International Journal of Electrical, Electronics and Computer Engineering **4**(2): 134-138(2015)

ISSN No. (Online): 2277-2626

Association Rule Mining -Various Ways: A Comprehensive Study

Udaiveer Singh Parmar*, Prof. Anand Motwani** and Prof. Anurag Shrivastava***

**Department of Computer Science & Engineering, NRI Institute of Research and Technology, Bhopal, (MP), INDIA*

***Assistant Prof & Head, Department of Computer Science & Engineering, NRI Institute of Research and Technology, Bhopal, (MP), INDIA*

****Assistant Professor, Department of Computer Science & Engineering, NRI Institute of Research and Technology, Bhopal, (MP), INDIA*

(Corresponding author: Udaiveer Singh Parmar)

(Received 23 October, 2015 Accepted 12 December, 2015)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Nowadays various effective algorithms, which are working on the principles of association rule mining, are available. Here, in this paper we have discussed several fundamentals of association rule mining. Association rule mining technique shows the connections in between items in a data set. The technique detectsfor relationships among items or transactions in data set for drawing inference and taking effective decisions within that domain.The paper offers a comparative study of distinct association rule mining methods.The main motive of this paper is to present theoretical survey offew existing algorithms like genetic, fuzzy, rough and etc. The theory behind association rules is discussedin starting sections while overview of previous research works done on this field is discussed later in the article.

Keywords: Data mining, Association Rule Mining.

I. INTRODUCTION

Data mining [1] refer as an activity of observing data with distinct perspectives and précising it into important information. Data mining is an analytical tool for observing data. It permits users to observe data, categorize it, and summarize the relationships among data.Data mining is all about searching correlations or patterns in large relational databases. It involves some techniques like anomaly detection, clustering, association rule learning, regression, summarization, classification etc.

Association rule learning searches for relationships among factors. For example a supermarket may analyzedata about how the customer purchases the various products in the supermarket and what are the certain patterns to do it. With the help of association rule mining, the online sellers or supermarket can identify which products are often bought together and this information may be used for marketing purposes. This phenomenonis known as market basket analysis. Clustering discovers the groups and structures in the data in some way or another similar wayin absence of implementing the familiar structures in the data. Classification generalizes known structures to apply classification category to new data. Take an example; an e-mail program may desire to classify an e-mail as "legitimate" or as "spam" mailon the basis of its contents.

Regression tries to detect a function which controls the data having minimum error. Summarization offers highly efficient representation of data set, which includes visualization and report generation.

II. ASSOCIATION RULE MINING

Here in data mining, association rule learning known as a famous and well researched technique for detecting interesting connections in between variables ofhuge databases. It identifiestough rules taken in databases through distinct measures of interestingness [2]. Depend on the theory of strong rules[3], we can say that important andbroadly-used example of association rule mining is Market Basket Analysis. To generate all effective association rules, compare given confidence & support with user-specified lessen support and minimum confidence.

$$Rule: \; X \Rightarrow Y \quad \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \end{cases}$$

Where X and Y are different set of attributes.

A. Kinds of Association Rule Mining

Based on the number of data dimensions involved, we can distinguish association rules on basis of dimensions:

- Single-dimensional association rule

An association rule is a single-dimensional, if the items or attributes in an association rule reference only one dimension. For example, if X is an itemset, then a single-dimensional rule could be rewritten as follows: buys(X, "bred"))→ buys(X, "milk").

- Multidimensional association rule:

If a rule references more than one dimension, such as the dimensions like study-level, income, and buys, then it is a multidimensional association rule. Let X an itemset, the following rule is an example of a multidimensional rule: Study-Level(X, "20…25")→income(X, "30K…. 40K")) buys(X, "performant computer"): Based on the types of values handled by the rule, we can distinguish two types of association rules:

- Boolean association rule:

A rule is a Boolean association rule, if it involves associations between the presence or the absence of items. For example, the following rule is a Boolean association rules obtained from market basket analysis: buys(X, "computer"))→buys(X, "scanner").

- Quantitative association rule: a rule is called quantitative association rule, if it describes associations between quantitative items or attributes. In these rules, quantitative values for items or attributes are partitioned into intervals. For example, the following rule is a quantitative association rules:

Study-Level(X, "20…25")→income(X, "30K….40K")→ buys(X, "performant computer").

