© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

Uloga i metodologija pretprocesiranja podataka

Kvaliteta podataka je presudni faktor o kome ovisi uspješno rudarenje podataka. Vrlo značajnu ulogu u kvaliteti podataka osim izvora podataka imaju i postupci čišćenja i pretprocesiranja podataka.
Iskustvo pokazuje da analitičar u prosjeku najviše vremena provede čisteći i pretprocesirajući podatke, i to do 80% vremena utrošenih na analize, da bi tek 20 % vremena nad pripremljenim podacima primjenjivao metode rudarenja podataka.
Jedna od glavnih prednosti sustava poslovne inteligencije očituje se u korištenju skladišta podataka kao izvora podataka za analize, što analitičaru znatno štedi vrijeme, iako ponekad mora direktno pristupiti izvornoj bazi podataka, ako neki podatak ne postoji u skladištu podataka, ili ako se ne pojavljuje na zadovoljavajućem stupnju granulacije.

Podaci u izvornom obliku mogu biti nekompletni, atributi mogu imati nedostajuće vrijednosti, ili može postojati nedostatak atributa. Isto tako može se pojaviti nekonzistentnost unutar samih podataka, primjerice nedosljednost u označavanju pojedinih kategorija ili grupa.

Govoreći o pretprocesiranju podataka možemo izdvojiti najznačajnije metodološke postupke pretprocesiranja u koje ubrajamo :

- Pronalaženje ekstremnih vrijednosti
- Dijagnostika nedostajućih vrijednosti i predviđanje nedostajućih vrijednosti
- Povezivanje relacijskih ključeva iz različitih izvora podataka
- Postizanje jednoobraznosti (konzistentnosti) u podacima
- Uzorkovanje
- Kategorizacija vrijednosti atributa
- Formiranje izvedenih atributa (eng. binning)
- Grupiranje (sažimanje podataka)
- Normiranje podataka

Često puta se unutar podataka pojavljuju ekstremne vrijednosti (eng. outliers), koje nužno ne moraju uvijek biti greške u podacima. Da bi se dijagnosticirao uzrok pojavljivanja ekstremnih vrijednosti potrebno je izvršiti dodatne analize, te nakon provedenih analiza odlučiti da li će se u analizu uči sa ekstremnim vrijednostima ili bez njih. Ponekad ekstremne vrijednosti u podacima mogu upućivati na vrlo poučnu devijaciju, te se ni u kom slučaju one ne smiju isključivati bez prethodnih razmatranja o uzrocima pojavljivanja takvih vrijednosti.

Kada se u populaciji podataka naiđe na nedostajuće vrijednosti, tada se u procesu analize podataka koriste metode predviđanja nedostajućih vrijednosti. U predviđanju nedostajućih vrijednosti često puta se koriste standardne metode rudarenja podataka.
Nedostajuće vrijednosti atributa možemo predvidjeti metodama kao što su neuralne mreže, regresijske metode, linearna interpolacija, Bayesove mreže, stabla odlučivanja i slično.

O analitičaru ovisi da li će predviđati nedostajuće vrijednosti (uz uvjet da je to opravdano), ili da će jednostavno zanemariti slogove koji su nepotpuni u bilo kojem segmentu.
U slučaju postojanja skladišta podataka, ovaj proces je znatno pojednostavljen, jer postoje već gotovi elementi sadržani u dimenzijskim kockama koji mogu ući direktno u proces analize.
Nije rijedak slučaj da se spomenute metode koriste i prilikom ETL procesa u skladištima podataka, posebice ako ju ona podloga za rudarenje podataka.
Problemi oko pretprocesiranja podataka još više dolaze do izražaja kada vršimo prikupljanje podataka iz različitih izvora, gdje često puta ne postoji jednoznačna definicija relacijskih ključeva, različiti sustavi kategoriziranja unutar podataka, nedoslijednost vrijednosti atributa vezanih uz iste relacijske ključeve i slično.
Za razliku od skladišta podataka koje teži ka transparentnosti i uglavnom unaprijed ima predefinirane atribute sa kojima će se vršiti obrada, te se u skladu sa tim ETL konstruira za čitav niz primjena, kod rudarenja podataka stvari su nešto kompliciranije. Naime, rudarenje podataka često puta se ne mora referencirati na iste izvore podataka, već uz standardne izvore podataka unutar poduzeća možemo koristiti čitav niz eksternih informacija sadržanima u različitim formatima. Upravo u takvim slučajevima dolazi do izražaja kompleksnost problematike, jer pridodjeljivanjem novih izvora podataka mogu se promijeniti kako i metodologija pretprocesiranja i čišćenja podataka, tako i ponderi važnosti pojedinih atributa ili skupine atributa s obzirom na ciljnu varijablu.
Imajući u vidu spiralni razvoj sustava, vidljivo je zašto upravo taj tip razvoja sustava odgovara ovoj problematici.

