Preprocesiranje podataka

Glava 3. Sadržaj

- Zašto preprocesiranje?
- Čišćenje podataka
- Integracija podataka
- Transformacija podataka
- Smanjivanje obima podataka

Zašto preprocesiranje?

- Podaci iz realnih sistema su:
  - Sa šumom, npr. sadrže greške
  - Nekompletni, npr. fale vrijednosti nekih atributa ili fale atributi u cjelini
  - Nekonzistentni, npr. nekonzistentna imena
  - Razlog: uglavnom veličina skupa podataka
- Kvalitetni podaci su neophodni za usješnu primjenu data mining algoritama

Tehnike preprocesiranja

- Čišćenje podataka (data cleaning)
  - Eliminacija nepoznatih vrijednosti, šuma, ne-konzistentnosti
- Integracija podataka (data integration)
  - Integracija različitih BP, kubova podataka ili datoteka
- Transformacija podataka (data transformation)
  - Normalizacija, agregacija
- Smanjivanje obima podataka (data reduction)
  - Agreriranje, eliminacija redudantnih atributa, klasterizacija
- Diskretizacija podataka (data discretization)
  - Hijerarhija koncepata, “rudarenje” na različitim nivoima abstrakcije

Tehnike preprocesiranja (2)

- Podaci za analizu su:
  - nekompletni, fale vrijednosti za neke atribute, ili samo sadrže agregatne vrijednosti
  - sa šumom (noisy), sadrže greške ili izuzetke
  - nekonzistentni, sadrže neslaganja u šifarnicima
- Razlozi: nedostupnost potrebnih atributa, mišljenje da u trenutku unosa podatak nije potreban, greška u funkcionisanju sistema ili ljudski faktor, prenos podataka, nekonzistentna imenovanja, izbrisani greškom ...

Tehnike preprocesiranja (3)

- Čišćenje podataka
  - dopisivanje vrijednosti koje fale
  - eliminacija šuma
  - identifikovanje i eliminacija izuzetaka
  - eliminisanje nekonzistentnosti
- Robusnost data mining algoritama vs. overfiting

Tehnike preprocesiranja (4)

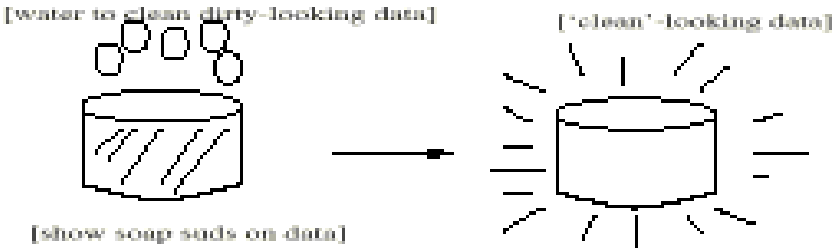
- Integracija podataka
  - obuhvatanje podataka iz više različitih izvora
  - nekonzistentnost, redudantnost (izvedeni atributi), isti koncepti različito imenovani, npr. customer_id, cust_id
- Uobičajeno je da se čišćenje i integracija sprovode prilikom izgradnje data warehouse sistema
  - moguće je primijeniti čišćenje podataka poslije integracije zbog eliminacije redudantnosti

Tehnike preprocesiranja (5)

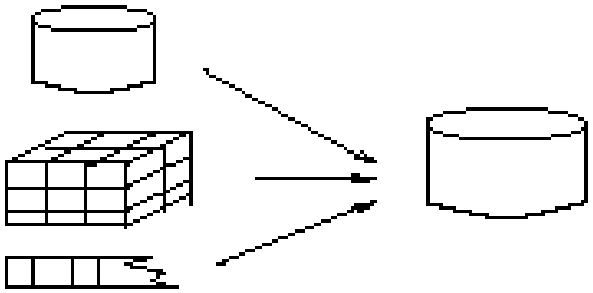
- Transformacija podataka
  - normalizacija
  - agregacija
- Redukcija podataka, formiranje manjeg skupa po obimu, ali sa istim rezultatima analize
  - agregacija (izračunavanje kocki podataka)
  - smanjivanje dimenzionalnosti (brisanje atributa)
  - redukcija brojnosti (klaster reprezentacija)
  - kompresija podataka
  - generalizacija, hijerarhija koncepata

Tehnike preprocesiranja (6)

Data Cleaning



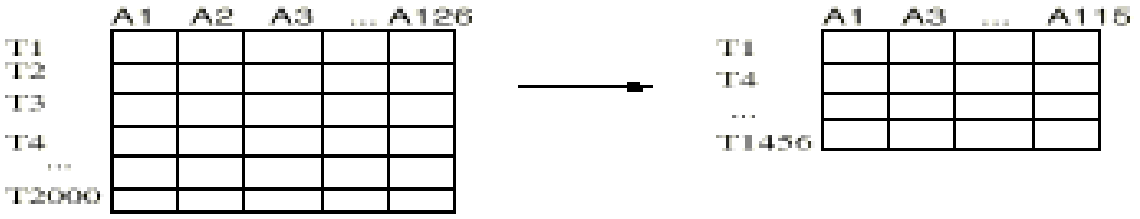
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Glava 3. Sadržaj

