A Survey of Genetic Algorithm for Association Rule Mining

Gagandeep Kaur
M.Tech Research Scholar, Department of Computer Science and Engineering,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India.

Shruti Aggarwal
Assistant Professor, Department of Computer Science and Engineering,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India.

ABSTRACT
In recent years, Data Mining is an important aspect for generating association rules among the large number of itemsets. Association Rule Mining is the method for discovering interesting relations between variables in large databases. It is considered as one of the important tasks of data mining intended towards decision making. Genetic algorithm (GA) based on evolution principles has found its strong base in mining Association Rules. Genetic algorithm is a search heuristic which is used to generate useful solutions to optimization and search problems. Genetic algorithm has proved to generate more accurate results when compared to other formal methods available. The fitness function used in Genetic Algorithm evaluates the quality of each rule. Many researchers have proposed genetic algorithm for mining interesting rules from dataset. This paper presents the survey of Genetic Algorithm for Association Rule Mining.

General Terms
Data Mining, Apriori Algorithm

Keywords
Association Rule Mining, Genetic Algorithm

1. INTRODUCTION
Data Mining is a process of extraction of useful information from huge amount of data. The development of information technology in various fields of human life lead to generate large amount of data. The data can be stored in various formats like records, documents, images etc. The data collected from different applications require proper mechanism of extracting knowledge from large repositories for better decision making. Data Mining aims at the discovery of useful information from large collections of data [1]. It is also known as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology. Data Mining tools search databases for hidden patterns, find information that users may miss. Usually, Data Mining and knowledge discovery from data are taken as synonyms, but in actual data mining is part of the knowledge discovery process. The knowledge discovery process [2, 3] comprises of following steps:

- **Data Cleaning:** In this phase, noise and irrelevant data is removed.
- **Data Integration:** In this phase, data collected from multiple data sources may be combined.
- **Data Selection**: In this phase, the data relevant to the user is retrieved from the database.
- **Data Transformation:** In this phase, selected data is transformed into appropriate forms.
- **Data Mining**: It is the essential step in which knowledgeable techniques are applied to extract data patterns.

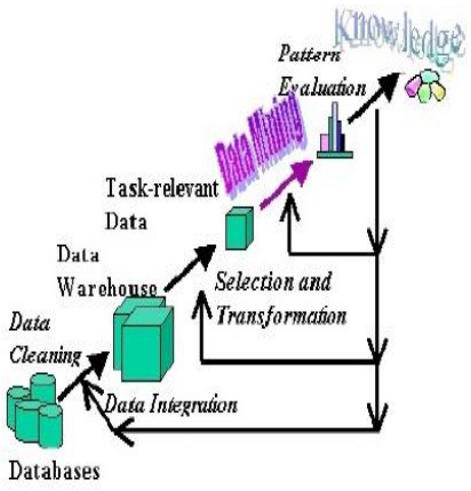

Fig 1: KDD Process [2]

- **Pattern Evaluation**: In this step, interesting patterns representing knowledge are identified based on given measures.
- **Knowledge Representation**: It is the last phase in which knowledge representation techniques are used to represent mined knowledge to the user.

2. ASSOCIATION RULE MINING
In Data Mining, Association Rule Mining is a popular and well researched method for discovering interesting relations between variables in large databases. Association Rule Mining technique was first introduced by Agrawal et al. in 1993, who developed Apriori algorithm for solving the ARM based problems. Association Rules are used in various areas such as telecommunication networks, market and risk management, inventory control etc. An association rule is an implication in the form of $X \Rightarrow Y$, where X and Y are collection of items and $X \cap Y = \emptyset$. Here X is called antecedent (left-hand-side or LHS) and Y is called consequent (right-hand-side or RHS) [4]. A rule may contain more than one item in antecedent and consequent part.

2.1 Measures of Association Rules
To select interesting rules from the set of all possible rules, there are two important basic measures: Support and Confidence. Usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimum

support and minimum confidence respectively. Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong [4].

