Survey on Techniques of Association Rule Mining

Syeda Baseer Unnisa Begum
Assistant Professor', Dept of Computer Science(MCA)
R.G.Kedia college of commerce, Opp: Chaderghat Bridge,Hyderabad, Telangana State.

Abstract- *Data mining has become considerably remarkable research topic in the field of computer science since the past decade. Data mining is a process in which extraction of proper, useful and potential knowledge and interesting facts from huge set of large data repositories. There are some approaches to carry out this process such as clustering, classification, time series analysis, and association rule mining etc. In these approaches association rule mining became more popular because of its wide range of applications. Association rule mining is a process of discovering the frequent patterns, associations and correlations between itemsets in information repositories and other databases. There are two stages in this process one is frequent set item generation in which we find all the frequent sets of items and association rules generated from the frequent itemsets. The objective of this paper is to study some of the recent association rule mining algorithms proposed by several researchers. The data used in this paper is secondary data collected from various sources such as Google search, scholar, and other open access journal papers.*

Keywords- Assocition Rule Mining, Dat Mining, Assocition rules, Frequent itemsets, Association Rule Mining Techniques.

I. INTRODUCTION

Data mining or Knowledge Discovery in Database (KDD) is a technique that analyzes and summarizes data from different contexts and discovers useful knowledge. Data mining lets users to analyze the data from many diversified aspects, classifies, summarizes data to find relationships among different data. Data mining is also termed as the process of finding correlations or patterns between several attributes of data from large relational databases. Data mining has several techniques (fig. 1) to achieve this such as classification, clustering, regression, time series analysis, prediction, association rules and numeration etc[27]. Data mining algorithms are further classified into descriptive mining and predictive mining algorithms. Descriptive mining is defined as Summarization or characterization of comprehensive tracts of data in data store where as Predictive mining is a process of performing presumptions on existing data, and deriving predictions from previous data [30].

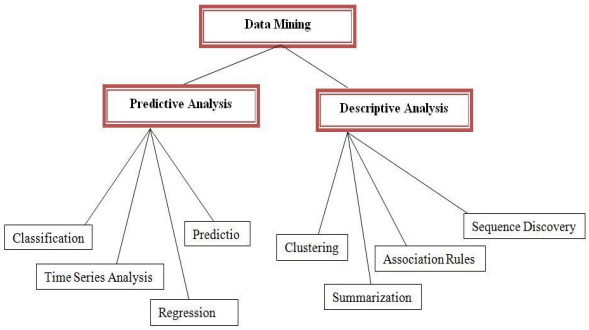

Fig. 1 Data mining Techniques.

Association Rule mining is categorized under descriptive mining analysis and became very popular technique among data mining techniques because of its vast variety of various applications domains such as insurance, stock market, super market, healthcare, tax inspection, sports and traffic management etc. It is the process of generating associations among fields of relational or transactional databases and building relationships between attributes. It is initially developed as combination of classification and association techniques[2]. Classification is the process of classifying data into different categories based on relationships among the data. It is also described as finding frequent patterns, correlations or associations between sets of objects in information repositories. Since the size of data is increasing rapidly, mining of association rules from massive data in the database is attentive among lot of organizations those help in solving several decision making problems. To generate association rules there are two steps. In first step, we discover frequent items from transactional database and the second step is generation of common association rules from those frequent items. Set of items in association rule are called itemsets and it occurs frequently more than a predefined minimum support[20]. Itemsets has two fundamental measures, one is support, and second one is confidence[9]. Support is defined as chances of occurrences of frequent itemsets in a transaction. Confidence is the probability of the rule's subsequent resultants which also contains previous consequents in the transaction. To term association rule is strong it should have minimum confidence. Conventional association rule mining algorithms are made up of binary attributes in databases[29]. Types of association rules described in [26] as positive association rule mining, negative association rule mining, constraint based association rule mining, multilevel

association rule mining. Association rules garneted gives detailed description of itemsets that are listed in dataset transactions[31]. Positive association rule mining is about finding positive relationships among itemsets and rules that are garneted from those positively related itemsets. These positive association rules classified into Boolean rule, checks the presence of itemset, quantitative rule, defines the associations among quantitative values which are partitioned into intervals, spatial rule is a rule that indicates particular association relationships between itemsets of spatial databases, and finally temporal rule, used to find significant relationships between itemsets of temporal databases. These algorithms are useful in decision making process.

Negative association rule mining is technique of finding negative association rules between the same itemsets that are not frequent, which are considered to be absent from data transactions. The negative rules are generated from infrequent itemsets. Negative rules also plays significant role in decision-making process. Constraint based association rule mining cost effective process that gives detailed account of only constraints while mining process and gives association rules based only on user interests. These constraints are knowledge based and data constraints. Multilevel association rule mining generates association rules from mining data at different abstraction levels. Some of the abstract levels are uniform minimum support for all level, reduced minimum support at each level, and item or group based minimum support level. Although association rule mining has vast variety of applications traditional algorithm is very slow. To overcome this problem many researchers contributed their work to improve traditional association rule mining algorithm. Next section (section III) describes the study of some of improved association rule algorithms proposed by several researchers.

II. BACKGROUND

Finding frequent patterns of itemsets is essential in association rule mining. Process of discovering associations and correlations between fields of huge datasets is called frequent item set mining. Initially frequent pattern mining was suggested by R.Agarwal to solve market basket analysis problem using association rule mining. Basically there are two approaches for frequent pattern algorithm, one is candidate generation approach follows breadth first approach, and second one is without candidate generation approach follows breadth first approach. Example for candidate generation approach is Apriori algorithm. Example for without candidate generation approach is FP-growth algorithm. Discovering the interesting associations and frequent patterns between itemsets of huge databases is called association rule mining.

Association rule mining described as follows: Let $X=\{I_1,I_2,I_3,\ldots\ldots,I_n\}$ constitutes as all itemsets and $T=\{T_1,T_2,T_3,\ldots\ldots,T_m\}$ are set of all data transactions, in which each transaction represents set of items, and Y be any subset of X, then association rule X=>Y illustrates a particular relationship among the itemsets X and Y. The confidence for the association rule X=>Y is determined by the percentage of transactions T that contains itemset Y in the transactions which contains itemset X[11][26][32]. Apriori algorithm analyzes candidates generation based approach and a compromise of two stages, first stage is generation of all frequent itemstets whose occurrence surpasses the minimum support. Second stage creates k+1 itemsets based k frequent itemsets which have been explored in previous stage. That is presumptive reduction process takes place to eliminate all the k+1 itemsets that are not frequent. Frequent pattern growth (FP-growth) algorithm reduces association rule problem of exploring large set of transactions and comparisons. FP-growth indexes frequently explored itemsets in a tree structure form, so that the just one data scan is require. However, there is a problem with exploring enormous candidate itemsets since either considering that the high memory requirements and number of input/output stores all itemsets.

III. LITERATURE SURVEY

J.S. Esther Sylvia Jebarani, et al. [4] proposed new rank based weighted association rule mining algorithm based on fuzzy c- means. This approach processes huge set of datasets with help of Limma statistical test measures and fundamentally measures as rank based weights. Limma process computes the p-value of each (item, and assigns weights to given items based on p-value measures. The Limma statistical test yields rank wise item lists. Limma gives a rank-wise item lists in consonance with their p-values taken from best case to worst case. After that, weights are assigned to each item according to their p-value ranks. And indexes these p-value ranks into the measures, so that the measure preference to each item by given data discretization process, that uses fuzzy C-means clustering algorithm. Maziyar Grami, et al. [5] proposed new algorithm for association rule mining based on heuristic genetic algorithm operates on prepared databases. The proposed heuristic genetic algorithm gives effective outcomes in lesser time and very useful when we want to obtain large set of resultants. Heuristic genetic algorithm solves problems using nature approach. In this method at first initial population/dataset is selected by the result of Apriori algorithm, which helps to produce more realistic results and after a micro genetic search takes place to produce results. Small form of genetic algorithm micro genetic and initial outcome of genetic algorithm will supplied as input to the micro genetic search. After that following steps takes

place in which each chromosome is described as the row of the database. In a way that population or chromosome contains binary array of items that are compromised of which consists of one (i.e. transaction takes place) and zero (i.e. transaction not takes place). Xiuli Yuan [11] proposed an improvement of apriori algorithm in which scanning of transaction database done only once and gives transaction identifier (TID) set for every item in transaction database. In the proposed technique candidate itemsets C_k are produced at first, after that elimination of L_{k-1} counts the time of every item occurred in L_{k-1}, and deletes the itemsets which are numbered less than k-1. For counting the support of the candidate itemsets C_k it uses overlap technique based on TID sets of L_{k-1} and L_1. Finally aborts the process of $| L_k| <= k$. Advantage if this algorithm is, it reduces the time and number of clients eliminated. This algorithm reduces the 98% time than traditional Apriori algorithm but consumes more space. In healthcare this algorithm can be used for analysis medical process and discovers efficient ways to enhance medical services and saves medical costs. Aashna Agarwal and Dr. Nirali Nanavati [13] suggested new association rule mining algorithm for solving multi objective optimization problem by employing hybrid Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) algorithm (GA-PSO) method. Multi objective optimization follows weighted sum approach and uses single confidence function for association rule mining. Generates number of important rules that are similar to the interests and leads to less diversity of lesser interest to the users. Further, user interference for these co-efficient specifications in the weighted sum method leads to two problems. At first, user need to familiar with corresponding significance of every parameter. The second problem is that the generation of rules with high confidence and lees support count. A good and frequently used algorithm to solve multi-objective optimization based on GA is Non-dominated Sorting Genetic Algorithm II (NSGA-II). It was most effective algorithm since it was implemented to improve the merge of properties and employs Pareto based non-domination fast sorting technique to discover the minimal solution. The main obstacle of NSGA-II algorithm is that finer rules that were generated during intermediate generations will be lost. There are no other new rules included in the resultant set because of elitism. To overcome this problem Aashna Agarwal and Dr. Nirali Nanavati [13] a hybrid version of GA and PSO algorithms, which follows assumptive population based approach. The algorithm creates new reliable resultants, for enhancing the search space. In GA, the chromosomes procreate with each other to produce successor chromosomes. Positions of particles affected by their own knowledge and sharing of data between large members, is the principal approach of PSO. GA operators have a low convergence when compared to PSO. GA integrated with crossover, mutation and feasibility for solving real-world problems which is not possible in PSO technique. The main intention here is to combine both GA and PSO techniques to solve multi optimization problem. With help of this hybrid algorithm, exploration of all high coverage rules is not required. So they can only explore potentially very useful knowledge. Users will be aware of rules that are less important since here mining takes place only based on support and confidence. This algorithm removes rules that have low support high confidence and infrequent itemsets.

The problem with the traditional apriori algorithm is it operation is a breadth-first search and bottom-up technique. In which the process of executing begins from the insignificant frequent itemsets and forwards till it make it to the biggest frequent itemset. Which will predominantly increases the time space and becomes slower. To overcome this problem Ashish Shah [14] proposed an improved approach that discovers frequent patterns by using bottom up and top down approaches. This algorithm checks frequentness of large data itemsets in a maintained list of all itemsets by counting support of itemsets. With this will get knowledge about the subset of frequent itemsets so that they can be removed from the maintained list. Which will increases the performance of the system. T.Bharathi, and Dr. A.Nithya in [17] proposed new enhanced algorithm based on ontology which is four step process. In first step, data is reduced from the database and is transformed to binary form which can help us to discover essential itemset in an easiest way. In the second step IR value is calculated and sets threshold for minimum values. Ontology tree is created on the based on analysis of domain ontology. In the third step, support value will be calculated as well as frequent item discovered from ontology tree during this stage. At fourth step, ontology tree is reduced on the basis of occurrences of frequent itemsets. This algorithm improves the scanning of database and quickly calculates the support values.

IV. CONCLUSION

Association rule mining is one of data mining or KDD processes, which finds the frequent patterns based on association rules. Mining with normal association rules is time and space consuming process. So several researchers proposed their own techniques for improvement of algorithms to increase the efficiency, reliability, and complexity and performnce. This paper provides study of those improved algorithms enhanced.

REFERENCES

[1] Sudhakar Singh, Rakhi Garg, P. K. Mishra, “A

Comparative Study of Association Rule Mining Algorithms on Grid and Cloud Platform “, IJETCAS, February 2014.

[2] Kavita Mittal, Dr.Prerna Mahajan, Dr.Gaurav Aggarwal, “A COMPARATIVE STUDY OF ASSOCIATION RULE MINING TECHNIQUES AND PREDICTIVE MINING APPROACHES FOR ASSOCIATION CLASSIFICATION”, IJARCS, Volume 8, No. 9, November-December 2017.

[3] Moushumi Sharma, Ajit Das, Nibedita Roy, “A Complete Survey on Association Rule Mining and Its Improvement”, IJIRCCE, Vol. 4, Issue 5, May 2016.

[4] J.S. Esther Sylvia Jebarani, Mr. S. Saravana Kumar, “A New Approach to Rank Based Weighted Association Rule Mining Based on Fuzzy C-Means Algorithm”, IJIREEICE, Vol. 4, Issue 6, June 2016.

[5] Maziyar Grami, Reza Gheibi, Fakhereh Rahimi, “A Novel Association Rule Mining Using Genetic Algorithm” , IKT, Hamedan, Iran, 2016.

[6] Abdelhamid Boudane, Said Jabbour, Lakhdar Sais, Yakoub Salhi, “A SAT-Based Approach for Mining Association Rules”, IJCAI, 2016.