- Zašto pre-procesiranje?
- Čišćenje podataka
- Integracija i transformacija podataka
- Smanjivanje obima podataka
- Diskretizacija i generisanje hijerarhije koncepata

Čišćenje podataka

- Osnovni zadaci čišćenja podataka
  - Eliminacija nepoznatih vrijednosti
  - Identifikovanje izuzetaka i eliminacija šuma
  - Rješavanje nekonzistentnosti

Nepoznate vrijednosti

- Podaci nijesu uvijek dostupni
- Mogući uzroci su:
  - Tehnička greška
  - Nekonzistentni sa ostalim podacima i izbrisani
  - Nijesu unijeti zbog greške korisnika sistema ili zato što su smatrani nebitnim u trenutku upisivanja
- Nepoznati podaci u nekim slučajevima moraju da se na neki način izvedu

Kako izvesti vrijednosti za nepoznate podatke?

- Ignorisati takve zapise
- Ručno dopuniti vrijednosti koje fale
- Koristiti novu vrijednost koja će zamijeniti onu koja fali
- Upisati srednju vrijednost za taj atribut
- Upisati srednju vrijednost za taj atribut ali u okviru iste klase
- Upisati najvjerovatniju vrijednost koja se može izvesti primjenom DT ili Bajesovim pravilom

Šum u podacima

- Šum je slučajna greška u podacima
- Eliminacija šuma:
  - Binning metoda (local smoothing)
    - Prvo se podaci sortiraju i podijele u binove
    - Onda se svaka vrijednost u binu zamijeni sa aritmetičkom sredinom, medijanom ili bližom graničnom vrijednošću
  - Klasterizacija
    - Podaci koji ostaju van klastera su šum
  - Regresija
    - Traži se funkcija koja odgovara podacima, podaci van funkcije su šum
  - Ovo su i metode za redukciju podataka

