Istraživanje podataka 1 - vežbe 11, 2020.

1 Pravila pridruživanja

1.1 Skupovi stavki u transakcijama

Skup podataka u kome se traže pravila pridruživanja se sastoji od transakcija. Jednu transakciju čine stavke koje se pojavljuju zajedno. Npr. ako skup podataka sadrži podatke o kupovini namirnica na pijaci voća, jedna transakcija je jedna kupovina jednog kupca. Primer skupa podataka je dat u tabeli 1.

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

Tabela 1: Skup podataka sadrži podatke o kupovini namirnica na pijaci voća, jedna transakcija je jedna kupovina jednog kupca.

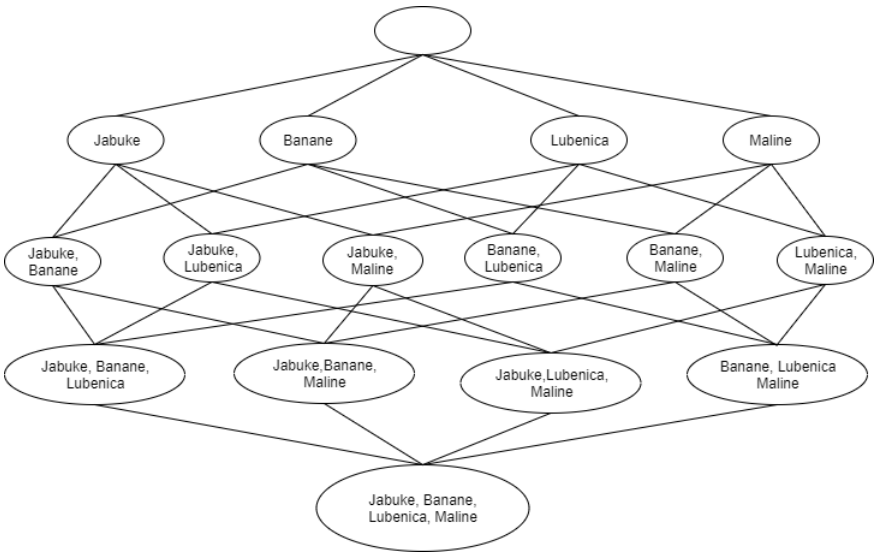
Skup stavki (eng. itemset) sadrži jednu ili više stavki. Skup stavki koji sadrži k stavki naziva se k-skup stavki.
Bitno svojstvo skupa stavki je podrška. Ako je ukupan broj transakcija u skupu N, a $\sigma(X)$ broj transakcija koje sadrže skup stavki X, podrška skupa stavki X se računa kao

$$sup(X) = \frac{\sigma(X)}{N}$$

Podrška za skup stavki $\{Jabuke, Banane\}$ u skupu iz tabele 1 je $\frac{2}{6}$

Skup stavki X za koji važi $sup(X) \geq min_{sup}$ se naziva **čest skup stavki**. min_{sup} je minimalna podrška koju zadaje korisnik.

Za pregledan način prikazivanja svih skupova stavki koje se pojavljuju u skupu podataka koirsti se *rešetka*. Skupovi stavki su podeljenji u $n + 1$ nivoa, gde je n broj stavki koje se javljaju u skupu. Nivoi su obeleženi od 0 do n. Na nivou k se predstavljaju k-skupovi stavki. 0-skup stavki je prazan skup, 1-skupovi stavki sadrže po jednu stavku, a n-skup stavki sadrži sve stavke koje se javljaju u skupu podataka. Skup stavki na nivou k je povezan sa svojim nadskupovima na nivou $k + 1$. Na slici 1 je prikazana rešetka sa skupovima stavki za skup podataka iz tabele 1.



Slika 1: Prikaz skupova stavki za skup podataka dat u tabeli 1 preko *rešetke*

1.2 Pravila pridruživanja u skupu podataka

Pri određivanju pravila pridruživanja u skupu podataka traže se pravila oblika:

$$telo \rightarrow glava$$

gde su *telo* i *glava* skupovi stavki. Pravilo se čita sa: *Ako se u transakciji pojave stavke iz tela, verovatno će se pojaviti i stavke iz glave.*

Npr. pravila mogu biti $Jabuke \rightarrow Banane$ ili $\{Jabuke, Mailine\} \rightarrow Lubenica$

Osnovne mere za određivanje kvaliteta pravila pridruživanja su:

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

Podrška određuje koliko se često pravilo pojavljuje u transakcijama skupa podataka, dok pouzdanost određuje koliko se često stavke iz Y pojavljuju u transakcijama koje sadrže stavke iz X.