2.1.1 Support

It is the probability of item or itemsets in the given transactional database:
Support(X)=n(X)/n where n is the total number of transactions in the database and n(X) is the number of transactions that contains the itemset X. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain that item [5].

2.1.2 Confidence

The confidence of a rule (X⇒Y)) is the percentage of transactions in database that contain X also contain Y. It is computed as follow:
Confidence(X⇒Y)= support(X and Y)/support(X)
Suppose the confidence of the association rule X⇒Y is 70%, it means that 70% of the transactions that contain X also contain Y [5].

2.2 Apriori Algorithm

A large number of association rule mining algorithms have been developed with different mining efficiencies. Apriori is a classic algorithm for mining all frequent itemsets and association rules learning. The frequent itemsets generated by Apriori can be used to determine association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers). Each transaction contains set of items called itemset. Apriori uses level-wise search where k-itemsets (an itemset that contains k-items) are used to explore (k+1)-itemsets. In the beginning, the set of frequent 1-itemsets is found. This set contains items that satisfy minimum support and is denoted by L1. In each subsequent pass, we begin with a set of itemsets found to be frequent in the previous pass. This set is used for generating new itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually frequent and they are used in the next pass. Therefore, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. An important property called Apriori property is used to reduce the search space which is described as: "All nonempty subsets of a frequent itemset must also be frequent" [6, 7].

3. GENETIC ALGORITHM

Genetic algorithm(GA) was introduced by John Holland in the 1970. GA is stochastic search algorithm based on the principles of natural selection and natural genetics, which has been successfully applied in many machine learning and optimization problems, to generate useful solutions. Genetic Algorithm works in an iteration manner by generating new populations of strings from old ones. Genetic algorithms are inspired by Darwin's theory of evolution [8]. The algorithm starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, the more fit individuals are selected from the current population, and genetic operators are applied to form a new population. The new population is then used in the next iteration of the algorithm. The algorithm terminates when a maximum number of generations has been produced or a satisfactory fitness level has been reached for the population. A standard genetic algorithm utilizes three genetic operators: reproduction (selection), crossover and mutation. The functions of genetic operators are as follows:

- **Selection**: Selection deals with the probabilistic survival of the fittest, in which fit chromosomes are chosen to survive. There are many methods to select best chromosomes like roulette wheel selection, rank selection, tournament selection etc [9].
- **Crossover**: It takes individual chromosomes from Parents, combines them to form new ones. Crossover can be single point crossover, multipoint crossover etc.
- **Mutation**: Mutation alters the new solutions in order to get the better solutions. It flips each bit in the individual with a pre-specified probability of mutation (0 becomes 1, 1 becomes 0).

4. RELATED WORKS

Numerous works have been carried out using Genetic Algorithm for mining Association Rules. This section describes the work done in the field of Association Rule Mining using Genetic Algorithm.

Anandhavalli M. et al. [8] used Apriori algorithm to generate frequent itemsets and then applied genetic algorithm in order to generate the rules which contain positive attributes, negation of attributes having single attribute or more than one attributes in the consequent part. The fitness function in this paper consists of confidence factor and completeness which are combination of true positives, true negatives, false positives and false negatives. Because it focused on negative attributes, most of the generated rules have lower support values. They use the Lens database from UCI to show the effectiveness of the proposed algorithm. The results reported in this paper are very promising since the discovered rules are of optimized rules. To minimize the complexity of the Genetic Algorithm and scanning of database, Baye's theorem can be applied on the generated rules.

Rupali Haldulakar et al. [10] explained that strong rule generation is an important area of Data Mining. In this paper, authors designed a novel method for generation of strong rule. Apriori algorithm is used to generate the rules. After that they use the optimization techniques. One of the best ways to optimize the rules is Genetic Algorithm. For the optimization of the rule set, they design a new fitness function that uses the concept of supervised learning. In which fitness function is divided into two classes c1 and c2, one class for discrete rule and another class for continuous rule. They use Abalone dataset obtained from UCI machine learning repository for their experimentation. The data set has 4177 samples. By using the new fitness function that uses the concept of supervised learning, authors got better result. To make genetic algorithm more effective and efficient, it can be incorporated with other techniques, so it can provide a best result.

