Istraživanje podataka

vežbe 1, 2, 3

Istraživanje podataka

- je proces automatskog otkrivanja korisnih informacija u velikom skladištu podataka
- Skup podataka – kolekcija objekata podataka (slogova, uzoraka, entiteta...)
- Atributi – svojstvo ili karakteristike objekata
- Vrednosti atributa - brojevi ili simboli koji su pridruženi atributu

Podela atributa prema osobinama i operacijama koje mogu da se primene

Za podelu se koriste operacije:

- Različitost: = i ≠
- Uređenje: <, ≤, > i ≥
- Aditivnost: + i –
- Multiplikativnost: * i /

Tip atributa	Opis	Primeri
Imenski (eng. Nominal)	Vrednost imenskog atributa su upravo različita imena, tj. imenski atributi pružaju samo mogućnost razlikovanja jednog od drugog objekta (=, ≠)	boja očiju, pol (muški, ženski)
Redni (eng. Ordinal)	Vrednosti rednih atributa pružaju dovoljno informacija za uređenje objekata (<, >)	redni brojevi zgrada u ulici
Intervalni (eng. Interval)	Za intervalne atribute, ima smisla razlika između vrednosti, tj. postoji jedinica mere takvih atributa (+, -)	datumi u kalendaru
Razmerni (eng. Ratio)	Kod razmernih atributa ima smisla i proizvod i količnik (*, /) tih atributa	količina novca, godine

Podela atributa prema broju vrednosti koji sadrže

- Diskretni atributi
  - Imaju konačan ili prebrojivo beskonačan skup vrednosti
  - Binarni atributi su specijalan slučaj diskretnih atriubuta
- Kontinuirani (neprekidni) atribututi
  - Skup vrednosti ovih atributa čine realni brojevi

Asimetrični podaci

- Jedino se prisustvo ne-nula vrednosti smatra značajnim
- Binarni atributi kod kojih su bitne ne-nula vrednosti se zovu asimetrični binarni atributi

Podaci

1. Za sledeće atribute odrediti da li su binarni, diskretni ili neprekidni. Takođe odrediti da li su kvalitativni (imenski ili redni) ili kvantitativni (intervalni ili razmerni).
   - starost u godinama
     diskretan, kvantitativni, razmerni
   - Vreme u oznakama AM ili PM
     binaran, kvalitativni, redni

Podaci

- osvetljenost merena aparatom

  neprekidan, kvantitativan, razmerni
- osvetljenost merena ljudskom procenom

  diskretan, kvalitativan, redni
- uglovi mereni u stepenima

  neprekidan, kvantitativan, razmerni
- Bronzane, srebrne i zlatne medalje osvojene na Olimpijadi

  diskretan, kvalitativan, redni

Podaci

– broj pacijenata u bolnici

  diskretan, kvantitativan, razmeran

– ISBN brojevi knjiga

  diskretan, kvalitativan, imenski

– sposobnost da se prenese svetlost opisana vrednostima: neproziran, delimično providan (prozračan), transparentan

  diskretan, kvalitativan, redni

Podaci

– rang u vojsci

  diskretan, kvalitativan, redni

– rastojanje od centra kampusa

  neprekidan, kvantitativan, razmerni

– broj u garderobi

  diskretan, kvalitativan, imenski

Tipovi skupova podataka

- Slogovi
  - Matrica podataka
    - skup numeričkih atributa
  - Podaci u dokumentima
    - atributi istog tipa, asimetrični
  - Transakcioni podaci
    - transakcija(objekat) – skup stavki
- Grafovi
- Podaci sa poretkom (eng. Ordered)
  - Prostorni podaci
  - Vremenski (zavisni) podaci
  - Redosledni podaci

Podaci

2. Koja veličina ima veću prostornu autokorelaciju: dnevna količina padavina ili dnevna temperatura?

   dnevna temperatura

Podaci

3. Zašto je matrica terma u dokumentima primer skupa podataka koji ima asimetrične diskretne ili asimetrične neprekidne osobine (atribute)?

