je pravilo točno). Npr.: v podatkovni bazi neke veleblagovnice se je ponavljal podatek, da kupci, ki kupijo polnozrnato žemljico, vzamejo v 90 odstotkih primerov zraven še navadni jogurt. Ker smo pri nastavljanju iskanja pravil določili, da nam računalnik prikaže vsa pravila, ki držijo v 80 odstotkih, dobimo izpis: »Če polnozrnata žemljica, potem navadni jogurt.« Taka ugotovitev nam seveda ne bi pomenila kaj prida, ker za to povezavo vemo že od prej, vendar pa so rezultati rudarjenja včasih presenetljivi in na ta način izvemo za povezave, ki bi jih drugače le stežka odkrili.

Uporaba te tehnike je precej zahtevna, saj je treba analizirati vse vzorce, ki se pojavijo v bazi podatkov. Baze v podatkovnih skladiščih so velike in zato se pojavi tudi veliko število najdenih vzorcev. Med njimi je ogromno neuporabnih.

4.9 Razvrščanje v skupine

Razvrščanje v skupine (angl. *clustering*) je postopek razvrščanja zapisov v skupine, ki so znotraj homogene, med seboj pa heterogene.

Razvrščanje v skupine je eno izmed najbolj raziskovanih področij v podatkovnem rudarjenju. Področja raziskave so podatkovno rudarjenje, statistika, strojno učenje (angl. *machine learning*), tehnologija podatkovnih prostorskih baz (angl. *spatial database technology*), biologija in trženje (Han & Kamber, 2006, str. 384).

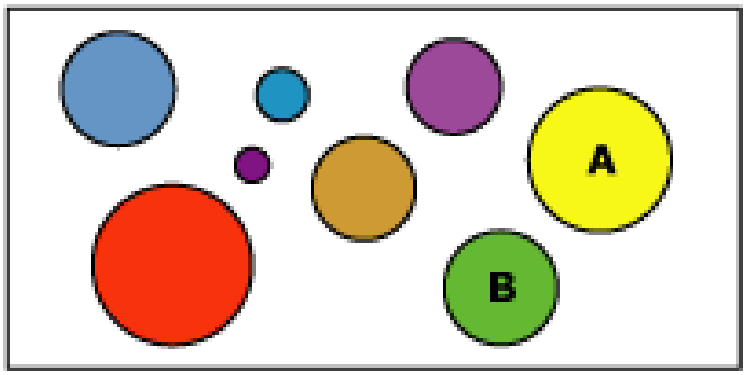
Primere uporabe razvrščanja v skupine lahko razdelimo glede na to, ali je njegov namen boljše razumevanje podatkov ali pa ima kakšno uporabno vrednost (Tan et al., 2006, str. 488–489).

Nekaj primerov uporabe razvrščanja v skupine:

- *Biologija*. Že od nekdaj se biologi ukvarjajo s taksonomijo (sistematičnim razvrščanjem živalskih vrst v kraljestva, razrede, družine …) in zato so poskušali ustvariti sistem, ki bi to naredil avtomatično. V zadnjem času se razvrščanje v skupine uporablja za grupiranje genov s podobnimi funkcijami.
- *Internetni iskalniki*. Svetovni splet sestavlja več milijard spletnih strani in spletni iskalniki podjetij, kot so Google, Yahoo in Microsoft, uporabljajo razvrščanje v skupine. Lep primer razvrščanja v skupine je iskalnik SearchPoint, ki uporablja ontološki prikaz skupin.
- *Podnebje*. Raziskovanje podnebnih sprememb je povezano z iskanjem vzorcev v atmosferi in oceanih. Razvrščanje v skupine je bilo uporabljeno za odkrivanje območij, ki imajo močan vpliv na klimo po svetu.
- *Psihologija in medicina*. Bolezni imajo velikokrat več variacij in razvrščanje v skupine so uporabili za identificiranje različnih tipov depresij.
- *Poslovni svet*. V poslovnem svetu je tehnika razvrščanja v skupine uporabljena za segmentacijo strank v manjše skupine, ki jih je nato lažje analizirati.

Razvrščanje v skupine je koristno za odkrivanje povezav med podatki v podatkovni bazi, ki bi jih sicer težko opazili. V Microsoftove knjižnice za razvijalce (angl. *Microsoft Developer Network*, v nadaljevanju MDSN knjižnica) (Microsoft Corporation, 2009a) je naveden primer, da lahko logično zaključimo, da ljudje, ki se vozijo v službo s kolesom, po navadi ne živijo zelo daleč od službe. Algoritem pa lahko najde skupine, ki niso tako očitne. Primera takih skupin sta skupina ljudi, ki se v službo raje vozijo s kolesom (B), in skupina ljudi, ki se v službo raje vozijo z avtom (A).

Slika 9: Skica najdenih skupin



Vir: Microsoft Corporation, SQL Server 2008 Books Online, Microsoft Clustering Algorithm, 2009.

4.9.1 Microsoftov algoritem za razvrščanje v skupine

V tej točki bom podrobno opisal Microsoftov algoritem za razvrščanje v skupine (angl. *Microsoft Clustering Algorithm*, v nadaljevanju MSCA), ki sem ga uporabil v nadaljevanju diplomske naloge. Opis algoritma je povzet iz MDSN knjižnice (Microsoft Corporation, 2009a).

Microsoftov algoritem za razvrščanje v skupine je segmentacijski algoritem SQL Server 2008 Analysis Services (SSAS). Algoritem razvršča primere (angl. *case*) v podatkovni bazi v skupine

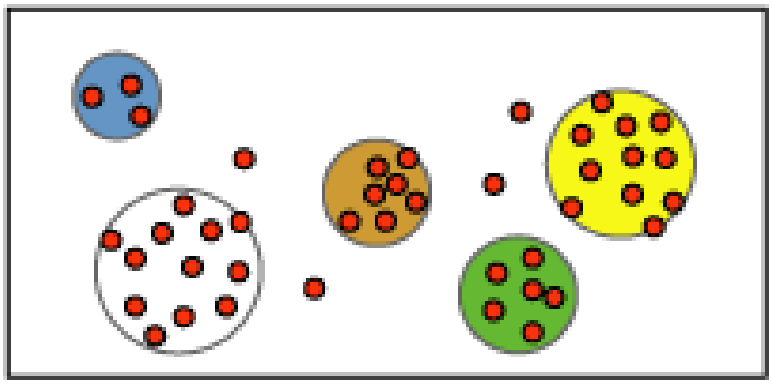
(angl. *clusters*), znotraj katerih imajo primeri podobne lastnosti. Za ta postopek uporablja iterativne tehnike. Dobljene skupine primerov ali zapisov so uporabne za raziskovanje podatkov, iskanje anomalij v podatkih in izdelavo napovedi.

Microsoftov algoritem za razvrščanje v skupine se od drugih algoritmov za podatkovno rudarjenje razlikuje v tem, da ni treba navesti napovednega stolpiča, ali z drugimi besedami, ni nam treba izbrati, katero spremenljivko bi radi napovedali. Algoritem izuri model izključno na podlagi povezav med zapisi in skupinami, ki jih najde.

4.9.2 Podrobna analiza delovanja algoritma za razvrščanje v skupine

MSCA najprej identificira vse povezave med zapisi v podatkovni bazi in nato formira skupine, ki temeljijo na njih. To si najlažje vizualiziramo z uporabo razsevnega diagrama, kot je prikazan na sliki 10.

Slika 10: Najdene skupine in posamezni zapisi



Vir: Corporation, SQL Server 2008 Books Online, Microsoft Clustering Algorithm, 2009.

Točke razsevnega diagrama predstavljajo vse obravnavane zapise, barvni krogi pa so najdene skupine, ki predstavljajo identificirane povezave med zapisi.

Po začetnem koraku identificiranja skupin algoritem preračuna, kako dobro skupine predstavljajo skupine podatkov. Nato poskuša na novo definirati skupine tako, da bi bolje zastopale skupine obravnavanih podatkov. Kako algoritem določi, kaj je »bolje«, bom podrobneje razložil v naslednjem razdelku. Algoritem ponavlja ta korak, dokler ne doseže stanja, v katerem ne more več izboljšati reprezentativnosti najdenih skupin.

4.9.3 Predstavitev implementacije Microsoftovega algoritma za razvrščanje v skupine

Sledi prikaz implementacije Microsoftovega algoritma za razvrščanje v skupine in predstavitev vseh nastavljivih parametrov algoritma, s katerimi vplivamo na vedenje modela.

Obravnavani algoritem nam daje na izbiro dve različni metodi kreiranja skupin in dodeljevanja zapisov vanje:

- Algoritem *optimizacije pričakovanja* (angl. *expectation maximization*, v nadaljevanju EM) pa je »mehka« metoda. »Mehka« metoda pomeni, da zapis pripada več skupinam hkrati, a z različno verjetnostjo.

- *K-means* je »trda« metoda. »Trda« metoda pomeni, da algoritem vsak zapis ali točko dodeli samo eni skupini in da se izračuna samo ena verjetnost članstva (angl. *membership probability*) za vsako točko. Več o tej metodi sledi v točki 4.9.3.2.

4.9.3.1 EM clustering

Ta metoda razvrščanja v skupine izboljšuje začetni model skupin in računa verjetnost, da neka točka oz. zapis pripada določeni skupini. Proces izboljševanja modela se ne konča, dokler verjetnostni model (angl. *probabilistic model*) ne ustreza obravnavanim podatkom.

Če se med procesom grajenja modela pojavijo prazne skupine, so le-te razpuščene in začetno stanje oz. seme (angl. *seed*) se ponovno določi na novih točkah. EM-algoritem se zatem ponovno zažene.

Rezultati EM-metode so verjetnosti. To pomeni, da vsak zapis ali točka pripada vsem skupinam, a z različno verjetnostjo. Ta lastnost EM-metode nam dovoljuje, da se skupine prekrivajo. Zato se lahko zgodi, da je seštevek vseh zapisov v skupinah večji kot začetno število zapisov.

EM-metoda je privzeta metoda v Microsoftovem orodju za iskanje skupin. Razlog za to so njene prednosti v primerjavi z metodo K-means. Prednosti so:

- Zahteva največ en pregled (angl. *scan*) podatkovne baze, ki vsebuje zapise, ki jih želimo uporabiti za izdelavo modela.
- Metoda deluje tudi v primeru majhnega pomnilnika (angl. *random access memory*) strojne opreme.