B. Goal of Association Rule Mining Methodologies

Association rule intent toclean the actual database in order to obtain the following goals:

1. There is no rule taken as sensitive from their owner's perspective and taken from the actual database at pre-specified thresholds of confidence and support. These rules should also declarefrom its sanitized database, when such database is owned at the same or some higher thresholds.

2. Every non sensitive rule that present at the time of mining the actual database at pre-specified thresholds of confidence and support are being successfully owed from the sanitized database at the same thresholds or higher.

3. There is no rule, which is not taken from the original database at the time of mining the data at pre-specified thresholds of confidence and support, which drawn from its sanitized counterpart while mining at the same or at higher thresholds.

The initial goal needs that every sensitive rulemust be taken from the sanitized database at the time of data mining under similar or higher levels of support and confidence like the actual database.

Secondgoal prescribed that there in the sanitized database no rules should be removed. That is, overall non sensitive rules which were mined from its actual database must be taken from its sanitized counterpart at the similar or higher levels of confidence and support.

The last goal describes that no false rules are called as ghost rules should be created when the sanitized database is taken at the similar or higher levels of confidence and support. A false (ghost) rule referred as the association rule which is not from the rules mined from the actual database.

A solution arrived from these three goals is corrected solution. Practical Mining solutions becomes the reasons of minimum possible modification in the actual database are called as ideal or optimal solutions. Non-exact butfeasible solutions are known as approximate.

III. LITERATURE SURVEY

There is lot of work done in Association Rule Mining (ARM). These works are categories into various groups which are given below:

1. Genetic based ARM
2. Fuzzy based ARM
3. Rough based ARM
4. Others

1. Genetic Based

Xin Li *et al* [5] proposed Frequent Itemsets Mining in Network Traffic Data. They think about the problem of frequent itemset mining problem in network traffic data, and propose an algorithm for mining frequent itemsets. They try to minimize the size of results and only maximal frequent itemsets are considered. To protect the privacy, intermediate mining results are encrypted using hashing method by different servers. The proposed algorithm is evaluated from the perspectives of accuracy and efficiency.

Mining of frequent itemsets using Genetic algorithm as proposed in [6]. This work carried out with logic of GA to improve the scenario of frequents itemsets data mining using association rule mining. The main benefit of using GA in frequent itemsets mining is to perform global search with less time complexity. This scheme gives better results in huge or larger data set. It is also simple and efficient.

Another frequent itemsets mining approach based on genetic algorithm for non binary dataset was proposed by G. VjiayBhaskar et al [7]. They present an efficient algorithm for generating significant association rules among database items.GA is used to improve the scenario and system can predict about negative attributes in generated rules. As per results obtained this scheme is simple and efficient one. The Time complexity of the algorithm is also less and suitable for non binary data sets.

2. Rough Set

The association rule is one of the main approaches of data mining but the number of rules discovered usually large while a small number of rules is actually useful from the user view. There are many proposed methods to choose the useful rules by using the utility measure as the interesting measure [8].These measure which is based on rough set theory is proposed by theJiye Li. Jiye Li include the Rule Important Measure - RIM [9] and the enhance rule important measure - ERIM [10].

To the problem with clear decision-making field by association rulesmining,improved R_Apriori algorithm can be form by integrating rough set theory with the Apriori algorithm.The problem raised in the preamble can be solved as follows:

About the problem of the efficiency of Apriori algorithm and the validity of the mining rules on account of the large amount of attributes set, we can first get the nuclear of attribute set by rough set attribute reduction, then the association rule mining set to the nuclear data.

In certain extent, it can improve the efficiency and effectiveness of mining; for inefficient Apriori algorithm raised from the needs of scanning all attribute sets to obtain each frequent attribute set.

Specifically, there are varieties of work under Rough Set Theory in Association Rule Mining; some of those are as follows:

(i) An Improved Apriori Arithmetic based on Rough set Theory [18]

Rough set theory in association with association rules algorithm is specially meant for search implicit rules especially from large data. There is lot of application of the association rule mining. But sometime it faces some issues like less efficient or inaccurate results. In this paper, we talk about R_Apriori algorithm which handles decision-making domain. In this process, first of all it consider or remembercode or very effective conditions in mind then based on the criticalness or importance of various conditions ,the R_apriorialgorithm reduces the number of attributes. Further it reduces tends to variety of efficiency of algorithm. The main ofR_Apriori algorithm is to solve the problems of Apriori algorithm to improve the efficiency of the algorithm.

