ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Market Basket Analysis using Apriori and Correlation Measures

Deepti Gupta [1], Harsh Arora [2]

Assistant Professor, Department of Computer Applications, IINTM, Janakpuri, New Delhi, India[1]

Assistant Professor, Department of Computer Applications, IINTM, Janakpuri, New Delhi, India [2]

ABSTRACT: Association Rule Mining is a branch of Data mining which aims at identifying associations between different items stored in large databases. Market Basket Analysis is one of the applications of Association Rule Mining which emphasizes on analysis of shopping baskets of customers to analyse their buying habits and mine such patterns from the massive amount of transactional records which can help retailers and shopping websites in developing their marketing strategies to increase their sales and hence maximize the profit. In this paper, Apriori algorithm is used to find the strong association rules which are then judged for their interestingness on the basis of correlation measures.

KEYWORDS: Market Basket Analysis, Association Rule Mining, Apriori, Correlation

I. INTRODUCTION

Data Mining is the knowledge discovery process to discover hidden patterns and to mine them to assist decision making by market analysts. Association Rule Mining(ARM) is used to extract associations and correlations among itemsets. ARM has find its applications in many fields like intrusion detection, bioinformatics, medical diagnosis and Web usage Mining. Market Basket Analysis is one of the applications that play a crucial role in decision making processes such as store layout design and cross marketing. For instance, shopping sites perform Market Basket analysis to analyse the purchasing history of the customers to recommend them those earlier bought products and provide discounts on combos of items that are likely to be ordered together by them in their next transaction. Also the introduction of point of sale has significantly increased the business use of Market Basket Analysis. Many algorithms like AIS, SETM have been worked upon to find frequent itemsets. Apriori is one of the vital algorithms that were discovered to find frequent patterns. Itemsets, subsequences or substructures that appear frequently in database are all forms of frequent patterns. [7] As data or the number of records in transactional or relational databases is increasing at a rapid rate, it is important to analyse this massive amount of transactions to understand the shopping behaviour of customers. The problem of mining frequent itemsets is a twostep procedure. First step is to find all the frequent itemsets that is itemsets that have support above the minimum threshold or minsupport and second step is to generate association rules from these frequent itemsets such as for each itemset l, generate each association rule s $\rightarrow$(l-s) for each s belongs to l. Rules that have confidence above the minconf factor are considered strong and their interestingness is determined using the null invariant correlation measures that is Kulczynski and Imbalance Ratio.

Paper is organized as follows. Section II describes the literature review by the various authors in this field. The flow diagram is used to represent the proposed methodology in Section III of the paper. Section IV presents an example to explain the significance and working of our methodology. Finally, the paper has been concluded in Section V.

II. RELATED WORK

The objective of mining frequent patterns is to find interesting associations rules which describe recurring relationships between various data items in a database. These association rules can be used by retailers and other

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

shopkeepers to display their items accordingly on the shelf to attract customers to buy related items in one transaction. An overview of the literature work done by the various authors in this area has been discussed as follows.

R Agarwal & R. Srikant in [1] has proposed the Apriori algorithm. They also compared the relative performance of the four association rule mining algorithms AIS, SETM, Apriori and AprioriTid for six synthetic and real time datasets on the basis of number of passes through the database, candidates generated, memory and data structure requirements. They described that AIS and SETM generates candidates on the basis of transactions in the database which results into unnecessary generation of candidates whereas Apriori and AprioriTid generates candidate itemsets using the itemsets found in the previous pass. They also proposed AprioriHybrid algorithm that uses Apriori in initial passes and then switches to AprioriTid for the later passes which incurs a cost.

R. Agarwal in [2] introduces syntactic and support constraints to discover large itemsets. He also incorporates pruning function optimization to prune and discard the candidate itemsets at early stage which will never turn out to be large. He then presented the results of mining the sales data of a retail company to show the effectiveness of his estimation procedure and pruning optimizations.

The author in [3] has described the limitations of Apriori algorithm that the candidate generation procedure is complex and it requires lot of time and memory to execute the Apriori. He then discussed about the FP Tree algorithm that overcomes the bottlenecks of Apriori by reducing the number of passes over the database. He also mentioned the concept of negative associations rules given by Brein which identifies the products that conflict or complement with each other.