Microsoftov algoritem ima še dva dodatni izbiri: razširljiv (**angl**. *scalable*) in neprilagodljiv (angl. *non-scalable*) EM. Privzeta nastavitev je razširljiva metoda. V tem primeru algoritem poskusi zgraditi model na prvih 50000 zapisih. Če to ni mogoče, uporabi še naslednjih 50000 zapisov. V neprilagodljivi opciji algoritem prebere vse zapise, ne glede na velikost. S to metodo sicer lahko dobimo točnejši model, vendar so lahko spominske zahteve bistveno večje. Če uporabimo razširljiv način, so vsi podatki v lokalnem pomnilniku (to je sicer mogoče tudi v primeru izbire neprilagodljivega načina, če imamo dovolj velik RAM oz. dovolj majhno bazo), tako da lahko algoritem izkoristi procesor računalnika veliko bolje kot pa v neprilagodljivem načinu. Razširljiv način ima še eno prednost pred neprilagodljivim. Trikrat hitrejši je tudi v primeru, da je pomnilnik dovolj velik za celotno bazo podatkov pri uporabi neprilagodljivega načina. V večini primerov pa točnost modela ne trpi, če izberemo za strojno opremo manj zahtevno možnost. Zaradi zgoraj naštetih razlogov, predvsem strojnih zahtev, tudi jaz nisem uporabil možnosti neprilagodljivega načina.

4.9.3.2 K-means clustering

Metoda K-means za razvrščanje v skupine je dobro poznana. Princip delovanja je minimizacija razlik znotraj skupin ali homogenizacija in maksimizacija razdalje med skupinami oz. heterogenizacija. Beseda »means« ali po slovensko »sredina« se nanaša na *centroide*, ki so naključno izbrani in nato z iterativnim postopkom prilagojeni oz. »premaknjeni« na način, da predstavljajo resnično sredino oz. povprečje vseh zapisov v skupini. »K« v K-means se nanaša

na poljubno število začetnih točk, ki so uporabljene za določanje začetnega stanja oz. semen (angl. *seed*).

K-means dodeli vsak zapis v samo eno skupino in ne dovoljuje dvoma, kam spada kakšen zapis. Članstvo v skupini je izraženo kot oddaljenost zapisa od *centroida.*

Tipično K-means uporabimo za izdelavo modelov z zveznimi atributi, ker se pri njih preprosto izračuna oddaljenost od središča oz. *centroida.* Microsoftov algoritem pa vendar omogoča uporabo metode K-means tudi za diskretne spremenljivke.

Pri uporabi obravnavane metode imamo tako kot pri metodi EM clustering na voljo dva načina vzorčenja podatkov. Eden je neprilagodljiv, ki prebere vse zapise v množici podatkov, drugi pa razširljiv, ki prebere samo prvih 50000 zapisov. Če prvih 50000 zapisov ne zadošča za izdelavo dobrega modela, algoritem prebere še naslednjih 50000 zapisov.

4.9.4 Vpogled v parametre Microsoftovega algoritma za razvrščanje v skupine

V tem razdelku bom predstavil različne nastavljive parametre Microsoftovega algoritma za razvrščanje skupine. Vse te parametre nastavimo pred začetkom samega procesa rudarjenja. Nastavitve vplivajo tako na učinkovitost kot tudi na točnost podatkovnega modela.

Hkrati bom tudi utemeljil svojo izbiro teh nastavitev za proces podatkovnega rudarjenja v tej diplomski nalogi.

CLUSTERING_METHOD

Tukaj lahko izberemo, katero metodo za algoritem bomo uporabili. Na voljo so štiri kombinacije metod, ki so bile podrobno predstavljene v prejšnji točki. Te so:

- scalable EM,
- non-scalable EM,
- scalable K-means,
- non-scalable K-means.

Privzeta metoda je prva.

CLUSTER_COUNT

Ta parameter določa približno število skupin, ki naj jih algoritem kreira. Če želeno približno število skupin ne more biti ustvarjeno, algoritem naredi toliko skupin, kot je mogoče z danimi podatki. Če ta parameter nastavimo na 0, algoritem uporabi hevristični način za določanje najustreznejšega števila skupin.

Privzeta vrednost je 10.

CLUSTER_SEED

Tukaj lahko vnesemo število, ki je potem uporabljeno pri naključni izbiri začetnega stanja ali semena.

S spreminjanjem te številke lahko vplivamo na razvoj začetnih skupin. Pozneje lahko primerjamo različne modele, ki so bili izdelani z različno vrednostjo tega parametra. Če imajo dobljeni modeli približno enake skupine, je to dober znak, da je dobljeni model relativno stabilen.

Privzeta vrednost je 0.

MINIMUM_SUPPORT

Minimalna podpora (angl. *minimal support*) določa, koliko zapisov oz. primerov (angl. *cases*) mora minimalno biti v posamezni skupini. Če je število zapisov v skupini manjše od določene vrednosti, se ta obravnava kot prazna in je zavržena.

Pri tem parametru je treba paziti, da ne vnesemo prevelike številke, ker lahko tako zgrešimo povsem legitimne in uporabne skupine.

Privzeta vrednost je 1.

MODELLING_CARDINALITY

Tu se določi število vzorčnih modelov, ki so zgrajeni med procesom izdelave modela. Manjše število sicer pripomore k hitrejšemu procesiranju, vendar ob tem tvegamo, da bo algoritem zgrešil dober vzorčni model.

Privzeta vrednost je 10.

STOPPING_TOLERANCE

Ta parameter pove, kdaj je dosežena konvergenca in algoritem neha graditi model. Konvergenca je dosežena, ko je celotna sprememba verjetnosti skupin manjša od v tem parametru določene stopnje, deljene z velikostjo modela.

Privzeta stopnja je 10.

SAMPLE_SIZE

Velikost vzorca (angl. *sample size*) določa, koliko zapisov bo algoritem prebral, ko bo gradil model. Ta parameter je relevanten samo v primeru, da smo izbrali eno od razširljivih metod. Če parameter nastavimo na 0, bo algoritem zajel celotno bazo. Tu moramo upoštevati strojne zmogljivosti naše opreme, predvsem velikost delovnega spomina. Če je množica podatkov prevelika in delovni pomnilnik premajhen, lahko pride do težav.

Privzeta vrednost je 50000.

MAXIMUM_INPUT_ATTRIBUTES

S tem parametrom določimo največje število atributov. Preveliko število atributov lahko resno ogrozi delovanje algoritma. Če je vrednost nastavljena na 0, pomeni, da ni maksimalnega števila atributov.

Privzeta vrednost je 225.

MAXIMUM_STATES

Tu gre za največje število različnih vrednosti posameznega atributa. Če ima atribut več različnih vrednosti, kot je dovoljeno, algoritem izbere najpopularnejše vrednosti in zavrže manj popularne.

Če je ta vrednost previsoka, lahko to bistveno vpliva na izvedbo procesa podatkovnega rudarjenja.

Privzeta vrednost tega atributa je 100.

5 Proces podatkovnega rudarjenja

Danes obstaja veliko metodologij podatkovnega rudarjenja. Eden izmed najbolj popularnih je tudi CRISP-DM (angl. *CRoss Industry Standard Process for Data Mining*), ki je prikazan na sliki 12. Zanj sem se določil ker ga je mogoče uporabiti v vseh panogah in z vsemi orodji in ker je vsebinsko podoben KDD procesu. Proces je iterativni postopek, ki vključuje več korakov (The CRISP-DM consortium, 2007).

5.1 Identifikacija in definiranje problema

Prvi korak vsakega projekta podatkovnega rudarjenja je razumevanje domenske problematike in formulacija problema oziroma vprašanja, na katerega želimo odgovoriti. Mimo tega koraka ne moremo, ker je podlaga za naslednje korake.

5.2 Razumevanje podatkov

Razumevanje podatkov se z zbiranjem podatkov in nadaljuje s spoznavanjem strukture in narave podatkov. Tu običajno zaznamo težave s kvaliteto, oblikujemo prvo sliko o strukturi podatkov in opazimo potencialno zanimive hipoteze o skritih informacijah.

5.3 Pridobivanje podatkov in njihova priprava

Drugi korak je pridobivanje in priprava podatkov. Kakovost podatkov, ki jih uporabimo za podatkovno rudarjenje, v veliki meri določa kakovost končnega produkta. Za podatkovno modeliranje velja znani rek »garbage in, garbage out«, kar v prevodu pomeni »smeti noter, smeti ven« (Rud, 2001, str. 51).

Podatkovne baze, ki so nam običajno na voljo, so sestavljene iz heterogenih virov in zelo pogosto vsebujejo manjkajoče ter protislovne podatke. Priprava podatkov vključuje integracijo podatkov, ugotavljanje pristranskosti podatkov, bogatenje podatkov, določanje strukture podatkov, odpravo manjkajočih vrednosti, redukcijo podatkov in transformacijo podatkov (Payle, 1999, str. 92–94).

Tu je morda dobro omeniti, da so nekateri strokovnjaki mnenja, da bi morali vsaj en podatkovni model izvesti s »surovimi« podatki in jih šele nato po potrebi očistiti. Pravijo, da so podatki na neki način »sveti« (Lukawiecki, 2008d). Priprava podatkov je ključni in zahteven del procesa izdelave podatkovnega modela. Lahko zavzame tudi 80 odstotkov napora. Več o tem v točki 7.5.

5.4 Modeliranje

Modeliranje je jedro celotnega procesa KDD. Tukaj z uporabo različnih inteligentnih metod iz podatkov izluščimo vzorce. Tu moramo najprej izbrati nalogo (npr. napovedno ali opisno), nato pa še tehniko, ki jo bomo uporabili. Najpogosteje uporabljene tehnike so razvrščanje v skupine, klasifikacija, regresija in asociativna pravila.

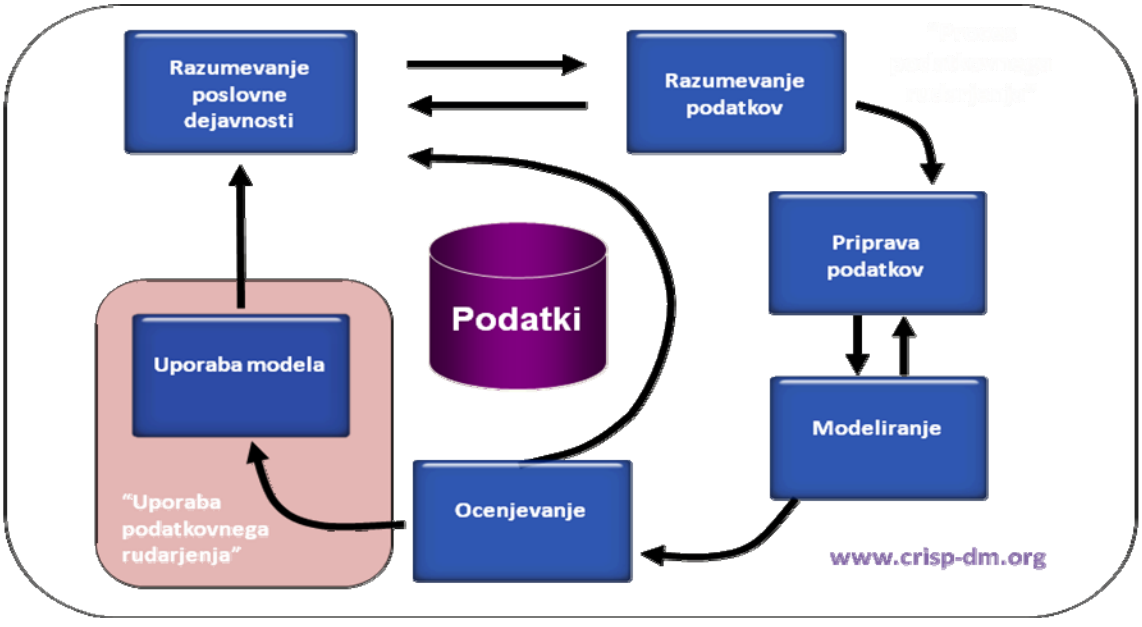
5.5 Ocenjevanje podatkovnega modela

Četrti korak postopka KDD je interpretacija dobljenega modela oziroma znanja, ki smo ga odkrili. Vsak dobljeni model ni uporaben ali koristen. Lahko je slab ali nesmiseln. Lahko nam ne pove nič novega. Kakovost podatkovnega modela lahko ocenimo na podlagi svoje presoje ali presoje domenskega strokovnjaka. Standardni način pa je, da uporabimo testno množico podatkov (običajno 30 odstotkov originalne množice podatkov) in z njim ocenimo točnost modela. Če s stopnjo točnosti nismo zadovoljni, ponovimo proces, v katerem lahko izberemo drugo tehniko modeliranja.