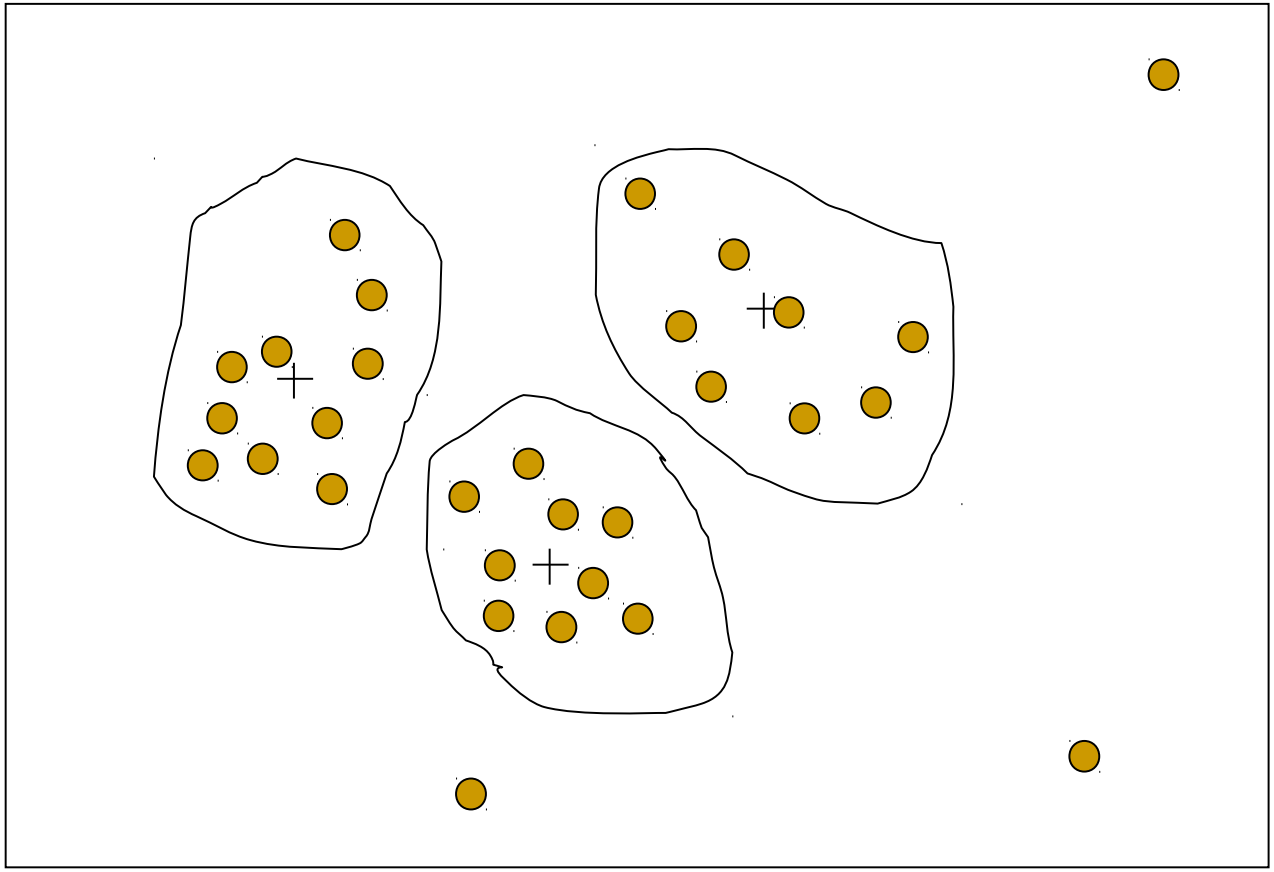
Binning metoda

- Equal-width binovi
  - Domen se dijeli na N jednakih intervala
  - Ako je A najmanja a B najveća vrijednost domena, tada je veličina intervala W=(B-A)/N
- Equal-depth binovi
  - Domen se dijeli na N djelova tako da svaki dio sadrži približno jednak broj primjeraka

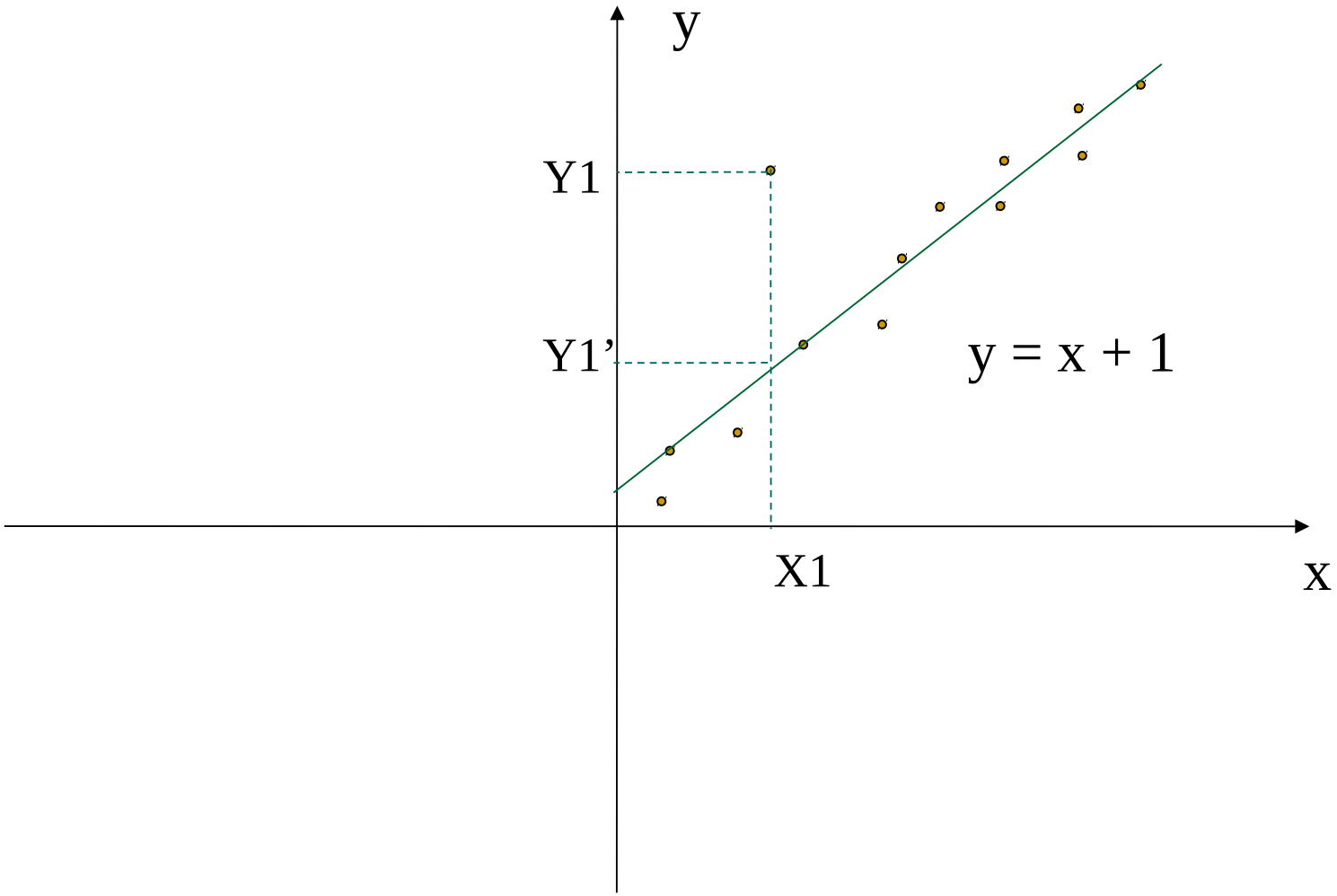
Binning metoda (2)

- Neka je sortirani niz podataka: 4, 8, 15, 21, 21, 24, 25, 28, 34
- Equal-depth binovi su:
  - Bin1 = 4, 8, 15
  - Bin2 = 21, 21, 24
  - Bin3 = 25, 28, 34
- Eliminacija šuma sa aritmetičkom sredinom
- Eliminacija šuma sa graničnim vrijednostima

