SVEUČILIŠTE U SPLITU
FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE
POSLIJEDIPLOMSKI DOKTORSKI STUDIJ ELEKTROTEHNIKE I INFORMACIJSKE TEHNOLOGIJE

Ivan Markić

Dubinsko pretraživanje, dohvaćanje i analiza digitalnih podataka

KVALIFIKACIJSKI DOKTORSKI ISPIT

Split, 5. ožujka 2015.

1. Sadržaj

2. Sažetak

Količina digitalnih podataka tijekom posljednjih godina kontinuirano se povećava. Ubrzani rast količine digitalnih podataka uzrokovan je širenjem računalnih mreža i softverskih sustava te razvojem društvenih mreža, objavljivanjem online video sadržaja, razvojem digitalne fotografije, različitih senzorskih mreža i mobilnih uređaja. Računalne mreže gotovo svakodnevno obrađuju enormne količine podataka, a izvori podataka gotovo su neiscrpni. Od interneta, poslovnih sustava, znanstvenih i inženjerskih sustava do svakog aspekta realnog svijeta. Možemo reći da živimo u svijetu informacija ili svijetu velikih podataka (Engl. Big Data).

Industrije poput telekomunikacija i medicine proizvode velike količine podataka i skupa s društvenim mrežama kroz razmjenu podataka spadaju u najveće generatore podataka. Novonastali trendovi donijeli su nove izazove pred analitičare podataka, programere, ali i pred administratore sustava. Ovakvu eksploziju podataka potrebno je podržati, kako razvojem alata, tako i načina kako ovakve podatke pretvoriti u korisno znanje.

Evolucijom tehnologija pretraživanja podataka, dohvaćanja i analize, stvorila se potreba za stvaranjem novih znanstvenih metodologija u navedenom području koje je dosta aktualno i dinamično u istraživačkom smislu.

S obzirom na trendove razvoja i sve veće količine dostupnih podataka, razvoj tehnologije pretraživanja, obrade i analize podataka (Engl. Data Mining) postaje jedno od zanimljivijih istraživačkih područja.

3. Uvod

Od svog nastanka baze podataka su među najpopularnijim područjima istraživanja računalne znanosti. Svoju su primjenu pronašle u gotovo svim granama industrije. Bilo to kupovanje u trgovinama, putničkim agencijama ili korištenje kreditnih kartica, potrebna je komunikacija s bazom podataka. Napretkom tehnologije posebno u područjima procesora, računalne memorije, spremišta podataka i računalnih mreža kontinuirano su rasle veličine, mogućnosti, ali i performanse sustava baza podataka. Sustav baze podataka nije ništa drugo nego mehanizam dizajniran da omogući spremanje, dohvat podataka, ali i da vodi brigu o konzistentnosti podataka. Baze podatka postoje u različitim formama zavisno od potreba grane industrije u kojoj se koriste. Prema svom logičkom dizajnu mogući logički formati baza podataka zasnovanih na određenom modelu su: *relacijske, mrežne, hijerarhijske* i *objektno orijentirane* baze podataka. Svi navedeni formati baza podataka u mogućnosti su spremati različite formate dokumenata poput binarnih datoteka, slika, video sadržaja, relacijskih podataka, višedimenzionalnih podataka, transakcijskih podataka, geografskih podataka itd. [1, 2, 3].

Nakon stvaranja relacijskih sustava baza podataka stvorila se i potreba za lakšim upravljanjem podacima. U tu svrhu napravljeni su alati za upravljanje sustavom baze podataka. Alati za upravljanje podacima u bazi podataka omogućavaju širok spektar mogućosti od pristupa podacima do kreiranja indeksa nad njima. Njihovom izradom stvoreni su uvjeti daljnjeg napretka klasičnih baza podataka prema naprednijim sustavima. Napredni sustavi baza podataka podrazumijevali su razvoj skladišta podataka, internet baza podataka, ali i naprednih metoda dubinske analize podataka. Razvojem naprednih sustava baza podataka sredinom 80-tih godina prošlog stoljeća otvorila su se nova područja istraživanja. Novi sustavi sadržavali su nove modele baza podataka kao što su prošireni relacijski model, objektno orijentirani, objektno relacijski i deduktivni. Objektno orijentirani sustavi baza podataka doslovno su procvali, uključujući prostorne, multimedijske, senzorske, znanstvene i inženjerske baze, baze znanja i baze poslovnih sustava.

Računalni hardver bio je glavni pokretač za razvoj baza podataka kakve imamo danas. Učestalim napredovanjem računalne opreme stvorili su se uvjeti za bržu obradu podataka te izradu podatkovnih skladišta s mnogo više prostora. Veliki broj baza i informacijskih repozitorija postao je dostupan širokom spektru korisnika za obradu i iznalaženje korisnog znanja iz velike količine podatka. Ovakvim slijedom razvoja tehnologije uslijedio je razvoj naprednih metoda dubinske analize podataka. Skladišta podataka (Engl. Data Warehouse) samo su jedan od novonastalih repozitorija podataka koji se sastoji od više različitih izvora podataka organiziranih u jedinstvenu shemu. Tehnologije u skladištu podataka omogućavaju manipulaciju s podacima poput čišćenja podataka, integraciju i pogled na podatke iz različitih kutova promatranja. Iako su tehnologije skladišta podataka omogućile izdvajanje korisnog znanja iz skupa podataka neka dublja analiza nije bila moguća. Stoga se javila potreba za

naprednijim metodama analize podataka, primjerice alata za dubinsku analizu koji omogućuju klasifikaciju, grupiranje, detekciju neočekivanih podataka ili anomalija kao i označavanje promjena podataka kroz vrijeme. Početkom novog stoljeća svijet je krenuo u globalnu informacijsku eru, a korištenje novih metoda koje su trebale biti djelotvorne i efikasne postao je glavni izazov za istraživače baza podataka. U obilju podataka sve je jasnije bilo da je svijet bogat podacima, ali jako siromašan pravim informacijama. Taj široki međuprostor koji se stvorio između informacija i podataka nagovještavao je sistematski razvoj alata za dubinsku analizu podataka. Novonastali skup alata za dubinsku analizu podataka trebao je biti u mogućnosti izdvojiti vrijedne informacije iz velike količine podataka u vrijedno korisno znanje koje treba biti u obliku pogodnom za ljudsko razumijevanje [4].

Jednostavno dohvaćanje podataka (Engl. Information Retrieval) iz izvora podataka nije bilo dovoljno da se korisnost podataka maksimizira već su bili potrebni alati i metode koje mogu automatski sažimati podatke, izdvojiti opće podatke ili otkriti uzorke u sirovim podacima. Jedini odgovor na ovakav problem bila je dubinska analiza podataka koja je ustvari izdvajanje skrivenih, ali vrijednih, često i prediktivnih informacija iz velikog skupa podataka čija je zadaća pomoći ljudima koji donose odluke u poslovnim organizacijama pri kvalitetnijem i bržem donošenju odluka [5].

4. Pregled literature

Zbog složenosti područja istraživanja osmišljena je strategija koja se koristila pri odabiru literature. Zbog velikog broja dostupnih izvora literature odabrani samo oni najrenomiraniji. Za odabir literature korištene su preporuke i stupnjevi rangiranja od strane za to odgovarajućih institucija, a koje se bave procjenom utjecaja odgovarajućeg časopisa ili članka.

Za izradu ovog rada korištena je dostupna literatura nastala između 1990 i 2015 godine, a koja sadržava informacije o odgovarajućem području istraživanja. Pored dostupnih izvora podataka korištene su i sve tehnike pretraživanja koje ti sustavi omogućuju.

Izvore literature moguće je podijeliti na tri dijela:

- Elektroničke baze podataka
- Časopisi
- Ostali izvori

Za odabir odgovarajućeg izvora podataka korišten je sustav razvijen od izdavača Elseviera. Osim što koristi pokazatelje otvorenog pristupa Elsevierov sustav nudi više opcija za procjenu znanstvenog časopisa koje se mogu podijeliti prema:

- članku (Engl. Source Normalized Impact per Paper - SNIP)
- izdanju (Engl. The Impact per Publication - IPP)
- rangu časopisa (Engl. SCImago Journal Rank - SJR).

Sve tri metode bazirane su na metodologijama vanjskih bibliometrika i koriste elektroničku bazu podataka Scopus. Elektronička baza Scopus jedna je od većih, kako sadržajno, tako i zbog mogućnosti praćenja rezultata pretrage, analiziranja i vizualizacije. Scopus elektronička baza podataka ima otprilike 18.000 časopisa. SJR pokazatelj dostupan je putem Internet poveznice „*http://www.scimagojr.com/*“ i najkorištenija je mjera za evaluaciju prestiža nekog časopisa kojeg je razvio profesor Félix de Moya. Baziran je na ideji da je citat dobiven iz važnijeg časopisa vrjedniji od citata dobivenog iz manje važnog časopisa. SJR je neovisan pokazatelj koji rangira časopis po njegovom prosječnom prestižu, i to prema pojedinom radu unutar njega, gdje se svakom citatu pridodaje određen faktor važnosti koji ovisi o važnosti časopisa iz kojeg dolazi. U obzir se uzima razdoblje od tri godine dok se samocitati djelomično isključuju. Sa SJR pokazateljem, temom koju obrađuje, ali i kvalitetom i ugledom časopisa postoji izravan utjecaj na vrijednost citata. Cijeli sustav zasnovan je na teoriji mreža konkretno mjeri centralnog vektora (Engl. Eigenvector Centrality). Mjera centralnog vektora utvrđuje važnost čvora u mreži koja se temelji na načelu da veza prema vrijednim čvorovima više doprinosi rezultatima čvora. SJR pokazatelj inspiriran je Googleovim Page Rank algoritmom, a razvijen je za izuzetno velike i različite mreže citata.

Moguće je koristiti i druge tipove pokazatelja za procjenu utjecaja znanstvenog časopisa poput Thomsonovog koji se još naziva Journal Impact Factor (JIF). JIP indikator je zatvorenog tipa i ne može se koristiti bez dopuštenja vlasnika (Thomson Reuters). Treća vrsta pokazatelja je EigenFactor Score (EF) kojim se izračunava ukupna važnost znanstvenog časopisa. Nedostatak EF je ta da je citiranost poistovjećena s kvalitetom znanstvenog časopisa, ali i davanje većeg faktora važnosti znanstvenim časopisima iz nekih zemalja kao što su npr. Sjedinjene Američke Države. [6, 7, 8, 9].

Pod skupom elektroničkih baza podataka korištene su one baze podataka koje su dostupne u centru za online baze podataka od kojih se mogu izdvojiti one što su projekt Ministarstva znanosti, obrazovanja i športa (MZOŠ), CARNet-a i IRB-a. Centar za online baze podataka dostupan je članovima znanstvene i istraživačke zajednice Republike Hrvatske, a ponajprije istraživačima, nastavnicima i studentima. Elektroničke baze podataka pretraživane su automatski pomoću sustava za pretraživanje ugrađenim u svaku pojedinu bazu i njima pripadajućih pomoćnih metoda za filtriranje.

Elektroničke baze podataka koje su korištene kao izvor literature za izradu ovog rada su:

- SpringerLink
- ACM Digital Library
- Thomson Reuters
- Wiley-BlackWell
- ScienceDirect
- Scopus
- ESSCO Host - Knjižnica Instituta Ruđer Bošković
- IEEE Xplore Digital Library
- Google Scholar
- Microsoft Academic Search

Od znanstvenih časopisa kao izvor informacija za izradu ovog rada mogu se izdvojiti:

- Computer (IEEE Computer society),
- Data & Knowledge Engineering (Elsevier),
- IEEE Transactions on Knowledge and Data Engineering (IEEE Computer Society)

Odabrani znanstveni časopisi dostupni su u organizaciji Fakulteta. Većina časopisa pregledavana je pojedinačno što znači pregledavanjem svakog pojedinog časopisa radi pronalaženja odgovarajućeg članka.

Izvori poput Internet tražilica (Google, Bing, Scribd, …) spadaju u kategoriju „ostalih“ i korišteni su samo u slučaju da pronađeni rezultati ne odgovaraju navedenom području istraživanja na početku procesa pretrage. Internet resursi poput Wikipedije nisu korišteni zbog

mogućeg mijenjanja sadržaja od strane zajednice, ali i često nedovoljne pouzdanosti specifičnih tema.

Prije pretrage elektronskih baza podataka definirane su ključne riječi (Data Mining - DM, Data Analytics - DA, Online Analytical Processing - OLAP, Data Warehouse - DW) prema kojima se vršilo pretraživanje. Nakon definiranja ključnih riječi pomoću logičkih operatora (OR, AND, NOT) pristupilo se filtriranju dobivenih rezultata. Filtriranjem rezultata sužen je izbor rezultata pretrage do ispunjavanja zahtjeva. Valjanost navedene metode provjerena je na način da su u obzir uzeti radovi koji su najkorišteniji. Testirane su ključne riječi iz odabranih radova i ključne riječi definirane prije pretrage čime je potvrđena valjanost načina pretraživanja.

Kako bi strategija odabira literature evaluirala, a samim time i kvaliteta rada bila zadovoljavajuća osmišljeni su kriteriji *prihvaćanja* i *odbijanja* pronađene literature. Pod kriterijem prihvaćanja pronađene literature podrazumijeva se uzimanje u obzir samo one literature koja obrađuje odgovarajuće području istraživanja. Osim ovog kriterija postoje i drugi kriteriji prihvaćanja poput datuma objave gdje se u obzir uzima ona literatura koja je novijeg datuma, ali i ona literatura koja je najviše citirana. Citiranost pronađene literature provjeravana je pomoću elektroničkih baza podataka. Kriteriji odbijanja pronađene literature su oni kriteriji koji isključuju one radove koji ne pripadaju području istraživanja, zatim nepotpuni radovi, skraćeni radovi, ali i radovi koji nisu na engleskom ili hrvatskom jeziku.

5. Sustav dubinske analize podataka

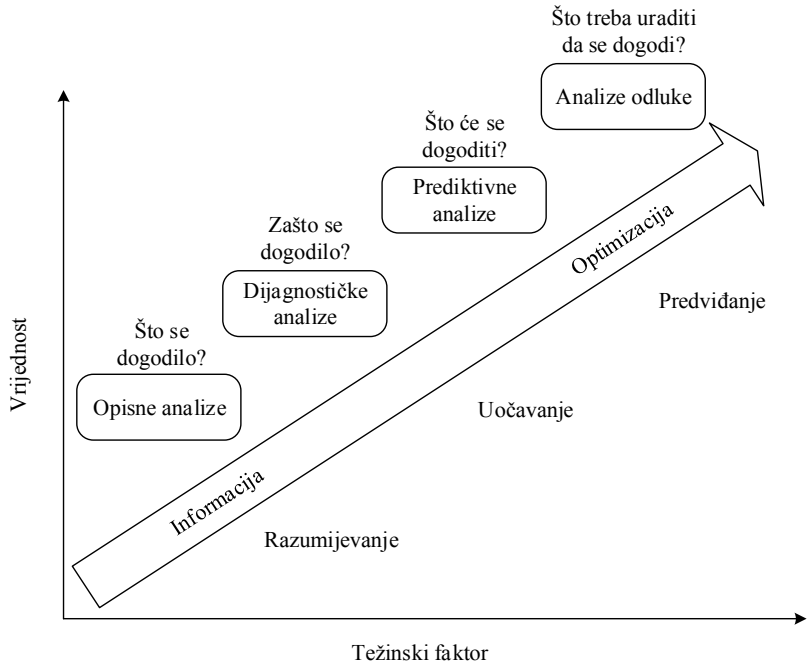
Mnoge industrije koriste podatke i metode za analizu podataka gotovo desetljećima. Financijski sektor, osiguravateljske tvrtke, biomedicina, farmakologija i drugi oduvijek koriste metode za analizu podataka, ali i predviđanje (projekciju) budućih rezultata. Ukoliko bi se postavilo pitanje što su to zapravo računovodstvena izvješća, analiza dobiti i gubitka, analiza za upravljanjem rizicima ili čak vremenska prognoza može se zaključiti da su to neke od metoda dubinske analize podataka koje nam pomažu da steknemo objektivne dojmove poslovnih očekivanja [10, 11].

5.1. Definicija pojma dubinske analize podataka

Dubinska analiza podataka je interdisciplinarna znanost i može se definirati na različite načine. Prvi i najjednostavniji način za opis dubinske analize podataka je proces izdvajanja malog podskupa uzoraka iz velikog skupa uzoraka. Također su mogući i drugi nazivi za dubinsku analizu podataka poput: izdvajanje znanja, analiza podataka prema uzorku, arheologija podataka, iskopavanje podataka itd. Dubinska analiza podataka je znanost koja je nastala prirodnom evolucijom informacijskih tehnologija, prvenstveno zahvaljujući razvoju sustava baza podataka. Razvoj sustava baza podataka uključuje razvoj nekoliko ključnih funkcionalnosti poput skupljanja podataka, upravljanja podacima što podrazumijeva spremanje i obradu transakcija te razvoj sustava skladišta podataka. Današnje baze podataka omogućuju pisanje upita i obradu transakcija kao učestalu praksu pri korištenju sustava za upravljanje bazom podataka, a napredne analize podataka postale su idući korak u njihovom korištenju [4, 12, 13, 14, 15, 16, 17].

