Association Rule Mining to Extract Knowledge from Online Store Transactions of a Turkish Retail Company: A Case Study

Elif Şafak Sivri[1]
Mustafa Cem Kasapbaşı[2]
Fettullah Karabiber[3]

Abstract
Data mining techniques have been implemented in many fields namely, marketing, insurance, finance, medicine, computer science and many more. In marketing it is used as a tool to cluster and classify customers so that their buying patterns, demographical information, market basket can be analyzed to help the CRM representative and decision makers [1]. In this study online store transactions of multi-branch Turkish Retail Company have been analyzed and many associations rules have been discovered. The analyzed volume of transactions of completed sales exceeds 14000 for a single season. At first data is cleaned from unrelated fields then presented to R studio to implement the Apriori algorithm[2] in order to extract knowledge and obtain association rules between goods. Results are proven be worthy over the conventional methodologies. The extracted data are tested successfully with a sample group of customers to validate the association rules which give unique insights about customer behaviors.

Keywords: *Association analysis; Apriori Analysis; Data Mining*

1. Introduction

Data mining techniques have been implemented in many fields namely, marketing, insurance, finance, medicine, education, computer science and many more. In marketing it is used as a tool to cluster and classify customers so that their buying patterns, demographical information, market basket can be analyzed to help the CRM representative and decision makers [1]. Data mining has been being used mainly for three purposes: classification, clustering and extracting association rules for given data sets. Market basket analysis is field of study of data mining which tries to emerge any meaningful relationship between sale transactions in a big data set. In this study It is aimed to extract meaningful association rules between sales transaction database of a Turkish Retail company. We acknowledge their contribution to this study for supplying their unique online store sales dataset.

[1] *Istanbul Commerce University, Computer Engineering Department, Istanbul, Turkey*
[2] *Istanbul Commerce University, Computer Engineering Department, Istanbul, Turkey, mckasapbasi@ticaret.edu.tr*
[3] *Yildiz Technical University, Computer Engineering Department, Istanbul, Turkey, fkara@ticaret.edu.tr*

A. Association Rules

In [2] R.Agrawal et al. proposed the association rules between database sales transactions of a supermarket to describe the frequent patterns and associations. Every transactions are represented in itemset groups then probabilities of these itemset are calculated to satisfy two metrics namely Support and Confidence Eq(1) and Eq(2) respectively. The implication of the form (ItemSet) X=>(ItemSet)Y is called an association rules Support means that If the probability of transaction on database contain X and Y itemset together s and percentage is S%, one could say, the rule X=>Y holds in the database with Support S. If the percentage of transactions in the database containing X that also contain Y is c%, the rule X=>Y has Confidence c. [3-4]. In order Rules to be accepted they are filtered by predetermined Support and confidence thresholds to finally state the rules. Lift is a metric used for interestingness and defined as eq (3).

$$Support(X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$Confidence(X \rightarrow Y) = P(Y|X) \quad (2)$$

$$Lift = P(Y|X) / P(Y) \quad (3)$$

2. Background Study and Methodology

A. Literature Review

Despite Market basket analysis is not a new concept it has been a point of attraction to many researcher since it gives good clues for decision makers and marketing strategies and those strategies are most of the time unique to the interested data sets.

The Apriori algorithm was developed to efficiently discover frequent itemsets. Name Apriori is derived from "prior" since algorithm uses data from previous step. This method relies on the assumption that subset of an itemset also caries the properties of the itemset [2, 11].

Since then, new several algorithms such as the FP-Growth algorithm have been proposed for this purpose [4]. Due to great success and widespread usage of Apriori algorithm, many variations of association rule algorithms have been proposed [3]. These different algorithms can be classified in three groups according the data they used for association nominal/ Boolean data [2, 5], ordinal data [6], and quantitative data [7, 8]. In [9] for an online store a recommendation system is established based on Bayesian Network and Association Rules. The detailed description of the Apriori Algorithm is given in [2].

