Association Rule Development for Market Basket Dataset

S. S. Bhaskar

Anandibai Raorane Arts, Commerce and
Science College Vaibhavwadi, Sindhudurg
India-416810

ABSTRACT

The Term Data mining is used to analyse a big dataset in Statistics. Data mining contains different kinds of approaches like classification, clustering and association. This research work focused on association rule only. association has two special characteristics, which are support and confidence. In this research work, the methodology of association has been studied and developed different rules for a real-life dataset of a super market**.** These rules are based on three items only.

Keywords

Data mining, association, Support, confidence, Lift

1. INTRODUCTION

Data mining is the process of extracting the knowledge from large amount of dataset. dataset may be relational, temporal, spatial, multimedia etc. Data mining is very popular technique used in statistics. In Data mining there are different kinds of approaches like classification, regression, cluster analysis and association. Each approach has a different technique, like classification based on decision tree induction and association based on interesting pattern generation. This research work focused on association rule only.

Association rule development is very important tool for mining process. it has two characteristics. first one is support and another is confidence. Support gives total number of transaction of any particular item occurring in dataset and confidence gives strength of a data in a dataset, one can say support is probability of A and B while confidence is conditional probability. Association rule is based on these two characteristics. Motivation to the problem and objectives are defined in Section 2. A brief review of literature is expressed in section 3. Methodology of the association tool is discussed in section 4. In this research work, a dataset has been collected from supermarket and patterns of the items are generated using association rule which are provided in section 5. Results and conclusions are reported in section 6.

2. MOTIVATION TO THE PROBLEM

In a supermarket there are various types of items which are placed in different places. When a customer comes in the supermarket to buy the products, he buys only those products which are visible easily and he doesn't buy other items which are related to those selected products. Thus, it is a loss of supermarket. To avoid this loss, the goal is to generate a relationship between the items using association rule.

Objectives

- To introduction and Characteristics of Association rule development method.
- To understand the methodology of association.
- To develop a relationship between variables using association rule for a real dataset.

3. REVIEW OF LITERATURE

A lot of research work has been done on association rule development by many researchers. Which are given bellow. Wiwik Novitasari et. al. obtained a Method of Discovering Interesting Association Rules from Student Admission Dataset (2015). V. Umarani discussed a Study on Incremental Association Rule Mining (2015). Guoqi Qian et. al. provided Boosting association rule mining in large datasets via Gibbs sampling (2016). Shang E et. al. explained Association Rules Mining and Statistic Test Over Multiple Datasets on TCM Drug Pairs (2017). Shalini Bhaskar Bajaj developed a method on ARAS: Efficient Generation of Association Rules Using Antecedent Support (2014).

4. METHODOLOGY

Association is used to identify the strong relationship between the variables. association rules are developed on transactional database with usual market-basket applications (see [7] V. Umarani 2015). Define a set of n variables, $I = \{i_1, i_2, ---- i_n\}$ called item set. And a set of m transactions $D = \{t_1, t_2, ----- t_m\}$ contains these items. Each transaction in D has a unique transaction ID and contains a subset of the item set I. suppose X and Y are the two item sets then association rule between item set X and Y is given in the form of implication which is given bellow.

$X \Longrightarrow Y$, where $X, Y \in I$

Where item set X is called antecedent and Y is called consequent (see [1] Guoqi Qian et. al. 2016).

The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than given thresholds. These rules are called strong rules, and the framework is known as the support-confidence framework for association rule mining (see [3] M.L. Antonie and O. R. Zaiane 2004).

There are three measures of association rule which are used to measure the strength of the association.

1. Support(S)
2. Confidence
3. Lift

Support(S)

support of an item set X is the proportion of the transactions t in the data set which contains item set X or it is the ratio of no. of transactions which contains itemset X to the total no. of transactions (see [8] Wiwik Novitasari et. al. 2015).

$$\text{Support(X)} = \frac{\text{no .of transaction which contains itemset X}}{\text{total no .of transaction}}$$

(see [5] S. B. Bajaj 2014).

