Istraživanje podataka 1

Uvod

Ne postoji stroga definicija **istraživanja podataka**, ali se najčešće koriste:

- proces koji uključuje prikupljanje podataka, njihovo čišćenje, obradu, analizu i dobijanje korisnih saznanja o njima
- pronalaženje skrivenih informacija u bazi podataka
- netrivijalno izdvajanje implicitnih, prethodno nepoznatih i potencijalno korisnih informacija iz baza podataka
- integralni deo otkrivanja znanja u bazama podataka (Knowledge Discovery in Databases - KDD)

Istraživanje podataka je **potrebno** jer:

- stalno se prikupljaju velike količine podataka
- postoji velika količina ravnih podataka (raw data) za obradu
- tradicionalne metode za analizu nisu pogodne zbog količine i prostorno-vremenske prirode podataka, a često veliki deo podataka nikada i ne stiže do analize
- postoje sakrivene informacije koje nisu odmah ili lako uočljive
- omogućava dobijanje konciznih i upotrebljivih informacija za neki cilj
- ima primenu u različitim oblastima kao što su nauka, medicina, inženjerstvo, poslovanje
- računari su snažniji

Poreklo istraživanja podataka može se naći u statistici, veštačkoj inteligenciji, mašinskom učenju, prepoznavanju obrazaca, tehnologijama bazama podataka i paralelnom i distribuiranom računarstvu. IP je **zasnovano na algoritmima**. Svaki algoritam pokušava da ukalupi podatke u neki model. Bira se model koji je najbliži karakteristikama podataka. Neke od oblasti koje su bliske IP su Big data, Predictive analytics, Data Science i slično. Faze procesa IP su skladištenje podataka, njihova transformacija, prečišćavanje, ponovno skladištenje i upotreba, istraživanje. **Izazovi i problemi u procesu IP** su razni:

- nema gotovih recepata, već imamo veliki broj problema i mogućih rešenja
- veliki broj različitih formata i tipova podataka, kao i složeni i heterogni podaci
- interpretacija i vizuelizacija rezultata
- velika količina ulaznog materijala (algoritmi moraju biti skalabilni)
- veliki broj atributa (dimenzionalnost)

- kvalitet podataka (nedostajući podaci, irelevantni podaci, ...)
- dostupnost ljudima koji nisu stručnjaci
- narušavanje privatnosti
- neautorizovano korišćenje podataka
- profilisanje (pogrešno kvalifikovanje)

Podaci

Podaci (data set) predstavljaju skup **objekata** i njihovih **atributa**. Atributi su svojstva ili karakteristike objekta. Skup atributa opisuje objekat. Vrednosti atributa su brojevi ili simboli koji su pridruženi atributu. Isti atribut može biti preslikan u različite vrednosti atributa (npr. dužina u metrima ili kilometrima). Različiti atributi mogu da budu preslikani u isti skup vrednosti, pri čemu osobine atributa mogu da budu različite (npr. godine i težina su celobrojne vrednosti, ali težina može da se smanjuje, a godine ne). Prema broju vrednosti koje mogu da sadrže, atributi mogu biti:

- **diskretni** - konačan ili prebrojivo beskonačan skup vrednosti. Najčešće se prikazuju kao celobrojne promenljive. **Binarni** atributi su specijalan slučaj diskretnih atributa koji imaju tačno dve vrednosti.
- **kontinuirani (neprekidni, kontinualni)** - skup vrednosti čine realni brojevi. Najčešće se prikazuju kao realni brojevi u pokretnom zarezu.
- **asimetrični** - kod njih je jedino bitno prisustvo ne-nula vrednosti. **Binarni asimetrični** atributi imaju tačno dve vrednosti, od kojih je samo jedna bitna. Na primer, za svakog studenta sa 0 ili 1 beležimo da li je slušao neki kurs i želimo da izmerimo sličnost između studenata po broju kurseva koje zajedno slušaju. Fakultet sadrži veliki broj kurseva i dva studenta će onda biti slični po svim kursevima koje ne sluša ni jedan od njih, pa su nam zbog toga zapravo bitni samo kursevi koje su oba studenta slušala.