Za dati skup transakcija cilj je izdvojiti pravila pridruživanja koja imaju

- podršku $\geq min_{sup}$
- pouzdanost $\geq min_{conf}$

Minimalni prag podrške min_{sup} i minimalni prag pouzdanosti min_{conf} zadaje korisnik.

Najjednostavniji način za određivanje pravila pridruživanja je (1) izdvajanje svih mogućih pravila pridruživaja, (2) računanje podrške i pouzdanosti za svako izdvojeno pravilo i (3) poređenje dobijenih vrednosti sa zadatim pragovima. Međutim, pristup grube sile za određivanje pravila

pridruživanja je izuzetno skup zbog velikog broja pravila koja se mogu izdvojiti iz skupa. Da bi se smanjio broj potrebnih izračunavanja, koristi se svojstvo da je $sup(X \rightarrow Y) = sup(X \cup Y)$, pa se prvo izdvajaju česti skupovi stavki, a zatim se za svaki čest skup stavki prave moguća pravila i proverava koje pravilo ima zadovoljavajuću pouzdanost. Npr. za skup stavki $\{Jabuke, Mailine, Lubenica\}$ moguća pravila (sa uslovom da ni glava ni telo nisu prazan skup) su:
$\{Jabuke\} \rightarrow \{Mailine, Lubenica\}$
$\{Mailine\} \rightarrow \{Jabuke, Lubenica\}$
$\{Lubenica\} \rightarrow \{Jabuke, Mailine\}$
$\{Jabuke, Mailine\} \rightarrow \{Lubenica\}$
$\{Jabuke, Lubenica\} \rightarrow \{Mailine\}$
$\{Mailine, Lubenica\} \rightarrow \{Jabuke\}$

Zbog toga se algoritam *Apriori*, a i drugi, za određivanje pravila pridruživanja sastoji iz dve faze:

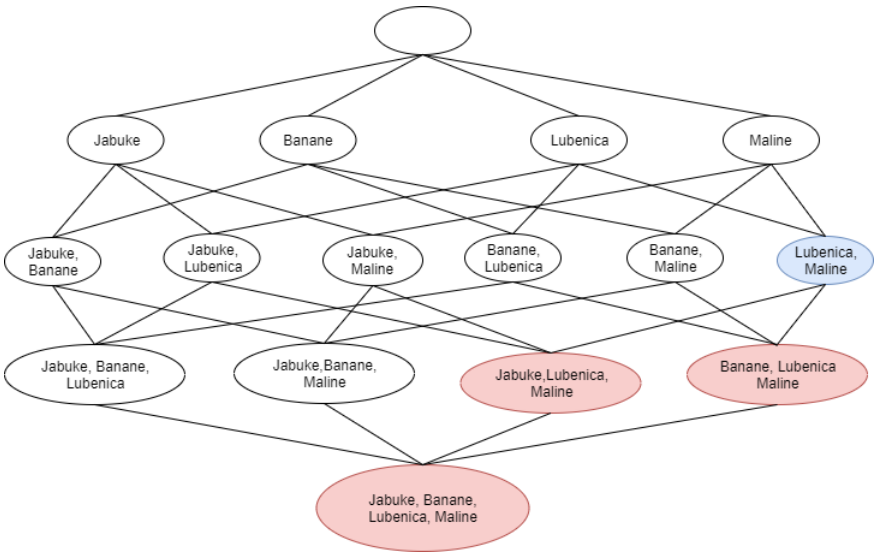
1. Generisanje čestih skupova stavki
2. Generisanje pravila na osnovu izdvojenih čestih skupova

1.3 Apriori algoritam

Jedan od najpoznatijih algoritama za izdvajanje pravila pridruživanja je Apriori algoritam. Apriori algoritam u fazi generisanja čestih skupova stavki koristi osobine podrške kako bi se smanjio broj skupova stavki za koje je potrebno izračunati podršku da bi se odredilo da li je skup stavki čest. Algoritam za smanjene kandidatskih skupova stavki za koje je potrebno izračunati podršku koristi:

- princip: *Ako je skup čest, onda i svi njegovi podskupovi moraju biti česti.*
- anti-monotonost podrške: $\forall X, X \subset Y : sup(Y) \leq sup(X)$

Npr. ako je skup stavki $\{Jabuke, Mailine, Lubenica\}$ čest onda su česti i skupovi stavki $\{Jabuke, Mailine\}$, $\{Jabuke, Lubenica\}$, $\{Mailine, Lubenica\}$, $\{Jabuke\}$, $\{Mailine\}$, $\{Lubenica\}$, jer u svakoj transakciji u kojoj se pojavljuje skup stavki $\{Jabuke, Mailine, Lubenica\}$, pojavljuje se i svaki od njegovih podskupova.
Takođe važi da ako neki od podskupova skupa stavki nije čest, onda ni skup stavki ne može biti čest. Npr. ako podskup $\{Mailine, Lubenica\}$ nije čest, tj. podrška mu je manja od zadatog praga, onda ni skup stavki $\{Jabuke, Mailine, Lubenica\}$ ne može biti čest, kao ni bilo koji drugi nadskup skupa stavki $\{Mailine, Lubenica\}$. Na slici 2 je preko rešetke označen skup stavki $\{Mailine, Lubenica\}$ plavom bojom, a crvenom bojom svi njegovi nadskupovi za skup transakcija dat u tabeli 1.



Slika 2: Prikaz nadskupova skupa stavki $\{Mailine, Lubenica\}$ preko rešetke za skup podataka dat u tabeli 1

U algoritmu 1 su dati koraci za generisanje čestih skupova stavki u Apriori algoritmu.

Algoritam 1 Generisanje čestih skupova stavki u Apriori algoritmu

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}

1.4 Mere kvaliteta pravila pridruživanja

Za računanje kvaliteta pravila pridruživanja, pored podrške i pouzdanosti, mogu se koristiti:

- Lift mera: $Lift = \frac{conf(X \to Y)}{sup(Y)}$ ili $Lift = \frac{conf_{posterior}}{conf_{prior}}$
  * Pravilo $X \to Y$ je zanimljivo ako je $Lift(X \to Y) \neq 1$

  Npr. ako je $conf(X \to Y)$=0.4 (u 40% transakcija u kojima se javlja X, javlja se i Y), a i $sup(Y)$=0.4 (u 40% transakcija u skupu podataka javlja se Y), vrednost mere $Lift$ je 1. Zaključak je da pravilo $X \to Y$ nije interesantno, jer se Y ne pojavljuje ni češće ni ređe u transakcijama koje sadrže X nego u celom skupu transakcija. Ako je $Lift > 1$, Y se pojavljuje u transakcijama koje sadrže X više od očekivanog broja pojavljivanja, a ako je $Lift < 1$, Y se pojavljuje u transakcijama koje sadrže X manje od očekivanog broja pojavljivanja, što takođe može biti interesantno. Za izdvajanje zanimljivih pravila prema

$Lift$ meri može se postaviti uslov da se izdvoje pravila za koja važi da vrednost Lift mere nije u opsegu od npr. 0,95 i 1,1.

- mogućnost raspoređivanja (Deployability): $\frac{\sigma(X)-\sigma(X\cup Y)}{N}$

1.5 Zadaci

1. Nacrtati rešetku skupova stavki koja odgovara skupu podataka datom u tabeli 1. Označiti čvorove u rešetki sledećim slovima:

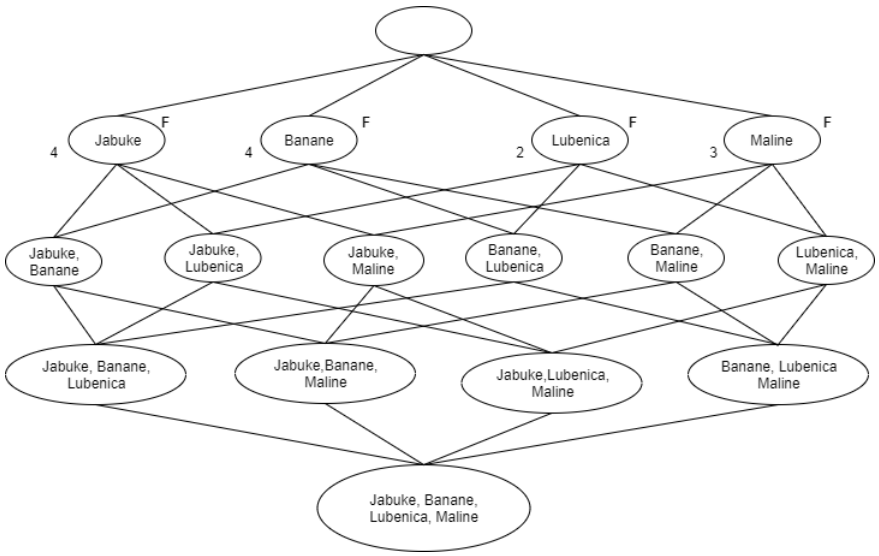
   - N: ako se skup stavki ne smatra kandidatom po Apriori algoritmu, tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
   - F: kandidat skupa stavki je čest po Apriori algoritmu
   - I: Ako se skup stakvi smatra retkim posle određivanja podrške

   Minimalna podrška za česte skupove je 0,3.

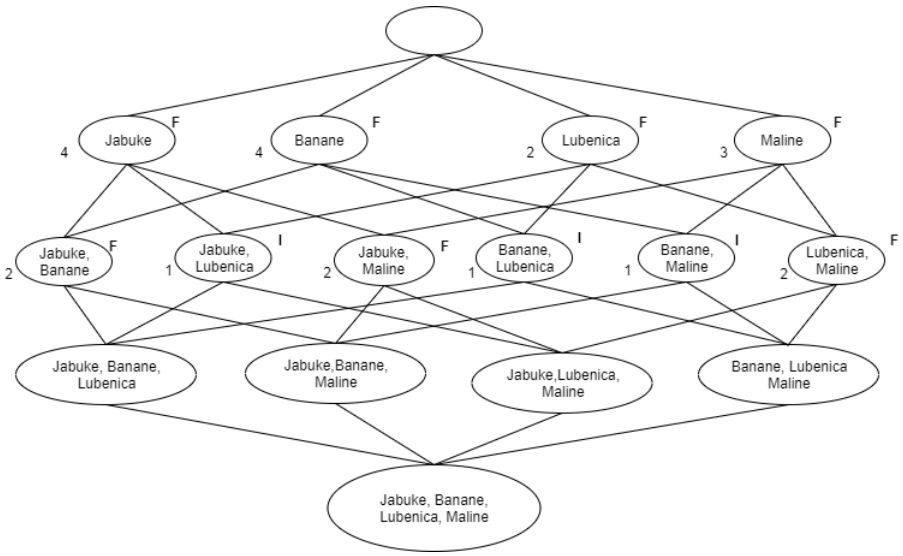
   Rešenje

   Pošto je $min_{sup} = 0,3$, skupovi stavki koji se pojavljuju u bar 2 transakcije su česti. Pored svakog čvora u rešetkama koje su deo rešenja je u gornjem desnom uglu označeno da li je pridruženi skup podataka čest (F), nije čest ali je bilo neophodno prebrojavanje u koliko transakcija se javlja jer su mu svi podskupovi česti (I), ili nije čest i nije bilo potrebno računati u koliko transakcija se javlja jer mu bar jedan podskup nije čest (N). U donjem levom uglu čvora je ispisan broj transakcija u kojima se javlja pridruženi skup stavki ukoliko je bilo potrebno izračunati tu vrednost.