Confidence
Confidence of association rule X⟹Y is the ratio of support of X∪Y to the support of X. where X∪Y means itemset X and Y occurs together (see [8] Wiwik Novitasari et. al. 2015).

Confidence(X⟹Y)

$$= \frac{\text{no .of transaction which contains itemset X and Y}}{\text{no .of transaction which contains item set X}}$$

(see [5] S. B. Bajaj 2014).

Lift (X⟹Y)

Lift is a simple correlation measure which is given bellow.

$$\text{Lift } (X \Longrightarrow Y) = \frac{P(X \cup Y)}{P(X) * P(Y)}$$

If Lift(X⟹Y) <1, then occurrence of X is negatively correlated with occurrence of Y.

If Lift(X⟹Y) >1, then occurrence of X is positively correlated with occurrence of Y.

If Lift(X⟹Y) =1, then occurrence of X is independent of occurrence of Y (see [8] Wiwik Novitasari et. al. 2015).

Example

To study the above procedure let's see an example. Consider an example of a dataset contains 5 transactions and 5 items i.e. 5 costumers with 5 products, which is given in table 1. In this data set, consider itemset X= {milk, bread} and Y= {butter}

Table 1

T. ID	Milk	Bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Let's obtain support and confidence of this dataset.

Thus, support (X) =2/5 support(Y) =2/5

support(XUY) 1/5= 1/5 Confidence(X⟹Y) = ½ = 0.5,

$$\text{Lift } (X \Longrightarrow Y) = \frac{1/5}{\frac{2}{5} * \frac{2}{5}} = 5/4 = 1.25 \quad \text{which is greater than 1.}$$

Then occurrence of X is positively correlated with occurrence of Y. i.e. occurrence of {milk & bread} is positively correlated with occurrence of {butter}.

Result: - If a customer buys milk and bread together then he also buys butter. Thus {milk & bread}⟹ {butter},

(see [6] Shang E. et. al. 2017).

5. REAL LIFE DATA ANALYSIS

To illustrate the above procedure, a dataset has collected from SAMATA BAZAR, PHONDA [4], and analyzed it using association tool. The dataset contains 20 transactions and 32 items. i.e. data of 20 costumers. out of 32 items, 14 items are selected which occurred more frequently. The selected items are classified into 4 groups. Number of transactions and Support of these items are calculated, which are given in table 2, 3, 4 and 5.

Table 2
Group 1

SN	items	Freq.	support
1	Tea powder	6	0.3
2	Sugar	4	0.2
3	Biscuit	5	0.25
4	Cardamom	3	0.15

Table 3
Group 2

SN	items	Freq.	support
5	Gram flour	4	0.2
6	Legumes	9	0.45
7	Coconut	10	0.5
8	Oil	5	0.25

Table 4
Group 3

SN	items	Freq	support
9	Soap	3	0.15
10	Toothbrush	3	0.15
11	Napkin	6	0.3
12	Toothpaste	2	0.1

Table 5
Group 4

SN	items	Freq.	support
13	Matchbox	3	0.15
14	Joss stick	7	0.35

Now item set with 2 items are considered and different item sets which contains 2 items are generated. also, number of transactions and Support of these item sets are calculated, which are given bellow in table 6, 7, 8 and 9.

Table 6
For group 1

Items		freq.	support
Tea powder	Sugar	2	0.1
Tea powder	Biscuit	2	0.1
Tea powder	Cardamom	2	0.1
Sugar	Biscuit	2	0.1

Sugar	Cardamom	2	0.1
Biscuit	Cardamom	1	0.05

Table 7
For group 2

items		freq.	support
Gram Flour	Legumes	2	0.1
Gram Flour	Coconut	2	0.1
Gram Flour	Oil	3	0.15
Legumes	Coconut	4	0.2
Legumes	Oil	2	0.1
Coconut	Oil	4	0.2

Table 8
For group 3

Items		frequency	support
Soap	Toothbrush	1	0.05
Soap	Napkin	1	0.05
Soap	Toothpaste	1	0.05
Toothbrush	Napkin	1	0.05
Toothbrush	Toothpaste	1	0.05
Napkin	Toothpaste	1	0.05

Table 9
For group 4

Items		frequency	support
Matchbox	Joss Stick	2	0.1

Now item sets with 3 items are generated, number of transactions, Support, confidence and Lift are calculated for these item sets. Which are given in table 10.