5.6 Uporaba podatkovnega modela

Zadnji korak procesa KDD je praktična uporaba odkritega znanja. Nekaj praktičnih uporab podatkovnega rudarjenja smo že našteli v točki 2.1**Error! Reference source not found.**. Zelo preprost in do neveščega uporabnika prijazen način je kar z uporabo Excela oz. dodatkov zanj. Uporaba podatkovnega modela tudi povečuje razumevanje poslovne dejavnosti oz. strank, kar zvišuje konkurenčnost organizacije.

Slika 11: Proces podatkovnega rudarjenja

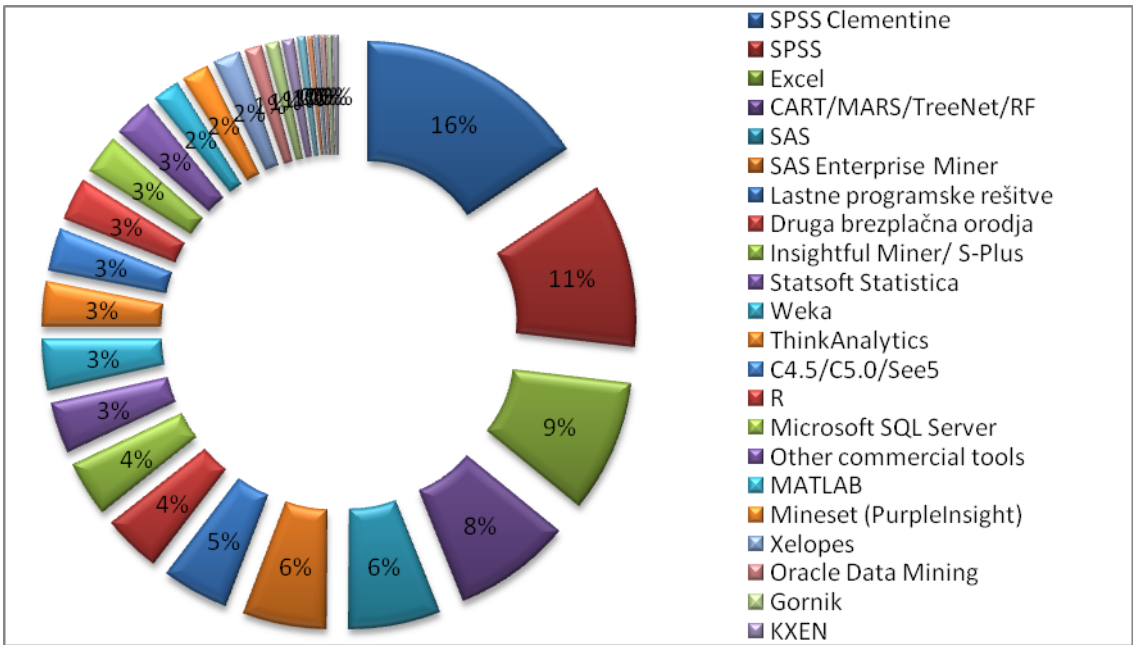


Vir: CRoss Industry Standard Process, 2007.

6 Izbira programskega orodja

Za izvedbo podatkovnega rudarjenja je na voljo veliko programskih rešitev. Na sliki 12 je prikazanih nekaj najbolj razširjenih programov za podatkovno rudarjenje in analitičnih orodij.

Slika 12: Popularnost programskih orodij za podatkovno rudarjenje in analitiko



Vir: KDnuggets, 2006.

Programska orodja lahko razdelimo na dve skupini. Prva skupina so programi, razviti za specifično nalogo v podatkovnem rudarjenju.

Imamo specializirane programe za klasifikacijo, razvrščanje v skupine in segmentacijo, tekstovno rudarjenje, statistične analize, spletno rudarjenje, rudarjenje po socialnih omrežjih, rudarjenje po avdio in video zapisih ter še veliko drugih (KDnuggets, 2006).

Druga skupina pa so programski paketi (angl. *program suite*) ali zbirke različnih programov, ki omogočajo izvedbo celotnega procesa podatkovnega rudarjenja in uporabo vseh glavnih tehnik. Produkti večjih proizvajalcev omogočajo tudi integracijo s podatkovnimi bazami in podatkovnimi skladišči organizacije.

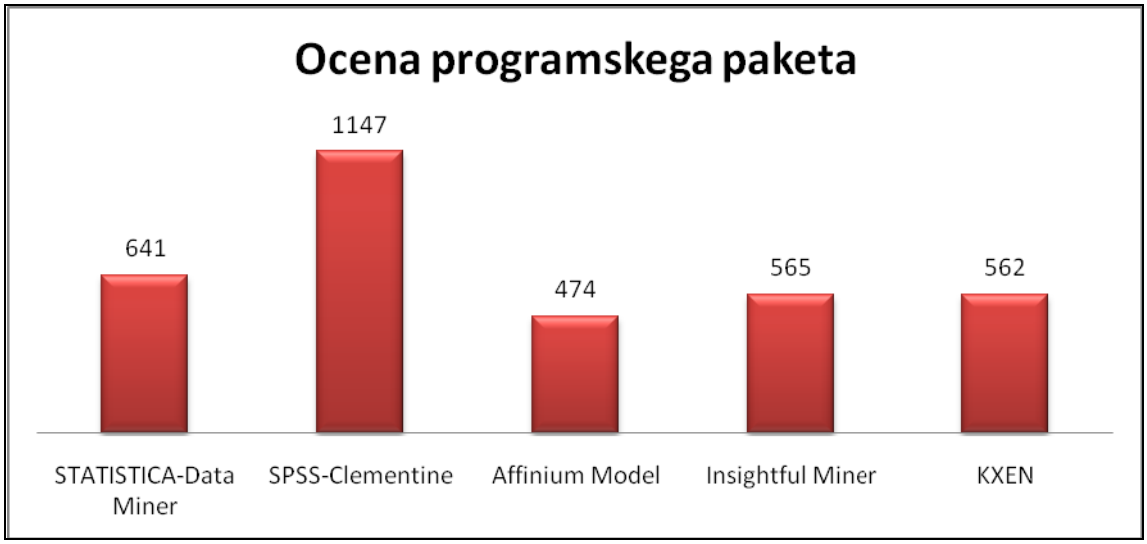
Izbira programskega paketa ni lahka naloga. Kljub splošnemu prepričanju najboljše orodje ni tisto, ki ima največ algoritmov za podatkovno rudarjenje, ali tisto, ki daje največjo točnost pri napovedih, ali pa tisto, ki ja najnaprednejše. Najpomembneje je, da je programski paket:

- preprost za uporabo,
- zagotavlja zadovoljiv nivo točnosti,
- omogoča izvedbo vseh nalog v procesu podatkovnega rudarjenja,
- je cenovno sprejemljiv.

Odgovori na zgoraj naštete parametre pa so drugačni glede na to, kdo jih uporablja in kakšne zahteve ima. Večina podjetij se bo verjetno zadovoljila z orodjem, ki dosega 80 odstotkov točnosti glede na najboljše orodje, če bo to občutno cenejše. Nekateri uporabniki pa bodo izbrali najtočnejše orodje ne glede na ceno.

Dr. Robert Nisbet v članku »How to Choose a Data Mining Suite« primerja 5 najbolj znanih programskih paketov za podatkovno rudarjenje. Rezultati so predstavljeni na sliki 13.

Slika 13: Ocena programskih paketov za podatkovno rudarjenje



Vir: R.A. Nisbet, How to Choose a Data Mining Suite, 2004, str. 5.

Sam sem se odločal med odprtokodnim programom Weka in Microsoftovim Microsoft SQL Serverjem 2008.

- Weka ima to prednost, da je brezplačno dostopna na internetu. Zaradi njene razširjenosti je tudi veliko virov s pomočjo za uporabnike. Witten in Frank sta tudi napisala knjigo z naslovom *Data Mining: Practical Machine Learning Tools and Techniques*, v kateri na praktičnih primerih opisujeta delo z Weko. Program ima tudi orodja za pripravo podatkov in veliko algoritmov za razvrščanje v skupine, regresijo, klasifikacijo in asociacijska pravila. Vsebuje tudi orodje za vizualizacijo, ki pa je po mojem mnenju pomanjkljivo. To in pa dejstvo, da morajo biti vsi podatki, s katerimi želimo modelirati, v formatu ARFF (Attribute-Relation File Format), ki je zelo okoren, sta največji pomanjkljivosti tega programa.
- Microsoft SQL Server 2008 je podatkovni strežnik, ki ima v sklopu SQL Server Analysis Services vgrajene zmogljivosti za podatkovno rudarjenje. Prednost tega programskega paketa je predvsem podpora uporabniku, ki jo ponuja. Na voljo je ogromno knjig, internetnih strani, člankov in forumov, ki razlagajo delovanje tega programa. Več o lastnostih uporabljenega orodja v točki 7.3. Slaba plat je seveda cena, ki znaša 5.999,00 USD za standardno različico.

7 Uporaba razvrščanja v skupine na primeru podjetja Merkur d.d.

Praktičen del diplomske naloge temelji na konkretnem primeru podjetja Merkur d.d. Glavni cilj tega poglavja je predstaviti celotni postopek podatkovnega rudarjenja na konkretnem primeru. V postopku bom uporabil tehniko razvrščanja v skupine in metodologijo CRISP-DM opisano v

točki 5. Skušal bom najti takšne skupine, ki bi v dejanskem poslovnem okolju imele dodano vrednost, denimo za oddelek trženja, ki bi tako lahko bolj prilagodil direktno trženje.

