IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Study on Market Basket Analysis using Apriori and Classification Rule Based Association Algorithm

Saroj A. Shambharkar[1*], Ruchi Bhajipale[2], Neha Nagpure[3], Himanshu Kanoje[4]

[1,2,3,4]Information Technology, Kavilkulguru Institute of Technology and Science, Nagpur, India

Corresponding author: sarojshambharkar123@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.723728 | Available online at: www.ijcseonline.org

Accepted: 12/Mar/2019, Published: 31/Mar/2019

Abstract— The proposed approach is performing market basket analysis using the Apriori algorithm and Classification rule Based Association algorithm (CBA) based on accuracy parameter. The motivation behind the approach is to know the customer's interest towards the products .The objective is to enhance the sales of the business. The approach is using online retail dataset. In the market basket analysis, the frequently purchased items by the customer were analyzed by the admin, the admin keep track of the items purchased by the customers. The admin can also know the purchasing habits of the customers, which will help to improve the quality and quality of the product. This task will be time consuming if performed manually, but the proposed approach reduces the time and improves performance and efficiency of analysis. The analysis helps admin to think about the different strategy to improve the sale. The proposed approach has been implemented in R language. The accuracy is more using Classification rule Based Association algorithm (CBA).

Keywords—Apriori,Classification rule Based Association algorithm, Data mining, Business Data Processing,R programming.

I. INTRODUCTION

Market Basket analysis is a data mining method focusing on discovering purchase patterns of the customers by extracting association or co-occurrences from a store's transactional data set. The market basket analysis in this paper is analysed using two approaches Apriori algorithm and Classification rule Based Association algorithm (CBA)**.**For example, when the person checkout items in a supermarket all the details about their purchase goes into the transaction data set. Then this huge data of many customers are analyzed to determine the purchasing pattern of customers. Also, decisions like object to stock more on cross selling, up selling, store shelf arrangement are determined. Some of the methods include association, classification, and clustering. In this paper, we primarily focus on association and classification.

As the population is increasing day by day, there is increase in competition in the business world. The Shopkeeper remains unaware about how to increase the sales. This has motivated us to develop some approach to increase the business sales. If shopkeeper wants to search how many products are sold there is no need to check one by one.

II. LITERATURE SURVEY

In this section, the author describes the previous research works in the form of title, problem statement, objectives. The objectives of the proposed approach are given below:

- To make more informed decisions about product placement, pricing, promotion and profitability.
- To learn more about customer behavior.
- To find out which products perform similarly to each other, we can know frequently purchased items by the customer.
- To determine which products should be placed near each other.
- To find out which products should be cross-sold.
- To find out if there are any successful products that are not having significant related elements.

In paper [1] authors Yanthy et al stated about the important goal in data mining is to reveal hidden knowledge from data and various algorithms have been proposed for, but the problem is that typically not all rules are interesting –only small fraction of the generated rules would be of interest to any given users. Hence numerous methods such as confidence, support, and lift have been proposed to determine the best or most interesting rules. However some algorithms are good at generating rules high in one measure but bad in other.

In paper [2] authors Rakesh Agarwal and R. Srikant proposed the Apriori algorithm. Apriori was the first associative algorithm proposed and future development in

association, classification, associative classification algorithms has used Apriori as part of the technique. Apriori algorithm is a level-wise, breadth-first algorithm which counts transactions Apriori algorithm uses prior knowledge of frequent item set properties. Apriori uses an iterative approach known as level-wise search, in which n-item sets are used to explore (n+1) - item sets. To improve the efficiency of the level-wise generation of frequent item sets Apriori property is used here. Apriori property insists that all non-empty subsets of a frequent item set must also be frequent. This is made possible because of the anti-monotone property of support measure - the support for an item set never exceeds the support for its subsets. A two-step process consists of join and prune actions are done iteratively. It is one of the Data Mining Algorithm which is used to find the frequent items/item set from a given data repository. The algorithm involves 2 steps

