Analiza podataka (Data Mining)

Seminarski rad u okviru kursa
Tehničko i naučno pisanje
Matematički fakultet

Anđela Milićević, Luka Dekić
mi16053@alas.matf.bg.ac.rs, mi16295@alas.matf.bg.ac.rs

Sažetak

Zbog sveprisutnog procesa globalizacije, savremene uslove poslovanja karakterišu neizvesnost, rizik i konkurencija. Preduzeća moraju svakodnevno da se bore za održavanje tržišnog učešća i ostvarivanje boljih rezultata. Usled intenzivnog razvoja informatičke infrastrukture skoro sve firme, posebno one veće, čuvaju velike količine podataka o poslovanju, klijentima i kretanjima u okruženju. Dnevni unos podataka koje velike firme pohranjuju u svoje baze, meri se u terabajtima. Međutim, to su sirovi podaci, neadekvatno strukturirani i različitih formata pa nemaju veliku upotrebnu vrednost. Zato ih je neophodno pripremiti, analizirati i pretvoriti u informacije koje preduzeću mogu obezbediti ostvarenje poslovnog uspeha. S obzirom na to da se radi o velikim količinama podataka, prosto je nemoguće da čovek sam vrši analize. One se prepuštaju za to posebno određenim tehnologijama. Najvaznija od njih je *Data Mining*, čija je svrha pronalaženje skrivenih obrazaca u podacima, povećavanje njihove upotrebljivosti i transformacija podataka u korisno znanje.

1 Uvod

Data Mining je proces ,,rovarenja" po sirovim informacijama uz pomoć računara i vađenja njihovog značenja. Zahvaljujući Data Miningu, moguće je predvideti trend tržišta ili ponašanje korisnika i na taj način obezbediti uspeh firme ili proizvoda. To se postiže analizom podataka iz raznih perspektiva i pronalaženjem veza i odnosa između naizgled nepovezanih informacija.

Jedan zanimljiv primer ilustruje prethodnu tvrdnju. Lanac supermarketa u Americi je, koristeći Oracleov softver za analizu podataka, otkrio da su muškarci koji su kupovali pelene četvrtkom i subotom najčešće kupovali i paket piva. Dublja analiza otkrila je da su ovi kupci sedmičnu kupovinu obavljali subotom. Četvrtkom su kupovali samo nekoliko proizvoda, a pivo su kupovali kako bi im se našlo za dolazeći vikend. Zahvaljujući ovoj informaciji, lanac supermarketa je povećao prihode tako što je vitrinu sa pivom pomerio bliže polici sa pelenama. Takođe, četvrtkom i subotom su pivo i pelene prodavani po punoj ceni, bez posebnih popusta. Verovatno bi svakom ljudskom ekspertu veza između muškaraca, pelena, piva i određenih dana u nedelji promakla, ali ne i nepristrasnoj logici kompjutera. Ovaj primer je preuzet sa [1].

2 Data Mining

2.1 Upoznavanje sa pojmom Data Mining

Postoji nekoliko definicija Data Mininga.

* Data Mining se može definisati kao proces pronalaženja skrivenih zakonitosti i veza među podacima. To je tehnika pretraživanja podataka u cilju identifikacije traženih uzoraka i njihovih međusobnih relacija. Jednostavno rečeno, Data Mining je postupak izdvajanja interesantnih, novih i potencijalno korisnih informacija, sadržanih u velikim bazama podataka.
* Data Mining je multidisiplinarno područje koje obuhvata baze podataka, ekspertne sisteme, teoriju informacija, statistiku, matematiku, logiku i čitav niz drugih oblasti.
* Data Mining se zove i Knowledge Discovery in Databases (KDD) tj. otkrivanje znanja u bazama podataka. To je proces analize koji omogućuje korisnicima da shvate sisteme i veze između njihovih podataka. On omogućava sagledavanje informacija na način koji ranije nije bio moguć.

2.2 Evolucija

Data Mining je nova tehnologija, koja se naglo razvila zahvaljujući razvoju računarske tehnologije. Iako kao zaseban pojam postoji tek od pre 15 godina, razvoj Data Mining-a započet je jos pedesetih godina proslog veka. Ove metode tada nisu nazivane Data Mining tehnikama, ali su se primenjivale u velikoj meri i to uglavnom u svrhu naučnih istraživanja i eksperimenata. Sa razvojem računarske tehnologije, preduzeća su mogla da čuvaju sve veće i veće količine podataka u svojim bazama, pa je omogućena komercijalna upotreba velikog broja DM tehnika u svrhe poslovnog odlučivanja.

