

Data Mining, Lecture 8

Association Pattern Mining

S. Nõmm

¹Department of Software Science, Tallinn University of Technology

02.11.2021

What Pattern is?

- Pattern recognition is the discipline whose goal is the classification of objects into a number of classes or categories. [S.Theodoridis]
- What Pattern is? Object? Sub set?

Market basket data

- Most popular example is *Supermarket data*. The goal is to determine *associations* between groups of items bought by customers.
- Discovered sets of items are referred to as *large itemsets*, *frequent itemsets*, or *frequent patterns*.
- Main applications include supermarket data (or shopping basket data in general), text mining, generalization to dependency-oriented data types.
- Within this chapter initial data will be referred as *transactions* and outputs as *itemsets*.

The Frequent Pattern Mining Model

- Let U be the d - dimensional universe of elements (goods offered by the supermarket) and \mathcal{T} is the set of transactions T_1, \dots, T_n . They said that transaction T_i is drawn on universe of items U .
- T_i may be represented by d -dimensional binary record.
- *itemset* is the set of items. *k-itemset* is the itemset containing exactly k -items.

The Frequent Pattern Mining Model

Definition

Support *The support of an itemset I is defined as the fraction of the transactions in the database $\mathcal{T} = \{T_1, \dots, T_n\}$ that contain I as the subset*

The support of the itemset I is defined by $sup(I)$. Not to be confused with supremum.

Definition

Frequent Itemset Mining Given a set of transactions $\mathcal{T} = \{T_1, \dots, T_n\}$ where each transaction T_i is drawn on the universe of elements U , determine all itemsets I that occur as a subset of at least a predefined fraction $minsup$ of the transactions in \mathcal{T} .

Predefined fraction $minsup$ is referred as *minimal support*.

Example: Market basket data set

tid	Set of items	Biary representation
1	{ Bread, Butter, Milk }	110010
2	{ Eggs, Milk, Yogurt }	000111
3	{ Bread, Cheese, Eggs, Milk }	101110
4	{ Eggs, Milk, Yogurt }	000111
5	{ Cheese, Milk, Yogurt }	001011

The Frequent Pattern Mining Mode

Definition

Frequent Itemset Mining: Set-wise Given as set of sets $\mathcal{T} = \{T_1, \dots, T_n\}$, where each transaction T_i is drawn on the universe of elements U , determine all sets I that occur as the subset of at least a predefined fractonminsup of the sets in \mathcal{T} .

Support Monotonicity Property *The support of every subset J of I is at least equal to the of the support of itemset I .*

$$\text{sup}(J) \geq \text{sup}(I) \quad \forall J \subset I$$

Downward Closure Property *Every subset of the frequent itemset is also frequent.*

Definition

Maximal Frequent Itemsets *A frequent itemset is maximala at a given minimum support level minsup , if it is frequent and no superset of its frequent.*

Association Rule Generation Framework

Informal definition If the presence of item set X in the certain transaction(s) leads (implies) presence of the set of items Y in the same transaction(s) then we talk about rule $(X \Rightarrow Y)$.

Definition

Confidence Let X and Y be two sets of items. The confidence of the rule $\text{conf}(X \Rightarrow Y)$ conditional probability of $X \cup Y$ occurring in a transaction, given that the transaction contains X

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

Definition

Association Rule Let X and Y be two sets of items. Then, the rule $X \Rightarrow Y$ is said to be an association rule at a minimum support of minsup and minimum confidence minconf if it satisfies following conditions.

- 1 $\text{sup}(X \cup Y) \geq \text{min sup}$
- 2 $\text{conf}(X \Rightarrow Y) \geq \text{minconf}$

Frequent Itemset Mining Algorithms

- Brute force algorithms.
- The Apriori algorithm.
- Enumeration-Tree Algorithms
- Recursive Suffix-Based Pattern Growth Methods

The Apriori Algorithm

begin

$k = 1;$

$\mathcal{F}_1 = \{ \text{All Frequent 1-itemsets} \};$

while $\mathcal{F}_k \neq \emptyset$

 Generate \mathcal{C}_{k+1} by joining itemset-pairs in \mathcal{F}_k ;

 Prune itemsets from \mathcal{C}_{k+1} that violate downward closure;

 Determine \mathcal{F}_{k+1} by support counting on (\mathcal{C}_{k+1}, T) and
 retaining from \mathcal{C}_{k+1} with support of at least minsup;

$k = k + 1;$

end

return $(\cup_{i=1}^k \mathcal{F}_i)$

end

Alternative Models: Interesting Patterns

- Collective strength
- Statistical Coefficient of Correlation
- χ^2 Measure
- Nonlinear relationships

Collective strength

- An itemset is said to be in *violation* of transaction, if some of the items are present in the transaction and others are not.
- The *violation rate* $v(I)$ of the itemset I is defined as the fraction of violations of the itemset I over all transactions.
- The collective strength $C(I)$ of the itemset I is defined as follows

$$C(I) = \frac{1 - v(I)}{1 - E[v(I)]} \cdot \frac{R[v(I)]}{v(I)}.$$

- The expected value of the $v(I)$

$$R[v(I)] = 1 - \prod_{i \in I} p_i - \prod_{i \in I} (1 - p_i)$$

where p_i is the fraction of transactions where the item i occurs.

Collective strength

- Let us consider *violation* to be an unfavorable event (prospective of establishing a high correlation among items)
- Collective strength may be expressed as follows:

$$C(I) = \frac{\text{Good events}}{E[\text{Good events}]} \frac{E[\text{Bad events}]}{\text{Bad events}}$$

- This leads us to the idea of *Negative Pattern Mining*. Determine patterns between the items or their absence.

Statistical Coefficient of Correlation

Covariance is the measure of the strength of correlation between two sets of random variables.

$$\text{cov}(X, Y) = \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Correlation coefficient is standardized

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

or in another form

$$\rho = \frac{E[XY] - E[X]E[Y]}{\sigma(X)\sigma(Y)}$$

Statistical Coefficient of Correlation

The Pearson correlation coefficient

$$\rho = \frac{E[XY] - E[X]E[Y]}{\sigma(X)\sigma(Y)}$$

May be rewritten in terms of *support* as follows

$$\rho_{ij} = \frac{\text{sup}(\{i, j\}) - \text{sum}(i) \cdot \text{sup}(j)}{\sqrt{\text{sup}(i) \cdot \text{sup}(j) \cdot (1 - \text{sup}(i)) \cdot (1 - \text{sup}(j))}}$$

Should we talk here about regression?

χ^2 measure

χ^2 test allows to assess if unpaired observations of two categorical variables are independent of each other or not.

$$\chi^2 = \sum_{i=1}^{\nu_1 \cdot \nu_2} \frac{(\mathcal{O}_i - E_i)^2}{E_i}$$

where ν_1 and ν_2 are the degrees of freedom (number of categories) in the first and in second variables respectively. In the case of binary data $\nu_1 \cdot \nu_2 = 2^{|X|}$.

Nonlinear

- $$y(x) = a_1x^n + a_2x^{n-1} + \dots + a_nx + b$$

- $$y(x) = f(x)$$

where $f(\cdot)$ is arbitrary nonlinear function