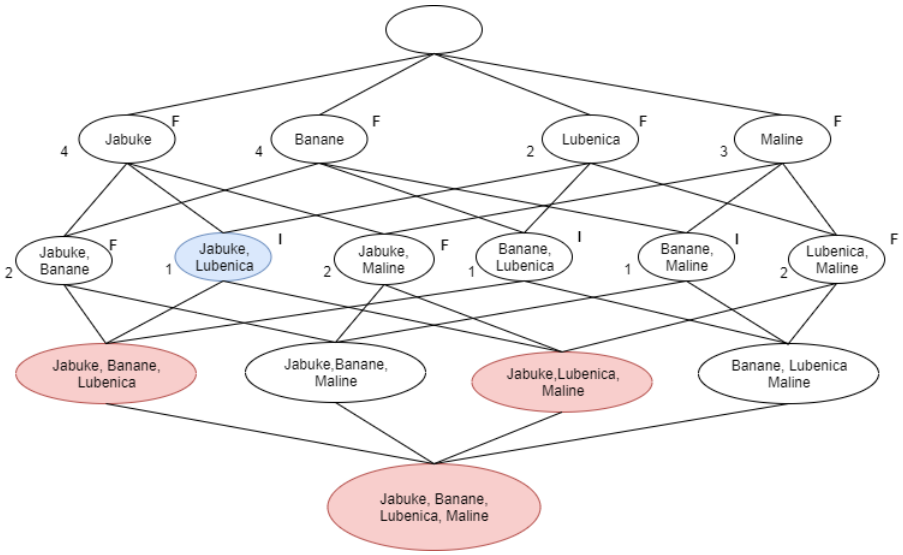
   Prvo se za 1-skupove stavki prebrojava u koliko transakcija se pojavljuju i za svaki 1-skup stavki određuje da li je F ili I.

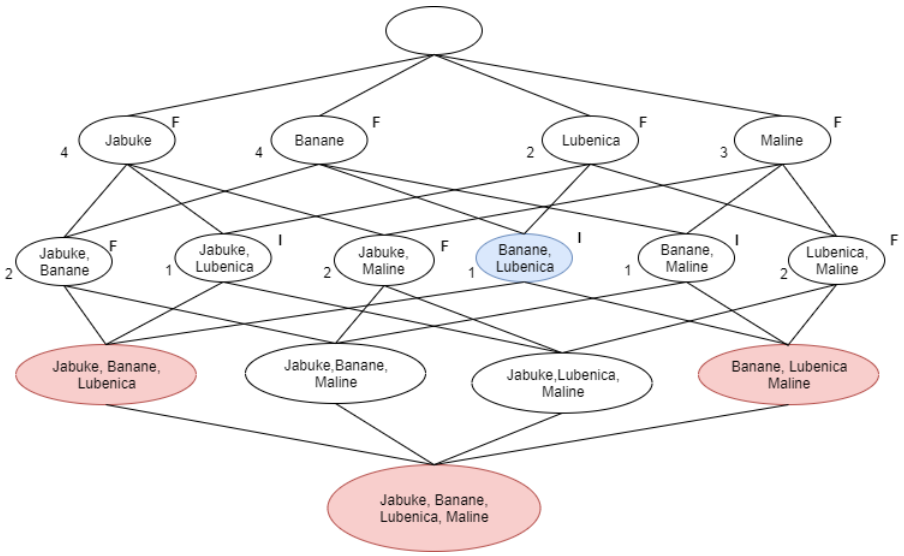


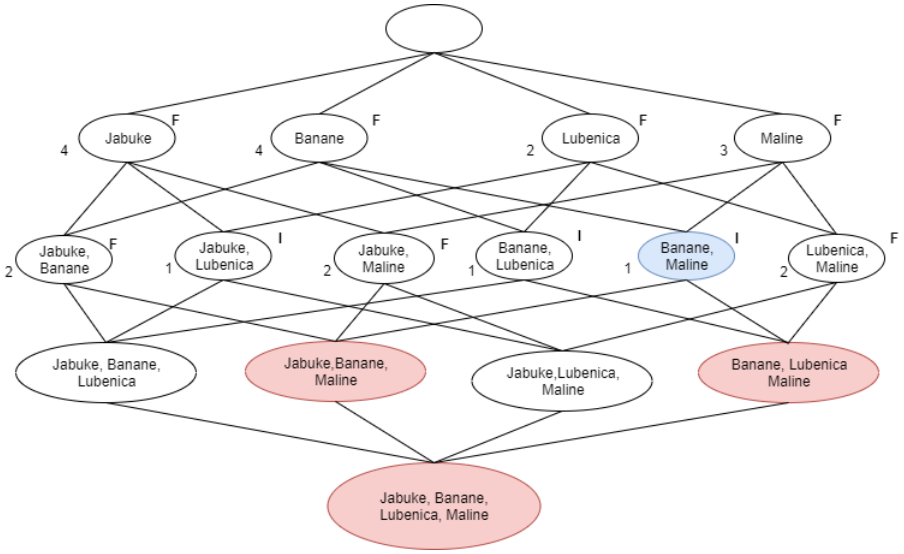
   Pošto su svi 1-skupovi stavki česti, za svaki 2-skup stavki se mora izračunati broj transakcija u kojima se javlja da bi se odredilo da li je skup čest ili ne.