M. Ramesh Kumar et al. [11] proposed a novel genetic algorithm based association rule mining algorithm . Fitness function is designed based on the two measures like all confidence and the collective strength of the rules other than the classical support and the confidence of the rules generated. The algorithm has been tested on the four data sets which are Adult, Chess, Wine, Zoo. The sample data sets have been taken from the UCI data repository for the testing of the algorithm. The environmental measure they had for testing is the population size for performing GA is 200, the selection rate is 10% and the crossover rate is 6% and the mutation rate is fixed as 1%. The fitness function is designed in such a way

that to prioritize the rules based on the user preference. Their approach significantly reduces the number of rules generated in four data sets they have used. The technique can be extended by incorporating other interesting measures in future work.

J.Malar Vizhi et al. [9] proposed a genetic algorithm to generate high quality association rules. Association rule mining problems can be considered as a multi objective problem rather than as a single objective one. They used multi-objective evolutionary framework for association rule mining. The fitness function used by them consists of four metrics. They are confidence, completeness, interestingness and comprehensibility. These metrics are combined as an objective fitness function. The fitness function is calculated as the arithmetic weighted average confidence, completeness, interestingness and comprehensibility. The dataset they used for their experimentation is real world Primary-tumor dataset. The Primary-tumor database contains 18 attributes and 339 instances. They used tournament selection and multipoint crossover methods which is flexible for large number of attributes in the database. They tested the approach only with the numerical and categorical valued attributes. The algorithm needs some modification in order to cope with continuous data.

Manish Saggar et al. [12] used the Genetic Algorithms to optimize the rules generated by Association Rule Mining(Apriori method). In general, the rule generated by Association Rule Mining technique do not consider the negative occurrences of attributes in them, but by applying Genetic Algorithms (GAs) over these rules, the system can predict the rules which contains negative attributes. The improvements applied by authors in GAs help the rule based systems used for classification. The authors first implemented Association Rule mining (using a-priori technique) and then applied the GAs to generate the rules which contains the negations in attributes and are of richer quality. The database they used is produced synthetically. The results generated when the authors applied their technique on the synthetic database, includes rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining. The approach needs some modification in order to minimize the complexity of Genetic Algorithms by using distributed computing.

Ashish Ghosh et al. [13] used a Pareto based Genetic Algorithm to solve the multi-objective rule mining problems. Association rule mining can be considered as a multi-objective problem rather than as a single-objective one. Authors used three measures to extract some useful and interesting rules from database. These three measures are comprehensibility, interestingness and the predictive accuracy. Using these three measures, some previously unknown, easily understandable rules can be generated. They used a variant of the Michigan approach to represent the rules as chromosomes, where each chromosome represent a separate rule. To improve the efficiency of the algorithm they used, some modifications are required. In this paper, authors used the random sampling method but by using other sampling methods, a good sample can be found. A perfect sample will improve the accuracy of the rules generated by the algorithm.

The Table below describes the merits and demerits of the different techniques on Association Rule Mining using Genetic Algorithm:

Table 1: Comparative study of Genetic Algorithm

Author(s) Name	Proposed Method/Merits	Demerits/ Future Work
Anandhavalli M., Suraj Kumar Sudhanshu	In this paper, Authors applied Genetic Algorithm on the frequent itemsets generated by Apriori algorithm to generate the rules containing positive attributes, negation of attributes having single attribute or more than one attributes in the consequent part. The results obtained are very promising since the discovered rules are optimized rules.	To minimize the complexity of the Genetic Algorithm and scanning of database, Baye's theorem can be applied on the generated rules.
Rupali Haldulakar, Prof. Jitendra Agrawal	A new fitness function is designed that uses the concept of supervised learning. The fitness function is divided into two classes c1 and c2, one class for discrete rule and another class for continuous rule. By doing this, authors got better results.	To make genetic algorithm more effective, it can be incorporated with other techniques, so that it can provide a best result.
M. Ramesh Kumar	The proposed Genetic Algorithm uses the fitness function having two measures like all confidence and the collective strength of the rules. This approach when applied on the different datasets significantly reduces the number of rules.	The technique can be extended by incorporating other interesting measures in future work.
J.Malar Vizhi, Dr. T.Bhuvaneswari	In this paper, Authors proposed a Genetic Algorithm, the fitness function of which consists of four metrics. They are confidence, completeness, interestingness and comprehensibility. These metrics are combined as an objective fitness	The algorithm needs some modification in order to cope with continuous data.

