UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

DIPLOMSKO DELO

UPORABA PODATKOVNEGA RUDARJENJA PRI ODKRIVANJU NEZAŽELENE ELEKTRONSKE POŠTE

Ljubljana, junij 2003 BLAŽ KONIČ

IZJAVA

Študent BLAŽ KONIČ izjavljam, da sem avtor tega diplomskega dela, ki sem ga napisal pod mentorstvom DR. JURIJA JAKLIČA in dovolim objavo diplomskega dela na fakultetnih spletnih straneh.

V Ljubljani, dne 9. junija 2003 Podpis:_____

KAZALO

1. UVOD

Izreden vpliv, ki ga imajo na sodobno družbo informacijske tehnologije, se med drugim odraža tudi v vedno večji količini hranjenih podatkov. Še pred dvemi desetletji je megabyte podatkov veljal za nekaj povsem nepredstavljivega. Ustanovitelj Microsofta William H. Gates je takrat ob predstavitvi prvega osebnega računalnika dejal: »640 kilobytev mora zadostovati vsakomur[1]« (History of Computing Industrial Era 1981, 2002). Kakšna zmota! Danes že povprečen dokument, elektronsko sporočilo ali slika neredko presegajo to mejo. Še posebej pa je porast količine hranjenih podatkov občuten v podatkovnih zbirkah podjetij. To pravzaprav niti ni presenetljivo, saj gre za logično posledico avtomatiziranja poslovnih procesov. Zamislimo si grosistično podjetje in preprost izdelek, ki vsebuje črtno kodo. Le-ta je prebrana in zabeležena na več mestih: npr. pri dobavi, kontroli, skladiščenju in odpošiljanju izdelka. Če podjetje dnevno proda nekaj deset- ali sto tisoč izdelkov, se količina hranjenih podatkov lahko izredno hitro poveča. Do podobnega sklepa lahko pridemo, če si zamislimo število ljudi, ki uporabljajo plačilne kartice. Pri tem imejmo v mislih, da banke zabeležijo podatke o vsaki posamezni transakciji. Vsak za sebe ve, kolikokrat mesečno uporabi plačilno kartico. Pomnožimo to število s številom vseh uporabnikov in dobili bomo nepredstavljivo veliko količino podatkov.

Tako zbrani podatki nudijo podjetju obilico potencialno koristnih informacij, ki lahko privedejo k uspešnejšemu poslovanju in učinkovitejšim poslovnim odločitvam v podjetju. Seveda pa je potrebno te informacije iz podatkov nekako izluščiti. Tu nam na pomoč priskočijo orodja in metode podatkovnega rudarjenja.

Ni pa to edino področje, kjer je možno uporabljati podatkovno rudarjenje. Za svojo diplomsko nalogo sem izbral njegovo uporabo na področju, ki je povprečnemu uporabniku vsekakor bolj poznano: nezaželena elektronska pošta oz. e-smetenje (ang. spam). Sama elektronska pošta je danes ena izmed najbolj uporabljanih internetnih storitev. Množična uporaba pa pogosto prinese s seboj razne oblike zlorab in e-pošta ni nobena izjema. Kdo izmed nas se še ni srečal z nezaželenimi in nenaročenimi sporočili, s smetenjem najrazličnejših ponudnikov pornografskih vsebin, organizatorjev denarnih verig ali pa vsiljivih ponudnikov česarkoli že?

[1] Mišljena je bila količina delovnega pomnilnika (RAM).

V mojem diplomskem delu bom poskušal povezati omenjeni področji: podatkovno rudarjenje in e-smetenje. Najprej bom podrobneje predstavil področje podatkovnega rudarjenja, različne možnosti uporabe in najpogosteje uporabljene tehnike. Posvetil se bom tudi smetenju, še posebej njegovim ekonomskim posledicam in škodi, ki jo povzroča. Predstavil bom načine boja proti smetenju in nazadnje s pomočjo orodij za podatkovno rudarjenje izdelal model za odkrivanje e-smeti. Za izdelavo modela bom analiziral preteklo e-pošto, iz te osnove »izluščil« pravila za določanje nezaželene pošte in ta pravila vključil v model, ki bo v prihodnje sposoben med prihajajočo e-pošto odkriti nezaželena sporočila. Model bom nazadnje tudi praktično preizkusil na množici še nerazvrščenih e-sporočil.

Množico razvrščenih (preteklih) in nerazvrščenih e-sporočil si bom izposodil iz naloge na tradicionalnem tekmovanju iz podatkovnega rudarjenja »Data Mining Cup 2003« oz. pri organizatorju tega tekmovanja, nemškem podjetju Prudsys.

2. PODATKOVNO RUDARJENJE

2.1 KAJ JE PODATKOVNO RUDARJENJE?

Obstaja več definicij tega pojma. Berry in Linoff pravita, da gre za proces avtomatskega ali polavtomatskega analiziranja velikih količin podatkov z namenom odkriti zanimive in uporabne vzorce ter pravila (Berry, Linoff, 2000, str. 7). Pri tem opozarjata, da nas ne sme zavesti besedica »avtomatsko«. Podatkovno rudarjenje (PR) namreč ne predstavlja črne škatlice, ki jo lahko kupimo na trgu, ampak gre bolj za disciplino, ki jo je potrebno zaobvladati. Podjetje SAS Institute na drugi strani pod PR razume »napredne metode za odkrivanje in modeliranje povezav v velikih količinah podatkov« (Data Mining Using Enterprise Miner Software, 2000, str. 5). Poenostavljeno bi lahko dejali, da PR zajema pridobivanje koristnih informacij iz velikih količin podatkov. Do teh informacij pridemo z »rudarjenjem« (kopanjem) po velikih količinah podatkov z uporabo umetne inteligence ter statističnih in matematičnih metod. Največja vrednost teh informacij je, da jih lahko uporabimo pri gradnji napovedovalnih modelov, t.j. modelov, ki bodo znali vnaprej napovedovati določene vrednosti v podatkih.

Še beseda ali dve o prevodu izvirnega angleškega izraza: podatkovno rudarjenje je prevod angleške besedne zveze »Data Mining«. Poleg omenjenega prevoda v slovenščini poznamo tudi izraz »izkopavanje podatkov«. Katerikoli izraz že uporabljamo, v vseh primerih govorimo o isti stvari.

2.2 UPORABLJENA TERMINOLOGIJA

Ko govorimo o podatkovnem rudarjenju, se ne moremo ogniti uporabi dveh različnih terminologij, ki se v precejšnji meri prekrivata. Gre na eni strani za statistično terminologijo in na drugi strani za izraze, udomačene na področju podatkovnih baz. Obe področji pa sta pomembni za PR: podatki, iz katerih želimo izluščiti nove informacije, so fizično shranjeni v podatkovnih bazah. Od tam se prenesejo v orodja za PR, ki temeljijo (tudi) na statističnih postopkih, in v večini primerov uporabljajo statistično terminologijo. V mojem diplomskem delu bosta uporabljeni obe terminologiji, odvisno pač od konteksta. V tabeli 1 so po vrsticah predstavljeni pojmi, ki imajo na različnih področjih uporabe različna imena, vendar (vsaj s stališča tega diplomskega dela) enak pomen.

Tabela 1: Sopomenke po različnih področjih uporabe

STATISTIKA	BAZE PODATKOV	SPLOŠNA RABA
opazovana enota	zapis, entiteta	vrstica
spremenljivka	atribut, polje	stolpec

Opazovano enoto razumemo kot sestavni del populacije, pri čemer ne gre za populacijo v dobesednem pomenu, ampak je ta pojem splošnejši in zajema npr. vsa podjetja v regiji, vse v istem mesecu izdelane avtomobile ali v mojem primeru – vsa e-sporočila, ki so prispela v podjetje. Za navedene primer so opazovane enote posamezno podjetje, posamezni avtomobil in posamezno e-sporočilo. **Spremenljivka** predstavlja posamezno lastnost opazovane enote. Pri avtomobilu bi lahko uporabili naslednje spremenljivke: *barva*, *moč motorja*, *število sedežev* ipd. (Blejec et al., 2003, str. 3).

Medtem, ko pri podatkovnih bazah ločimo precej različnih tipov atributov (npr. cela števila, cela števila s predznakom, realna števila, realna števila z dvojno natančnostjo…), orodja za PR uporabljajo dva glavna tipa spremenljivk: intervalni in kategorični (opisni) (Data Mining Using Enterprise Miner Software, 2000).

Intervalna spremenljivka je spremenljivka številskega tipa, pri kateri je smiselno izračunati njeno povprečno vrednost v več opazovanih enotah – npr. povprečna višina plače ali povprečna temperatura.

Kategorične spremenljivke pa so tiste, pri katerih izračunavanje povprečja ni smiselno, npr. *velikost* (velika, srednja, majhna) ali *barva* (rdeča, rumena, zelena). Ta

tip spremenljivk lahko razdelimo naprej še na dva podtipa. Če vrednosti spremenljivke lahko smiselno uredimo po vrsti (npr. *velikost*), govorimo o **ordinalnih** spremenljivkah. Nasprotno pri **nominalnih** spremenljivkah takšno razvrščanje ni smiselno (npr. *barva*). Opozorim naj še, da tudi kategorične spremenljivke lahko vsebujejo številske vrednosti. Primer sta *poštna številka* in *telefonska številka*. Izračunavanje povprečne poštne številke ali povprečne telefonske številke namreč ne bi imelo nobenega smisla.

V tem diplomskem delu se bomo srečali tudi z dvema posebnima vrstama kategoričnih spremenljivk, ki ju uporablja programski paket SAS Enterprise Miner: unarna (ang. unary) in binarna (ang. binary). **Unarna** spremenljivka ima v vsaki opazovani enoti v populaciji enako vrednost (primer je spremenljivka *rojstni planet*, ki bi se lahko uporabljala pri popisu prebivalstva) in je kot taka za PR bolj ali manj neuporabna. Lahko pa se seveda pojavlja v podatkih. **Binarna** spremenljivka pa je tista, ki v celotni populaciji zavzema natanko dve vrednosti (npr. *spol*).

Natančno lahko opredelimo še dva, za podatkovno rudarjenje pomembna pojma (Berson, Smith, Thearling, 2000, str. 110):

- **model** je opis izvirnih (zgodovinskih) podatkov, ki ga je mogoče aplicirati na nove podatke z namenom napovedovanja manjkajočih oz. pričakovanih vrednosti
- **vzorec** je dogodek ali kombinacija dogodkov v podatkih, ki se pojavlja bolj pogosto, kot bi bilo pričakovati.

2.3 KAKO NAM PODATKOVNO RUDARJENJE LAHKO POMAGA?

Podatkovnega rudarjenja ne smemo jemati kot čarobno paličico, ki bo bdela nad podatkovnimi bazami, opazovala, kaj se v njih dogaja in uporabnika obveščala, ko najde kak zanimiv vzorec. Prav tako PR ni zamenjava za analitika: še vedno je potrebno poznati probleme oz. področje delovanja / poslovanja, razumeti podatke in poznati analitične metode. Orodja za PR kot taka so zgolj v pomoč analitiku – odkrivajo zanimive vzorce in pravila, ki se pojavljajo v podatkih. Ne povejo pa, kakšna je za podjetje vrednost teh odkritih informacij. Za takšne ocene so še vedno potrebni ljudje.

Pri PR govorimo o dveh različnih pristopih. Prvi pristop je **usmerjeno** podatkovno rudarjenje, kjer vemo, kaj točno iščemo. Najpogostejša oblika tega pristopa je

izdelava napovedovalnih modelov. Ciljna (odvisna) spremenljivka je vnaprej določena in model mora kar se da dobro prepoznati povezave med njo in ostalimi podatki. Nasprotno pri **neusmerjenem** PR naš cilj ni določen. V tem primeru PR v podatkih odkrije vzorce in nam samim prepusti odločitev o njihovi (ne)pomembnosti. Sledijo aktivnosti PR, pri čemer so prve tri primeri usmerjenega PR, zadnje tri pa spadajo pod neusmerjeno PR (Berry, Linoff, 2000, str. 8).

Klasifikacija
Pod klasifikacijo razumemo analiziranje lastnosti opazovane enote, na osnovi česar opazovano enoto razvrstimo v enega izmed vnaprej definiranih razredov.

Primer: Banka prejme prošnjo za odobritev kredita svojemu komitentu. Na osnovi podatkov o njegovem preteklem poslovanju določi njegovo boniteto: NIZKA (visoko tveganje), SREDNJA (zmerno tveganje) ali VISOKA (nizko tveganje).

Ocenjevanje
Če so pri klasifikaciji rezultati kategorične spremenljivke (DA/NE, NIZKA/SREDNJA/VISOKA…), imamo pri ocenjevanju opravka s spremenljivkami intervalnega tipa. Na osnovi vhodnih podatkov ocenimo vrednost neke spremenljivke, kot npr. dohodek. Pogosto je ocenjevanje uporabljeno kot osnova za klasifikacijo. Pri tem za vrednosti ocenjevane spremenljivke določimo intervale, ki pripadajo posameznemu razredu.

Primer: Podjetje se odloči za oglaševalsko kampanjo, v kateri bo potencialnim kupcem poslalo nov katalog. Na osnovi podatkov o kupcih podjetje oceni, kakšna je verjetnost, da se po kupec odzval na akcijo. Rezultat je ocenjena verjetnost nakupa na intervalu med 0 in 1. Podjetje se nato odloči, kako visoko verjetnost nakupa bo zahtevalo za prejemnike kataloga in vsi potencialni kupci, katerih verjetnost nakupa je enaka ali večja, prejmejo katalog.