2.3 Neophodna infrastruktura

Softver za Data Mining danas je pristupačan i za velike mejnfrejm računare i za samostalne PC platforme. Dva osnovna uslova za izbor odgovarajućeg softvera jesu veličina baze podataka i kompleksnost upita. Velika baza podataka sa sobom povlači veliki broj podataka koji treba skladištiti i održavati, pa samim tim zahteva moćniji sistem. Kompleksnost upita i njihov veliki broj takođe povećavaju potrebu za procesorskom moći. Rad s velikim bazama podataka značajno ubrzava paralelno procesiranje. Stotine paralelno vezanih (običnih) kompjutera mogu da postignu performanse jednog superkompjutera. [1]

2.4 Faze u procesu Data Mininga-a

Životni ciklus jednog Data Mining projekta sastoji se iz osam koraka:

1. *Sakupljanje podataka* – Poslovni podaci su uskladišteni u brojnim sistemima, na Internetu, u bazama podataka kompanija i prvi korak predstavlja prenos relevantnih podataka u bazu podataka gde se podaci analiziraju. Ponekad postoji i skladište podataka što olakšava dalji rad, ali u velikom broju slučajeva podaci koji su sakupljeni mogu biti nedovoljno korisni za analizu, pa se zbog toga neophodni

podaci moraju sakupiti iz drugih izvora. Nakon što se sakupe, podaci se mogu semplovati da bi se smanjila veličina trenutnog skupa podataka. U mnogim slučajevima, obrasci koji su pronađeni na skupu od 50 000 kupaca su isti kao i oni pronađeni na skupu od 1 000 000 kupaca.

2. *Filtriranje podataka i transformacija* – Ovo je najintenzivniji korak u Data Mining projektu kad su resursi u pitanju. Cilj filtriranja podataka je odstranjivanje irelevantnih i suvišnih informacija iz skupa podataka. To podrazumeva uklanjanje duplih i nepotpunih podataka, njihovu tranformaciju i jedinstven sistem podataka, izabiranje podgrupa podataka, odredjivanje broja promenjivih sa kojima je moguće raditi. Cilj transformacije podataka je promena izvornog podatka u drugačiji format tipa podataka. Postoje razne tehnike koje se mogu primeniti za korak filtriranja i transformaciju podataka, a neke od njih su transformacija tipova podataka, neprekidna transformacija kolona, grupisanje, rad sa vrednošću koja nedostaje, brisanje abnormalnih slučajeva.
3. *Kreiranje i izbor modela* – Tek kada se podaci filtriraju i kada se promenljive transformišu u pogodne tipove podataka, može se započeti sa kreiranjem modela. Pre kreiranja modela treba da razumemo cilj Data Mining projekta i vrstu zadatka koji će se koristiti. Za svaki Data Mining problem postoji nekoliko odgovarajućih algoritama. Preciznost algoritma zavisi od prirode podataka kao što su: broj stanja atributa koji se koriste za predviđanje, prenos vrednosti svakog atributa, veza između atributa itd. U ovom početnom delu projekta potrebno je sastaviti tim poslovnih analitičara koji su eksperti u određenoj oblasti.
4. *Procena kvaliteta modela* – U delu kreiranja modela mi kreiramo skup modela koristeći algoritme i tehnike DM-a , ali nakon kreiranja moramo izvršiti i evaluaciju tog modela. Postoji nekoliko popularnih alata za evaluaciju kvaliteta modela. Najpoznatiji je lift dijagram. On koristi već istreniran model kako bi predvideo vrednosti koje će se dobiti iz skupa podataka koji se testira. Na osnovu vrednosti koje se dobiju i verovatnoće on grafički prikazuje model na dijagramu.
5. *Kreiranje izveštaja* – Nakon kreiranja modela i evaluacije kvaliteta tog modela, vrši se kreiranje izveštaja. Postoje dva osnovna tipa izveštaja, izveštaji o pronađenim obrascima i izveštaji o predviđenim vrednostima modela.
6. *Ocenjivanje modela* – U mnogim Data Mining projektima, pronalaženje obrazaca i modela je samo pola posla. Konačni cilj je upotreba tog modela za predviđanje. Predviđanje se još naziva i scoring u DM terminologiji. Da bismo dobili predviđene vrednosti moramo da imamo već istrenirani model i skup novih podataka.
7. *Integracija Data Mining modela u aplikaciju* – Sve više poslovnih aplikacija uključuje Data Mining komponentu. Na primer CRM (Customer Relationship Management) aplikacije mogu imati Data Mining osobine koje grupiše kupce u segmente, ERP (Enterprise Resource Planning) aplikacije mogu imati Data Mining osobine koje im koriste da predvide obim proizvodnje. On-line knjižara može dati potencijalnim kupcima preporuke knjiga. Integrisanje Data Mining osobina, pogotovo komponente za predviđanje, u aplikacije jedan je od bitnijih koraka Data Mining projekta. Ovo je ključni korak za uvođenje Data Mining-a u masovnu upotrebu.