Klasterizacija



Regresija



Glava 3. Sadržaj

- Zašto pre-procesiranje?
- Čišćenje podataka
- Integracija i transformacija podataka
- Smanjivanje obima podataka
- Diskretizacija i generisanje hijerarhije koncepata

Integracija podataka (2)

- Redudantnost u podacima
  - Isti atribut sa različitim imenima
  - Izvedeni atributi
- Korelaciona analiza
  - Određuje uticaj jednog atributa na drugi
  - Pozitivna i negativna korelacija

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} \qquad \overline{A} = \frac{\sum A}{n} \qquad \sigma_A = \sqrt{\frac{\sum (A - \overline{A})^2}{n-1}}$$

Integracija podataka (3)

- Prepoznavanje duplikata na nivou torke (duplication at the tuple level)
- Konflikti u podacima (data value conflicts)
  - za isti entitet vrijednosti atributa iz raznih izvora se razlikuju, npr. težina izražena u kg i britanskim funtama, cijena sa porezom i taksama ili bez, semantika u podacima

Transformacija podataka

- Eliminacija šuma
  - bining, klasterisanje, regresija
- Agregacija (konstrukcija kocki podataka)
- Generalizacija (hijerarhija koncepata)
- Normalizacija
  - kod algoritama klasifikacije i klasterizacije, za povećanje brzine i sprečavanje dominacije nekih atributa
- Konstrukcija atributa

Normalizacija

- min-max normalizacija

$$v' = \frac{v - \mathbf{min}_A}{\mathbf{max}_A - \mathbf{min}_A}\left(\mathbf{new}_{\mathbf{max}_A} - \mathbf{new}_{\mathbf{min}_A}\right) + new_{min_A}$$

- z-score normalizacija

$$v' = \frac{v - mean_A}{\mathbf{stand}_{dev_A}}$$

- decimalno skaliranje
  - j je najmanji cijeli broj takav da je $Max(|v'|) < 1$

$$v' = \frac{v}{10^j}$$

Glava 3. Sadržaj

- Zašto pre-procesiranje?
- Čišćenje podataka
- Integracija i transformacija podataka
- Smanjivanje obima podataka
- Diskretizacija i generisanje hijerarhije koncepata

Smanjivanje obima podataka

- Data warehouse sistemi mogu da sadrže terabajte podataka.
- Redukcija podataka je smanjenje skupa podataka u veličini, ali bez gubitka informacija bitnih za analizu
- Tehnike
  - Agregacija
  - Redukcija dimenzionalnosti
  - Kompresija podataka
  - Redukcija brojnosti
  - Diskretizacija i generisanje hijerarhije koncepata

Agregacija

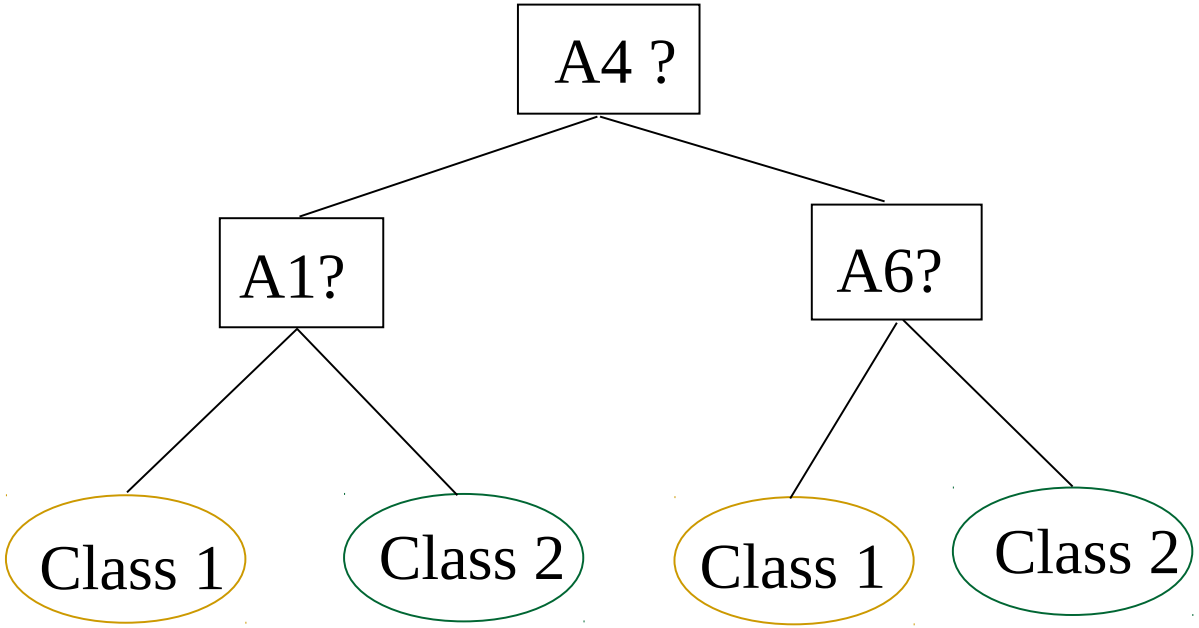
- Kombinovanje dva ili više atributa (objekata) u jedan atribut (objekat)
- Cilj
  - Smanjivanje obima podataka smanjivanjem broja atributa ili objekata
  - Promjena nivoa na kojem se podaci prikazuju
    - mjesec, kvartal, godina itd.
  - Agregirani podaci su statistički “stabilniji”

