

# TEHNICI DE EXPLORARE A DATELOR -DATA MINING -

CURS 8

# Reguli de asociere

- aparute în 1993 [AIS93]
  - au devenit una din cele mai populare metode de analiză în procesul descoperirii de cunoștințe din bazele de date
- originea în analiza datelor referitoare la coșul de piață, unde sunt generate reguli de forma :
  - “*Un client care cumpără produsele  $x_1, x_2, \dots, x_n$  va cumpăra de asemenea produsul  $y$  cu probabilitatea  $c\%$* ”.
- nu se limitează la analiza dependențelor din aplicațiile de vânzare cu amănuntul,
  - sunt aplicabile cu succes unei game mult mai largi de probleme,
    - experimente științifice
    - monitorizarea unor sisteme fizice precum rețelele de telecomunicații
    - domeniul medical
    - stabilirea profilului clientilor

# *Reguli de asociere*

- formă de explorare a datelor care își propune descoperirea de legături interesante între attribute din datele conținute în baze sau depozite de date.
- cerințe majore:
  - eficiență,
  - scalabilitate,
  - utilitate și
  - trebuie să fie ușor înțelese.
- esența cercetărilor în ceea ce privește explorarea regulilor de asociere constă în găsirea celor mai rapide și eficiente metode de determinare a regulilor de asociere care să aducă un plus consistent de informație

## Definirea formală a problemei

- $I = \{I_1, \dots, I_m\}$  o mulțime de attribute numite articole
- O submulțime  $X$  de  $k$  articole a mulțimii  $I$  se numește mulțime de articole de dimensiune  $k$  (k-itemset ).
- Fie baza de date  $D = \{T_1, T_2, \dots, T_n\}$ , o mulțime de seturi de tranzacții,
  - fiecare tranzacție  $T_i$ ,  $i \in \{1, \dots, n\}$  este o mulțime de articole.
  - o tranzacție  $T$  conține o mulțime de articole  $X$  dacă  $X \subseteq T$ .
- Fiecare mulțime de articole are asociată o anumită semnificație statistică numită *suport* sau *frecvență* (fracțiunea tranzacțiilor din baza de date  $D$ , care conține setul  $X$ ):

$$s(X) = \frac{|\{T \in D \mid X \subseteq T\}|}{|D|}$$



## Definirea formală a problemei - continuare

- O regulă de asociere - o implicație de forma  $X \rightarrow Y$ , unde  $X \subset I$ ,  $Y \subset I$  și  $X \cap Y = \emptyset$ .
- $X$  - *antecedentul* regulii
- $Y$  - *consecventul*
- Regula are *gradul de încredere (confidența)*  $c$  - fracțiunea tranzacțiilor care conțin  $X$  și de asemenea conțin  $Y$ , din totalul tranzacțiilor care îl conțin pe  $X$ .

$$c(X, Y) = \frac{s(X \cup Y)}{s(X)} = \frac{|\{T \in D \mid X \cup Y \subseteq T\}|}{|\{T \in D \mid X \subseteq T\}|}$$

- Suportul unei reguli  $X \rightarrow Y$  este definit ca:  
$$sup(X \rightarrow Y) = sup(X \cup Y)$$
- Pentru cele doua mărimi statistice se impun praguri minime
  - minsup
  - minconf
- Fiind dată o mulțime de articole  $I$ , o bază de date  $D$ , alcătuită din mulțimi de articole, și valorile pragurilor minime pentru suport respectiv pentru confidență (*minsup*, *minconf*), să se găsească toate regulile de forma  $X \rightarrow Y$  din  $D$ , care au suportul  $s(X \cup Y) \geq \text{minsup}$  și confidența  $c(X, Y) \geq \text{minconf}$ .
  - *reguli de asociere booleene*

# Exemplu

TID	Articole
1	paine, lapte
2	Bere, alune, paine, oua
3	Bere, fursecuri, alune, lapte
4	Bere, paine, alune, lapte
5	fursecuri, paine, alune, lapte

Regula: {alune, lapte} → bere

$$\text{Suport} = \frac{|(\text{alune, lapte, bere})|}{|D|} = 0.4$$

$$\text{Confidenta} = \frac{\text{sup}(\text{alune, lapte, bere})}{\text{sup}(\text{alune, lapte})} = 0.66$$