Napovedovanje
Napovedovanja mogoče niti ne bi bilo potrebno omenjati posebej, saj ga je moč razumeti kot klasifikacijo ali kot ocenjevanje. Kljub temu Berry in Linoff razlikujeta napovedovanje od omenjenih dveh aktivnosti. Če uporabimo PR za klasifikacijo telefonskih priključkov kot primarno namenjenih dostopu do interneta ali govorni komunikaciji, ne pričakujemo, da bomo kasneje lahko preverili pravilnost klasifikacije. Ta je lahko pravilna ali napačna – vzrok negotovosti je samo v nepopolnem znanju. Dejansko so se na klasifikacijo vezani dogodki že zgodili

(telefonska linija se uporablja za povezavo s ponudnikom dostopa do interneta ali pa se ne uporablja).

Pri napovedovanju pa se opazovane enote klasificirajo na osnovi pričakovanega dogajanja v prihodnosti ali na osnovi pričakovanih prihodnjih vrednosti spremenljivk. Pravilnost klasifikacije lahko preverimo samo tako, da počakamo in vidimo, kaj se je zares zgodilo.

Kot primer lahko navedemo napovedovanje, kateri uporabniki bodo v naslednje pol leta prekinili naročnino na določeno storitev.

Opisovanje in vizualizacija[2]

Včasih nam podatkovno rudarjenje lahko pomaga že s tem, da opiše, »kaj se dogaja« v veliki in zapleteni podatkovni bazi in s tem prispeva k razumevanju celotne situacije. Dober opis dogajanja pogosto predoči, kje iskati razlage za takšno dogajanje. Če je le možno, je zaželena tudi vizualna predstavitev, saj tudi tu velja pregovor, da »slika pove več kot tisoč besed«.

Primer: Relativno enostavna ugotovitev, da v ZDA ženske podpirajo Demokratsko stranko v večji meri kot moški, je povzročila izredno povečanje zanimanja in novih študij s strani novinarjev, sociologov, ekonomistov in seveda tudi politikov.

Asociacije

Pri iskanju asociacij gre za ugotavljanje, katere stvari spadajo skupaj. Preprost primer je ugotavljanje artiklov, ki se pogosto znajdejo skupaj v nakupovalni košarici. Trgovske verige ta podatek lahko uporabijo za ustrezno razvrščanje artiklov na prodajne police.

Zadnja aktivnost, jo lahko uvrstimo pod neusmerjeno podatkovno rudarjenje, je **razvrščanje v skupine**. Ker se zanjo uporablja ena sama točno določena tehnika, sem namesto aktivnosti podrobneje predstavil kar tehniko samo (glej poglavje 2.6.2).

[2] V primerjavi s prej omenjenimi področji gre tu za relativno enostavne metode opisne statistike (povprečja, deleži), ki nam pogosto služijo za prvi vpogled v podatke.

2.4 PODROČJA UPORABE

Podatkovno rudarjenje ni specifično povezano z nobenim področjem, ampak je uporabno na vseh področjih, kjer obstaja velika količina podatkov in kjer se iz podatkov splača naučiti nekaj novega. Ne bi nas smelo presenetiti, da npr. vojaške obveščevalne službe uporabljajo tehnike PR za obdelovanje satelitskih posnetkov z namenom klasifikacije objektov na zemeljskem površju – ali gre za tanke (vojaški cilji) ali za traktorje (civilni cilji). Seveda si od takšnega prepoznavanja generali obetajo koristi. Napad na civilne cilje bi verjetno izzval vsaj močno negodovanje, če že ne protivojne demonstracije. Poleg tega bi trpela učinkovitost napadov, saj je s tanki oborožen sovražnik bistveno nevarnejši od tistega, ki so mu ostali samo še traktorji.

Podobno pa velja za uporabo PR na **poslovnem področju**, le da je pravilo, kdaj se nekaj splača, strožje. Pridobivanje novih znanj se splača toliko časa, dokler je vrednost teh znanj višja od stroškov njihovega pridobivanja. V resnici je definicija še strožja: investicija v pridobivanje novega znanja mora zagotavljati višjo povrnitev vloženih sredstev (ang. Return On Investment, ROI), kot če bi ta sredstva investirali kam drugam (Berry, Linoff, 2000, str. 11). V tem smislu nam PR lahko pomaga na dva načina. Lahko pripomore k **višjemu dobičku**:

- z zmanjšanjem stroškov ali
- z zvišanjem prihodkov.

Obstaja pa še tretji vpliv: s pomočjo kateregakoli od omenjenih dveh mehanizmov lahko napove zvišanje dobička v prihodnosti, kar **zviša ceno delnic** podjetja.

Prvi možnost, da uporaba PR podjetju zniža stroške, se ponuja že na samem začetku življenjskega cikla izdelka: v fazi raziskav in razvoja. Tipičen primer je **farmacevtska industrija**, za katero so značilni izredno visoki stroški za raziskave in razvoj. Ti globalno na letnem nivoju znašajo preko 25 milijard USD (Chemical and Laboratory Supply Online – The Industry, 2003). Razvoj zdravil si poenostavljeno lahko predstavljamo kot lijak, v katerega širši del vstopa na milijone kemičnih spojin, na drugi strani pa se pojavi samo nekaj varnih in uporabnih zdravil. Da bi se potencialno zdravilo prebilo skozi celoten proces mora najprej biti sposobno se vezati na neko ciljno molekulo. Spojine, pri katerih je ta pogoj izpolnjen se imenujejo spojine vodnice. Te nadalje čaka vrsta preizkusov: dokazati morajo, da imajo želene učinke, telo jih mora biti sposobno absorbirati, preživeti morajo v sovražnem okolju živega organizma, ne smejo biti strupene in imeti morajo dokazljive učinke na živalih. Nazadnje mora spojina vodnica še skozi vrsto kliničnih testov, ki dokažejo njeno

varnost in učinkovitost na človeškem organizmu. Izredno malo spojin se prebije skozi vse teste in postane glavna učinkovina zdravila – približno ena od 10.000. Stroški za ta proces (»rešetanje«) lahko znašajo tudi do 800 milijonov USD[3] (The Pharmaceutical Landscape, 2003).

V tem primeru obstaja veliko vhodnih podatkov, na katerih lahko temeljijo napovedi. Dandanašnji je »rešetanje« zelo napredno: avtomatizirani sistemi iz znanih reagentov sestavljajo razne kombinacije spojin. Te spojine posebni roboti testirajo na snoveh, ki vsebujejo ciljne molekule. Roboti zaznajo, kdaj se spojina veže na ciljno molekulo in tiste, ki se vežejo samo na ciljne molekule, so verjetne spojine vodnice.

Tako avtomatizirano okolje ustvarja podatke, idealne za podatkovno rudarjenje – veliko vhodnih spremenljivk in preprosto ciljno spremenljivko binarnega tipa (DA / NE). Z uporabo tehnik PR lahko farmacevtska podjetja zožijo raziskave samo na spojine, ki bodo verjetno pripeljale do koristnih zdravil in s tem prihranijo velike vsote denarja.

Podatkovno rudarjenje se učinkovito uporablja tudi na področju **proizvodnje**. Veliko sodobnih proizvodnih linij (npr. proizvodna linija za pivo na Pivovarniški ulici v Ljubljani) je nadzorovanih z računalniško vodenimi metodami nadzora. Senzorji beležijo pritisk, temperaturo, vlažnost, hitrost, barvo in še mnogo ostalih spremenljivk, pomembnih za posamezni proizvodni proces. Računalniški programi spremljajo vrednosti, ki jih zaznavajo senzorji in prilagajajo proizvodne dejavnike spremembam teh vrednosti. Vse to z namenom, da proizvodnja ostane v okviru dovoljenih odstopanj. Podatki so ponovno zelo primerni za PR: ogromno vhodnih podatkov in rezultat, npr. DOBER ali SLAB izdelek. Na osnovi teh podatkov lahko izdelamo model, ki bo ciljal na DOBER izdelek, zanj določil dovoljena odstopanja in temu primerno prilagajal proizvodno linijo. V tem primeru se lahko bistveno znižajo proizvodni stroški, do katerih bi prišlo ob morebitni napaki na proizvodni liniji (zaustavitev proizvodnje, uničena serija izdelkov, ponovni zagon…).

Večina najuspešnejših uporab podatkovnega rudarjenja je znanih s področja **trženja**, pri čemer gre v večini primerov za analizo podatkovnih zbirk, ki vsebujejo informacije o strankah (ang. database marketing). Na tem področju se PR izkaže z obema dejavnikom zviševanja dobička – tako z nižanjem stroškov kot z višanjem prihodkov. PR tržnikom pomaga pri analizi podatkov o strankah in določanju, katera

[3] Mišljen je razvoj originalnih in ne generičnih zdravil.

stranka se bo verjetno odzvala na novo ponudbo. S tem podjetje zmanjša stroške za oglaševalsko kampanjo, saj jo usmeri zgolj na verjetne kupce. Na strani višanja prihodkov pa PR lahko pomaga pri identificiranju kupcev, ki bi verjetno posegli po luksuznih izdelkih ali storitvah (prestižni avtomobili, luksuzne počitnice, najdražja zavarovanja ipd.).

Podatkovno upravljanje igra pomembno vlogo tudi pri **upravljanju odnosov s strankami** (ang. Customer Relationship Management – CRM). Gre za pojem, ki uteleša večino tistega, kar razumemo pod trženjskim principom 1 : 1 (ang. one-to-one marketing). CRM je koncept, ki stranko postavlja v središče pozornosti. Med seboj povezuje vse aktivnosti podjetja, ki so usmerjene na stranko. Spremlja se njeno obnašanje in na osnovi informacij o stranki podjetje skuša izboljšati poslovanje tako z obstoječimi kot tudi z novimi strankami. CRM najprej zahteva poznavanje strank, kdo so, kaj hočejo in česa ne marajo. Šele na tej osnovi podjetje lahko gradi naprej, ustvarja lojalnost strank, prepoznava njihovo nezadovoljstvo še preden presedejo h konkurenci… (Berson, Smith, Thearling, 1999, str. 42). Vse te dejavnosti zahtevajo ustrezno podporo informacijske tehnologije, kjer pomembno vlogo igra PR. Šele z njim lahko podjetje ogromne količine podatkov o strankah prevede v neko razumljivo in uporabno sliko.

2.5 POSTOPEK PODATKOVNEGA RUDARJENJA

Postopek podatkovnega rudarjenja sestoji iz štirih procesov (Berry, Linoff, 2000, str. 43). Potrebno je:

- določiti probleme oz. področje, kjer analiza podatkov obeta koristi,
- z uporabo PR pretvoriti podatke v uporabne informacije,
- ukrepati na osnovi pridobljenih informacij ter
- ovrednotiti rezultate.

2.5.1 DOLOČANJE PROBLEMOV ALI PODROČIJ ZA ANALIZO

Pri določanju problemov in področij je pomembno, da analitiki razumejo potrebe podjetja. Nikakor nas ne sme zavesti občutek, da je za podatkovno rudarjenje dovolj samo tehnično poznavanje metod in algoritmov. Prav tako pomembno je poznavanje širše poslovne problematike. Eno brez drugega ne gre.

Primer: Operater mobilne telefonije opaža, da vse več njegovih uporabnikov prehaja h konkurenčnemu ponudniku. Zato želijo na osnovi preteklih podatkov za vsakega

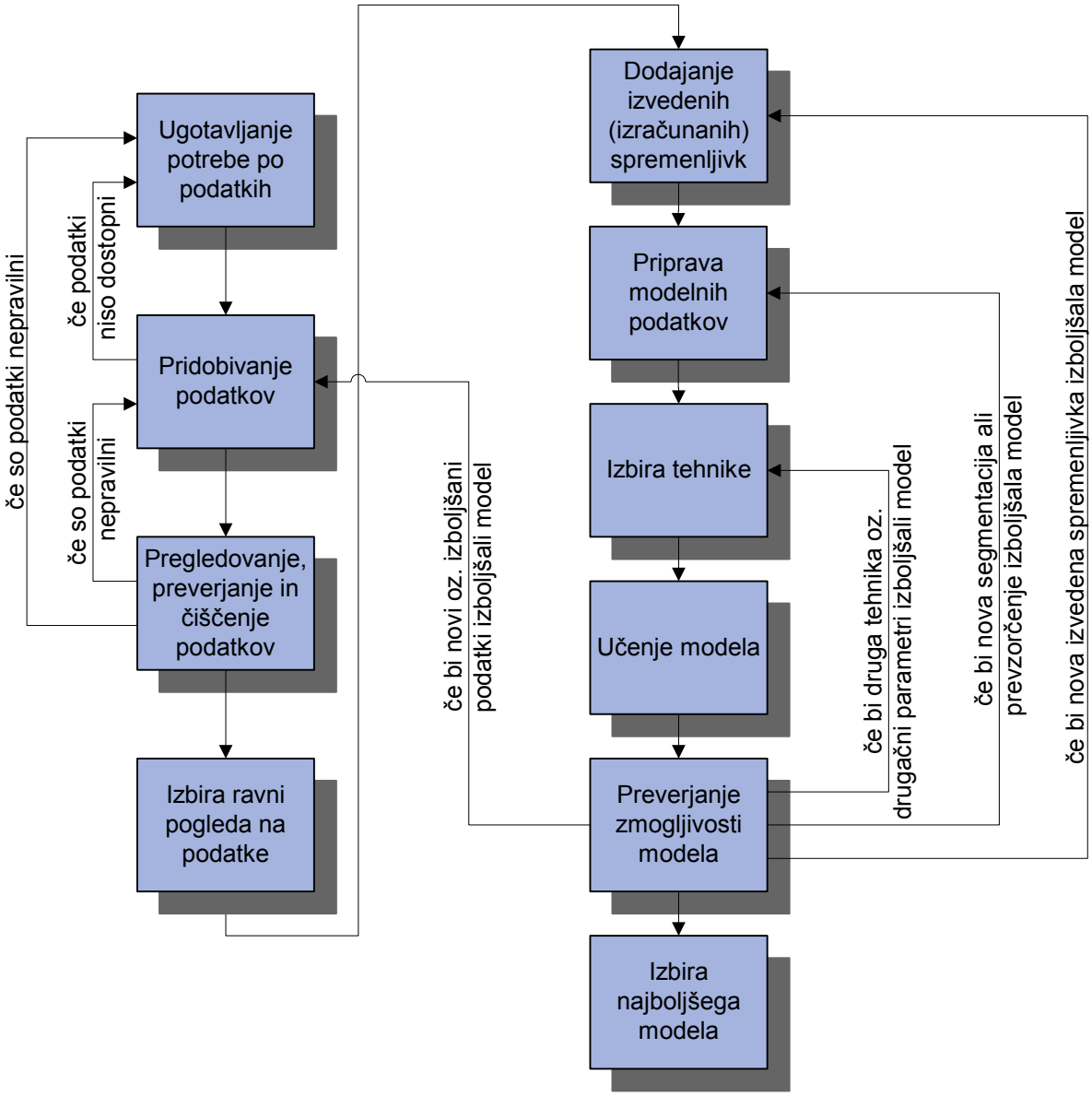
uporabnika oceniti verjetnost, da bo v naslednjih mesecih prestopil h konkurenci. Uporabnikom z največjo verjetnostjo odhoda oddelek za trženje nameni posebno tržno intervencijo, ki naj bi čim več uporabnikov prepričala, da ostanejo pri trenutnem operaterju. Rezultati te akcije pa so slabi. Izkaže se, da je podjetje želelo obdržati stranke, ki ne glede na stroške veliko uporabljajo njegove storitve in mu prinašajo največji delež prihodkov. Med uporabniki z visoko verjetnostjo odhoda pa je bilo takšnih manj kot 10%. Večina je bila »slabih« uporabnikov, ki operaterju ne prinašajo prihodkov in so sploh nagnjeni k menjavi operaterjev – pogosto prestopijo k tistemu, čigar ponudba je v danem trenutku najboljša (beri: čigar subvencionirana cena novega telefonskega aparata je najnižja). V tem primeru gre za očitno nerazumevanje problematike. Rezultati bi bili veliko boljši, če bi se pri gradnji modela usmerili na pravi tržni segment (prej omenjene najboljše uporabnike).