a. Pruning
b. Joining

The Apriori property is the important factor to be consider before proceeding with the algorithm Apriori property states that If an item X is joined with item Y,
Support (XUY) =min (Support(X), Support(Y)).
Basically when we are determining the strength of an association rule i.e. how string the relationship is between the transaction of the items we measure through the use of the support and confidence. The support of an item is the number of transaction containing the item. Those items that do not meet the minimum support are excluded from the further processing. Support determines how often a rule is applicable to a given data set.
Confidence is defined as the conditional probability that a transaction containing the LHS will also contain the RHS.
Confidence (LHS->RHS->)
P (RHS/LHS) =P (RHS∩LHS)/P (LHS) =support (RHS∩LHS)/support (LHS)
Confidence determines how frequently item in RHS appears in the transaction that contains LHS. While determining the rules we must measure these two components as it is very important to us. A rule that has very low support may occur simply by chance. Confidence on the other hand, measures the reliability of the inference made by the rule.

In paper [6] author J. Han and others presented a new association rule mining approach that does not use candidate rule generation called FP-growth that generates a highly condensed frequent pattern tree (fptree) representation of the transactional database. Each database transaction is represented in the tree by at most one path. FP-tree is smaller in size than the original database the construction of it requires two database scans, where in the first scan, frequent item sets along with their support in each transaction are produced and in the second scan, FP-tree is constructed. The mining process is performed by concatenating the patterns with the ones produced from the conditional FP-tree. One constraint of FP-growth method is that memory may not fit FP-tree especially in dimensionally large database.

The author Liu proposed CBA the first Associative Classification (AC) algorithm.CBA implements the famous Apriori algorithm in order to discover frequent rule items.
The Apriori algorithm consists of three main steps:

a. Continuous attribute in the training data set gets discredited.
b. Frequent rule items discovery
c. Rule generation

CBA selects high confidence rules to represent the classifier. Finally, to predict a test case CBA applies the highest confidence rule whose body matches the test case. Experimental result designated that CBA drives higher quality classifiers with regards to accuracy that rule induction and decision tree classification approaches.

In paper [14] the authors Phani Prasad J, Murlidher Mourya stated that there are lots of case studies about the association Rules and existing data mining algorithms usage for market basket analysis but focuses on Apriori algorithm and concludes that the algorithm can be modified and it can be extended in the future work which also decrease the time complexity. Author also clearly states the De-merits of the algorithm but claims that there is the way to improve the efficiency of the algorithm.

The authors Zhixin et al. in their paper recommended an improved classification technique based on Predictive Association Rules. Classification Dependent Predictive Association Rules (CPAR) is one of the types of association classification method which integrates the benefits of associative classification and conventional rule-based classification. For generation of the rule, CPAR is more efficient than the conventional rule-based classification, since most of the replicate calculation is ignored and multiple literals can be selected to create multiple rules at the same time. Although the benefit mentioned above avoids the replicate calculation in rule generation, the prediction processes have the disadvantage in class rule distribution inconsistency and interruption of inaccurate class rules. Further, it is ineffective in instances that satisfy no rules. To avoid these difficulties, the author recommends Class Weighting Adjustment, Center Vector-based Pre-classification and Post-processing with Support Vector Machine (SVM).

The authors Wang and others were suggested a novel rule weighting approach in Classification Association Rule Mining. Classification Association Rule Mining (CARM) is the newest classification rule mining technique that built an association rule mining based classifier by using Classification Association Rules (CARs). The specific CARM algorithm which is used is not regarded, a similar set

of CARs is continually produced from the data, and a classifier is commonly presented as a structured CAR list, depending on a selected rule ordering approach. Several number of rule ordering approaches have been recognized in the recent past, which can be categorized as rule weighting, support-confidence and hybrid. In this approach, an alternative rule-weighting method, called CISRW (Class Item Score based Rule Weighting) and a rule-weighting based rule which orders mechanism depending on CISRW. Later on, two hybrid techniques are added and developed by merging support-confidence and CISRW.

The author Vladimir Bartik in his paper "association based classification for relational data its use in web mining the classification according to the mining association rules is a better and human understandable classification scheme. The intention of the author is to force an alteration of the fundamental association based classification technique that can be used in gathering data from the Web pages. The alteration of the technique and necessary discretization of numeric characteristics are given.

In paper entitled as "Association Rule Mining and Website's Design Improvement" the authors Omari et al. developed a new temporal measure for interesting frequent item set mining. Frequent item set mining helps in searching for powerfully associated items and transactions in large transaction databases. This measure is based on the fact that interesting frequent item sets are typically covered by several recent transactions. This minimizes the cost of searching for frequent item sets by minimizing the search interval. Additionally, this measure can be used to enhance the search approach implemented by the Apriori Algorithm.

The authors Qiang et al. presented an association classification method based on the compactness of rules. Associative classification provides maximum classification correctness and strong flexibility. Simultaneously, this associative classification undergoes a over fitting because the classification rules satisfied the least support and lowest confidence are returned as strong association rules return to the classifier. An innovative association classification technique is based on the presentation of rules, it extends Apriori Algorithm which considers the interestingness, importance and overlapping relationship among rules. Experimental observation proves that the proposed approach has better classification accuracy in comparison with CBA and CMAR.

The authors Sen Guo et al. presented a mechanism called R_Apriori for learning rules from large datasets. The existing rough set based methods are not valid for large data sets owing to its high time and space complexity. Large data sets are separated into numerous parts, in combination with Apriori Algorithm. Implicated rules are obtained in liner relation to the size of the data set. The experimental result shows that this method is better than the existing ones. Apriori Algorithm is one of the classic and best algorithm for learning association rule and its process. Mining association rules is based on Apriori Algorithm and application. In the data mining research, mining association rules is significant and it can be used effectively. This section deals with Apriori Mining Algorithms and the survey done by various researchers on this algorithm. It is clear from the review that the major shortcoming of association rules data-mining is that the support-confidence framework often generates too many rules. Although Apriori algorithm can identify meaningful item sets and construct association rules, it suffers from the disadvantage of generating numerous candidate item sets that must be repeatedly compared with the entire database.

The author Andrej Trnka in his paper "Classification and Regression Trees as a part of Data Mining in Six Sigma Methodology" described the implementation of market basket analysis to Six Sigma methodology. The methods of data mining provide great deal opportunities in the market sector. One of them is market basket analysis. By implementing this into Six Sigma, the results can be improved and the performance level of the process can be changed.

The authors Cunningham et al. in their paper "Market Basket Analysis in library circulation Data" provided a model for library circulation data and applied the Apriori tool for the task of detecting subject classification categories that co-occur in transaction records of the library borrowed books from the university. The results of the paper provide insight into the degree of "scatter" that the classification scheme foster in a particular collection of documents.

The authors Rastogi et al. presented in their paper the optimized association approach on association rules that contain instantiated attributes. To determine the relationship between two items such that the support and confidence of the optimized rule are maximized. He presented effective techniques for pruning the search space while computing optimized association rules for both categorical and numerical data.

The authors Neesha et al. studied the various advancements in the field of data mining. In their paper, they described these advancements starting from the year 2008; a novel frequent pattern generation algorithm had been proposed in order to tackle the data imbalance problem. In 2009, an experiment was performed to compare three association rule mining algorithms: Apriori, Predictive Apriori, and Tertius, on the basis of predictions made on the status of the heart using heart disease data. The results of the experiment showed that Apriori was best suited for this type of data.

In 2010, a new algorithm SC-BF Multilevel was introduced as a better version of Apriori algorithm which was faster and efficient since it required an only a single scan of the database for mining frequent item sets. In 2012, three data mining techniques were applied upon a specific set of data consisting of students' enrolment for the likeliness of the courses to be learned. The techniques involved clustering (k-means algorithm), classification (AD Tree algorithm) and association rule (Apriori algorithm). The comparison depicted that the combined approach is better than using association rule mining alone for such kind of task.

The work carried out in the year 2013 involved following things:

(1) Comparative analysis was done of three algorithms Apriori, Predictive Apriori and Tertius. The author discussed after results, the limitations of the Apriori algorithm and ways to improvise it. This showed that Apriori was faster than the other two algorithms.

(2) The association rule mining technique was applied to a data set consisting of crimes against women. A comparison was made between Apriori and Predictive Apriori on the same data set in which Apriori was found better and faster.

(3) A comparison was made among Apriori, FP-Growth and Tertius algorithm on a super-market data using Weka tool. The results depicted that FP-Growth performed better than the other two algorithms.

Through each research study has its own specific purpose.

The purpose of research is to find the answer to the questions through the application of market basket analysis.

IV. APRIORI AND CLASSIFICATION RULE BASED ASSOCIATION ALGORITHM

The proposed system having following features:-

1. This proposed approach is based on taking online retail dataset .
2. This proposed system shows the information about the products sold and also the customer behavior regarding the product.
3. The shopkeeper can easily find out which products are sold together and which are not sold.

The proposed approach is using the algorithms Apriori and CBA (Classification rule Based Association) for performing market basket analysis. Market basket analysis is applied to various fields of the retail sector in order to boost sales and generate revenue by identifying the needs of the customers and make purchase suggestions to them. Market basket analysis used in the cross selling, product placement, affinity promotion, fraud detection, customer behaviour etc.

The Apriori algorithm finds the most frequent items purchased in the transaction.

In the system the input is a file of data containing the online retail data as shown below.

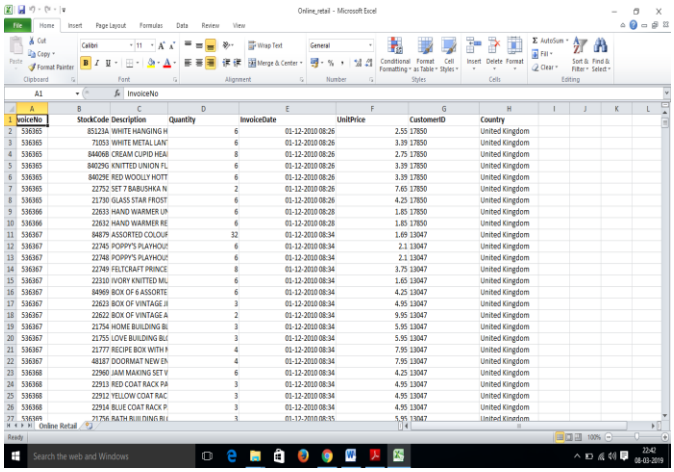

Figure 1. Online retail dataset

The analysis of dataset is performed by R programming using the algorithms Apriori and Classification rule Based Association. R is utilized as a statistical programming environment for solving problems. R tools can also operate as a general matrix calculation toolbox. R provides the freedom of selecting and editing tools to interact with the native console. While scripting in R, user don't need to type commands but rather call functions to achieve results. The R Console allows command editing. The developed R scripts can be executed in the selected editor. User could download the data and save them in a local file, or just cut and paste the data from the browser to an editor such as Notepad, and then save them. By making the use of Classification Rule Based Association algorithm, the analysis is more accurate which can reduce the complexity of the Apriori algorithm.

The online retail dataset is analysed to form the frequent item sets and generate the association rules using Apriori algorithm.

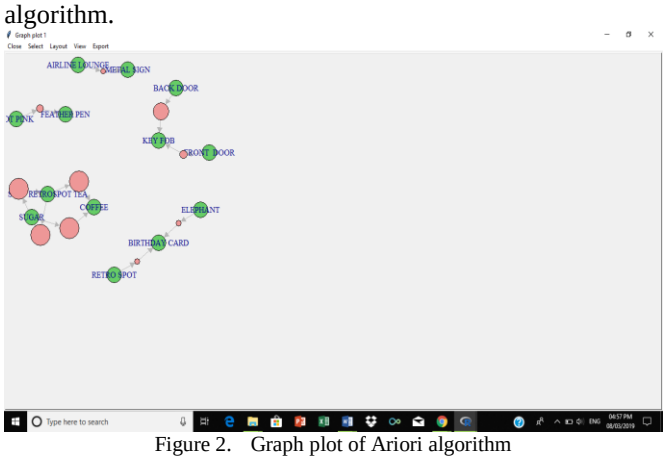

Figure 2. Graph plot of Ariori algorithm

The rules generated using Apriori can be reduced in number by using Classification Rule Based Association in R language, as a result we reduce complexity to some extent. And the comparison between this two based on accuracy obtained will be the final result of the proposed approach.In

the below figure the graph having one central item shown in the graph is connected to those items which the customer purchased along with the centrally located item.

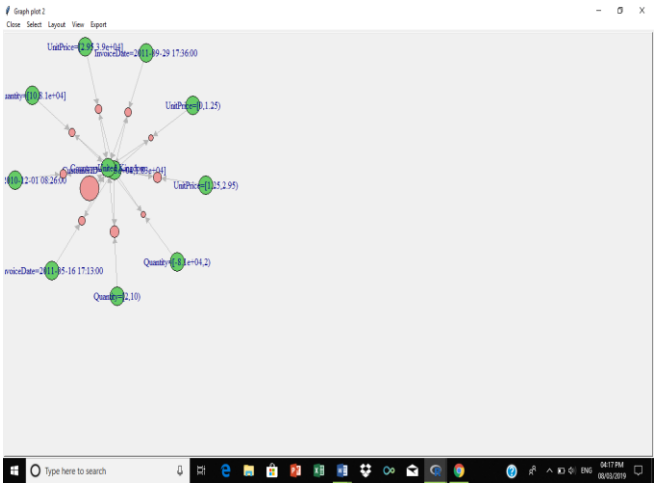

Figure 3. Graph plot of CBA

Following figure shows the scatter plot of apriori algorithm which relates the support and confidence.The dots showing the 10 rules .

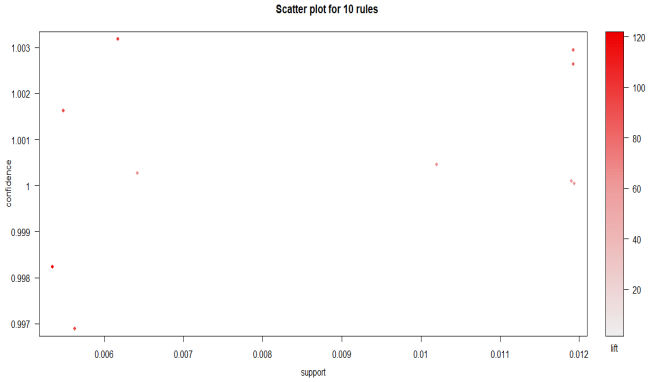

Figure 4. Scatter plot for Apriori Algorithm

The figure 4 showing the item frequency plot to represents which items are frequently bought by the customer in particular transaction. The figure 4 is showing top 20 frequent items in the transaction.

V. RESULT AND DISCUSSION

The main aim of proposed work is finding frequent items and association rules from that items . The result shows the analysis of Apriori algorithm and CBA (Classification rules Based Association) algorithm.The accuracy of the two algorithms were compared.The accuracy which we get from Apriori algorithm is 0.7 and the accuracy of CBA we get is 0.8.

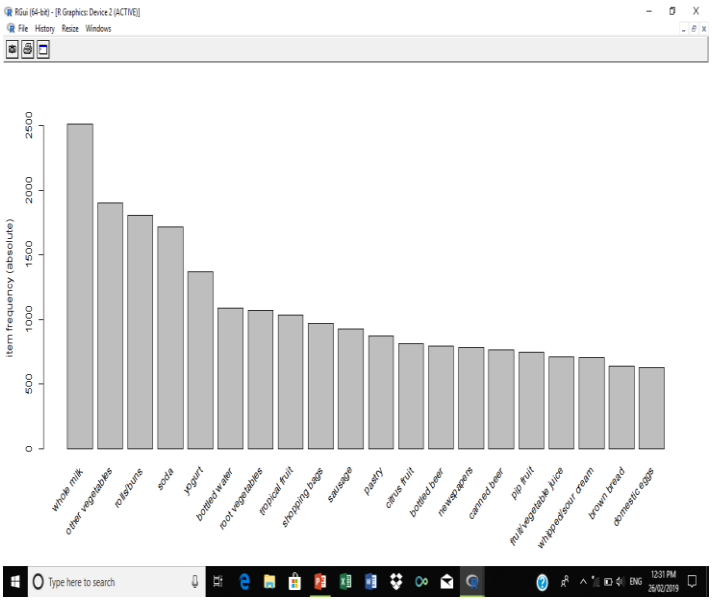

Figure 5. Frequency plot of items by CBA

VI. CONCLUSION AND FUTURE SCOPE

The main purpose of the system is to compare the Apriori and CBA for finding the best algorithm for market basket analysis. Accordingly the retailers can arrange the items in the store for best selling. The result obtained shows the most frequent items purchased by the customers in the transaction. From the obtained results we conclude that the CBA is more accurate as compared to Apriori algorithm.

On the basis of Apriori and CBA algorithm to analyse the data included in the model. Also, the proposed model will be updated by customer survey and questionnaire and can enhance our structure model by surveying customers and generating new results that will give very productive output.

ACKNOWLEDGMENT

We thank Mr. Akshay Ghatole, Assistant Manager, National Payment Corporation of India (NPCI) for his time to time guidance and suggestion.

REFERENCES

[1] W. Yanthy, T. Sekiya, K. Yamaguchi,"*Mining Interesting Rules by association and Classification Algorithms*", FCST **09**.

[2] R. Agrawal and R. Srikant," *Fast Algorithms for Mining Association Rules in Large Databases"*, Journal of Computer Science and Technology, vol. **15**.

[3] X. Yin, J. Han, "*CPAR: Classification based on Predictive Association Rules"*, Proceedings of the Third SIAM International Conference on Data Mining, pp 331-335, **2003**.

[4] Gourab Kundu , Sirajum Munir, Md. Faizul Bari, Md. Monirul Islam, and K. Murase, *"A Novel Algorithm for Associative Classification"*, 14th International Conference, ICONIP **2007**, Kitakyushu, Japan, pp **453-459** , November 13-16, **2007**.

[5] J. Han, J. Pei, and Y. Yin, *"Mining frequent patterns without candidate 1-12, generation"*, Proc of the ACM SIGMOD International Conference on, vol. **1**, pp **2000**.

[6] J. Han, J. Pei, Y. Yin, and R. Mao, *"Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach"*.

[7] Phai Prasad J, Murlidher Mourya, *"A Study On Market basket Analysis Using Data"*.

[8] Yen-Liang Chen, Kwei Tang, Ren-Jie Shen, Ya-Han Hu, *"Market basket analysis in a multiple store environment"*, SciVerse ScienceDirect, Volume **40**, Issue **2**, August **2005**, Pages **339-354** .

[9] Raorane A.A, Kulkarni R.V, and Jitkar B.D, *"Association Rule – Extracting Knowledge Using Market Basket Analysis"*, Research Journal of Recent Sciences, Vol. **1(2)**, **19-27**, Feb. (**2012**).

[10] Christian Borgelt, *"An Implementation of the FP-growth Algorithm"*.

[11] Sotiris Kotsiantis, Dimitris Kanellopoulos, *"Association Rule Mining: A Recent Overview"*, GESTS International Transactions on Computer Science and Engineering", Vol. **32**(1), **2006**, pp. **71-82.**

[12] J. Han, H. Pei, and Y. Yin. *"Mining Frequent Patterns without Candidate Generation"*, In Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA **2000**.

[13] https://en.wikipedia.org/wiki/Association_rule_learning.

[14] Phani Prasad, MurlidherMourya, *"A Study on Market Basket Analysis Using a Data Mining Algorithm"*, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Vol **3**, Issue **6**, June **2013**.

[15] Akanksha Singh, K. K. Singh, *"Data Mining and Data Warehousing"*, India: Umesh Publications**, 2011-2012**.

[16] Harpreet Kaur, Kawaljeet Singh, *"Market Basket Analysis of Sports Store using Association Rules"*, International Journal of Recent Trends in Electrical & Electronics Engg.,ISSN: 22316612, Dec**. 2013**.

[17] Neesha Sharma, C. K. Verma, "Association Rule Mining: An Overview", IJCSC, Volume **5**, Number **1**, March **2014**, pp.**10-15**, ISSN-**0973-7391**.