Prema operacijama koje se mogu koristiti, atributi mogu biti (svaki sledeći tip podržava operacije prethodnog tipa):

Tip	Operacije	Dodatne operacije	Primer
Imenski	**različitost ($=,\ \neq$)**	**modus, entropija, korelacija, $\mathcal{X}^2$ test**	**pol, poštanski kod, boja očiju**
Redni	**uređenje ($<,\ \leq,\ >,\ \geq$)**	**medijana, percentil, korelacija ranga**	**ocene, redni brojevi nečega**
Intervalni	**aditivnost ($+,\ -$)**	**srednja vrednost, standardna devijacija**	**datumi, temperatura**
Razmerni	**multiplikativnost ($\cdot,\ /$)**	**geometrijska sredina, harmonijska sredina**	**dužina, masa, količina**

Kvalitativni (kategorički) atributi su imenski i redni, a **kvantitativni (neprekidni) atributi** su intervalni i razmerni. Kvalitativni atributi nemaju većinu svojstava brojeva.

Nezavisni (tabelarni) podaci su podaci koji međusobno nisu povezani. Najčešće su to multidimenzionalni ili tekstualni podaci. Za njihovo čuvanje pogodne su baze podataka. Sastoje se od slogova (objekata), a svaki slog se sastoji od polja (atributa). Datoteke se u bazama mogu čuvati tako što se u redovima čuvaju dokumenti, u kolonama reči, a u poljima broj pojavljivanja određene reči u određenom dokumentu. **Skup multidimenzionalnih podataka** je skup od n slogova $\overline{X_1}$, ..., $\overline{X_n}$ takvih da svaki od slogova $\overline{X_i}$ sadrži skup od d osobina označenih sa $(x_i^1, \ldots, x_i^d)$. **Retki podaci** su podaci gde postoji mali broj podataka koji su značajni (ne-nula) podaci. Na primer, to je čest slučaj pri čuvanju dokumenata kroz tabele jer se mnoge reči pojavljuju samo u jednom ili par dokumenata, a u ostalima ne. **Zavisni podaci** su podaci kod kojih postoji **implicitna** ili **eksplicitna zavisnost**. Na primer, implicitna zavisnost se uočava kod senzora koji se ponašaju slično i ako se u nekom trenutku uoči veliko odstupanje ono je od interesa za istraživanje. Eksplicitna zavisnost je npr. prisutna kod grafova poseta veb sajtu, uticaja lekova na druge lekove ili bolesti i slično.

Podaci sa poretkom su podaci gde atributi imaju odnose koji podrazumevaju vremenski ili prostorni redosled. **Vremenske serije** imaju implicitnu zavisnost od prethodnih merenja. To su na primer EKG, temperatura i slično. **Sekvencijalni podaci** sastoje se od skupa koji predstavlja sekvencu objekata, npr. sekvenca slova ili sekvenca reči. Redosled određuje pozicija u sekvenci. To su npr. DNK i RNK sekvence. **Prostorni podaci** određuju prostorne lokacije. Kao kod vremenskih serija, ovde imamo prostornu/fizičku korelaciju. Svi ovi podaci sadrže više atributa koji prikazuju ponašanje, kao i jedan ili više **kontekstualnih podataka** (vreme, lokacija ili pozicija) koji određuju sam kontekst. Postoje i **prostorno-vremenski podaci** gde razlikujemo dva tipa:

- i prostorni i vremenski atributi mogu biti kontekstualni (npr. merenje varijacije temperature mora kroz vreme)
- vremenski atribut je kontekstualan, a prostorni modelira ponašanje (npr. analiza trajektorija)

Grafovski (mrežni) podaci su podaci kod kojih su vrednosti pridružene čvorovima u grafu. Objekte predstavljamo kao čvorove, a njihove međusobne odnose kao grane. Granama možemo pridružiti i usmerenja i težine za dodatni opis odnosa. Koriste se npr. za predstavljanje veb stranica ili hemijskih jedinjenja.

