

Föreläsning 9 - Associationsanalys

Josef Wilzén

Agenda

- Introduktion
- Associationsanalys
- Algoritmer
- Intresse mått

Introduktion

Introduktion

- Målet med associationsanalys:
 - Utvinna intressanta samband (el. mönster) som finns i stora datamängder
- Exempel: Shoppingtransaktioner
 - En återförsäljare är intresserad av att veta sina kunders köpbeteende och vill använda det i marknadsföringen.

$\{Diapers\} \rightarrow \{Beer\}$

Table 6.1. An example of market basket transactions.

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Introduktion

- Andra tillämpningsområden:
 - Bioinformatik
 - Medicinsk diagnos
 - Web mining
 - Vetenskaplig dataanalys

Representation av data

TI D	Item
1	Bread
1	Milk
2	Bread
2	Diapers
2	Beer
2	Eggs
3	Milk
...	...

Table 6.2. A binary 0/1 representation of market basket data.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Definitioner

- Supportnivå (σ):

- Låt:

- $I = \{i_1, i_2, \dots, i_d\}$ är attributmängden (enheter i materialet),

- $T = \{t_1, t_2, \dots, t_N\}$ är objektmängden (transaktioner),

- X är en godtycklig delmängd av I

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \subseteq T\}|$$

- Kan beskrivas som antalet transaktioner (t_i) i objektmängden (T) som innehåller X .

Definitioner

- Associationsregel (R) är ett uttryck som har följande form:

$$X \rightarrow Y$$

där X och Y är disjunkta enhetsmängder enligt $X \cap Y = \emptyset$

- Två mått på styrkan av en regel är **support** och **konfidens**

Definitioner

- Support (s)

$$s(X \rightarrow Y) = p(X, Y) = \frac{\sigma(X \cup Y)}{N}$$

- Konfidens (c)

$$c(X \rightarrow Y) = p(Y|X) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Utvinning av regler

- Givet en transaktionsmängd, hitta regler som har support $\geq \text{minsup}$ och konfidens $\geq \text{minkonf}$
 - Dessa trösklar finns till för att ge en ”objektiv” bedömning av intressanta mått
- Problemet: Att använda sig av Brute-Force är inte genomförbart då komplexiteten ökar exponentiellt med antalet enheter enligt:

$$|R| = 3^d - 2^{d+1} + 1$$

där d = antalet enheter

Utvinning av regler

- Många av dessa regler är också icke-intressanta då även för små datamängder uppfyller ungefär 80% inte vanliga konfidens- och supportgränser
- Notera att reglerna $\{A, B\} \rightarrow \{C\}$, $\{A\} \rightarrow \{B, C\}$ och $\{A, C\} \rightarrow \{B\}$ har samma support. Detta medför att problemet kan delas upp i:
 1. Framkalla frekventa enhetsmängder
 - Hitta alla möjliga X som uppfyller $s(X) \geq \text{minsup}$
 2. Framkalla intressanta regler
 - Utvinna regler ur alla X från steg 1 enligt $X_1 \rightarrow X_2$ där $X_1 \cup X_2 = X$ och $X_1 \cap X_2 = \emptyset$ samt uppfyller $c(X_1 \rightarrow X_2) \geq \text{minconf}$

Exempel

- Transaktioner från en stormarknad

	items	transactionID
[1]	{rice,tomato souce,tunny,water}	460202000107
[2]	{tomato souce}	460202000213
[3]	{brioches}	460202000312
[4]	{brioches,tunny,water,yoghurt}	460202000404
[5]	{biscuits}	460202000671
[6]	{frozen fish}	460202000893
[7]	{brioches,tunny,yoghurt}	460202001036
[8]	{coffee,coke,frozen fish,tunny}	460202001067
[9]	{biscuits}	460202001098
[10]	{biscuits,frozen vegetables,rice}	460202001142

Exempel

- Associationsregler:

```
> inspect(head(rules, n = 5))
```

	lhs	rhs	support	confidence	lift	count
[1]	{crackers}	=> {biscuits}	0.05549024	0.5989848	1.381498	236
[2]	{crackers}	=> {water}	0.05219845	0.5634518	1.207234	222
[3]	{frozen fish}	=> {frozen vegetables}	0.05619563	0.5759036	1.866858	239
[4]	{frozen fish}	=> {water}	0.05478486	0.5614458	1.202936	233
[5]	{juices}	=> {brioches}	0.10463202	0.5604534	1.650698	445

- Tolkning: Om en kund köper crackers, köper denne också biscuits i ca. 59.8 procent av fallen.

