ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Implementation of Data Mining Techniques to Perform Market Analysis

B.Sabitha[1], N.G.Bhuvaneswari Amma[2], G.Annapoorani[3], P.Balasubramanian[4]

PG Scholar, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, India[1]

Faculty, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, India[2, 4]

Assistant Professor, University College of Engineering, BIT Campus, Tiruchirappalli, India[3]

ABSTRACT: Market analysis is an important component of analytical system in retail companies to determine the sales for different segments of customers to improve customer satisfaction and to increase profit of the company which has different channels and regions. These issues for a leading shopping mall is addressed using frequent item set mining and decision tree technique. The frequent item sets are mined from the market basket database using the efficient Apriori algorithm and hence the association rules are generated. The decision tree can be constructed using ID3 and C4.5 algorithm.

KEYWORDS: Association Rules, Frequent Item sets, Apriori, Decision tree, Market Basket Analysis

I. INTRODUCTION

One of the challenges for companies that have invested heavily in customer data collection is how to extract important information from their vast customer databases and product feature databases, in order to gain competitive advantage. Several aspects of market basket analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for one-to-one marketing, purchasing patterns in a multi-store environment to improve the sales [1]. Market basket analysis has been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them. Informed decision can be made easily about product placement, pricing, promotion, profitability and also finds out, if there are any successful products that have no significant related elements [2]. Similar products can be found so those can be placed near each other or it can be cross-sold. A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together. Market basket analysis gives retailer good information about related sales on group of goods basis and also it is important that the retailer could know in which channel and in which region the products can be sold more and which session (i.e) morning or evening [3].

Market basket analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Market basket analysis determines the products which are bought together and to reorganize the supermarket layout and also to design promotional campaigns such that products' purchase can be improved [11]. Association rules are derived from the frequent item sets using support and confidence as threshold levels [4]. The sets of items which have minimum support are known as Frequent Item set [2]. The support count of an item set is defined as the proportion of transactions in the data set which contain the item set. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. Association rules derived depends on confidence [5].

II. RELATED WORK

A number of approaches have been proposed to implement data mining techniques to perform market analysis. Loraine et al. in their work proposed a market basket analysis using frequent item set mining. They compared Apriori with K-Apriori algorithm to find the frequent items [1]. Vishal et al. implemented data mining in online

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

shopping system using Tanagra tool. They made decision about the placement of product, pricing and promotion [2].

Sudha and Chris et al. proposed the impact of customers perception and crm on indian retailing in the changing business scenario using data mining techniques[3][4]. Comparing to the works discussed above, our work is different by using apriori and decision tree to perform market basket analysis.

III. SYSTEM ARCHITECTURE

A. Customer data set:

The Wholesale customer data provided by the UCI Machine Learning Repository is used for analysis of this work [7]. The dataset has 8 continuous and 1 numeric input attributes namely channel, region, fresh, milk, grocery, frozen, detergents, delicatessen and session.

It also has the predicted attribute i.e) the class label. Here the channel1 represents horeca (hotel/restaurant/café), channel2 represents retail shops. Region1 represents Lisbon, region2 represents Oporto, region3 represents the others. The description of the dataset is tabulated in Table 1.

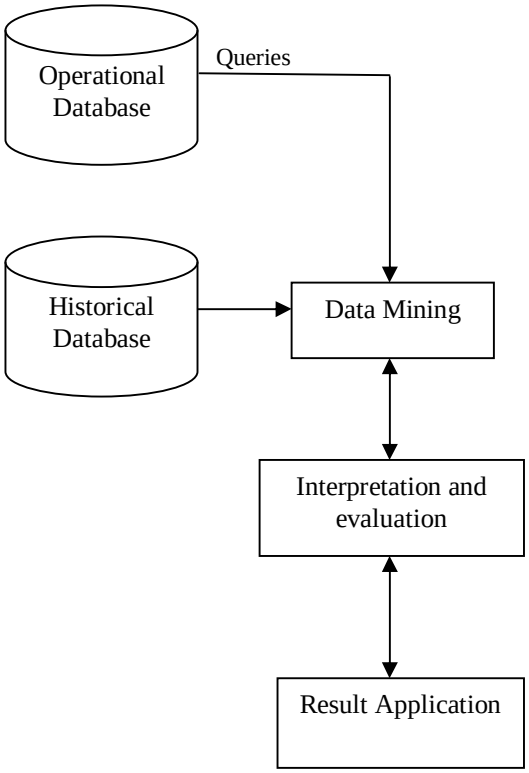

Fig.1 Block Diagram of Proposed system.