Automatizirani proces koji počinje od izdvajanja znanja i analize iz različitih perspektiva, a koji završava rezimiranjem podataka u korisnu informaciju poznat je kao dubinska analiza podataka. Dubinska analiza podataka omogućuje da se spoje velike količine raznovrsnih podataka i pruža organizacijama znanje koje mogu iskoristiti za predviđanje budućih trendova i ponašanja, pri tome smanjujući troškove, povećavajući prihode i poboljšavajući procese [18].

Na slici 1 prikazana je linearna funkcija na kojoj je prikazana vrijednost težinskog faktora razumijevanja informacije, uočavanja i predviđanja.



Slika 1 Prikaz težinskog faktora razumijevanja, uočavanja i predviđanja [19]

Primjenjujući odgovarajuće računalne postupke i koristeći matematičke metode teorije vjerojatnosti i statistike, ali i suvremene računalne alate moguće je analizom ranijih podataka napraviti podatkovne modele[1] za projekciju budućih događaja. Pogrešno je mišljenje da se dubinskom analizom podataka izdvajaju podaci. Dubinskom analizom izdvajaju se modeli koji se primjenjuju na nove podatke u svrhu dobivanja odgovarajućih rezultata. U gotovo svim metodama dubinske analize podatka postoji prebrojavanje skupa podataka S te uspoređivanje tako dobivenih veličina radi dobivanja potencijalnih informacija ili hipoteza koji se nalaze u skupu podataka.

$$S = \{a_1, a_2, a_3, .., a_n \mid n \in N\} \tag{1}$$

Rezultati dubinske analize podataka ne moraju nužno biti relevantni što znači da se relevantnost svake analize mora utvrditi ili ljudskom analizom ili statističkom provjerom [20].

Dubinskom analizom podataka primjenjujući odgovarajuću algoritme nad podacima moguće je predvidjeti trendove, identificirati uzorke, kreirati pravila i preporuke tj. podacima dati novu dimenziju. Ovakvi uzorci ne mogu biti pronađeni tradicionalnim metodama poput strukturnog upitnog jezika (Egl. Structured Query Language - SQL) ili osnovnih statističkih metoda poput *Sum* (zbrajanje), *Avg* (prosjek), *Count* (brojanje), *Max* (maksimalna vrijednost) i *Min*

[1] Model – Model je pojednostavljena, svrsishodna prezentacija realnog svijeta [31].

(minimalna vrijednost) iz razloga što su veze između podataka jako kompleksne, ali i zbog velike količine podataka. Kada se uzorci i trendovi prikupe može se reći da je definiran *model podataka*.

Model dubinske analize podataka može se primijeniti na razne scenarije kao što su prema [21]:

- Predviđanje (procjena prodaje, predviđanje serverskih ispada).
- Rizik i vjerojatnost (odabir najboljih kupaca za buduću marketinšku kampanju, primjena modela vjerojatnosti za dijagnozu budućih izlaza).
- Preporuke (određivanje proizvoda koji se prodaju skupa, izrada preporuka za prodaju proizvoda).
- Pronalaženje sekvenci (analiza odabira kupaca u košarici, predviđanje sličnih događaja).
- Grupiranje (odvajanje događaja od kupaca u grupe, analiza i predviđanje veza među njima).

Dubinska analiza podataka ima ugrađene metode za otkrivanje uzoraka iz promatranih podataka koje služe izradi modela znanja potrebnih korisniku. Iako predstavljaju interesantno znanje potrebna je odluka eksperta kao finalni korak dubinske analize podataka.

Postoje dva primarna matematička formalizma u ugrađenim modelima:

- statistički i
- logički

Statistički pristup omogućuje nedeterminističke efekte u modelima i najraširenija je pristup koji se koristi u dubinskoj analizi podataka dok je *logički* čisti deterministički pristup. Većina se metoda bazira na pristupu pokušaja i pogrešaka, uzimanja ili ostavljanja. Shvaćanje dubinske analize podataka i modela indukcije[2] pojednostavljuje ponašanje algoritama i omogućuje korisniku lakše shvaćanje procesa otkrivanja znanja [22, 13].

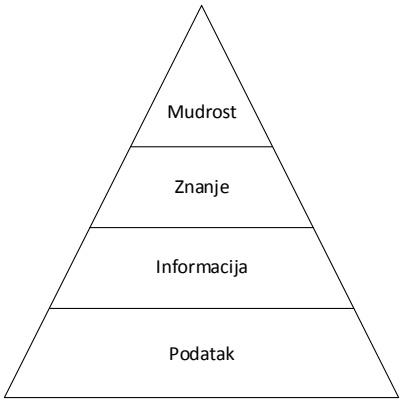
5.2. Životni ciklus otkrivanja znanja

Dubinska analiza podataka je sinonim za popularno korišteni pojam procesa izdvajanja znanja iz podataka ili KDD (Engl. Knowledge Discovery From Data). Proces izdvajanja znanja uključuje sve, od sakupljanja podataka, postavljanja upita nad podacima i kreiranja modela do prezentacije i integracije s drugim izvorima znanja [21].

[2] Indukcija je pojam koji se koristi kada neki proces započinje s generalnim pravilima i određenim činjenicama iz kojih proizlaze nove činjenice. Suprotno pojmu indukcije je pojam dedukcije [31].

Proces izdvajanja znanja uključuje korištenje tehnika za analizu velikih količina informacija i automatsko učenje znanja koje nam može poslužiti za shvaćanje dostupnih informacija [23, 24].

Informacija je podatak kojemu je dano značenje na osnovu veza s drugim podacima. Njegova svrha je korisnost. Ovo značenje može biti korisno, ali i ne mora. Postoji hijerarhijska shema koja predočava znanje, od sirovih podataka, preko informacije i znanja do mudrosti tzv. DIKW shema (Engl. Data, Information, Knowledge, Wisdom) prikazana na slici 2. Prema DIKW shemi podatak je najosnovnija razina, informacija podatku daje kontekst, znanje definira kako ga upotrijebiti, a mudrost (Engl. Wisdom) kada i zašto ga upotrijebiti [25].



Slika 2. DIKW hijerarhija ([25])

Osnovni problem s kojim se suočava izdvajanje znanja je proces pretvaranja podataka iz forme koja je manje kompaktna u formu koja je kompaktnija, apstraktnija ili korisnija. Komponenta izdvajanja znanja dubinske analize podataka počiva uglavnom na poznatim tehnikama strojnog učenja, prepoznavanja uzoraka i statistike kako bi se pronašli modeli dubinske analize podataka. Proces izdvajanja znanja fokusira se na cjelokupni proces, počevši od toga kako su podaci spremljeni u izvoru podataka pa do korištenja algoritama, tj. njihove učinkovitosti. Može se reći da je proces izdvajanja znanja interdisciplinarna znanost koja objedinjuje sve aspekte dubinske analize podataka [22, 26].

Da bi životni ciklus otkrivanja znanja iz skupa podataka bio uspješan potrebno je definirati problem, zatim definirati poslovne zahtjeve, ali i ciljeve koji analizom trebaju biti ispunjeni.

Prema [21] potrebno je odgovoriti na sljedeća pitanja:

- Što tražimo i koje sve tipove veza između podataka pokušavamo pronaći?
- Hoće li se rješenje problema koji tražimo reflektirati na poslovne procese?
- Kakve tipove podataka imamo i kakve tipove informacija imamo u svakoj koloni (atribute), kakve su veze između tablica i kakav izlaz ili atribute želimo predvidjeti?
- Želimo li napraviti model dubinske analize podataka kojim ćemo predvidjeti buduće događaje ili samo želimo tražiti interesantne uzorke i asocijacije?

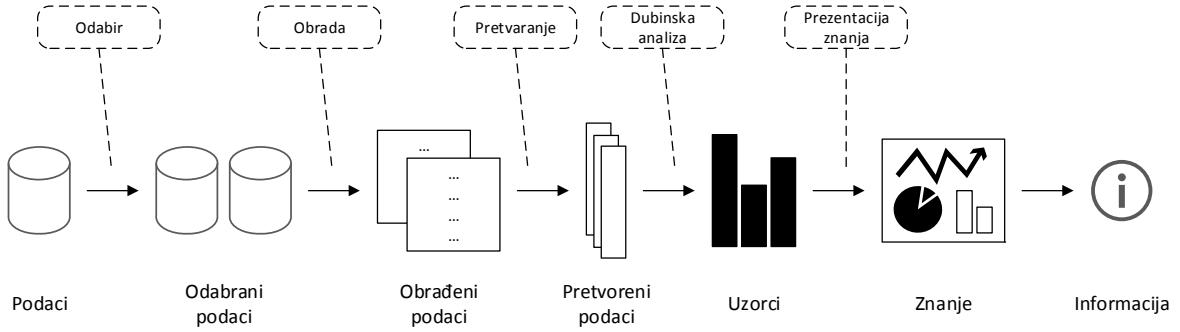
- Trebamo li očistiti dohvaćene podatke, koristiti neke od agregatnih funkcija za obradu korisnih podataka?
- Određuju li podaci točno poslovne procese itd.?

Da bi se odgovorilo na ova pitanja potrebno je uraditi studiju dostupnosti podataka, istražiti potrebe poslovnih procesa uzimajući u obzir dostupne podatke. Ukoliko dostupni podaci ne podržavaju potrebe poslovnih procesa potrebno je redefinirati poslovne zahtjeve i vratiti se na početak procesa. Rezultat dubinske analize je model podataka i mora biti moguće dobiveni model implementirati u poslovne procese pri čemu kvaliteta modela podataka ovisi isključivo o kvaliteti ulaznih podataka [21].

Dubinska analiza podataka je samo jedan od koraka procesa otkrivanja znanja. Proces uključuje i druge korake ciklične prirode što podrazumijeva da se je u svakom trenutku moguće vratiti na početak procesa ukoliko se zaključi da sakupljeni podaci nisu dovoljni za izradu modela dubinske analize podataka. Životni ciklus otkrivanja znanja počinje s razumijevanjem objektivnih zahtjeva iz poslovne perspektive, pretvaranjem tog znanja u definiciju problema dubinske analize podataka i dizajniranjem plana u svrhu postizanja cilja, a sastoji se prema [4, 5, 14, 27] od:

- Odabira (Odabir odgovarajućih podataka za analizu)
- Obrade (Čišćenje i prilagodba podataka)
- Pretvaranja (Pretvaranje obrađenih podataka u odgovarajući format pogodan za analizu)
- Dubinske analize (Otkrivanje znanja)
- Prezentacije znanja (Provjere uzoraka dobivenih dubinskom analizom podataka i prezentacije stečenog znanja).

Na slici 3 prikazan je proces otkrivanja znanja u dubinskoj analizi podataka.

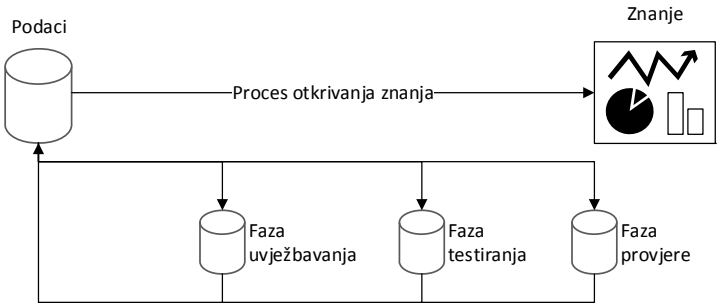


Slika 3. Proces otkrivanja znanja ([14])

Sustav dubinske analize podataka je centar procesa i ponavlja se dok se ne dobiju prihvatljivi rezultati prikazani na slici 4.

Prema [2] proces otkrivanja znanja u dubinskoj analizi podataka koja je dio ukupnog procesa sastoji se od tri faze:

1. uvježbavanja,
2. testiranja i
3. provjere modela.



Slika 4. Proces otkrivanja znanja u procesu dubinske analize podataka

5.3. Odabir, obrada i pretvaranje podataka

U procesu odabira podataka kombiniraju se podaci s više izvora podataka u jedan konsolidirani izvor. Novi izvor podataka može sadržavati nepotpune i netočne podatke, ali i podatke različitih formata [21]. Osim netočnih i nedostajućih podataka moguće su i nekonzistentne prezentacije iste vrijednosti kao i duplicirani podaci [28].

Proces čišćenja podataka naziva se još i eliminacija i smanjenje šuma ili eliminacija svojstava i njegova je zadaća da se iz novog skupa podataka uklone šumovi kao i nepravilni i nepotpuni podaci. Proces se može izvršiti korištenjem alata za čišćenje tzv. ETL alata (Engl. Extract, Transform, Load).

Podatke je moguće promatrati s nekoliko aspekata prikazanih tablicom 1. Prvi aspekt je sa stajališta *strukturiranosti*, tj. podjela podataka na *strukturirane* i *nestrukturirane* vrste podatka. Drugi aspekt promatranja podataka je sa stajališta *pokretljivosti* i to *statične* i *dinamičke* vrste podatka. Treći aspekt je *tip podataka*. Podatke je moguće prema tipu podijeliti na *jednake* (Engl. Homogeneous) i *različite* (Engl. Heterogeneous). Jednakost podrazumijeva podatke istog tipa dok različitost podrazumijeva podatke različitih tipova. Čišćenje podataka ovisi o određenoj vrsti podataka. Tako je uvijek jednostavnije čistiti statične naspram dinamičkih vrsta podatka kao i jednake naspram različitih vrsta podataka. [29, 30, 31].

Podaci	
Strukturirani	**Nestrukturirani**
Statični i dinamički	
Jednaki i različiti	
– Relacijske baze podataka – Skladišta podataka – Transakcijski podaci	– Internet – Socijalne mreže – Senzorske mreže – Multimedija – Elektronička pošta – Tokovi podataka

Tablica 1 Bitni aspekti promatranja podataka

Upoznavanje vrijednosti postojećih podataka među najvažnijim je čimbenicima za donošenje odluka kada se izrađuje model dubinske analize podataka. Tehnike upoznavanja podataka podrazumijevaju izračun minimalnih i maksimalnih vrijednosti, srednje i standardne devijacije kao i pretragu ostalih distribucija podataka.

Primjer:

Pregledajući maksimume, minimume ili standardnu devijaciju moguće je odrediti koji podaci ne predstavljaju kupce kao niti poslovni proces. Ako se ovakva situacija dogodi, potrebno je pronaći više uravnoteženih podataka i pregledati pretpostavke koje su osnova očekivanja dubinske analize podataka. Standardna devijacija i druge distribucijske vrijednosti predstavljaju korisnu informaciju o stabilnosti i točnosti rezultata. Velika vrijednost standardne devijacije pokazuje da nam povećanje količine podataka može pomoći da popravimo model dubinske analize podataka. Podaci koji puno odstupaju od vrijednosti standardne devijacije mogu biti iskrivljeni ili mogu prikazivati realnu sliku stvarnih problema. Takve je podatke teško uklopiti u model.

U ovom koraku otkrivaju se skrivene veze između podataka i definira se sami problem istraživanja. Određuju se kolone koje su najvažnije za korištenje u analizama tj. odabiru se odgovarajući podaci koji se analiziraju. Odabrane kolone često se nazivaju atributi i najbolje opisuju promatrano područje analiziranja. Nepotpuni i pogrešni podaci koji se pojavljuju odvojeno, a zapravo su čvrsto vezani mogu utjecati na konačni ishod analize podataka na način koji nije očekivan [21].

Još jedna vrlo važna tehnologija koje poboljšava kvalitetu podatka je *deduplikacija*. Deduplikacijom se identificiraju grupe približno sličnih entiteta. Problem se može promatrati u kontekstu problema grupiranja grafa gdje je svaki čvor entitet i između kojih postoji granica ukoliko je stupanj sličnosti između dva čvora dovoljno jak. Funkcija koja definira stupanj sličnosti između dva entiteta bazirana je na funkciji tekstualne sličnosti, kao što je primjerice

razlika dvije riječi „Analiza“ i „Anliza“. Mogućnost da se vrši takva usporedba na velikom broju parova poznato je kao *neizrazito združivanje* (Engl. Fuzzy Matching) i od velike je važnosti za deduplikaciju. Svi moderni alati za dubinsku analizu podataka podržavaju neizrazito združivanje i to kao dio alata za ETL [28].

5.4. Pripremanje podataka

Skupovi podataka sastoje se od podatkovnih objekata, a podatkovni objekt predstavlja *entitet* (u bazi prodaje to može biti kupac, proizvod) i opisan je skupom *atributa*. Podatkovni objekti obično se nazivaju i primjeri, instance ili jednostavno objekti. Ako se podatkovni objekti nalaze u bazi podataka onda su to parovi podataka, a to su redci baze podataka koji odgovaraju podatkovnim objektima i njima pripadajuće kolone koje odgovaraju atributima [4].

Priprema podataka podrazumijeva prikaz podataka u obliku tablice. Kada se podaci pripreme u obliku tablice smatra se da su podaci pogodni za analizu. Preporuke su da bude što je više moguće primjera tj. redaka tablice kako bi analiza bila relevantnija i kako bi se pouzdanije otkrili složeniji odnosi koji predstavljaju novost promatranog područja [20].