Algoritmer

Frekventa enhetsmängder

- Aprioriprincipen:

Om en enhetsmängd är frekvent, då kommer alla dess delmängder vara frekventa.

alternativt:

Om enhetsmängd inte är frekvent, då kommer alla dess supermängder inte heller vara frekventa

- Den senare formuleringen är grunden till att stora delar av enhetsmängderna kan hoppas över

Frekventa enhetsmängder

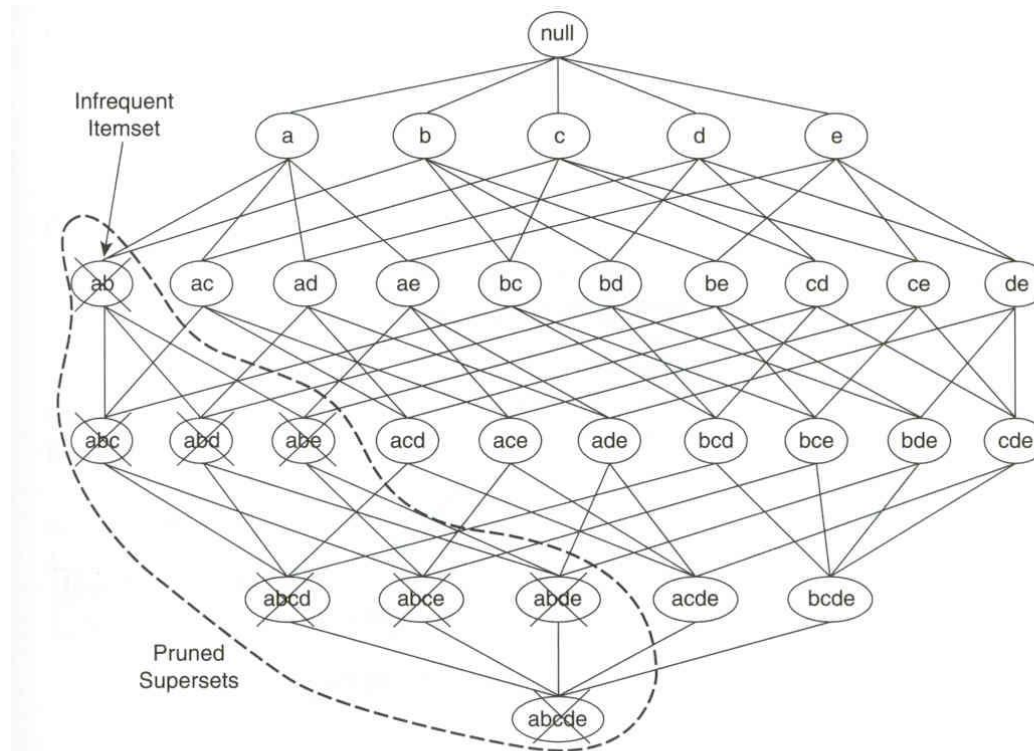


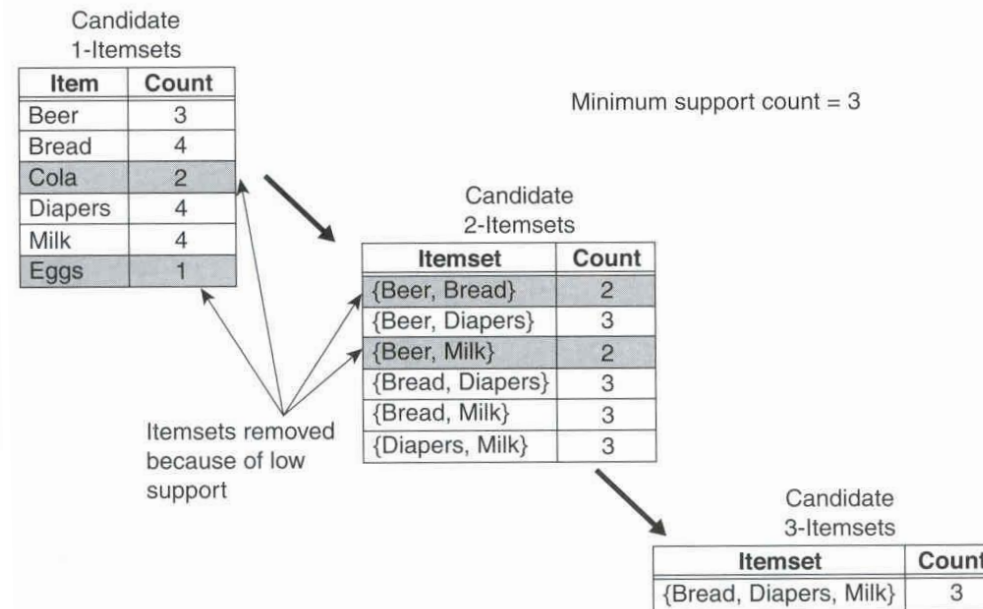
Figure 6.4. An illustration of support-based pruning. If $\{a, b\}$ is infrequent, then all supersets of $\{a, b\}$ are infrequent.