Table.1 Summary of market analysis

Table 1. Dataset Description

Attribute	Description
Channel	1. Horeca 2. Retail
Region	1. Lisbon 2. Oporto 3. Others
Fresh	Annual spending on fresh products
Milk	Annual spending on milk products
Grocery	Annual spending on grocery products
Frozen	Annual spending on frozen products
Detergents	Annual spending on detergents products
Delicatessen	Annual spending on delicatessen products

B. Association Rules:

Association rules are of the form if X then Y. Frequent patterns is patterns (such as item sets, subsequences, or substructures) that appear in a data set frequently [6]. Frequent pattern mining searches for recurring relationships in a given data set. Association rules are not always useful, even if they have high support, confidence and lift > 1. Association rules can also be improved by combining purchase items. Items often fall into natural hierarchies. In This

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Section, frequent item set can be generated using apriori algorithm and associate outliers also be generated according to the given support count and confidence level.

C. Decision tree:

Decision tree induction is the learning of decision trees from class-labeled training tuples. Decision tree algorithms, such as ID3, C4.5, and CART, were originally intended for classification. Decision tree induction constructs a flow chart like structure where each internal (non leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. The topmost node in a tree is the root node. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery [13].

Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. Decision trees are the basis of several commercial rule induction systems. At each node, the algorithm chooses the "best" attribute to partition the data into individual classes [13].

ID3:

ID3 uses information gain as its attribute selection measure. The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by jCi,Dj/jDj [4]. A log function to the base 2 is used, because the information is encoded in bits. Info (D) is just the average amount of information needed to identify the class label of a tuple in D[8].

InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. This can be measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term D_j acts as the weight of the jth partition. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$Gain(A) = Info(D) - Info_A(D)$$

C4.5:

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, it is often referred to as a statistical classifier. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set *S*=s1, *s2*... of already classified samples.

Each sample s_i consists of a p-dimensional vector ($x1,i,x2,i,...,xp,i$), where the xj represent attributes or features of the sample, as well as the class in which s_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

It applies a kind of normalization to information gain using a "split information" value defined analogously with Info (D) as

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

This value represents the potential information generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A [9].

Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D [5]. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning. The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}.$$

The attribute with the maximum gain ratio is selected as the splitting attribute[15].

Here also the error rate and the confusion matrix of ID3 can be found and for the given dataset the decision tree can be generated by

- Channel < 1.5000
  - Region < 2.5000 then session = morning(54.02 % of 87 examples)
  - Region >= 2.5000 then session = evening(58.77 % of 211 examples)
- Channel >= 1.5000
  - Region < 1.5000 then session = evening (72.22 % of 18 examples)
  - Region >= 1.5000 then session = morning(56.45% of 124 examples)

This is the simple decision tree for three attributes channel, region and session.

If we construct the decision tree for the whole dataset it becomes very efficient with the accuracy of 72.22% maximum [10].

IV. SIMULATION RESULTS

The whole dataset was given to the data mining tool like Tanagra. Then frequent item set is found using apriori algorithm in the association technique [12]. This paper is mainly focused to find out whether the products can be sold more at morning session or evening session. For this, it uses two decision tree algorithms called ID3 and C4.5. Using ID3 the dataset parameters can be splitted and also found the error rate with confusion matrix [13]. Using C4.5 algorithm, the decision tree can be constructed for the given confidence level and minimum size of leaves [6].

Table2. Statistical analysis of wholesale customer data

Attribute	Min	Max	Mean	Standard deviation
Fresh	3	112151	12000.30	12647.329
Milk	55	73498	5796.27	7380.377
Grocery	3	92780	7951.28	9503.163
Frozen	25	60869	3071.93	4854.673
Detergents	3	40827	2881.49	4767.854
Delicatessen	3	47943	1524.87	2820.106

The statistical analysis of the whole dataset is given in Table 3. It gives the mean and accuracy of the product sold in two sessions.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Table 3. Statistical analysis of session:

Description of session							
Session = morning				Session = evening			
Examples		[47.5%]209		Examples		[52.5%] 231	
Att – Desc	Test value	Group	Overall	Att – Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Channel	1.54	1.36(0.48)	1.32(0.47)	Fresh	0.80	12462.13 (14302.89)	12000.30 (12647.33)
Milk	0.39	5941.77 (7921.05)	5796.27 (7380.38)	Grocery	0.54	8186.32 (9870.68)	7951.28 (9503.16)
Delicatessen	0.26	1562.14 (1931.80)	1524.87 (2820.11)	Region	0.31	2.55 (0.78)	2.54 (0.77)
Frozen	0.13	3102.73 (5626.43)	3071.93 (4854.67)	Detergents	0.06	2894.07 (4868.62)	2881.49 (4767.85)
Detergents	-0.06	2867.59 (4665.58)	2881.49 (4767.85)	Frozen	-0.13	3044.07 (4043.96)	3071.93 (4854.67)
Region	-0.31	2.53 (0.77)	2.54 (0.77)	Delicatessen	-0.26	1491.15 (3435.49)	1524.87 (2820.11)
Grocery	-0.54	7691.50 (9096.26)	7951.28 (9503.16)	Milk	-0.39	5664.62 (6869.39)	5796.27 (7380.38)
Fresh	-0.80	11489.85 (10530.36)	12000.30 (12647.33)	Channel	-1.54	1.29 (0.45)	1.32 (0.47)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