   U i-tom redu i j-toj koloni matrice čuva se broj poljavljivanja j-tog terma u i-tom dokumentu. Kako većina dokumenata sadrži mali deo svih mogućih reči, 0 vrednosti, koje nemaju značaja u opisu i poređenju dokumenata, će se pojavljivati u velikom broju. Zato matrica ima asimetrične diskretne osobine. Ako se upotrebi normalizacija nad termima i dokumentima, onda matrica ima asimetrične neprekidne atribute.

Šum i elementi van granice

- Šum predstavlja modifikaciju originalnih vrednosti
- Elementi van granica su objekti sa karakteristikama koje su značajno različite od najvećeg broja objekata u skupu podataka

Podaci

4. Napraviti razliku između šuma i elemenata van granica.
   – Da li je šum interesantan ili poželjan? Elementi van granica?

     Šum – nije, elementi van granica – jesu

   – Da li objekti koji spadaju u šum mogu biti elementi van granica?

     Da

Podaci

- Da li su objekti koji spadaju u šum uvek elementi van granice?

  Ne

- Da li su elementi van granice uvek objekti koji spadaju u šum?

  Ne

- Da li šum može da pretvori očekivanu vrednost u neobičnu i obrnuto?

  Da

Bliskost-sličnost i različitost

- Sličnost
  - Numerička mera koliko su dva objekta slična
  - Što dva objekta više liče jedan na drugi sličnost im je veća
  - Često se meri vrednostima u intervalu [0,1]
- Različitost
  - Numerička mera koliko su dva objekta različita
  - Što dva objekta više liče jedan na drugi različitost im je manja
  - Najmanja različitost je često 0; gornja granica varira
  - Kao sinonim koristi se i termin rastojanje
- Blizina (eng. proximity) označava ili sličnost ili različitost

Sličnost i različitost za jednostavne atribute

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Različitost između objekata podataka

- Rastojanje Minkovskog

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

gde su r parametar, n broj dimenzija (atributa), a p_k i q_k su vrednosti k-tih atributa objekata p i q

Rastojanje Minkovskog

- r = 1 Menhetn (L1 norma) rastojanje
  - Hamingovo rastojanje
- r = 2 Euklidsko rastojanje
- $r \rightarrow \infty$. “supremum” (Lmax norma) rastojanje
  - Predstavlja maksimum razlike između odgovarajućih komponenti vektora.

Mera sličnosti za binarne podatke

- p i q - binarni vektori
  - M01 = broj atributa koji su 0 u p i 1 u q
  - M10 = broj atributa koji su 1 u p i 0 u q
  - M00 = broj atributa koji su 0 u p i 0 u q
  - M11 = broj atributa koji su 1 u p i 1 u q

Mera sličnosti za binarne podatke

- Jednostavno uparivanje koeficijenata (eng. SMC)

  SMC = broj uparenih / broj atributa

  = (M11 + M00) / (M01 + M10 + M11 + M00)
- Žakardovi (Jaccard) koeficijenti
  - asimetrični binarni atributi

  J = broj parova 11/ broj ne oba-su-nula vrednosti atributa = (M11) / (M01+ M10+ M11)

Mera sličnosti

- Kosinusna sličnost
  - d1 i d2 - dva vektora

  $\cos(d1, d2) = (d1 \bullet d2) / ||d1|| ||d2||$

  gde • označava skalarni proizvod vektora, i || d || označava dužinu vektora d
  - asimetrični podaci
  - najčešća mera sličnosti dokumenata

Mera sličnosti

- Prošireni Žakardovi koeficijenti (Koeficijenti Tanimoto)

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

  - asimetrični podaci

Mera sličnosti

- Korelacija
- Korelacija dva objekta koji imaju binarne ili neprekidne atribute je mera linearnog odnosa između njihovih atributa
- x i y vektori

  corr(x,y)=kovarijansa(x,y)/(standardna devijacija(x) * standardna devijacija(y))

Mera sličnosti

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

kovarijansa(x,y)= $s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$

standardna devijacija(z)= $s_z = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$

srednja vrednost od z= $\bar{z} = \frac{1}{n} \sum_{k=1}^{n} z_k$

Podaci

5. Sledeći atributi su korišćeni za opis članova krda azijskih slonova: težina, visina, dužina kljove, površina uveta. Koju meru bliskosti treba koristiti za poređenje ili grupisanje slonova?

