Volume: 6 Issue: I Month of publication: January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1270

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

Comparative Study and Analysis of Wholesale Customer's Dataset Using Association Rule Mining

Vijayakumar. M [1], R. Porkodi[2]
[1]*PG Student, Department of computer science, Bharathiar University, Coimbatore.*
[2]*Assistant professor, Department of computer science, Bharathiar University, Coimbatore.*

Abstract: Association Rule Mining (ARM) has always been the area of interest for many researchers for a long time and continues to be the same. It is one of the important tasks of the data mining concept. It aims at discovering relationships among various items in the database. The datamining is the computing process of discovering patterns in large datasets. It is based on complex algorithms that allow segmentation of data to identify pattern and trends, detect anomalies, and predict the probability of various situational outcomes. The Data mining trends includes: Distributed Data Mining (DDM), Multimedia Data Mining, Spatial and Geographic Data Mining, Time series and sequence data mining. This paper is based on Association rule mining and its methodology. The main objective of this paper is to present a review on the basic concepts of ARM technique and its algorithms. In this paper, the association rule mining algorithms namely Apriori, Predictive Apriori and Filtered Associator is being implemented in the Whole sale customer's dataset and the performance of these algorithms are compared and analyzed deeply.

Keywords - Association rule mining, Weka, Apriori, Predictive Apriori, Filtered Associator.

I. INTRODUCTION

Data mining is defined as the computing process of extracting interesting information or patterns from large information repositories such as: relational database, data warehouses etc. The list of areas where data mining is widely used is: Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Intrusion Detection and other Scientific Applications. The main objective of the data mining concept is to extract information from a large data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data preprocessing model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization and online updating. Data mining is referred as the analysis step of the "Knowledge Discovery in Databases" process (KDD).Recently, there are several major data mining techniques that have been developed and used in data mining projects including classification, clustering, association rules, prediction, sequential patterns and decision tree. The Association rule is one of the best-known data mining techniques. In association rules, a pattern is discovered based on a relationship between items in the same transaction. The association rule technique is used in market based analysis to identify a asset of products that customers frequently purchase together. The main focus of this paper is to find the number of occurrence and frequent items in the data set by using various algorithms in association rule mining. The paper deals with finding the number of occurrence of the items, so that the preferred items of the customer can be easily detected and can increase the stock of the large number of sold product. The combinational product preferred by the customer can be detected by means of association rule mining algorithms. This paper will focus on the analysis of data from the data set called wholesale customer's data set [1].

The section I discuss about the introduction of data mining and the Association rule mining. Section II gives the literature review of the various journals. Section III explains the methods that are used in Association rule mining algorithms. The results and discussion are explained in Section IV and Section V concludes this analysis work.

II. LITERATURE REVIEW

Many journals and articles concerning association rule mining algorithms were studied from year 2013 to 2016. Some compared association rule mining algorithms while some modified the existing algorithms to improve the performance.

Anjana Gosain and Maneela Bhugra [2] presented apaper that described different algorithms given by various researches to generate association rules among quantitative data. They have performed a comparative analysis of different algorithms such as Apriori, CT-Apriori, FP-Tree for association rules based on various parameters. The comparative study depicts the advantages and disadvantages of the algorithms.

Dhiren R. Patel and Khyati B. Jadav[3] presented a paper about the algorithms developed by researchers for Boolean and Fuzzy association rule mining such as Apriori,FP-tree, Fuzzy FP-tree etc. It concluded that Apriori algorithm combined with other fuzzy association rule mining algorithms can overcome most of the problems faced by the algorithms.

Khyati B. Jadav and Jignesh Vania [4] presented a survey paper based on the existing approaches such as heuristic approach in association rule hiding, along with some open challenges. The future scope is to find hybrid technique to reduce the side effects.

Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang [5] studied the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework. They proposed an attack model such as item based and set based attack on background knowledge and devise a scheme for privacy preserving outsourced mining. The future scope is to improve the attack model to minimize the vulnerabilities.

T. Karthikeyan and N. Ravikumar [6] presented a paper on theoretical survey of the existing algorithms such as Apriori, AIS, Genetic algorithm etc. The concepts behind association rules are provided at the beginning followed by an overview to some of the previous research works done on this area. The advantages and limitations are discussed and concluded with an inference.

Gurneet Kaur et al [7] presented a paper to give an idea on the basic concepts of ARM technique such as AIS, Apriori, and FP-Growth along with the recent, related work that has been done in this field. The advantages and disadvantages along with some issues of the algorithms are determined. The future scope is to solve the issues related to this field.

Xu He, Fan Min, and William Zhu [8] presented a paper to study the impact of discretization approaches and granular association rules on mining semantically richer and stronger rules from numerical data on basis of association rules. It consists of evaluation and comparison of discretization approaches to granular association rule mining. The future scope is to develop more appropriate discretization approaches.

Prof. Ashish Mishra, Kuldeep Tripathi and Neelkamal Upadhyay [9] presented a paper that provides a survey of association rule hiding methods for privacy preservation and it surveys current existing techniques such as heuristic approach, border based approch, exact approach, reconstruction based approach and cryptographic approach for association rule hiding.

J. L. Dominguez and J. Mata [10] presented a paper based on deterministic method for extraction of quantitative association rules. Several experiments have been performed using Apriori algorithm. It states that the Apriori algorithm is best suited to work under circumstances.

P. K. Mishra, Sudhakar Singh, Pankaj Singh and Rakhi Garg [11] presented a paper representing the journey of ARM algorithms started from sequential algorithms, and through parallel and distributed, and grid based algorithms to the current state-of-the-art, along with the motives for adopting new machinery.ARM algorithms have redesigned on Map Reduce framework for processing large scale data.

III. METHODOLOGY

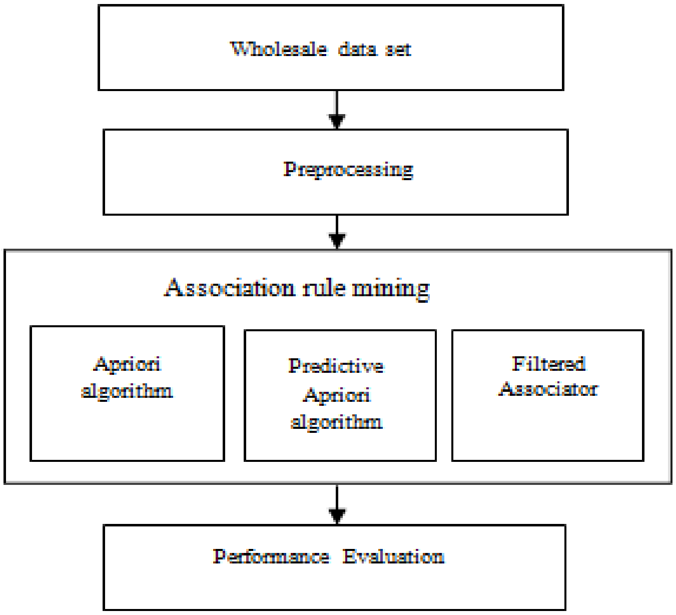

Fig.1 flowchart

The flow chart fig.1 determines the flow of data in the process of determining the wholesale dataset undergoing the preprocessing stage and then the preprocessed data is further subjected to various association rule algorithms such as Apriori, Predictive Apriori, and Filtered Associatorin datamining to obtain the desired results using the Weka tool[12].

A. *Association Rule Mining*

Association rule mining is a method that is used to find frequent patterns, correlations, associations, or causal structures from large data sets which are found in various kinds of databases like relational databases, transactional databases, and data repositories. Association rules are being used widely in various areas such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc. Association rules are the statements that find the relationship between data in any database. The association rule mining can be viewed as a two step process:

1) Finding all the frequent item sets.
2) Generate strong association rules from the frequent item sets

The association rule mining is calculated by means of support and confidence. The support and confidence is determined in order to find out the association rules that satisfy the predefined minimum support and confidence from a given database [13] [14].

3) *Support:* Support is a process to find the frequent occurrence of the items in the dataset. The support is calculated by means of [15]:

$$\mathrm{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

4) *Confidence* Confidence is a process to find the combination of the items in the dataset. The confidence is calculated by means of [16]:

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

B. *There are a few commonly used terms that must be defined, they are*

1) *Itemset*: An Itemset is a set of items. A k-Itemset is an Itemset that contains k number of items.
2) *Frequent Itemset*: This is an itemsets that has minimum support.
3) Some other interestingness measures are:
4) Expected Predictability: The frequency of occurrence of the item Y is said to be its expected predictability
5) Lift: It is the ratio of confidence or predictability to expected confidence or expected predictability i.e. the number of transactions that include the consequent or the right hand side of the rule divided by the total number of transactions.

C. *Association Rule Mining Algorithms*

1) *Apriori algorithm:* The Apriori algorithm is an influential algorithm that is used for mining frequent item sets for Boolean association rules. Apriori algorithm is mainly used to determine the frequent item set mining in the transactional databases. It uses bottom up approach .It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. The Apriori Algorithm displays the result such as the schema, relation, instances and the attributes along with their minimum support for instances, minimum metrics (confidence), the number of cycles, the size of the generated set of large Itemset and the best rules are determined [17].
2) *Predictive Apriori Algorithm:* N Predictive Apriori algorithm use larger support and traded with higher confidence, and calculate the expected accuracy in Bayesian framework. The result of this algorithm maximizes the expected accuracy for future data of association rules. This algorithm generates association rules as expected number of rules by user. The Predictive Apriori Algorithm displays the result such as the schema, relation, instances and the attributes along with their minimum support for instances, minimum metrics (confidence), the number of cycles, the size of the generated set of large Itemset and the best rules are determined [18].
3) *Filtered Associato:* The Filtered Associator algorithm is a class for running an arbitrary Associator on data that has been passed through an arbitrary filter. Like the Associator, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. It includes option such as Associator with which we

can consider the Apriori, Predictive Apriori and Filtered Associator algorithm, class index and filter to get the result. The Filtered Associator displays the result such as the schema, relation, instances and the attributes along with their minimum support for instances, minimum metrics (confidence), the number of cycles, the size of the generated set of large Item set and the best rules are determined [19].

IV.RESULTS AND DISCUSSIONS

The Wholesale customer's dataset is compiled from data collected from Weka dataset. Only 8 attributes and 440 instances from the database are considered for the wholesale required for the annual spending. The following attributes with nominal values are considered: fresh, milk, grocery, frozen, detergents and paper, delicatessen, channel and region.

1.Wholesale customer's Dataset

ATTRIBUTE	*DESCRIPTION*	*POSSIBLE VALUES*
Channel	*Customer Channel*	*Numeric*
Region	*Customer Region*	*Numeric*
Fresh	*Annual Spending (M.U) On Fresh Products*	*Numeric*
Milk	*Annual Spending (M.U) On Milk Products*	*Numeric*
Grocery	*Annual Spending (M.U) On Grocery Products*	*Numeric*
Frozen	*Annual Spending (M.U) On Frozen Products*	*Numeric*
Detergents And Paper	*Annual Spending (M.U) On Detergents And Paper Products*	*Numeric*
Delicatessen	*Annual Spending (M.U) On Delicatessen Products*	*Numeric*

The Wholesale customer's dataset is imported from the UCI repository and then subjected to preprocess. The preprocessing is done by choosing the Numeric-to-Nominal under the attributes category of the filters. The Numeric to Nominal option is selected in order to convert all the numeric values to nominal. Each item is selected and the results such as label, count, missing percent, distinct values and unique percent is determined.

The Apriori Algorithm displays the result such as the schema, relation, instances and the attributes along with the details such as thesize of the generated large item sets. The minimum support 0.1 is taken as a constant value and the confidence values are changed to 0.1, 0.5, 0.9 and the results are compared. It produces the rules count as 6, 4 and 2 respectively. The number of cycles performed is 18.The Filtered Associator Algorithm displays the result such as the schema, relation, instances and the attributes along with the details such as thesize of the generated large item sets. The minimum support 0.1 is taken as a constant value and the confidence values are changed to 0.1, 0.5, 0.9 and the results are compared. It produces the rules count as 6, 4 and 2 respectively. The number of cycles performed is 18.

The Predictive Apriori Algorithm displays the result such as the schema, relation, instances and the attributes along with the number of cycles performed is 1. The minimum support 0.1 is taken as a constant value and the confidence values are changed to 0.1, 0.5, 0.9 and the results are compared. It produces the rules count as 16, 14 and 12 respectively.

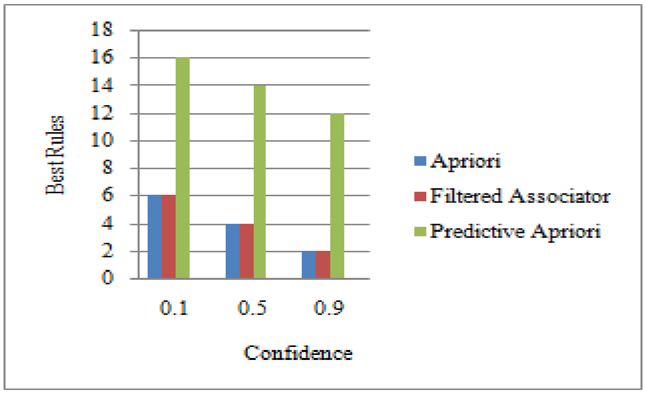

Fig.2Confidence Vs Rules

The fig.2 shows the rules determined by the association rule algorithms for various confidence values. From this bar graph the Predictive Apriori algorithm is determined as the best algorithm producing the highest number of best rules than the Apriori and Filtered Associator algorithms.