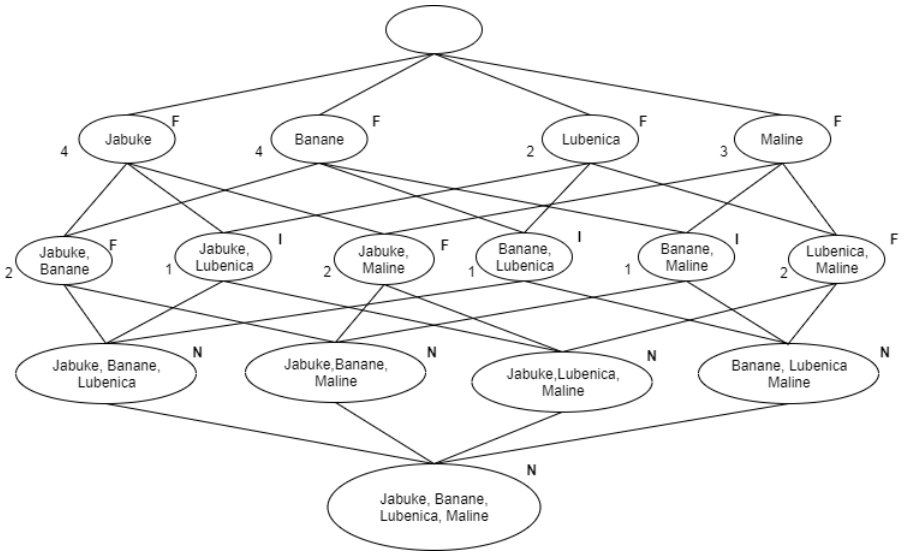
Za 2-skupove stavki koji nisu česti (označeni sa I) mogu se odmah pronaći svi nadskupovi i označiti kao retki, bez računanja broja transakcija u kojima se javljaju. Na narednim rešetkama je po jedan nečest 2-skup stavki označen plavom bojom, a crvenom bojom njegovi nadskupovi za koje se zna da nisu česti bez računanja broja transakcija u kojima se javljaju i koji mogu da se označe sa N.







Pošto svaki 3-skup stavki ima podskup koji nije čest, ni on ne može biti čest. Isto važi i za 4-skup stavki. Rešetka sa pridruženim oznakama za svaki čvor izgleda:



2. Nacrtati rešetku skupova stavki koja odgovara datom skupu podataka. Označiti čvorove u rešetki sledećim slovima:

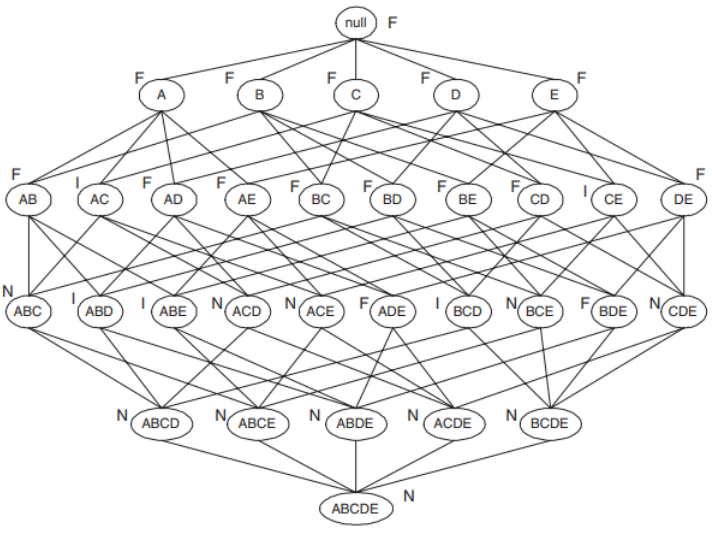
   - N: ako se skup stavki ne smatra kandidatom po Apriori algoritmu. Tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
   - F: kandidat skupa stavki je čest po Apriori algoritmu.
   - I: Ako se skup stakvi smatra retkim posle određivanja podrške.

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	$\{b,d\}$

   Minimalna podrška za česte skupove je 0,3.

- 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)?