Model dubinske analize podataka sličan je kontejneru koji određuje kolone korištene za ulaz, atribute koje predviđamo i parametre koji govore algoritmima kako da obrađuju podatke [21].

Atribut je podatkovno polje koje predstavlja oznaku (karakteristiku) ili značajku podatkovnog objekta. Postoje i drugi nazivi za atribut poput imenica: *dimenzija, značajka i promjenljiva* (Engl. Variable). Izraz dimenzija najčešće se koristi u skladištima podataka, značajka kod strojnog učenja, a promjenljiva kod statističkih metoda evaluacije. Razlikuje se više vrsta atributa: *nominalni, binarni, redni* ili *numerički*.

Nominalni atribut odnosi se na imena. Vrijednosti nominalnih atributa su simboli ili imena stvari. Svaka vrijednost predstavlja neku vrstu kategorije, šifru ili stanje. Ovakve vrijednosti u računalnim znanostima nazivaju se enumeracije.

Primjer:

Za dati atribut „*vrsta_osiguranja“* koji opisuje objekt police mogući su nominalni atributi poput: imovina, motorna vozila, životno osiguranje, itd.

Binarni atributi su nominalni atributi, ali samo s dvije vrijednosti i to 0 ili 1. Nula obično znači da je atribut odsutan, dok jedan znači da je atribut prisutan. Binarni atributi često se nazivaju i Booleovi atributi ako odgovaraju vrijednostima Istina (Engl. True) ili Laž (Engl. False).

Redni atribut je atribut sa mogućim vrijednostima koji imaju smišljen redoslijed ili prioritet između njih dok veličina između uzastopnih vrijednosti nije poznata.

Nominalni, binarni i redni atributi su kvalitativni, što znači da ne predstavljaju trenutnu veličinu ili količinu.

Primjer:

Vrsta kupca koji može biti: ključni, srednji i normalni kupac.

Numerički atributi su kvantitativni što znači da mjere kvantitetu, a predstavljeni su cjelobrojnim ili realnim vrijednostima. Numerički atributi mogu biti *intervalni* i *razmjerni*. Intervalni se mjere na skali jednakih jediničnih veličina gdje je značajna razlika između dvije vrijednosti.

Primjer:

Razlika između temperature od 100 stupnjeva i 90 stupnjeva je ista kao razlika između 90 stupnjeva i 80 stupnjeva.

Razmjerni atributi su numerički atributi s naslijeđenom nultom točkom, imaju istu definiciju kao intervalne promjenljive ali imaju i točnu definiciju nule [4, 16].

Primjer:

Temperatura apsolutne nule od -273.15 °C.

5.5. Provjera modela dubinske analize podataka

Provjerom modela testira se njegova sposobnost obavljanja zadaće dubinske analize podataka. Potreba je identificirati stvarne i zanimljive uzorke predstavljajući znanje bazirano na potencijalnim interesima. Također je bitno poznavanje domene istraživanja, a to znači razumjeti problem istraživanja i značenje dobivenih podataka kao i uvjeta pod kojim su podaci skupljeni. Kreiranje modela dubinske analize podataka ukoliko se uspoređuje sa stvarnim činjenicama ili trenutnim znanjem može biti dobra podloga za potvrdu kako je postupak stvaranja modela potpuno u skladu s realnim očekivanjima. Ukoliko se rezultati dubinske analize ne podudaraju s realnim očekivanjima jedan od mogućih problema je da ulazni skup podataka nije dovoljno dobro pripremljen ili da odabrani atributi ne odgovaraju domeni istraživanja. Preporuka je dobivene modele testirati statističkim rezultatima, međutim, jedina prava provjera dobivenog modela je provjera napravljena na nezavisnim podacima koji su prikupljeni paralelno procesu dubinske analize podataka. Ukoliko se modeli slažu onda je to dobra osnova za otkrivanje novog znanja [20, 21].

Budući da proces dubinske analize podataka uključuje široki spektar aktivnosti uključujući i provjeru dobivenih rezultata nužno je u svrhu izrade kvalitetnih modela uključivanje eksperta za izdvajanje znanja. Više je vrsta pristupa provjere modela dubinske analize podataka od kojih je potrebno izdvojiti *aktivno dubinsko analiziranje* (Engl. Active Mining) i *korištenje ontologija*. U aktivnom dubinskom analiziranju ekspert se aktivno uključuje u manipuliranje

procesima dubinske analize dok se kod ontološkog pristupa postojeći modeli podataka povezuju s bazama koje predstavljaju formalnu prezentaciju znanja [20, 32].

U filozofskom smislu, ontologija je znanost o stvarima koje postoje. U području umjetne inteligencije izraz ontologija obično se odnosi na jednu ili dvije odgovarajuće stvari. Bez ontologija tj. konceptualizacije znanja nemoguće je na kvalitetan način predstaviti znanje. Od svog nastanka ontologije su postale popularne pogotovo za istraživače na području umjetne inteligencije. Izraz se uskoro proširio na dosta područja poput inteligentne integracije informacija, povrata informacije, ali i upravljanja znanjem. Postale su popularne zbog mogućnosti dijeljenja i shvaćanja domene promatranja u komunikaciji između ljudi i računala. Ontologije se mogu definirati kao formalne i eksplicitne specifikacije dijeljenih koncepata. Konceptualizacija se odnosi na nešto što je apstraktno, apstraktni opis realnih fenomena u svijetu u kojem živimo. Uloga ontologija u procesu otkrivanja znanja je ta da one predstavljaju rječnik pojmova i veza koje model podataka sadržava [33, 34].

Budući da je skladište podataka presudni dio svakog sustava za donošenje odluka u kojem su spremljeni očišćeni i integrirani podaci za istraživanje znanja potrebno je u svrhu poboljšanja ovih procesa implementirati inteligentne metode dubinske analize s podrškom za korisničke ontologije. Ontologije pomažu izgraditi korisne modele dubinske analize koji sprječavaju generiranje pogrešnih uzoraka, otkrivanju koncepta proširenih pravila, ali i za izradu aktivnog mehanizma za ponovno otkrivanje znanja [35].

Izgradnja ontologija iz rezultata dubinske analize podataka mogu se podijeliti u dvije faze: *faza dubinske analize* i *faza izgradnje ontologija*. Faza dubinske analize odnosi se na proces dubinske analize podataka uključujuću pripremu podataka, odabir i izdvajanje znanja, a drugi dio je izgradnja ontologija iz izdvojenog znanja što predstavlja izlaz dubinske analize podataka. Većina alata za dubinsku analizu podataka ima ugrađene metode za generiranje ontologija na osnovu izlaznih rezultata dubinske analize podataka koje zatim prezentira u XML (Engl. Extensible Markup Language) i OWL (Engl. Web Ontology Language) formatu [36, 37].

5.6. Prezentacija znanja

U procesu predstavljanja stečenog znanja koriste se različite tehnike vizualizacije i prezentacije. Najčešći oblik prezentacije znanja je grafički čime je osiguran najbolji prijenos stečenog znanja na ljude koji koriste modele dubinske analize podataka [21]. Jedan od popularnijih načina prezentacije znanja je histogram, a moguće je koristiti neke od metoda poput dijagrama pite (Engl. Pie) ili dijagrama šipke (Engl. Bar). Tumačeći podatke s histograma može se točno odrediti koji su najučestaliji podaci koji se pojavljuju u bazi podataka, a također je moguće pomoću histograma odrediti važne statističke vrijednosti kao što je prosjek te maksimalna i minimalna vrijednost [38].

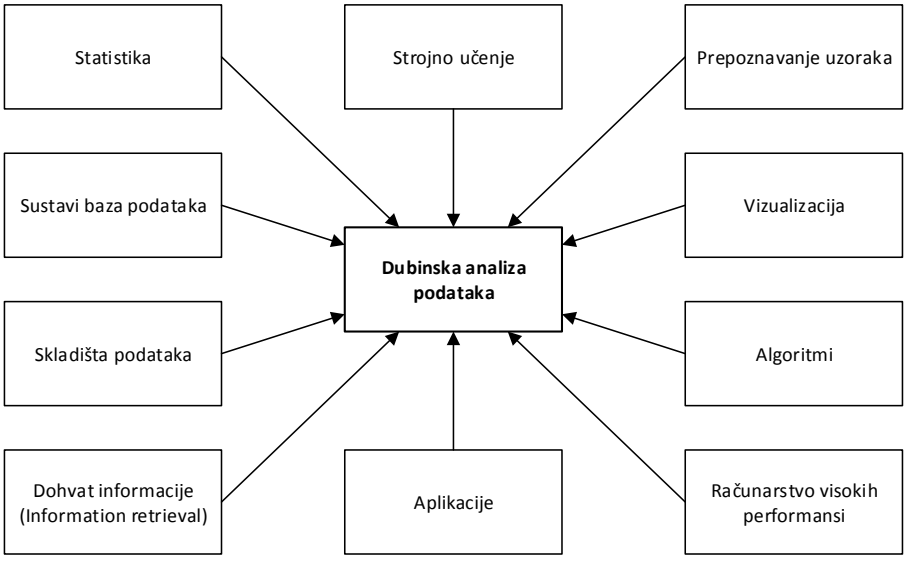
Najpopularnije tehnike prezentacije stečenog znanja su ontologije. Kada su ontologije formalizirane na logički način za dani skup činjenica mogu se koristiti za izdvajanje novih činjenica, ali i za provjeru konzistentnosti. Vrlo su korisne posebno kada se radi o kompleksnim i različitim problemima kao i velikoj količini znanja. Osim rješavanja problema, ontologije budući da koriste formalnu prezentaciju znanja služe i za opisivanje svih koncepata i tehnika nekog procesa kao i nuđenju sugestija u novom projektu jer imaju mogućnost pamćenja ranijeg znanja u svrhu iskorištenja u budućnosti [39].

5.7. Korištenje tehnika iz drugih domena istraživanja

Dubinska analiza podataka kao aplikacijski usmjerena tehnologija objedinjuje više vrsta tehnika iz drugih domena istraživanja. Tehnike koje posebno utječu na razvoj metoda dubinske analize podataka su prema [4]:

- statistika
- strojno učenje
- prepoznavanje uzoraka
- baze podataka i skladišta
- dohvat informacije
- vizualizacija
- algoritmi
- računarstvo visokih performansi itd.

Na slici 5 nalazi se grafički prikaz korištenih tehnika u dubinskoj analizi podataka.



Slika 5. Tehnike koje se koriste u dubinskoj analizi podataka ([4])

Najvažniji dijelovi sa slike 5 objašnjeni su dalje u nastavku.

5.7.1. Statistika

Statistika ili statističke metode nisu metode dubinske analize podataka, ali se statistika koristila čak i prije nego je izraz „analiza“ uveden u poslovne procese. Statistika odgovara na mnoga pitanja poput kolika je vjerojatnost da će se neki događaj dogoditi, koji je uzorak pouzdaniji ili koja je najveća razina sažetosti podataka koja nam govori kakvi se podaci nalaze u bazi podataka [38].

Statistički model je matematička funkcija koja opisuje ponašanje objekta u odredišnoj klasi u uvjetima različitih promjenljivih i njihovih povezanosti s mogućim distribucijama klase, a najviše se koristi za modeliranje podataka i njihovih klasa. Izrada statističkog modela specifična je za korištenje tehnika dubinske analize podataka poput *označavanja* i *klasifikacije* (tehnike dubinske analize podataka objašnjene su dalje u nastavku) koje su ujedno i zadaće dubinske analize podataka. *Prediktivna statistika* modelira podatke na način da izdvaja slučajna ili nepouzdana zapažanja i oslikava zaključke procesa ili skupova koji se istražuju. Mnogo je primjena statističkih metoda poput evaluacije rezultata dubinske analize, ali i za izradu modela šuma i nedostajućih podataka na velikom skupu podataka. Primjerice, nakon što je model klasifikacije ili predikcije[3] dobiven dubinskom analizom podataka, model se testira statističkom hipotezom. *Test statističke hipoteze* (potvrdna analiza podataka) daje statističke odluke koristeći eksperimentalne podatke. Statistička hipoteza iskazuje se na način da može biti vrednovana statističko-analitičkim postupcima. Statistička hipoteza matematički je izraz koji predstavlja polaznu osnovu na kojoj se temelji kalkulacija statističkog testa. Testiranje hipoteze je statistički postupak kojim se određuje da li i koliko pouzdano raspoloživi podaci podupiru postavljenu pretpostavku. Testiranje hipoteza, je u biti postupak kvantifikacije impresija o specifičnoj hipotezi.

Rezultati se nazivaju statistički značajni ako su se desili slučajno, a pouzdanost modela se mjeri pouzdanošću koja se povećava ukoliko model klasifikacije ili predikcije sadržava istinu. Često je glavni izazov kako statistiku prilagoditi velikom skupu podataka. Mnoge statističke metode su vrlo kompleksne za izračun i zahtijevaju pažljivo dizajniranje algoritama. Dizajnirani algoritmi moraju biti efikasni čime se smanjuje potreba za velikom količinom resursa, primjerice procesorske snage, itd. Ovi izazovi posebno dolaze do izražaja kod web pretraživača gdje se proces dubinske analize podatka događa kontinuirano i u realnom vremenu [4, 31].

Kako bi obrada podataka prije dubinske analize bila uspješna potrebno je steći cjelokupnu sliku baze podataka. Ovo se obično radi općim statističkim opisima koji uvelike pomažu pri

[3] Predikcija – Predikcija znači predviđanje budućih događaja. U znanosti koja se bavi podacima predikcija znači procjenu nepoznatih vrijednosti. Vrijednost predikcije može biti buduća vrijednost, ali također može biti trenutna ili neka prošla vrijednost iz razloga što dubinska analiza podataka uglavnom analizira povijesne podatke, modeli se često prave i testiraju događajima iz prošlosti [31].

detektiranju nedostajućih podataka, ali i šumova. Ove vrijednosti prema [38, 40] mogu uključivati:

- Maksimalnu i minimalnu vrijednost prediktora
- Aritmetičku sredinu (Engl. Arithmetic Mean)
- Medijan (vrijednost koja dijeli bazu podataka na dva dijela s gotovo jednakim brojem zapisa)
- Mod (Najčešća vrijednost prediktora)
- Varijancu (mjera raspršenosti na osnovu prosječne vrijednosti).

5.7.2. Strojno učenje

Strojno učenje (Engl. Machine Learning - ML) je područje računalnih znanosti koje se vrlo brzo razvija i koje izučava sposobnosti računala da uče ili poboljšaju svoje performanse na osnovu podataka. Glavno područje istraživanja je mogućnost automatskog prepoznavanja kompleksnih uzoraka i donošenja inteligentnih odluka baziranih na podacima. Problemi koji se susreću u strojnom učenju usko su vezani s dubinskom analizom podataka [4].

Izgradnja modela dubinske analize podataka omogućuje shvaćanje dostupnih podataka i otkrivanje logičkih veza između podataka. Velika količina podataka se interpretira kao konceptualizacija veza između podataka što ovu metodu učenja čini jednom od najpopularnijih metoda koja se koristi u znanstvenim istraživanjima. Rezultat strojnog učenja je veliki broj dobivenih modela, a točnost se utvrđuje na osnovu prediktivne točnosti konstruiranih modela nad podacima koji nisu korišteni u procesu učenja. Dobiveni model ne mora ujedno biti i najbolji rezultat strojnog učenja [20].

Strojno učenje sastoji se od više vrsta metoda strojnog učenja. Metode strojnog učenja prema [31] su:

- nadgledano učenje
- nenadgledano učenje
- polunadgledano učenje
- aktivno učenje.

5.7.2.1 Nadgledano učenje

Ako je određeni cilj osiguran tj. kada je problem dubinske analize definrian učenje je nadgledano. Promatrajući u domeni dubinske analize podataka nadgledano učenje podrazumijeva da postoje ciljni podaci. Nije dovoljno da informacije postoje načelno, one moraju postojati u podacima [31].

Nadgledano učenje (Engl. Supervised learning) je naziv za klasifikaciju i proizlazi iz opisanih primjera (klasificiranih) testnog skupa [4, 31]. Cilj ove vrste strojnog učenja je dobiti prediktivni model na osnovnu vrijednosti ostalih primjera [20].

Primjer:

Problem prepoznavanja poštanskog broja iz skupa rukom napisanih poštanskih brojeva koji se koriste kao testni primjeri, a koji nadgledaju učenje modela klasifikacije.

