WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Association Rules

Marzena Kryszkiewicz

HUMAN CAPITAL
Project is co-financed by European Union within European Social Fund

Association Rules - Informally

- Let item {*fish*} occur in 5% of sales transactions and set {*fish, white wine*} occur in 4% of them. This information allows us to derive an *association rule* stating that:

4 out of 5 *customers; that is,* 80% of *customers who buy fish also buy white wine*.

- In order to derive such rules we need to know how many transactions support respective *sets of items* (or *itemsets*).

Support of Itemsets

- Let dataset D be a set of *transactions*, where each transaction is a subset of items in *I*.
- *Support of an itemset X*, denoted by *sup*(*X*), is the number of transactions in D that contain all items in *X*; that is,

$$sup(X) = |\{T \in D \mid X \subseteq D\}|.$$

Example: Supports of Itemsets

Example dataset D

Id	*Transaction*
T_1	*ABCDEG*
T_2	*ABCDEF*
T_3	*ABCDEH*
T_4	*ABDE*
T_5	*ACDEH*
T_6	*BCE*

- *sup*(*ABC*) = 3.
- *sup*(*EH*) = 2.
- Supports of all supersets of *EH* are not greater than 2 either.
- Supports of all subsets of *EH* can be greater than 2.

Relative Support of Itemsets

- *Relative support of an itemset X,* denoted by *rSup*(*X*), is the ratio of the number of the transactions in D that contain all items in *X* to the number of all transactions in D:

$$\mathbf{rSup(X) = sup(X) / |D|}.$$

- **Remark:** *rSup*(*X*) can be regarded as an estimation of the probability of the occurrence of itemset *X* in D.

Example: Relative Supports

Example dataset D

Id	*Transaction*
T_1	*ABCDEG*
T_2	*ABCDEF*
T_3	*ABCDEH*
T_4	*ABDE*
T_5	*ACDEH*
T_6	*BCE*

- *rSup*(*ABC*) = 3/6 = 50%,
- *rSup*(*EH*) = 2/6 ≈ 33%.

Frequent Itemsets

- X is defined a *frequent itemset* if
  $sup(X) > minSup$,
  where *minSup* is the user-defined threshold value.
- **Basic property of itemsets**:Supports of supersets of an itemset X are not greater than $sup(X)$.

7

Example: In(frequent) Itemsets

Example dataset D

Id	Transaction
T_1	$ABCDEG$
T_2	$ABCDEF$
T_3	$ABCDEH$
T_4	$ABDE$
T_5	$ACDEH$
T_6	BCE

- $sup(ABC) = 3$, $sup(EH) = 2$.
- Let $minSup = 2$. Then: ABC is frequent, EH is not frequent.
- Supports of all supersets of EH are not greater than 2 either, hence supersets of EH are not frequent.
- However, supports of subsets of EH can be greater than 2. Thus, it may happen that subsets of EH are frequent.

8

Association Rules (ARs)

- An *association rule* is an expression associating two itemsets:
  $X \rightarrow Y$,
  where $\emptyset \neq Y \subseteq I$ and $X \subseteq I \setminus Y$.
- X is called an *antecedent* of $X \rightarrow Y$.
- Y is called a *consequent* of $X \rightarrow Y$.
- $X \rightarrow Y$ is said to be *based on* $X \cup Y$, and $X \cup Y$ is called the *base* of $X \rightarrow Y$.

9

Support of Association Rule

- *Support* of $X \rightarrow Y$ is defined as the number of transactions that contains the base of $X \rightarrow Y$; that is,
  $\mathbf{sup(X \rightarrow Y) = sup(X \cup Y)}$.
- *Relative support* of $X \rightarrow Y$ is defined as the relative support of its base:
  $\mathbf{rSup(X \rightarrow Y) = rSup(X \cup Y)}$.

10

Confidence of Association Rule

- *Confidence* of $X \rightarrow Y$ is defined as the ratio of the number of transactions that contain the base $X \cup Y$ to the number of transactions containing the antecedent X:
  $\mathbf{conf(X \rightarrow Y) = sup(X \rightarrow Y) / sup(X)}$.
- **Remark:** $conf(X \rightarrow Y)$ can be regarded as an estimation of the conditional probability that Y occurs in a transaction T provided X occurs in T.

11

Example: Association Rules

Example dataset D

Id	Transaction
T_1	$ABCDEG$
T_2	$ABCDEF$
T_3	$ABCDEH$
T_4	$ABDE$
T_5	$ACDEH$
T_6	BCE

$sup(ABC) = 3$, $sup(A) = 5$.

Hence:

- $sup(\{A\} \rightarrow \{BC\}) = sup(\{ABC\}) = 3$,
- $conf(\{A\} \rightarrow \{BC\}) = sup(\{ABC\}) / sup(\{A\}) = 3/5$.

12

Strong Association Rules

- *Strong association rules* (***AR***) are defined as those rules in *AR* whose support is above *minSup* and confidence is above *minConf*; that is,

$$\boldsymbol{AR} = \{X \rightarrow Y \in AR \mid sup(X \rightarrow Y) > minSup \wedge conf(X \rightarrow Y) > minConf\},$$

where $minSup \in [0, |D|)$ and $minConf \in [0, 1)$.

13

Strong ARs and Frequent Itemsets

$$\boldsymbol{AR} = \{X \rightarrow Y \in AR \mid \boldsymbol{sup(X \rightarrow Y) > minSup} \wedge conf(X \rightarrow Y) > minConf\}$$

$$= \{X \rightarrow Y \in AR \mid \boldsymbol{sup(X \cup Y) > minSup} \wedge conf(X \rightarrow Y) > minConf\}$$

$$= \{X \rightarrow Y \in AR \mid \boldsymbol{(X \cup Y) \text{ is frequent}} \wedge conf(X \rightarrow Y) > minConf\}$$

14

Discovery of Strong Association Rules

AR is discovered in two steps:

- Find frequent itemsets ***F*** and their supports in dataset D.
- Generate ***AR*** only from ***F***: Let $Z \in \boldsymbol{F}$, $Z \neq \emptyset$ and $Y \subseteq Z$. Then, any candidate rule $Z \setminus Y \rightarrow Y$ is a strong association one if:

$$sup(Z) / sup(Z \setminus Y) > minConf.$$

15

References

- Agrawal R., Imielinski T., Swami A.: Mining Associations Rules between Sets of Items in Large Databases. In: Proc. of the ACM SIGMOD Conference on Management of Data, Washington, USA (1993) 207–216
- Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499
- Fernando Berzal Galiano, Ignacio J. Blanco, Daniel Sánchez, María Amparo Vila Miranda: A New Framework to Assess Association Rules. IDA 2001: 95-104

16