2.5.2 PRETVORBA PODATKOV V UPORABNE INFORMACIJE

Bistvo podatkovnega rudarjenja leži v spreminjanju podatkov v koristne informacije. Temu delu postopka se bom v pričujočem diplomskem delu najbolj posvetil. Navsezadnje celoten praktični del temelji na tem delu postopka.

Slika 1 prikazuje osnovne korake, ki nas privedejo do uporabnih informacij. Z drugimi besedami – prikazuje izgradnjo modela, ki bo na osnovi pridobljenih informacij sposoben napovedovati bodoče dogajanje ali stanje.

Slika 1: Proces izgradnje modela za podatkovno rudarjenje



Vir: Berry, Linoff, 2000, str. 48.

Ugotavljanje potrebe po podatkih in njihovo pridobivanje

Prvi korak v procesu izgradnje modela je ugotavljanje, kakšne podatke sploh potrebujemo. Pogosto se zgodi, da so to enostavno podatki, ki so pač na voljo. V splošnem velja, naj jih bo rajši preveč kot premalo. Seveda morajo biti izbrani podatki popolni. Če želimo med strankami odkriti verjetne kupce novega izdelka, je potrebno imeti podrobne podatke o vsaki stranki posebej. Pogosto je potrebno kombinirati podatke iz različnih virov, lahko tudi zunanjih. Pri izgradnji napovedovalnih modelov moramo poskrbeti, da podatki vsebujejo želene izide (ciljne spremenljivke).

Pregledovanje, preverjanje in čiščenje podatkov

Iluzorično je pričakovati, da bodo izbrani podatki idealni. Običajno je prej nasprotno. Pogosto manjkajo vrednosti v nekaterih poljih (npr. starost kupca). Manjkajoče vrednosti so lahko velik problem pri uporabi nekaterih tehnik, npr. pri nevronskih mrežah. Nadalje se je potrebno vprašati o pravilnosti danih podatkov – vrednosti polj morajo biti smiselne in verjetne. Če podatki kažejo, da ima stranka rojstni datum nekje v prihodnosti, tu verjetno nekaj ni v redu.

Kritični in za poslovanje pomembni podatki so na zgoraj omenjene težave bolj ali manj imuni. Večjo težavo predstavljajo podatki, ki uporabljajo poredko ali skoraj nikoli. Gre za podatke, ki jih podjetje hrani in zbira »za vsak slučaj«. Pogosto se celotna količina podatkov prvič uporabi ravno pri procesu podatkovnega rudarjenja. Seveda tudi tu velja pri podatkih splošno uveljavljeno načelo, da na osnovi slabih podatkov ne moremo pričakovati dobrih in uporabnih rezultatov (ang. Garbage In, Garbage Out - GIGO). Na osnovi slabih podatkov nam ne bo nikoli uspelo zgraditi uporabnega modela.

Izbira ravni pogleda na podatke

Vsi algoritmi, ki se uporabljajo pri podatkovnem rudarjenju, uporabljajo skrajno preprost pogled na podatke. Ta sestoji iz ene same tabele, z vrsticami kot opazovanimi enotami in s stolpci kot spremenljivkami. Vendar v večini poslovnih okolij podatki niso shranjeni na tak način. Kar je dobro za PR, je neprimerno ali vsaj neoptimalno za večino drugih namenov. Zato je potrebno podatke »prevesti« v primerno obliko.

Odgovor na vprašanje, kaj naj predstavlja posamezna vrstica podatkov, je odvisen od želenih rezultatov oz. od načina uporabe teh rezultatov. Ker je večina poslovne uporabe PR osredotočene na stranko (kupec, dobavitelj, naročnik, komitent…), ta največkrat tudi predstavlja opazovano enoto.

Dodajanje izvedenih spremenljivk

Izvedene spremenljivke so izračunane iz ostalih spremenljivk. Pogosto lahko njihova uporaba bistveno pripomore k boljšemu modelu. Nekaj primerov tovrstnih spremenljivk:

- celotno število in skupna vrednost transakcij,
- rast prodaje od začetka do konca obdobja ali
- delež vrednosti nakupov, ki odpade na stranko.

Priprava modelnih podatkov
Na tej stopnji pridemo do podatkov, ki bodo dejansko uporabljeni pri gradnji modela (modelni podatki). Kaj je poleg prej omenjenih korakov še potrebno storiti?

Vzemimo za primer izdelavo napovedovalnega modela, ki ga gradimo na osnovi preteklih podatkov. Ugotoviti želimo, katere transakcije s plačilnimi karticami so goljufije (posledica kraje kartice). Na osnovi preteklih podatkov o transakcijah je vidno, da je delež goljufij manjši od 1%. Skoraj vsak model, zgrajen na tej osnovi, bi bil več kot 99% točen – če enostavno ne bi napovedal nobene goljufije. Zato je potrebno delež redkih izidov (v tem primeru goljufij) povečati. Najenostavnejša in najbolj priljubljena metoda je **prevzorčenje**[4] (ang. oversampling). Pri tem iz celotne populacije transakcij izberemo vzorec, ki vsebuje precej večji delež (npr. 20% - 30%) goljufivih transakcij kot celotna populacija. Ta vzorec se v naslednjem koraku uporabi za učenje modela.

Modelne podatke je potrebno še razdeliti na učni del, potrjevalni del in testni del. Več o tem v poglavju 4.4.1.

Izbira prave tehnike in učenje modela
Izbira tehnike in množice nastavitev je dolgotrajen postopek. Najboljša kombinacija ni znana vnaprej, zato običajno na osnovi istih podatkov izdelamo več modelov. Več o različnih tehnikah, ki se uporabljajo pri gradnji modela, v poglavju 2.6.

Preverjanje zmogljivosti modela
Nazadnje nam preostane še ocenjevanje zmogljivosti modela. Preveriti moramo, kako se model obnaša na nikoli prej videnih podatkih. Pri tem se uporablja potrjevalni del modelnih podatkov, ki niso bili uporabljeni pri učenju modela.

Zmogljivosti modela najlažje preverimo z matricami učinkovitosti modela in grafičnimi prikazi koristnosti. Nekaj teh je natančneje predstavljenih v poglavju 4.4.5.

Izbira najboljšega modela
Zmogljivosti modela primerjamo med seboj in se odločimo za najboljšega.

[4] Statistiki bi to imenovali stratificirano naključno vzorčenje.

2.5.3 UKREPANJE NA OSNOVI PRIDOBLJENIH INFORMACIJ

Namen PR je, da nam na osnovi pridobljenih rezultatov omogoča ustrezno ukrepanje. Če so bili ti rezultati uporabljeni za gradnjo napovedovalnega modela, se zgrajeni model aplicira na še nevidene podatke. Pri tem gre lahko za zgolj enkratno uporabo (npr. za eno oglaševalsko kampanjo), lahko pa se model uporablja periodično (npr. za vsako kampanjo) ali pa je celo za uporabljen za obdelavo podatkov v realnem času (npr. dobičkonosnost vsake stranke se preverja in določa ob vsaki njeni transakciji – nakupu). V tem primeru je treba upoštevati, da običajno od zajemanja modelnih podatkov do apliciranja novega modela na nevidene podatke mine nekaj časa. Razmere v okolju se do tedaj lahko že spremenijo in na preteklih podatkih naučen model se na trenutnih podatkih lahko izkaže precej slabše. Takrat govorimo o nizki **stabilnosti modela**[5]. Nasprotno je stabilen model tisti, ki se podobno dobro obnese tako na modelnih kot na nevidenih podatkih.

Lahko pa so rezultat PR nova dejstva, ki omogočijo bolj jasen vpogled v poslovanje podjetja ali v obnašanje strank. Ta vpogled moramo nato omogočiti vsem zainteresiranim, ki bi od njega lahko imeli koristi.

Zgodi se tudi, da ves trud okrog PR razkrije slabe podatke, na osnovi katerih ni mogoče ni mogoče priti do želenih rezultatov. V tem primeru so edini ukrepi usmerjeni v izboljšanje, čiščenje in popolnjevanje podatkov.

2.5.4 OVREDNOTENJE REZULTATOV

Dobljene rezultate je potrebno primerjati z dejanskim stanjem, kar velja še posebej pri napovedovanju bodočih dogodkov. Je stranka sprejela ponudbo? Je bilo elektronsko sporočilo res nezaželeno? Edini način, da dobimo odgovor na takšna in podobna vprašanja je, da napovedi primerjamo s tem, kar se je v resnici zgodilo. Pri tem lahko uporabljamo enake tehnike primerjave, kot pri preverjanju zmogljivosti modelov. Razlika je le v tem, da tu uporabljamo nevidene podatke, ki pri gradnji modela še niso bili znani.

Pri napovedovalnih modelih so realni rezultati običajno slabši od napovedanih. Razlog je v že omenjenem časovnem razmaku med nastankom modelnih in nastankom nevidenih podatkov. Običajno velja: večji kot je razmak, slabši so

[5] Vzrok nizke stabilnosti je lahko tudi t.i. preveliko prileganje – glej poglavje 4.4.1.

rezultati. Ko ti niso več zadovoljivi, je model potrebno ponovno zgraditi. Celoten postopek podatkovnega rudarjenja se ponovi še enkrat od začetka.

2.6 TEHNIKE PODATKOVNEGA RUDARJENJA

V tem poglavju je predstavljenih pet najpogosteje uporabljanih tehnik PR.

2.6.1 NAJBLIŽJI SOSED

Ideja pri tej tehniki je preprosta in ljudem lahko razumljiva: pri reševanju problemov se pogosto naslonimo na rešitve podobnih problemov, ki smo jih že rešili. Podobno pot ubira tudi tehnika najbližjega soseda. Da bi napovedala vrednost spremenljivke pri izbrani opazovani enoti se nasloni na pretekle podatke in izmed podobnih enot izbere tisto, ki ji je po značilnostih »najbližja«. Vrednost (istovrstne) spremenljivke pri tej enoti nato enostavno priredi iskani spremenljivki.

Za zgled vzemimo ocenjevanje cen nepremičnin. V primeru, da želimo oceniti ceno nepremičnine, se zgledujemo po cenah podobnih sosednjih nepremičnin. Če je npr. po velikosti podobna sosednja nepremičnina vredna 20 milijonov SIT, obstaja verjetnost, da je vrednost naše nepremičnine podobna. Ta verjetnost je še bistveno večja, če imajo tudi ostale nepremičnine te velikosti v bližji soseščini podobno vrednost. Vsekakor večja, kot če bi bila vrednost teh sosednjih nepremičnin v povprečju 5 milijonov SIT. Seveda je naveden primer zelo poenostavljen, vendar vseeno prikaže princip najbližjega soseda. »Bližina« je v tem primeru lahko predstavljiva, saj gre za geografsko razdaljo.

Običajno pri tej tehniki ne iščemo samo enega »najbližjega soseda«, ampak več. Pri tem upoštevamo povprečje vrednosti spremenljivk, če gre za intervalne spremenljivke oz. najpogostejšo vrednost (modusni razred), če gre za kategorične spremenljivke. S tem se izognemo nepravilni napovedi v primeru, da je »najbližji sosed« izjema, ki se ne podreja splošnim zakonitostim v podatkih.

2.6.2 RAZVRŠČANJE V SKUPINE

Gre za razvrščanje opazovanih enot v skupine (ang. **clustering**). V isto skupino se razvrstijo enote, ki so si med seboj čim bolj podobne, same skupine pa naj bi bile med seboj čim bolj različne. Pri tem skupine niso znane vnaprej, ampak se obdelujejo »surovi« podatki. Običajno z namenom ustvariti boljši pogled na podatke in

ugotoviti, kaj se v bazi podatkov dogaja. Prav tako ni vnaprej znano, na osnovi katerih spremenljivk se bo izvedlo razvrščanje v skupine. Zato je pogosto potrebno izključiti nepomembne spremenljivke, ki na razvrščanje nimajo (oz. ne smejo imeti) nobenega vpliva. Ko so skupine definirane (odkrite), jih lahko uporabimo za klasifikacijo novih podatkov.