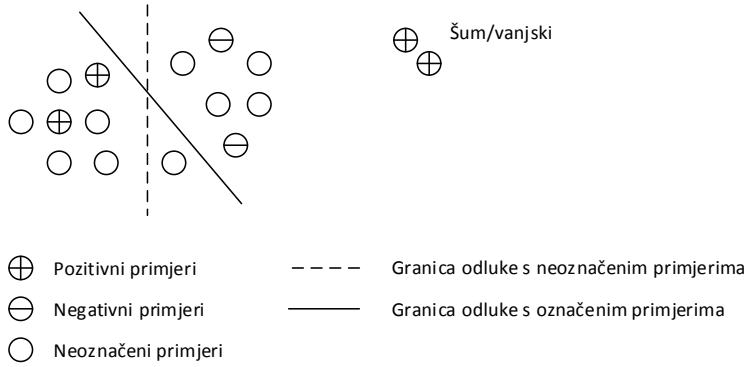
5.7.2.2 Nenadgledano učenje

Nenadgledano učenje (Engl. Unsupervised learning) je sinonim za grupiranje (Engl. Clustering) i suprotnost je nadgledanoj metodi strojnog učenja. Nenadgledana metoda zapravo rješava problem grupiranja. Kada je definicija problema dubinske analize podataka nepoznata tada se problem naziva nenadgledan. Učenje je nenadgledano sve dok su klase ulaznih primjera neoznačene ili nisu klasificirane. Ovu vrstu strojnog učenja možemo koristiti i za otkrivanje klasa unutar podataka. Sve dok podaci nisu označeni ovaj model strojnog učenja ne može nam opisati semantička značenja skupine ili grupe [4, 31].

5.7.2.3 Polunadgledano učenje

Polunadgledano učenje (Engl. Semi-supervised learning) je tehnika strojnog učenja koja koristi označene i neoznačene primjere klasa dok je u procesu spoznavanja (učenja) modela. Prvi pristup koristi označene primjere za učenje modela klase i neoznačene za definiranje granica između klasa [4].

Na slici 6 prikazan je slučaj dvije klase gdje se primjeri koji pripadaju jednoj klasi promatraju kao pozitivni, a oni koji pripadaju drugoj klasi promatraju kao negativni. Isprekidana linija predstavlja granicu odluke ukoliko ne uzimamo u obzir neoznačene primjere i dijeli pozitivne i negativne primjere. Koristeći neoznačene primjere možemo precizirati granicu odluke u punu liniju, a tim više možemo detektirati i šumove poput dva pozitivna primjera u desnom kutu prikazane slike.



Slika 6. Polunadgledano učenje

5.7.2.4 Aktivno učenje

Aktivno učenje (Engl. Active learning) je pristup strojnog učenja koji omogućava korisnicima da imaju aktivnu ulogu u procesu učenja na način da postoji mogućnost da se od korisnika zatraži da označi primjer koji može biti iz skupa neoznačenih primjera. Cilj aktivnog učenja je da se optimizira kvaliteta modela s čestim upitima za označavanjem primjera prema korisniku. Broj zahtjeva za označavanjem neoznačenih klasa koje korisnik treba označiti mora biti ograničen [4].

Strojno učenje se fokusira na točnost modela. Točnost podrazumijeva stavljanje većeg naglaska na učinkovitost i stabilnost metoda dubinske analize podataka na velikim skupovima podataka, a manjeg na podršku za kompleksne tipove podataka ili otkrivanje novih alternativnih metoda.

5.7.3. Dohvat informacije

Dohvat informacije (Engl. Information Retrieval - IR) je tehnika pretraživanja dokumenata ili informacija u dokumentima. Najčešće se koristi na tekstualnim dokumentima koji se mogu nalaziti bilo gdje na Internetu. Dohvat informacije podrazumijeva:

- da su podaci pretrage nestrukturirani
- da su upiti sastavljeni uglavnom od ključnih riječi bez kompleksne strukture (za razliku od SQL upita u sustavima baza podataka)

Česti pristup u povratu informacija je prilagodba vjerojatnostnom modelu. Primjerice tekstualni dokumenti mogu biti promatrani kao skupina riječi koje se pojavljuje u dokumentu. Jezični model dokumenta je funkcija gustoće vjerojatnosti koju generira skupina riječi u dokumentu. Sličnosti između dva dokumenta mogu se mjeriti sa sličnošću jezičnih modela. Nadalje, predmet u skupu tekstualnih dokumenata može se modelirati kao distribucija vjerojatnosti preko popisa riječi, koji se zove predmetni model. Tekstualni dokument, koji može sadržavati jedan ili više predmeta može se smatrati kao skupina više predmetnih modela. Spajajući modele dohvata informacija s tehnikama dubinske analize možemo pronaći glavni predmet u skupini dokumenta, glavni predmet za svaki dokument i predmet sadržan u dokumentu [4].

Između dubinske analiza podataka, strojnog učenja i dohvata informacije ne postoji jasna granica. Spona između ove tri tehnike prikazana je na slici 7. Cilj dohvata informacije je pronaći nešto što već postoji u podacima na najbrži mogući način. Strojno učenje je tehnika generaliziranja postojećeg znanja do novih podataka što je preciznije moguće. Dubinska analiza podataka otkriva ono što je skriveno u podacima, nešto što nismo znali ranije [41].

Slika 7 Spona dubinske analize podataka, dohvata informacije i strojnog učenja

5.8. Tipovi sustava

Dubinska analiza podataka svoju primjenu je pronašla u gotovo svim područjima koji na bilo koji način generiraju ili sadrže podatke. Glavna zadaća dubinske analize podataka je otkrivanje zanimljivih uzoraka i znanja iz skupa podataka koji može biti različit, od internet podataka, grafova i mreža, geografskih podataka, tekstualnih i mnogih drugih [4].

Sustavi dubinske analize podataka mogu se klasificirati prema sljedećim kategorijama [5]:

- izvoru podataka (audio, video, tekst i slično)
- podatkovnom modelu (relacijski, objektni, objektno orijentirani, hijerarhijski)
- vrsti dobivenog znanja (označavanje, podjela, asocijacija, klasifikacija, grupiranje)
- korištenim tehnikama (strojno učenje, neuronske mreže, algoritmi, statistika vizualizacija, skladišta podataka)

Opći i najpopularniji formati pohrane podataka nad kojima se mogu izvršavati metode dubinske analize su relacijske baze podataka i skladišta podataka (Engl. Data Warehouse) [4]. Razlika između podataka u bazi i u skladištu podataka je da su podaci u bazi podataka u strukturiranoj formi dok podaci u skladištu podataka ne moraju biti u strukturiranoj formi. Prema [29] struktura podataka određuje kompatibilnost podataka za obradu.

Dubinska analiza teksta uključuje algoritme za analizu teksta s leksičkog i gramatičkog aspekta. Tekst se prelama u specifične strukture gdje se uzorci i opće informacije grupiraju i klasificiraju koristeći metode za dubinsku analizu podataka. *Dubinska analiza interneta* (Engl. Web) sastoji se od tri dijela: analize sadržaja, analize strukture interneta i analizu korištenja interneta. *Dubinska analiza slike* ima zadaću pronaći prostorne uzorke koji nisu eksplicitno spremljeni u slici, primjerice, boja, teksture, oblici, udaljenosti objekata na slici. Dubinska analiza *geoprostornih koordinata* ima svrhu otkrivanja zanimljivih uzoraka iz velike količine podataka nastale promatranjem zemlje [2, 42].

5.8.1. Baze podataka

Drugi naziv za bazu podataka je sustav za upravljanje bazom (Engl. Database Managament System – DBMS) koji se sastoji od skupa logički povezanih podataka, skupa programa koji omogućuju pristup bazi podataka, stvaranje datoteka te unos i organizaciju podataka. Sustav za upravljanje bazom podataka prema [43] odgovoran je za:

- pronalaženje i izdvajanje potrebnih informacija
- upravljanje integritetom
- sigurnosnim sustavom baze podataka
- kreiranje novih baza podataka
- određivanje sheme podataka (shema podataka je logička struktura podataka)
- definiranje baze podataka korištenjem opisnog jezika podataka (Engl. Data Definition Language – DDL)
- pisanjem upita nad bazom podataka koristeći jezik za manipulaciju (Engl. Data Manipulation Language - DML)
- spremanje velikih količina podataka itd.

Relacijske baze podataka su skupovi tablica gdje je svaka određena svojim jedinstvenim imenom. Svaka tablica sastoji se od skupa atributa (kolone ili polja) i parova (zapisa ili redaka). Svaki par u relacijskoj tablici definiran je jedinstvenim ključem opisanim svaki sa svojim skupom atributa. Semantički model podataka poznatiji je i kao Entitet-Veza model (Engl. Entity-Relationship – ER) koji se obično sastoji od relacijskih baza podataka. Entitet-Veza model prikazuje bazu podataka kao skup entiteta i veza između njih. Primjer sheme relacijske baze može biti skup tablica „Osoba“, „Proizvod“, „Zaposlenik“, „ProdajnoMjesto“, „Prodaja“. Tablice su definirane atributima gdje su atributi tablice „Prodaja“ ključevi koji definiraju vezu s drugim tablicama i referenciraju se na određene podatke iz tih tablica [4, 44, 45].

Na slici 8 dat je primjer skupa tablica relacijske baze podataka.

Osoba

Šifra	Ime	Prezime	Adresa
1	Ivan	Markić	Ljubuški
…	…	…	…

Polica

Šifra	Naziv	Cijena
1	AO	100,00
…	…	…

Zaposlenik

Šifra	Naziv	Vrsta
1	Ivan Markic 1	Suradnik
…	…	…

Prodajno mjesto

Šifra	Naziv
1	Ljubuški
…	…

Prodaja

Osoba	Polica	Zaposlenik	ProdanoMjesto	Količina	UkupnaCijena
1	1	1	1	1	100
…	…	…	…	…	…

Slika 8. Primjer skupa tablica u relacijskoj bazi podataka

Svaki zapis koji se nalazi u transakcijskoj bazi podataka je jedna transakcija, primjerice kupovina, zakup avionskih karata ili klik na web stranicu. Svaka transakcija određena je transakcijskim brojem i listom stavki koji je sačinjavaju. Dubinskom analizom ovakvih podataka možemo dobiti odgovore na pitanja tipa „Koji su proizvodi prodani skupa?“ ili „Koliko je puta prodan proizvod P1?“. Također se dubinskom analizom transakcijskih podataka mogu otkriti tzv. učestali uzorci koji mogu biti primjerice skupovi proizvoda koji su prodani zajedno.

U svrhu podrške za ovakve tipove problema razvijene su metodologije koje definiraju pravila izvršavanja transakcija. Jedna od njih je i ACID filozofija. Definicija ACID filozofije prema [46] podrazumijeva:

- Atomnost (Engl. Atomicitty) - izvođenje transakcije pod uvjetom sve ili ništa. Transakcija će u svakom slučaju biti dostupna za izvršavanje. Ne postoji nepoznati slučaj za transakciju, mogući slučajevi su uspješnost izvršavanja ili greška.
- Konzistentnost (Engl. Consistency) - podrazumijeva poštivanje svih pravila baze poput jedinstvenih ključeva.

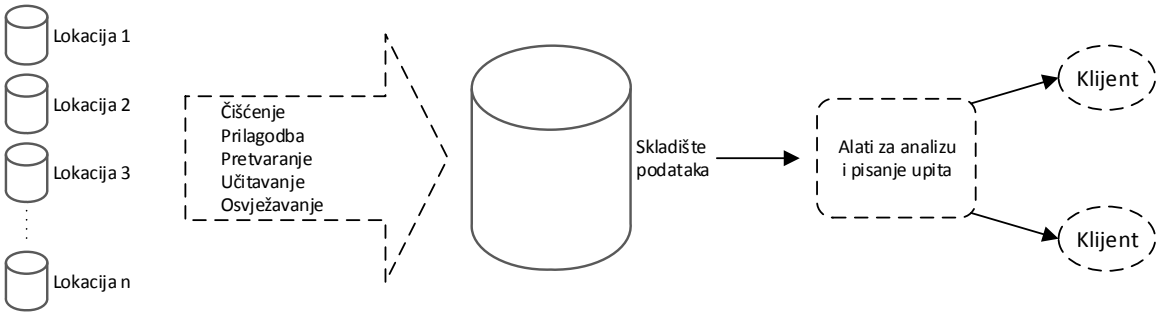
- Izolacija (Engl. Isolation) - svaka transakcija mora biti u mogućnosti da se izvrši bez obzira na broj transakcija koje se izvode u tom trenutku.
- Održivost (Engl. Durability) - podrazumijeva da se uvjeti koji se odnose na transakciju baze podataka ne smiju nikada biti izgubljeni.

Podacima u relacijskoj bazi podataka pristupa se pomoću upita baze podataka pisanih relacijskim jezikom poput SQL-a. Relacijski jezici za pristup podacima često koriste i gotove funkcije integrirane u DBMS poput agregatnih kao što je *Sum, Avg, Count, Max i Min*. Dubinskom analizom relacijskih baza podataka moguće je istraživati trendove ili uzorke podataka, primjerice predviđanja rizika za kupce prema ulaznim parametrima poput godina starosti ili ranijim informacijama, detektirati devijacije poput broja proizvoda koji su prodani ispod očekivanog broja u odnosu na ranije razdoblje [4, 47].

Istraživanje sustava baza podataka fokusirano je na kreiranje, održavanje i korištenje baze podataka za krajnje korisnike tj. zaposlenike neke organizacije. Znanstvenici koji se bave sustavima baza podataka odredili su opće principe modela, jezike za pisanje upita, obradu upita i njihovu optimizaciju, spremanje, indeksiranje i metode pristupa. Većina zadaća dubinske analize podataka zahtjeva podršku velikim skupovima podataka i obradu u realnom vremenu te moraju stvoriti zadovoljstvo kod korisnika u naprednim analizama podataka. Većina sustava baza podataka ima ugrađene analitičke sposobnosti koristeći skladišta podataka. Skladište podataka sadržava podatke iz više izvora i različitih vremenskih obilježja, konsolidira podatke u višedimenzionalni prostor i djelomično ih materijalizira u obliku kocke podataka koja predstavlja višedimenzionalni model podataka [4].

5.8.2. Skladišta podataka

Skladište podataka je repozitorij informacija skupljenih iz različitih izvora i spremljenih kao jedna cjelina. Na slici 9 prikazana je osnovna arhitektura skladišta podataka. Izrada skladišta podataka prolazi kroz nekoliko procesa poput čišćenja, prilagodbe, transformacije, učitavanja i periodičnog osvježavanja podataka. Podaci koji se nalaze u skladištu podataka organizirani su na način da se tiču glavnih subjekata specifične domene istraživanja. Primjeri subjekata domene istraživanja mogu biti „kupac“, „proizvod“, „dobavljač“ i „aktivnost“. To su obično arhivski podaci zadnjih šest ili dvanaest mjeseci rezimirani (Engl. Summarized) prema odgovarajućim grupama. Skladište podataka je modelirano kao višedimenzionalna struktura podataka, zvana *kocka* (Engl. Cube) koja se sastoji od *dimenzija* gdje svaka dimenzija odgovara atributima ili skupu atributa organiziranih u shemu. Svaka *ćelija* sadržava vrijednost ili neku agregatnu funkciju kao što je brojanje ili sažimanje [4]. Unaprijed sažeti i materijalizirani podaci uvelike ubrzavaju izvođenje upita za donošenje odluka, a komprimirajući podatke u skladištu podataka postižu se i mnoge druge prednosti poput smanjenja opsega podataka koji se može poslati komunikacijskim kanalima [28, 30].



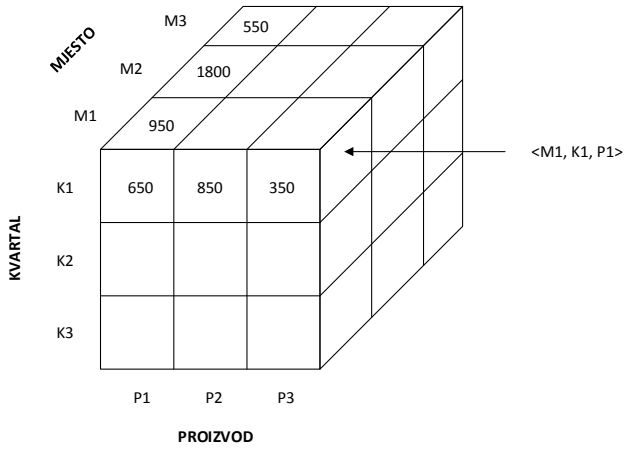
Slika 9. Arhitektura skladišta podataka

Primjer:

Na slici 10 prikazana je višedimenzionalna podatkovna kocka s tri dimenzije. Dimenzije je moguće promatrati kao osi Kartezijevog trodimenzionalnog koordinatnog sustava. Prva dimenzija (os x) je „mjesto“ s vrijednostima: M1, M2 i M3, druga dimenzija (os z) je „vrijeme“ s kvartalnim vrijednostima: K1, K2 i K3 i treća dimenzija (os y) je „proizvod“ s vrijednostima: P1, P2 i P3. Svaka ćelija kocke sadržava neku agregatnu vrijednost. Koristeći ovakav način analize moguće je dobiti informacije poput ukupne prodaje za prvi kvartal (Q1) i to za svaki proizvod koji odgovara prodaji proizvoda P1 u mjestu M1. Tražena vrijednost nalazi se spremljena u ćeliji u formatu tripleta $\langle M1, K1, P1\rangle$ [4, 28].