Frekventa enhetsmängder

- Låt oss utnyttja Aprioriprincipen med en supporttröskel på 60%.

Table 6.2. A binary 0/1 representation of market basket data.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1



Apriorialgoritmen

Algorithm 6.1 Frequent itemset generation of the *Apriori* algorithm.

```

1:  $k = 1$ .
2:  $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$ .    {Find all frequent 1-itemsets}
3: repeat
4:    $k = k + 1$ .
5:    $C_k = \text{apriori-gen}(F_{k-1})$ .    {Generate candidate itemsets}
6:   for each transaction  $t \in T$  do
7:      $C_t = \text{subset}(C_k, t)$ .    {Identify all candidates that belong to  $t$ }
8:     for each candidate itemset  $c \in C_t$  do
9:        $\sigma(c) = \sigma(c) + 1$ .    {Increment support count}
10:    end for
11:  end for
12:   $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$ .    {Extract the frequent  $k$ -itemsets}
13: until  $F_k = \emptyset$ 
14: Result =  $\bigcup F_k$ .
```

Beräkningskomplexitet

- Följande faktorer påverkar komplexiteten:
 - Support (lägre *minsup* → fler frekventa enhetsmängder)
 - Antalet enheter (dimensionalitet)
 - Antalet transaktioner
 - Genomsnittliga transaktionsvidden (sparsity i binär framställning)

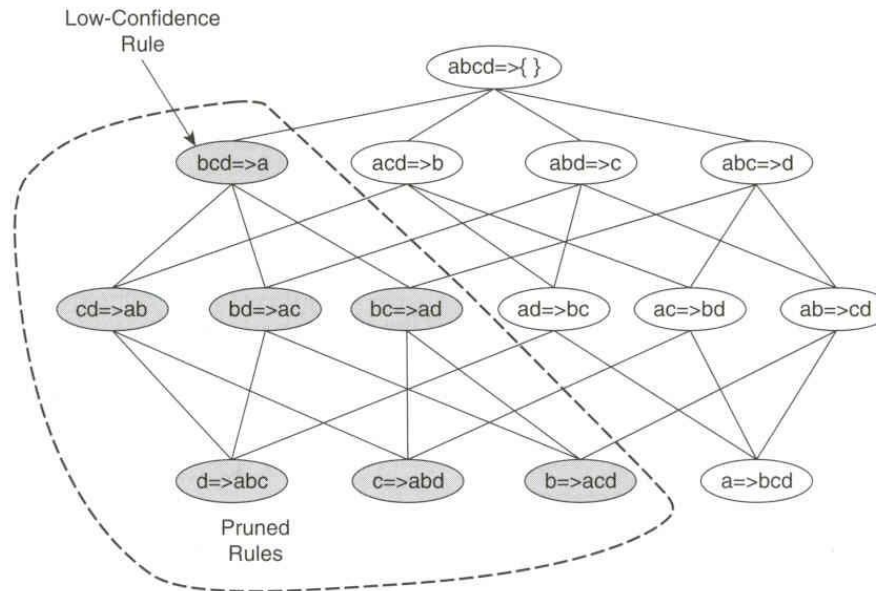
Regelframkallning

- Låt Y vara en frekvent enhetsmängd från L_k .
 - Alla möjliga regler som kan utvinnas följer formen:
 $X_i \rightarrow (Y - X_i)$ där $X_i \subset Y$.
 - Eftersom Y är frekvent, så är X och $(Y - X)$ frekventa. Det betyder att vi vet dess support från tidigare L_n .
 - Sålunda, vi kan beräkna konfidensen:

$$c(X_i \rightarrow (Y - X_i)) = \frac{\sigma(Y)}{\sigma(X_i)}$$

Regelframkallning

- Aprioriprincipen tillämpad på konfidens kan beskrivas som:
 - Om $X \rightarrow (Y - X)$ inte uppfyller konfidenströskeln, då kommer $X' \rightarrow (Y - X')$ inte heller göra det om $(X' \subset X)$