U djelokrug pretprocesiranja podatka spada i metodologija uzorkovanja.
Iako na temelju osobnog iskustva uvijek treba težiti zahvatiti kompletnu populaciju, dio te populacije izdvojiti kao testni uzorak, a ostatak populacije koristit za treniranje modela, ponekad je fizički nemoguće zahvatiti cjelokupnu populaciju.

U takvim slučajevima koristimo metode uzorkovanja s težnjom da uzorak što reprezentativnije predstavlja populaciju. Isto tako metoda uzorkovanja koristi se i kod cijepanja osnovne populacije na uzorak za učenje modela i testni uzorak.

Postoji čitav niz statističkih metoda za uzorkovanje i ocjenjivanje reprezentativnosti uzorka, te ih možemo koristiti u pretprocesiranjima podataka s ciljem rudarenja podataka.

Vrlo važan postupak koji se koristi u pretprocesiranju podataka je formiranje kategorija na temelju podataka. Tako primjerice možemo kategorizirati i dodijeliti vrijednost atributa kategorija dobi “od 10 do 20” sve one slogove čija se vrijednost atributa dob nalazi unutar zadanih granica.

Isto tako možemo formirati kategorije mlad, star na temelju vrijednosti atributa dobi.
Postoji metodologija koja je vrlo slična opisanoj metodologiji (eng. binning),a može se ilustrativno prikazati kao formiranje novih kategorija koje obuhvaćaju vrijednosti sortiranog niza, pri čemu svaka od kategorija sadrži jednak broj vrijednosti definiranog niza. Isto tako vrijednosti tih kategorija mogu biti prosjeci u okviru izdvojenih vrijednosti, ili pak vrijednosti neke druge statističke funkcije.

© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

Daljnja metodologija pretprocesiranja može se primjenjivati s ciljem sažimanja podataka, na način da se velika populacija podataka sažima grupiranjem po određenim kriterijima.
Ovo se može ilustrirati primjenom SQL upita sa GROUP BY klauzulom:

SELECT stupanj_obrazovanja, AVG(dob) AS pdob, AVG(prihodi) as pprihodi FROM tablica GROUP BY stupanj_obrazovanja

Pirkazan SQL upit iz tablice sažima podatke na temelju stupnja obrazovanja, računajući prosjek dobi i prihoda prema stupnju obrazovanja.

Normiranje podataka kao postupak je dosta interesantno kod korištenja metoda poput klasteriranja, neuronskih mreža gdje je potrebno izbjeći preveliki utjecaj pojedine varijable koja gravitira kao visokim apsolutnim iznosima.