   Svi atributi su numerički, ali mogu imati različit opseg vrednosti (zavisno od skale na kojoj su mereni). Nisu asimetrični i veličina atributa je važna. Euklidsko rastojanje, pri čemu se vrši standardizacija da sredina bude 0 i standardna devijacija 1.

Podaci

6. Ako je data dokument-term matrica u kojoj je tf_{ij} frekvencija *i*-te reči (terma) u *j*-tom dokumentu i *m* je broj dokumenata. Ako je data transformacija nad promeljivom

$$tf'_{ij} = tf_{ij} * \log(\frac{m}{df_i})$$

gde je df_i broj dokumenata u kojima se term *i* pojavljuje (dokument frekvencija terma). Ova transformacija je poznata kao inverzna dokument frekvencija.

Podaci

– Šta je rezultat ove transformacije ako se reč pojavljuje u jednom dokumentu? U svakom dokumentu?

  Ako se reč pojavljuje u svakom dokumentu ima težinu 0, a ako se pojavljuje u jednom dokumentu ima težinu log m.

– Koji je cilj ove transformacije?

  Razlikovanje dokumenta po rečima koja se retko pojavljuju.

Podaci

7. Upoređivanje mera sličnosti i razlika
   - Izračunati Hamingovo rastojanje i Žakardov koeficijent za vektore

     x=0101010001

     y=0100011000

     Hamingovo rastojanje = broj različitih bitova=3

     J= broj parova 11/ broj ne oba-su-nula vrednosti atributa=2/5=0.4

Podaci

– Ako se poredi koliko su slična dva organizma različitih vrsta preko broja gena koji dele, koju meru treba koristiti, Hamingovo rastojanje ili Žakardov koeficijenat radi poređenja genetskog sklopa dva organizma? (Svaki organizam je predstavljen kao binarni vektor, gde je svaki atribut 1 ako organizam sadrži određeni gen, a u suprotnom je 0).

Žakardov koeficijent je bolji za poređenje genetskog sklopa dva organizma, jer se dobija podatak koliko gena dele.

Podaci

– Ako se porede dva organizma iste vrste (npr. dva čoveka), da li je bolje koristiti Hamingonovo rastojanje ili Žakardov koeficijent? Dva čoveka imaju preko 99,9% istih gena.

  Hamingonovo rastojanje, jer nas zanimaju razlike.

Podaci

8. Za vektore x i y izračunati navedene mere sličnosti ili razlike:
   - x=(1,1,1,1), y=(2,2,2,2) kosinusna sličnost, korelacija, Euklidsko rastojanje

     cos(x,y)=1, corr(x,y)=0/0, Euklidsko(x,y)=2
   - x=(0,1,0,1), y=(1,0,1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje, Žakardov koeficijent

     cos(x,y)=0, corr(x,y)=-1, Euklidsko(x,y)=2, Žakard(x,y)=0

Podaci

- x=(0,-1,0,1), y=(1,0,-1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje

  cos(x,y)=0, corr(x,y)=0, Euklidsko(x,y)=2

- x=(1,1,0,1,0,1), y=(1,1,1,0,0,1) kosinusna sličnost, korelacija, Žakardov koeficijenat

  cos(x,y)=0,75, corr(x,y)=0,25, Žakard(x,y)=0,6

- x=(2,-1,0,2,0,-3), y=(-1,1,-1,0,0,-1) kosinusna sličnost, korelacija

  cos(x,y)=0, corr(x,y)=0

Podaci

9. Ako mera sličnosti ima vrednosti u intervalu [0,1], kako bi transformisali vrednost sličnost u vrednost različitosti u intervalu $[0, \infty]$?

   d=-log s

Podaci

10. Bliskost je obično definisana između para objekata.
    - Kako se može izračunati razlika između dva skupa tačaka u Euklidskom prostoru?

      Npr. računanjem centroida između skupa tačaka.
    - Kako se može definisati bliskost između dva skupa objekata?

      Prosečna vrednost bliskosti parova iz različitih grupa, ili najmanja ili najveća bliskost parova iz različitih grupa.