Performance Based Association Rule-Mining Technique Using Genetic Algorithm

Amit Kumar Barai
Dredging Corporation in India Limited, Government of India Undertaking (A Miniratna Category –I PSU), Dredge House, Port Area, Visakhapatnam, india
E-mail: amit.barai0@gmail.com

***Abstract*—Data Mining is the process of obtaining high level knowledge by automatically discovering information from data in the form of rules and patterns. The data mining seeks to discover knowledge that is comprehensible and accurate. And Association rule mining (ARM) is a well established technique of data mining that identifies significant correlations between items in transactional data. To measure like support count, comprehensibility, used for evaluating a rule can be thought of as different objectives of association rule mining difficulty. In the paper work we improved the association rule-mining problem using genetic algorithm (GA). In this research work, we used the random sampling method. The perfect sample improved the correctness of the rules generated by the algorithm. Our proposed technique shows the effectiveness than existing works.**

***Index Terms*—data mining techniques, association rule mining, genetic algorithm.**

I. Introduction

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Data mining can be achieved by Association, Classification, Clustering, Predictions, Sequential Patterns, and Similar Time Sequences.

In Association, the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns. In Classification, the methods are intended for learning different functions that map each item of the selected data into one of a predefined set of classes. Given the set of predefined classes, a number of attributes, and a "learning (or training) set," the classification methods can automatically predict the class of other unclassified data of the learning set.

Cluster analysis takes ungrouped data and uses automatic techniques to put this data into groups. Clustering is unsupervised, and does not require a learning set. It shares a common methodological ground with Classification.

Prediction analysis is related to regression techniques. The key idea of prediction analysis is to discover the relationship between the dependent and independent variables, the relationship between the independent variables.

Sequential Pattern analysis seeks to find similar patterns in data transaction over a business period.

Existing algorithms for mining association rules are mainly worked on a binary database, and termed as market basket database. For preparing the market basket database, each and every record of the original database is represented as a binary record where the fields are defined by a unique value of each attribute in the original database. Fields of this binary database are often termed as an item. A database having a huge number of attributes and each attribute containing a lot of distinct values, the number of items will be huge. For storing of this binary database, to be used by the rule mining algorithms, is one of the limitations of the existing algorithms [1], [2].

Another aspect of these algorithms is that they work in two phases. In the first phase is for frequent item-set generation. The frequent item-sets are detected from all-possible item-sets by using a measure called support count (SUP) and a user-defined parameter called minimum support. Support count of an item set is defined by the number of records in the database that contains all the items of that set. If the value of minimum support is too much high, and number of frequent item sets generated will be less, and thereby resulting in generation of few rules. Again, if the value is too small, then almost all possible item sets will become frequent and thus a huge number of rules may be generated. Selecting better rules from them may be another problem. After detecting the frequent item-sets in the first phase, the second phase generates the rules using another user-defined parameter called minimum confidence [3], [4] and [6].

In the paper we improved the association rule-mining problem using genetic algorithm and we used the random sampling method for overcoming the rule.

II. Background Techniques

A. Association Rules

Association rules are if and then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a

Manuscript received January 20, 2014; revised July 5, 2014.

doi: 10.12720/jiii.3.2.114-118

consequent (then). Association rule is expressed as X=>Y, where X is the antecedent and Y is the consequent. Each association rule has two quality measurements, support and confidence. Support implies frequency of occurring patterns, and confidence means the strength of implication [1]-[3] and [9].

B. Genetic Algorithm

Genetic Algorithm (GA) was developed by Holland in 1970. This incorporates Darwinian evolutionary theory with sexual reproduction. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. GA process is an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings.

The most important biological terminology used in a genetic algorithm is:

- The chromosomes are elements on which the solutions are built.
- Population is made of chromosomes.
- Reproduction is the chromosome combination stage.
- Mutation and crossover are reproduction methods.
- Quality factor (fitness) is also known as performance index, it is an abstract measure to classify chromosomes.
- The evaluation function is the theoretical formula to calculate a chromosome's quality factor [4] and [6] and [10].

C. Genetic Operators

The GA maintains a population of n chromosomes (solutions) with associated fitness values. Parents are selected to mate, on the basis of their fitness, producing offspring via a reproductive plan (mutation and crossover). Consequently highly fit solutions are given more opportunities to reproduce (selected for next generation), so that offspring inherit characteristics from each parent. As parents mate and produce offspring, room must be made for the new arrivals since the population is kept at a static size (population size). In this way it is hoped that over successive generations better solutions will thrive while the least fit solutions die out. The representation scheme, Population Size, Crossover rate, Mutation rate, and fitness function and selection operator are the GA operators [4]-[6].