Od metoda normiranja podataka koje se najviše koriste u rudarenju podataka spominjemo:

- Min-max normiranje
- Z-skaliranje
- Decimalno skaliranje

Min-max normiranje svodi se na linearnu transformaciju izvornog raspona podataka na novi raspon, najčešće između 0-1.

$$y' = \frac{y - \min}{\max - \min}\left(\max' - \min'\right) + \min'$$

gdje oznake formule predstavljaju :

min' – nova, normirana minimalna vrijednost
max'- nova, normirana maksimalna vrijednost
y' - nova normirana vrijednost atributa
min – minimalna vrijednost originalnog niza
max - maksimalna vrijednost originalnog niza
y - Izvorna vrijednost atributa

Ova metoda je primjenjiva kada poznajemo minimalnu i maksimalnu vrijednost originalnog niza, u slučaju da ne znamo minimalnu i maksimalnu vrijednost originalnog nita, tada koristimo Z skaliranje.

Formula Z sklaliranja glasi:

$$y' = \frac{y - srednja\_vrijednost}{st\_devijacija}$$

Korisna metoda normiranja je i decimalno skaliranje koju možemo izraziti formulom:

© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

$$y' = \frac{y}{10^n}$$

y – originalna vrijednost
n - broj znamenaka maksimalne apsolutne vrijednosti

Normiranje se može provesti direktno na relacijskoj tablici, ili pak na kreiranom "view-u" u bazi podataka na način kao što je to slučaj kod DB2 baze:

UPDATE tablica
SET z=(placa-srednja_vrijednost)/stdev

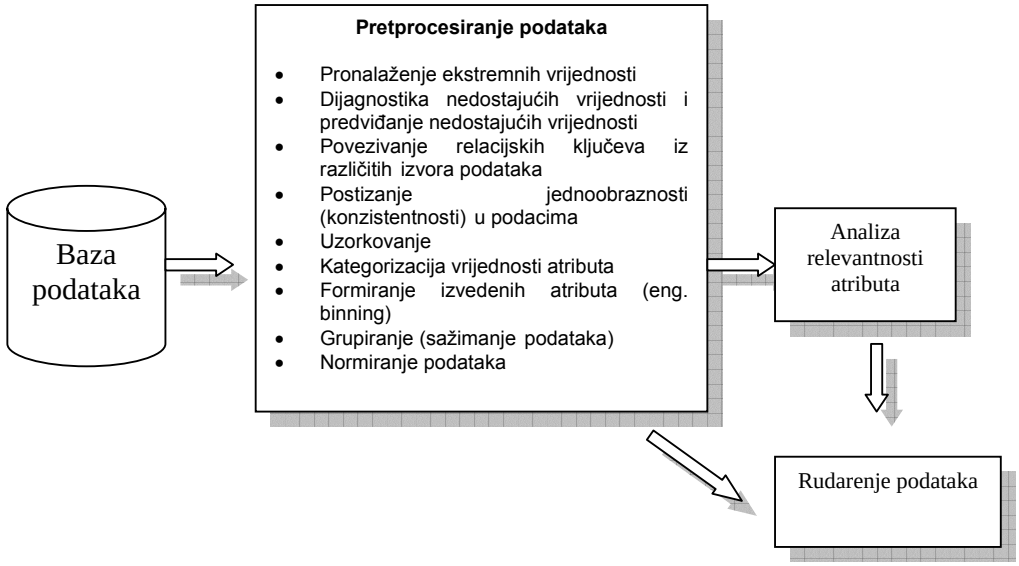
Još veća efikasnost ovog koncepta se postiže parametrizacijom u okviru spremljenih procedura na razini baze podataka (eng. storage procedures) koje se parametrizirano pozivaju prilikom pretprocesiranja podataka. Kreiranje ovakve vrste procedura uvelike pomaže kod kontinuiranog jednoobraznog pretprocesiranja koje se provodi u određenim vremenskim intervalima.
SQL može biti od velike koristi kod pretprocesiranja podataka, posebice ako se optimiziraju upiti koji pristupaju velikim količinama podataka s ciljem modifikacije i kreiranja izvedenih tablica za potrebe pretprocesiranja podataka.