Triplet se može promatrati kao točka u Kartezijevom trodimenzionalnom koordinatnom sustavu pa vrijedi izraz:

$$T(x, y, z) \Leftrightarrow \langle M1, K1, P1\rangle \tag{2}$$



Slika 10. Višedimenzionalna kocka podataka

5.9. Pregled alata za dubinsku analizu podataka

Povijesno gledajući razlikuju se dvije vrste alata za dubinsku analizu. Prva vrsta je bila zasnovana na klasičnim statističkim metodama gdje je s vremenom taj model konvergirao iz postupka dokazivanja postavljenih hipoteza statističkim metodama u model koji generira nove hipoteze. Primjeri uključuju metode iz Bayesove teorije, regresije i opće analize komponenti. Druga vrsta metoda proizašla je iz područja umjetne inteligencije kao što su stabla odlučivanja, sustavi bazirani na pravilima i drugi. Međutim na nekoliko područja se ove dvije vrste preklapaju, primjerice kod neizrazite logike (Engl. Fuzzy), neuronskih mreža i metoda iz područja računalne inteligencije (Engl. Computational Intelligence). Većina istraživačkih prototipova zasniva se na skriptno orijentiranim matematičkim jezicima kao što su MATLAB (komercijalni) i R (besplatni). Ovi matematički programi nisu izvorno fokusirani na dubinsku analizu podataka, ali sadrže ugrađene matematičke funkcije kao i funkcije za vizualizaciju koje podržavaju implementaciju algoritama dubinske analize podataka. Opće akcije koje podržava većina alata su predviđanje budućih vrijednosti, pronalaženje učestalih uzoraka ili pronalaženje sličnih vremenskih serija grupiranjem (Engl. Clustering).

Analiza vremenskih serija (Engl. Time Series) igra važnu ulogu u većini dostupnih alata za dubinsku analizu podataka, uključujući predviđanje burzovnih tržišta, predviđanje potrošnje energije i drugo. Alati za dubinsku analizu podataka sa stajališta interakcije s korisnikom, tj. korisničkim sučeljem mogu se podijeliti na tri tipa [48, 49]:

1. tekstualno sučelje koje je teže za održavanje, ali lakše za automatizaciju zadaća,
2. grafičko sučelje sa strukturom izbornika
3. grafičko sučelje gdje korisnik definira i odabire sve moguće parametre od algoritama do tijeka izvođenja.

Prema [50] priprema podataka oduzima od 60% do 90% vremena dubinske analize podataka i doprinosi od 75% do 90% uspjehu projekta. Većina alata koji omogućavaju pripremu podataka podržavaju i sve ostale korake dubinske analize podataka koristeći metodu usmjerenih grafova. U pristupu usmjerenih grafova čvorovi grafa predstavljaju određene zadaće, a usmjerena veza između njih predstavlja tijek procesa od jedne zadaće do druge. Spomenuti pristup usmjerenih grafova koriste sljedeći besplatni alati WEKA (Engl. Waikato Enviroment for Knowledge Analysis), RapidMiner, KNIME (Engl. Konstanz Information Miner) i R [15, 2].

WEKA je jedan od najkorištenijih i najstarijih besplatnih programskih alata za strojno učenje i dubinsku analizu podataka, a referentan je za većinu drugih alata. Razvijen je na sveučilištu Waikato na Novom Zelandu i sadržava golem skup „state-of-the-art“ algoritama strojnog učenja napisanih u Javi. WEKA sadržava alate za regresiju, klasifikaciju, grupiranje, asocijativna pravila, vizualizaciju i predprocesiranje podataka. Idealan je alat za razumijevanje

procesa dubinske analize pa je zbog toga i vrlo dobro prihvaćen među akademskim djelatnicima i istraživačima baza podataka [15, 51, 52, 2].

Programski alat RapidMiner razvijen je u Njemačkoj na Dortmundskom Sveučilištu. Trenutno je za njegov razvoj zadužena kompanija Rapid-I. RapidMiner je napredni analitički alat za dubinsku analizu podataka, strojno učenje i prediktivnu analizu. Nudi podršku za dubinsku analizu teksta, interneta, slika i podataka otvorenih veza (Engl. Open Link Data). Moguće je također vršiti i predradnje poput procesa ETL kao i vizualizirati rezultate dubinske analize podataka [50, 53, 2].

Programski alat KNIME razvijen je na sveučilištu Kostanz u Njemačkoj. Za prezentiranje grafova koristi datoteke gdje je svaka datoteka određena za odgovarajući operator. Unutar svake datoteke nalazi se posebna XML datoteka s informacijama o operatorima kao primjerice parametrima i rezultatima spremljenim u binarnim datotekama. Ukoliko se na nekom koraku dogodi greška nema potrebe proces dubinske analize pokretati iz početka jer je svaki korak zapisan u svojoj datoteci [50, 54, 2].

R/Rattle je široko rasprostranjeni programski jezik za statističku analizu podataka. Ima mnogo ugrađenih metoda i funkcija iz područja statistike, ali i iz područja dubinske analize podataka. R podržava nelinearno modeliranje, statističke testove, analizu vremenskih serija, klasifikaciju kao i grupiranje, regresiju, metodu slučajnih šuma (Engl. Random Forest) i mnogo drugih. Rattle je korisničko sučelje za dubinsku analizu podataka koristeći R programski jezik čime je olakšano učenje programskog jezika R i korištenja njegovih funkcija [2, 55, 56, 57].

5.9.1. Zadaće alata dubinske analize podataka

Alati za dubinsku analizu podataka podržavaju višestruke zadaće. Od bitnijih se mogu izdvojiti:

- Zadaće *nadgledanog učenja s poznatim izlaznim promjenljivim* na skupu podataka, uključujući klasifikaciju (predviđanje klasa), neizrazitu (Fuzzy) klasifikaciju, regresiju (predviđanje realnih vrijednosti izlaznih promjenljivih uključujući iznimne slučajeve predviđanja budućih vrijednosti u vremenskim serijama na osnovu trenutnih i prošlih vrijednosti).
- Zadaće *nenadgledanog učenja bez poznatih izlaznih vrijednosti* nekog skupa podataka, uključujući grupiranje (pronalaženje i opisivanje grupa sličnih primjera u podacima koristeći neizrazite (Fuzzy) algoritme), asocijativno učenje (pronalaženje stavki grupe koje se često događaju u primjerima).
- Zadaće *polu nadgledanog* učenja gdje su izlazne promjenljive poznate samo na nekim primjerima.

Svaka od navedenih zadaća sastoji se od skupa pod zadaća koje se mogu ili ne moraju izvoditi samostalno. Možemo napravit općenitu podjelu zadaća alata za dubinsku analizu podataka prema [48] na:

- čišćenje i filtriranje podataka
- izdvajanje svojstava iz vremenskih serija
- transformacija svojstava (matematičke operacije, reduciranje dimenzija s kombinacijom linearnih i nelinearnih kombinacija)
- provjera (Engl. Evaluation) svojstava
- izračun sličnosti
- detekcija najsličnijih elemenata (K najbliži susjed)
- provjera modela (statistički)
- stapanje i optimizacija modela.

5.9.2. Kategorizacija

Općenita podjela alata za dubinsku analizu je na komercijalne i besplatne (Engl. Open Source), a dodatne kategorizacije moguće je napraviti prema [48] na sljedeće kategorije:

- Paketi alata dubinske analize (Engl. Data Mining Suites – DMS) fokusiraju se na dubinsku analizu podataka i uključuju većinu metoda. Podržavaju dubinsku analizu gotovo svih vrsta izvora podataka. Mogu se izdvojiti alati poput IBM SPSS Modeler, SAS Enterprise Miner, DataEngine, DataDetective, GhostMiner, Knowledge Studio, STATISTICA, itd.
- Alati za poslovnu inteligenciju (Engl. Bussiness inteligence - BI) čiji osnovni fokus nije na dubinskoj analizi, ali sadržava opće funkcionalnosti dubinske analize podataka posebno statističke metode u poslovnim aplikacijama. Posebno su razvijene izvještajne mogućnosti kao i podrška za učenje. Većina alata ovog tipa su komercijalna poput IBM Cognos 8 BI, Oracle DataMining, SAP Netweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM, itd.
- Matematički alati čiji fokus nije na dubinskoj analizi podataka, ali sadrže sve moguće algoritme i tehnike vizualizacije. Komercijalni alati su: MATLAB i R-PLUS, a besplatni alati su R i Kepler.
- Integrirajući paketi su prošireni skupovi većine besplatnih algoritama kao i samostalnih programskih paketa baziranih većinom na programskom jeziku Java, a to su: KNIME, WEKA s grafičkim sučeljem, KEEL i TANAGRA.
- Biblioteke (Engl. Libraries) dubinske analize podataka implementiraju metode dubinske analize kao skup funkcija. Ove funkcije moguće je ugrađivati i u druge programske alate koristeći za to predviđene API-je (Engl. Application Programming Interface). Biblioteke su obično pisane u JAVI ili C++. Primjeri biblioteka otvorenog

pristupa su WEKA, MLC++, JAVA Data Mining Package i LibSVM za potporne vektore. Primjer komercijalne biblioteke je Neurofusion.

- Razna rješenja koja su namijenjena specifičnim problemima kao primjerice problemima za dubinsku analizu teksta (GATE), obradi slike (ITK, ImageJ), otkrivanju droga (Molegro Data Modeler), analizi mikroskopskih slika (CellProfilerAnalyst) ili dubinskoj analizi gena (Partek Genomics Suite, MEGA). Prednost ovih rješenja je dobra podrška za izdvajanje svojstava, metode provjere i vizualizacija kao i formati uvoza.

5.9.3. Odabir softverskog alata

Instalacija i implementacija alata za dubinsku analizu podatka zahtjeva velika financijska ulaganja. Pogrešan odabir alata može uzrokovati financijske gubitke, ali i gubitak vremena. Sami proces odabira alata za dubinsku analizu ovisi od više faktora. Neki od faktora mogu biti poteškoće u pristupu primjenjivosti programskih paketa za poslovne potrebe organizacija zbog dostupnosti velikog broja alata na tržištu, postojanje nekompatibilnosti između različitog računalnog hardvera ili čak manjak tehničkog znanja osoba zaduženih za donošenje odluka. Ispravno odabran alat za dubinsku analizu podataka nudi značajne dijagnostičke metode za rješavanje problema i poboljšanje izlaznih rezultata [58].

5.10. Primjena dubinske analize podataka

Dubinska analiza podataka se primjenjuje u raznim područjima znanosti i industrije poput financijske, medicinske, reklamne, financijske, telekomunikacijske, itd. Korištenjem aplikacija za dubinsku analizu podataka organizacije nastoje poboljšati sposobnosti donošenja odluka prvenstveno poboljšanjem poslovnih procesa i zadržavanjem konkurentnosti. Aplikacije koje se bave analizom podataka imaju mogućnosti vizualizacije dobivenih rezultata i sposobnosti detaljne obrade podataka raznim metodama, od matematičkih do empirijskih. Dvije opće skupine aplikacija koje su u potpunosti integrirale tehnike dubinske analize podataka su *područje poslovne inteligencije* i *Internet tražilice* [4, 18, 52, 5].

Internet tražilice su specijalizirani računalni serveri koji pretražuju informacije na internetu. Rezultat pretraživanja (zvani pogodci) obično se korisniku prikazuju kao lista. Rezultati mogu biti Internet stranice, slike ili drugi tipovi datoteka. Razlikuju se od Internet direktorija po tome što su Internet tražilice mješavina algoritama i interakcije korisnika dok se sadržaj u Internet direktorijima održava obično od strane samog korisnika [4].

Jedno od područja gdje je dubinska analiza podataka također pronašla svoju primjenu je računarstvo u oblaku (Engl. Cloud Computing) zbog velikih računalnih resursa koji su na raspolaganju. To je koncept nuđenja servisa na Internetu. Davatelji usluga u oblaku nude servise u više formi poput Software-as-a-Service (SaaS) i Platform-as-a-Service (PaaS) [18].

5.10.1. Poslovna inteligencija

Poslovna inteligencija (Engl. Business Intelligence) je često korišten pojam u poslovnoj literaturi. Postoji mnogo definicija ovog pojma. Jedna od mogućih definicija je da poslovna inteligencija predstavlja korištenje programske inteligencije za poslovne aplikacije čineći ih tako inteligentnijim. U tom smislu se područje poslovne inteligencije posebno razvilo tijekom zadnja tri desetljeća. Moderna globalizacija tržišta zahtjeva da tvrtke za uspjeh budućeg poslovanja imaju korisne informacije o kupcima, tržištima, dobavljačima, resursima i konkurenciji, analizi šteta u primjerice industriji osiguranja ili otkrivanju prijevara u financijskim institucijama. Poslovna inteligencija prikazuje povijesne, trenutne i buduće poglede na poslovne aktivnosti i predstavlja skup tehnologija za odlučivanje koje omogućuju djelatnicima koji se bave traženjem znanja, korisnicima i analitičarima da brže donesu odluke koje će poboljšati poslovanje [28].

Neki od primjera korištenja su razna izvješća, OLAP (Engl. Online Aanalytical processing), upravljanje performansama, natjecateljska inteligencija, testovi i analitika predviđanja. Osnovne tehnike poslovne inteligencije su klasifikacija i predikcija. Grupiranje predstavlja bitnu ulogu u vezi s kupcem i grupira pojedine kupce zajedno prema sličnosti, a koristeći označavanje mogu se detaljnije razumjeti značajke kupaca i razviti prilagođene programe kao primjerice njihovo nagrađivanje [4].

Ubrzan razvoj ovakvih tehnologija rezultirao je povećanjem količine podataka zbog čega su trebale biti razvijene metode i tehnologije za rad s istim. Razvijene metode i tehnologije za rad s velikim količinama podataka kombiniraju dohvaćanje podataka, njihovo spremanje i upravljanje stečenim znanjem. Implicitna definicija sustava poslovne inteligencije podrazumijeva dostupnost informacije u pravo vrijeme, na pravom mjestu i u odgovarajućem formatu [59].

6. Tehnike dubinske analize podataka

Cilj procesa izdvajanja znanja definiran je očekivanim namjerama korištenja nekog sustava. Razlikujemo dvije vrste pristupa pri ostvarivanju cilja. Metodologija prvog pristupa zasnovana je na *provjeri postavljenih hipoteza* dok je drugi pristup zasnovan na *istraživanju* u kojem sustav autonomno pronalazi nove uzorke.

Pristup provjere prema [22] dijeli se dalje u dvije skupine funkcija:

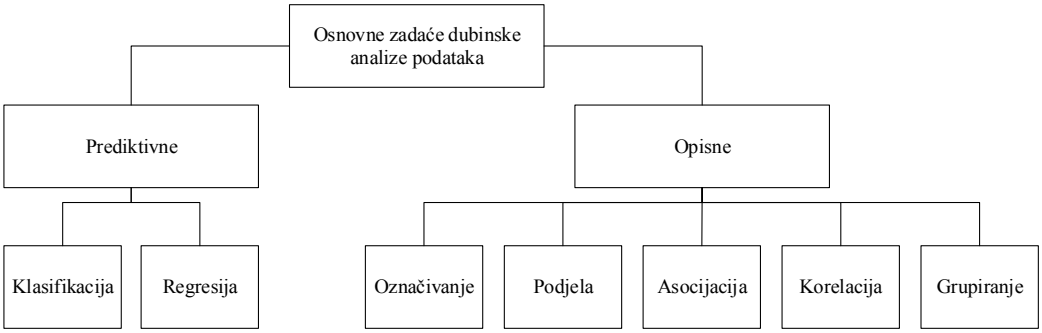
- opisne i
- prediktivne.

Opisne funkcije označavaju svojstva podataka u odgovarajućem skupu prikazujući ih razumljivim korisniku, a prediktivne funkcije indukcijom na skupu podataka vrše predikciju tj. predviđanje ponašanja entiteta u budućnosti. Korištenje različitih funkcija na testnom skupu podataka i potrebe za željenim izlazom uvelike određuje odabir funkcije [60, 14, 61, 4].

Nakon što su podaci koje želimo dubinski analizirati spremni za analizu potrebno je odrediti koji su sve mogući tipovi uzoraka koje rezultati dubinske analize mogu sadržavati. Slika 11 prikazuje osnovnu klasifikaciju tehnika istraživanja uzoraka u procesu dubinske analize podataka prema vrsti pristupa ostvarivanja cilja. Za određivanje odgovarajućeg uzorka moguće je prema [4, 61] koristiti sljedeće tehnike:

- označavanje i podjelu
- asocijaciju i korelaciju
- klasifikaciju i regresiju
- analizu grupiranjem
- vanjsku analizu.

Uzorci dobiveni korištenjem navedenih tehnika trebali bi vrijediti na novim podacima uz uvjet da uzorci budu kvalitetni, potencijalno korisni i razumljivi i to ako ne odmah, onda nakon dodatne obrade [22].



Slika 11. Osnovne zadaće dubinske analize podataka ([4])

6.1. Označavanje i podjela

Povezivanje podataka s klasama ili konceptima zbog opisa kako u općim tako i u specifičnim okolnostima moguće je vršiti analogno. Primjeri klasa su „računala“ i „printeri“ dok su primjeri koncepata „veliki kupci“ i kupci koji imaju unaprijed definiran budžet. Opisivanje klasa ili koncepata na ovakav način naziva se *klasno ili konceptno opisivanje*. Opisi klasa ili koncepata mogu se dobiti koristeći *označavanje, podjelu ili obje metode zajedno*.

