Istraživanje podataka

Vežbe 13

21. maj 2021

Outline

Outline

Pravila pridruživanja

- Podaci su skup transakcija.
- Jednu transakciju čine stavke koje se pojavljuju zajedno.

Id transakcije	Stavke u transakciji
1	{*Jabuke*, *Lubenica*, *Lubenica*, *Maline*}
2	{*Jabuke*, *Maline*}
3	{*Jabuke*, *Banane*}
4	{*Jabuke*, *Banane*}
5	{*Banane*}
6	{*Banane*, *Maline*, *Lubenica*}

Pravila pridruživanja

- Skup stavki (eng. itemset) sadrži jednu ili više stavki
- Pravila pridruživanja su oblika:

$$telo \rightarrow glava$$

gde su *telo* i *glava* skupovi stavki. Npr. $Jabuke \rightarrow Banane$ ili $Jabuke, Mailine \rightarrow Lubenica$

Pravila pridruživanja - mere kvaliteta

Za pravilo pridruživanja $X \rightarrow Y$ i broj transakcija N

- Broj transakcija koje sadrže stavke X: $\sigma(X)$
- Podrška (Support) : $sup(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
- Pouzdanost (Confidence): $conf(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

Pravila pridruživanja

Za dati skup transakcija cilj je pronaći pravila pridruživanja koja imaju

- podršku $\geq min_{sup}$ (minimalni prag podrške)
- pouzdanost $\geq min_{conf}$ (minimalni prag pouzdanosti)

Skup stavki S za koji važi $sup(S) \geq min_{sup}$ se naziva čest skup stavki.

Pravila pridruživanja

1. Generisanje čestih skupova stavki
2. Generisanje pravila

Generisanje čestih skupova stavki

Apriori princip: Ako je skup čest, onda i svi njegovi podskupovi moraju biti česti.