Shelia A. Abaya in [4] has proposed a modified Apriori algorithm which uses two factors set size and set size frequency to improve the efficiency of Apriori in terms of execution time and number of database passes. Set Size is the number of items per transaction and set size frequency is the number of transactions in the database having count of items equal to set size.

III. THE PROPOSED METHODOLOGY

Apriori algorithm is one of the major algorithms for identifying the frequent itemsets in the database and mining the association rules between these itemsets to help the market analysts to analyse the purchasing patterns of their customers which can help companies to increase their sales and hence the profits. It is based on Apriori property that each non-empty subset of a frequent itemset must also be frequent. It is named as Apriori because it is a level wise search approach that uses prior knowledge of frequent itemsets. The definition of Apriori has been discussed below along with the notations in Fig1.

Notations used:

L_k :- Set of large itemsets having minimum support
C_k: -Set of itemsets that are candidates to be frequent

Fig. 1. Notations Used in Implementation of Apriori

The two step procedure of Apriori is as follows:

1. Join Step :- The set of candidate k itemsets C_k is generated by performing self join of L_{k-1} as p with L_{k-1} as q on condition $p.item_1 = q.item_1, p.item_2 = q.item_2 \ldots\ldots p.item_{k-1} < q.item_{k-1}$

2. Prune Step: - The itemsets in C_k are pruned to compute large itemset L_k on the basis of minsupport. The support of an itemset is the probability of transactions that contain the two itemsets together and can be calculated as :

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

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

The itemsets which have support count less than minsupport threshold value will not be considered further to be part of frequent itemset. All the association rules are generated next once the frequent itemsets are discovered and then filtered to find strong ones among them using the confidence factor which the conditional probability of Y is given X is already purchased.

$$\text{Confidence}(X \rightarrow Y) = P(Y \mid X) \quad [2]$$

The complete proposed methodology is presented using a flow chart given below. The Apriori algorithm is combined with Null Invariant Correlation measures to find strong and interesting rules.

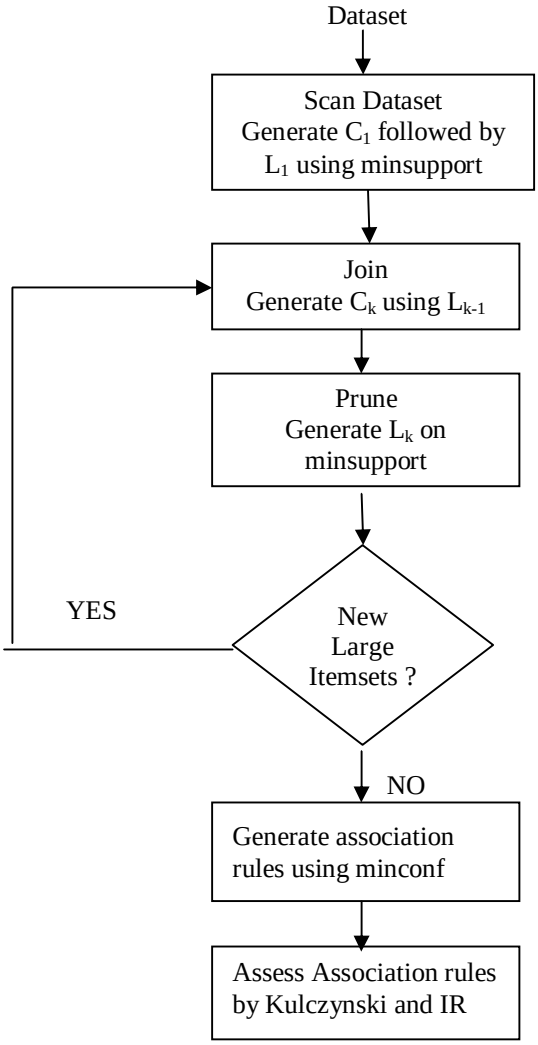

Fig. 2. The Proposed Methodology describing use of Apriori with Correlation Measures

Rules generated from Apriori having confidence above threshold are considered strong. But are the strong rules really interesting? The interestingness of the rules is measured using Null Invariant Pattern Evaluation measures Kulczynski and Imbalance Ratio in our proposed methodology. Null Invariant means measures which are not influenced by null transactions that are transactions not having the itemset being examined.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Given two itemsets X & Y, the Kulczynski(or Kulc) measure can be computed using conditional probabilitiesP(X|Y) and P(Y|X) mentioned in equation 3.

$$Kulc(X, Y) = (P(X|Y) + P(Y|X))/2 \quad [3]$$

The value of this measure ranges between 0 and 1. Kulc= 0.5 signifies neutral or balanced skewness whereas more the value is far away from 0.5; the closer is the relationship between the two itemsets.