7.1 Podjetje Merkur

Merkur sestavljajo tri divizije z 20 podjetji v 8 državah, 87 trgovskih centrov in franšiznih prodajaln ter dve spletni trgovini (Merkur d.d., 2008, str. 9).

Merkur sestavljajo poleg delniške družbe Merkur še tri podjetja v Sloveniji, tj. Bofex, Sava Trade in Kovinotehna, ter podjetja v tujini (Zagreb, Praga, Skopje, München, Milano, Varšava, Sarajevo, Rusija in Beograd) (Merkur - Novinarsko središče).

Hčerinsko podjetje Bofex s svojima blagovnima znamkama Big Bang in BOF ponuja kupcem izdelke akustike, zabavne elektronike, bele tehnike in male gospodinjske aparate. Kovinotehna pa se ukvarja z upravljanjem nepremičnin (Merkur - Novinarsko središče).

Podjetja v tujini so večinoma specializirana za uvoz in izvoz izdelkov iz dela prodajnega programa delniške družbe Merkur. Poslovanje v tujini je pretežno usmerjeno v oskrbo industrijskih podjetij in v nabavo blaga za matično podjetje, vse bolj pa postaja pomembno tudi pri širjenju maloprodajne mreže na tujih trgih (Merkur - Novinarsko središče).

Raznovrsten prodajni program podjetij Merkur Group sestavlja več kakor 700 skupin blaga s približno 200.000 izdelki. Skupine blaga združujejo v naslednje strateške programe:

Metalurški izdelki: pločevina, nerjavna pločevina, nosilci in profili, varilno-tehnični materiali, palična jekla in žice, cevi, orodna jekla, betonsko jeklo, armaturne mreže in krivljena armatura, barvasta metalurgija.

Gradbeni material in les: cement in apno, opeka in strešne kritine, izolacije, suho montažni gradbeni elementi, izdelki za urejanje vrta in okolice, les in lesni izdelki, stavbno pohištvo, stenske in talne obloge.

Tehnični izdelki: okovje, vijaki in pritrdilna tehnika, strojno in vpenjalno orodje, brusni materiali, stroji in pribor, merilno orodje, logistična oprema, ročno orodje, tehnični izdelki iz gume, osebna zaščitna sredstva.

Energetika in inštalacije: elektroinštalacije, razsvetljava, vodniki in kabli, stikalna tehnika, energetska oprema, vodovodne inštalacije, sanitarna keramika, keramične ploščice, oprema za wellness, ogrevanje, prezračevanje in klimatizacija ter drugi inštalacijski in elektromateriali.

Izdelki za široko potrošnjo: akustika in videotehnika, mali gospodinjski aparati, bela tehnika, ogrevalna telesa, gospodinjski pripomočki, računalniška in biro-oprema, telekomunikacije, zeleni program in kmetijsko-gozdarski program, drugi izdelki široke potrošnje.

Kemija in papir: barve in kemični izdelki, fasadni sistemi, industrijske kemikalije, plastični granulati, grafični papirji in materiali, embalažni materiali (Merkur d.d., 2008, str. 9).

7.2 Izbrana baza podatkov

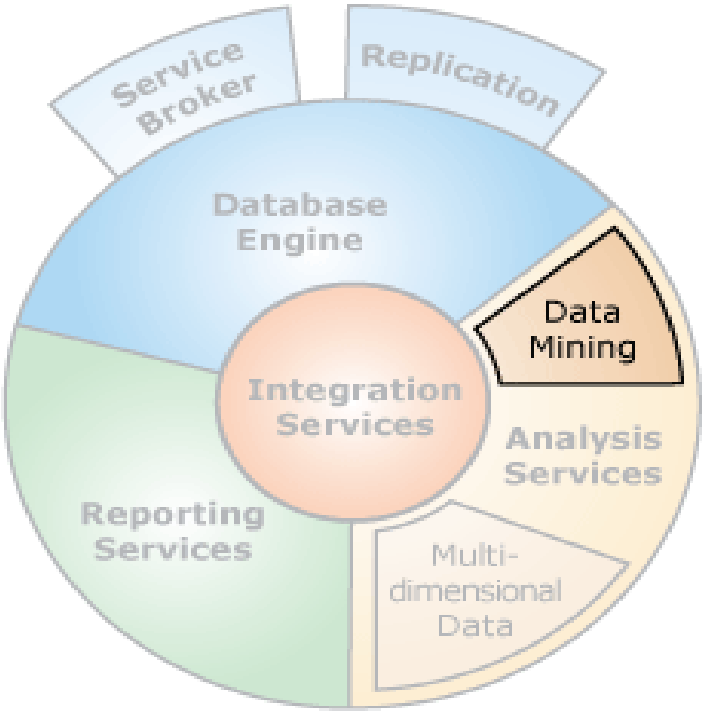
Za osnovo praktičnega dela diplomske naloge sem uporabil podatke o maloprodaji podjetja Merkur, d. d. Podjetje hrani podatke o maloprodaji v vseh svojih centrih. Zajemajo podatke o vrsti, skupini in nazivu blaga, bruto vrednosti nakupa, količini kupljenega blaga, datumu nakupa, lokaciji nakupa in številki Merkurjeve kartice pri posameznem nakupu. Te podatke o nakupih pa povežejo z informacijami o kupcih oziroma imetnikih Merkurjeve kartice. Te informacije so datum rojstva kupca, pošta njegovega prebivališča, število računov, bruto vrednost prodaje in število kupljenih izdelkov. Podrobneje o uporabljenih podatkih v točki 7.5.1.

7.3 Program, izbran za podatkovno rudarjenje

V točki 6 sem opisal, kakšna je izbira orodij za podatkovno rudarjenje in kakšne so razlike med njimi.

Odločil sem se za Microsoftovo rešitev Microsoft® SQL Server® 2008 (v nadaljevanju MSSQL). Zanj sem se odločil, ker ima širok nabor algoritmov, je združljiv z drugimi programi, kot sta Access in Excel, predvsem pa ponuja rešitve za vse korake v projektu podatkovnega rudarjenja.

Slika 14: Pregled komponent MSSQL 2008



Vir: Microsoft Corporation, 2009.

MSSQL 2008 je relacijski podatkovni strežnik. Na sliki 14 vidimo, da ga sestavlja več enot, in enota za podatkovno rudarjenje je samo ena od njih. Pri svojem delu s podatki sem uporabljal samo tri enote, in sicer: podatkovno bazo (Database Engine), kjer sem hranil vse podatke, integrirane storitve (Integration Services), kjer sem opravil pripravo podatkov, opisano v točki 7.5, in pa enoto za podatkovno rudarjenje (Analysis Services) za samo podatkovno rudarjenje.

7.4 Razumevanje poslovnega problema

Razumevanje poslovnega problema je ključnega pomena za nadaljnji potek projekta podatkovnega rudarjenja. Tu določimo kaj želimo doseči kot rezultat podatkovnega rudarjenja in od tega je odvisna uporabljena tehnika, zbrani podatki, orodje in ostalo. Poslovni problem v mojem primeru je sicer namišljen in tudi pri tipu, količini in kvaliteti podatkov sem bil omejen ne to kar je bilo na voljo.

Poslovni problem s katerim se sooča podjetje Merkur je, da ne poznajo svojih kupcev dovolj dobro. Pridobljeno znanje bi nato radi izkoristili v trženju. Odgovor na ta poslovni projekt se najlažje zagotovi z uporabo razvrščanja v skupine.

7.5 Priprava podatkov

V celotnem procesu podatkovnega rudarjenja je priprava podatkov ključnega pomena. Nekateri avtorji omenjajo pripravo podatkov kot jedro procesa podatkovnega rudarjenja. Kakor koli že, danes so podatkovne baze polne šuma (angl. *noise*), manjkajočih podatkov in nekonsistentnosti zaradi svoje velikosti in dejstva, da so pogosto sestavljene iz zelo raznolikih virov. Podatkovna baza lahko vsebuje polja, ki so zastarela ali odvečna, manjkajoča polja, izjeme (angl. *outlier*), podatke v oblikah, ki niso ustrezne za proces rudarjenja, ali pa vrednosti, neskladne z vsakdanjo logiko (npr. negativno število za starost osebe). Priprava podatkov je ključni korak v postopku KDD in nekatera podjetja menijo, da korak priprave podatkov porabi od 50 do 80 odstotkov razpoložljivih virov za posamezni projekt podatkovnega rudarjenja (Han & Kamber, 2006, str. 47–51).

Pripravo podatkov lahko razdelimo na štiri segmente (Han & Kamber, 2006, str. 48–51):

- analiza podatkov,
- čiščenje podatkov
- integracija in transformacija podatkov,
- redukcija podatkov.

V nadaljevanju bom na kratko opisal vsakega od njih in predstavil njihovo aplikacijo na podatke podjetja Merkur, d. d..

7.5.1 Analiza podatkov

Analiza podatkov se uporablja za splošen pregled podatkov, s katerimi nameravamo izvesti podatkovno rudarjenje. Zanima nas, kakšna je porazdelitev podatkov, njihova struktura, razpon vrednosti, koliko zapisov manjka, ali podatki vsebujejo izjeme in podobno.

Podatki podjetja Merkur, d. d., ki sem jih uporabil v tej diplomski nalogi, so bili v dveh delih. Eden je vseboval podatke o imetnikih kartice zvestobe, drugi pa podatke o nakupih v letu 2005.

Podatki o imetnikih kartice so bili sestavljeni iz atributov, podanih v tabeli 1.

Tabela 1: Imena in opisi atributov imetnikov kartic

Naziv atributa	Opis atributa
MKZM stev. kartice ID	Številka kartice
MKZM posta ID	Številka pošte
MKZM posta DESC	Naziv pošte
MKZM datum rojstva ID	Datum rojstva imetnika kartice
Število prodajnih dokumentov	Število prodajnih dokumentov (računov)
BtProdVred	Bruto prodajna vrednost prodaje
StevProdPost	Število postavk na računih
OE	Organizacijska enota

Analiza zgornjih podatkov je pokazala naslednje:

- Velika večina kupcev, 63 odstotkov, je Ljubljančanov.
- Letnica rojstva je normalno porazdeljena. Vsebuje kar nekaj manjkajočih in nelogičnih vnosov, kot je na primer datum rojstva 19. 1. 2058.
- Bruto vrednost nakupov je logaritemsko normalno porazdeljena (angl. *log-normal*), kar pomeni, da je skoraj 84 odstotkov vseh nakupov v razredu od 0 do 80.000 SIT (cene so v SIT ker so podatki še iz obdobja ko v Sloveniji nismo imeli evra).
- Tako kot bruto vrednost nakupov ima tudi število postavk na računih izrazito logaritemsko normalno porazdelitev. Kar 94 odstotkov kupcev je imelo od 0 do 16 nakupov.
- Organizacijske enote so neenakomerno zastopane. Enota 229 (PE Vič) ima prevladujoč delež, in sicer 67 %.
- Številka kartice je najboljši kandidat za ključ, čeprav vsebuje kar nekaj podvojenih vrednosti.