Programski paketi koji se koriste u rudarenju podataka imaju već unaprijed definirane module za pretprocesiranje podataka.
Programski paketi poput SPSS-a, koji u sebi sadrže skriptni jezik pružaju mogućnost brzog i efikasnog programiranja modula pretprocesiranja podataka.
Zbog kompleksnosti područja nije rijedak slučaj da se pretprocesiranje vrši uz pomoć modula isprogramiranih u nekim od programskih jezika, te se tek završni procesi pretprocesiranja obvljaju u okviru skriptnih jezika.

Jedna od neizostavnih metodologija koja prati sve etape rudarenja podataka od čišćenja, preko pretprocesiranja i rudarenja podataka je vizualizacija podataka. Vizualizacija podataka može na jednostavan i efikasan način ukazati na osnovne smjernice daljnje analize u bilo kojoj etapi.

SPSS programski paket sadrži vrlo moćne alate za vizualizaciju podataka, pri čemu prednjači modul interaktivne grafike koji omogućava kreiranje 3D vizualnih modela podataka koji se mogu rotirati u prostoru.
Od velike koristi u "upoznavanju" podataka mogu biti i standardne metode deskriptivne statistike, koje nam mogu pomoći kod procjenjivanja osnovne karakteristike cjelokupne populacije kao što je homogenost populacije, stupanj disperzije populacije, tendencije "nagiba" populacije s obzirom na jednu ili niz promatranih varijabli i slično.
Isto tako analiza korelatornih odnosa i krostabulacija među varijablama može nam dati smjernice za dublje analize.
Ovakav pristup omogućava temeljno upoznavanje populacije, te s obzirom na cij analize uspješnije građenje modela rudarenja podataka.
Spomenuti metodološki postupci neki su od najčešćih postupaka koji se provode nad izvornim oblikom podataka s ciljem "uvođenja reda" među podacima. Nakon provođenja ovakvog seta postupaka moguće je još temeljitije upoznavati i

pretprocesirati podatke metodama relevantnosti atributa o kojima će biti riječi u poglavlju o analizi relevantnosti atributa .
Slika x.4. prikazuje odnos analize relevantnosti atributa nakon provođenja pretprocesiranja podataka.



Slika 1 Odnos analize relevantnosti atributa nakon provođenja pretprocesiranja podataka.

Iz slike 1 vidljivo je da metodika pretprocesiranja podataka stvara preduvjete za analize relevantnosti atributa.

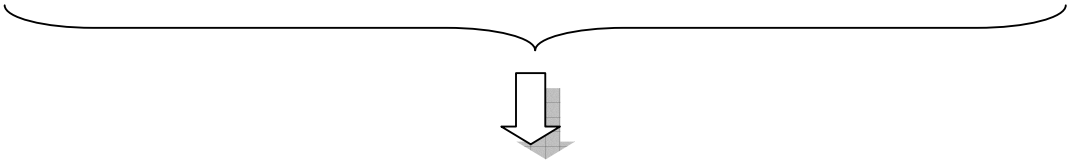
Kao što je vidljivo iz slike, podaci se mogu rudariti nakon primjene spomenute metodologije pretprocesiranja podataka, no analiza relevantnosti atributa nam daje korisne informacije relevantnosti s obzirom na ciljnu varijablu. Takvo saznanje nam može pomoći prilikom modeliranja modela rješenja metodama rudarenja podataka.
Analiza relevantnosti atributa nije preduvjet za uspješno rudarenje podataka, kao što je to slučaj sa pretprocesiranjem podataka. U nekim radovima vezanima za rudarenje podataka analiza relevantnosti atributa promatra se kao sastavni dio pretprocesiranja podataka.
Razlog posebnog razmatranja analize relevantnosti atributa u našem slučaju proizlazi iz razloga njene korisnosti u upoznavanju osnovne populacije nad kojom vršimo analize. Iskustvo pokazuje da ovakve vrste analiza mogu značajno pridonijeti u razumijevanju odnosa među atributima, a samim time i izbora adekvatne metode rudarenje podataka.