Najznačajniji gradivni blokovi u IP su:

1. **podaci** - najčešće su prethodno prikupljani iz nekog drugog razloga. Te podatke možemo smestiti u matricu sa n redova i d kolona, gde je n broj slogova, a d broj atributa. Potrebno je izvršiti i **preprocesiranje**, tj. pripremu i prečišćavanje podataka. **Otkrivanje anomalija** je određivanje redova u matrici koji su jako različiti od ostatka redova. **Element van granica (outlier)** je podatak koji je u značajnoj meri različit od

ostalih podataka. To su npr. spam poruke, upadi u računarski sistem, zloupotreba kartica i slično.

2. **pravila pridruživanja** - tehnike koje otkrivaju odnose među podacima. Na primer, imamo podatke o transakcijama gde su u vrstama transakcije, a u kolonama stavke u obliku **retke binarne baze** (koristimo 0 i 1 gde 1 znači da je u transakciji i kupljena stavka j). U datoj binarnoj matrici posmatrajmo podskupove kolona A takve da sve vrednosti u tim kolonama u jednoj vrsti imaju vrednost 1. Tada A nazivamo skupom stavki, a sa $\#(A)$ označavamo broj pojavljivanja skupa stavki A. Sa $A \Rightarrow B$ označavamo da je skup stavki B pridružen skupu stavki A, odnosno u našem primeru to bi značilo "ako korisnik kupi stavke A, kupiće i stavke B". U opštem slučaju, elementi matrice ne moraju nužno biti binarne vrednosti. Zadatak je naći pravila pridruživanja koja povezuju atribute pri čemu su nam interesantna pravila koja zadovoljavaju neki nivo određenih svojstava. Svojstva koja opisuju pravila pridruživanja su:
- podrška (support): Podrška pravila pridruživanja $A \Rightarrow B$ je količnik broja transakcija koje sadrže A i B u odnosu na ukupan broj transakcija N. Opisuje koliko često važi posmatrano pravilo pridruživanja.

$$sup(A \Rightarrow B) = \frac{\#(A \cup B)}{N}$$

- pouzdanost (confidence): pouzdanost pravila pridruživanja $A \Rightarrow B$ je količnik broja transakcija koje sadrže A i B u odnosu na broj transakcija koje sadrže A. Opisuje verovatnoću da se desi B ako već važi A.

$$conf(A \Rightarrow B) = \frac{\#(A \cup B)}{\#(A)}$$

Бр. транс.	хлеб	млеко	пелене	пиво	јаја	кола
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Na primer, za $\{mleko,\ pelene\} \Rightarrow \{pivo\}$ važi:

$$sup = \frac{\#(\{mleko,\ pelene,\ pivo\})}{N} = \frac{2}{5},\ conf = \frac{\#(\{mleko,\ pelene,\ pivo\})}{\#(\{mleko,\ pelene\})} = \frac{2}{3}$$

3. **klasifikacija** - određivanje relacija između kolona. Na osnovu atributa objektima dodeljujemo klase. Naziva se i **klasifikacija pod nadzorom**.
4. **klasterovanje** - grupisanje vrsta po sličnosti. Naziva se i **klasifikacija bez nadzora**.

5. **analiza i vizuelizacija rezultata**

Mere sličnosti

Mere sličnosti pružaju način za određivanje sličnosti ili različitosti objekata, atributa i slično. Podaci mogu biti različitog tipa, strukture, raspodele, dimenzionalnosti i tako dalje. **Blizina (proximity)** označava i sličnost i različitost dva objekta. **Sličnost** predstavlja numeričku meru koliko su dva objekta ili atributa slična i najčešće uzima vrednosti iz intervala [0, 1], gde 0 označava da objekti nisu nimalo slični, a 1 da su objekti isti. **Različitost (rastojanje)** predstavlja numeričku meru koliko su dva objekta ili atributa različite i najčešće uzima vrednosti iz intervala [0, $+\infty$), gde 0 označava da su objekti isti, a vrednosti veće od nule opisuju u kojoj meri su oni različiti. Primer nekih funkcija sličnosti i različitosti atributa p i q:

Tip	Sličnost	Različitost
Nominalni	$s = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases}$	$s = \begin{cases} 1, & p \neq q \\ 0, & p = q \end{cases}$
Redni - vrednosti se preslikavaju u skup [0, $n-1$] gde je n broj vrednosti	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$	$d = \frac{\lvert p-q \rvert}{n-1}$
Intervalni ili razmerni	$s = -d$, $s = \frac{1}{1+d}$, $s = 1 - \frac{d-d_{min}}{d_{max}-d_{min}}$	$d = \lvert\ p - q\ \rvert$