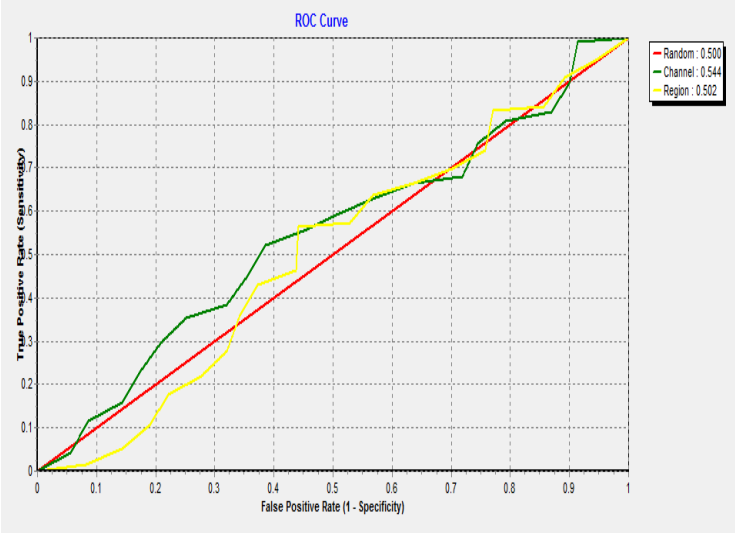

Fig.2 ROC curve

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

The ROC curve of our work is shown in Fig. 2. Here the positive value should be taken as morning and the result becomes nearly true positive is little bit higher than the false positive [14]. This diagram illustrates at what channel and region our products sends more in the morning and whether it gets true positive or not.

V. CONCLUSION

In this paper, a framework for Decision tree and frequent item set is developed for the analysis of wholesale data. The wholesale customer dataset is taken and analyzed to know the session at which the product can be sold more using decision tree algorithm like ID3 and C4.5. The data in the dataset is preprocessed to make it suitable for classification. The preprocessed data is used for classification and we obtained high classification accuracy.

REFERENCES

1. Loraine Charlet Annie M.C.1 and Ashok Kumar D, “Market Basket Analysis for a Supermarket based on Frequent Itemset Mining”, IJCSI International Journal of Computer Science Issues, Vol. 9, No. 3, pp.257-264, 2012.
2. Vishal jain, Gagandeep singh narula & Mayank singh, “Implementation of data mining in online shopping system using Tanagra tool”, International journal of computer science And engineering Vol. 2, No. 1, 2013.
3. Sudha vemaraju, “Changing waves in indian retailing: Impact of customers perception and crm on indian retailing in the changing business scenario”, International Journal of Multidisciplinary Research , Vol.1, No.8, 2011.
4. Chris Rygielski, Jyun-Cheng Wang b, David C. Yen, “Data mining techniques for customer relationship management”, Technology in Society, 2002.
5. P Salman Raju, Dr V Rama Bai, G Krishna Chaitanya, “Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries”, International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, No.1, 2014.
6. Bharati M Ramager, “Data Mining techniques and Applications”, International Journal of Computer Science and Engineering, Vol. 8, No.12, 2009.
7. P. Nancy, and Dr. R. Geetha Ramani, “A Comparison on Data Mining Algorithms in Classification of Social Network Data”,International Journal of Computer Applications, Vol.32, No.8, 2011.
8. Sheikh, L Tanveer B. and Hamdani, "Interesting Measures for Mining Association Rules", IEEE Conference-INMIC , 2004.
9. Sonali Agarwal, Neera Singh, Dr. G.N. Pandey, “Implementation of Data Mining and Data Warehouse in E-Governance”, International Journal of Computer Applications, Vol.9, No.4, 2010.
10. Chen, Y.-L., Tang, K., Shen, R.-J., Hu, Y.-H.:”Market basket analysis in a multiple store environment”, Journal of Decision Support Systems, 2004.
11. Berry, M.J.A., Linoff, G.S.: “Data Mining Techniques: for Marketing, Sales and Customer Relationship Management” (second edition), Hungry Minds Inc., 2004.
12. C. Rygielski, J. C. Wang, and D. C. Yeh, "Data mining techniques for customer relationship management," Technology in Society, vol. 24, 2002.
13. J. Han and M. Kamber, “Data Mining : Concepts and Techniques”, San Francisco: Morgan Kaufmann Publisher, 2006.
14. H. Jantan, A. R. Hamdan, Z. A. Othman, and M. Puteh, "Applying Data Mining Classification Techniques for Employee's Performance Prediction," 5th International Conference in Knowledge Management, 2010.
15. Rastogi, R.. and kyuseok Shim, “Mining optimised association rules with Categorical and numerical attributes”, IEEE transactions on Knowledge and Data Engineering, vol.14, No.2, pp.425-439, 2002.