Označavanje (Engl. Characterization) podatka je sažimanje (Engl. Summarization) općih oznaka klase (karakteristika). Odgovarajući podaci za klase korisnika najčešće se skupljaju pomoću izvođenja upita baze podataka. Moguće je koristiti i statističke metode, zatim metode iz OLAP sustava te objektno orijentirane metode indukcije. Nakon uspješnog označavanja potrebno je rezultate vizualizirati u obliku razumljivom korisniku.

Primjer:

Opisivanje oznaka nekih proizvoda čija se prodaja u prošloj godini povećala za 10%.

Rezultati opisa također mogu biti predstavljeni kao opće veze ili pravila formi koje se još nazivaju i pravila označavanja.

Podjela (Engl. Discrimination) podataka je usporedba općih mogućnosti podatkovnog objekta ciljne klase prema općim mogućnostima objekta jedne ili više suprotnih klasa. Ciljne i suprotne klase mogu biti određene od strane korisnika dok se odgovarajući podatkovni objekti dohvaćaju izvođenjem upita prema bazi podataka. Opisi podjele podataka izražavaju se pravilima zvanim *pravilima podjele*.

Primjer:

Usporedba dvije grupe kupaca od kojih su prva grupa oni koji kupuju više od dva puta mjesečno, a druga grupa su oni koji kupuju rjeđe. Rezultat analize je iznos od 80% onih koji kupuju često i to u dobi od 20 do 40 godina, a koji su visoko obrazovani kupci. Nadalje 60% kupaca koji rijetko kupuju su stariji ili jako mlađi, a nisu visoko obrazovani. Iz primjera se

mogu definirati dvije dimenzije „zanimanje“ i „ukupni_ulaz“ koje nam mogu pomoći da pronađemo razlike između dvije klase [4].

6.2. Asocijacije i korelacije

Dubinska analiza učestalih skupova stavki (Engl. Frequent Itemset Mining - FIM) je glavni i osnovni problem analize asocijativnih pravila. Početna faza je otkrivanje asocijativnih pravila, ali je rješenje problema generalizirano za otkrivanje i drugih uzoraka, primjerice učestalih sekvenci, epizoda i učestalih podgrafova [62].

Učestali uzorci kao što im samo ime govori su uzorci koji se često pojavljuju u skupu podataka i mogu se podijeliti na *učestale skupove stavki, podsekvence (zvane sekvencijalni uzorci) i podstrukture*.

Za otkrivanje učestalih uzoraka koristi se asocijacija i korelacija. *Učestali skupovi stavki* odnose na one stavke koje se nalaze zajedno u transakcijskom skupu.

Primjer:

Proizvod „proizvod_1“ i njegov dodatak „proizvod_2“ su skupovi koje kupuje većina kupaca prilikom kupovine.

Učestala događanja podsekvenci poput uzorka kupaca koji kupuje „proizvod_1“, a nakon njega „proizvod_2“ te dodatnu opremu uz „prozvod_2“ primjeri su (učestalih) sekvencijalnih uzoraka. *Podstrukture* se odnose na različite strukturne forme (primjerice, grafovi, stabla ili rešetke) koje se kombiniraju sa skupovima stavki ili pod sekvenci. Ako se pod struktura događa često naziva se *(učestali) strukturni uzorak*. Analiza učestalih uzoraka otkriva nam zanimljive asocijacije i korelacije između podataka i smanjuje probleme pri izvođenju dubinske analize podataka. Učinkoviti algoritmi glavni su uvjet uspjeha dubinske analize asocijativnih pravila, a problem dubinske analize učestalih uzoraka je prvenstveno jedan od problema analize asocijativnih pravila.

Nekoliko je metoda za pronalaženja učestalih skupova [4, 21, 2]:

- A priori algoritam
- rast uzoraka (Engl Frequent Pattern Growth, FP-Growth)
- vertikalni format podataka (Engl. Vertical Data Format).

A priori algoritam baziran je na činjenici korištenja prioritetnog znanja o svojstvima učestalih stavki. Koristi iterativni pristup u literaturi poznat pod nazivom *pretraga mudrijeg nivoa* (Engl. Level-Wise) gdje se k skupova stavki koristi za istraživanje $k + 1$ skupova stavki. Prvo se pronađe jedan skup učestalih stavki (tzv. Kandidat) skenirajući bazu i brojeći svakog od njih, sakupljajući ih jednog po jednog dok se ne zadovolji prag minimalne podrške. Proces se izvršava dok se ne pronađe k skupova stavki tj. dok se skup kandidata ne isprazni, a to znači

dok se ne skenira cijela baza podataka. Nakon toga vrši se podrezivanje onih skupova koji ne zadovoljavaju pragove minimalne podrške.

Rast uzoraka koriti strategiju „podijeli pa vladaj“. Prvi korak je sažimanje baze podataka predstavljajući učestale uzorke strukturom stabla koje sadržava informacije o vezama (asocijacijama). Proces se nastavlja podjelom sažete baze podataka na skupove *uvjetnih* baza podataka (poseban skup projecirane baze podataka) gdje je svaka novonastala baza podataka povezana s učestalom stavkom ili dijelom uzorka. Ovakvim načinom svaka se baza podataka koja je povezana dijelom uzorka analizira posebno. Prednost ovog algoritma je skeniranje baze podataka samo dva puta što uvelike smanjuje utrošak izračuna, ne postoji odabiranje skupa kandidata i postojanje strategije „podijeli pa vladaj“ što uvelike smanjuje prostor pretraživanja. Obje metode (A priori i Rast uzoraka) pogodne su za primjenu na podacima u skupu transakcija horizontalnog formata podataka, međutim *Vertikalni format podataka* uključuje obje metode, A priori i rast uzoraka [63, 62, 14].

Propozicijskom logikom u izrazu (3) opisana su klasifikacijska pravila u kojima je definiran zahtjev korisnika za informacijom o tome koji se proizvodi prodaju zajedno ili pojedinačno.

Primjer pravila iz transakcijske baze je:

$$kupovina(X, "proizvod\_1") \Rightarrow kupovina(X, "proizvod\_2") \quad [podrška = 25\%, pouzdanost = 50\%] \tag{3}$$

U primjeru promjenljiva X predstavlja kupca, a faktor *pouzdanosti* je vjerojatnost od 50% da nakon što je kupac kupio „proizvod_1“ da će kupiti i „proizvod_2“. Faktor *podrške* od 25% je vjerojatnost da će ova dva proizvoda biti prodana skupa tj. u jednoj transakciji. Ova asocijativna pravila povlače jedan atribut ili predikat koji se ponavlja, npr. „kupovina“. Asocijativna pravila koja sadrže samo jednu asocijaciju poput „kupovine“ su *jednodimenzionalna* asocijativna pravila.

Pravilo prikazano izrazom (4) indicira na kupce studije gdje je njih 30% od 20 do 30 godina, sa prihodom od 10K do 20K novčanih jedinica koji su kupili „proizvod_1“. Parametar pouzdanosti od 70% je vjerojatnost da kupci s ovim godinama i prihodom kupuju „proizvod_1“. Asocijativna pravila uključuju više atributa, npr. „godine“, „prihod“ i „kupovina“. Ukoliko se atributi usporede s višedimenzionalnim bazama podataka, gdje se svaki atribut odnosi na dimenziju, može se reći da je ovo pravilo *višedimenzionalno* asocijativno pravilo [4].

$$godine(X, "20 \ldots 30") \wedge prihod(X, 10\text{K} \ldots 20\text{K}") \Rightarrow \text{kupovina(X, "proizvod\_1")} \quad [podrška = 30\%, pouzdanost = 70\%] \tag{4}$$

Neka je $I = \{I_1, I_2, \dots, I_n)$ skup stavki, a D skup valjanih podataka tj. transakcija gdje je T transakcija koja nije prazan skup takav da $T \subseteq I$. Svaka je transakcija označena oznakom zvanom TID. Ako je A skup stavki sadržan u T, tj. $A \subseteq T$. Asocijativnim pravilima vrijedi implikacija $A \Longrightarrow B$, gdje je $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ i $A \cap B \neq \emptyset$. Pravilo $A \Longrightarrow B$ sadržava skup transakcija D s *podrškom* (Engl. Support) s gdje je s postotak transakcija D koji se ne nalazi u $A \cup B$. Tvrdnja se uzima kao vjerojatnost $P(A \cup B)$. Pravilo $A \Longrightarrow B$ je *pouzdanost* (Engl. Confidence) c u skupu transakcija D, gdje je c postotak transakcija D koje sadrže A i B. Navedena tvrdnja se uzima kao uvjetna vjerojatnost, $P(B|A)$.

Vrijede sljedeći izrazi:

$$podrška(A \Longrightarrow B) = P(A \cup B) \tag{5}$$

$$pouzdanost(A \Longrightarrow B) = P(B|A) \tag{6}$$

Kako asocijativna pravila ne bi bila tretirana kao nebitna potrebno je da budu zadovoljeni pragovi minimalne podrške i pouzdanosti. Pravila koja zadovoljavaju minimalne pragove podrške i pouzdanosti nazivaju se *jakima*. Dogovorno se uzima vrijednosti pouzdanosti od 0% do 100% umjesto od 0 do 1.0. Broj transakcija koje se nalaze u nekom skupu stavki naziva se *apsolutna podrška*. Drugi nazivi za apsolutnu podršku su *frekvencija*, *broj podrški* ili samo *broj* skupa stavki. Ako relativna podrška iz izraza (5) zadovoljava minimalni prag podrške te ako apsolutna podrška zadovoljava minimalni prag broja podrški skupa stavki I, tada je I *učestali* skup stavki.

Iz uvjeta izraza (6) vrijedi sljedeća definicija za provjeru jakosti asocijativnih pravila:

$$pouzdanost(A \Longrightarrow B) = P(B|A) = \frac{podrška(A \cup B)}{podrška(A)} = \frac{broj\_podrški(A \cup B)}{broj\_podrški(A)} \tag{7}$$

Gornja definicija predstavljena izrazom (7) prikazuje provjeru jakosti asocijativnih pravila. Pouzdanost pravila $A \Longrightarrow B$ može se izvesti iz broja podrški od A i $A \cup B$. Jednom kada su brojevi podrške od A, B i $A \cup B$ pronađeni, bez problema se mogu izvesti asocijativna pravila za $A \Longrightarrow B$ i $B \Longrightarrow A$ skupa s provjerom jakosti pravila [4].

6.3. Klasifikacija i regresija

6.3.1. Klasifikacija

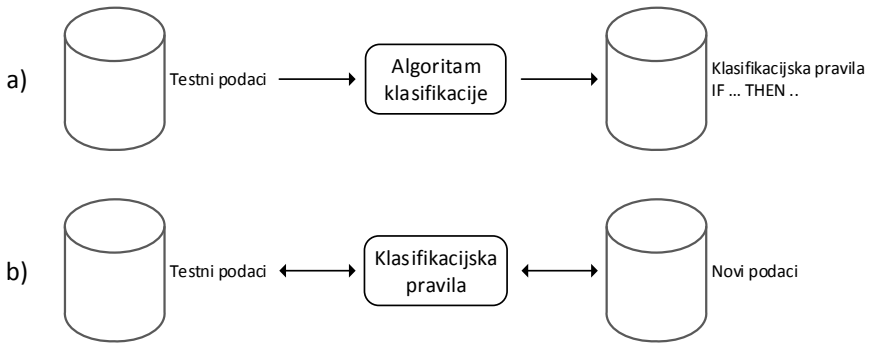
Klasifikacija (Engl. Classification) je nadgledana metoda strojnog učenja tj. proces pronalaženja modela (ili funkcije) koja opisuje i razlikuje podatkovne klase ili koncepte. Dobiveni modeli su bazirani na analizi skupa testnih podataka i koriste se za predviđanje oznake

klase objekta, ali samo za one klase čija je oznaka nepoznata. Dobiveni model može se prezentirati u više formi kao što su [4]:

- klasifikacijska pravila
- stabla odlučivanja
- matematičke formule
- neuronske mreže.

Proces klasifikacije sastoji se iz dva dijela. Prvi dio procesa klasifikacije je *učenje* u kojem se vrši konstrukcija modela tj. klasifikacijskim se algoritmima analizira skup testnih podataka. Proces prikazan na slici 12a napravljen je od parova baza podataka i njihovih povezanih oznaka klasa. Drugi dio procesa *klasifikacije* prikazan na slici 12b je dio u kojem se koriste testni podaci da se odredi točnost klasifikacijskih pravila koja se mogu primijeniti na novom skupu podataka nakon kojeg bi se predvidjeli nazivi klasa za nove podatke [14, 16].

Zadaća klasifikacije u procesu dubinske analize podataka je napraviti model koji će odrediti pripadnost klasi. Novi model koji se može promatrati kao bodovni primjenjuje se na pojedinačne klase, pa umjesto predikcije, predstavlja vjerojatnost da neki objekt pripada određenoj klasi [31].



Slika 12. Proces klasifikacije

Neka je X par koji se predstavlja kao atributni vektor od n dimenzija, $X = (X_1, X_2, \dots, X_n)$ prikazujući n mjeru napravljenu od parova n atributa baze podataka, odnosno $A_1, A_2, \dots, A_n$. Svaki par X pripada klasi definiranoj s atributom baze podataka zvanim *atribut oznake klase*. Sa stajališta klasifikacije podatkovni parovi nazivaju se i primjeri, instance, podatkovne točke ili objekti. U primjerima učenja gdje su oznake klase poznate, proces se naziva nadgledano učenje dok u primjerima gdje su oznake klase nepoznate proces se naziva nenadgledano učenje [4].

Sustavi koji izvode konstruiranje klasifikacija najučestaliji su alati za izvođenje dubinske analize podataka. Takvi sustavi uzimaju kao ulaz skupove slučajeva gdje svaki pripada malom

skupu klasa opisanih skupom atributa, a izlaz su klasifikatori koji točno mogu predvidjeti klase kojima će novi slučajevi pripadati [64].

Primjer korištenja klasifikacije kao dio procesa izdvajanja znanja uključuje klasifikaciju trendova na financijskim tržištima. Klasifikaciju većinom koriste bankovne institucije za donošenje odluka o davanjima zajma. Također se klasifikacija koristi kod otkrivanja prijevara u financijskim institucijama, osiguravajućim kućama i slično. Upravo je to čini jednom od najkorištenijih metoda za pravljenje modela u dubinskoj analizi podataka [22, 12].

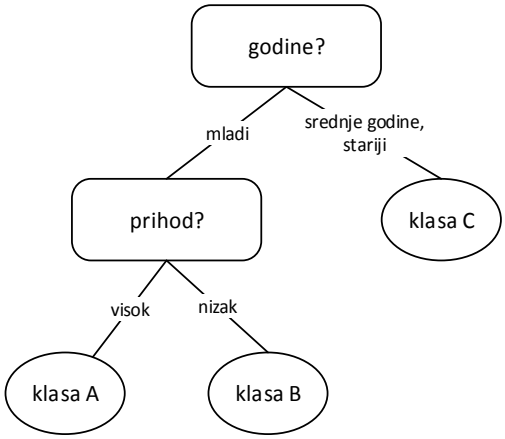
Klasifikacijska pravila (Engl. Classification Rules) poznata su još kao i „*if-else"* pravila. Kreiraju se tumačeći stablo odlučivanja gdje svako pravilo vrijedi samo za pojedini čvor. Ovim načinom svakom se čvoru pridružuje pojedina klasa. *„If-else"* pravilo prikazano je prema [65] izrazom:

$$IF\ uvjet\ THEN\ zaključak \tag{8}$$

Ukoliko pravila iz izraza (8) primijenimo na primjer kupca dobijemo prikaz klasifikacijskog modela IF-ELSE pravilima sljedećim izrazom:

$$\begin{gathered} godine(X, \text{"mladi"})\ \text{AND prihod(X, "}visok\text{")} \longrightarrow \text{klasa(X, "}A\text{")} \\ godine(X, \text{"mladi"})\ \text{AND prihod(X, "}nizak\text{")} \longrightarrow \text{klasa(X, "}B\text{")} \\ godine(X, \text{"srednje godine"}) \longrightarrow \text{klasa(X, "}C\text{")} \\ godine(X, \text{"stariji"}) \longrightarrow \text{klasa(X, "}C\text{")} \end{gathered} \tag{9}$$

Stablo odlučivanja (Engl. Decision Tree) je prikaz tijeka, slično strukturi stabla. Stablo odlučivanja prikazano je na slici 13 gdje svaki čvor koji nije list obilježava vrijednost atributa, svaka grana predstavlja testni izlaz, a svaki list sadržava oznaku klase. Čvor na vrhu naziva se korijenski čvor (Engl. Root) [4]. Stabla odlučivanja među najpopularnijim su postupcima strojnog učenja zbog jednostavnosti prezentacije rezultata indukcije u kojem je najbitnije da stablo sadržava samo važne atribute. Za konstrukciju klasifikatora stabla odlučivanja nije potrebno prethodno poznavanje domene istraživanja ili posebna parametrizacija [20].