Funkcija rastojanja d je **metrika** ako važi:

1. **pozitivna određenost**: $(\forall\ p,\ q)\ d(p,\ q) \geq 0$ i $d(p,\ q) = 0\ \Leftrightarrow\ p = q$
2. **simetrija**: $(\forall\ p,\ q)\ d(p,\ q) = d(q,\ p)$
3. **nejednakost trougla**: $(\forall\ p,\ q,\ r)\ d(p,\ r) \leq d(p,\ q) + d(q,\ r)$

Ako je funkcija d metrika i važi $(\forall\ p,\ q,\ r)\ d(p,\ r) \leq \max\{d(p,\ q),\ d(q,\ r)\}$ onda je d **ultrametrika**.

Kod kvantitativnih podataka računamo rastojanje dve tačke u n-dimenzionalnom prostoru $X = (x_1,\ \ldots,\ x_n)$ i $Y = (y_1,\ \ldots,\ y_n)$. Najčešće korišćena mera je **rastojanje Minkovskog (** $L_p mera$**)**:

$$d(X,\ Y) = (\sum_{i=1}^{n} \mid x_i - y_i \mid^p)^{\frac{1}{p}}$$

Specijalni slučajevi rastojanja Minkovskog:

- **Hamingovo (Menhetn) rastojanje -** za $p = 1$ odnosno

$$d(X,\ Y) = \sum_{i=1}^{n} q_i,\ q_i = \begin{cases} 1,\ x_i \neq y_i \\ 0,\ x_i = y_i \end{cases}$$

- **Euklidskog rastojanje** - za $p = 2$ odnosno

$$d(X,\ Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Supremum rastojanje** - za $p \to \infty$ odnosno:

$$d(X,\ Y) = \max_{1 \leq i \leq n} \mid x_i - y_i \mid$$

Rastojanje Minkovskog nije pogodno za upotrebu kod retkih višedimenzionalnih podataka sa nepoznatom raspodelom, šumovima, kao ni ako postoje lokalno irelevantni atributi zbog šuma koji se akumulira pri izračunavanju. **Mahalanobisovo rastojanje** je:

$$d(X,\ Y) = \sqrt{(X - Y)\Sigma^{-1}(X - Y)^T}$$

gde je Σ matrica kovarijansi podataka. Ovo rastojanje je korisno kada su atributi u korelaciji, imaju različite opsege vrednosti i raspodela podataka je približno normalna. Sa druge strane, računanje samog rastojanja je dosta skupo. Ukoliko želimo nekim atributima da dodelimo veću važnost, koristimo težine. **Rastojanje Minkovskog sa težinama**:

$$d(X,\ Y) = (\sum_{i=1}^{n} a_i \times \mid x_i - y_i \mid^p)^{\frac{1}{p}}$$

Kod binarnih atributa moguće je koristiti posebne mere bliskosti. Označimo sa M_{01} broj atributa koji su 0 u X i 1 u Y, sa M_{10} broj atributa koji su 1 u X i 0 u Y, sa M_{00} broj atributa koji su 0 u X i 0 u Yi sa M_{11} broj atributa koji su 1 u X i 1 u Y. Neke od mera sličnosti su:

- **Jednostavno uparivanje koeficijenata (SMC)**:

$$SMC = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- **Žakardov koeficijent** - kod asimetričnih atributa:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

  Prošireni Žakardovi koeficijenti (koeficijenti Tanimotoa) su varijanta Žakardovih koeficijenata za atribute sa prebrojivim i neprekidnim vrednostima:

$$T(X,\ Y) = \frac{X \circ Y}{||X||^2 + ||Y||^2 - X \circ Y}$$

- **Kosinusna sličnost** - kada je M_{00} veliko, ali može da se koristi i kod nebinarnih vektora:

$$\cos(X, Y) = \frac{X \circ Y}{||X|| \cdot ||Y||}$$

Korelacija dva objekta koji imaju binarne ili neprekidne atributa je mera linearnog odnosa između njihovih atributa. **Pirsonov koeficijent korelacije**:

$$\rho_{X,\,Y} = \frac{cov_{X,\,Y}}{\sigma_X \sigma_Y},\ cov_{X,\,Y} = \frac{\sum_{i=1}^{n}(x_k - \overline{X})(y_k - \overline{Y})}{n-1}$$

Ako korelacija ima vrednost 1 (-1) objekti imaju **perfektno pozitivan (negativan) linearni odnos**, odnosno $X = aY + b$.

