DATA MINING

Reguli de asociere

- Reguli de asociere = asocieri (relatii, dependente) interesante in seturi mari de date

- **Market Basket Analysis**
  - una dintre cele mai intuitive aplicatii ale regulilor de asociere

  Ex: se analizeaza cosul de cumparaturi al fiecarui client care face cumparaturi la un anumit magazin, intr-o anumita perioada de timp
  - Se observa ca apare o regula de asociere de forma: cine cumpara paine, de obicei cumpara si lapte
  - o astfel de observatie ar putea fi folosita pentru maximizarea profitului:
    - Nu se va oferi reducere la ambele produse in acelasi timp
    - Produsele vor fi asezate pe rafturi in capete opuse ale magazinului, pentru a obliga prin asta clientii sa vada mai multe produse si probabil sa cumpere mai mult

- Notatii:
  - D – setul de tranzactii
  - I – setul de produse (items)
  - Fiecare tranzactie T reprezinta un set de produse din I
  - Fie A si B seturi de items (ex: A={rosii, fasole}, B={ceapa})
- O regula de asociere are forma: if A then B (A=>B)
- Nu exista item-uri care sa apartina si lui A si lui B

- Pentru a masura calitatea unei reguli se foloseste suportul si confidenta

- **SUPORT**
  - Proportia de tranzactii din D care contin si A si B

$$\text{support} = P(A \cap B) = \frac{\text{number of transactions containing both } A \text{ and } B}{\text{total number of transactions}}.$$

- **CONFIDENTA**
  - Procentul de tranzactii din D care contin A si contin si B

$$\text{confidence} = P(B|A) = \frac{\text{number of transactions containing both } A \text{ and } B}{\text{number of transactions containing } A}.$$

- O regula de asociere este interesanta daca satisface valori minime pentru suport si confidenta (aceste praguri trebuie specificate de catre un expert in domeniul aplicatiei, depinzand de domeniul aplicatiei)

- Exemplu:
  - daca doresti sa determini ce produse sunt cumparate impreuna, se poate stabili un suport minim de 20% si o confidenta minima de 70%
    - Cumpara (x, lapte)=>cumpara (x, paine) [20%, 70%]
  - un analist pentru descoperirea fraudelor sau a teroristilor va stabili probabil un suport minim de 1%

For the given transactions, find all rules such that LHS $= \{A, B\}$, RHS $= \{C\}$, with minimum support $= 50\%$ and minimum confidence $= 50\%$.

TID	Transactions
1000	A, B, C
2000	A, C
3000	A, D
4000	B, E, F

Support is the probability that a transaction contains $\{A, B, C\}$, and *confidence* is the conditional probability that a transaction containing $\{A, B\}$ also contains C.
Rule $A \wedge B \Rightarrow C$ [support 25%, confidence 100%] does not satisfy the minimum confidence. Two (shorter) strong association rules are generated as:

$A \Rightarrow C$ [support 50%, confidence 66.6%]
$C \Rightarrow A$ [support 50%, confidence 100%]

Referinta figura: K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, *Data Mining. A Knowledge Discovery Approach*, Springer, 2007.

Tipuri de reguli de asociere

- Unidimensionale
  - cumpara (x, lapte)=>cumpara (x, paine) [25%, 60%]
- Multidimensionale
  - licenta (x, informatica) SI curs_optional (x, Data Mining) => nivel_studiu (x, doctorat) [1%, 75%]

- Booleene
- Cantitative

- Pe un singur nivel
- Pe mai multe niveluri
  - Exemplu: cumpara (x, lapte slab) => cumpara (x, paine neagra) [2.5%, 60%]

- Exemplu de regula multidimensionala cantitativa:
  - varsta (x, “18-25”) SI venit (x, “<1000”) => cumpara (x, paine) [0.5%, 50%]
  - Variabilele continue (varsta, venit) sunt discretizate

Metode pentru generarea regulilor de asociere unidimensionale, pe un singur nivel, booleene

- **itemset** = set de items
- **k-itemset** = set de k items (ex: {Bere, Oua} este un 2-itemset)
- **Frecventa unui itemset (support count)** = numarul de tranzactii din D care contin acel itemset
- un itemset este frecvent daca
    - Frecventa >=suport minim *nr total tranzactii in D

(exista definitii alternative)

- **Algoritm naiv**

  m items si n tranzactii =>2^m itemsets =>

  =>O(2^m *n) teste

  (complexitatea creste exponential cu numarul de items)

- **Algoritm Apriori**
  - Proprietatea apriori: toate subseturile nevide ale unui itemset frecvent sunt de asemenea frecvente

    SAU
  - Daca un itemset nu este frecvent, atunci orice supraset nu este frecvent

1. Determin toate 1-itemsets
2. Determin L1, multimea acelor 1-itemsets frecvente
3. Genereaza 2-itemsets din L1
4. Determina L2, multimea acelor 2-itemsets frecvente
5. …

- La fiecare iteratie, acele k-itemsets care nu satisfac suportul minim sunt scoase si doar cele care raman sunt folosite pentru a genera itemsets pentru iteratia k+1

- Generarea Ck=multimea de k-itemsets bazata pe L(k-1)

1. Pentru fiecare itemset frecvent Fi din L(k-1), cauta fiecare item i care nu apartine lui Fi dar apartine altor (k-1)-itemsets frecvente in L(k-1)
   - Adauga i la Fi pentru a crea un k-itemset
   - Se sterg k-itemsets duplicate
2. Daca (k-1)-itemsets frecvente din L(k-1) au (k-2) items in comun, atunci se creeaza un k-itemset adaugand cei 2 items diferiti la cei doi (k-2) items comuni

- Generarea regulilor de asociere din itemsets frecvente
  - Suportul minim este asigurat prin folosirea itemset-urilor frecvente
  - Pentru confidenta: fiecare itemset frecvent Fi este folosit pentru a genera reguli de asociere
    - Se genereaza toate subseturile nevide Y ale lui Fi
    - Pentru fiecare Y, se calculeaza confidenta pentru regula “Y=>(Fi-Y)”

- Alte masuri pentru calitatea regulilor de asociere (suportul si confidenta sunt masuri obiective)
  - Masura de corelatie (trebuie sa fie pozitiv corelate)

$$\text{correlation}(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

Daca corelatia>1 => pozitiv corelate (cresterea uneia determina cresterea celeilalte)

Daca corelatia<1 => negativ corelate (aparitia uneia inhiba aparitia celeilalte)

Daca corelatia=1 => independente (nu exista corelatie intre ele)