V idealnem primeru bi skupino sestavljale samo opazovane enote z enakimi vrednostmi spremenljivke (oz. spremenljivk), na osnovi katerih je bilo razvrščanje sploh opravljeno. Druga skrajnost je, če so vrednosti teh spremenljivk tako različne, da se opazovane enote razdrobi do skrajnosti na tak način, da vsaka skupina vsebuje zgolj eno opazovano enoto. Seveda takšno razvrščanje izgubi svoj smisel, saj s tem ne dosežemo ničesar. Ker v prvem primeru opisanega razvrščanja podatki običajno ne omogočajo, drugi pa je neuporaben, je končni rezultat nekje med obema skrajnostma. Običajno je končno število skupin odvisno od uporabnikovih želja, saj večina programskih orodij vsaj do neke mere omogoča izbiranje števila skupin. Pri tem gre za tehtanje (ang. trade-off) med večjim številom bolj homogenih skupin in manjšim številom bolj heterogenih skupin.

Razvrščanje podatkov v skupine pa je uporabno tudi pri iskanju izjemnih vrednosti spremenljivk (ang. outliers). Njihovo prepoznavanje nam lahko precej olajša razumevanje situacije v podatkih. Poleg tega jim je treba nameniti posebno pozornost, saj lahko »izkrivijo« model.

2.6.3 INDUCIRANA PRAVILA

Uporaba induciranih pravil se verjetno še najbolj približa tistemu, kar si pod pojmom podatkovno rudarjenje predstavlja povprečen laik, t.j. iskanje koristnih informacij v množici podatkov. Pri tem koristne informacije predstavljajo nova, doslej neznana pravila, ki veljajo med podatki. Pravila so tipa ČE-POTEM: če je izpolnjen pogoj A, potem pogosto velja B.

Uporaba induciranih pravil je lahko velik zalogaj. Potrebno je namreč sistematično analizirati vse vzorce, ki se pojavljajo v podatkovni bazi, in za vsak vzorec določiti stopnjo zaupanja in verjetnost, da se bo ta vzorec ponovil. Kljub temu pa gre za zelo enostavna pravila. Tako bi npr. pri analiziranju nakupov lahko ugotovili, da je 65% kupcev ob nakupu okrasnih rastlin kupilo tudi umetno gnojilo. Ta vzorec se je pojavil pri 6% vseh nakupov.

Zgornje pravilo lahko shematsko zapišemo na sledeč način:
OKRASNA RASTLINA ⇒ UMETNO GNOJILO
Pri tem »OKRASNA RASTLINA« predstavlja levi del zapisa oz. LHS (left-hand side), »UMETNO GNOJILO« pa desni del oz. RHS (right-hand side). LHS lahko sestoji iz enega ali več pogojev, ki morajo biti izpolnjeni, medtem ko RHS običajno predstavlja en sam dogodek.

Deleža nakupov, ki vsebujejo (tudi) omenjena artikla, ni težko izračunati. Enostavno seštejemo vse tovrstne nakupe in dobljeno število primerjamo s številom vseh nakupov. Dobljenemu rezultatu rečemo **stopnja pokritja** ali **podpora** (ang. coverage ali support) in v našem primeru znaša 0,06 oz. 6%.

Naslednji kazalec je **stopnja zaupanja** (ang. accuracy ali confidence). Gre za pogojno verjetnost, v našem primeru je to verjetnost, da je kupec ob nakupu okrasne rastline kupil tudi umetno gnojilo. Ta znaša 0,65 oz. 65%.

Opozoriti velja, da so najdena pravila v resnici opisi razmerij, ki veljajo v podatkovni bazi. Odkrite povezave in vzorci ne temeljijo nujno na vzročno-posledični zvezi. Z drugimi besedami: LHS ni nujno **vzrok** za RHS. Poglejmo si primer: PR lahko odkrije, da pri moških z mesečno plačo od 200.000 SIT do 300.000 SIT, ki so naročniki določene revije, obstaja verjetnost, da bodo kupili nov izdelek[6] (stopnja zaupanja je relativno velika). Ponudnik tega izdelka lahko to informacijo izkoristi in svojo trženjsko kampanjo usmeri na ljudi, ki ustrezajo temu vzorcu. Vendar na osnovi tega pravila ne sme sklepati na dejanski vzrok za nakup izdelka.

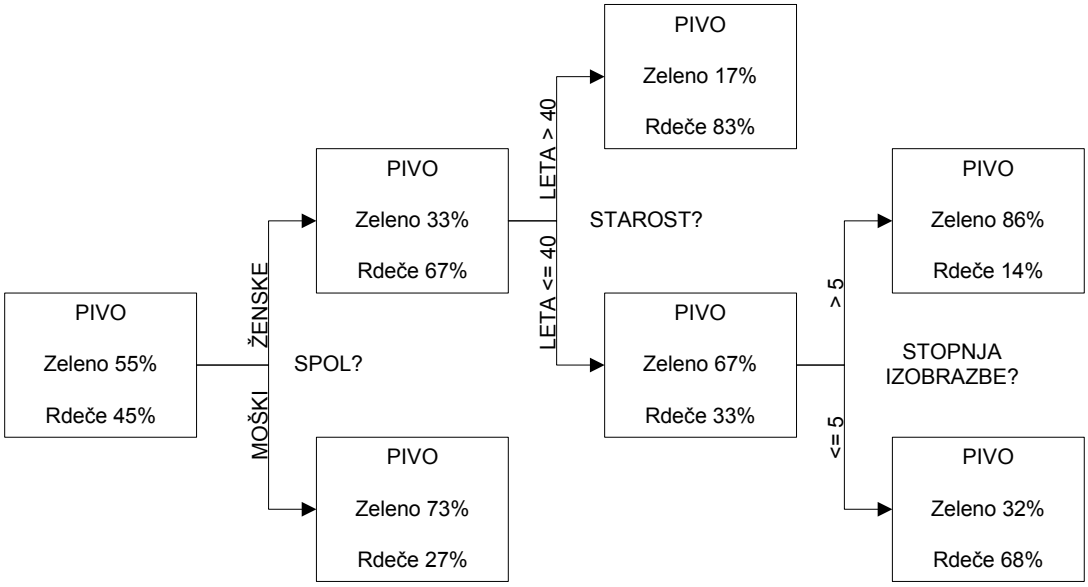
2.6.4 ODLOČITVENA DREVESA

Poimenovanje odločitvenih dreves izhaja iz dejstva, da je to tehniko možno predstaviti v sliki, ki spominja na drevo. Vsako vozlišče predstavlja vprašanje, vsak list pa segment populacije (Berson, 2000, str. 156). Proces izgradnje odločitvenega drevesa se imenuje rekurzivno particioniranje. Opazovane enote so na osnovi vrednosti ene izmed neodvisnih spremenljivk razdeljeni na segmente (particije). Novonastali segmenti se potem delijo naprej in naprej. Odločitev, na osnovi katerih spremenljivk in njihovih vrednosti naj se te delitve izvajajo, temelji na preizkušanju vseh možnih kombinacij. Za delitev se nato izbere najboljša. Na sliki 2 je predstavljen grafični prikaz odločitvenega drevesa. Zaradi boljše izkoriščenosti prostora je koren

[6] Primer pravila s tremi pogoji v LHS.

drevesa (izhodiščna točka) na desni strani, drevo pa »raste« v levo (običajnejši prikaz je od zgoraj navzdol).

Slika 2: Primer odločitvenega drevesa



V tem primeru gre za klasifikacijo pivopivcev na dva tabora: na ljubitelje piva v zeleni embalaži in na ljubitelje piva v rdeči embalaži, pri čemer so podatki izmišljeni in ne odražajo dejanskega stanja na trgu. Navdušenje nad eno ali drugo vrsto piva želimo razložiti z ostalimi podatki, ki jih imamo o posameznem pivopivcu. Pri gradnji drevesa bomo uporabljali odgovore na preprosto vprašanje: katero pivo je pivopivcem ljubše? Najprej pri vseh obdelovanih pivopivcih preverimo najljubše pivo. Rezultat je malenkost v prid »zelenemu« pivu, mi pa želimo dobiti bolj homogene segmente. Zato populacijo razdelimo na osnovi spola. Vidimo, da sta novonastala segmenta že bolj homogena. Če ženske nadalje razdelimo po letih, segmenti postanejo še »čistejši«. Mlajše od 40 let pa razdelimo še po stopnji izobrazbe in model še malenkost izboljšamo.

Ta proces bi se lahko nadaljeval še naprej, vendar ga je potrebno na določeni stopnji ustaviti. V skrajnem primeru bi namreč prišli do situacije, ko bi vsak segment sestavljal en sam pivopivec. Dobljeni model bi ustvaril popolnoma homogene segmente, vendar bi bil neuporaben za naš prvotni namen – klasifikacijo potrošnikov. Ta pojav imenujemo **preveliko prileganje** (ang. overfitting). Pri odločitvenih drevesih se z njim lahko borimo na dva načina. Lahko vnaprej omejimo rast na osnovi izbranih kriterijev (npr. največje dovoljeno število segmentov), kar

imenujemo tehnika »bonsai«, lahko pa pri že zgrajenem drevesu odstranimo neuporabne veje. Slednjo tehniko imenujemo obrezovanje (ang. pruning).

Pridobljene informacije o pivopivskem trgu potem lahko uporabimo za klasifikacijo – za novega potrošnika na osnovi njegovih značilnosti (spol, starost, izobrazba…) ugotovimo, v kateri segment se uvršča. Nato preverimo, katero pivo »prevladuje« v tem segmentu, in na osnovi tega ocenimo potrošnikov okus.

Pri odločitvenih drevesih je precej enostavno razumeti, kako je zgrajen model, ki ga je tudi enostavno grafično prikazati. Pogosto namreč ljudje ne zaupajo tehnikam, ki jim niso razumljive. Tega se pač pri drevesih ni treba bati, kar je verjetno eden glavnih razlogov za njihovo široko uporabo.

2.6.5 NEVRONSKE MREŽE

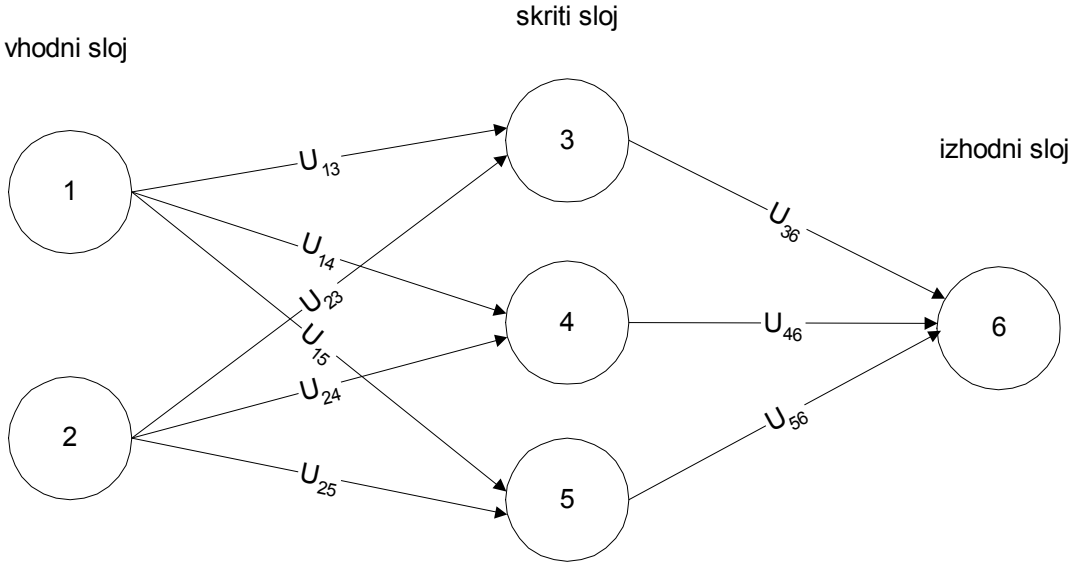
Nevronske mreže so za področje podatkovnega rudarjenja zanimive zato, ker omogočajo modeliranje kompleksnih problemov z velikim številom odvisnih spremenljivk. Svoje ime dolgujejo zgodovinskemu razvoju, ki se je pričel s tezo, da je računalnik mogoče s posnemanjem delovanja človeških možganov »naučiti razmišljati«. Sprva se je razvoj na tem področju odvijal znotraj laboratorijev za umetno inteligenco, danes pa se nevronske mreže uporabljajo tudi na drugih področjih – tudi na področju podatkovnega rudarjenja.

Nevronska mreža sestoji iz vozlišč in povezav med njimi ter je razdeljena na tri sloje ali več (glej sliko 3):

- vhodni sloj (ang. input layer), kjer vsako vozlišče predstavlja neodvisno (vhodno) spremenljivko,
- eden ali več skritih slojev (ang. hidden layer) in
- izhodni sloj (ang. output layer), ki ga sestavljajo ciljne (odvisne) spremenljivke.

Vsako vozlišče (shematsko prikazano s krogom) v vhodnem sloju je povezano z vsakim vozliščem v skritem sloju. Ta so naprej lahko povezana v vozlišči v naslednjem skritem sloju (če jih je več) ali pa z vozlišči v izhodnem sloju.

Slika 3: Primer nevronske mreže



Vir: Introduction to Data Mining and Knowledge Discovery, 1999, str. 12.

Po vhodnem sloju vsako vozlišče prevzame vhodne vrednosti predhodnih vozlišč, jih pomnoži s pripadajočimi utežmi U_{xy}, sešteje, nad njihovo vsoto izvede neko numerično operacijo (aktivacijska funkcija) in rezultat preda preda vozliščem (vozlišču) v naslednjem sloju. Vrednost, ki se npr. prenese iz vozlišča 4 v vozlišče 6 (glej sliko 3), dobimo takole:
*AF ([U_{14} * vrednost vozlišča 1] + [U_{24} * vrednost vozlišča 2])*, pri čemer AF predstavlja aktivacijsko funkcijo, ki jo izvedemo nad zapisano vsoto.