Slika 13. Prikaz klasifikacijskog modela stablom odlučivanja

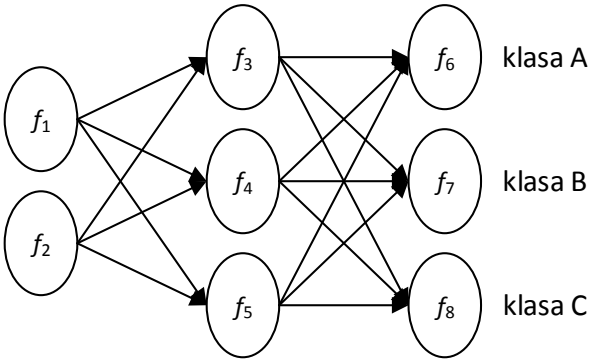
Neki od algoritama koji se koriste pri odabiru atributa kojim se testira svaki čvor koji nije list u strukturi stabla, a koji se razlikuju po mjerama odabira su ID3 (Engl. Iterative Dichotomiser) te C4.5 koji je poboljšana verzija ID3 i CART (Classification and Regression Trees) [4, 64].

Postoji i algoritam podrezivanja stabla (Engl. Pruning Tree), a koristi se za poboljšanje točnosti na način da se odstrane čvorovi koji su ujedno i šumovi u podacima. Opća svrha algoritma C4.5 je konstruiranje klasifikatora koji su predstavljeni strukturom stabla i koji su prikazani u obliku razumljivih skupova pravila. Ako je dat skup testnih podataka (slučajeva) S, algoritam C4.5 koristi već poznat metodu „podijeli pa vladaj“ na način da ako svi slučajevi u S pripadaju istoj klasi ili je S mali, listovi strukture stabla označeni su s najučestalijom klasom u S. S druge strane uzima se test baziran na jednom atributu s dva ili više izlaza. Tada se taj test napravi kao korijen stabla s jednom granom za svaki izlaz iz testa, podijeli se S u odgovarajuće pod skupove $S_1, S_2, \ldots$ prema izlazu za svaki test. Ova procedura primjeni se za svaki podskup rekurzivno. C4.5 koristi dvije vrste kriterija za rangiranje mogućih testova. Napredovanje informacije (Engl. Information gain) koji minimizira ukupnu entropiju[4] podskupa $\{S_i\}$ i standardno napredovanje koje dijeli napredovanje informacije prema podacima dobivenim od testnih izlaza. Stablo odlučivanja generirano CART algoritmom je binarno rekurzivna procedura pogodna za obradu kontinuiranih i nominalnih atributa kao ciljeva i prediktora. Podaci su obrađeni u sirovom obliku, a njihovo spremanje nije potrebno, a niti preporučljivo. Stabla mogu narasti do maksimalne veličine i to bez uporabe pravila zaustavljanja, a stablo se zatim podrezuje natrag prema korijenu (podjela po podjela) preko troškova složenosti podrezivanja (Engl. Cost Complexity Pruning). Sljedeća podjela koja će biti podrezana je ona koja najmanje pridonosi ukupnoj uspješnosti stabla na testnim podacima pri čemu je moguće uklanjanje više od jedne podjele u jednom trenutku. Postupak generira stabla koja su nepromjenljiva pod bilo

[4] U teoriji informacije entropija je mjera neodređenosti pridružena slučajnoj promjenljivoj.

kojim redoslijedom čuvajući transformaciju atributa prediktora. CART algoritam nastoji generirati ne samo jedno stablo, već i više podrezanih stabala od kojih su svi kandidati za optimalno stablo. Idealno stablo se evaluira po ocjenjivanju prediktivnih performansi svakog stabla u redoslijedu podrezivanja. Ocjenjivanje se mjeri na neovisnim testnim podacima i ako takvih podataka nema CART algoritam nastavlja s odabirom najboljeg stabla u sekvenci. CART je oštra suprotnost s metodama kao što su C4.5 koji stvaraju željene modele na temelju mjera trening podataka. CART mehanizam uključuje automatski balansiranje klase, automatsku podršku za nedostajuće vrijednosti, te omogućuje učenje osjetljivo na troškove performansi [64].

Neuronske mreže (Engl. Neural Network) pri korištenju za klasificiranje služe za prikaz rezultata klasifikacije kao skupine neurona – slične procesorske jedinice s težinom veze između jedinica [4]. Na slici 14 dat je prikaz klasifikacijskog modela. Kako bi se razumjeli načini na koje neuronske mreže mogu detektirati uzorke u bazi podataka može se povući analogija s načinom na koji ljudi uče. Ranije poznati podaci svjesno se u ovom slučaju primjenjuju na neuronsku mrežu i to jedan po jedan. Tehnike pronalaženja modela ne razlikuju se puno od tehnika korištenih u statistici ili drugim algoritmima dubinske analize podataka. Izlazni modeli neuronskih mreža su uvijek numerički što zahtjeva da i prediktori budu numeričke vrijednosti. Često su i ekspertima nerazumljivi pa trebaju biti pretvoreni u kategoričke vrijednosti kao primjerice plava boja treba biti prikazana nekom brojčanom vrijednosti [38].



Slika 14. Prikaz klasifikacijskog modela neuronskim mrežama

Razlikuje se više vrsta metoda za konstruiranje klasifikacijskog modela kao što su:

- Bayesova klasifikacija (Engl. Bayesian classification)
- potporni vektori (SVM - support vector machines)
- K najbliži susjed.

Metode koje se mogu koristiti za povećanje točnosti modela klasifikacije su primjerice metoda pakiranjem (Engl. Bagging), jačanjem (Engl. Boosting) i metoda slučajnih šuma (Engl. Random forest) [4].

Bayesova klasifikacija bazirana je na Bayesovom teoremu (Engl. Bayes Thorem) koja se bavi predviđanjem vjerojatnosti da dati par pripada određenoj klasi. Jednostavni Bayesovi klasifikatori zovu se još i *Naivni Bayesovi* (Engl. Naïve Bayes) klasifikatori. Naivna Bayesova klasifikacija polazi od pretpostavke da su vrijednosti atributa date klase neovisni od vrijednosti drugih atributa. Ova pretpostavka naziva se uvjetna neovisnost koja je uvedena da se pojednostavi izračun, tj. da se odredi kao „naivna“ [4]. U biti, ova metoda pokušava promijeniti klasifikaciju kako bi se povećala uvjetna vjerojatnost da grupa odgovara stvarnoj strukturi podataka pod uvjetom dostupnih podataka [66].

Bayesov teorem nazvan je prema Thomasu Bayesu i jednostavna je i učinkovita metoda koja se često koristi kod problema klasifikacije. Neka je X podatkovni par. U Bayesovoj terminologiji X je „dokaz“. Obično je opisan s napravljenim mjerenjem na skupu od n atributa. Neka je H neka hipoteza da par X pripada određenoj klasi C. Problem klasifikacije bavi se određivanjem vjerojatnosti P od H ili $P(H|X)$ da hipoteza H posjeduje *dokaz* ili podatkovni par X. Drugim riječima, tražimo vjerojatnost da par X pripada klasi C, obzirom da znamo opise atributa od X [4, 64, 67].

Neka je $P(H|X)$ je *posteriorna vjerojatnost* od H uvjetovanog na X.

Primjer:

Podatkovni par opisan vrijednostima godina starosti i prihoda u nekom vremenskom periodu, gdje je X 40-to godišnji kupac s prihodom od 1.000 novčanih jedinica. Pretpostavimo da je H hipoteza da će kupac kupiti „proizvod_1“. U tom slučaju se $P(H|X)$ odnosi na vjerojatnost da će kupac X kupiti proizvod_1 s obzirom poznavanje njegove dobi i prihoda.

Suprotno posterirornoj vjerojatnosti je *A priori vjerojatnost* $P(H)$. A priori vjerojatnost $P(H)$ predstavlja ranije znanje o hipotezi koje može utjecati na točnost te hipoteze.

Primjer:

Vjerojatnost da će kupac kupiti „proizvod_1“ unatoč godinama, prihodu ili bilo kojoj informaciji koja se može vezati za taj proizvod.

Posteriorna vjerojatnost P od H ili $P(H|X)$ je bazirana na većem broju informacija nego A priori vjerojatnost $P(H)$, koja je neovisna od X i predstavlja vjerojatnost da je hipoteza H točna ako postoji par podataka X.

Slično je vjerojatnost P od X ili $P(X|H)$ A priori vjerojatnost od X uvjetovanog na H tj. vjerojatnost pojavljivanja para podataka X ako je hipoteza H točna.

Primjer:

Ukoliko znamo da će kupac X kupiti proizvod_1, pitanje je kolika je to vjerojatnost da je kupac 40 godina star i da ima prihod od 1.000 novčanih jedinica.

$P(X)$ je A priori vjerojatnost od X i predstavlja vjerojatnost pojavljivanja parova podataka X u podacima.

Primjer:

Analogno ranije navedenom primjeru to je vjerojatnost da je osoba iz skupa kupaca starosti 40 godina i ima prihod od 1.000 novčanih jedinica.

Bayesov teorem predstavlja način izračuna posteriorne vjerojatnosti P od H ili $P(H|X)$ iz A priori vjerojatnosti $P(H)$, $P(X|H)$ i $P(X)$ što prema [4] daje sljedeći teorem prikazan izrazom:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{10}$$

Iz izraza (10) vidljivo je da posteriorna vjerojatnost raste ako rastu $P(X|H)$ i $P(H)$, a pada ako raste $P(X)$. Povećanjem vjerojatnosti pojave podatkovnih parova X neovisno o pojavi hipoteze H smanjuje se vrijednost pojave podatkovnih parova X kao dokaza točnosti hipoteze H. U strojnom učenju to podrazumijeva određivanje najboljih hipoteza iz nekog prostora H promatrajući skup testnih podataka.

Mnogo je načina za odabir ispravnog klasifikatora. *Potporni vektori* (Engl. Support Vector Machines - SVM) su prepoznati kao najbolji kandidati za to jer imaju mogućnost da se generaliziraju u višedimenzionalnom prostoru bez potrebe za prethodnim znanjem. Potporni vektori koriste linearne modele za implementaciju nelinearnih granica klasa pretvarajući ulazne vektore nelinearno u osobine višedimenzionalnih prostora. Bazirani su na jakoj vezi s statističkom teorijom učenja što znači da imaju ugrađene metode za minimiziranje rizika [68, 69].

U slučaju zadaće učenja dvije klase, cilj potpornih vektora je pronaći najbolju funkciju klasifikacije za razlikovanje članova dvije klase u testnom skupu podataka. Mjera koncepta najbolje funkcije klasifikacije može biti realizirana geometrijski. Za linearno razdvojene skupove podataka, funkcija linearne klasifikacije odgovara funkciji $f(x)$ koja prolazi sredinom dvije klase razdvajajući ih pri tome na dva dijela. Jednom kada se odredi ova funkcija instance novih podataka mogu se klasificirati jednostavnim testiranjem funkcije $f(x_n)$ gdje x_n pripada pozitivnim klasama ako je $f(x_n) > 0$ [64].

Izrada modela predikcije na osnovu *najbližih susjeda* zajedno s grupiranjem je jedna od najstarijih metoda za izradu modela predikcije. Metoda najbližih susjeda i grupiranje su poprilično slične. Osnova metode najbližih susjeda je određivanje vrijednosti predikcije na osnovu sličnih prediktora ili susjeda. Ova metoda najsličnija je ljudskom razmišljanju zato što

detektira najčešće primjere. Objekti koji nose oznaku „bliski“ imati će slične vrijednosti predikcije te ako je poznata vrijednost predikcije jednog objekta, moguće je predvidjeti i vrijednost njegovog najbližeg susjeda. Sposobnost metode K najbližeg susjeda je mogućnost usporedbe s K najbližih susjeda, a ne samo s najbližim susjedima [38].

Metoda najbližih susjeda pronalazi grupe od K objekata u testnom skupu koje su najbliže testnom objektu i bazira dodjeljivanje oznaka na prevlasti određene klase u susjedstvu. Tri su ključna elementa ovog pristupa. Prvi element je skup označenih objekata, primjerice skup spremljenih zapisa, udaljenost ili mjera sličnosti za izračun udaljenosti između objekata i vrijednost od K, a to je broj najbližih susjeda. Da bi se klasificirali neoznačeni objekti izračunava se udaljenost od označenih objekata, identificira se K najbliži susjed, a oznake klasa najbližih susjeda koriste se za označavanje oznake klase objekta [64].

6.3.2. Regresija

Regresija (Engl. Regression) pokušava za svaku klasu procijeniti ili predvidjeti numeričku vrijednost promjenljivih koje joj pripadaju. Regresija je slična klasifikaciji, a razlikuje se prema tome što klasifikacija predviđa da li će se nešto dogoditi, dok regresija predviđa koliko puta će se puta nešto dogoditi [31].

Regresijski model sa svojim metodama evaluacije pogodan je za područja gdje su klasifikacijske oznake predikcije kategoričke (diskretne, nesortirane). Regresija se ponajprije koristi da se predvide nedostajuće ili nedostupne numeričke vrijednosti prije nego same oznake klasa. Predviđanje se odnosi na obje vrste predviđanja, numeričko i klasno predviđanje. Regresijska analiza je statistička metodologija koja se prije svega koristi za numeričko predviđanje i kao takva obuhvaća i identifikaciju trendova nad dostupnim podacima [4, 12].

Više je različitih tipova regresije u statistici, ali opća ideja je kreiranje modela koji preslikava vrijednosti od prediktora na način da je mogućnost nastanka greške kod predviđanja najmanja. Jednostavna forma regresije je linearna regresija koja sadržava jednog prediktora i jednu predikciju. Forma se može prikazati jednostavnom linearnom funkcijom definiranom izrazom:

$$f(x) = a \cdot X + b \tag{11}$$

Iz izraza (11) X os predstavlja vrijednosti predikcije, a Y os vrijednosti prediktora. Zadaća regresije je pronaći liniju između svake vrijednosti prediktora i predikcije, a linija gdje je distanca između vrijednosti osi X i Y najmanja se izabire za model predikcije.

Ako realno stanje preslikamo kao linearnu funkciju, onda se to definira kao što je prikazano izrazom (12):

$$Predikcija = a \cdot Prediktor + b \tag{12}$$

Tipičan primjer klasifikacije i regresije je klasifikacija velikog skupa proizvoda neke trgovine baziranih na tri tipa odziva prodajne kampanje: „dobar“, „blag“ i „nikakav“ odziv. Naš zadatak je izvesti model za svaki od ove tri klase na osnovu opisa stavki kao što je „cijena“, „brand“, „mjesto_izrade“, „tip“ i „kategorija“. Rezultat klasifikacije trebao bi maksimalno razlikovati svaku klasu jednu od druge, dajući jasnu sliku skupa podataka. Ukoliko je rezultat predstavljen u obliku stabla odlučivanja moguće je otkriti i druge mogućnosti koje nam mogu pomoći u razlikovanju objekata pojedine klase i pomoći nam da razumijemo i učinkovitije dizajniramo buduće prodajne kampanje. Ovo povlači sljedeću mogućnost da umjesto kategoričkih predikcija za svaki proizvod promatramo iznos povrata ulaganja kojeg će svaki proizvod generirati u nadolazećem periodu prodaje na osnovu prethodnih podataka [4, 12, 2].

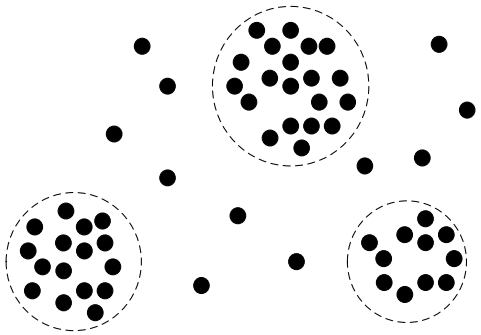
6.4. Analiza grupiranjem

Analiza grupiranjem (Engl. Cluster Analysis) pokušava podijeliti skup u grupe bazirajući se pri tome na sličnosti između objekata skupa. Objekti skupa grupirani na taj način imaju međusobno veću sličnost nego s objektima druge grupe. Postoje dvije vrste analize grupiranjem, *tradicionalna i konceptualna*. Tradicionalna analiza identificira grupe sličnih objekata, ali ne može opisati pojedinu klasu dok konceptualni način grupiranja čini isto to koristeći odgovarajuća pravila [66, 12, 16, 2].

Za razliku od klasifikacije i regresije koje analiziraju testne skupove označenih klasa, analiza grupiranjem analizira podatkovne objekte bez konzultiranja oznake klase. U većini slučajeva podaci s označenim klasama ne postoje na početku. Analiza grupiranjem može se koristi za označavanje naziva klase grupe podataka. Objekti se grupiraju prema principu maksimalne i minimalne sličnosti između klasa. To znači da objekti u istoj grupi imaju najvišu sličnost pri međusobnoj usporedbi dok su različiti s objektima iz druge grupe. Svaka grupa se može promatrati kao klasa objekta iz koje je moguće izvesti pravila. Ova vrsta analize također olakšava formiranje taksonomija, a to su skupine poslužitelja u hijerarhiji klasa koji grupiraju slične događaje [4, 14].