# O analiză critică a măsurilor pentru determinarea și evaluarea regulilor de asociere

- Regulile care se pot extrage din tabelele relaționale sunt de forma
  - $C1 \rightarrow C2$ 
    - unde  $C1$  și  $C2$  sunt condiții asupra tuplurilor din tabelele respective.
- Aceste reguli pot fi:
  - *exacte*, ceea ce presupune că toate tuplurile care satisfac  $C1$  satisfac de asemenea și  $C2$ ;
  - *puternice*, atunci când majoritatea tuplurilor care satisfac  $C1$  satisfac și  $C2$  ;
  - *aproximative* dacă doar o mică parte din tuplurile care satisfac  $C1$  satisfac și  $C2$ .



- initial fost realizat un *model suport-confidență* pentru explorarea regulilor de asociere din bazele de date.
  - este utilizat în general pentru depistarea anumitor tipuri de dependențe între articolele reprezentate într-o bază de date.
  - modelul măsoară **nesiguranța** unei reguli de asociere prin doi factori: *suport și confidență*.
    - nu furnizează un mod de evidență al corelației între două mulțimi de articole.
    - suportul este limitat la un **rol pur informativ** deoarece el reprezintă numărul tranzacțiilor care conțin un anumit itemset dar nu și numărul articolelor din set.
- au fost propuse noi măsuri, legate tot de suportul și confidența regulilor de asociere, dar care conduc la modele diferite de explorare a regulilor de asociere.



# Dezavantajele celor două măsuri clasice

- Confidența reprezintă o *măsură a preciziei* unei reguli
- Piatetsky- Shapiro susține că orice măsură pentru precizie (MP) ar trebui să verifice trei proprietăți specifice pentru a separa regulile puternice de cele slabe
- **P1.**  $MP(X \rightarrow Y) = 0$  când  $\sup(X \rightarrow Y) = \sup(X) * \sup(Y)$ . Această proprietate cere ca orice măsură pentru precizie să testeze independența.
- **P2.**  $MP(X \rightarrow Y)$  să crească monoton în raport cu  $\sup(X \rightarrow Y)$  atunci când alți parametri rămân neschimbați.
- **P3.**  $MP(X \rightarrow Y)$  să descrească monoton funcție de  $\sup(X)$  sau de  $\sup(Y)$  când ceilalți parametri rămân constanți
- *confidența nu verifică toate aceste proprietăți*

## 1. Confidența nu verifică proprietatea P1.

$i_1$	$i_2$	$i_3$	$i_4$	itemset	Support
1	0	1	0	{i1}	1/2
0	0	0	1	{i2}	2/3
0	1	1	1	{i1,i2}	1/3
0	1	1	1		
1	1	1	1		
1	1	1	1		

Exemplu de itemseturi

Suportul a trei itemseturi

- $\text{sup}(\{i1\}) * \text{sup}(\{i2\}) = 1/3 = \text{sup}(\{i1,i2\})$ 
  - $\rightarrow i1$  și  $i2$  sunt independente statistic
    - prin urmare, confidența unei reguli care conține doar cele doua itemuri ar trebui să fie 0.
  - $\text{conf}(\{i1\} \rightarrow \{i2\}) = (1/3) / (1/2) = 2/3 \neq 0.$

- 2. Confidența verifică proprietatea P2
- P3 este verificată doar pentru  $\text{sup}(X)$ , și nu este verificată și pentru  $\text{sup}(Y)$  - acesta nu apare deloc în definiția confidenței.
- $\Rightarrow$  confidența nu este aptă să detecteze nici **independența statistică**, nici **dependențele negative** dintre articole deoarece nu ia în considerare și suportul pentru consecventul unei reguli.
- Un principiu acceptat în explorarea regulilor de asociere este acela că, *un itemset este cu atât mai „bun” cu cât suportul său este mai ridicat.*
  - valabil numai parțial
  - itemset-urile cu suport foarte ridicat pot fi și o sursă a regulilor înșelătoare.