Redukcija dimenzionalnosti

- Smanjivanje broja atributa je eliminacija redundantnih i nerelevantnih atributa
- Metoda selekcije podskupa atributa + heuristike (greedy algoritmi, information gain)
  - Selekcija unaprijed
  - Selekcija unazad
  - Kombinovana selekcija
    - kriterijum zaustavljana za prethodne 3 metode
  - Drveta odlučivanja

Primjer sa drvetom odlučivanja

Početni skup atributa:
{A1, A2, A3, A4, A5, A6}

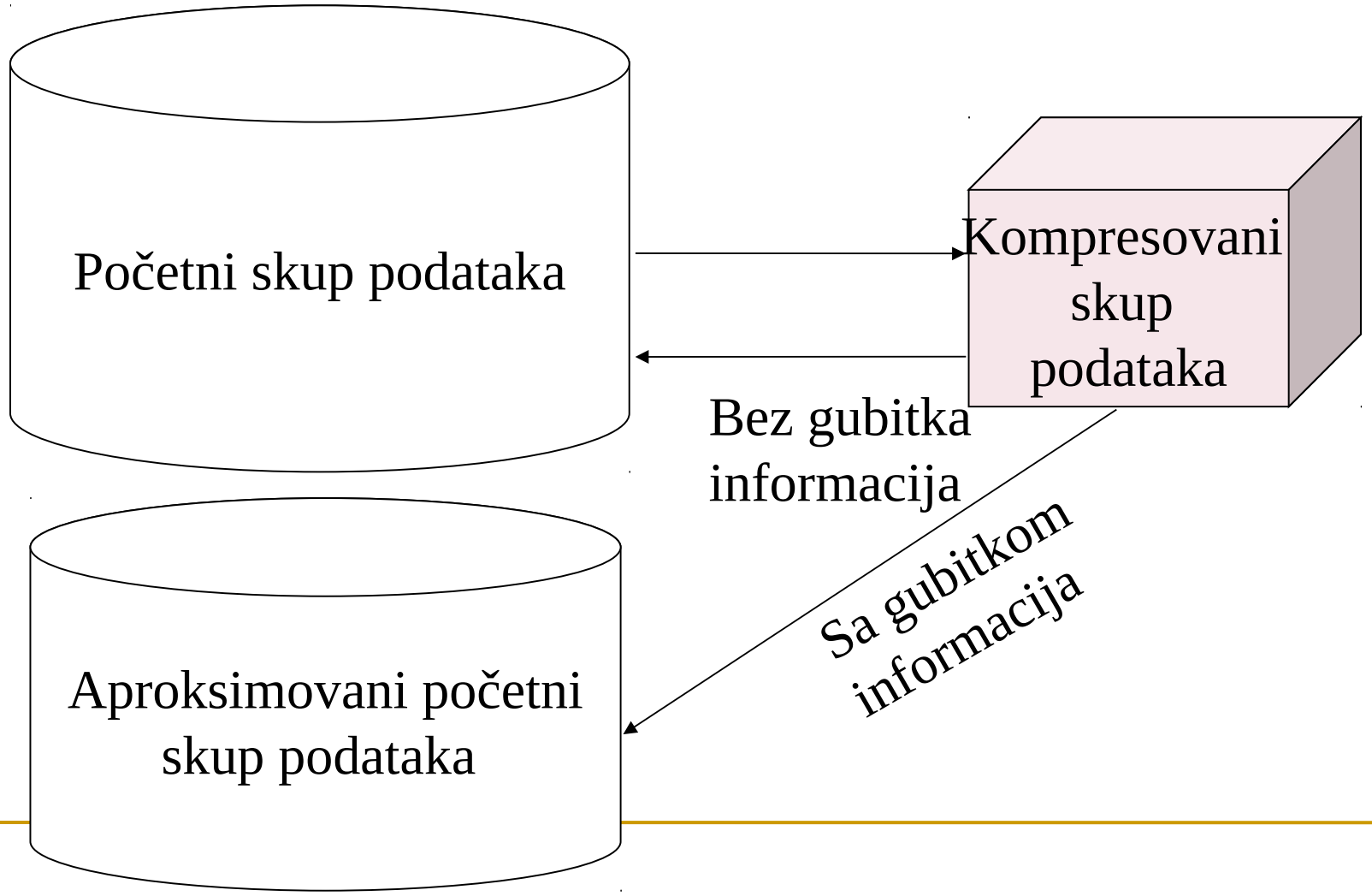


-------> Smanjeni skup atributa: {A1, A4, A6}

Kompresija podataka

- Kompresijom se smanjuje polazni skup podataka
  - Sa gubitkom informacija
  - Bez gubitka informacija
- Najčešće tehnike
  - Diskretna wavelet transformacija
  - Primarna analiza komponenti