Uteži (U_{xy}) so neznani parametri, katerih vrednost se določi v procesu učenja modela. V model posamično vstopajo opazovane enote. Na osnovi primerjave med napovedano in dejansko vrednostjo ciljne spremenljivke pri posamezni opazovani enoti se ustrezno popravijo vrednosti uteži. Postopek je namenjen manjšanju napake napovedi in se ponovi za vsako opazovano enoto. Ko je obdelana še zadnja po vrsti, v model spet vstopi prva. Odvisnost napake od število ponavljanj (kolikokrat se obdela celotni set modelnih podatkov) je vidna s slike 4.

Nevronske mreže se pogosto izkažejo kot najboljša izbira, vendar so modeli zaradi kompleksnosti težko razumljivi. Zato je pri njihovi uporabi zelo težko ali celo nemogoče obrazložiti dobljene rezultate. To dejstvo pogosto preprečuje njihovo uporabo na področjih, kjer je bolj kot zmogljivost modela cenjena njegova razumljivost (npr. odobritev / zavrnitev prošnje za posojilo).

2.7 ETIČNI VIDIK PODATKOVNEGA RUDARJENJA

Uporaba podatkovnega rudarjenja v določenih primerih lahko sproži etična vprašanja, sploh kadar se PR uporablja nad osebnimi podatki. Pogosto namreč uporaba PR privede do diskriminacije – kateremu komitentu bo banka odobrila posojilo ali komu bo poslana posebna ponudba. Nekatere vrste diskriminacije, kot npr. spolna, verska ali rasna, ne samo da niso etične, ampak so tudi kaznive. Vendar tudi v tem primeru ni čiste ločnice med sprejemljivim in nesprejemljivim (mogoče celo kaznivim). Uporaba podatkov o spolu in rasi je pri diagnosticiranju bolezni nedvomno etična, kar pa ne moremo reči za rudarjenje po teh podatkih z namenom ugotavljanja plačilne discipline posameznika. Četudi skušamo ravnati etično in tovrstne informacije izključimo iz PR, lahko v podatkih še vedno obstajajo implicitno. Tako bi lahko iz PR načrtno izključili informacije o nacionalni pripadnosti oseb, vendar bi vseeno obstajala nevarnost, da model temelji na narodnosti. Pogosto se namreč dogaja, da se ljudje naselijo v okolici, kjer živi več pripadnikov njihove narodnosti. Če bi bil v gradnjo modela vključen tudi kraj bivanja, bi zgrajen model vseeno vsaj deloma temeljil tudi na nacionalnosti.

Pri zbiranju podatkov bi se zato moralo uporabljati načelo, da je ljudi, ki so pripravljeni izdati svoje osebne podatke, potrebno vnaprej seznaniti z načini uporabe teh podatkov – kako in zakaj se bodo uporabljali njihovi osebni podatki. Vendar podjetja navadno to skrčijo v izjavo »da se bodo podatki uporabljali za interne potrebe podjetja«. Podjetje običajno tudi zagotovi, da osebnih podatkov ne bo posredovalo tretji osebi. Vendar prav PR predstavlja potencialen etični konflikt. Z njegovo uporabo se podjetju lahko odprejo popolnoma nove dimenzije vpogleda v te podatke, ki kar kličejo po nadaljnji obdelavi. Seveda z nameni, ki niti približno niso bili predstavljeni osebam, ki so zaupale svoje osebne podatke. Zastavlja se vprašanje, ali lastništvo nad podatki dopušča njihovo obdelavo na poljuben način za poljubne namene? Če je bil namen zbiranja podatkov eksplicitno določen, vsekakor ne.

3. SMETI V ELEKTRONSKI POŠTI

3.1 KAJ JE TO?

Z angleško besedo spam označujemo nezaželeno elektronsko pošto. Gre torej za »smeti«, kakršnih smo vajeni tudi v realnem svetu, saj nam poštne nabiralnike vsakodnevno polnijo razni oglasi, brezplačne revije, katalogi in podobno. Pomembna razlika je v tem, da v prvem primeru govorimo o nekaj tisoč poslanih primerkih, pri

e-pošti pa ta številka zlahka doseže na desetine milijonov. V večini primerov nezaželena e-pošta vsebuje komercialno oglaševanje pogosto sumljivih izdelkov, ponudb za hitro bogatenje, pornografskih strani ipd.

S tem, ko govorimo o nezaželeni pošti oz. o e-smeteh, v prvi vrsti mislimo na nenaročeno pošto, ki se odpošilja na množico naslovov. V tem smislu bo ta pojem tudi uporabljan v moji diplomski nalogi in ne bo zajemal individualnih, čeprav nezaželenih sporočil (nadlegovanje bivšega dekleta preko e-pošte, opomini za plačilo, vabila na dolgočasne zabave…). Za ostalo, torej »zaželeno« e-pošto, bom uporabljal termin »običajna e-pošta«.

3.2 ŠKODA, KI JO POVZROČA E-SMETENJE

Nevarnosti e-smetenja se najlažje zavemo, če pogledamo ocenjene stroške, ki jih ta pojav povzroča: samo v ZDA imajo podjetja letno zaradi e-smetenja 8,9 milijarde USD stroškov. Od tega odpade štiri milijarde USD na zmanjšanje produktivnosti v podjetjih. V svetovnem merilu pa so celotni stroški ocenjeni na 13 milijard USD. (Study: Spam costs businesses $13 billion, 2003).

Zakaj se tako razburimo, ko prejmemo e-sporočilo, ki ga nismo zahtevali? Razlogov je več, nekaj najbolj očitnih je predstavljenih v nadaljevanju.

3.2.1 ZASTONJKARSTVO

Ena izmed značilnosti e-smetenja je, da prejemnik plača veliko več kot pošiljatelj. Ameriški ponudnik dostopa do interneta AOL (America On-Line) navaja, da je samo eden izmed največjih pošiljateljev e-smeti dnevno njihovo omrežje »zasuval« z 1,8 milijona sporočil, dokler ga niso zaustavili s sodno prepovedjo (Information about spam, 2003). Predpostavimo, da je vsak uporabnik povprečno v 10 sekundah spoznal, da gre za e-smet, in sporočilo zavrgel. V tem primeru so samo uporabniki AOL-a v enem dnevu potrošili 5000 ur povezave v internet. Resnici na ljubo ta stroškovni vidik zgublja na pomembnosti, saj vedno več uporabnikov za povezave v internet plačuje pavšalne zneske. Vendar so kljub temu zneski za omenjenih 5000 ur povezave še vedno bistveno višji od stroškov, ki jih ima pošiljatelj.

3.2.2 POPLAVA E-SMETI

Veliko sporočil, ki jih uvrščamo med e-smeti, vsebuje možnost, da svoj naslov odstranimo iz seznama prejemnikov sporočil[7]. Dokler je količina prejetih sporočil majhna, bi lahko na vsako nezaželeno sporočilo odgovorili na način, ki obljublja odstranitev iz seznama prejemnikov. Kaj pa, če količina e-smeti naraste in začnemo dobivati na svoj e-naslov nekaj sto nezaželenih sporočil dnevno? Verjetno se nihče izmed nas ne bi več trudil z »odnaročanjem« nečesa, česar si sploh nismo želeli prejemati. E-smeti tako lahko preplavijo naš e-nabiralnik, da ni več uporaben za vsakdanje delo. Na žalost to ni neverjeten scenarij – e-smetenje pač vsakomur omogoča z nizkimi (oz. zanemarljivimi) stroški doseči ogromno potencialnih strank. Prodajate avto? Zakaj ne bi poslali ponudbe na nekaj deset tisoč naslovov?

Seveda ne smemo poazbiti, da tudi »obvladljiva« količina e-smeti v elektronski pošti za podjetja predstavlja velik strošek. Za prepoznavanje, sortiranje in brisanje nezaželene pošte zaposleni porabljajo vedno več časa, ki bi moral sicer biti namenjen njihovim delovnim nalogam. Da v tem primeru močno trpi produktivnost poslovanja, je jasno vsakomur.

3.2.3 PORABA OMREŽNIH IN SISTEMSKIH ZMOGLJIVNOSTI

Pošiljanje ogromnih količin e-smeti oz. pošiljanje enega sporočila na ogromno število naslovnikov ustvarja velik promet v omrežju, zaradi česar je lahko oteženo ali upočasnjeno vsakodnevno delo. Prav tako e-smeti zasedajo dragocen pomnilniški prostor.

3.2.4 NEKORISTNA VSEBINA E-SMETI

Velika večina tovrstne e-pošte vsebuje / oglašuje popolnoma neuporabne stvari. Poleg že omenjenih vsebin so tu še ponudbe za programsko opremo, ki omogoča e-smetenje (lahko skupaj z bazo elektronskih naslovov) in ostala šara, katere oglaševanje v ostalih medijih se ne splača, ker ga je dejansko potrebno plačati. Opozoriti velja še na vsebine, ki so v nekaterih državah lahko prepovedane – npr. otroška pornografija.

[7] Kot bomo videli kasneje, ta možnost ne pomeni vedno tistega, kar obljublja.

3.2.5 GOLJUFANJE E-SMETILCEV

Večina programske opreme, ki omogoča e-smetenje je na voljo skupaj z množico naslovov e-pošte. Ti naslovniki naj bi dovolili pošiljanje oglasov (kakršnihkoli že) na njihov e-naslov. Vendar temu ni tako. Povečini gre za nedolžne žrtve, katerih naslovi so bili zbrani iz mnogoterih virov. Nadalje večina ponudnikov laže glede možnosti, ki naj bi prejemniku omogočala, da prekine pošiljanje e-smeti na svoj naslov. Navadno je potrebno klikniti na povezavo v sporočilu ali odgovoriti na sporočilo in v naslov ali telo sporočila vpisati nekaj v smislu: REMOVE, PLEASE_REMOVE ali ODJAVI! Izkušnje kažejo, da v večini primerov smetilec ne odstrani pošiljatelja iz seznama naslovnikov. Še več, pošiljatelj s tem smetilcu izda, da je sporočilo prebral (večina ljudi pač takšnih sporočil ne prebira) in nedolžni naslovnik je zato pogosto deležen še več smeti med svojo e-pošto. »Aktivni« naslovniki so med smetilci iz razumljivih razlogov še posebej cenjeni.

3.3 NAČINI OBRAMBE

Prvi in hkrati najpomembnejši nasvet pri boju proti nezaželeni pošti je, da svojega e-naslova ne izdamo komurkoli. Seveda je v poplavi internetnih strani, ki zahtevajo prijavo s pravim e-naslovom to praktično nemogoče. Prav tako moramo »izdati« svoj e-naslov pri uporabi novičarskih skupin, zahtevah po raznih informacijah, internetnih nakupih in dražbah, pri uporabi nekatere programske opreme ipd. Zato je boljša ideja, da si poleg prvega (službenega) računa e-pošte ustvarimo še dodatnega, katerega naslov lahko brez zadržkov uporabljamo povsod[8]. Pri tem nam na pomoč priskoči množica ponudnikov brezplačne e-pošte (Yahoo Mail, email.si ipd.). Novoustvarjeni račun bo zelo verjetno kmalu postal tarča smetilcev, vendar je to vsekakor manj moteče, kot če bi smetje dobivali na službeni e-naslov. V najslabšem primeru lahko brezplačni račun enostavno prenehamo uporabljati in si ustvarimo novega.

Lahko pa posežemo po namenskih rešitvah za boj proti e-smetenju. Pri tem se uporabljata dva pristopa: lahko se odločimo za zaščito na nivoju **e-poštnega odjemalca** ali na **nivoju strežnika** za e-pošto. V prvem primeru protismetno zaščito namestimo na osebni računalnik, kar je primerno predvsem za domačo rabo. V drugem primeru pa se zaščita namesti na strežnik za elektronsko pošto in z eno namestitvijo pokrijemo vse uporabnike e-pošte na tem strežniku. Ta način je

[8] Službenih računov običajno ni dovoljeno uporabljati v zasebne namene, vendar je to določilo pogosto kršeno.

primeren za podjetja z lastnim strežnikom za e-pošto, vse pogosteje pa ga uporabljajo tudi komercialni ponudniki e-pošte (predvsem ponudniki dostopa do interneta, ki so obenem tudi ponudniki e-pošte).

Omeniti velja še, da večina odjemalcev za elektronsko pošto pravzaprav ponuja neke vrste orožje za boj proti smetenju: gre za enostavno funkcijo sortiranja, ki na osnovi pravil prihajajočo pošto razvrsti v ločene mape. V našem primeru bi lahko izdelali pravilo, ki vsa sporočila od določenega pošiljatelja (smetilca) izbriše ali jih loči od ostale pošte. Pogosto je omogočeno tudi enostavno sortiranje na osnovi vsebine – sporočila lahko sortiramo na osnovi besed ali besednih zvez, ki jih vsebuje sporočilo. Za prvo silo to zadostuje, za kaj več pa kaže uporabiti zgoraj omenjene specializirane programe.

Pri boju proti e-smetenju se moramo zavedati, da popolne zaščite ni. Še več, kar izgleda na prvi pogled kot zelo dobra zaščita, je lahko dvorezen meč. Pri merjenju uspešnosti protismetnih programov se namreč uporabljata dva kriterija:

- delež odkrite nezaželene pošte (učinkovitost) ter
- delež napačno ocenjene pošte (običajna pošta, označena za nezaželeno).

V želji povečati učinkovitost, torej povečati prvi delež, se zelo verjetno poveča tudi drugi delež. Zato pošte, spoznane za e-smeti, ne kaže brisati, ampak jo je zgolj potrebno ločiti od običajne. S tem ima uporabnik še vedno možnost, da prebere običajno sporočilo, čeprav je napačno klasificirano kot e-smet.