(ii) Association Rules Mining Algorithm Based on Rough Set [19]

This work puts forward AssociationRule Mining through Rough Set theory, which applies the improved Apriori algorithm. In the association rules mining based on Rough set theory, this process is done by using Decision Table. The use of rough set theory has its own advantages and these advantages are as follows: (i) it eliminates redundant attributes, (ii) it reduces the number of attributes. This deduction is done in single shot whenscanning Decision table and this is done as in decision attribute sets.

3. Fuzzy Set

Srikant and Agrawal [11] used equidepth partitioning to mine quantitative rules. They separate intervals by their relative ordering and quantities equally. Miller and Yang applied Birch clustering [12] to identify intervals and proposed a distance-based association rule to improve the semantics of intervals. Lent,et al [13] presented a geometric-based algorithm to perform clustering for numerical attributes. Finally, Guha, et al [14] proposed an efficient clustering algorithm called CURE.

Another trend to handle this problem is based on fuzzy theory. In contrast to quantitative clustering, fuzzy linguistic-based approaches focus on qualitative filtering. For instance, PradnyaMuley and Anniruddha Joshi [15] introduced fuzzy linguistic summaries on different attributes. Hirota and Pedrycz [lo, 211 proposed a context sensitive fuzzy clustering method based on fuzzy C-means to construct rule-based models. However, context-sensitive fuzzy C-means method cannot handle data consisting of both numerical and categorical attributes. To solve the qualitative knowledge discoveryproblem, [16]Peng Chen, Hongye Su, LichaoGuo and Yu Qu [16] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, [17].

Saroj and NishantPrabhat proposed the F-APACS method [17] to discover fuzzy association rules. They utilized adjacent difference analysis and fuzziness in finding the minimum support and confidence values instead of having them supplied by a user.

To generate fuzzy association rules, all sets of items that have a support above a user specified threshold,should be determined first. Itemsets with at least a minimum support are called frequent or large itemsets. The process alternates between the generation of candidate and frequent itemsets until all large itemsets are identified. The above process is used to calculate the fuzzy support value of itemsetZ and its corresponding set of fuzzy sets F. This way, the problem of mining all fuzzy association rules converts to generating each rule whose confidence is larger than the user specified minimum confidence. Explicitly, each large itemset, say L, is used in deriving all association rules.

4. Others

By Idheba Mohamad *et al.* [4] "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Item sets" here the author explains the Association rule mining taken as the crucial process in data mining. The general concept of association rules is to mine the positive frequent patterns from the overall transaction database. by, mining the negative patterns has drawn the attention of researchers in this sector too. The motive of this survey is to generate latest model for mining interesting negative and positive association rules from the transactional data set. The model introduced here is integration in between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to introduces a latest approach (PNAR_IMLMS) for mining both negative and positive association rules regarding the various interesting frequent and infrequent item sets mined through the IMLMS model. The results displayed that the PNAR_IMLMS model offers effective results as compare to previous model.

As infrequent item sets become more significant for mining the negative association rules that play an important role in decision making, this study proposes a new algorithm for efficiently mining positive and negative association rules in a transaction database. The IMLMS model adopted an effective pruning method to prune uninteresting item sets. An important measure VARCC is taken into action in order to overlook creating uninteresting rules which might be discovered, at the time of mining positive and negative association rules.

To the algorithm of the positive and negative association rules, this paper use linked list to implement the algorithm of the positive and negative association rules. To the algorithm of the positive and negative association rules, there are more research to do. These research work can be done in various domains like…(l) How to optimize the searching space? (2) To the mine the rules, there are further to do the research and use relativity.(3) In the positive and negative association rules, Apriori algorithm displace and improvement and so on.

IV. CONCLUSION

In this paper, the authors have discussed about the algorithm aspects of the association rule mining. Firstly, we have gone through various aspect of association rule mining along with it'stypes and various goals. In the later part, we have examined the various ARM work based on various aspect like genetic, fuzzy, rough and others.

REFERENCES

[1]. Gaurab Tewary, "Effective Data Mining for Proper Mining Classification using Neural Networks," *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.**5**, No.2, March 2015.