8. *Upravljanje modelom* – Održavanje statusa Data Mining modela predstavlja pravi izazov. Svaki Data Mining model ima svoj životni ciklus. U nekim oblastima primene, obrasci su relativno stabilni i modeli ne zahtevaju učestalo ponovno treniranje modela. Ali u mnogim oblastima obrasci se menjaju često. Trajanje jednog Data Mining modela je ograničeno. Nova verzija modela se mora praviti često. Određivanje preciznosti i kreiranje novih verzija ovog modela bi trebalo biti postignuto korišćenjem automatizovanih procesa.

2.5 Data Mining tehnike

Analitičke tehnike koje se koriste u Data Miningu u najvećem broju slučajeva su odavno poznate matematičke tehnike i algoritmi. Iako je DM mlada tehnologija, koriste se ranija saznanja. Ono što je povezalo ta saznanja i velike baze podataka jeste pojeftinjenje prostora za skladištenje podataka i procesorske snage. Kako bi se problemi što brže i efikasnije rešavali, poslednjih godina razvijen je veliki broj tehnika, algoritama i metoda. Sve su one svrstane pod nazivom *Data Mining tehnike* i mogu se podeliti u dve grupe:

1) Discovery data mining – tehnike za otkrivanje novih znanja
2) Predictive data mining – tehnike za predviđanja

Neke tehnike su detaljnije objašnjene u nastavku. Pored njih postoji niz drugih algoritama na kojima se temelje modeli za Data Mining, a neki od njih dati su u tabeli 2.5.

fuzzy logika	(fuzzy logic)
memorijski zasnovano rasuđivanje	(memory based reasoning)
klastering	(clustering)
analiza potrošačke korpe	(market basket analysis)

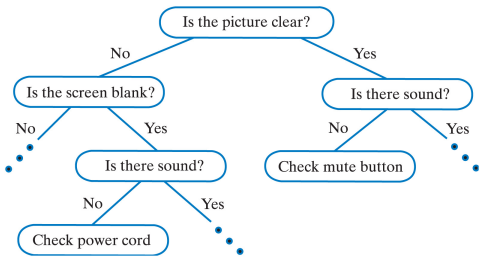
2.5.1 Metoda najbližeg suseda (Nearest neighbor classification)

Nearest neighbour classification jedna je od najstarijih tehnika koja se primenjuje u Data Miningu za klasifikaciju podataka. Zbog načina rada, koji je sličan ljudskom načinu razmišljanja, ova metoda je jedna od najjednostavnijih. Zasniva se na traženju podataka koji imaju najsličinija svojstva i poznato ponašanje. Podatak koji ima najsličnija svojstva je najbliži sused, pa se pretpostavlja da će se slično i ponašati. Pitanje algoritma je kako odrediti ko je najbliži sused. Jedan od najjednostavnijh načina je upotreba euklidske geometrije u n-dimenzionom prostoru. [2]

2.5.2 Stablo odlučivanja (Decision Tree)

Decision Tree je veoma popularan metod za klasifikaciju i odlučivanje. To je tehnika odlučivanja utemeljena na odnosima između strategija i stanja. Koristi se za rešavanje problema u finansijama, bankarstvu, osiguranju. Korišćenjem serije pitanja i pravila za kategorizaciju podataka, predviđaju se ishodi. Stablo odlučivanja nastaje grananjem kao posledica ispunjenja uslova klasifikacijskih pitanja. Svako pitanje će podeliti podatke u podskupove koji su homogeniji nego viši skup. Ako pitanje ima dva odgovora, tada će kao odgovor na pitanje nastati dva podskupa.

