Istraživanje podataka - zadaci za vežbe 4

Zadaci

1. 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	$\{b,d\}$

   - 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.
   - 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)?

Praktični zadaci

1. Primenom pravila pridruživanja proveriti da li postoji zavisnost među podacima u skupu *transactions*. Diskutovati dobijene rezultate. (*KNIME, SPSS Modeler*)

2. Primenom pravila pridruživanja proveriti da li postoji zavisnost medju upisanim izbornim predmetima. Diskutovati dobijene rezultate. (*SPSS Modeler*)

Mere

Pravila pridruživanja su oblika: $X \to Y$, a broj transakcija je N.

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

  *U alatu SPSS Modeler:

  - Podrška (Support): $sup(X \to Y) = \frac{\sigma(X)}{N}$
  - Podrška pravila (Rule Support): $sup(X \to Y) = \frac{\sigma(X \cup Y)}{N}$
- Pouzdanost (Confidence): $conf(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$
  U literaturi se pojavljuju:
  - $conf_{prior}$ je pouzdanost za pravilo $empty \to Y$
  - $conf_{posterior}$ je pouzdanost za pravilo $X \to Y$
- Lift: $Lift = \frac{conf(X \to Y)}{sup(Y)}$ ili $Lift = \frac{conf_{posterior}}{conf_{prior}}$
  * Pravilo $X \to Y$ je zanimljivo ako je $Lift(X \to Y) \neq 1$
- Razlika u pouzdanosti (Confidence Difference): $conf_{diff} = conf_{posterior} - conf_{prior}$
- Odnos pouzdanosti (Confidence Ratio): $conf_{ratio} = 1 - \frac{min(conf_{posterior}, conf_{prior})}{min(conf_{posterior}, conf_{prior})}$
- Obim raspoređivanja (Deployability): $\frac{\sigma(X) - \sigma(X \cup Y)}{N}$