Kompresija podataka (2)



Diskretna wavelet transformacija

- DWT je linearna tehnika transformisanja signala
  - vektor D transformiše u D', vektor koeficijenata, dužine oba vektora su jednake
- Dovoljno je pamtiti nekoliko najznačajnijih koeficijenata
- Inverzna transformacija
- Složenost za brzu DWT je O(n)

Primarna analiza komponenti

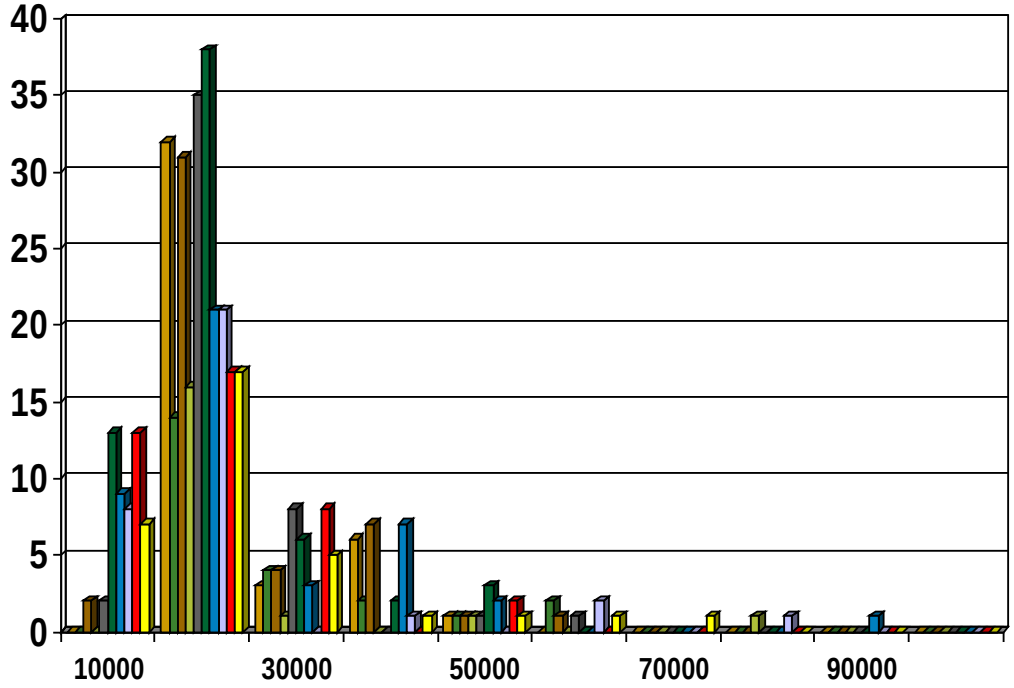
- Polazni skup sadrži n vektora sa k atributa
- Ideja je da se pronađe c<k međusobno ortogonalnih vektora, koji će da čine normiranu bazu za ulazni skup
- Svaki vektor iz polaznog skupa može da se prikaže linearnom kombinacijom vektora iz primarne komponente
- Redukcija dimenzionalnosti

Redukcija brojnosti

- Parametarizovane metode
  - Podaci se opisuju modelom, procijene se parametri modela i sačuvaju se, originalni podaci se brišu
  - Regresija, $Y = a + b*X$
- Neparametarizovane metode
  - Histogrami
  - Metode klasterizacije
  - Metode uzoraka

Histogrami

- Domen atributa A se dijeli na disjunktne podskupove ili bakete
- Baketi se prikazuju na x osi, na y osi se prikazuje broj primjeraka u baketu
- Singlton baketi



Histogrami (2)

- Metode za formiranje baketa:
  - equiwidth
  - equidepth
  - V-optimal
    - za dati broj baketa ovo particionisanje je sa najmanjom histogram varijansom
  - Max-Diff