[2]. Shital KSomawar and Prof. VanitaBabanne, "Privacy Clustered Mining of Association Rules in Distributed Database by Using Fuzzyfication," *International Journal of Innovative Research in Computer and Communication Engineering*," Vol. **3**, Issue 5, May 2015.

[3]. HimaniBathla and KavitaKathuria, "Association Rule Mining: Algorithms Used*," JCSMC*, Vol. **4**, Issue. 6, June 2015.

[4] IdhebaMohamad Ali O. Swesi, Azuraliza Abu Bakar, AnisSuhailis Abdul Kadir, "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets *"* in *proceeding of 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.*

[5]. X. Li, X. Zheng, J. Li, and S. Wang —Frequent itemsets mining in network traffic data, 2012 Fifth *International Conference on Intelligent Computation Technology and Automation*, pp. 394- 397, 2012.

[6]. S. Ghosh, S. Biswas, D. Sarkar, and P. P. Sarkar —Mining frequent itemsets using genetic algorithm, *International Journal of Artificial Intelligence & Applications* (IJAIA), Vol.**1**, No.4, pp. 133 – 143, October 2010.

[7]. G. Vijay Bhasker, K. Chandra Shekar, and V. Lakshmi Chaitanya —Mining frequent itemsets for non binary data setusing genetic algorithm, *International Journal of Advanced Engineering Sciences and Technologies (IJAEST)*, ISSN: 2230-7818, Vol. **11**, Issue No. 1, pp. 143 – 152, 2011.

[8]. K. Suresh and V. Pattabiraman, "An Improved Utility Itemsets Mining With Respect To Positive And Negative Values Using Mathematical Model," *International Journal of Pure and Applied Mathematics*, Volume **101** No. 5 2015, 763-772.

[9]. Jiye Li, Nick Cercone ,"Discovering and Ranking Important Rules. Granular Computing," *IEEE International Conference on* Volume **2,** (2005).

[10]. Jiye Li, Nick Cercone, W. H . Wong, Lisa Jing Yan,"Enhancing Rule Importance Measure Using Concept Hierarchy. Faculty of Computer Science and Engineering," York University, 2009.

[11]. R. Srikant and R. Agrawal. "Mining quantitative association rules in large relational tables," *Proc. of ACM SIGMOD*,pp1-12, 1996.

[12]. R.J. Miller and Y. Yang, "Association Rules over Interval Data," Proc. Of ACM SIGMOD, pp.452-461, 1997.

[13]. B. Lent, A. Swami and J. Widom "Clustering Association Rules," *Proc. of IEEE ICDE!*, pp.220-23 1, 1997.

[14]. S. Guha, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," Information Systems, Vo1.**26**, No.1, pp.35-58,2001.

[15]. Pradnya Muley and Anniruddha Joshi ,"Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence, " *International Journal of Innovative Research in Advanced Engineering (IJIRAE),* ISSN: 2349-2163 Issue 4, Volume **2** (April-2015).

[16]. Peng Chen, Hongye Su, LichaoGuo and Yu Qu , "Mining fuzzy association rules in data streams," IEEE, 16-18 April 2010.

[17]. Dr. Saroj and NishantPrabhat,"A Genetic-Fuzzy Algorithm to Discover Fuzzy Classification Rules for Mixed Attributes Datasets," *International Journal of Computer Applications* **34**(5):15-22, November 2011.

[18]. Chen Chu-Xiang, Shen Jian-jing, Chen Bing, Shang Chang-xing and Wang Yun-cheng "An Improvement Apriori Arithmetic based on Rough set Theory," Circuits, Communications and System (PACCS), *IEEE, 2011 Third Pacific-Asia Conference on,17-18 July 2011*.

[19]. XUN Jiao, XU Lian-cheng and QILin "Association Rules Mining Algorithm Based on Rough Set,*"2012 International Symposium On Information Technology In Medicine And Education, 2012 IEEE.*

[20]. N K Kameswara Rao and G P SaradhiVarma,"A hybrid Algorithm for Epidemic Disease Prediction with Multi Dimensional Data," *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume **4**, Issue 3, March 2014.

[21]. Divya Tomar and Sonali Agarwal, "A Survey on Pre-processing and Post-processing Techniques in Data Mining," *International Journal of Database Theory and Application,* Vol.**7**, No.4 (2014), pp.99-128.