Table 10

Items			frequency	support	Confidence	Lift
Tea powder	Sugar	Biscuit	2	0.1	1	**4**
Tea powder	Sugar	Cardamom	1	0.05	0.5	**3.333333**
Sugar	Biscuit	Cardamom	1	0.05	0.5	**3.333333**
Biscuit	Cardamom	Tea powder	1	0.05	0.5	**3.333333**

Items			frequency	support	Confidence	Lift
Gram Flour	Legumes	Coconut	2	0.1	1	**2**
Gram Flour	Legumes	Oil	2	0.1	1	**4**
Gram Flour	Coconut	Oil	3	0.15	1.5	**6**
Legumes	Coconut	Oil	2	0.1	0.5	**2**

Items			frequency	support	Confidence	Lift
Soap	Toothbrush	Napkin	1	0.05	1	**3.333333**
Soap	Toothbrush	Toothpaste	1	0.05	1	**10**
Toothbrush	Napkin	Toothpaste	1	0.05	1	**10**
Napkin	Toothpaste	Soap	1	0.05	1	**6.666667**

6. CONCLUSION

Lift of all item sets are greater than 1 therefore,

In group 1

- occurrence of {Tea Powder, Sugar} ⟹ {Biscuit}
- {Tea Powder, Sugar} ⟹ {Cardamom}
- {Sugar, Biscuit} ⟹ {Cardamom}
- {Biscuit, Cardamom} ⟹ {Tea Powder}

In group 2

- {Gram Flour, Legumes} ⟹ {Coconut}
- {Gram Flour, Legumes} ⟹ {Oil}
- {Gram Flour, Coconut} ⟹ {Oil}
- {Legumes, Coconut}⟹ {Oil}

In group 3

- {Soap, Toothbrush} ⟹ {Napkin}
- {Soap, Toothbrush} ⟹ {Toothpaste}
- {Toothbrush, Napkin} ⟹ {Toothpaste}
- {Napkin, Toothpaste} ⟹ {Soap}

Therefore, these items are to be placed **nearly such that they are visible easily to the customer.**

7. REFERENCES

[1] Guoqi Qian, Calyampudi Radhakrishna Rao, Xiaoying Sun, and Yuehua Wu, Boosting association rule mining

in large datasets via Gibbs sampling, PubMed.gov, US National Library of Medicine National Institutes of health, volume-113(18):4958-63, (2016).

[2] Han, J. and Kamber, M. and Pei, J. Data Mining: Concepts and Techniques. Morgan Gaufmann, (2012)

[3] Maria-Luiza Antonie and Osmar R. Zaiane, Mining Positive and Negative Association Rules: An Approach for Confined Rules, PKDD 2004, LNAI 3202, pp. 27–38, 2004. c Springer-Verlag Berlin Heidelberg 2004.

[4] Samata Bazar, Phonda, Sindhurdurg, India.

[5] Shalini Bhaskar Bajaj, ARAS: Efficient Generation of Association Rules Using Antecedent Support, 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 978-1-4799-5148-2/14/$31.00, (2014).

[6] Shang E, Duan J, Fan X, Tang Y and Ye L, Association Rules Mining and Statistic Test Over Multiple Datasets on TCM Drug Pairs, International Journal of Biomedical Data Mining, ISSN: 2090-4924, Volume 6, (2017).

[7] V. Umarani, A Study on Incremental Association Rule Mining, International Journal of Computer Science and Information Technologies, Vol. 6 (4), 3961-3964, (2015).

[8] Wiwik Novitasari, Arief Hermawan, Zailani Abdullah, Rahmat Widia Sembiring and Tutut Herawan, A Method of Discovering Interesting Association Rules from Student Admission Dataset, International Journal of Software Engineering and Its Applications Vol. 9, No. 8, 51-66 (2015).