	function to generate high quality Association Rules.	
Manish Saggar, Ashish Kumar Agarwal	The proposed Genetic Algorithm when applied on the synthetic database, produced the desired rules, i.e. rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining.	The approach needs some modifications in order to minimize the complexity of Genetic Algorithms by using distributed computing.
Ashish Ghosh, Bhabesh Nath	Authors used a Pareto based Genetic Algorithm to solve the multi-objective rule mining problems using three measures. These measures are comprehensibility, interestingness and the predictive accuracy. They used these three measures to extract some useful and interesting rules from the database.	To improve the efficiency of the algorithm, some modifications are required. Authors used the random sampling method but by using other sampling methods, a good sample can be found. A perfect sample will improve the accuracy of the rules generated by the algorithm.

5. CONCLUSION

Genetic Algorithms are used in the discovery of high quality rules because they perform global search and its complexity is less compared to other algorithms. In recent years, lots of work has been carried out using genetic algorithm for mining Association Rules. This paper surveyed the existing works on Genetic Algorithm in mining Association Rules. Less work is done in the field of association rule mining using multi-objective genetic algorithm. More interesting measures can be used in order to discover more interesting rules.

6. REFERENCES

[1] Venkatadri.M and Dr. Lokanatha C. Reddy, "A Review on Data mining from Past to the Future", International Journal of Computer Applications, Volume 15– No.7, pp. 19-22, February 2011.

[2] Gurjit Kaur and Lolita Singh, "Data Mining: An Overview", International Journal of Computer Science and Technology, Volume 2, Issue 2, pp. 336-339, June 2011.

[3] Vibha Maduskar and Prof. Yashovardhan Kelkar, "Survey on Data Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 2, pp. 275-279, February 2012.

[4] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", International Transactions on Computer Science and Engineering, Volume 32 (1), pp. 71-82, 2006.

[5] Shanta Rangaswamy and Shobha G, ''Optimized Association Rule Mining using Genetic Algorithm'', Journal of Computer Science Engineering and Information Technology Research, Volume 2, Issue 1, pp 1-9, Sep 2012.

[6] Sanjeev Rao and Priyanka Gupta, "Implementing Improved Algorithm over Apriori Data Mining Association Rule Algorithm", International Journal of Computer Science and Technology, Volume 3, Issue 1, pp. 489-493, Jan. - March 2012.

[7] Han J. and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2nd Edition.

[8] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., "Optimized Association Rule Mining using Genetic Algorithm", Advances in Information Mining, Volume 1, Issue 2, pp. 1-4, 2009.

[9] J.Malar Vizhi and Dr. T.Bhuvaneswari, "Data Quality Measurement With Threshold Using Genetic Algorithm", International Journal of Engineering Research and Applications, Volume 2, Issue 4, pp. 1197-120, July-August 2012.

[10] Rupali Haldulakar and Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm", International Journal on Computer Science and Engineering, Volume 3, No. 3, pp. 1252-1259, Mar 2011.

[11] M. Ramesh Kumar and Dr. K. Iyakutti, "Application of Genetic algorithms for the prioritization of Association Rules", International Journal of Computer Applications, pp. 35-38, 2011.

[12] Manish Saggar, Ashish Kumar Agarwal and Abhimanyu Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms", IEEE International Conference on Systems, Man and Cybernetics, pp. 3725-3729, 2004.

[13] Ashish Ghosh, Bhabesh Nath, "Multi-objective Rule Mining using Genetic Algorithms", Information Sciences, pp. 123–133, 2004.