Koliko pitanje ima odgovora toliko će podskupova nastati. Samim tim vrši se klasifikacija pojedinih podataka. Predviđanje ponašanja pojedinog klijenta može se izvesti na temelju njegovog pripadanja pojedinom skupu (u koji je svrstan na osnovu niza pitanja i uslova), za koji se zna kako će se ponašati. Prilikom izgradnje stabla odlučivanja važno je znati postaviti pravo pitanje. Pitanje je bolje ako će se njime bolje organizovati podaci, odnosno ukoliko će se nakon toga stvoriti podskupovi koji su homogeniji. Na slici 1 prikazan je jedan primer stabla odlučivanja. [2]



Slika 1: Stablo odlučivanja

2.5.3 Neuronske mreže (Neural networks)

Ova tehnika Data Mining-a zamišljena je da deluje slično ljudskom mozgu. Kao što ljudski mozak nakon učenja izvlači određene pretpostavke na osnovu ranijih zapažanja, tako i ove mreže predviđaju promene i dešavanja u sistemu nakon procesa učenja. DM na osnovu ove tehnike počinje ,,učenjem" mreže pomoću podataka koji su već poznati, a koji se odnose na vrednost koju želimo prognozirati. Nakon toga znanje se proverava, sve dok rezultati provere ne budu zadovoljavajući. Ceo proces se svodi na sledeće: prvo se neuronskoj mreži daju određeni podaci za koje već znamo izlazne vrednosti, na osnovu ovih podataka neuronske mreže prepoznaju obrasce i pravila, zatim se na osnovu ovih obrazaca i funkcija istražuju gomile podataka koje preduzeća imaju u svojim bazama. [2]

Neuronske mreže su najkomplikovanija metoda, ali daju najtačnije modele. Nuronske mreže nastale su pokušajima imitiranja rada mozga i nervnog sistema čoveka. Osnovna ćelija neuronskih mreža je neuron. Neuron svoj izlaz temelji na kombinaciji niza ulaza pomnoženih sa odgovarajućim težinama. Neuronska mreža sastoji se od niza neurona koji su međusobno povezani. Prilikom projektovanja neuronske mreže potrebno je odrediti strukturu (broj neurona i njihove međusobne veze). Da bismo stvorili model predviđanja upotrebom neuronskih mreža potrebno je definisati težine pojednih veza. To se postiže treningom neuronske mreže. Daju joj se test podaci i zatim se koriguje odgovor koji daje, ako je netačan. Neuronska mreža će tada korigovati težine pojednih veza između neurona. Ako je prethodni neuron dao tačan odgovor vezi prema njemu, težina će se povećati, dok će se u suprotnom smanjiti. S vremenom neuronska mreža uči, pa sa povećanjem broja treninga daje sve tačnije rezultate.

2.6 Primena

Data Mining najvecu primenu ima u trgovini, radi poboljšanja prodaje nekih proizvoda. Pored toga, primenjuje se i u bankarstvu, gde je, na primer, moguće na osnovu ranijih slučajeva odrediti da li pojedinac spada u rizičnu grupu kada je u pitanju davanje kredita. U medicini može da se odredi koju terapiju treba prepisati pacijentu. Elektrane ili telefonske kompanije mogu da predvide kada će i koliki biti vrhunac opterećenja, kako bi ga izbegle...

U poslednje vreme javlja se i pojam *Data Warehousing*, koji podrazumeva centralizaciju svih podataka u jedno veliko ,,skladište". Centralizacija podataka dramatično ubrzava pristup podacima i njihovu analizu. Podaci koji se nalaze u ovim skladištima mogu biti dostupni svima. Zato Data Mining, pored svih uzbudljivih i neograničenih mogućnosti, sa sobom nosi i potencijalne opasnosti. Najveći problem je pitanje privatnosti.

3 Zaključak

Potencijali koji leže u Data Mining tehnologiji su ogromni i već se uveliko primenjuju. U svetu u kojem je znanje moć, sposobnost pretvaranja sirovih podataka u informacije je od neprocenjive vrednosti. Zato je vazno prepoznati pravi značaj Data Mining-a, a dalji razvoj računarske tehnologije će ga sve više i više uvoditi u svakodnevni život.

Literatura

[1] http://www.sk.rs/2005/05/skpr01.html

[2] Principles of Data Mining, Max Bramer, Springer-Verlag London limited, 2007