D. Genetic Algorithm for Association Rule Mining

Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. The evolutionary process of a GA is a highly simplified and stylized simulation of the biological version. It starts from a population of individuals randomly generated according to some probability distribution, usually uniform and updates this population in steps called generations. In each generation, multiple individuals are randomly selected from the current population based upon some application of fitness, bred using crossover, and modified through mutation to form a new population [7] and [8].

III. Related Works

Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad *et al*. proposed to optimize the rules generated by Association Rule Mining (apriori method), using Genetic Algorithms. In general the rule generated by Association Rule Mining technique do not consider the negative occurrences of attributes in them, but by using Genetic Algorithms (GAs) over these rules the system can predict the rules which contains negative attributes. The main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. The improvements applied in GAs are definitely going to help the rule based systems used for classification.

In this paper the authors have tried to use the enormous robustness of GAs in mining the Association Rules. The results generated when the technique applied on the synthetic database, includes the desired rules, i.e. rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining. [2].

Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M. K *et al*. is to find all the possible optimized rules from given data set using genetic algorithm. The rule generated by association rule mining algorithms like priori, partition, pincer-search, incremental, border algorithm etc, does not consider negation occurrence of the attribute in them and also these rules have only one attribute in the consequent part. By using Genetic Algorithm (GAs) the system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The major advantage of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach.

They have dealt with a challenging association rule mining problem of finding optimized association rules. The frequent itemsets are generated using the Apriori association rule mining algorithm. The genetic algorithm has been applied on the generated frequent itemsets to generate the rules containing positive attributes, the negation of the attributes with the consequent part consists of single attribute and more than one attribute.

The results reported in this paper are very promising since the discovered rules are of optimized rules [3].

Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona *et al.* presented a novel association rule mining approach that can efficiently

discovered the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance.

The developed approach adopts the philosophy of Apriori approach with some modifications in order to reduce the time execution of the algorithm. First, the idea of generating the feature of items is used and; second, the weight for each candidate itemset is calculated to be used during processing. The feature array data structure is built by storing the decimal equivalent of the location of the item in the transaction. Transforming here means reorganizing and transforming a large database into manageable structure to fulfill two objectives: (a) reducing the number of I/O accesses in data mining, and (b) speeding up the mining process. There is one mandatory requirements for the transforming technique, that the transaction database should be read only once within the whole life cycle of data mining. By storing the appearing feature of each interested item as a compressed vector separately, the size of the database to be accessed can be reduced greatly.

This paper is to improve the performance of the conventional Apriori algorithm that mines association rules by presenting fast and scalable algorithm for discovering association rules in large databases. The approach to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding new features to the Apriori approach. The proposed mining algorithm can efficiently discover the association rules between the data items in large databases. In particular, at most one scan of the whole database is needed during the run of the algorithm. Hence, the high repeated disk overhead incurred in other mining algorithms can be reduced significantly. They demonstrated the effectiveness of the algorithm using real and synthetic datasets. They developed a visualization module to provide users the useful information regarding the database to be mined and to help the user manage and understand the association rules [4].

IV. Proposed ARM Techniques Using Genetic Alhorithm

Association Rule Mining is computationally and Input/ Output intensive. The number of rules grows exponentially with the number of items. Because data is increasing in terms of the dimensions (number of items) and size (number of transactions), one of the main attributes needed in an Association Rule Mining algorithm is scalability: the ability to handle massive data stores. Sequential algorithms cannot provide scalability, in terms of the data dimension, size, or runtime performance, for such large databases.

In the present work we solve the association rule-mining problem with genetic algorithm. The first task for this is to represent the possible rules as chromosomes, for which a suitable encoding/decoding scheme is required. For this, two approaches can be adopted. In the previous approach each chromosome represents a set of rules, and this approach is more suitable for classification rule mining; as we do not have to decode the consequent part and the length of the chromosome limits the number of rules generated. The other approaches where each chromosome represents a separate rule. We have to encode the antecedent and consequent parts separately; and thus this maybe an efficient way from the point of space utilization since we have to store the empty conditions as we do not known a priori which attributes will appear in which part. So we followed a new approach that is better than this approach from the point of storage requirement. With each attribute we associate two extra tag bits. If these two bits are 00 then the attribute next to these two bits appears in the antecedent part and if it is 11 then the attribute appears in the consequent part. And the other two combinations, 01 and 10 will indicate the absence of the attribute in either of these parts. In this way we can handle variable length rules with more storage efficiency, adding only an overhead of bits.