In this study a clustering algorithm namely Expectation Maximization (EM) is used since such clustering algorithms are used to classify grouped data according to their resembling. In our study resembling are extracted using EM clustering algorithms. EM Clustering algorithms [12];

- In every iteration of EM algorithm value of probability function increases.
- Under general regularity conditions it has confident convergence
- Calculations are relatively easy and applicable
- Cost of iteration is low when compared with cost of other methods
- Missing value guess also available

b. Methodology

Blok diagram of the proposed methodology has been given in Fig. 1.

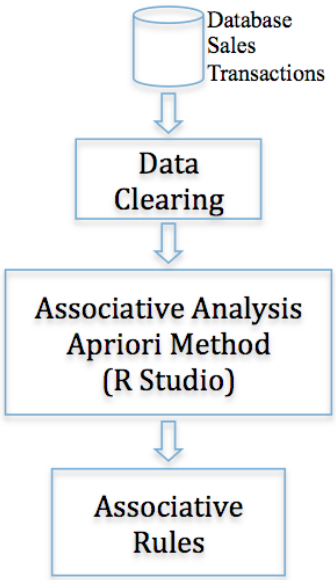

Fig.1.*Block Diagram of Determining Associative Rules*

There are more then 14000 sales transaction for a single season. Therefore first step of a data mining is to select raw data form transection and clearing them with either projection of filtering. When the processed raw data is read to analyze it is feed to analyses engine to further process and produce knowledge. Figure 1 summarizes our process.

Clustering algorithms (such as Expectation Maximization EM algorithm) are used to classify grouped data according to their similarities. Since our data has some missing parts we have chosen EM clustering algorithm to discover the demographical representation of the database transactions. EM clustering Algorithm is chosen especially when there is a data dropouts (missing data) or clustering in such a way that the number of underlying data points is unknown [10]. EM (Expectation Maximization) clustering algorithm is utilized according acquired discount amount, gender and marital status. Result are represented as in Table I.

Table.1. *Result of EM Algorithm*

Set Number	Acquired Discount Amount	Gender	Marital Status	Percentage
Set 1	10%	Female	Married	14%
Set 2	10%	Female	Single	10%
Set 3	10%	Male	Married	4%
Set 4	10%	Male	Single	2%
Set 5	15%	Female	Married	28%
Set 6	15%	Female	Single	13%
Set 7	15%	Male	Married	6%
Set 8	15%	Male	Single	4%
Set 9	20%	Female	Married	9%
Set 10	20%	Female	Single	3%
Set 11	20%	Male	Married	1%
Set 12	20%	Male	Single	1%
Set 13	25%	Female	Married	2%
Set 14	25%	Female	Single	1%
Set 15	25%	Male	Married	1%
Set 16	25%	Male	Single	1%

Table 1 can be interpreted as; In set 1 %10 discount amount married females group account for %14 of all transactions.

Second Set represents %10 discount amount single female group which accounts for %10 of whole groups. Likewise Third Set represents %10 discount amount married men group which accounts for only %4 of the universal set. Table 1 can be interpret likewise.

Moreover after two iteration two distinct group of products are obtained %93 in one group %7 is in other group.

3. Results

At the end of the study for some of the selected Itemsets (X and Y) depicted in Table 2 support, confidence and lift values found to be 100%. Support level is defined as supp(X) an itemset X is defined as the proportion of transactions in the data set which contain the itemset. Confidence level defined as: Conf(XàY)=supp(X U Y)/ supp(X) meaning X item sold with Y item.

Table.2. *Association Rules for Support, Confidence Lift %100*

Itemset X	Itemset Y
Tight pants	T-Shirt
Deep Slits Silk Tunic	Hand Portfolio
Pearl Leather Bracelet	Rubber based Loafer and 3 lines long leather pearl necklace
Crepe Dress with Scarf	Flared dress Scarf Crepe and 3 lines long leather pearl necklace
Ruffled Chiffon blouses	Jean Pant Love
Jean Pant Love	Jean Pant The London Pant
Black Jean	Dark Grey jean pant
Knitting Multicolored	Knitting beige
Knitting	Classical skirt
Classical silk Shirt	V-neck sleeveless dress
Lid Pocket Classic Coats	Box Collar Sweater
Boot	Silk Cashmere Sweater
Christmas ornament	Decorative Gift Package
Sweatshirt	Scarves and shirts
Knitting Tube Skirt	V-neck Sleeveless blouse, Crepe cigarette pants
Lace Tube Skirt	Embroidered Sleeveless blouses

