A Review on Association Rule Mining with Genetic Algorithm
Mrinalini Rana* and Palvinder Singh Mann **

*Research Scholar, Department of Computer Science and Engineering, D.A.V.I.E.T.
**Asstt. Professor, Department of Computer Science and Engineering, D.A.V.I.E.T.
Jalandhar, Punjab, India.

(Received 27 April 2013 Accepted 30 April 2013)

Abstract: Data mining includes number of techniques like clustering, classification, sequential patterns, association rules and etc. Association rule mining is a technique for mining interesting rules from large databases for further analysis. Association rule mining includes two-step approach for extracting the rules. These two steps require many database scan and use support and confidence as the threshold value to evaluate the rules.
Genetic algorithm can be used to overcome the disadvantages. This paper discusses a simple, effective and time saving algorithm for association rule generation.

Keywords: Rule mining, Apriori algorithm, Support, Confidence, Genetic Algorithm, Fitness function.

1. Introduction:

Data mining is an important step in knowledge discovery in databases (KDD). Data mining includes number of techniques for mining the rules and association rule mining is one of the techniques for data mining.
Association rule is an implication of expression X → Y where both X and Y are the items from the database called as itemsets. Association rule mining includes a two step approach and requires support and confidence as a threshold value for evaluating the rules.
Support is fractions of transaction that contain both X and Y. Confidence is also called as predictive accuracy, estimates that how much the rule is relevant.
First step, Generate frequent item-set: generate all item-sets whose "support>=minsup". Second step, Rule generation: Generate high confidence rules from each frequent itemset. For calculating frequent item-set then the candidate item-sets requires a lot of database scan.

Apriori is a classical association rule mining algorithm. It generates frequent itemsets based on threshold value minimum support and then generates rules from minimum confidence. Both of the threshold value is user specified even before the starting of the algorithm. So, why the algorithm is named as a-priori, threshold values are defined by the user before the starting of the algorithm.
Other interestingness measures are [7]:

- \( \text{Lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)} = \frac{\text{support}(A \rightarrow B)}{\text{support}(A) \cdot \text{support}(B)} \)
- \( \text{Conviction}(A \rightarrow B) = \frac{\text{support}(A) \cdot \text{support}(B \text{ not } A \rightarrow B)}{\text{support}(A \rightarrow B)} \)
- \( \text{Leverage}(A \rightarrow B) = \frac{\text{support}(A \rightarrow B) \cdot \text{support}(B)}{\text{support}(A) \cdot \text{support}(B)} \)

To overcome the disadvantages Genetic algorithm can be used for association rule mining. Genetic algorithm uses fitness function as threshold value to evaluate the rules [1] [2] [9] [10].

Working principle of Genetic Algorithm [18]:
Step I [Start] Generate random population of chromosomes, that is, suitable solutions for the problem.
Step II [Fitness] Evaluate the fitness of each chromosome in the population.
Step III [New population] Create a new population by repeating following steps until the new population is complete.
  a) [Selection] Select two parent chromosomes from a population according to their fitness. Better the fitness, the bigger chance to be selected to be the parent.
  b) [Crossover] With a crossover probability, cross over the parents to form new offspring, that is, children. If no crossover was performed, offspring is the exact copy of parents.
  c) [Mutation] With a mutation probability, mutate new offspring at each locus.
  d) [Accepting] Place new offspring in the new population.
Step IV [Replace] Use new generated population for a further run of the algorithm.
Step V [Test] If the end condition is satisfied, stop, and return the best solution in current population.
Step VI [Loop] Go to step 2.

2. Association Rule Mining Algorithms:

To mine the association rules numbers of algorithm are being introduced. The basic algorithm for association rule mining is Apriori Algorithm. Apriori algorithm is used to operate on databases containing transactions. Many improved Apriori algorithm was introduced. Apriori was implemented with different data structures like hash, trees, link list and etc.

Following are the various algorithms for association rule mining:

2.1 Apriori Algorithm [6]:

The Apriori algorithm is a classical association rule mining. It was first introduced by Agrawal and Srikan in 1994. The algorithm starts with a dataset containing transactions and generates frequent item sets, having at least a user specified threshold. In the algorithm of Apriori, an item set X of length k is frequent if and only if every subset of X, having length k - 1, are also frequent. Apriori supports downward closure.

