

Big Data Analytics
– Week 5: Association Rules –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Association Rules: Basic Notions

Items $I = \{i_1, i_2, \dots, i_m\}$: a set of literals (denoting items)

Itemset X : Set of items $X \subseteq I$

Database D : Set of *transactions* T , each transaction is a set of items $T \subseteq I$

Transaction T *contains* an itemset X : $X \subseteq T$

The *support* of an itemset X is defined as: $\text{support}(X) = |\{T \in D | X \subseteq T\}|$

Frequent itemset: an itemset X is called frequent iff $\text{support}(X) \geq \text{minSup}$

Association rule: An association rule is an implication in the form $A \Rightarrow B$ where $A \subset I$ and $B \subset I$ are two itemsets with $A \cap B = \emptyset$

Note: simply enumerating all possible association rules is not meaningful!
→ What are the interesting association rules w.r.t. D ?

Interestingness of Association Rules

Interestingness of an association rule:

Quantify the interestingness of an association rule with respect to a transaction database D :

- Support: frequency (probability) of the entire rule with respect to D

$$\text{support}(A \Rightarrow B) = P(A \cup B) = \frac{|\{T \in D | A \cup B \subseteq T\}|}{|D|} = \text{support}(A \cup B)$$

“probability that a transaction in D contains the itemset $A \cup B$ ”

- Confidence: indicates the strength of implication in the rule

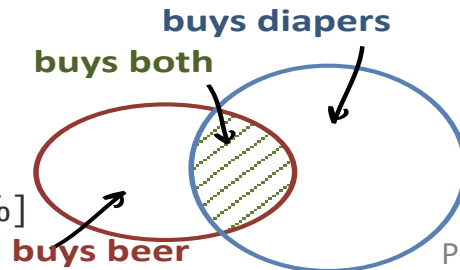
$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{|\{T \in D | A \cup B \subseteq T\}|}{|\{T \in D | A \subseteq T\}|} = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

“conditional probability that a transaction in D containing the itemset A also contains itemset B ”

- Rule form: “*Body* $\Rightarrow$ *Head* [*support*, *confidence*]”

Association rule examples:

- buys diapers $\Rightarrow$ buys beers [0.5%, 60%]
- major in CS $\wedge$ takes DB $\Rightarrow$ avg. grade A [1%, 75%]



Mining of Association Rules

Given a database D , determine all association rules having a $support \geq minSup$ and a $confidence \geq minConf$ (so-called *strong association rules*).

Key steps of mining association rules:

- e.g. *Apriori*, *FP-growth* 
- 1) Find *frequent itemsets*, i.e., itemsets that have at least $support = minSup$
 - 2) Use the frequent itemsets to generate association rules
 - n frequent items yield $2^n - 2$ association rules

Example

transaction ID	items
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

$minSup = 50\%$



frequent items	support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

For $minSup = 50\%$ and $minConf = 50\%$ we obtain 2 rules:

- $A \Rightarrow C$: $support(A \Rightarrow C) = 50\%$, $confidence(A \Rightarrow C) = 66.67\%$
- $C \Rightarrow A$: $support(C \Rightarrow A) = 50\%$, $confidence(C \Rightarrow A) = 100\%$

Generating Rules

For each frequent itemset X

- For each nonempty subset Y of X , form a rule $Y \Rightarrow (X - Y)$
- Delete those rules that do not have minimum confidence

Computation of the confidence of a rule $Y \Rightarrow (X - Y)$

$$\text{confidence}(Y \Rightarrow (X - Y)) = \frac{\text{support}(X)}{\text{support}(Y)}$$

Store the frequent itemsets and their support in a hash table in main memory $\rightarrow$ no additional database access

Example: $X = \{A, B, C\}$, $\text{minConf} = 60\%$

- $\text{conf}(A \Rightarrow B, C) = 2/2$; ✓ $\text{conf}(B, C \Rightarrow A) = 1/2$ ✗
- $\text{conf}(B \Rightarrow A, C) = 2/4$; ✗ $\text{conf}(A, C \Rightarrow B) = 1$ ✓
- $\text{conf}(C \Rightarrow A, B) = 2/5$; ✗ $\text{conf}(A, B \Rightarrow C) = 2/3$ ✓

itemset	support
{A}	2
{B}	4
{C}	5
{A, B}	3
{A, C}	2
{B, C}	4
{A, B, C}	2

Interestingness Measurements

Objective measures

- Two popular measurements:
- **support** and
- **confidence**

Subjective measures

[Silberschatz & Tuzhilin, KDD95]

- A rule (pattern) is interesting if it is
- **unexpected** (surprising to the user) and/or
- **actionable** (the user can do something with it)

Criticism to Support and Confidence

Example 1 [Aggarwal & Yu, PODS98]

Among 5000 students

- 3000 play basketball (=60%)
- 3750 eat cereal (=75%)
- 2000 both play basket ball and eat cereal (=40%)

Rule *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%] is **misleading** because the overall percentage of students eating cereal is 75% which is higher than 66.7%

Rule *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%] is far **more accurate**, although with lower support and confidence

Observation: *play basketball* and *eat cereal* are **negatively correlated**

Other Interestingness Measures: Correlation

Lift is a simple correlation measure between two items A and B :

$$\text{corr}_{A,B} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}$$

! The two rules $A \Rightarrow B$ and $B \Rightarrow A$ have the same correlation coefficient

take both $P(A)$ and $P(B)$ in consideration

$\text{corr}_{A,B} > 1$ the two items A and B are positively correlated

$\text{corr}_{A,B} = 1$ there is no correlation between the two items A and B

$\text{corr}_{A,B} < 1$ the two items A and B are negatively correlated