# Mărimi suplimentare pentru măsurarea calității regulilor

- s-a stabilit că o regulă  $X \rightarrow Y$  nu este de interes dacă
$$\text{sup}(X \rightarrow Y) \approx \text{sup}(X) * \text{sup}(Y)$$
- In concordanță cu interpretarea în termenii teoriei probabilităților :
- $\text{Sup}(X \cup Y) = P(X \cup Y)$  și
$$\text{Conf}(X \cup Y) = P(X|Y) = P(X \cup Y) / P(X)$$
- In aceste condiții argumentarea lui Piatetski-Shapiro poate fi scrisă astfel:

$$P(X \cup Y) \approx P(X) P(Y)$$

- $\Rightarrow X \rightarrow Y$  nu poate fi extrasă ca o regulă dacă
$$P(X \cup Y) \approx P(X) P(Y).$$

(de fapt, și în teoria probabilităților relația  $P(X \cup Y) \approx P(X) P(Y)$  indică faptul că  $X$  și  $Y$  sunt independente)

- O definiție statistică a dependenței între mulțimile  $X$  și  $Y$  este dată o nouă măsura denumită *interes sau lift*, a cărei expresie este:

$$\text{int}(X, Y) = \frac{P(X \cup Y)}{P(X)P(Y)}$$

cu posibila extensie pentru mai mult de două mulțimi

- exprimă **deviația gradului de încredere al regulii de la probabilitatea apriori pentru  $Y$ ,  $\text{sup}(Y)$ .**
- In termenii măsurilor clasice pentru regulile de asociere interesul are expresia:

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

- Cu cât valoarea interesului este mai depărtată de 1 cu atât dependența dintre mulțimile de articole este mai mare.

- Similar celorlalte măsuri, se poate stabili o valoare de prag numită *minint* ( $1 > \text{minint} > 0$ )
- dacă 
$$\left| \frac{P(X \cup Y)}{P(X)P(Y)} - 1 \right| \geq \text{minint}$$
- atunci se poate spune despre regula  $X \rightarrow Y$  că este interesantă.
- se pot considera trei cazuri distincte:
  - dacă  $P(X \cup Y)/P(X)*P(Y) = 1$  atunci  $P(X \cup Y) = P(X)*P(Y)$  și  $X$  și  $Y$  sunt *independente*.
  - dacă  $P(X \cup Y)/P(X)*P(Y) > 1$  sau  $P(X \cup Y) > P(X)*P(Y)$  se spune că  $Y$  *depinde pozitiv de X*
  - dacă  $P(X \cup Y)/P(X)*P(Y) < 1$  sau  $P(X \cup Y) < P(X)*P(Y)$  se spune că  $Y$  *este negativ dependent de X* sau ca  $\neg Y$  *este dependent pozitiv de X*.



- se poate defini o altă formă de interpretare a regulilor de interes, astfel:
  - dacă  $\left| \frac{P(X \cup Y)}{P(X)P(Y)} - 1 \right| \geq \text{minint}$  atunci  $X \rightarrow Y$  este o regulă interesantă.
  - dacă  $\left| \frac{P(X \cup Y)}{P(X)P(Y)} - 1 \right| \leq \text{minint}$  atunci  $X \rightarrow \neg Y$  este o regulă de interes

# Definitia regulilor de asociere interesante

- Fie  $I$  o mulțime de articole în baza de date  $D$ .
- Fie  $X, Y \subseteq I$ , mulțimi de articole astfel încât:
  - $X \cap Y = \emptyset$ ,
  - $P(X) \neq 0$  și  $P(Y) \neq 0$ ,
- fie valorile minime de prag:
  - $\text{minsup} > 0$ ,
  - $\text{minconf} > 0$  și
  - $\text{minint} > 0$ , date de către utilizatori sau experți.
- Regula  $X \rightarrow Y$  poate fi extrasă ca o regulă validă și interesantă dacă:
  - $P(X \cup Y) \geq \text{minsup}$
  - $P(X|Y) \geq \text{minconf}$
  - $|P(X \cup Y) - P(X)P(Y)| \geq \text{minint}$  sau  $\left| \frac{P(X \cup Y)}{P(X)P(Y)} - 1 \right| \geq \text{min int}$