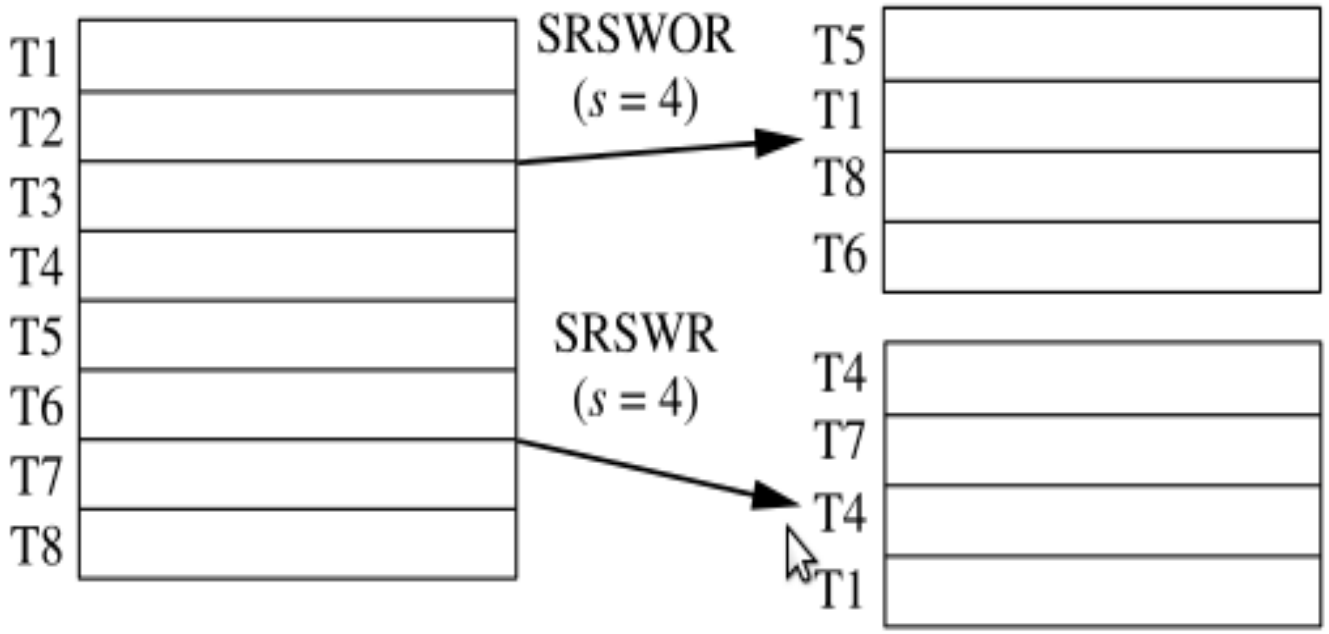
Metode klasterizacije

- Generišu se klasteri sličnih objekata
  - funkcija rastojanja
  - kvalitet klasterizacije
    - dijametar klastera, maksimalno rastojanje bilo koja dva objekta u klasteru
    - rastojanje cetroida, prosječno rastojanje objekata do centroida
- Polazni skup se zamjenjuje klaster reprezentacijom
- Hijerarhija klastera

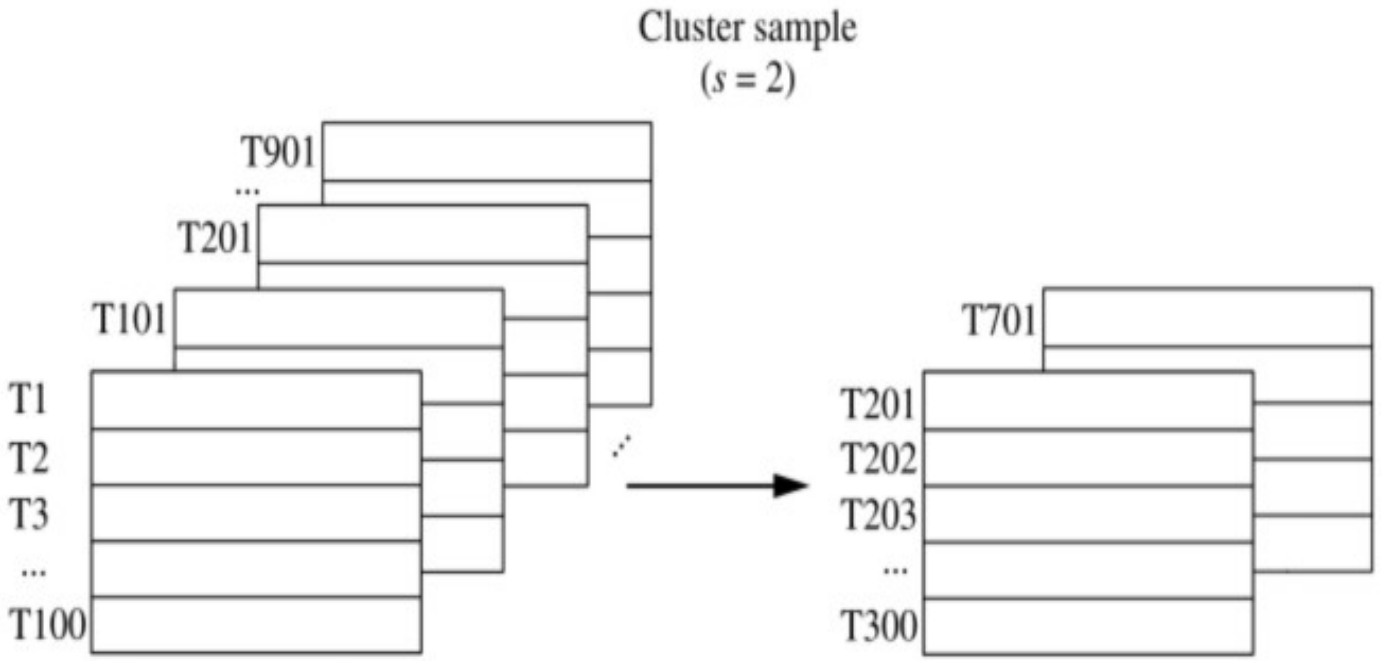
Metoda uzoraka

- Originalni skup zamjenjuje se po veličini mnogo manjim uzorkom
- Izabiranje reprezentativnog uzorka
  - SRSWOR veličine n (simple random sample without replacement)
  - SRSWR veličine n (simple random sample with replacement)
  - Klaster uzorak
  - Stratum uzorak