Regelframkallning

- Pseudokod för algoritmen:
 1. Låt Y vara en frekvent enhetsmängd
 2. Utvinn enhetsmängder av storlek 1 ur Y . Låt oss nämna dem som E_{i1} .
 3. Beräkna konfidensen av reglerna som ser ut $(Y - E_{i1}) \rightarrow E_{i1}$. Kasta dem som har låg konfidens. Nu innehåller E_{i1} endast högkonfidenta regler.
 4. Utvinn enhetsmängder av storlek 1 ur Y och slå de samman till E_{i1} vilket leder till E_{i2} .
 5. Repetera steg 3 med E_{i2} .
 6. osv.

Utvärdering av regler

Intressanta och ointressanta regler

- Regler som uppfyller trösklarna kan vara av två sorter
 - Ex. $\{Bread\} \rightarrow \{Butter\}$ och $\{Diapers\} \rightarrow \{Beer\}$
- De verkliga kommersiella databaser är väldigt stora och kan sluta med tusentals eller miljontals associationsregler, även med väldigt höga support- och konfidenströsklar.
- Att utforska dem alla är inte möjligt. Istället används:
 - Subjektiva intressemått
 - Objektiva intressemått

Objektiva intressemått

- Ett statistiskt resonemang man kan utgå ifrån är att “Mönster som innehåller oberoende mängder är ointressanta.”
- Mått som vi redan har använt är:
 - Support
 - Konfidens
- Nu vill vi bygga vidare med fler mått

Objektiva intressemått

- Förväxlingsmatris (Confusion matrix)

Table 6.7. A 2-way contingency table for variables A and B .

	B	\overline{B}	
A	f_{11}	f_{10}	f_{1+}
\overline{A}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Objektiva intressemått

- Lift:

$$Lift(A \rightarrow B) = \frac{c(A \rightarrow B)}{s(B)} = \frac{s(A, B)}{s(A) * s(B)} = \frac{P(A, B)}{P(A) * P(B)}$$

- För binära variabler används Intressefaktor:

$$I(A, B) = \frac{N f_{11}}{f_{1+} f_{+1}}$$

$$I(A, B) \begin{cases} = 1, & \text{if } A \text{ and } B \text{ are independent;} \\ > 1, & \text{if } A \text{ and } B \text{ are positively correlated;} \\ < 1, & \text{if } A \text{ and } B \text{ are negatively correlated.} \end{cases}$$

Exempel

- Intressanta regler med avseende på Lift:

```
> inspect(head(sort(rules, by = "lift"), n = 5))
```

	lhs	rhs	support	confidence	lift	count
[1]	{coffee,juices}	=> {brioches}	0.05713614	0.6567568	1.934340	243
[2]	{frozen fish}	=> {frozen vegetables}	0.05619563	0.5759036	1.866858	239
[3]	{biscuits,juices}	=> {brioches}	0.06912767	0.6282051	1.850247	294
[4]	{juices,water}	=> {brioches}	0.06395486	0.6167800	1.816597	272
[5]	{tomato souce,tunny}	=> {frozen vegetables}	0.05219845	0.5536160	1.794610	222

Objektiva intressemått

- Andra intressemått:

Table 6.11. Examples of symmetric objective measures for the itemset $\{A, B\}$.

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$
Odds ratio (α)	$(f_{11} f_{00}) / (f_{10} f_{01})$
Kappa (κ)	$\frac{N f_{11} + N f_{00} - f_{1+} f_{+1} - f_{0+} f_{+0}}{N^2 - f_{1+} f_{+1} - f_{0+} f_{+0}}$
Interest (I)	$(N f_{11}) / (f_{1+} f_{+1})$
Cosine (IS)	$(f_{11}) / (\sqrt{f_{1+} f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11} + f_{00}}{f_{1+} f_{+1} + f_{0+} f_{+0}} \times \frac{N - f_{1+} f_{+1} - f_{0+} f_{+0}}{N - f_{11} - f_{00}}$
Jaccard (ζ)	$f_{11} / (f_{1+} + f_{+1} - f_{11})$
All-confidence (h)	$\min \left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}} \right]$

Länkar

- <http://r-statistics.co/Association-Mining-With-R.html>
- <https://datascienceplus.com/visualize-market-basket-analysis-in-r/>
 - Network plots

www.liu.se