Kod kategoričkih podataka sličnost dva objekta možemo definisati kao:

$$d(X,\ Y) = \sum_{i=1}^{n} S(x_i,\ y_i)$$

gde je $S(x_i,\ y_i)$ neka mera sličnosti nad pojedinačnim atributima. Najjednostavnija mera bi bila $S(x_i,\ y_i) = \begin{cases} 1,\ x_i = y_i \\ 0,\ x_i \neq y_i \end{cases}$. Ova mera ne uzima u obzir relativnu frekvenciju atributa.
Označimo sa $p_k(x)$ broj slogova u kojima k-ti atribut uzima vrednost x. Mere koje uključuju učestalost su:

- **Inverzna učestalost pojavljivanja**

$$S(x_i,\ y_i) = \begin{cases} \frac{1}{p_k(x_i)^2},\ x_i = y_i \\ 0,\ x_i \neq y_i \end{cases}$$

- **Pojavljivanje je dobro**

$$S(x_i,\ y_i) = \begin{cases} 1 - p_k(x_i)^2,\ x_i = y_i \\ 0,\ x_i \neq y_i \end{cases}$$

Sličnost dva dokumenta najbolje se ocenjuje ako se koriste reči koje su zajedničke. Za normalizaciju uparivanja reči u slučajevima kada ima reči koje se retko javljaju i koje se često javljaju koristi se funkcija $id_i = \log \frac{n}{n_i}$ gde je n_i broj dokumenata u kojima se javlja reč i, a n ukupan broj dokumenata. Za smanjenje mogućnosti da pojava neke česte reči utiče na sličnost dokumenata mogu da se koriste i funkcije $f(x_i) = \sqrt{x_i}$ i $f(x_i) = \log x_i$.
Normalizovana frekvencija za i-tu reč se zatim definiše kao $h(x_i) = f(x_i) id_i$. Nad ovim frekvencijama je moguće koristiti kosinusno ili Žakardovo rastojanje kao i u slučaju binarnih podataka:

$$\cos\left(X,\ Y\right) = \frac{\sum_{i=1}^{n} h(x_i) \times h(y_i)}{\sqrt{\sum_{i=1}^{n} h(x_i)^2} \times \sqrt{\sum_{i=1}^{n} h(y_i)^2}}$$

$$J(X,\ Y) = \frac{\sum_{i=1}^{n} h(x_i) \times h(y_i)}{\sum_{i=1}^{n} h(x_i)^2 + \sum_{i=1}^{n} h(y_i)^2 - \sum_{i=1}^{n} h(x_i) \times h(y_i)}$$

Sličnost dva sloga $X = (X_n, X_c)$ i $Y = (Y_n, Y_c)$ sa **mešanim atributima** je:

$$S(X,\ Y) = \lambda S_{Num}(X_n, Y_n) + (1 - \lambda) S_{Cat}(X_c,\ Y_c)$$

gde λ određuje relativnu važnost kategoričkih i numeričkih atributa, a S_{Num} i S_{Cat} su neke mere sličnost za numeričke i kategoričke atribute.

Mere kod diskretnih podataka:

- **Edit rastojanje** - različitost dve niske po tome koliko je potrebno za transformaciju niske $X = (x_1,\ \ldots,\ x_m)$ u nisku $Y = (y_1,\ \ldots y_n)$. Za prvih i simbola iz X i prvih j simbola Y cena transformacije je:

  $$Edit(i,\ j) = \min \begin{cases} Edit(i-1,\ j) + C_b \\ Edit(i,\ j-1) + C_u \\ Edit(i-1,\ j-1) + I_{ij} C_z \end{cases}$$

  gde su $C_b,\ C_u,\ C_z$ redom cena brisanja, umetanja i zamene, a I_{ij} indikator jednakosti i-tog simbola X i j-tog simbola Y.
- **Najduža zajednička podniska (LCSS)** - sličnost na osnovu najduže zajedničke podniske. Za prvih i simbola iz X i prvih j simbola Y najduža zajednička podniska je:

  $$LCSS(i,\ j) = \max \begin{cases} LCSS(i-1,\ j-1) + 1,\ x_i = y_i \\ LCSS(i-1,\ j),\ x_i\ nije\ upareno \\ LCSS(i,\ j-1),\ y_i\ nije\ upareno \end{cases}$$