Anti-monotonost podrške: $\forall X, X \subset Y : sup(Y) \leq sup(X)$

Generisanje čestih skupova stavki - Apriori algoritam

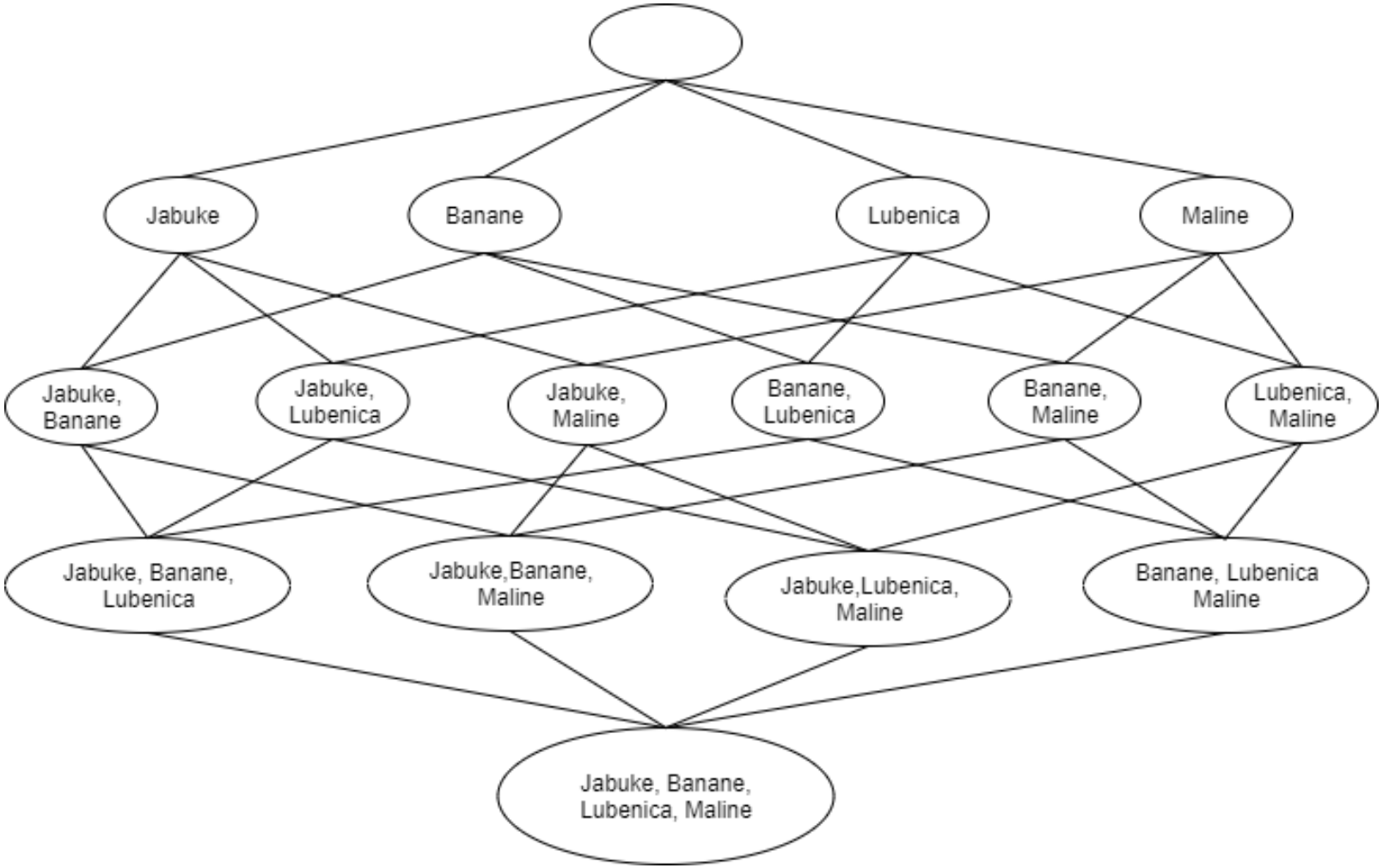
Generisanje čestih skupova stavki

1. Identifikacija skupova stavki dužine 1 koji su česti.
2. Za k od 2 do n, gde je n broj stavki u skupu podataka izvršiti
   1. Generisanje kandidata
      - Generisanje kandidata dužine k (skup stavki dužine k) na osnovu k-1 čestih skupova stavki - spajanjem dva česta skupa stavki dužine k-1 za koje važi da se prvih k-2 stavki poklapaju kada su im stavke sortirane leksikografski.
   2. Čišćenje kandidata
      - Eliminisanje k skupova stavki iz kandidata ako važi da postoji neki podskup dužine k-1 koji nije čest.
      - Računanje podrške za preostale kandidate i eliminisanje kandidata dužine k čija je podrška manja od sup_{min}

Generisanje čestih skupova stavki

- Rešetka skupa stavki
- *Maksimalno čest skup stavki* - čest skup čiji ni jedan nadskup nije čest.

Primer rešetke



Lift mera

Za računanje kvaliteta pravila pridruživanja, pored podrške i pouzdanosti, može se koristiti Lift mera:

$$Lift = \frac{conf(X \rightarrow Y)}{sup(Y)}$$

* Pravilo $X \rightarrow Y$ je zanimljivo ako je $Lift(X \rightarrow Y) \neq 1$

Outline

Zadatak 1

Dat je skup podataka

Id transakcije	Stavke u transakciji
1	$\{Jabuke, Lubenica, Maline\}$
2	$\{Jabuke, Maline\}$
3	$\{Jabuke, Banane\}$
4	$\{Jabuke, Banane\}$
5	$\{Banane\}$
6	$\{Banane, Lubenica, Maline\}$

Zadatak 1

- Nacrtati rešetku skupova stavki koja odgovara datom skupu podataka. Označiti čvorove u rešetki sledećim slovima:
  - *N*: ako se skup stavki ne smatra kandidatom po Apriori algoritmu. Tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
  - *F*: kandidat skupa stavki je čest po Apriori algoritmu.
  - *I*: Ako se skup stakvi smatra retkim posle određivanja podrške.

Minimalna podrška za česte skupove je 0,3.