2.2 FP tree:

FP tree is also one of the other popular association rule mining algorithm. FP tree uses tree as a data structures and tries to overcome the disadvantages of Apriori algorithm.

2.3 Partition Algorithm for Mining Frequent Itemset (PAFI)[25]:

The PAFI is the algorithm proposed by D. Kerana Hanirex, Dr. A. Kumaravel as an improved version of Apriori to generate more efficient rules. The algorithm PAFI divides the database into different partitions with some transactions in each partition. This approach reduced the multiple scan of databases.

2.4 Two Dimensional Approach Algorithm for Mining Frequent Itemsets (TDFI) [25]:

As above algorithm this algorithm is also introduced by D. Kerana Hanirex, Dr. A. Kumaravel and the algorithm specifies that the problem of mining frequent itemset is to find all itemsets that are greater than the user specified minimum support count (i.e.) minsup. The support count for each partition is calculated and then the algorithm finds the frequent itemsets. This reduced number of scans and the complexity.

2.5 Genetic algorithm:

Genetic algorithm initially developed by John Holland, In university of Michigan in approximately 1970's, to understand the processes of natural systems and to design artificial systems containing natural system properties and robustness. It uses chromosomes to represent the itemsets as individuals and fitness function to evaluate the rules. Genetic operators are Selection
operator, Crossover operator, Mutation operator.

2.6 Multi-Objective Genetic Algorithm:

Multi-Objective Genetic algorithm uses Pareto Optimality theory and it is used in data mining within databases that include numeric attributes.

3. Literature survey of Association Rule Mining Algorithms:

The literature survey indicates that efforts are made to develop many association rule mining algorithms. For market-basket analysis it is essential to provide more efficient rules. Moreover, it is very important to have lesser meaningful rules rather than having more mixed (relevant and irrelevant) rules.

Ashish Ghosh, Bhabesh Nath(2004) discussed: Multi-objective rule mining using genetic algorithms[1]. In this paper, they used measures like support count, comprehensibility and interestingness for evaluating interesting rules as their different objectives for mining association rule problem. Using these measures they were implementing genetic algorithm to generate some relevant rules from any market basket analysis type database. Based on their experimentation the proposed algorithm has been found for suitable for large databases.

Manish Saggar, Ashish Kumar Agrawal, Abhimanyu Lad (2004) discussed: Optimization of Association Rule Mining using Improved Genetic Algorithms [2]. In this paper, their main objective is to use genetic algorithm in the discovery of high level prediction rules that perform a global search and perform better with attributes than greedy rule induction algorithms (used in data mining). This improvement in genetic algorithm may helps in rule based system for classification.

Bilal Alata, S. Erhan Akin (2005) discussed: An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules [3]. In this paper, they proposed a genetic algorithm as a search strategy for not only positive association rule but also for negative association rules. The proposed algorithm performs a database independent approach test which directly mined without generating frequent itemsets. Proposed algorithm also not relies on minimum support and minimum confidence.

Virendra Kumar Shrivastava, Dr. Parveen Kumar, Dr. K. R. Pardasani(2010) discussed: Extraction of Interesting Association Rules using GA Optimization[4]. In this paper, they used genetic algorithm for extracting association rules. They used measures like support, confidence, interestingness, and completeness.

Ashish Ghosh, Bhabesh Nath(2004) discussed: Multi-objective rule mining using genetic algorithms[1]. In this paper, they used measures like support count, comprehensibility and interestingness for evaluating interesting rules as their different objectives for mining association rule problem. Using these measures they were implementing genetic algorithm to generate some relevant rules from any market basket analysis type database. Based on their experimentation the proposed algorithm has been found for suitable for large databases.

Manish Saggar, Ashish Kumar Agrawal, Abhimanyu Lad (2004) discussed: Optimization of Association Rule Mining using Improved Genetic Algorithms [2]. In this paper, their main objective is to use genetic algorithm in the discovery of high level prediction rules that perform a global search and perform better with attributes than greedy rule induction algorithms (used in data mining). This improvement in genetic algorithm may helps in rule based system for classification.

Bilal Alata, S. Erhan Akin (2005) discussed: An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules [3]. In this paper, they proposed a genetic algorithm as a search strategy for not only positive association rule but also for negative association rules. The proposed algorithm performs a database independent approach test which directly mined without generating frequent itemsets. Proposed algorithm also not relies on minimum support and minimum confidence.