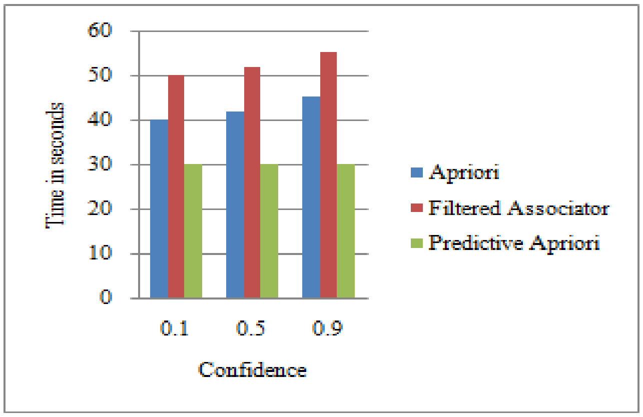

Fig.3Confidence Vs Time

The fig.3 shows the time taken to process the association rule algorithms for various confidence values to produce the desired best rules. The Predictive Apriori algorithm takes 30 seconds to run the experimental dataset whereas Apriori and Filtered Associator algorithms consume 45 and 55 seconds respectively.

V.CONCLUSION

Association rule mining is used for finding frequent patterns, co-relation among the items in the database. The objective of this study is to evaluate and investigate three selected algorithms based on the association rules. This paper reviewed the research done by the various author in this field. The extensive survey has been conducted in association rule mining area and analyzed various association rule mining algorithms used so far. In this paper I have implemented three association rule mining algorithms using Wholesale customer’s dataset and compared the performance of the association rule algorithms using support and confidence metrics.It is observed that thePredictive Apriori algorithm is the best as it produces the best rules and takes less time to execute the dataset. The second best algorithm is the Apriori algorithmproducing the less number of best rules but takes more seconds than Predictive Apriori algorithm. The third best algorithm is the Filtered Associator followed by the Apriori Algorithmproducing the least number of best rulesand it takes too more seconds to execute. Hence, the time complexity is an issue to run the algorithms.

REFERENCES

[1] Big-datamadesimple.co
[2] Anjana Gosain and Maneela Bhugra “A Comprehensive Survey Of Association Rules On Quantitative Data In Data Mining” Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013

[3] Sudhakar Singh, Pankaj Singh, Rakhi Garg and P. K. Mishra" Mining Association Rules in Various Computing Environments: A Survey" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 8 (2016) pp 5629-564

[4] T. Karthikeyan and N. Ravikumar2 "A Survey on Association Rule Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 201

[5] Gurneet Kaur" Association Rule Mining: A Survey" Gurneet Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2320-232

[6] Khyati B. Jadav, Jignesh Vania and Dhiren R. Patel, PhD" A Survey on Association Rule Hiding" International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 201

[7] Kuldeep Tripathi, Neelkamal Upadhyay and Prof. Ashish Mishra" A Survey of Association Rule Hiding Approaches" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555Vol. 5, No1, February 201

[8] Khyati B. Jadav, Jignesh Vania and Dhiren R. Patel, PhD" A Survey on Association Rule Hiding Methods" International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 201

[9] Xu He, Student Member, IEEE, Fan Min, Member, IEEE, and William Zhu, Member, IEEE" Comparison of Discretization Approaches for Granular Association Rule Mining" Canadian Journal Of Electrical And Computer Engineering, Vol. 37, No. 3, Summer 201

[10] J. L. Dominguez and J. Mata" Comparison of Standard Discretization with a New Method for Quantitative Association Rules" IEEE Latin America Transactions, Vol. 14, No. 4, April 201

[11] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang" Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases" IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 201

[12] Agrawal R., Imielinski, T., and Swami, "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.

[13] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm

[14] https://en.wikipedia.org/wiki/Apriori_algorithm

[15] AnisSuhailis Abdul Kadir, Aurelia Abu Bakar and Abdul RazakHamdan, "Frequent Absence and Presence Itemset for Negative Association Rule Mining ", IEEE, 2011.

[16] Guimei Liu, Haojun Zhang and Limsoon Wong, "Controlling False Positives Iin Association Rule Mining" In Proceedings of the VLDB Endowment ACM, 2011.

[17] https://www.techopedia.com/definition/30306/association-rule-mining

[18] http://zyberlord-biju.blogspot.in/2011/06/rule-mining-algorithm-tertius.html

[19] G. Piatetsky-Shapiro, Proc. AAAI-91 Workshop on Knowledge Discovery in Databases, Anaheim, California, July 1991.