In Table 3 association rules of itemsets having support and confidence %50 and lift value 100% are given;

Table.3. *Association Rules for Support and Confidence %50, Lift %100*

Itemset X	Itemset Y
Goose feathered coat	Double-sided thin-card
Knitting Tube Skirt	V-neck Sleeveless blouses
plain T-shirt	Long Sleeve T-Shirt
Tights Pants	Low-Shoulder loose T-shirt

As it can be seen from Tables II-III Association rules have been generated with the established model. Any row can be read as Customers who bought X itemset also bought Y itemset with related Support, Confidence and Lift amounts. We can interpret Table II like If a customer buys a tight pant, s/he will probably buy a T shirt. As an other example, If a person buys a Christmas ornament, s/he would buy a Decorative gift package. The Support, confidence and Lift metrics of those decisions is %100.

A. Evaluation of the Results

Second and important phase of the study was to evaluate weather this rules are valid. In order to confirm the correctness of the association rules a relatively small model campaign have been carried out on the selected target customers who only bought X itemset with a support and Confidence and Lift Value %100. The sample set of customer are chosen according to clustering results of table 1 which helped us to select number of customers to represent whole transactions. In this pilot study the chosen customers have been informed that there would be a discount in the corresponding Y itemset for a short period of time. It has been inspected that these customers shopped the Itemsets Y in the given period of time. %78 of customers shopped the Itemsets Y in the given period of time. %12 percent of the customer did not respond to the pilot campaign, %7 percent were out of city and the rest %3 did not give any reason for not participating the campaign.

4. Conclusion

In this study online store database transaction of Turkish retail company have in analyzed to perform market basket analysis. Apriori algorithm is utilized to perform associative analysis, which is implemented in R studio. As it could be inferred from Table 2 and Table 3 many meaningful associative rules is established to better understand the consumer behaviors. In order to evaluate and test the association rule small promotion campaign is carried out and result shows that the association values are valid.

REFERENCES

[1] Timor M. ,EZERCE A. , GURSOY U. T., "Müşteri Profili ve Alişveriş Davranişlarını Belirlemede Kümeleme ve Birliktelik Kuralları Analizi: Perakende sektöründe bir uygulama" , İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadı Enstitüsü Dergisi, February 2011 22 68

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499

[3] J. Singh, H. Ram, Dr. J.S. Sodhi, Improving Efficiency of Apriori Algorithm Using Transaction Reduction International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013

[4] Cheng-Hsiung Weng, Mining fuzzy specific rare itemsets for education data, Knowledge-Based Systems, Volume 24, Issue 5, July 2011, Pages 697-708, ISSN 0950-7051, http://dx.doi.org/10.1016/j.knosys.2011.02.010.

[5] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of ACM SIGMOD, 1993, pp. 207–216.

[6] Y.L. Chen, C.H. Weng, Mining association rules from imprecise ordinal data, Fuzzy Sets and Systems 159 (4) (2008) 460–474.

[7] Y.L. Chen, C.H. Weng, Mining fuzzy association rules from questionnaire data, Knowledge-Based Systems 22 (1) (2009) 46–56.

[8] M. Delgado, N. Marin, D. Sanchez, M.A. Vila, Fuzzy association rules: general model and applications, IEEE Transactions on Fuzzy Systems 11 (2) (2003) 214–225

[9] S. S. Weng, S. C. Liu, T. H. Wu, Applying bayesian network and association rule analysis for product recommendation, International Journal of Electronic Business Management 2011

[10] Moon, T.K., "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE* , vol.13, no.6, pp.47,60, Nov 1996 doi: 10.1109/79.543975

[11] G. Gürgen, "Birliktelik kuralları ve sepet analizi uygulaması", yüksek lisans tezi, Marmara Universitesi, istatistik Anabilim dalı

[12] T. SERVİ, "Çok Değişkenli Karma Dağilim Modeline Dayali Kümeleme Analizi", Çukurova Üniversitesi Fen Bilimleri Enstitüsü, PhD. Thesis, 2009