Neke mere su zasnovane na teoriji informacija. **Entropija** događaja X sa n mogućih ishoda sa verovatnoćama ishoda $p_1,\ \ldots,\ p_n$ je:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Pripada intervalu $[0,\ \log_2 n]$ i meri nepredvidljivost nekog događaja. Mala entropija znači veća sigurnost, a velika entropija znači veća nepredvidljivost. **Zajedničke informacije** za događaje X i Y definišemo sa:

$$I(X,\ Y) = H(X) + H(Y) - H(X,\ Y)$$

Mere zasnovane na gustini mere stepen blisosti objekata u nekoj oblasti. Najčešće se koriste:

- **Euklidska gustina** - broj tačaka po jedinici površine ili zapremine.
- **Gustina verovatnoće** - procena distribucije podataka na osnovu izgleda
- **Graf zasnovane gustine** - povezanost

Priprema podataka

Priprema (preprocesiranje) podataka je bitno jer kvalitet podataka direktno utiče na rezultate. Izvorni podaci mogu biti u različitim formatima, mogu postojati nedostajući i nekonzistentni podaci i slično.

Izdvajanje karakteristika je tehnika prepocesiranja koja iz ravnih podataka izdvaja karakteristike. Na primer, izdvajanje nekih piksela koji su značajni na slikama kako bismo razlikovali pse i mačke.

Prenosivost tipova podataka je tehnika preprocesiranja koja predstavlja transformaciju podataka iz jednog tipa u drugi. Potrebna je jer neke karakteristike onemogućuju primenu gotovih alata, a pojedini algoritmi rade samo sa određenim tipovima podataka. U ovom procesu moguć je gubitak informacija. **Diskretizacija** je transformacija neprekidnih u kategoričke atribute. Obično se primenjuje na atribute u klasifikaciji ili pravilima pridruživanja. Prvo se bira broj kategorija n. Zatim se interval brojeva deli na n podintervala i sve vrednosti iz jednog podintervala preslikavaju se u istu kategoričku vrednost. Između dobijenih vrednosti ne postoji uređenje. Intervali se mogu izabrati na više načina:

- **jednake širine intervala** - ako vrednosti atributa upadaju u interval $[m,\ M]$ on se deli na n jednakih delova. Ovakva podela je nekorektna ako je distribucija elemenata po intervalima neravnomerna.
- **jednaki** $\log$**-intervali** - ako su podintervali oblika $[a,\ b]$ onda je vrednost $\log b - \log a$ jednaka za svaki interval. Ovakva podela je nekorektna ako je distribucija elemenata po intervalima neravnomerna. Ako distribucija elemenata može da se modelira funkcijom f mogu se birati intervali $[a,\ b]$ tako da je vrednost $f(b) - f(a)$ jednaka za svaki interval.
- **jednak broj elemenata u intervalu** - vrednosti atributa se prebroje i dobijeni broj k se podeli sa n. Vrednosti atributa se sortiraju i u svaki interval se smešta $\frac{k}{n}$ elemenata. Čvor Binning u SPSS Modeleru radi na ovaj način.

Binarizacija je transformacija neprekidnih i diskretnih atributa u binarne. Obično se primenjuje na atribute u analizi zasnovan na pravilima pridruživanja. Ako kategorički atribut ima n vrednosti formira sa n binarnih atributa tako da svaki od njih odgovara jednoj vrednosti kategoričke promenljive.

Predstavljanje tekstualnih podataka preko retkih numeričkih vektora nije pogodno za najveći broj metoda IP, a ograničen je i broj mera koje se mogu koristiti. **Latentna semantička analiza (LSA)** prevodi tekst u neretku reprezentaciju manje dimenzije. Dokument $X = (x_1,\ \ldots,\ x_d)$ se posle transformacije skalira funkcijom:

$$\frac{X}{\sqrt{\sum_{i=1}^{d} x_i^2}}$$

Za ovako dobijene podatke moguće je primeniti Euklidsko rastojanje.

Podaci se iz vremenskih serija transformišu u diskretne niske **SAX algoritmom (simbolička aproksimacija agregata)**. U prvom koraku se serija deli u prozore veličine w za koje se

računa prosečna vrednost atributa. U drugom koraku se srednje vrednosti vremenskih serija diskretizuju pomoću tehnike sa intervalima koji imaju isti broj elemenata. Pretpostavka je da vrednosti u vremenskim serijama imaju normalnu raspodelu. Na osnovu toga računaju se srednja vrednost i standardna devijacija vrednosti vremenskih serija iz prozora, a za određivanje granica intervala koriste se kvantili normalne raspodele. Diskretizacija se najčešće vrši u 3 do 10 intervala.

Podaci se iz vremenskih serija transformišu u numeričke podatke putem **diskretne transformacije talasićima (DWT)**, a može da se koristi i **diskretna Furijeova transformacija (DFT)**. Na ovaj način se omogućava upotreba algoritama koji rade sa multidimenzionalnim podacima. Osobina ovih metoda je da dobijeni koeficijenti nisu zavisni kao u originalnim podacima.

Diskretne niske mogu se transformisati u numeričke podatke u dva koraka:

- Diskretne niske se konvertuju u skup binarnih vremenskih serija čiji je broj jednak broju različitih simbola.
- Svaka serija se konvertuje u multidimenzionalni vektor pomoću transformacije talasićima. Osobine iz ovih vektora se kombinuju i formira se multidimenzionalni slog.

Čišćenje podataka je tehnika preprocesiranja koja obuhvata:

- rad sa **nedostajućim podacima**: Podaci mogu da nedostaju jer informacije nisu prikupljene ili neki atributi nisu primenljivi u svim slučajevima. Moguće je obraditi slog sa nedostajućim podacima na više načina.
  - Ceo slog se odbacuje.
  - Nedostajuća vrednost se procenjuje i unosi - **imputacija**.
  - Neki algoritmi mogu da obrađuju slogove sa nedostajućim vrednostima, pa ih nije potrebno menjati ili izbacivati.
- rad sa **nekorektnim podacima**: Neki podaci mogu biti nekonzistentni.
- rad sa **dupliranim podacima**: Najčešće se javljaju kod spajanja podataka iz više izvora. Ovakvi podaci se najčešće eliminišu, ali ne uvek.
- **skaliranje**: Transformacija promenljive označava transformaciju koja se primenjuje na sve vrednosti te promenljive. U statistici se često koriste funkcije $\sqrt{x}$, $\log x$ i $\frac{1}{x}$ radi transformacije podataka koji nemaju normalnu raspodelu u podatke koji je imaju. U IP postoje i drugi razlozi, npr. primena $\log$ funkcije na vrednosti iz opsega [1, 1000000000] kako bi se dobili bolji odnosi kod poređenja. Primenom nekih transformacija moguće je promeniti prirodu podataka, npr. sa $\frac{1}{x}$. **Standardizacija** je skaliranje tako da srednja vrednost bude 1, a standardna devijacija 0. Ako atribut a ima srednju vrednost μ_a i standardnu devijaciju σ_a, onda se njegove vrednosti normalizuju izrazom $\frac{x-\mu_a}{\sigma_a}$. Za normalnu raspodelu dobijene vrednosti najčešće se nalaze u intervalu [-3, 3]. **Normalizacija** je skaliranje vrednosti u određeni opseg. Za svođenje na interval [0, 1] primenjuje se **min-maks skaliranje**: $y = \frac{x-min}{max-min}$.

Redukcija podataka je tehnika preprocesiranja koja za cilj ima smanjenje količine podataka, što omogućava efikasniju primenu algoritama. Postoje različite tehnike redukcije:

1. **Agregacija** je kombinovanje dva ili više atributa (ili objekata) u jedan. Na ovaj način smanjuje se broj atributa ili objekata, menja se skala i dobijaju stabilniji podaci. Nedostatak agregacije je mogući gubitak nekih informacija.
2. **Uzimanje uzorka** koristi se jer obrada kompletnog skupa podataka koji je od interesa može biti skup ili vremenski zahtevan. Uzorak je **reprezentativan** ako ima aproksimativno iste osobine kao i originalni skup podataka. Korišćenjem uzoraka koji su reprezentativni dobija se efekat skoro isti kao da je rađeno na kompletnom skupu podataka. Veličina uzorka treba da bude dovoljno velika da se ne naruši struktura objekta ili uklone interesantne osobine. **Prost (jednostavan) slučajan uzorak** je uzorak gde svaki element ima jednaku verovatnoću da bude izabran. Uzorkovanje može biti **sa vraćanjem** ili **bez vraćanja**. **Pristrasno uzorkovanje** podrazumeva da su neki podaci važniji od drugih, odnosno imaju veću verovatnoću da budu izabrani. **Stratifikovano uzorkovanje** podrazumeva da se podaci dele u više delova, a zatim se bira slučajni uzorak iz svakog od tih delova.
3. **Izbor karakteristika** je jedan od načina za smanjenje dimenzionalnosti. Podrazumeva eliminaciju redudantnih i irelevantnih karakteristika. Često se formiraju i novi atributi koji uključuju važne karakteristike zbog efikasnije obrade.
4. **Redukcija pomoću rotacije osa** predstavlja automatsko uklanjanje koordinatnih osa pomoću rotacije. Koriste se algoritmi **PCA (Principal Component Analysis)** i **SVD (Singular Value Decomposition)**. PCA je tehnika linearne algebre koja pronalazi nove atribute koji su linearne kombinacije originalnih atributa, koji su ortogonalni jedni na druge i obuhvataju maksimalno raznovrsne podatke. Korisna je jer smanjuje broj dimenzija, nalazi obrasce u podacima velike dimenzionalnosti i omogućava vizuelizaciju podataka velike dimenzionalnosti. Osnovna ideja je da se podaci rotiraju u sistem sa osama gde je najveći broj varijansi pokriven najmanjim brojem dimenzija. Novi sistem sa osama zavisi od korelacije između atributa. Najčešće se primenjuje posle oduzimanja srednje vrednosti od svake tačke. Za matricu podataka D reda $m \times n$ može da se formira matrica kovarijansi C gde je c_{ij} kovarijansa i-te i j-te kolone. Kovarijansa je mera kako se atributi menjaju u paru, a ako je $i = j$ tada je kovarijansa jednaka varijansi atributa. Cilj PCA je nalaženje transformacije podataka za koju važi:
   - Svaki par novodobijenih atributa ima kovarijansu 0.
   - Atributi su uređeni u opadajućem redosledu u odnosu na veličinu varijanse koja je pokrivena od strane atributa.
   - Između atributa postoji ortogonalnost, tako da svaki naredni atribut pokriva što je moguće veći broj preostalih varijansi.
     Transformacija se vrši upotrebom sopstvenih vrednosti matrice kovarijanci C. Neka važi:
   - λ_i su nenegativne sopstvene vrednosti C uređene u redosledu $\lambda_1 \geq \ \ldots \ \geq \lambda_m$.
   - $U = [u_1, \ \ldots, \ u_n]$ je matrica sopstvenih vektora C uređena tako da i-ti vektor odgovara i-toj najvećoj sopstvenoj vrenosti.

- Matrica D je prethodno pripremljena tako da je srednja vrednost svakog od atributa jednaka 0. U tom slučaju važi $C = D^T D$.
  Tada važi:
- Matrica $D' = DU$ je tražena transformacija matrice podataka.
- Novi atribut je linearna kombinacija starih atributa, a težina linearne kombinacije i-tog atributa su komponente i-tog sopstvenog vektora.
- Varijansa novog i-tog atributa je λ_i. Zbir varijansi originalnih jednak je zbiru varijansi novih atributa.
- Novi atributi nazivaju se **glavne komponente**. Prvi novi atribut je prva glavna komponenta, i tako dalje.