Zadatak 2

Dat je skup podataka

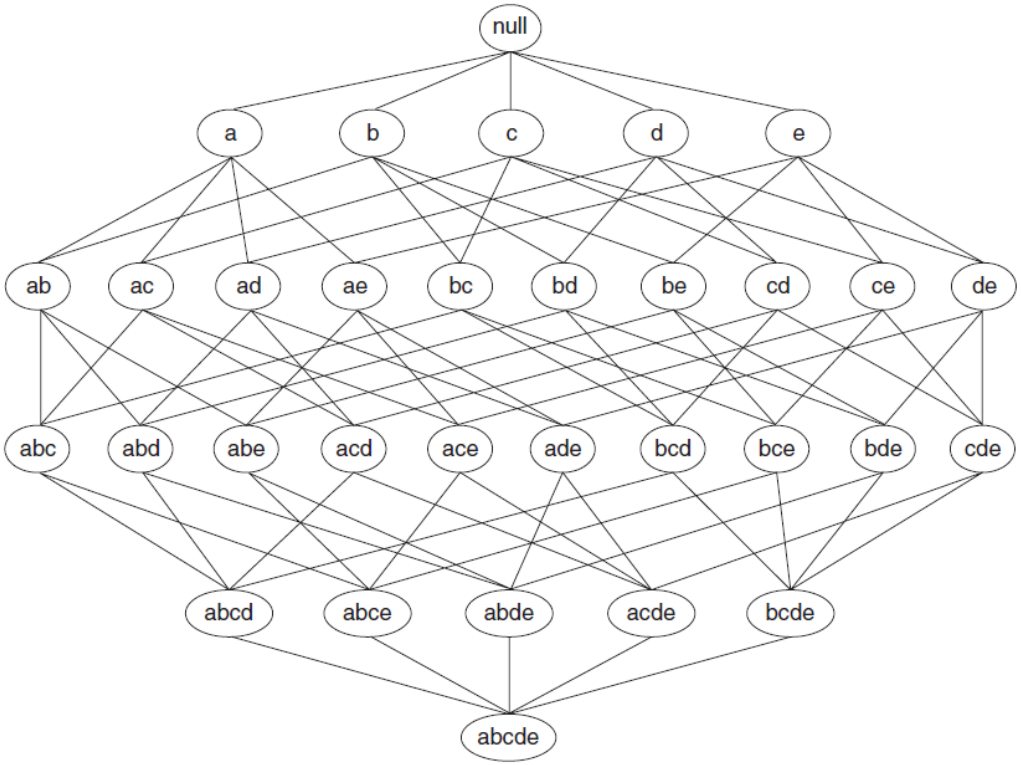
Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Zadatak 2

- Nacrtati rešetku skupova stavki koja odgovara datom skupu podataka. Označiti čvorove u rešetki sledećim slovima:
  - *N*: ako se skup stavki ne smatra kandidatom po Apriori algoritmu. Tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
  - *F*: kandidat skupa stavki je čest po Apriori algoritmu.
  - *I*: Ako se skup stakvi smatra retkim posle određivanja podrške.

Minimalna podrška za česte skupove je 0,3.

Zadatak 2



Transaction ID	Items Bought
1	$\{a,b,d,e\}$
2	$\{b,c,d\}$
3	$\{a,b,d,e\}$
4	$\{a,c,d,c\}$
5	$\{b,c,d,e\}$
6	$\{b,d,e\}$
7	$\{c,d\}$
8	$\{a,b,c\}$
9	$\{a,d,e\}$
10	$\{b,d\}$

Zadatak

- Koliki je procenat čestih skupova stavki?
- Koliki je odnos čišćenja Apriori algoritma za ovaj skup podataka? Odnos čišćenja je definisan kao procenat skupova stavki koji nisu generisani za vreme generisanja kandidata ili su eliminisani u koraku čišćenja kandidata.
- Koliki je odnos *lažnog alarma* (procenat kandidatskih skupova stavki koji su obeleženi kao retki posle prebrojavanja podrške)?

Outline

Format podataka

- Transakcioni format - jedan red u tabeli - jedna stavka u transakciji. Kolone: id transakcije i jedna stavka.

Id transakcije	Stavka
1	A
1	B
1	C
2	A
2	B
3	B
3	D

Format podataka

- Tabelarni format (korpa). Jedna stavka - jedna binarna kolona. Jedna transakcija - jedan red. Ukoliko se stavka javlja u transakciji, pridružena kolona stavke ima vrednost *tačno* u redu koji odgovara toj transakciji.
  - Uloge atributa: Input, Target, ili Both

Id transakcije	A	B	C	D
1	T	T	T	F
2	T	T	F	F
3	F	T	F	T

Parametri u SPSS modeleru, algoritam Apriori

- Minimalna podrška tela pravila
  *U alatu SPSS Modeler:
  - Podrška (Support): $sup(X \rightarrow Y) = \frac{\sigma(X)}{N}$
  - Podrška pravila (Rule Support): $sup(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
- Minimalna pouzdanost pravila
- Maksimalan broj uslova u telu pravila
- Uzeti u obzir samo vrednosti **tačno** u binarnim atributima

Praktični zadaci

1. Primenom pravila pridruživanja proveriti da li postoji zavisnost među podacima u skupu *transactions*.
2. Primenom pravila pridruživanja proveriti da li postoji zavisnost među upisanim izbornim predmetima.

Association Rules

U odnosu na čvor **Apriori** čvor **Association Rules**:

- ne radi sa podacima u transakcionom obliku
- ima više opcija za postavljanje ograničenja koja se koriste pri izdvajanju zanimljivih pravila pridruživanja