Rešenje



Odgovori:

- $\frac{16}{32} = 50\%$
- $\frac{11}{32} = 34,4\%$
- $\frac{5}{32} = 15,6\%$

2 Pravila pridruživanja u SPSS Modeleru

2.1 Format podataka

Skup podataka za određivanje pravila pridruživanja se može zadati u transakcionom ili tabelarnom formatu.

- Transakcionoi format - jedan red u tabeli sadrži podatke o jednoj stavci u jednoj transakciji. Tabela sadrži dve kolone: id transakcije i jedna stavka. Npr. podaci za transakcije {A,B,C}, {A,B}, {B,D} u tabelarnom obliku izgledaju:

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

- Tabelarni format (korpa). Jednoj stavci iz skupa podataka je dodeljen jedan binarni atribut. Jedna transakcija je predstavljena jednim redom. Ukoliko se stavka javlja u transakciji, pridružen atribut te stavke ima vrednost *tačno* u redu koji odgovara toj transakciji, a inače *netačno*.

  Atributi mogu imati dodeljene uloge:

  - *Input* - stavka pridružena atributu može biti samo u telu pravila
  - *Target* - stavka pridružena atributu može biti samo u glavi pravila
  - *Both* - stavka pridružena atributu može se pojaviti u telu ili glavi pravila

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

2.2 Algoritam Apriori

Algoritmu *Apriori* odgovara čvor **Apriori**, koji se nalazi u paleti *Modeling* u delu *Association*.

2.2.1 Parametri u SPSS modeleru, algoritam Apriori

- Odeljak *Fields*
  - *Use predefined roles* - korišćenje uloga atributa koje su dodeljene pre korišćenja čvora Apriori
  - *Use custom field assignments* - zadavanje uloga atributima. Ako nije izabrana opcija *Use transactional format*, podaci su u tabelarnom obliku. U listu *Consequents* se dodaju stavke koje mogu da se pojave u glavi pravila, a u listu *Antecedents* se dodaju stavke koje mogu da se pojave u telu pravila.

    Ako je izabrana opcija *Use transactional format* podaci su u transakcionom obliku i potreno je definisati koji atribut sadrži podatak o identifikaciji transakcije (opcija *ID*), a koji atribu sadrži podatak o stavci koja se javlja u transakciji (opcija *Content*).

- odeljak *Model*
  - *Minimum antecedent support* - minimalna podrška **tela** pravila

    *Obratiti pažnju da je u alatu SPSS Modeler:
    - Podrška (Support): $sup(X) = \frac{\sigma(X)}{N}$
    - Podrška pravila (Rule Support): $sup(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
  - *Minimum rule confidence* - minimalna pouzdanost pravila
  - *Maximum number of antecedents* - maksimalan broj stavki u telu pravila
  - *Only true values for flags* - uzeti u obzir samo vrednosti **tačno** u binarnim atributima. Ako ova opcija nije izabrana može se izdvojiti i pravilo oblika $A{=}T$ & $B{=}F$ $\rightarrow$ $C{=}T$. Ova opcija je dostupna samo za tabelarni oblik.

- odeljak *Expert* - koristi se za zadavanje praga za neku od mera kvaliteta, npr.:
  - Razlika u pouzdanosti (Confidence Difference): $conf_{diff} = conf_{posterior} - conf_{prior}$
    - $conf_{prior}$ je pouzdanost za pravilo $empty \rightarrow Y$
    - $conf_{posterior}$ je pouzdanost za pravilo $X \rightarrow Y$
  - Odnos pouzdanosti (Confidence Ratio): $conf_{ratio} = 1 - min(\frac{conf_{posterior}}{conf_{prior}}, \frac{conf_{prior}}{conf_{posterior}})$

2.3 Model dobijen primenom algoritma Apriori

Model koji se dobija kao rezultat korišćenja čvora **Apriori** je prikazan u obliku dijamanta u radnom toku. Podaci koji su dostupni preko modela su:

- odeljak *Model* prikazuje tabelu sa izdvojenim pravilima pridruživanja. Jedan red u tabeli sadrži podatke o jednom pravilu pridruživanja. Za svako pravilo pridruživanja može se prikazati:
  - *Consequent* - glava (posledica) pravila
  - *Antecedent* - telo (uzrok) pravila
  - *Rule ID* - identifikator pravila
  - *Instances* - broj instanci u skupu podataka koje zadovoljavaju telo pravila