[7] Mr. Sudhir M. Gorade, Prof. Ankit Deo, Prof. Preetesh Purohit, “A Study of Some Data Mining Classification Techniques”, IRJET, Volume 04, Issue 04, April 2017.

[8] Pandya Jalpa P, Morena Rustom D, “A Survey on Association Rule Mining Algorithms Used in Different Application Areas”, IJARCS, Volume 8, No. 5, May-June 2017.

[9] Surbhi K. Solanki, Jalpa T. Patel, “Survey on Association Rule Mining”, Fifth International Conference on Advanced Computing & Communication Technologies, 2015.

[10] T. Karthikeyan, N. Ravikumar, “A Survey on Association Rule Mining”, IJARCCE, Vol. 3, Issue 1, January 2014.

[11] Xiuli Yuan, “An improved Apriori algorithm for mining association rules”, AIP Conference Proceedings 1820, August 2017.

[12] Ling Yuan, Wasan Itwee, and Shikang Wei, “Association Rule Mining Technique with Optimized FP-Tree Algorithm”, IJASCSE, VOLUME 6, ISSUE 5, 2017.

[13] Aashna Agarwal, Prof(Dr.) Nirali Nanavati, “Association rule mining using hybrid GA-PSO for multi-objective optimization”, ICCICR, 2016.

[14] Ashish Shah, “Association Rule Mining with Modified Apriori Algorithm using Top down Approach”, 2nd, iCATccT.

[15] Omer M. Soysal, “Association rule mining with mostly associated sequential patterns”, Expert Systems with Applications, 2015.

[16] Dr. S.Vijayarani, Ms. S.Sharmila, “COMPARATIVE ANALYSIS OF ASSOCIATION RULE MINING ALGORITHMS”,

[17] T.Bharathi, Dr. A.Nithya, “Enhanced Way of Association Rule Mining With Ontology”, ijecs, Volume 5, Issue 10, October 2016.

[18] Sushil Kumar Verma, R.S. Thakur, “Fuzzy Association Rule Mining based Model to Predict Students’ Performance”, IJECE, Vol. 7, No. 4, August 2017.

[19] Wei Han, Julio Borges, Peter Neumayer, Yong Ding, Till Riedel and Michael Beigl, “Interestingness Classification of Association Rules for Master Data”, ICDM 2017.

[20] Zailani Abdullaha, Tutut Herawanb, Noraziah Ahmadb, Mustafa Mat Derisc, “Mining significant association rules from educational data using critical relative support approach”, WCETR 2011.

[21] Hemant Kumar Soni, “Multi-objective Association Rule Mining using Evolutionary Algorithm”, ijarcsse, Volume 7, Issue 5, May 2017.

[22] Mayur Bhosale, Tushar Ghorpade, Rajashree Shedge, On Demand Recommendation using Association Rule Mining Approach”, SCOPES 2016.

[23] Kiburm Song , Kichun Lee, “Predictability-based collective class association rule mining”, Expert Systems With Applications, February 2017.

[24] Michael Hahsler, Probabilistic Approach to Association Rule Mining”,

[25] G. Silambarasan, Dr. T. Anand, Dr. V. Chandrasekar, “Rank-Based Weighted Rule Mining Using Post Mining Methods with Ontology Support”, IJAER, Volume 12, Number 2017.

[26] M. Shridhar, M. Parmar, Survey on Association Rule Mining and Its Approaches”, IJCSE, Volume-5, Issue 3, Mar 2017.

[27] Deepashri.K.S, Ashwini Kamath, “Survey on Techniques of Data Mining and its Applications”, IJERMT, Volume 6, Issue 2, February 2017.

[28] Shanthi S, “Survey on Web Usage Mining using Association Rule Mining”, ijicse, Volume 4 Issue 3, May-June 2017.

[29] Gosain A, and Bhugra M, A comprehensive survey of association rules on quantitative data in data mining”, IEEE CICT,JeJu Island, April 2013.

[30] Qiankun Zhao and Sourav S. Bhowmick, “Association rule mining: A Survey”, Technical Report, CAIS, Nanyang Technological University, Singapore, 2003.

[31] K.P. Kumar and S. Arumugaperumal, “Association Rule Mining and Medical Application; A Detailed Survey”, International Journal of Computer Application(0975-8887), Volume 80, number 17, October 2013.

[32] Lazcorreta E, Botella F, Fernández-Caballero A, “Towards personalized recommendation by two-step modified Apriori data mining algorithm”, Expert Systems with Applications, 2008.