Kako učinkovito protismetno zaščito pa je danes možno dobiti na trgu? Na preizkusu revije Monitor (Klančar, 2003, str. 79) je velika večina programov uspela najti več kot polovico nezaželene pošte. Pri najboljših je odstotek najdene tovrstne pošte celo malenkost nad 90%. Zanimiva je primerjava rezultatov med osebnimi in strežniškimi programi. Morda kar malce preseneti ugotovitev, da so slednji v povprečju dosegli celo manjšo učinkovitost. Vendar so na drugi strani napačno ocenili bistveno manj sporočil kot osebni programi. Slednja ugotovitev vsekakor ni presenetljiva: strežniške protismetne rešitve so uporabljajo predvsem v resne, poslovne namene, kjer si ne moremo privoščiti »izgubljene« pošte.

3.4 KAJ PRAVI ZAKON?

Zakon o varstvu potrošnikov (2003) v 45.a členu navaja:

»Podjetje lahko uporablja sistem klicev brez posredovanja človeka, faksimile napravo in elektronsko pošto samo z vnaprejšnjim soglasjem posameznega potrošnika, ki mu je sporočilo namenjeno«. V nadaljevanju tega člena zakon od ponudnikov tudi zahteva, da prenehajo s pošiljanjem e-sporočil naslovnikom, ki jih ne želijo več prejemati.

V 77. členu Zakona o varstvu potrošnikov (2003) je določena tudi kazen za ravnanje v nasprotju z 45.a členom istega zakon: posameznik, ki stori prekršek v zvezi s samostojnim opravljanjem dejavnosti, se kaznuje za prekršek z denarno kaznijo najmanj 1,000.000 tolarjev, pravna oseba pa z denarno kaznijo najmanj 3,000.000 tolarjev.

Pravno gledano je torej e-smetenje v Sloveniji prepovedano, vendar so sodni ukrepi možni zgolj proti domačim smetilcem. Tudi pri nas pa večina e-smeti prihaja iz tujine, kamor roka našega zakona ne seže.

4. S PODATKOVNIM RUDARJENJEM PROTI E-SMETENJU

4.1 TEKMOVANJE DATA MINING CUP 2003

Podatki, ki jih bom uporabil za praktičen prikaz uporabe tehnik podatkovnega rudarjenja, izhajajo z letošnjega tekmovanja Data Mining Cup (DMC). DMC spada v sklop dogodkov, ki se odvijajo pod okriljem konference Data Mining User Event Days v Chemnitzu (Nemčija). Tekmovanje je odprtega tipa in je namenjeno študentom iz celega sveta. Letos se ga je udeležilo 514 tekmovalcev iz 199-ih univerz / 38-ih držav (Data Mining Cup, 2003). Sama naloga je bila registriranim tekmovalcem na voljo od 15. aprila, časa za reševanje pa je bilo mesec dni. V tem času nas je 30% tekmovalcev oddalo rešitev naloge. Omenim naj še, da nista bila predpisana ne način reševanja, ne programsko orodje.

4.2 SCENARIJ NALOGE

Scenarij naloge na DMC temelji na dejanskih težavah, ki jih je imelo neimenovano podjetje z nezaželeno e-pošto. V nadaljevanju je ta naloga podrobneje predstavljena.

Podjetje opaža, da je velik del prihajajoče pošte v resnici nezaželeno oglaševanje, ki nima nobene zveze s poslovanjem podjetja. Velika količina delovnega časa, ki ga zaposleni namenijo razvrščanju (prepoznavanju) in brisanju teh e-smeti, je v podjetju s 120 zaposlenimi razkrila velike možnosti racionalizacije. Zato je podjetje nekaj časa zbiralo vso prihajajočo pošto, e-smeti ločeno od ostale pošte, in vsako sporočilo opisalo z množico atributov. Tako je bilo shranjenih in obdelanih skupno 8000 e-sporočil, skupaj z informacijo ali je sporočilo e-smet ali ne.

Moja naloga je bila, da na osnovi teh 8000 e-sporočil izdelam model, ki bo v prihodnje zmožen avtomatsko filtrirati prihajajočo pošto. Zaustaviti bo moral sporočila, spoznana za e-smeti, dostaviti pa zgolj zaželeno e-pošto. Z drugimi besedami: model bo uporabljen za **klasifikacijo** prihajajoče pošte.

Cilj je minimizirati število e-smeti, ki bodo prišla skozi filter. Pri tem je potrebno upoštevati naslednjo omejitev: med zaustavljenimi sporočili je dovoljen največ 1% običajnih sporočil, ki ne spadajo med e-smeti (glede ne vsa običajna sporočila).

Model bom apliciral na množico 11.177 e-sporočil, ki so opisana z enakimi atributi kot prvih 8000 e-sporočil, vendar brez informacije, ali gre za e-smet ali ne. Klasifikacija teh 11.117 sporočil je tudi rešitev naloge na DMC.

Pri tem si bom pomagal s programskim paketom SAS Enterprise Miner Release 4.1 podjetja SAS Institute, Inc.

4.3 VEČ O PODATKIH

Vsako e-sporočilo je opisano z 834 atributi, vključno s serijsko številko sporočila (atribut *id*, šestmestno celo število) in z informacijo, ali gre za e-smet ali ne (atribut *target* z vrednostjo »YES« ali »NO«). Vsi ostali atributi so binarnega tipa in lahko vsebujejo vrednosti 0 ali 1.

Vsi zgoraj omenjeni atributi binarnega tipa so skladni z atributi, ki jih uporablja odprto-kodni projekt SpamAssassin, pod okriljem katerega poteka razvoj istoimenskega programa za odkrivanje oz. filtriranje nezaželene e-pošte (SpamAssassin.org, 2003). V tabeli 2 je navedenih nekaj primerov teh atributov.

Tabela 2: Izvleček atributov, s katerimi je opisano e-sporočilo

ime atributa	opis
GENUINE_EBAY_RCVD90	sporočilo je poslano s strani portala eBay
FWD_MSG	gre za posredovano sporočilo (ang. forward)
SUBJ_ALL_CAPS	"tema" (ang. "subject") sporočila vsebuje samo velike črke
SUBJ_FREE_CAP	"tema" (ang. "subject") sporočila vsebuje besedo FREE
HTML_70_80	70% - 80% sporočila je napisano v HTMLju
HTML_WIN_OPEN	sporočilo vsebuje programsko kodo, ki odpre novo okno
HOTMAIL_FOOTER4	sporočilo vsebuje nogo (podpis), značilno za Hotmail
DATE_YEAR_ZERO_FIRST	nepravilen datum v sporočilu – letnica se začne z 0
SORTED_RECIPS	naslovniki sporočila so abecedno urejeni po naslovih
PURE_PROFIT	sporočilo obljublja čisti dobiček (ang. pure profit)
FREE_PORN	sporočilo obljublja brezplačno pornografijo
SIGNATURE_LONG_DENSE	sporočilo vsebuje dolg podpis, brez praznih vrstic

Vir: SpamAssassin.org, 2003.

Podatki so popolni in ne vsebujejo manjkajočih vrednosti.

4.4 IZDELAVA MODELA PO METODOLOGIJI SEMMA

SEMMA (**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) je postopek, ki ga je za uporabo tehnik podatkovnega rudarjenja razvilo podjetje SAS Institute (Data Mining Using Enterprise Miner Software, 2000). Sam postopek je implementiran v njihovo programsko opremo, med drugim tudi v SAS Enterprise Miner.

Kot pove že ime, metodologija predvideva pet korakov: določanje modelnih podatkov (ang. sample), pregledovanje podatkov (ang. explore), spreminjanje podatkov (ang. modify), izdelava modela (ang. model) in ovrednotenje modela (ang. assess).

4.4.1 DOLOČANJE MODELNIH PODATKOV

Na začetku moramo najprej določiti vhodne oz. **modelne podatke**. Če delamo z izredno velikimi količinami podatkov, lahko izberemo vzorčenje in delamo na vzorcu celotne populacije. Podatke nadalje razdelimo na tri dele: na **učni** (ang. training set), **potrjevalni** (ang. validation set) in **testni** (ang. test set) del. Pri tem je

potrebno opozoriti, da poimenovanje teh treh delov modelnih podatkov ni enotno. Uporabljena imena so skladna s tistimi, ki jih uporablja SAS Enterprise Miner. V večini literature pa so ta imena malenkost drugačna, kot je razvidno iz tabele 3.

Tabela 3: Razlika v izrazoslovju med podjetjem SAS Institute in ostalimi avtorji

SAS INSTITUTE	OSTALI AVTORJI	PREVOD
Training Set	Training Set	učni del
Validation Set	Test Set	potrjevalni del
Test Set	Evaluation Set	testni del

Vir: SAS Institute Inc., 2000a, str. 50 in Berry, Linoff, 2000, str. 184.

Učni del predklasificiranih podatkov se uporablja za »učenje« modela. Algoritmi PR z njegovo pomočjo odkrijejo zanimive vzorce in pravila, ki bodo uporabni pri napovedovanju. Potrjevalni del poskrbi za fino prilagajanje modela in preprečuje, da bi si model »zapomnil« podatke iz učnega dela (overfitting). Testni del se pri gradnji modela ne uporablja, pač pa je uporabljen za ugotavljanje zmogljivosti modela. Da bi bila ocenjena zmogljivost čim bolj verodostojna, ta del podatkov med gradnjo modela ni viden in zato za model predstavlja še nevidene podatke.

V mojem primeru je modelne podatke predstavljala tekstovna datoteka z 8000 predklasificiranimi zapisi o prejetih e-sporočilih. Število zapisov ni bilo preveliko, tako da vzorčenja nisem uporabil in sem od začetka uporabljal vseh 8000 zapisov.

Podatke sem razdelil na sledeč način:

- učni del 40% opazovanih enot (3200 sporočil),
- potrjevalni del 30% opazovanih enot (2700 sporočil) ter
- testni del 30% opazovanih enot (2700 sporočil).

Uporabljeno razmerje je kar tisto, ki ga kot privzeto uporablja SAS Enterprise Miner[9]. Pri razvrščanju opazovanih enot na te tri skupine sem uporabil enostavno slučajno vzorčenje.

[9] Seveda je to razmerje možno spremeniti, vendar uporaba drugačnih razmerij ni prinesla praktično nikakršnih izboljšav.

Enterprise Miner že takoj po uvozu podatkov dodeli vsaki spremenljivki svojo **vlogo**. Najbolj pogosto uporabljene vloge so:

- *target* – ciljna spremenljivka, katere napovedovanje bo naloga modela,
- *input* – spremenljivka, od katere je odvisna vrednost ciljne spremenljivke,
- *rejected* – spremenljivka, ki nima vpliva na ciljno spremenljivko ter
- *id* – identifikacijska spremenljivka, ki enolično določa opazovano enoto[10].

Seveda uporabnik te vloge kasneje lahko spremeni. Vendar samodejno dodeljevanje vlog precej olajša delo. V mojem primeru je bilo od 832 (vse, razen *id* in *target*[11]) spremenljivk kar 384 takšnih, ki na ciljno spremenljivko nisi imele nobenega vpliva. Njihova vloga je bila torej nastavljena na *rejected*. Razlog za to je enostaven: vrednost vsake od teh spremenljivk je bila za vsako od opazovanih enotah enaka oz. so bile to unarne spremenljivke. To ne pomeni nujno, da te spremenljivke v splošnem nimajo vpliva na ciljno spremenljivko. Pomeni samo to, da na osnovi danih podatkov na ta vpliv ne moremo sklepati in jih zato tudi ne vključimo v model.

4.4.2 PREGLEDOVANJE PODATKOV

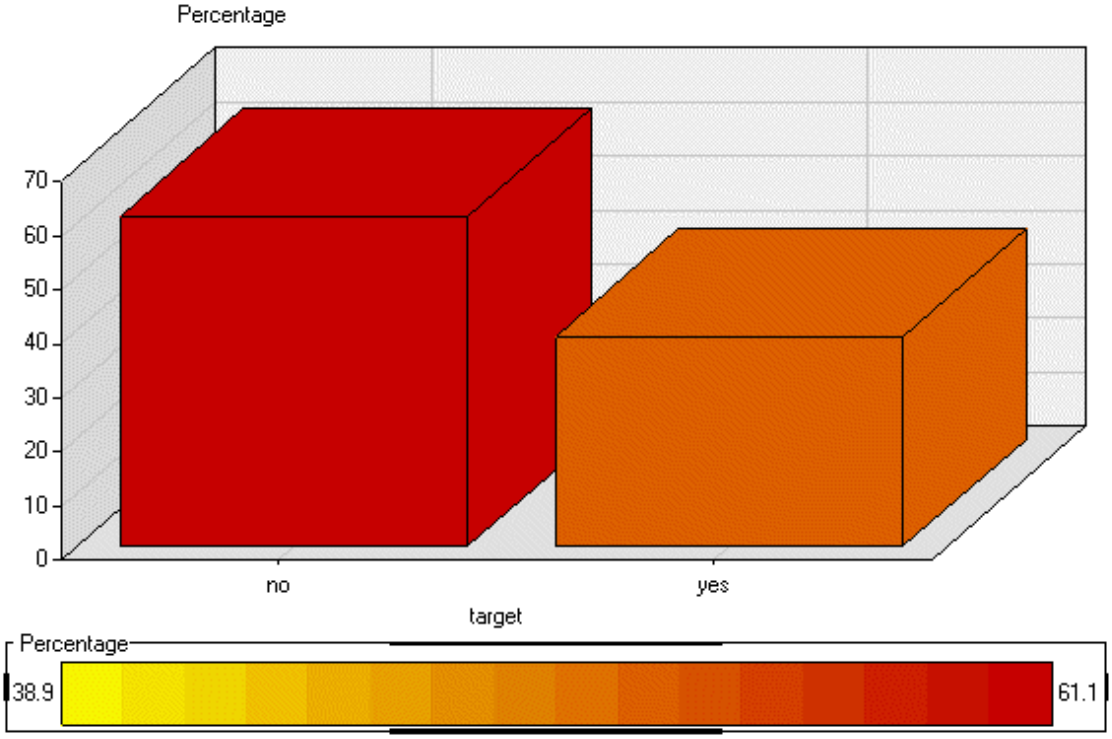
V tem koraku pregledamo podatke s statističnimi metodami, lahko jih tudi grafično prikažemo (opisovanje in vizualizacija). Uporabljamo lahko tehniko induciranih pravil in določamo, katere vhodne spremenljivke bomo uporabili za iskanje ciljne spremenljivke. Ugotavljamo odvisnosti med spremenljivkami (npr. *POŠTA* in *POŠTNA ŠTEVILKA*), odstranimo spremenljivke z velikim deležem manjkajočih vrednosti in na osnovi tega določimo, katere vhodne spremenljivke ne vplivajo na ciljno spremenljivko. Slednje nato izključimo iz modela.