Metoda uzoraka (2)



Metoda uzoraka (3)



Metoda uzoraka (4)

Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Glava 3. Sadržaj

- Zašto pre-procesiranje?
- Čišćenje podataka
- Integracija i transformacija podataka
- Redukcija podataka
- Diskretizacija i generisanje hijerarhije koncepata

Hijerarhija koncepata

Redukcija podataka grupisanjem koncepata nižeg nivoa u koncepte na višem nivou hijerarhije

- problemi generalizacije
- podaci postaju izražajniji, lakši za razumijevanje, zahtijevaju manje prostora za čuvanje i pogodniji su za mnoge algoritme
- definisanje hijerarhije na nivou šeme; automatsko generisanje hijerarhije

Diskretizacija i hijerarhije koncepata za numeričke atribute

- Binning
- Histogrami
- Klasterizacija
- Entropija (information-based measure)
  - koristi informacije o raspodjeli klasa
- Segmentacija prirodnim particionisanjem
- Prethodne metode rekurzivnom primjenom generišu hijerarhiju

Entropija

- Polazi se od skupa primjeraka S i atributa A koji se diskretizuje
- Svaka vrijednost v iz domena atributa A proizvodi binarnu diskretizaciju: A<v i A>=v
- Bira se vrijednost koja maksimizuje informacionu dobit

$$E(S,T)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2)$$

$$Ent(S_1)=-\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Kriterijum zaustavljanja $Ent(S)-I(S,v)>\delta$

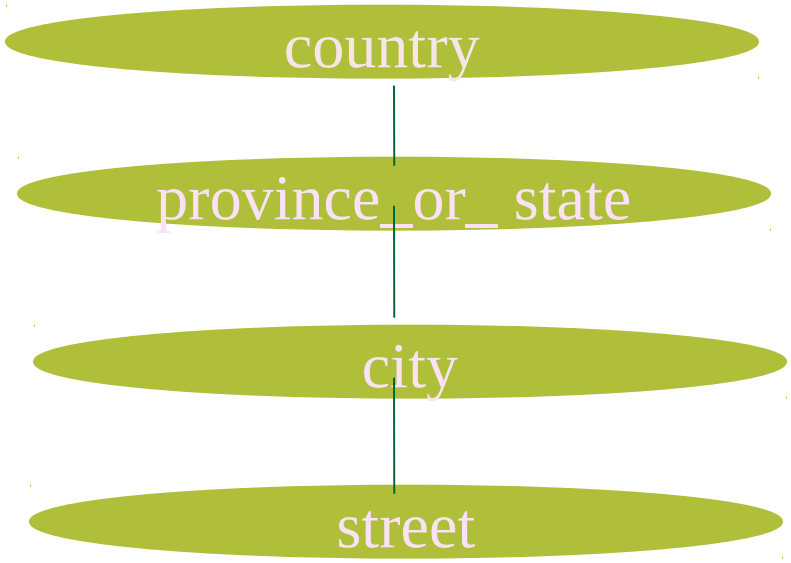
Segmentacija prirodnim particionisanjem

- Generiše intervale koji su “prirodni”
- Pravilo 3-4-5 dijeli polazni skup na 3, 4 ili 5 približno jednaka intervala, pa se rekurzivno primjenjuje nivo po nivo
- Algoritam
  - Ako interval sadrži 3, 6, 7 ili 9 različitih vrijednosti u odnosu na najtežu cifru dijeli se na 3 intervala
  - Ako interval sadrži 2, 4 ili 8 razli ... na 4 intervala
  - Ako interval sadrži 1, 5 ili 10 razli ... na 5 intervala

Hijerarnija koncepata za diskretne atribute

- Na nivou šeme baze podataka definiše se uređenje skupa atributa
- Specifikacija dijela hijerarhije eksplicitnim grupisanjem podataka
  - intermediate levels
- Specifikacija skupa atributa ali ne i uređenja
  - automatsko generisanje na osnovu brojnosti
- Specifikacija samo dijela skupa atributa
  - metapodaci definišu semantički povezane atribute

Primjer automatskog generisanja hijerarhije



15 različitih vrijednosti

65 različitih vrijednosti

3567 različitih vrijednosti

674,339 različitih vrijednosti

Zaključak

- Pre-procesiranje podataka je važan korak za data warehouse i data mining
- Pre-procesiranje obuhvata
  - Čišćenje i integracija podataka
  - Transformacija podataka
  - Redukcija podataka i selekcija atributa
  - Diskretizacija i hijerarhija koncepata
- Predloženi su brojni algoritmi ali je ovo i dalje aktivna oblast istraživanja