Imbalance ratio assesses the imbalance of two itemsets and is defined as :

$$IR(X, Y) = sup(X) - sup(Y)/sup(X) + sup(Y) - sup(X \cup Y) \quad [4]$$

IR=0 can be interpreted as perfectly balanced and IR=1 means highly skewed or very interesting rule. Kulczynski and IR along with confidence and support can help to identify more interesting rules and can support wise decision making by the analysts.

IV. THE EXPERIMENTAL RESULTS

We have used an example to explain our proposed methodology. Suppose we are interested in analysing the purchasing patterns of customer at an electronic store that sells computer games and videos. Considering the dataset of 10,000 transactions, the following given contingency table shows the probability of customer purchasing the videos and computer games.

Table I. The Contingency Table

	Games	$\overline{\text{Games}}$	Total(row)
Videos	4000	3500	7500
$\overline{\text{Videos}}$	2000	500	2500
Total(col)	6000	4000	10000

If the following given association rule is discovered with support and confidence 40% and 66% respectively:

Buys (A,"computer games") => Buys (A,"videos")

Then we will find if this strong rule is really interesting or misleading which could result into unwise decisions.

As discussed above the formula to compute Kulczynski measure and Imbalance ratio, using the above contingency table, the Kulc and IR are as follows:

$$Kulc = (4000 / 7500 + 4000 / 6000) / 2 = 0.59$$

$$IR = | 6000-7500 | / 6000+7500-4000 = 0.157$$

It can be observed from the above computed values that although the association rule buys(A,"computer games") → buys (A,"videos") is having the desired support and confidence which proves it be strong according to Apriori, but the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

rule is not interesting as proved by Kulc and IR values. In fact, the computer games and videos are negatively correlated which means that the increase in sales of computer games will result into decreased sale of videos or vice-versa.

V. CONCLUSION

With the advancements in technology and increase in count of transactions in datasets every minute throughout the world, it has become significant to perform Market Basket Analysis to assist wise decision making. Apriori algorithm alone is not effective and sufficient and can lead to misleading association rules. So use of null invariant measures Kulcyznski and IR along with Apriori has been proposed in our methodology supported by experimental results on an electronic store dataset to identify strong and interesting association rules and hence enhance the accuracy of analysis. The size of candidate sets generated with Apriori can be reduced using hash based techniques to improve the efficiency of Apriori. Also the quantity of items purchased in each transaction is not taken into account for analysis and can be considered as part of future work.

REFERENCES

[1] Rakesh Agarwal , Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules" , In Proceedings, 20th International Conference Very Large Databases,487-499
[2] Rakesh Agarwal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of 1993 ACM SIGMOD International Conference on Management of Data,pp.207-2016,Washington, D.C, May 1993
[3] Sotiris Kotsiantis, Dimitris Kanellopoulos , "Association Rules Mining : A Recent Overview" , GESTS International Transactions on Computer Science and Engineering , Vol. 32(1),2006,pp.71-82
[4] Sheila A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation", International Journal of Scientific & Engineering Research Volume 3,Issue 7, July-2012
[5] S. Pradeep Kumar, Mrs C. Grace Padma, "A Technical Analysis of Market Basket by using Association Rule Mining and Apriori Algorithm", International Journal of Innovative Research in Science, Engineering and Technology, Vol 6,Issue 3, March 2017
[6] Adalbert F.X. Wilhelm, "Visual Exploration of Association Rules", School of Humanities and Social Sciences, Jacobs University Bremen
[7] Jiawei Han, Micheline Kamber, Jian Pei, " Data Mining Concepts and Techniques" , 3rd Edition