Sama vizualizacija v mojem primeru ni bila koristna. Natančneje, z njeno uporabo niso postale zakonitosti v podatkih nič bolj jasne. Uporabljal sem namreč veliko število spremenljivk binarnega tipa, katerih grafično prikazovanje ni bilo potrebno. Uporabil sem zgolj grafičen prikaz razmerja med nezaželeno in običajno pošto, ki je prikazan na sliki 3. Iz nje je razvidno, da je znašal delež nezaželene pošte 38,9%.

[10] V podatkovni bazi bi to polje imenovali »ključ«.

[11] Zgolj po naključju sta imeni teh dveh spremeljivk enaki njunima vlogama.

Slika 3: Razmerje med običajno in nezaželeno pošto v testnih podatkih



Iz modela nisem izključil nobene dodatne spremenljivke poleg že omenjenih.

4.4.3 SPREMINJANJE PODATKOV

V tretjem koraku pripravimo podatke za analizo. Dodajamo lahko nove spremenljivke, ki jih izvedemo (izračunamo) iz že obstoječih. Iščemo lahko izjemne vrednosti spremenljivk, ki bistveno odstopajo od večine vrednosti. Te vrednosti imajo lahko velik vpliv na natančnost modela, zato jih je priporočljivo poiskati in prepoznati. Naslednje pomembno opravilo je vstavljanje manjkajočih vrednosti. Pogosto se namreč zgodi, da vrednosti spremenljivk manjkajo. Smiselna nadomestitev teh vrednosti omogoči, da za predvidevanje uporabimo tudi opazovane enote z manjkajočimi vrednostmi. Sploh pri tehnikah, ki se »ne razumejo« z manjkajočimi vrednostmi, kot so npr. nevronske mreže.

V tem delu nam Enterprise Miner omogoča tudi razvrščanje v skupine in druge oblike analiziranja strukture podatkov.

Pri gradnji mojega modela se s to fazo nisem ukvarjal. Razlog je v tem, da so bili podatki »pisani na kožo« podatkovnemu rudarjenju. Dodatnih spremenljivk nisem

dodajal, saj je zelo težko iz spremenljivk binarnega (v tem primeru logičnega) tipa izpeljevati nove spremenljivke. Poleg tega je bilo precej spremenljivk slabo opisanih, nekatere pa sploh ne. Zato bi bilo izpeljevanje v tem primeru nesmiselno in tvegano početje.

Kot sem že omenil, so bili podatki popolni – brez manjkajočih vrednosti. Podatki prav tako niso vsebovali izjemnih vrednosti. Razlog je pač v samem tipu spremenljivk (binarni).

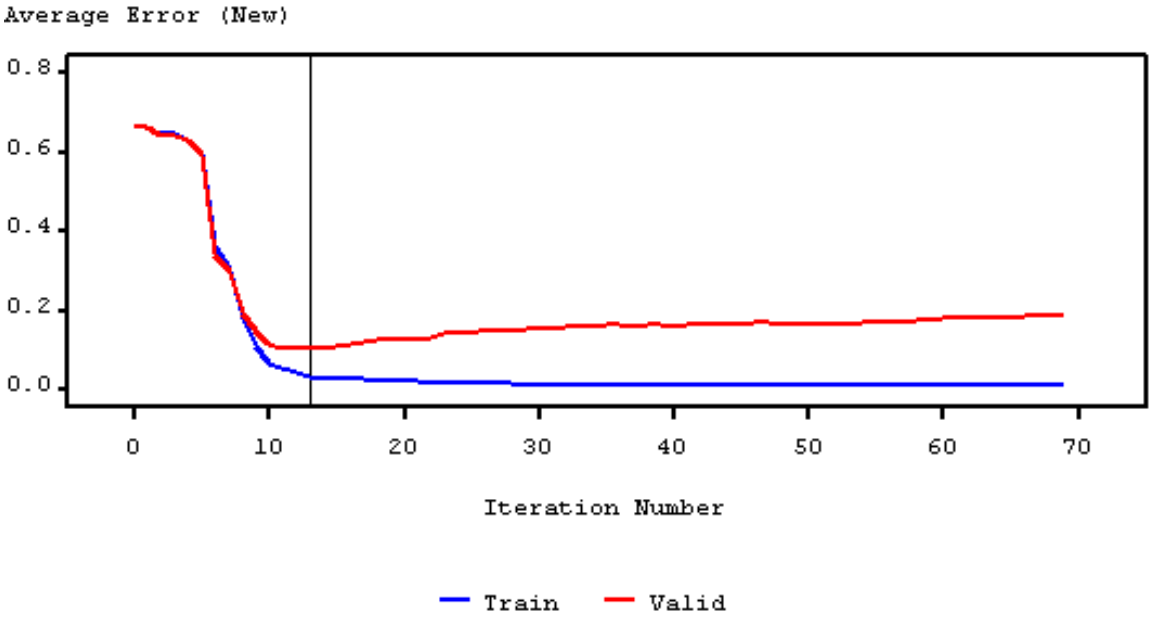
4.4.4 IZBIRA TEHNIKE IN IZDELAVA MODELA

Srčika celotnega procesa je izdelava napovedovalnega modela. Pri tem se lahko v Enterprise Minerju naslonimo na regresijske modele, na tehnike odločitvenih dreves in na nevronske mreže. Sam sem preizkusil vse tri možnosti. Na moje veliko presenečenje so se vse tri tehnike izkazale zelo dobro. Kot bomo videli kasneje so bile razlike med njimi minimalne, prav tako so dodatna nastavljanja parametrov, izbiranje alternativnih algoritmov in nasploh »piljenje« modela prinesli bore malo koristi. Glavni razlog je verjetno že omenjeno dejstvo, da so bili podatki zelo primerni za uporabo tehnik podatkovnega rudarjenja. Vseeno se je v končni fazi kot najboljši obnesel model, izdelan z **nevronskimi mrežami** (glej naslednje poglavje 4.4.5).

Sam postopek izdelave modela poteka v dveh delih: najprej model nastavimo, potem pa računalniku prepustimo, da ga »izračuna« oz. da ga izvede nad modelnimi podatki. Ta proces je računsko lahko izredno zahteven in traja tudi po več ur. To trajanje pa je koristno vnaprej omejiti: v primeru gradnje izbranega modela sem ta čas omejil na dve uri. Zakaj je bilo to potrebno, se lepo vidi na sliki 4. Po določenem času se model ne izboljšuje več in bi bilo nadaljnje vztrajanje nesmiselno. Še več, model se lahko z večanjem števila ponavljanj začne slabšati. Razlog je v tem, da se model »zapomni« učni del podatkov, saj z naraščajočo kompleksnostjo ne odkriva več splošnih zakonitosti v podatkih, ampak tudi čisto naključne (preveliko prileganje). Takšen model bi sicer zelo dobro opisal modelne podatke, vendar bi bil bolj ali manj neuporaben na še nevidenih podatkih. Zato se pri gradnji modela uporablja tudi potrjevalni del podatkov. Na grafu je t.i. preveliko prileganje lepo vidno kot naraščanje povprečne napake (slabšanje modela) pri potrjevalnem delu podatkov, medtem ko se povprečna napaka pri učnem del podatkov zmanjšuje. Da nas to zmanjševanje ne zavede, poskrbi prav potrjevalni del podatkov. Z njegovo

pomočjo lahko ugotovimo optimalno kompleksnost modela, ki zagotavlja najmanjšo povprečno napako (vertikalna črta na sliki 4).

Slika 4: Natančnost modela pri učnem (Train) in potrjevalnem (Valid) delu modelnih podatkov v odvisnosti od števila ponavljanj (primer nevronske mreže)



4.4.5 OVREDNOTENJE MODELA

Zadnji korak je namenjen ovrednotenju posameznega modela in primerjanju modelov med seboj. Pri tem se uporablja testni del modelnih podatkov. Ko izberemo najboljši model, ga uporabimo na novih, nevidenih podatkih (ang. scoring). Enterprise Miner tudi omogoča, da shranimo programsko kodo izbranega modela in jo izvajamo izven samega Enterprise Minerja. Za tovrstno izvajanje kode še vedno potrebujemo SASovo osnovno programsko okolje (base SAS), ki pa je na voljo za praktično vse platforme.

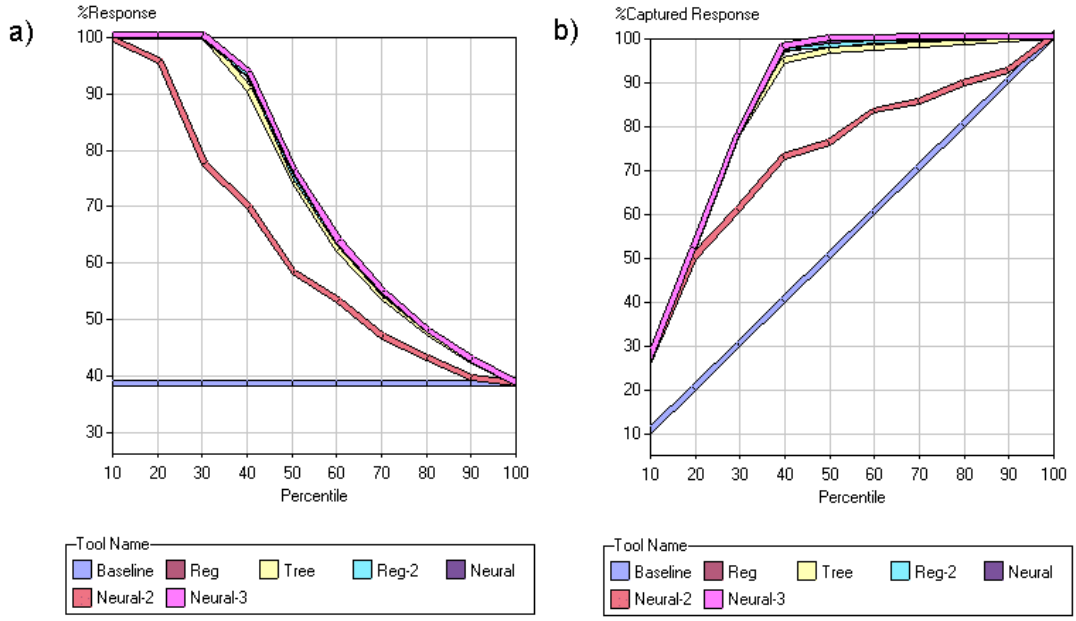
Pri ocenjevanju modelov si najlažje pomagamo z grafi. Najprej si oglejmo prikaz kumulativnega deleža odkritih e-smeti (ang. Cumulative %Response), ki je prikazan na sliki 5 (a). V splošnem za ta tip grafa razvrstimo opazovane enote v kvantile (percentile) glede na pričakovano verjetnost odziva, nato pa narišemo dejanski (kumulativni) odstotek »odzivnikov«. V našem primeru je odziv, ki nas zanima, nezaželena pošta. »Odzivnik« je torej e-sporočilo, ki spada med nezaželeno pošto. Za vsako e-sporočilo model napove verjetnost, da spada med e-smeti. Sporočila se

potem rangirajo po tej verjetnosti, od najvišje proti najnižji. Boljši je model, katerega krivulja leži višje.

Iz grafa je najprej razvidno, da so si modeli po zmogljivosti zelo podobni. Izjema je edino model Neural-2 (črta rdeče barve, glej legendo), ki sem ga vključil z namenom, da grafično prikažem razlike med dobrimi in slabimi modeli. Vidimo, da ostali modeli v zgornjih 30% - 40% zajamejo skoraj vso nezaželeno pošto. Modra vodoravna črta (Baseline) predstavlja odstotek nezaželene pošte, ki bi jo odkrili z naključnim izbiranjem.

Za razlago slike 5 (b) (ang. Cumulative %Captured Response) si namesto vprašanja »Kakšen delež pošte v kvantilu spada med e-smeti?« zastavimo vprašanje »Kakšen delež celotne nezaželene pošte se nahaja v kvantilu?«. Odgovor na slednje vprašanje dobimo iz omenjenega grafa. Diagonalna ravna črta predstavlja naključen izbor, površina med njo in krivuljo izbranega modela pa zmogljivost modela oz. koristi, ki nam jih uporaba modela prinese glede na naključen izbor. Spet velja, da je boljši model tisti, čigar krivulja leži više.