Podatki o nakupih imetnikov kartice pa so vsebovali atribute v tabeli 2:

Tabela 2: Atributi tabele o nakupih imetnikov kartic

Naziv atributa	Opis atributa
Blg naziv ID1	Šifra vrste blaga
Blg naziv ID2	Šifra skupine blaga
Blg naziv ID3	Šifra blaga
Blg naziv DESC	Naziv blaga
Artikel naziv ID	Šifra artikla
Artikel naziv DESC	Naziv artikla
MKZM stev kartice ID	Številka Merkurjeve kartice zaupanja
MKZM posta ID	Številka pošte
MKZM posta DESC	Naziv pošte
MKZM datum rojstva ID	Datum rojstva imetnika Merkurjeve kartice
Dat prometnega dokumenta ID	Datum računa
Stev pdok ID	Številka računa
BtProdVred	Bruto prodajna vrednost postavke računa
KolProd	Količina artikla

Analiza podatkov o nakupih kupcev je pokazala naslednje:

- Za razvrščanje v skupine je primeren samo atribut »vrsta blaga«, ker ima le 100 različnih vrednosti. Agregata »šifra blaga« in »skupina blaga« sta preveč razdrobljena za uporabo pri razvrščanju s skupine. Uporabiti moramo atribut »vrsta blaga«, ki je na najvišjem nivoju.
- Atribut »vrsta blaga« je samo v obliki šifre, ki nam ne koristi, razen če je ne povežemo z vsebinskim šifrantom, ki šifre prevede v opise skupin.
- 97 % prodanih artiklov je imelo ceno med 0 in 25.000 SIT, preostali 3 % pa so ekstremi.
- Kar četrtino artiklov predstavljajo izdelki s popustom.

7.5.2 Čiščenje podatkov

Namen čiščenja podatkov je odstranitev šuma in nezaželenih podatkov iz množice podatkov, ki jih bomo uporabili za podatkovno modeliranje (MacLennan, Tang, Crivat, 2009, str. 10).

Šum so vsi vplivi napak, ki vplivajo na pravo vrednost atributov. Lepa analogija je šum, ki ga včasih lahko slišimo po radiu (Payle, 1999, str. 93). Šum predstavlja vse, kar ni produkt radijske postaje. Šum je zelo težko v celoti odpraviti, saj ne obstaja nobena tehnika, ki bi to omogočala. Han in Kamber (2006, str. 63) predlagata dve tehniki za glajenje (angl. *smoothing*) podatkov. Prva je vezava v razrede (angl. *binning*), druga pa regresija, vendar ju nisem uporabil, ker bi bil vložen napor v primerjavi s koristmi prevelik..

Nezaželeni podatki so prisotni v skoraj vseh podatkovnih bazah in to je veljalo tudi za podatke podjetja Merkur, d. d..

V atributu »letnica rojstva« je bilo 2381 zapisov z neznano vrednostjo. Čeprav je to predstavljajo le 1,89 odstotka celotne populacije, sem se vseeno odločil odpraviti pomanjkljivost. To je mogoče doseči na več načinov. Zapis z manjkajočim atributom lahko ignoriramo, vstavimo manjkajočo vrednost ročno, lahko nadomestimo vse manjkajoče zapise z globalno konstanto, npr. »nepoznan«. Manjkajoči atribut lahko nadomestimo s povprečno vrednostjo vseh enakih atributov, lahko pa nadomestimo manjkajoči atribut s povprečno vrednostjo razreda ali pa izberemo najverjetnejšo vrednost manjkajočega atributa (Han & Kamber, 2006, str. 62). V konkretnem primeru sem manjkajoče vrednosti nadomestil s povprečno vrednostjo atributa.

V atributu »naziv izdelka« je bilo 25 odstotkov kupljenih artiklov zabeleženih kot »popust«. Ker bi to močno vplivalo na rezultat razvrščanja v skupine, sem se odločil te zapise odstraniti.

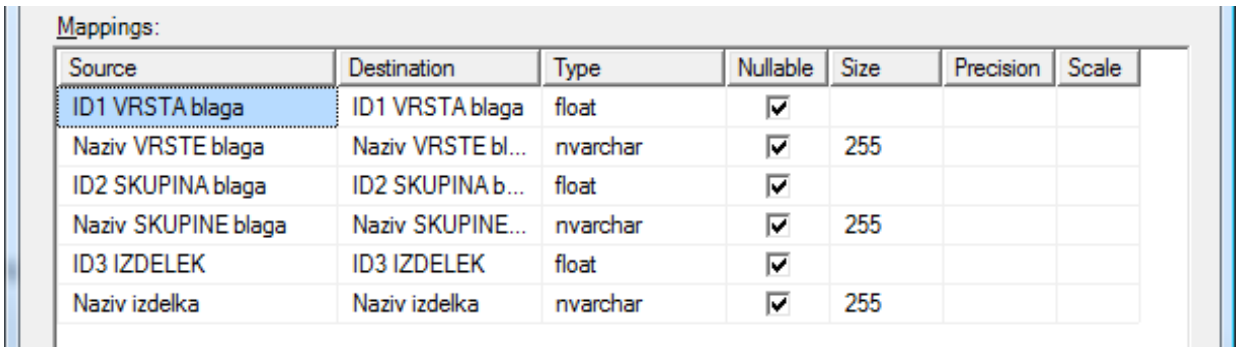
V atributu »prodajna cena« sem odstranil vse zapise, ki so imeli negativni predznak. Ti zapisi so bili posledica popusta oziroma vračila in za analizo skupin niso relevantni.

7.5.3 Integracija in transformacija podatkov

Integracija podatkov je potrebna skoraj pri vsakem KDD. Integracija podatkov je na kratko poenotenje podatkov, ki prihajajo iz različnih virov, v konsistentno celoto. Podatki iz različnih virov se lahko razlikujejo po poimenovanju atributov, nivoju podatkov, merskih enotah, formatu zapisa in podobno (Han & Kamber, 2006, str. 67–70).

Za potrebe te diplomske naloge sem vse podatke uvozil v SQL-strežnik in tako poenotil format, v katerem so bili vsi zapisi. Predvsem je šlo za uvoz podatkov iz MS Accessa in MS Excela. V procesu uvoza podatkov sem poenotil tudi tip podatkov, tako da je bila njihova obdelava v SQL Server 2008 Integration Services (v nadaljevanju SSIS) in v SQL Server 2008 Analysis Services sploh mogoča. Korak pri uvozu podatkov je prikazan na sliki 15.

Slika 15: Korak pri uvozu podatkov iz MS Accessa v MS SQL Server 2008



Transformacijo podatkov uporabimo za ureditev podatkov v format, primeren za podatkovno rudarjenje. Podatke podjetja Merkur, d. d., sem transformiral z uporabo naslednjih tehnik:

*Generalizacija atributov (*angl. *generalization)*

V primeru, da so podatki prepodrobni in zaradi tega preveč razdrobljeni, uporabimo generalizacijo. To tehniko sem uporabil pri atributu »naziv izdelka«, ki je imel 202 različni vrednosti. Najvišji nivo generalizacije je bil atribut »naziv skupine«, ki ima samo 61 različnih vrednosti in je zato veliko primernejši za razvrščanje v skupine.

*Izpeljava novih atributov (*angl. *attribute construction)*

Za boljše in lažje razumevanje podatkov si pomagamo z izpeljavo novih atributov iz že obstoječih podatkov. Precej Merkurjevih podatkov je v originalu neustreznih za analizo, zato sem izpeljal kar nekaj atributov. Izpeljani atributi so naslednji:

- »Starost kupca« – starost kupca je zelo zanimiv podatek, saj vemo, da starost kar precej vpliva na človekove navade. Ta atribut je izpeljan iz datuma rojstva kupca.
- »Mesec rojstva« in »dan rojstva« sta izpeljana iz datuma rojstva.
- »Povp. vrednost nakupa« – povprečna vrednost nakupa je zelo zanimiv podatek, izpeljan iz bruto vrednosti vseh nakupov, deljene s številom vseh nakupov.
- »Povp. vrednost kupljenega izdelka« – atribut, izpeljan iz bruto vrednosti vseh nakupov, deljene s številom prodajnih postavk.
- »Povp. število kupljenih artiklov na nakup« – atribut, izpeljan iz števila prodajnih postavk, deljenega s številom nakupov.
- »Ime dneva« – ime dneva nakupa je izpeljano iz datuma nakupa. Zanimivo bi bilo videti, ali se nakupi spreminjajo glede na dan nakupa. Ali je kakšna posebna skupina kupcev, ki kupuje v petek?

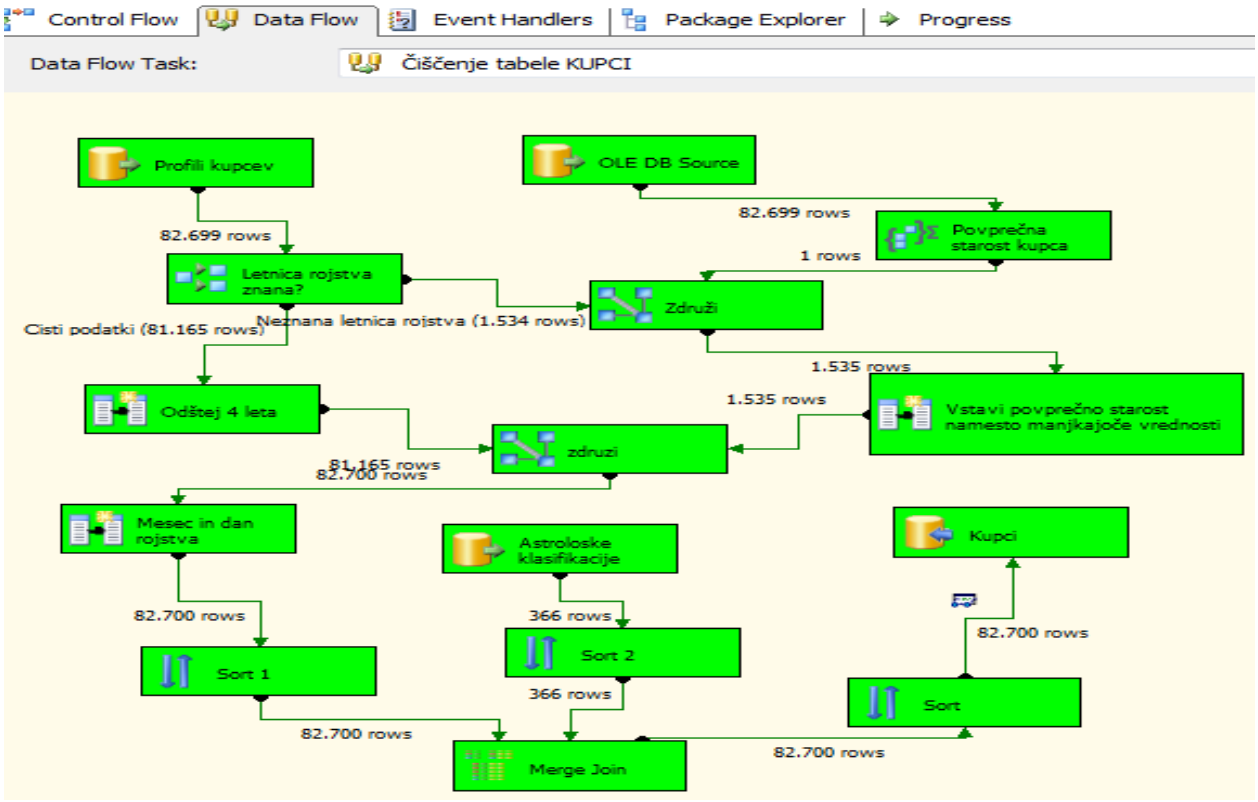
- »Dan nakupa«, »mesec nakupa« in »leto nakupa« so atributi, izpeljani iz datuma nakupa. Zanimivo bi bilo videti, ali je kakšna skupina »sezonskih« kupcev.