- *Support* - podrška za **telo** pravila
- *Confidence* - pouzdanost
- *Rule Support* - podrška pravila
- *Lift*
- *Deployability* - mogućnost raspoređivanja

Iznad tabele se nalaze dugmići i padajući meni za izbor šta i kako da se prikaže u tabeli:

- *Sort by* - za izbor kolone u tabeli po kojoj će se urediti podaci u opadajućem ili rastućem poretku
- *Show/hide criteria menu* - za izbor kolona koje će biti prikazane u tabeli
- *Show filters* - za zadavanje uslova koje moraju da zadovoljavaju pravila prikazana u tabeli. Moguće je zadati:
  * stavke koje mogu/ne smeju da se pojavljuju u glavi pravila
  * stavke koje mogu/ne smeju da se pojavljuju u telu pravila
  * opseg mogućih vrednosti za mere
    · pouzdanost
    · podrška tela
    · Lift

  Nakon izdvajanja pravila pridruživanja koja zadovoljavaju željene uslove, moguće je napraviti novi model samo sa tim pravilima klikom na dugme *Generate*, a zatim izborom opcije *Filtered Model* u padajućoj listi. Nakon zadavanja imena modela, novi model predstavljen čvorom u obliku dijamanta se dodaje u radni tok.

- odeljak *Settings* se koristi za postavljanje uslova koji moraju da važe pri dodeli najboljih pravila pridruživanja za svaku transakciju. Opcije su:
  - *Maximum number of predictions* - broj pravila koja će biti dodeljena svakom redu u tabeli sa skupom podataka. Pri dodeli pravila biraju se najbolja pravila prema izabranoj meri preko liste *Rule Criterion*.
  - *Allow repeat predictions* - da li je pri izboru najboljih pravila dozvoljeno ponavljanje glave ili ne. Ako jeste, onda je moguće imati više najboljih pravila sa istom glavom, a ako nije onda izdvojena najbolja pravila moraju imati različite glave.
  - *Ignore unmatched basket items* - da li je dozvoljeno povezati pravilo sa transakcijom i ako se u pravili ne pojavljuju sve stavke iz transakcije. Ako je izabrana opcija, onda pravilo *X & Y -> Z* može biti povezano sa transakcijom *{X, Y, W}*.
  - *Check that predictions are not in basket* - glava pravila ne sme da se pojavljuje u transakciji
  - *Check that predictions are in basket* - glava pravila mora da se pojavljuje u transakciji
  - *Do not check basket for predictions* - glava pravila može, a i ne mora da se pojavljuje u transakciji

Klikom na dugme *Preview*, za svaki red se prikazuje zadati broj najboljih pravila prema zadatim uslovima. Za jedno pravilo se prikazuje: glava pravila, vrednost zadate mere za izbor najboljih pravila i ID pravila.

Napomena: u transakcionom obliku se prikazuju podaci za svaki red, iako je jedna transakcija podeljena u više redova. Pri određivanju najboljih pravila za jedan red (jedan red sadrži jednu stavku jedne transakcije), koriste se sve stavke iste transakcije o kojima postoje podaci u redovima iznad tog reda koji se obrađuje. Npr. za transakciju {A, B, C} sa id 1, podaci u transakcionom obliku su

Id transakcije	Stavka
1	A
1	B
1	C

Pri određivanju najboljih pravila, prvo se traže najbolja pravila samo za stavku A pri obradi reda **1,A** , zatim pri obradi reda **1,B** traže se najbolja pravila za stavke A i B, jer je podatak o stavci A za transakciju 1 prethodno učitan. Na kraju, pri obradi reda **1,C**, traže se najbolja pravila za sve stavke transakcije A, B i C.

3 Praktični zadaci u SPSS Modeleru

1. Primenom pravila pridruživanja proveriti da li postoji zavisnost među podacima u skupu *transactions*. Rešenje:

   - Radni tok: **pravila_pridruzivanja_transactions.str**
   - Komentari: **ipVezbe11Primer1.pdf**

2. Primenom pravila pridruživanja proveriti da li postoji zavisnost među upisanim izbornim predmetima.

   Rešenje:

   - Radni tok: **pravila_pridruzivanja_izborni_predmeti.str**
   - Komentari: **ipVezbe11Primer2.pdf**