Virendra Kumar Shrivastava, Dr. Parveen Kumar, Dr. K. R. Pardasani(2010) discussed: Extraction of Interesting Association Rules using GA Optimization[4]. In this paper, they used genetic algorithm for extracting association rules. They used measures like support, confidence, interestingness, and completeness.

Amy H.L. Lim a, Chien-Sing Lee a, Murali Raman (2012) discussed: Hybrid genetic algorithm and association rules for mining workflow best practices [5]. In this paper, they used correlation measure instead of traditional support and confidence in genetic algorithm to driven data dynamically. They used correlation as fitness function to support upward closure in association rule. Downward closure is opposite to upward closure.

Jesmin Nahar et al (2013) discussed: Association rule mining to detect factors which contribute to heart disease in males and females [6]. In this paper, they investigate the sick and healthy factors related to heart diseases in males and females. They used UCI Cleveland dataset, a biological database. They were comparing three rule generation algorithms Apriori, Predictive Apriori, Tertius. They conclude that Apriori algorithm is fastest so they again generate rules for sick and healthy data with Apriori algorithm.

Dong Gyu Lee et al (2013) discussed: Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction[7]. In this paper, they proposed association rule mining algorithm that can generate association rules related to hypertension and diabetes from AMI patients having age
45 years old or lesser. They used risk factors like glucose, cigarette smoking, creatinine, triglyceride which are related to hypertension and diabetes. To increase the interestingness they were using measures lift, convergence, and leverage measures.

B. Minaei-Bidgoli, R. Barmaki, M. Nasiri(2013) discussed: Mining numerical association rules via multi-objective genetic algorithms[8]. In this paper, they proposed a multi-objective genetic algorithm, method for association rule mining, for numerical data. They used measures like confidence, interestingness, and comprehensibility as multiple objectives for genetic algorithm method.

Basheer Mohamad, Al-Maqaleh(2013) discussed: Discovering Interesting Association Rules: A Multi-objective Genetic Algorithm Approach[9]. In this paper, they proposed a multi-objective genetic algorithm for generating association rules. They used measures like support, confidence and simplicity/comprehensibility. The proposed algorithm generates interesting association rules without specifying, minimum support, minimum confidence, user specified threshold.

Bettahally, N. Keshavamurthy, Asad M. Khan, Durga Toshniwal(2013) discussed: Privacy preserving association rule mining over distributed databases using genetic algorithm[10]. In this paper, they compare traditional frequent pattern mining algorithm i.e. Apriori algorithm with proposed genetic algorithm in local search. In Apriori algorithm population is formed in only single iteration but in genetic algorithm population is formed in every new generation. Proposed algorithm also addresses various types of partitions such as horizontal, vertical, and arbitrary.

Various different techniques or methods are being used to generate interesting association rules. But genetic algorithm can be a better algorithm for generating rules. Different interestingness measures can be used as to increase strength of fitness function to generate more relevant/interesting association rules.

To overcome the classical (Apriori algorithm) association rule mining algorithm disadvantages Genetic algorithm is used to find frequent items and to develop a global association rules. Or the genetic algorithm can also optimized association rule generated from Apriori algorithm. Firstly apply Apriori to generate frequent itemsets and than secondly to optimize the rules genetic algorithm can be applied.

4. Conclusion:

In this paper, an effort is made to propose a simple, effective and time saving algorithm for association rule generation in which multi fitness functions can be used as the threshold value for rule evaluation. Some more interestingness measures like lift, leverage, convergence can be used to enhance the fitness function except support and confidence.

5. Future Work:

In future work we can work on, Association Rule Mining based classifiers increases the effectiveness of the rules. Sampling techniques like regression-based sampling or cluster-based sampling will improve the correctness of the rules generated by the algorithm. Genetic algorithm complexity can be reduced by distributed computing. Combination of support, confidence, lift, leverage and conviction can be used to evaluate the interestingness. Genetic algorithm association rule mining can be improved by improving the chromosome representation and fitness function parameters.

Acknowledgement:

The authors are greatly indebted to the Department of Computer Science and Engineering, D.A.V Institute of Engineering and Technology, Jalandhar for providing excellent lab facilities that make this work possible.

References:


