Istraživanje podataka 1

Uvod

Istraživanje podataka je proces automatskog otkrivanja korisnih informacija u velikom skladištu podataka. **CRISP-DM (Cross-Industry Standard Process for Data Mining) metodologija**:

1. **Razumevanje posla** - utvrđivanje poslovnih ciljeva i ciljeva istraživanja podataka.
2. **Razumevanje podataka** - prikupljanje, opisivanje, upoznavanje i provera kvaliteta podataka.
3. **Priprema podataka** - odabir, čišćenje, konstruisanje, formatiranje podataka.
4. **Modeliranje** - odabir tehnika, izgradnja i procena modela.
5. **Evaluacija** - procena rezultata.
6. **Razvoj** - integracija dobijenih znanja u svakodnevne poslovne procese kako bi se rešio originalni poslovni problem.

Skup podataka predstavlja kolekciju objekata kao što su slogovi, uzorci, entiteti i slično. **Atributi** su svojstva ili karakteristike objekata, a vrednosti atributa su brojevi ili simboli koji su pridruženi atributu. Prema operacijama koje se mogu koristiti, atributi mogu biti:

1. **kvalitativni**
   - **imenski (nominal)** - operacije poređenja ($=$, $\noteq$)
   - **redni (ordinal)** - operacije uređenja ($<$, $\leq$, $>$, $\geq$)
2. **kvantitativni**
   - **intervalni (interval)** - operacije aditivnosti ($+$, $-$)
   - **razmerni (ratio)** - operacije multiplikativnosti ($\cdot$, $/$)

Prema vrednosti koje sadrže, atributi mogu biti:

1. **diskretni** - konačan ili prebrojivo beskonačan skup vrednosti. **Binarni atributi** su specijalan slučaj diskretnih, gde su moguće samo dve vrednosti.
2. **kontinualni (neprekidni)** - neprebrojivo beskonačan skup vrednosti, tj. skup vrednosti čine realni brojevi.

Asimetrični (retki) podaci su podaci gde se jedino prisustvo ne-nula vrednosti smatra značajnim.

Skupovi podataka mogu biti:

1. **slogovi** - matrica podataka, podaci u dokumentima, transakcioni podaci, ...

2. **grafovi**
3. **podaci sa poretkom** - prostorni, vremenski i redosledni podaci
4. **slike, video i audio zapisi, ...*

Šum predstavlja modifikaciju originalnih vrednosti. **Elementi van granica (outliers)** su objekti sa karakteristikama koju su značajno različite od najvećeg broja objekata u skupu podataka. **Ekstremne vrednosti** su objekti sa karakteristikama koje se veoma razlikuju od najvećeg broja objekata u skupu podataka, odnosno vrednosti koje su gotovo nemoguće u realnom slučaju.

SPSS (Statistical Package for the Social Sciences) koristi se za učitavanje i manipulaciju podacima, kao i izvoz rezultata. Operacije koje se mogu primeniti nad podacima su predstavljene kao čvorovi, a niz povezanih operacija (čvorova) naziva se **tok podataka (data stream)**. Vezama između čvorova određuje se pravac toka podataka.

Za učitavanje podataka iz baze podataka koristi se čvor **Database**. Opcije:

- **Data** - učitavanje podataka iz tabele ili rezultata upita.
- **Filter** - odabir atributa.
- **Types** - informacije o atributima:
  1. **Measurement levels** - tip upotrebe atributa:
     - **Default** - nepoznat
     - **Continuous** - neprekidan
     - **Categorical** - kategorički, nakon čitanja vrednosti prelazi u Flag, Nominal ili Typeless
     - **Flag** - binarni
     - **Nominal** - imenski
     - **Ordinal** - redni
     - **Typeless** - atributi koji imaju jednu vrednost ili imenski atributi sa više vrednosti od dozvoljenog
  2. **Values** - interval ili moguće vrednosti atributa:
     - **Read** - informacije se učitavaju pri izvršavanju čvora.
     - **Read+** - informacije se učitavaju i dodaju već definisanim ako postoje.
     - **Pass** - ne učitavaju se informacije.
     - **Current** - ostaju već definisane vrednosti.
     - **Specify** - posebno definisanje vrednosti.
  3. **Missing** - definisanje načina obrade nedostajućih vrednosti.
  4. **Check** - definisanje akcije za objekte koji imaju vrednost koja ne pripada definisanom intervalu ili listi vrednosti u Values.
     - **None** - ne menja se vrednost.
     - **Nullify** - postavlja se na null.

- **Coerce** - vrednost se prebacuje u legalnu. Za flag u false, za nominal i ordinal u prvu vrednost iz skupa, za neprekidne u gornju/donju granicu ako je vrednost veća/manja od gornje/donje granice ili u srednju vrednost ako je null.
- **Discard** - ceo slog se odbacuje.
- **Warn** - prijavljuje se broj slogova sa nepravilnim vrednostima.
- **Abort** - prijavljuje se greška.

Data Audit je čvor za upoznavanje sa podacima. Prikazuje sumarne statistike za atribute i grafike sa distribucijom vrednosti po atributima. Prikazuje i izveštaj o nedostajućim vrednostima, elementima van granica, ekstremnim vrednostima i omogućava definisanje akcija za obradu tih vrednosti. Za uzorak $x_1, x_2, \ldots, x_n$ moguće su sledeće statistike:

- **srednja vrednost (mean)**: $\mu = frac\sum_i x_i n$
- **varijansa (variance)**: $\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n-1}$
- **standardna devijacija (standard deviation)**: σ
- **iskrivljenost (skewnes)** - mera asimetrije distribucije. Normalna raspodela je simetrična i ima vrednost asimetrije nula. Raspodela sa pozitivnom asimetrijom ima dugi desni rep, a raspodela sa negativnom asimetrijom ima dugačak levi rep.
- **mod (mode)** - vrednost koja se najčešće pojavljuje u skupu podataka.
- **medijana (median)** - vrednost koja deli slučajeve na pola nakon sortiranja.
- **percentil** - za izabrani broj p važi da je barem p% vrednosti u skupu manje ili jednako toj vrednosti. **Prvi, drugi i treći kvartil** su redom 25, 50. i 75. percentil. **Interkvartilni raspon** je razlika trećeg i prvog kvartila.

Nedostajuće vrednosti su Null ili sistemski nedostajuće vrednosti. Označene su sa $null$. Prazne niske i beline su regularne vrednosti i mogu se definisati kao **blanko vrednosti**. Moguće je i prisustvo korisnički definisanih nedostajućih vrednosti u nekim čvorovima. Metode obrade nedostajućih vrednosti mogu biti:

- **Fixed** - zamena vrednošću koja je zadata konstanta ili rezultat izabrane statistike.
- **Random** - zamena izborom slučajne vrednosti.
- **Expression** - zamena rezultatom izraza.
- **Algorithm** - zamena korišćenjem vrednosti predviđene modelom dobijenog algoritmom C&RT.

Mere bliskosti

Sličnost predstavlja numeričku meru koliko su dva objekta slična i najčešće se meri vrednostima u intervalu [0, 1]. **Različitost (rastojanje)** je numerička mera koliko su dva

objekta različita. Donja granica je često 0, a gornja granica varira. **Blizina (proximity)** označava ili sličnost ili različitost. Sličnost i različitost kod različitih atributa:

- imenski: imamo samo operacije = i $\noteq$ pa je različitost 0 ako su objekti isti, a sličnost 1 i obrnuto.
- redni: iako nemamo operacije sabiranja i oduzimanja možemo objektima dodeliti brojeve od 0 do $n-1$, gde je n broj objekata, poređane po njihovom uređenju. Različitost sada dobijamo kao $d = \frac{|p-q|}{n-1}$, a sličnost kao $s = 1 - d$.
- kvantitativni: moguće je koristiti više vrsta mera različitosti. **Rastojanje Minkovskog** je

$$(\sum_{k=1}^{n} |p_k - q_k|)^{\frac{1}{r}},$$

  gde je r parametar, n broj atributa, a p_k i q_k su vrednosti k-tog atributa objekata p i q. Za $r = 1$ dobijamo **Menhetn rastojanje**, za $r = w$ **Euklidsko rastojanje**, a za $r \to \infty$ **supremum rastojanje** koje predstavlja maksimum razlike između odgovarajućih komponenti vektora. Za ocenu bliskosti ponovo je moguće koristiti nekoliko različitih mera, npr. $s = -d$ ili $s = \frac{1}{1+d}$, gde je d odgovarajuće rastojanje (različitost).

Normalizacija predstavlja svođenje opsega vrednosti sa $[a,\ b]$ na opseg $[0,\ 1]$. Na ovaj način se svi atributi svode na isti segment i nijedan neće biti bitniji od drugih (npr. ako merimo dužinu u metrima, a težinu u gramima, težina će imati mnogo veći uticaj iako su oba atributa bitna za nas). Normalizaciju možemo izvesti prema sledećoj formuli:

$$y = \frac{x - a}{b - a},$$

gde je x stara vrednost atributa, a y normalizovana vrednost. **Standardizacija** predstavlja transformaciju podataka tako da oni imaju očekivanu vrednost 1 i standardnu devijaciju 0. Standardizacija, kao i normalizacija, uklanja razlike u skalama atributa i poboljšava performanse modela. Vrši se po formuli:

$$Y = \frac{X - \mu}{\sigma}$$

Kod binarnih atributa moguće je koristiti posebne mere bliskosti. Označimo sa M_{01} broj atributa koji su 0 u p i 1 u q, sa M_{10} broj atributa koji su 1 u p i 0 u q, sa M_{00} broj atributa koji su 0 u p i 0 u qi sa M_{11} broj atributa koji su 1 u p i 1 u q. Za meru različitosti možemo koristiti **Hamingovo rastojanje**:

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

Mere sličnosti:

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

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

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

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

Još neke mere sličnosti kod različitih vrsta atributa:

- **Kosinusna sličnost** (korisno kod asimetričnih atributa, najčešća mera sličnosti dokumenata):

$$\cos(p, q) = \frac{p \circ q}{||p|| \cdot ||q||}$$

- **Korelacija**:

$$r = \frac{cov(p,\ q)}{\sigma_p \cdot \sigma_q},\ cov(p,\ q) = \frac{\sum_{k=1}^{n}(p_k - \overline{p})(q_k - \overline{q})}{n-1},$$

  pri čemu je $cov(p,\ q)$ kovarijansa vektora p i q, a σ_p i σ_q redom njihove standardne devijacije.

Primer: **Dokument-term matrica** je matrica tf u kojoj je tf_{ij} frekvencija i-te reči (terma) u j-tom dokumentu. Neka je m broj dokumenata. **Inverzna dokument frekvencija** je transformacija:

$$tf'_{ij} = tf_{ij} \cdot \log \frac{m}{df_i},$$

gde je df_i broj dokumenata u kojima se term i pojavljuje. Cilj ove transformacije je razlikovanje dokumenata po rečima koje se retko pojavljuju. Ako se neka reč pojavljuje u svakom dokumentu ona će nakon transformacije imati težinu 0, a ako se neka reč pojavljuje samo u jednom dokumentu imaće težinu $\log m$.

Bliskost možemo definisati i između dva skupa objekata. Na primer, rastojanje između dva skupa tačaka u Euklidskom prostoru možemo dobiti računanjem rastojanja između **centroida** tih skupova, a bliskost između dva skupa nekih objekata možemo dobiti kao najmanju, najveću ili prosečnu vrednost bliskost parova iz tih skupova.

Diskretizacija je transformacija neprekidnog atributa u kategorički atribut. Na primer, visinu pretvaramo iz neprekidnog opsega u kategorije niski, srednji i visoki. **Binarizacija** je transformacija atributa u jedan ili više binarnih atributa.

SPSS:

- **Set Globals** - računa statistike atributa koje se mogu kasnije koristiti globalno u drugim čvorovima.
- **Binning** - diskretizacija.
- **Reclassify** - spajanje nekoliko različitih vrednosti atributa u jednu novu ili već postojeću vrednost.

Klasifikacija - stabla odlučivanja

Ulazni podaci su slogovi oblika $(x,\ y)$, gde je x skup ulaznih atributa, a y je ciljni atribut, odnosno klasa. Cilj **klasifikacije** je pronaći funkciju f (**model klasifikacije**) koja preslikava skup atributa x u jednu od predefinisanih oznaka klasa y. Podaci se dele na **trening** i **test skup**. **Stratifikovana podela** podrazumeva da se podaci uzimaju redom jedan po jedan za date skupove, a ne da je prvih k u trening skupu, a preostali u test skupu. Neke od mera za ocenu modela su **preciznost (accuracy)** koja predstavlja udeo slogova čija klasa je dobro predviđena modelom, kao i **stopa greške (error rate)** tj. udeo slogova čija klasa nije dobro predviđena modelom.

Model klasifikacije može se predstaviti kao **stablo odlučivanja** koje ima:

- **unutrašnje čvorove** - sadrže uslov nad test atributom koji služi za podelu slogova koji imaju različite karakteristike tako da se dobiju čistije grupe slogova. Grane koje izlaze iz unutrašnjeg čvora odgovaraju mogućim vrednostim test atributa.
- **listove** - svakom listu dodeljena je jedna klasa.

Slog se klasifikuje tako što počevši od korena drveta odlučivanja primenjujemo test uslov nad slogom i pratimo granu koja odgovara dobijenom rezultatu. Ukoliko se pri spuštanju niz drvo naiđe na unutrašnji čvor, postupak se ponavlja, a ako se naiđe na list slogu se dodeljuje klasa koja je pridružena tom listu.

Označimo sa $p(j|t)$ relativnu frekvenciju klase j u čvoru t. Najčešće korišćene **mere nečistoće** su:

- **Ginijev indeks**

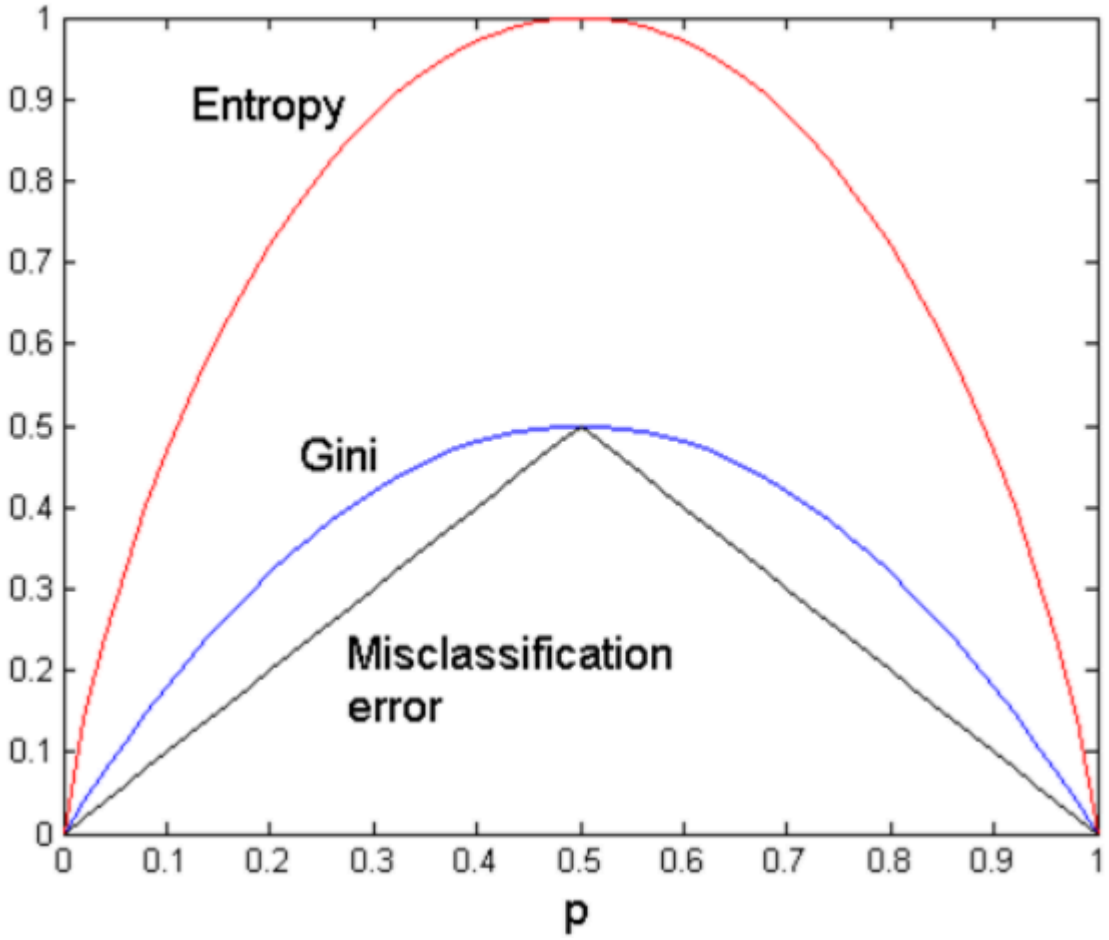
$$1 - \sum_j p(j|t)^2$$

- **Entropija**

$$-\sum_j p(j|t) \log_2 p(j|t)$$

- **Greška klasifikacije**

$$1 - \max_j p(j|t)$$



Ako je vrednost neke mere 0 to znači da je taj čvor **čist**. Ako je vrednost 0.5 (1 kod entropije) to znači da je čvor potpuno nečist, odnosno imamo šanse 50:50 da pogodimo pravu klasu.

Informaciona dobit je

$$I(p) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j),$$

gde je I neka od prethodnih mera, $I(p)$ mera posmatranog čvora, $I(v_j)$ mera j-tog deteta posmatranog čvora, N ukupan broj slogova, $N(v_j)$ broj slogova koji se nalaze u j-tom detetu posmatranog čvora. Kada biramo po kom atributu pravimo sledeću podelu čvora, želimo da dobit bude što veća, odnosno minimizujemo izraz $\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$.

SPSS:

- **C5.0** - klasifikacija sa korišćenjem informacione dobiti i entropije. Za numeričke atribute koristi binarnu podelu, a za kategoričke po jednu granu za svaku moguću vrednost (moguće je i grupisanje).
- **Partition** - podela podataka na trening i test skup. Opcije:
  - **Boosting** - m puta pravi model, gde svaki naredni pokušava da ispravlja greške prethodnog.
  - **Cross-validation** - **unakrsna validacija**, tj. podaci se dele na k grupa i model se k puta trenira pri čemu svaki put sledeća grupa predstavlja test skup, a preostali

podaci trening skup.

- **Analysis** - ispisuje informacije o modelu. Moguće je uključiti i **matrice konfuzije** koje prikazuju koliko instanci koje klase je tačno klasifikovano, a koliko nije i u koju klasu su pogrešno raspoređene.