Primjer:

Analiza grupiranjem može se provesti nad kupcima za detektiranje jednakih podskupova kupaca. Grupe prikazane na slici 15 mogu predstavljati tri različite skupine kupaca iz tri različita mjesta i kao takve mogu biti ciljne grupe promatranja za primjerice odjel marketinga.



Slika 15. Prikaz primjera grupiranja u 2-D formi

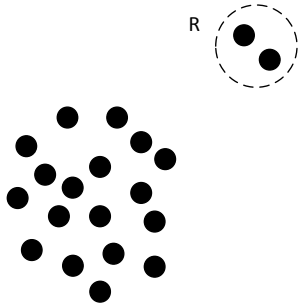
Analiza grupiranjem često se koristi u značenju segmentacije. Segmentacija je u poslovnom svijetu ptičji pogled na neki poslovni sustav. Informacije koje se dobiju grupiranjem često se koriste za označavanje (Engl. Tagging) osoba u bazi podataka. Jednom kada su podaci označeni moguće je dobiti pogled s visokog nivoa na ono što se dešava u bazi podataka [38].

Jedan od najosnovnijih algoritama nenadgledanoga učenja koji rješava probleme grupiranja je *k-means* algoritam. *K-means* algoritam je algoritam koji dijeli skup podatka u k grupa. Algoritam djeluje na skupu vektora dimenzije d, $D = \{x_i | i = 1, \dots, N\}$ gdje x_i označava i-tu točku podataka. Algoritam se inicijalizira biranjem k točaka kao osnovnih k grupa ili težišta. Glavna ideja je da se definira skup od k težišta, jedan za svaku grupu. Težišta trebaju biti postavljeni na različite načine (udaljene jedne od drugih) jer se zbog različitog položaja dobije drugačiji rezultat. Sljedeći korak je da se svaka točka koja pripada određenom skupu podataka poveže do najbližeg težišta. Kada više nema točaka grupiranje je gotovo, a rezultat je podjela podataka. Svaki predstavnik grupe se pomiče na centar (Engl. Mean) i postupak se ponavlja sve dok su kretanja težišta onemogućena, tj. dok se grupe ne formiraju [64, 70].

6.5. Vanjska analiza

Skupovi podataka mogu sadržavati objekte koji se ne slažu s generalnim ponašanjima ili modelom podataka, a koji se zovu *vanjski* objekti (Engl. Outlier Analysis). Ova analiza poznata je još i pod nazivom analiza anomalija, otklanjanje šumova i kategorizirana je kao metoda iznimki. Vanjski objekti mogu se detektirati koristeći statističke testove koji pretpostavljaju mogući model podatka ili koristeći mjeru udaljenosti gdje se udaljeni objekti iz drugih grupa označavaju kao vanjski [4]. Na slici 16 dat je primjer objekata označenih kao vanjski objekti.

Šum (Engl. Noise) se definira kao skup podatkovnih objekata koji iz nekog razloga odstupaju od ostalih skupova. Uzroci stvaranja šumova mogu biti u procesu skupljanja podataka ili grešaka bilo koje vrste. Svako postojanje šuma unutar skupa može smanjiti kvalitetu induciranih modela, a njegovo otklanjanje može otkriti specifičnije koncepte koji kada se interpretiraju mogu predstavljati važno novo znanje [20, 12].



Slika 16. Primjer objekata označenih kao vanjski

Vanjska analiza skupa s analizom grupiranja je vrlo važna metoda iznimki u dubinskoj analizi podataka, ali iznimka nije ista izolirana točka. Neki se podaci mogu pojaviti kao normalne izolirane točke koje mogu biti nebitne korisnicima, ali za donositelje odluka mogu biti prava prilika ili znak [71].

Primjer:

Vanjskom analizom može se otkriti neovlašteno korištenje kreditne kartice detektirajući neočekivano velike iznose u usporedbi s normalnom potrošnjom korisnika po kreditnoj kartici, a moguće je detektirati i mjesta gdje se pojavljuju neočekivane vrijednosti. Drugi primjer je obračun poreza gdje osobe koje obračunavaju porez mogu prema neočekivanim vrijednostima detektirati neke greške ili probleme.

6.6. Učinkovitost metoda

Uspješnost i učinkovitost metoda dubinske analize mjeri se pomoću dva faktora:

1. Prvi faktor je vrijeme izračuna
2. Drugi faktor je robusnost algoritama

Ukoliko odgovara tvrdnja da ako vrijeme izvođenja algoritma raste brže nego linearna ovisnost kvadrata broja podataka koji se traže zaključuje se da taj algoritam nije pogodan za obradu velikih količina podataka. Na vrijeme izvođenja može se djelovati smanjivanjem opsega promatranja, ali i kompresijom podataka. Algoritmi trebaju biti robusni da se mogu nositi s nepotpunim i oštećenim podacima. Problem je taj što oštećeni podaci proizvode umjetne uzorke. Sustav ne bi trebao razmatrati ovakve primjere kao dio normalne analize već ih ako je to moguće treba detektirati te izvijestiti o tome u posebnom koraku [66].

7. Zaključak

U ovom radu predstavljen je SOTA (State-Of-The-Art) područja dubinske analize podataka i otkrivanja znanja. Obrađen je u biti pregled područja ovih brzo rastućih računalnih grana znanosti. Budući da dubinska analiza podataka obrađuje široko područje, opisani su koraci dubinske analize podataka od sakupljanja, pripreme, obrade i analize podataka pa do korištenja dobivenih rezultata. U radu su opisani načini i metode korištenja i evaluacije dostupne literature, definicije pojmova iz područja dubinske analize podataka i otkrivanja znanja. Opisane su metode evaluacije dobivenog modela podataka kao i metode prezentacije dobivenog znanja. Uz tehnike koje se koriste u procesu dubinske analize podataka iz drugih domena znanosti, dat je pregled besplatnih, ali i komercijalnih alata za dubinsku analizu podataka koji podržavaju sve korake procesa. Kako dubinska analiza podataka svoju primjenu nalazi u mnogim granama industrije opisana je primjena dubinske analize podataka kao podrška poslovnim sustavima ili popularno korišteni pojam poslovne inteligencije. Zadnji dio opisuje tehnike dubinske analize podataka i njima pripadajuće algoritme. Na kraju je opisana učinkovitost metoda dubinske analize podataka i opisani su osnovni uvjeti po kojima je metoda učinkovita ili ne.

8. Popis slika

9. Popis tablica

10. Reference

[1] S. Neeraj, F. C. Raul, Abhishek, I. Abhishek, N. Chaitali, M. Adi-Cristina, N. Mallarswami and D. Mirela, Database Fundamentals, First Edition, Markham: IBM Corporation, 2010.

[2] A. Lausch, A. Schmidt i L. Tischendorf, »Data mining and linked open data – New perspectives for data analysis in environmental research,« *Ecological Modelling 295,* p. 5–17, 2015.

[3] T. Connolly and C. Begg, Database Systems - A Practical Approach to Design Implementation, and Management, Fourth Edition ed., Harlow: Pearson Education Limited, 2005.

[4] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition, Waltham: Elsevier Inc., 2012.

[5] N. Padhy, P. Mishra and R. Panigrahi, "The Survey of Data Mining Applications And Feature Scope," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),* vol. 2, 2012.

[6] Scopus, »The SCImago Journal & Country Rank,« Scopus, [Mrežno]. Available: http://www.scimagojr.com/. [Pokušaj pristupa 22 12 2014].

[7] Elsevier, »Journal Metrics from Elsevier,« Elsevier B.V. , [Mrežno]. Available: http://www.journalmetrics.com/. [Pokušaj pristupa 22 12 2014].

[8] V. P. Guerrero-Bote i F. Moya-Anegón, *A further step forward in measuring journals' scientific prestige: The SJR2 indicator,* Elsevier Ltd, 2012.

[9] Institut Ruđer Bošković, »Knjižnica Instituta Ruđer Bošković,« Institut Ruđer Bošković, [Mrežno]. Available: http://lib.irb.hr/. [Pokušaj pristupa 6 1 2015].

[10] S. Earley, "Big Data And Predictive Analytics: What's New?," *Computer,* 2014.

[11] V. Grbavac, Informatika: kompjutori i primjena, Zagreb: Sveučilište u Zagrebu, 1994.

[12] E. Ngai, Y. Hu, Y. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems,* 2011.

[13] J. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda and C. Carlsson, "Past, present, and future of decision support technology," *Decision Support Systems,* 2002.

[14] F. Jianhua and L. Deyi, "An Overview of Data Mining and Knowledge Discovery," *J. of Comput. Sci. & Technol.,* vol. 13, 1998.

[15] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi and E. M. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods,"

(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, 2011.

[16] M. J. Zaki and W. J. Meira, Data Mining and Analysis Fundamental Concepts and Algorithms, New York: Cambridge University Press, 2014.

[17] S. Kabira, S. Ripon, M. Rahman i T. Rahman, »Knowledge-Based Data Mining Using Semantic Web,« u *2013 International Conference on Applied Computing, Computer Science, and Computer*, Bangladesh, 2013.

[18] M. Revels i H. Nussbaumer, *Data Mining and Data Warehousing in the Airline Industry,* Kentucky, 2013.

[19] Rose Business Technologies, »Rose Business Technologies,« [Mrežno]. Available: http://www.rosebt.com/. [Pokušaj pristupa 2 2 2015].

[20] Institut Ruđer Bošković, »Otkrivanje znanja dubinskom analizom podataka - Priručnik za istraživače i studente,« 2014. [Mrežno]. Available: http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf.

[21] Microsoft, "Data Mining (SSAS)," Microsoft, 2014. [Online]. Available: http://msdn.microsoft.com/en-us/library/bb510516.aspx.

[22] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *American Association for Artificial Intelligence,* 1996.

[23] S. Seshasai, A. Gupta and A. Kumar, "An integrated and collaborative framework for business design: A knowledge engineering approach," *Data & Knowledge Engineering,* p. 157–179, 2005.

[24] D. Arnott and G. Pervan, "Eight key issues for the decision support systems discipline," *Decision Support Systems,* p. 657–672, 2008.

[25] G. Jifa i Z. Lingling, »Data, DIKW, Big data and Data science,« *2nd International Conference on Information Technology and Quantitative Management, ITQM,* p. 814–821, 2014.

[26] B. Denkena, J. Schmidt i M. Krüger, »Data mining approach for knowledge-based process planning, 2014,« u *2nd International Conference on System-Integrated Intelligence: Challenges for Product and Production Engineering*, 2014.

[27] B. M. Ramageri, "Data Mining Techniques and Applications," *Indian Journal of Computer Science and Engineering,* Vols. No. 4 301-305, 2010.

[28] S. Chaudhuri, U. Dayal and V. Narasayya, "An Overview of Business Intelligence Technology," *Communications of the acm,* 2011.

[29] M. PhridviRaj and C. GuruRao, "Data mining – past, present and future – a typical survey on data streams," *The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013),* 2014.

[30] C. Date, An Introduction to Database Systems Eight Edition, Perason Education Inc., 2004.

[31] F. Provost and T. Fawcett, Data Science for Business, O'Reilly, 2013.

[32] E. A. Yokome i F. L. Arantes, »Meta-DM: An ontology for the data mining domain,« *Revista de Sistemas de Informacao da FSMA,* 2011.

[33] R. Studer, V. R. Benjamins i D. Fensel, »Knowledge Engineering: Principles and methods,« u *Data & Knowledge Engineering 25*, 1998.

[34] B. Chandrasekaran, J. R. Josephson i V. R. Benjamins, »What Are Ontologies, and Why Do We Need Them?,« *IEEE Intelligent Systems,* 1999.

[35] A. Kazi i P. D. Kurian, »An Ontology Based Approach to Data Mining,« *International Journal of Engineering Development and Research,* svez. 2, br. 4, 2014.

[36] H. Gorskis i Y. Chizhov, »Ontology Building Using Data Mining Techniques,« *Information Technology and Management Science,* 2012.

[37] A.-E. Elsayed, S. R. El-Beltagy, M. Rafea i O. Hegazy, *Applying data mining for ontology building,* 2007.

[38] A. Berson, S. Smith and K. Thearling, "An Overview of Data Mining Techniques," in *Building Data Mining Applications for CRM*, McGraw-Hill Companies, 1999, p. 488.

[39] F. Pinto, M. F. Santos and A. Marques, "Ontology based Data Mining – A contribution to Business Intelligence," in *10th WSEAS Int. Conference on mathematics and computers in business and economics*, 2009.

[40] G. J. Myatt i W. P. Johnson., Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, Hoboken, New Jersey: John Wiley & Sons, Inc., 2014.

[41] O. Maimon i L. Rokach, Data Mining and Knowledge Discovery Handbook, New York: Springer, 2013.

[42] X. Wu, X. Zhu, G.-Q. Wu i W. Ding, »Data Mining with Big Data,« *IEEE transactions on knowledge and data engineering,* svez. 26, 2014.

[43] H. Garcia-Molina, J. D. Ullman and J. Widom, Database Systems: The Complete Book, New Jersey 07458: Prentice Hal, 2002.

[44] T. Kraska and B. Trushkowsky, "The New Database Architectures," *IEEE Internet Computing,* 2013.

[45] H. J. Park, P. H. Kim, M. Marsico i N. Rasheed, »Data Mining Strategies for Real-Time Control in New York City,« u *The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014)*, New York, 2014.

[46] Eric Brewer, "CAP twelve years later: How the "Rules" Have Changed," *IEEE Computer Society,* 2012.

[47] M. Yannakakis, "Perspectives on Database Theory," *IEEE,* 1995.

[48] R. Mikut and M. Reischl, "Data mining tools," John Wi ley & Sons , Inc. WIREs Data Mining Knowl Discov, 2011.

[49] N. M. Adams, »Perspectives on data mining,« *International Journal of Market Research,* svez. 52, br. 1, 2010.

[50] P. M. Goncalves, R. S. Barros and D. C. Vieira, "On the Use of Data Mining Tools for Data Preparation in Classification Problems," *IEEE/ACIS 11th International Conference on Computer and Information Science,* 2012.

[51] University of Waikato, "WEKA the University of Waikato," University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 15 12 2014].

[52] S. Natek i M. Zwilling, »Student data mining solution–knowledge management system related to higher education institutions,« *Expert Systems with Applications 41,* p. 6400–6407, 2014.

[53] RapidMiner, »RapidMiner,« RapidMiner, [Mrežno]. Available: https://rapidminer.com/. [Pokušaj pristupa 15 12 2014].

[54] KNIME, "KNIME (Konstanz Information Miner)," KNIME.com Headquaters, [Online]. Available: http://www.knime.org/. [Accessed 15 12 2014].

[55] »Rattle,« [Mrežno]. Available: http://rattle.togaware.com/. [Pokušaj pristupa 22 12 2014].

[56] G. J. Williams, »Rattle: A Data Mining GUI for R,« *The R Journal,* svez. 1, 2009.

[57] R. Gentleman i R. Ihaka, »R Project,« Statistics Department of the University of Auckland, [Mrežno]. Available: http://www.r-project.org/. [Pokušaj pristupa 22 12 2014].

[58] N. Bhargava, A. Aziz and R. Arya, "Selection Criteria for Data Mining Software: A Study," *IJCSI International Journal of Computer Science Issues,* vol. 10, no. 3, 2013.

[59] S. Negash, »Business Intelligence,« *Communications of the Association for Information Systems,* 2004.

[60] S. K. Mohamada i Z. Tasira, »Educational data mining: A review,« *The 9th International Conference on Cognitive Science,* 2013.

[61] A. Peña-Ayala, »Educational data mining: A survey and a data mining-based analysis of recent works,« *Expert Systems with Applications 41,* p. 1432–1462, 2014.

[62] W. Song, B. Yang and Z. Xu, "An improved algorithm for mining frequent itemsets," *Knowledge-Based Systems,* 2008.

[63] S. Nasreen, M. A. Azamb, K. Shehzada, U. Naeemc and M. A. Ghazanfar, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey," in *The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)*, 2014.

[64] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, "Top 10 algorithms in data mining," Springer, 2008.

[65] I. H. Witten, E. Frank and M. A. Hall, Data Mining - Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[66] D. N. Bissantz and D. J. Hagedorn, "Data Mining," *Business & Information Systems Engineering,* 2009.

[67] I. Pavlić, Statistička teorija i primjena, Zagreb: Grafički zavod Hrvatske, 1970.

[68] M. A. Mohd Shukran, M. Adib Khairuddin and K. Maskat, "Recent Trends in Data Classifications," in *International Conference on Industrial and Intelligent Information (ICIII 2012)*, 2012.

[69] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose, "Detection of financial statement fraud and feature selection using data," *Decision Support Systems, Elsevier,* 2011.

[70] A. Likas, N. Vlassis i J. J. Verbeek, »The global k-means clustering algorithm,« *Pattern Recognition 36,* p. 451 – 461, 2003.

[71] B. Liu, G. Xu, Q. Xu and N. Zhang, "Outlier Detection Data Mining of Tax Based on Cluster," *International Conference on Medical Physics and Biomedical Engineering,* 2012.