*Združevanje (*angl. *aggregation)*

Tabela, ki je vsebovala podatke o kupcih, je imela 126.664 zapisov. Po analizi podatkov se je izkazalo, da je 35 odstotkov zapisov z isto številko kartice (atribut »MKZM stev kartice ID«), to pa zato, ker so nekateri kupci kupovali v različnih trgovskih centrih. Odločil sem se združiti zapise tako, da se vsak kupec v tabeli pojavi samo enkrat. To je bilo potrebno zaradi Microsoft SQL Serverja 2008, saj program zahteva, da so v glavni tabeli unikatni zapisi. Na koncu sem v tabeli »Kupci« dobil 82.699 unikatnih zapisov.

Transformacijo podatkov sem izvajal v SSIS. Slika 16 prikazuje del procesa obdelave.

Slika 16: Podatkovni tok v SSIS pri transformaciji podatkov



7.5.4 Redukcija podatkov

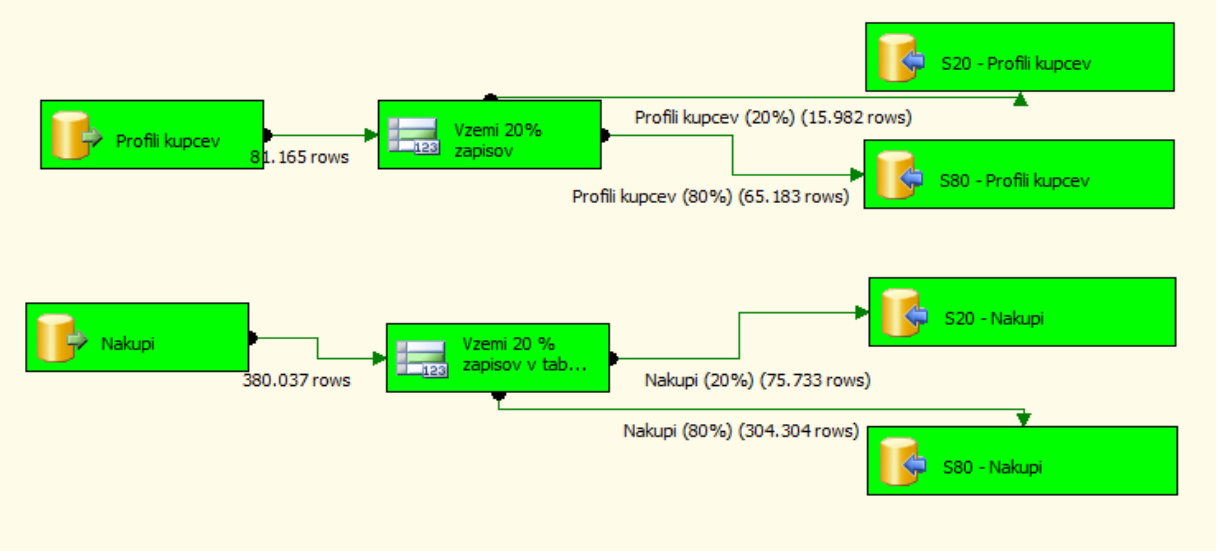
Kot je bilo že omenjeno, so podatkovne baze lahko zelo velike in polne atributov, ki jih ne potrebujemo za učinkovito izgradnjo modela. Zaradi tega uporabljamo reduciranje podatkov. To je treba izvesti na tak način, da se ohranijo struktura, razmerja in relacije med podatki.

Podatke lahko reduciramo na več načinov. Lahko reduciramo dimenzije, izberemo samo najpomembnejše atribute, jih agregiramo v podatkovne kocke ali pa jih diskretiziramo.

Pri praktičnem delu diplomske naloge sem opazil, da je procesiranje vseh zapisov zelo neučinkovito, saj so posamezne operacije vzele tudi do dve uri. Zato sem se odločil za naključno vzorčenje. V SSIS sem naključno izbral 20 odstotkov zapisov iz obeh tabel. Tako dobljene tabele

so bile občutno manjše, a so še vedno vsebovale enake povezave in strukturo. Proces vzorčenja je prikazan na sliki 17.

Slika 17: Vzorčenje podatkov



7.6 Izbrana tehnika

Tehnika razvrščanja v skupine spada v skupino nenadzorovanega (angl. *unsupervised*) podatkovnega rudarjenja. Nenadzorovano je zaradi tega, ker uporabnik ne določi ciljnih razredov oziroma ciljnega atributa.

Tehnika razvrščanja v skupine se pogosto uporablja kot prvi korak v procesu podatkovnega rudarjenja. Z njo lahko najdemo homogene skupine in jih nato uporabimo za nadzorovano modeliranje (npr. za klasifikacijo). Tehnika je opisana že v točki 4.9 in je zato tu ne bomo ponovno opisovali.

Razvrščanje v skupine sem si izbral, ker sem hotel odkriti, če med Merkurjevimi kupci obstajajo različne skupine kupcev, ki bi pozneje omogočale bolj prilagojeno Merkurjevo direktno trženje.

7.7 Potek dela v SSAS in izbrani parametri

Delo v SSAS poteka po korakih.

Prvi korak je **določanje vira** podatkov. V konkretnem primeru so bili vsi podatki uvoženi na SQL-strežnik, kar je najboljša izbira. Komunikacija med programoma je hitra in preprosta.

Nato sledi **definicija podatkovnega pogleda**. Tu nastavimo relacije med uvoženimi tabelami (če jih je več), preimenujemo atribute, če želimo prijaznejša imena, definiramo glavne ključe in raziščemo strukturo podatkov.

Določanje glavnega ključa je zelo pomembno, saj s tem programu povemo, kateri zapisi so za nas predmet analize. Podatke podjetja Merkur bi lahko na primer analizirali po imetniku kartice, številki računa, nazivu pošte ipd. Ker nas zanimajo kupci, smo si za glavni ključ izbrali številko kartice. Pogoj, da je atribut primeren za glavni ključ, pa je, da se zapisi ne podvajajo. Prav zaradi tega smo morali v fazi priprave podatkov (točka 7.5.3) združiti zapise po številki kartice.

Tretji korak je **definicija strukture** za podatkovne modele. Tu določimo, katere atribute in katere definicije podatkovnega pogleda želimo uporabiti. Definiramo tudi tip in vsebino podatkov .

Končno pridemo do ključnega: **izbire algoritma**, **nastavitve parametrov** in določanja **lastnosti atributom**.

Izbrani atributi in njihove nastavitve so podani v tabeli 3. V uporabljenem programskem orodju imamo možnost diskretizacije atributov, kar znatno izboljša razvrščanje, ker lahko tvorimo smiselne razrede npr. nizko, srednje, visoko. S takimi razredi imamo pri analizi lažje delo, saj so lastnosti razredov očitnejše.

Tabela 3: Vhodni atributi in nastavitve

Vhodni atribut	Nastavitve atributa
Bruto vrednost nakupov EUR	Diskretiziran na 5 razredov z metodo skupin
Povp. vrednost kupljenega izdelka EUR	Diskretiziran na 5 razredov z metodo enakih površin
Povp. vrednost nakupa EUR	Diskretiziran na 5 razredov z metodo enakih skupin
Število Nakupov	Diskretiziran na 5 razredov z avtomatično metodo

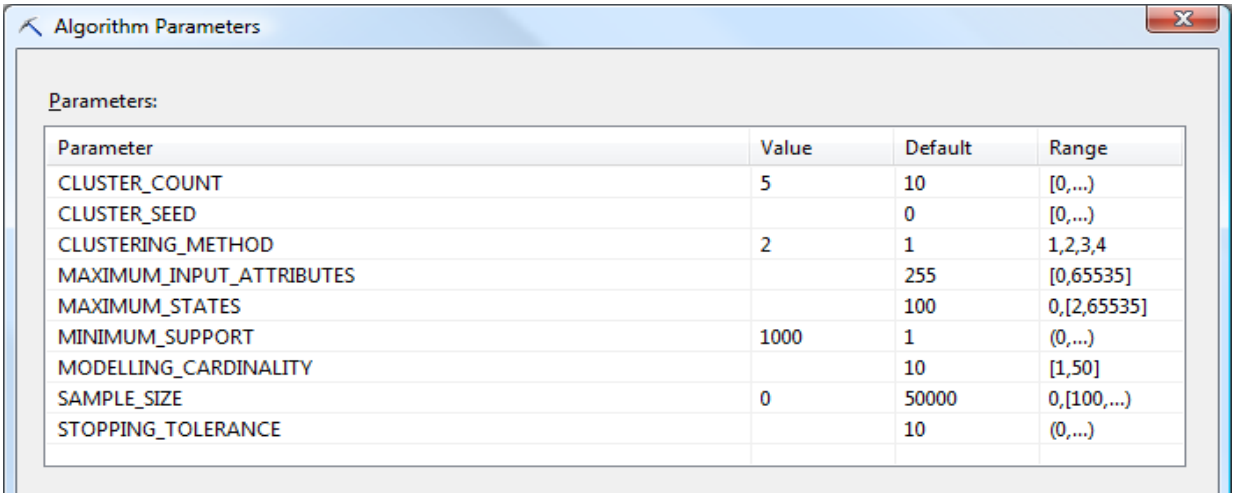
Microsoftov algoritem za razvrščanje v skupine sem že podrobno opisal v točki 4.9.1 zato bom tu samo povzel izbrane nastavitve.

Med štirimi možnimi tehnikami za razvrščanje v skupine sem uporabil Non-scalable EM metodo. EM metodo sem izbral zaradi njenih prednosti pred K-means metodo v teoriji in pa tudi na podlagi praktičnega preizkusa. EM metoda je v primeru Merkurjevih podatkov ustvarila skupine, ki jih je bilo mogoče razložiti, kar pa ne morem trditi za K-means metodo. Non-scalable možnost je sicer bolj zahtevna za strojno opremo, ki jo uporabljamo, vendar sem v postopku priprave podatkov zmanjšal količino le-teh na raven, ki dovoljuje njeno uporabo.

Določil sem tudi število skupin in sicer 5. Po mnogih poskusih sem opazil, da tako dobim najbolj homogene skupine.

Mejo, ki predstavlja minimalno podporo sem nastavil na 1000. Ker namen razvrščanja v skupine ni odkrivanje osamelcev (angl. *outlier*), je to razumno.

Slika 18: Nastavitve parametrov algoritma

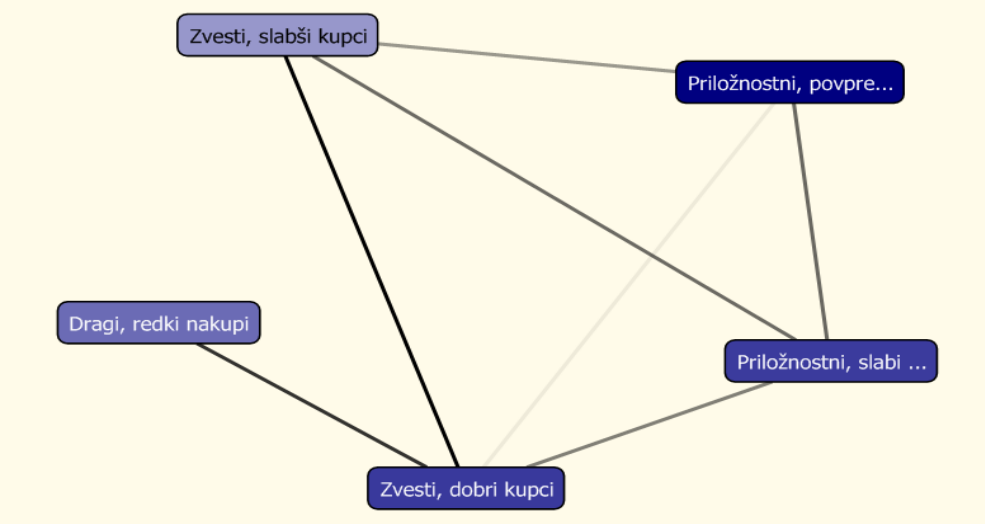


Parameter	Value	Default	Range
CLUSTER_COUNT	5	10	[0,...)
CLUSTER_SEED		0	[0,...)
CLUSTERING_METHOD	2	1	1,2,3,4
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_STATES		100	0,[2,65535]
MINIMUM_SUPPORT	1000	1	(0,...)
MODELLING_CARDINALITY		10	[1,50]
SAMPLE_SIZE	0	50000	0,[100,...)
STOPPING_TOLERANCE		10	(0,...)

7.8 Analiza rezultatov

Rezultat razvrščanja v skupine je prikazana na spodnji sliki. Skupine sem preimenoval glede na njihove najbolj izrazite lastnosti. Za slabe kupce sem štel tiste, ki potrošijo malo za dobre pa tiste, ki potrošijo veliko.

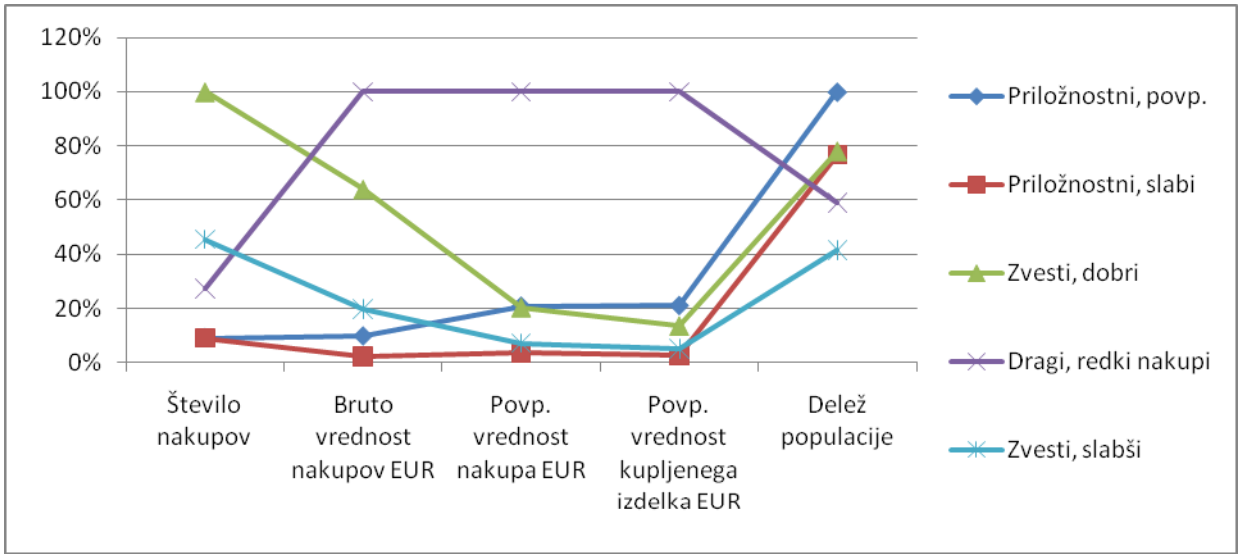
Slika 19: Grafični prikaz razvrščanja v skupine



V grobem lahko obravnavano populacijo kupcev razdelimo na dva dela in sicer na priložnostne in zveste kupce. Naprej lahko razdelimo skupine na podlagi koliko denarja porabijo letno za vse nakupe in koliko denarja porabijo za posamezen nakup.

Povprečja vrednosti atributov skupin so prikazana na sliki 20. Tu lahko hitro vidimo, da imajo priložnostni kupci prevladujoč delež in da so v povprečju zvesti kupci bolj donosni kupci, saj kupujejo več in dražje izdelke.

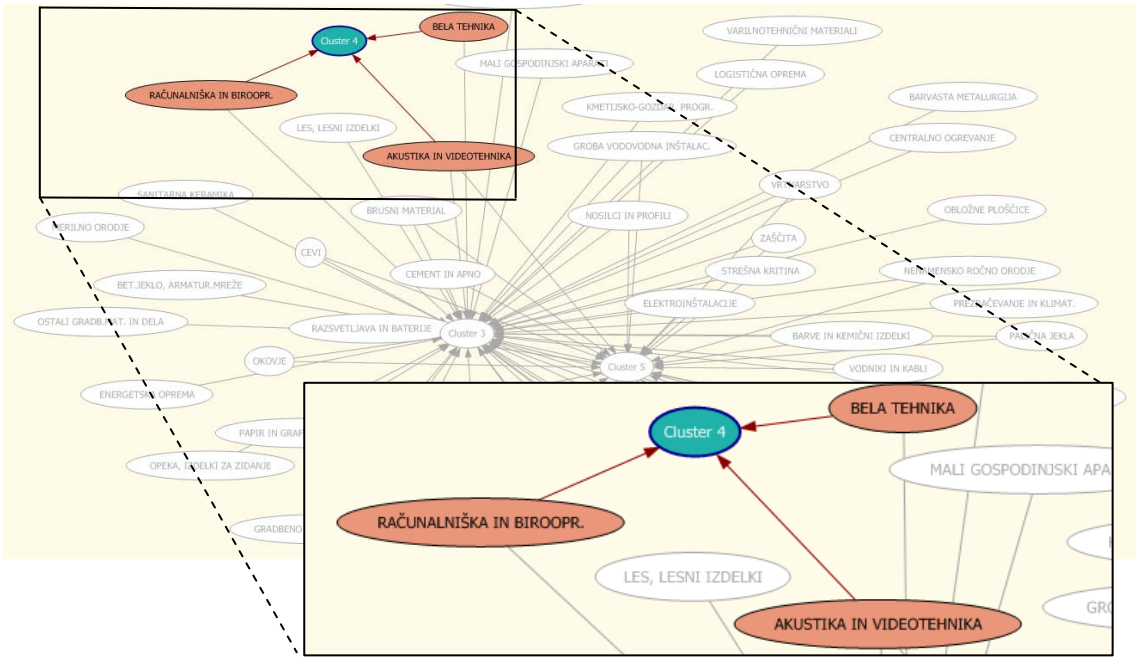
Slika 20: Povprečne vrednosti atributov znotraj skupin



Opazimo lahko tudi, da skupina kupcev poimenovana »Dragi, redki nakupi« zelo izstopa. Ta skupina je zanimiva tudi s tržnega vidika, saj je najbolj dobičkonosna. Smiselno bi bilo bolj intenzivno in prilagojeno trženje za to skupino.

Da bi bolj še poglobili naše razumevanje dobljenih skupin si lahko pomagamo z analizo nakupne košarice glede na skupino. Povedano drugače, zanima nas, ali lahko določene skupine izdelkov napovejo skupino kupcev. Grafični prikaz rezultata analize nakupne košarice je prikazan na spodnji sliki. Tehnika nakupne košarice oziroma asociativnih pravil, ki je bila opisana v točki 4.8.

Slika 21: Analiza nakupne košarice.



Če združimo in povzamemo vse kar smo spoznali o skupinah bi lahko rekli sledeče:

- Skupina »Zvesti, dobri« predstavlja kupce, ki so zvesti in kupujejo izdelke višje vrednosti. Na podlagi analize nakupne košarice vidimo, da to skupino najbolje napovejo skupine izdelkov kot so *nerjavni program, tv pločevina in trakovi, suho montažni gradbeni elementi, varilnotehnični materiali, barvasta metalurgija, cevi, vijaki, betonsko jeklo, armaturne mreže* ipd. Zaključimo lahko, da gre za profesionalce oz. obrtnike.
- Skupina »Zvesti, slabši kupci« predstavljajo skupino kupcev, ki kupuje manj pogosto kot obrtniki, a jih lahko še vedno štejemo med zveste kupce. Na podlagi analize nakupne košarice je velika verjetnost, da gre za domače mojstre, vrtičkarje in gospodinje. To skupino najbolje napovedujejo naslednje skupine izdelkov: *vrtnarstvo, strešna kritina, barve in kemični izdelki, gospodinjski pripomočki, razsvetljava in baterije, elektroinštalacije, vodniki in kabli, vrtno orodje, papir in grafični material* ipd.
- Skupina »Dragi, redki nakupi« so kupci, ki kupujejo najdražje izdelke, predvsem s področja *bele tehnike, računalniške opreme* in *akustike in videotehnike*. Za to skupino je še značilno, da večino nakupov opravi v soboto. Kupci v tej skupini so verjetno zaposleni na višje plačanih mestih, ki med tednom nimajo časa za nakupovanje.

Sklep

Podjetja v tujini kot tudi pri nas delujejo v vedno bolj konkurenčnem okolju. Z odpravo meja, liberalizacije trgov, odpravo zaščit domačih podjetij in drugimi znaki globalizacije bodo podjetja prisiljena bolj učinkovito izkoriščati svoje resurse in sprejemati informirane odločitve, ki temeljijo na analizi podatkov. Eno od teh orodij je podatkovno rudarjenje.

Slovenska podjetja začenjajo počasi posnemati svoje tekmece iz bolj razvitih ekonomij pri uvajanju podatkovnega rudarjenja v svoje poslovanje. Največji problem s katerim se soočajo, je pridobivanje kvalitetnih podatkov, se pravi podatkov, ki so konsistentni in vsebujejo relevantne parametre. To bi izpostavil kot ključni problem pri uvajanju projektov podatkovnega rudarjenja v slovenskih podjetjih – pomanjkanje kvalitetnih podatkov. Čeprav so orodja in tehnike uporabljene v podatkovnem rudarjenju veliko bolj robustne kot na primer statistične metode in dovoljujejo določeno mero nekonsistentnosti v podatkih, še vedno velja pravilo »smeti notri, smeti ven« . Če želijo podjetja imeti kvalitetne informacije o svojih kupcih, organizaciji, proizvodnih procesih in podobno, bodo morala misliti na to že v samem snovanju svojih informacijskih sistemov.

V diplomski nalogi sem opredelil podatkovno rudarjenje in najbolj poznane tehnike ter orodja, ki so trenutno na voljo. Podrobno sem predstavil programsko rešitev podjetja Microsoft, SQL Server 2008 za podatkovno rudarjenje in delovanje algoritma za razvrščanje v skupine. V drugem delu sem podal primer praktične izvedbe podatkovnega rudarjenja od faze zbiranja podatkov do končne analize.

Podatkovno rudarjenje je močno orodje, ki podjetju med drugim omogoča boljše razumevanje kdo so njihove stranke in kakšne so njihove želje ter navade. To koristi tako kupcu kot podjetju. Kupec dobi prilagojeno ponudbo, ki ga dejansko zanima in mu olajša iskanje. Na drugi strani pa

ima podjetje manjše stroške z reklamo, večjo prodajo in kar danes postaja najpomembnejše – zadovoljno stranko.

Podatki, ki sem jih uporabil za podatkovno rudarjenje v praktičnem delu, so bili žal zelo skromni kar se tiče informacij o kupcih. Poskušal sem izpeljati čim več novih atributov kot so različna povprečja in celo dodajal nove, vendar ti atributi niso nosili nobenih novih informacij, ki bi pomagale definirati skupine. Menim, da bi že samo spol kupcev močno izboljšal kvaliteto dobljenega modela. Kljub temu sem uspel identificirati nekaj zanimivih skupin, ki bi bile lahko dejansko komercialno zanimive za podjetje Merkur in bi njihovim kupcem nudile določeno dodano vrednost. Prav dejstvo, da lahko kljub malo informacijam z uporabo tehnik podatkovnega rudarjenja najdemo dokaj homogene skupine kupcev, priča o potencialu te tehnologije.

Tekom izvedbe praktičnega dela diplomske naloge sem opazil, da sta za uspešno izvedbo podatkovnega rudarjenja poleg pravih podatkov ključna dva elementa. Najprej si moramo zastaviti pravo vprašanje. Pravo vprašanje v smislu, da nam bo odgovor nanj koristil in da smo nanj sploh zmožni odgovoriti. Zavedati se moramo omejitev, ki jih predstavljajo podatki, ki so nam na voljo. Drugi ključen element oziroma faza podatkovnega rudarjenja pa je priprava podatkov. Pravilna priprava vzame večino časa in je tudi najbolj pomembna, saj predstavlja temelje na katerih gradimo modele. V tej fazi moramo najti ravnotežje med tem, da podatke transformiramo na način, ki bodo ustrezale uporabljeni tehniki in med tem, da ne popačimo morebitnih odvisnosti med posameznimi spremenljivkami.

Trendi v podatkovnem rudarjenju gredo v smeri raziskovanja njegove uporabe na novih področjih, izboljšave prilagodljivosti in interaktivnosti, integraciji s podatkovnimi skladišči in podatkovnimi bazami, standardizaciji programskih jezikov za podatkovno rudarjenje, odkrivanju novih načinov vizualizacij ter novih metod za obdelavo kompleksnih oblik podatkov.

Literatura in viri

1. Auerbach, K. (marec 2004). *Winter Corporation articles*. Najdeno 6. oktobra 2008 na spletnem naslovu http://wintercorp.com/rwintercolumns/archives.html

2. Berry, M., & Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* New York: John Wiley & Sons, Inc.

3. Berry, M., & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York: John Wiley & Sons, Inc.

4. Berry, M., & Linoff, G. (2007). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* New York: John Wiley & Sons, Inc.

5. Berson, A., Smith, S., & Thearling, K. (2009). *Customer Acquisition and Data Mining.* Najdeno 15. februarja 2009 na spletnem naslovu http://www.thearling.com/text/chapter10/chapter10.htm

6. Brachman, R., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communications of the ACM , 39*, 42-48.

7. College of Saint Benedict and Saint John's University. *Band Structure.* Najdeno 15. januarja 2008 na spletnem naslovu http://www.physics.csbsju.edu/QM/i/band.r.lattice.kx.gif

8. Dumas, M., Aldred, L., Governatori, G., Hofstede, A., & Russell, N. (2002, 7. maj). *A Probabilistic Approach to Automated Bidding in Alternative Auctions.* Najdeno 3. oktobra 2009 na spletnem naslovu http://www2002.org/CDROM/refereed/279/

9. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine , 13* (3), 37-54.

10. *Goddard Space Flight Cente*. (2009). Najdeno 26. januarja 2009 na spletnem naslovu http://lupus.gsfc.nasa.gov/

11. Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Technigues* (2nd ed.). Amsterdam: Morhan kaufman.

12. Hartigan, J. A. (1975). *Clustering Algorithms (Probability & Mathematical Statistics).* New York City: Wiley.

13. International Society for Bayesian Analysis. (2009). *Who was the Reverend Thomas Bayes*. Najdeno 20. junija 2009 na spletnem naslovu http://www.bayesian.org/resources/bayes.html

14. *Internet Archive*. (2009). Najdeno 20. februarja 2009 na spletnem naslovu http://www.archive.org/index.php

15. KDnuggets. (Maj 2006). *Data Mining (Analytic) Tools*. Najdeno10. januarja 2006 na spletnem naslovu http://www.kdnuggets.com/polls/2006/data_mining_analytic_tools.htm

16. Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining: Introduction to principales and algorithms.* West Sussex: Horwood publishing limited.

17. Lukawiecki, R. (22. maj 2008a). *Technet Spotlight On Demand Video: Introduction to Data Mining.* Najdeno 11. aprila 2009 na spletnem naslovu http://technet.microsoft.com

18. Lukawiecki, R. (22. maj 2008b). *Technet Spotlight On Demand Video: Using Data Mining in Your IT Systems (Part 1).* Najdeno 11. aprila 2009 na na spletnem naslovu http://technet.microsoft.com

19. Lukawiecki, R. (22. maj 2008c). *Technet Spotlight On Demand Video: Using Data Mining in Your IT Systems (Part 2).* Najdeno 11. aprila 2009 na na spletnem naslovu http://technet.microsoft.com

20. Lukawiecki, R. (22. maj 2008d). *Technet Spotlight On Demand Video: Working with Data Mining.* Najdeno 11. aprila 2009 na na spletnem naslovu http://technet.microsoft.com

21. Lyman, P., & Varian, H. R. (27. oktober 2003). *How Much Information? 2003.* Najdeno 15. decembra 2008 na spletnem naslovu http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

22. MacLennan, J., Tang, Z., & Crivat, B. (2009). *Data Mining with Microsoft SQL Server 2008.* Indianapolis: Wiley Publishing, Inc.

23. Merkur d.d. (2008). Konsolidirano letno poročilo Merkur Group in letno poročilo družb Merkur d.d., in Mersteel, d.o.o., za leto 2008. Ljubljana: Merkur d. d.

24. *Merkur - Novinarsko središče*. Najdeno 15. maja 2009 na spletnem naslovu http://www.merkur.eu/slo/novinarsko_sredisce0/index.html?geoip=0\\\\

25. Microsoft Corporation. (2009a). *SQL Server 2008 Books Online*. Najedno 3. marca 2009 na spletnem naslovu http://msdn.microsoft.com/en-us/library/bb510517%28v=SQL.100%29.aspx

26. Microsoft Corporation. (2009b). *SQL Server 2008 Books Online*. Najdeno 7. aprila 2009 na spletnem naslovu http://msdn.microsoft.com/en-us/library/ms174879.aspx

27. Miller, G. J., Brautigam, D., & Gerlach, S. V. (2006). *Business Intelligence Competency Centers: A Team Approach to Maximizing Competitive Advantage.* New Jersey: John Wiley & Sons, Inc.

28. Nisbet, R. A. (marec 2004). *How to Choose a Data Mining Suite.* Prevzeto 20. januar 2009 iz Information management: http://www.information-management.com/specialreports/20040323/1000465-1.html

29. Payle, D. (1999). *Data preparation for data mining.* San Francisco: Morgan Kaufman.

30. Rud, O. P. (2001). *Data mining cookbook: Modeling data for marketing, risk, and customer relationship management.* New York: John Wiley & Sons, Inc.

31. Shapiro, G. P. (2000). Knowledge Discovery in Databases: 10 years after. *SIGKDD Explorations,* februar, 2000, Vol. 1, No. 2 , 59-61.

32. *SQL Server Overview* . (April 2009). Najdeno 15. junija 2009 na spletnem naslovu http://msdn.microsoft.com/en-us/library/ms166352%28SQL.100%29.aspx

33. Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining* (2nd ed.). Boston: Addison-Wesley.

34. *The CRISP-DM consortium*. (2007). Najdeno 6. septembra 2008 na spletnem naslovu http://www.crisp-dm.org/Process/index.htm

35. Support vector machine (b.l.) V *Wikipedia, the free encyclopedia*. Najdeno 20. novembra 2009 na spletnei strani http://en.wikipedia.org/wiki/Support_vector_machine.

36. Winter Corporation. (2005, 14. september). *2005 Winter TopTen Award Winners.* Najdeno 20. decembra 2008 na spletnem naslovu http://www.wintercorp.com/VLDB/2005_TopTen_Survey/2005TopTenWinners.pdf.

37. Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed.). Amsterdam: Morgan Kaufman.