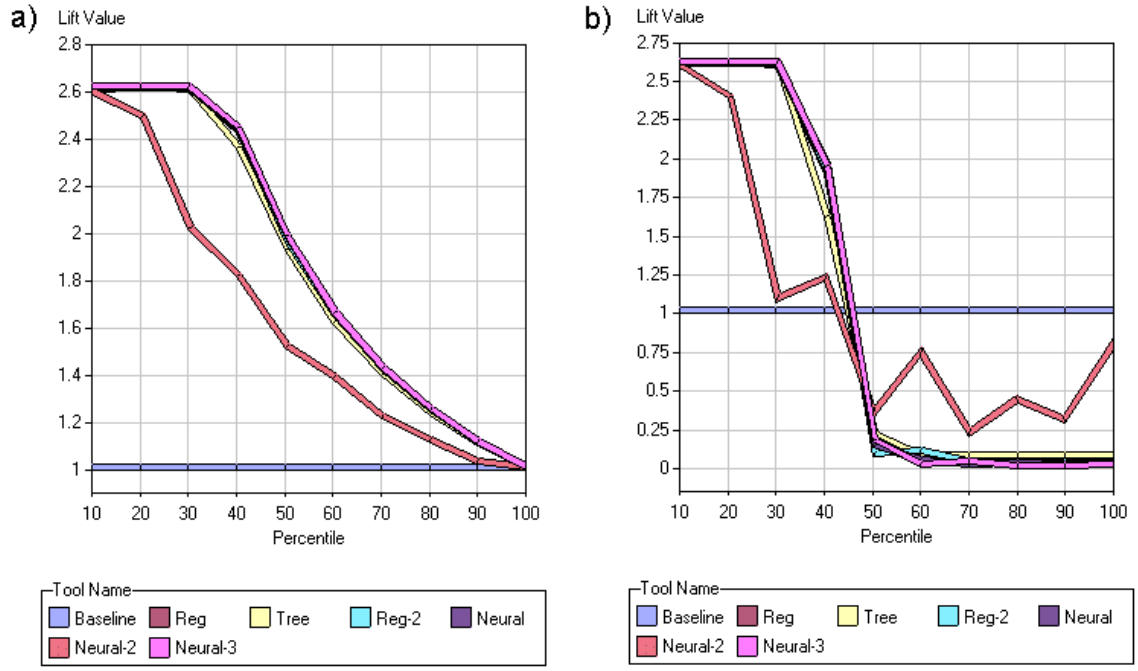
Slika 5: Kumulativni delež odkritih e-smeti (a) in kumulativni prikaz koristi (b)



Grafično pa lahko prikažemo tudi t.i. **dvig** (ang. lift value). Gre za kazalec, ki pove, koliko je model boljši v primerjavi z naključnim izbiranjem. Iz grafa na sliki 6 (a) je razvidno, da je v prvih kvantilih kumulativni delež nezaželene pošte približno 2,6-krat večji kot v celotnem vzorcu. Ta rezultat izračunamo tako, da odstotek najdenih

e-smeti delimo z dejanskim odstotkom e-smeti v testnem delu modelnih podatkov. Slika 5 (b) prikazuje enako informacijo, le da na grafu prikazane vrednosti dviga po kvantilih niso kumulativne ampak trenutne. Lepo se vidi, da je v drugi polovici kvantilov ta vrednost okrog ničle, saj ti vsebujejo zanemarljiv delež e-smeti.

Slika 6: Zmogljivost modela glede na naključni izbor



Kot že rečeno, se je najbolje obnesel model, zgrajen s tehniko nevronskih mrež. V tabeli 4 je predstavljena matrika natančnosti za izbrani model. Ciljna spremenljivka je *target*, njuni vrednosti sta predstavljeni kot »DA« (nezaželena pošta) in »NE« (običajna pošta).

Tabela 4: Matrika natančnosti modela

		DEJANSKA VREDNOST	
		DA	NE
NAPOVEDANA VREDNOST	DA	96,20%	0,74%
	NE	3,80%	99,26%

Iz tabele 4 je razvidno, da je model v testnih podatkih našel 96,20% vseh nezaželenih sporočil. Še pomembnejši podatek pa je, da je samo 0,74% običajne pošte napačno

prepoznal kot nezaželeno. Ali je ta delež res majhen? Smo lahko z njim zadovoljni? Če nič drugega, model vsaj pri testnih podatkih izpolnjuje 1% pogoj pri DMC-ju. Da je boljši od marsikatere komercialne programske opreme, bi bila verjetno preveč drzna trditev. Dasiravno te trditve matrika natančnosti ne zavrača. Za dokončno potrditev pa bo potrebno vsaj počakati rezultate klasifikacije še nevidenih podatkov[12], če že ne model neposredno primerjati s specializirano protismetno programsko opremo.

Izbrani model, ki se je izkazal kot najboljši, sem nato apliciral na neklasificirane podatke. V tem primeru je vhodne podatke predstavljala datoteka z 11.177 zapisi, ki je vsebovala enake atribute kot datoteka z modelnimi podatki, manjkal pa je seveda atribut *target*. Rezultat klasifikacije ni bila neposredna razvrstitev e-sporočil na nezaželeno in običajno pošto, ampak je model za vsako e-sporočilo ocenil verjetnost, da spada med nezaželeno pošto.

Za začetek sem določil, da vsako sporočilo z verjetnostjo nad 0,9 spada med nezaželeno pošto. Na drugi strani sem vsako sporočilo z verjetnostjo pod 0,1 označil kot običajno. »Sumljivih« sporočil, pri katerih je omenjena verjetnost znašala med 0,1 in 0,9, je bilo 163. Vprašanje je bilo, kje potegniti mejo. Nazadnje sem se z ozirom na pravilo, ki je dovoljevalo 1% običajne pošte med e-smetmi, odločil za verjetnost 0,5. Vsa sporočila z večjo verjetnostjo sem označil za nezaželena, ostala pa za običajna. V absolutnih številkah to pomeni, da sem med e-smeti razvrstil 4305 sporočil. Delež e-smeti je torej znašal 38,5% (4.305/11.177) in je za 0,4 odstotne točke manjši od deleža e-smeti v modelnih podatkih.

Primerjava deležev e-smeti med modelnimi in neklasificiranimi podatki bi lahko napeljevala k temu, da bi kazalo sprostiti kriterij za uvrstitev sporočila med e-smeti. Zahtevano verjetnost bi lahko zmanjšal (< 0,5), s tem »povečal« število nezaželenih sporočil in se približal vrednosti 38,9%. Takšno razmišljanje temelji na dejstvu, da so e-sporočila iz modelnih in neklasificiranih podatkov naključno izbrana iz iste populacije. Delež e-smeti naj bi bil torej v obeh skupinah podatkov bolj ali manj enak. Vendar bi bilo vztrajanje pri »izenačevanju« deležev v obeh skupinah tvegano dejanje, saj bi s tem med e-smeti »zajel« tudi sorazmerno veliko običajnih sporočil.

[12] Rezultati tekmovanja v času pisanja diplomskega dela še niso bili znani.

5. SKLEP

V zaostrenih konkurenčnih pogojih sodobnega sveta so postale informacije vredne več kot kdajkoli prej. Pravočasne in pravilne informacije lahko pomenijo mejo med preživetjem in propadom. Tako pomembnim dejavnikom je potrebno posvetiti vso pozornost in skoraj paradoksalno je dejstvo, da informacije pogosto ležijo na dosegu roke. Resda na prvi pogled vse niso vidne, ampak vseeno se je potrebno zavedati njihovega obstoja. Njihovo odkritje v vsakem primeru predstavlja korak k učinkovitejšemu in uspešnejšemu poslovanju. Zato je ključnega pomena, da se podjetje zave, kakšno bogastvo informacij skrivajo gomile podatkov, ki jih hrani v svojih strežnikih in se do teh informacij tudi dokoplje. To je lahko odločilni dejavnik v boju s konkurenco.

Majhen, vendar pomemben vidik izboljšanja uspešnosti poslovanja podjetja je tudi boj proti nezaželeni e-pošti, natančneje proti škodi, ki jo ta povzroča. Iluzorično je pričakovati, da se bo ta problem rešil sam od sebe. Nasprotno, predvidevanja kažejo, da se bo problem le še stopnjeval in da brez učinkovitega boja e-pošta lahko postane popolnoma neuporabna storitev. Osebno v to sicer ne verjamem, vendar bojazen obstaja. Bojazen, ki jo že in jo še bodo spretno izkoriščali razni ponudniki protismetne programske opreme.

Na koncu lahko samo še ugotovim, da je cilj mojega diplomskega dela dosežen: izdelal sem model, ki v boju proti nezaželeni pošti obeta soliden rezultat. Dokazal sem, da podatkovno rudarjenje predstavlja učinkovito orodje pri odkrivanju koristnega znanja iz ogromne količine podatkov. Obenem sem potrdil, da se s primernim pristopom lahko obvarujemo pred e-poštnim smetenjem. Nevarnost, da bodo e-smeti začele odvračati uporabnike od te priljubljene storitve, se mi zdi manjša, kot pred pisanjem tega dela. Vseeno obstaja grenak priokus, da bo za nemoteno uporabo elektronske pošte prej ali slej potrebno vložiti nekaj truda – bodisi s strani uporabnikov, sistemskih administratorjev ali koga tretjega.

LITERATURA

1. Berry Michael J. A., Linoff Gordon: Mastering Data Mining : The Art and Science of Customer Relationship Management. New York : John Wiley & Sons, Inc., 2002. 494 str.
2. Berson Alex, Smith Stephen, Thearling Kurt: Building Data Mining Applications for CRM. New York (etc.) : McGraw-Hill, 2000. 510 str.
3. Blejec Matjaž et al.: Statistika. Piran : Gea College, Visoka šola za podjetništvo, 2003. 150 str.
4. Damij Talib, Grad Janez, Jaklič Jurij: Izbrane teme iz informacijske tehnologije. Ljubljana : Ekonomska fakulteta, 1995. 316 str.
5. Frank Eibe, Witten Ian H.: Data Mining. San Francisco : Morgan Kaufmann Publishers, 2000. 371 str.
6. Han Jiawei, Kamber Micheline: Data Mining: Concepts and Techniques. San Francisco (etc.) : Morgan Kaufmann Publishers, 2001. 550 str.
7. Kennedy Ruby L. et al.: Solving Data Mining Problems through Pattern Recognition. Upper Saddle River (N.J.) : Prentice Hall PTR, 1997. 317 str., pril. CD-ROM
8. Klančar Matjaž: Vnaprej izgubljen boj? Monitor, Ljubljana, februar 2003, str. 62-79.
9. Sarka Dejan: Cluster Analysis Algorithm. [URL: http://sql.reproms.si/data/podatki/dejan/Cluster_Analysis_WP.doc], januar 2001.
10. Sarka Dejan: Decision Trees Algorithm. [URL: http://sql.reproms.si/data/podatki/dejan/Decision_Trees_WP.doc], marec 2001.
11. Zhengxin Chen: Intelligent Data Warehousing : From Data Preparation to Data Mining. Boca Raton : CRC Press LLC, 2002. 243 str.

VIRI

1. Chemical and Laboratory Supply Online – The Industry. [URL: http://www.american.edu/carmel/eg8814a/theindustry.html], 2.6.2003.
2. Data Mining Cup. [URL: http://www.data-mining-cup.com], 15.4.2003.
3. Data Mining Using Enterprise Miner Software: A Case Study Approach, First Edition. Cary, NC, USA : SAS Institute Inc., 2000. 105 str.
4. Dealing effectively with spam. [URL: http://www.gfi.com/mes/wpeliminatespam.htm], 14.5.2003.
5. Getting Started with Enterprise Miner Software, Release 4.1. Cary, NC, USA : SAS Institute Inc., 2000a. 129 str.
6. History of Computing Industrial Era 1981. [URL: http://www.thocp.net/timeline/1981.htm] , 17.10.2002.
7. Information about spam. [URL: http://spam.abuse.net/overview/], 12.5.2003.
8. Introduction to Data Mining and Knowledge Discovery, Third Edition. Falls Road (Potomac) : Two Crows Corporation, 1999. 36. str.
9. SpamAssassin.org. [URL: http://spamassassin.org], 15.4.2003.
10. Study: Spam costs businesses $13 billion. [URL: http://www.cnn.com/2003/TECH/biztech/01/03/spam.costs.ap/], 5.1.2003.
11. The Pharmaceutical Landscape. [URL: http://www.whoswho-sutter.com/pagine/pharma.htm], 8.6.2003.
12. Zakon o varstvu potrošnikov - uradno prečiščeno besedilo (ZVPot-UPB1). [URL: http://objave.uradni-list.si/bazeul/URED/2003/014/B/525663109.htm], 29.1.2003.

PRILOGE

PRILOGA 1: Slovarček tujih izrazov in kratic

- accuracy – stopnja zaupanja
- AOL (America On-Line) – ameriški ponudnik dostopa do interneta
- byte – zlog, enota za količino podatkov, običajno sestavljena iz 8 bitov. Večje enote: kB (kilobyte) – tisoč bytev, MB (megabyte) – milijon bytev.
- confidence – glej accuracy
- coverage – stopnja pokritja, podpora
- CRM (Customer Relationship Management) – upravljanje odnosov s strankami
- data mining – podatkovno rudarjenje
- DMC (Data Mining Cup) – tradicionalno študentsko tekmovanje v podatkovnem rudarjenju
- eBay – ameriški ponudnik, ki omogoča elektronske dražbe preko interneta
- GIGO (Garbage In, Garbage Out) – pravilo, da na osnovi slabih podatkov ne bomo dobili uporabnih rezultatov
- Hotmail – ponudnik brezplačne spletne e-pošte
- HTML (HyperText Markup Language) – hipertekstni označevalni jezik, standardni jezik za oblikovanje spletnih strani
- lift value - dvig
- Microsoft Corp. – ameriško računalniško podjetje
- outliers – izjemne vrednost spremenljivk
- overfitting – preveliko prileganje
- oversampling – prevzorčenje
- PR – podatkovno rudarjenje
- Prudsys – nemško podjetje, ki se ukvarja z razvojem programske opreme
- RAM (Random Access Memory) – delovni pomnilnik
- ROI (Return On Investment) – povrnitev vloženih sredstev na enoto investicije
- SAS Institute, Inc. – ameriško podjetje, ki se ukvarja z razvojem programske opreme
- SEMMA (Sample, Explore, Modify, Model, Assess) – postopek za uporabo tehnik podatkovnega rudarjenja
- spam – nezaželena elektronska pošta
- SpamAssassin – programska oprema za filtriranje nezaželene e-pošte

- support – glej coverage
- test set – testni del modelnih podatkov
- trade-off – tehtanje med dvema možnostima
- training set – učni del modelnih podatkov
- validation set – potrjevalni del modelnih podatkov
- Yahoo Mail – storitev, ki omogoča uporabo brezplačne spletne e-pošte