Krajnji cilj čišćenja i pretprocesiranja podataka pri tradicionalnom rudarenju podataka je formiranje jedinstvene tablice nad kojom primjenjujemo metode rudarenja podataka.
Shematski takav proces možemo prikazati slikom x.5

© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

Šifra korisnika	Datum i godina rođenja	Adresa	Poštanski broj	Grad
8080	10.06.1972	Ulica lipa 14	10 000	Zagreb
1508	09.09.1943	Miroslava Krleže 31	42 000	Varaždin
2812	17.07.1956	E.A. Poe-a 19	44 000	Sisak
0505	22.07.1976	Tolstojeva 14	47 000	Ogulin
...	...	...	...	...

Šifra korisnika	Broj računa	Datum transakcije	Iznos	Način plaćanja
0505	234	14.07.2003	423,22	MC
1508	235	14.07.2003	28,21	Gotovina
8080	236	15.07.2003	311,16	VISA
0505	237	15.07.2003	421,19	MC
...	...	...	...	...



Dobni razred	Županija	Prosječni mjesečni iznos transakcija prema načinu plaćanja	Način plaćanja	Pauza u dolasku dulja od mjesec dana	Registriran ponovni dolazak nakon pauze
21-25	Zagrebačka	400-600	MC	NE	-
46-50	Sisačko-Moslovačka	600-700	MC	NE	-
61-65	Bjelovarsko-Bilogorska	200-300	Visa	DA	DA
21-25	Zagrebačka	100-200	Gotovina	DA	NE
...	...	...	...	...	...

Slika 2. Shematski prikaz pretprocesiranja podataka i formirane tablice za rudarenje

Iz slike je vidljivo da na temelju cilja analize pripremamo podatke za pretprocesiranje. Same funkcije agregiranja, grupiranja i formiranja razreda ovise o cilju analize što se vrlo transparentno vidi iz slike x.5
Dakle, predprocesiranje podataka treba provoditi u skladu sa ciljevima analize, jer ako prtprocesiranje nije provedeno u skladu sa ciljevima analize može prouzrokovati gubitak detaljnosti podataka i njihove reprezentativnosti.
Isto tako, kao što se vidi iz kolone pretprocesirane tablice, postoji čitav niz izvedenica atributa ili njihovih kombinacija. Tako na osnovu korištenje metoda analiza vremenskih serija iz datuma transakcije možemo ekstrahirati podatke o pauzama u dolascima klijenata, te o njihovom ponovnom dolasku nakon pauze nedolaženja kao što je to prikazano na slici.
Isto tako u realnim analizama moraju se obuhvatit kategorije kao što su učestale pauze, neprekidni dolasci i slično, kako bi model bio potpuniji, a segmentacija preciznija.

© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

Prikazana tablica može biti primjerice podesna za otkrivanje pravila koja skupina klijenata je prekinula korištene usluga više od mjesec dana (eng. churn) korištenjem primjerice stabla odlučivanja.
Kao rezultat obrade možemo hipotetski dobiti informaciju da su to svi oni klijenti koji imaju prosječni iznos transakcija između 100-300 kuna u periodima kada su koristili naše usluge. Isto tako hipotetski možemo dobiti informacije da su to klijenti koji pretežno plaćaju gotovinom i imaju prosječni iznos transakcija između 100-300 kuna.
Isto tako hipotetski to mogu biti korisnici MC kartice.

Na što nas ovakve hipotetski rezultati analize mogu navesti ?
Prvo i osnovno moramo tražiti odgovor na pitanja zašto ?

Zašto klijenti koji imaju prosječni iznos transakcija između 100-300 kuna u periodima kada su koristili naše usluge u primjerice 85% slučajeva prestaju bit naši klijenti ?

To može implicirati pitanja tipa:

- Da li konkurencija daje veće popuste na gotovinska plaćanja ?
- Koje su karakteristike klijenata koji su se vratili nakon duže pauze?
- Koje artikle su ti klijenti prije kupovali kod nas ?
- Da li konkurencija nudi isti asortiman artikala koje su ti klijenti prije kupovali kod nas ?

Kada primjerice dijagnosticiramo koje artikle su ti klijenti prije kupovali kod nas, možemo vidjeti da li su porasle cijene tim artiklima ili skupinama artikala u odnosu na konkurenciju.
Na ovaj način možemo potvrđivati ili odbacivati hipoteze o uzorcima, kombinirajući interne i eksterne podatke.

Također možemo formirati Bayesovu mrežu koja će ne osnovu modela procjenjivati kod kojih je klijenata najveći rizik od duže pauzu sa povratkom nakon te pauze, i kod kojih je klijenata najveći rizik od potpunog prekida poslovnog odnosa.

Ako se hipotetski ispostavi da korisnici MC kartice u 80 % slučajeva imali dužu pauzu sa povratkom nakon te pauze, postoji hipotetska mogućnost (što treba istražiti) da je primjerice konkurencija pokrenula kampanju kao što je to sudjelovanje u nagradnoj igri ako se određeni iznos potroši preko MC u njihovim prostorima, te postoji usprkos kampanji lojalnost tog segmenta klijenata.

Isto tako možemo na temelju analize lojalnosti vidjeti koliki je broj klijenata u potpunosti odustao od korištenja naših proizvoda i usluga privučen reklamnom kampanjom iz spomenutog primjera, te na uzorku podataka tih bivših klijenata saznati njihove karakteristike kako bismo u budućnosti izbjegli ovakve situacije.

Ovakve situacije nam mogu biti poticaji za analizu dobiti/gubitka ako i naše poduzeće organizira ovakvu nagradnu igru.
Daljnja strategija može biti preusmjerena na analizu tržišta, te “preotimanje” klijenata konkurentskoj firmi na temelju procijenjene kampanje.

© dr.sc. Goran Klepac
www.goranklepac.com
goran(AT)goranklepac.com

Ovaj jednostavan primjer može se promatrati sa aspekta povećanog priljeva novih klijenata. U tom slučaju želimo u skladu sa koncepcijom unapređenja odnosa sa klijentima otkriti njihove karakteristike, segmentirati ih i pristupati im na temelju otkrivene sličnosti sa postojećim klijentima.
U skladu s tim treba provesti adekvatno pretprocesiranje podataka koje za krajnji cilj ima formiranje jedinstvene tablice za rudarenje podataka.

Ako nakon provedenog pretprocesiranja podataka metodama rudarenja podataka otkrijemo određenu skupinu artikala koju preferira mlađa populacija klijenata koji žive u Zagrebačkoj županiji, možemo novopristiglim klijentima sa sličnim karakteristikama ponuditi rabat na te artikle.
Na taj način možemo povećati koeficijent obrtaja te vrste robe, a sa druge strane povećavamo lojalnost klijenata.
Nadalje, segmente klijenata možemo pratiti s obzirom na vrijeme, te isto tako možemo pratiti potrošnju određenih tržišnih segmenata određenih tipova preferirane robe.
Trendovskom analizom koja može biti uklopljena u skladišta podataka sa mogućnošću “dril up” i “drill down” procesa, možemo pratiti pad ili rast interesa s obzirom na granulaciju tržišnih segmenata, što može biti putokaz za promjenu strategija vezanih uz odnose sa klijentima.
Ovdje dolazi do izražaja sinergija rudarenja podataka i skladišta podataka, te njihova međuzavisnost.
Postoji čitav niz odgovora i putokaza koje nam mogu pružiti podaci.
Prateći metodologiju spiralnog pristupa, ako analiza ne da zadovoljavajuće rezultate, možemo ponovo pretprocesirati podatke uvrštavajući u model analize i neke nove atribute, koji će nas dovesti do cilja analize.

Na žalost, receptura ne postoji. Da postoji, pristup rudarenju podataka ne bi se temeljio na spiralnom pristupu. Zbog kompleksnosti materije koju smo pokušali prikazati kroz ovaj jednostavan primjer poželjno je široko multidisciplinarno obrazovanje analitičara koji rudari podatake.