The next step is to find a suitable scheme for encoding/decoding the rules to/ from binary chromosomes. Since the positions of attributes are fixed, we need not store the name of the attributes. We have to encode the values of different attribute in the chromosome only. For encoding a categorical valued attribute, the market basket encoding scheme is used. As discussed earlier this scheme is not suitable for numeric valued attributes. For a real valued attribute their binary representation can be used as the encoded value. The range of value of that attribute will control the number of bits used for it. Decoding will be simply the reverse of it. The length of the string will depend on the required accuracy of the value to be encoded. Decoding can be performed as:

$$Value = Minimum\ value - (maximum\ value - minimum\ value)\ X\ ((\sum (2^{i-1}\ X\ ith\ bit\ value)) / (2^n - 1))$$

where $1 <= i <= n$ and n is the number of bits used for encoding; and minimum & maximum are minimum and maximum values of the attribute.

Using these encoding schemes values of different attributes can be encoded into the chromosomes. Since in the association rules an attribute may be involved with different relational operators, it is better to encode them also within the rule itself.

To handle this situation we used another bit to indicate the operators involved with the attribute. Equality and not equality are not considered with the numerical attribute. In this way the whole rule can be represented as a binary string, and this binary string will represent one chromosome or a possible rule.

After getting the chromosomes, various genetic operators can be applied on it. Presence of large number of attributes in the records will results in large chromosomes, thereby needing multi-point crossover. There are some difficulties to use the standard multi-objective GAs for association rule mining problems. In case of rule mining problems, we need to store a set of better rules found from the database. If we follow the

standard genetic operations only, then the final population may not contain some rules that are better and were generated at some intermediate generations. It is better to keep these rules. For this task, a separate population is used. In this population no genetic operation is performed. It will simply contain only the non- dominated chromosomes of the previous generation. The user can fix the size of this population. At the end of first generation, it will contain the non-dominated chromosomes of the first generation. After the next generation, it will contain those chromosomes, which are non-dominated among the current population as well as among the non-dominated solutions till the previous generation.

A. Definitions

1) Support

The rule X ⇒ Y holds with support s if s% of transactions in D contains X ∪ Y. Rules that have a s greater than a user-specified support is said to have minimum support.

2) Confidence

The rule X ⇒ Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

3) Itemset: An itemset is a set of items. A k-itemset is an itemset that contains k number of items.

4) Frequent itemset: This is an itemset that has minimum support.

5) Candidate set: This is the name given to a set of itemsets that require testing to see if they fit a certain requirement.

6) Chromosome: A chromosome (also sometimes called a genome) is a set of parameters which define a proposed solution to the problem that the genetic algorithm is trying to solve. The chromosome is often represented as a simple string, although a wide variety of other data structures are also used. We have to redefine the Chromosome representation for each particular problem, along with its fitness, mutate and reproduce methods.

7) Fitness: Fitness (often denoted ω in population genetics models) is a central idea in evolutionary theory. It can be defined either with respect to a genotype or to a phenotype in a given environment. In either case, it describes the ability to both survive and reproduce, and is equal to the average contribution to the gene pool of the next generation that is made by an average individual of the specified genotype or phenotype. If differences between alleles at a given gene affect fitness, then the frequencies of the alleles will change over generations; the alleles with higher fitness become more common.

The most important part of Genetic Algorithm is a design of Fitness Function:

$$f(x) = M/N$$

where

M= Support (*x*) with condition is support (*x*) > Minsupport and *N*= support (*x*) with condition is support < minsupport

Support is the Support of New rule generated through genetic operation. Normal case the value of (Support(x) < minsupport) is rejected for the better performance of genetic algorithm. We have used class-learned classifier for the prediction for rejected those value near to the Maximum value.

B. The Suggested Approach will Work as Follows

Step1: Load a sample of records from the database that fits in the memory.
Step2: Generate N chromosomes randomly.
Step3: Decode them to get the values of the different attributes.
Step4: Scan the loaded sample to find the support of antecedent part, consequent part and the rule.
Step5: Find the confidence, comprehensibility and interestingness values.
Step6: Rank the chromosomes depending on the non-dominance property.
Step7: Assign fitness to the chromosomes using the ranks, as mentioned earlier.
Step8: Bring a copy of the chromosomes ranked as 1 into a separate population, and store them if they are non-dominated in this population also. If some of the existing chromosomes of this population become dominated, due to this insertion, then remove the dominated chromosomes from this population.
Step9: Select the chromosomes, for next generation, by roulette wheel selection scheme using the fitness calculated in Step 7.
Step10: Replace all chromosomes of the old population by the chromosomes selected in Step 9.
Step11: Perform multi-point crossover and mutation on these new individuals.
Step12: If the desired number of generations is not completed, then go to Step 3. Otherwise next step.
Step13: Decode the chromosomes in the final stored population, and get the generated rules.

V. Results Analysis

Association rule mining is one of the important tasks of data mining intended towards decision support. It is the process of finding some relations among the attributes/attribute values of a huge database. In the huge collection of data stored in a database, the relationships among various attributes may exist.

Sample Size	Number of Generations	Number of Rules Generated
500	100	14
	150	19
	200	20
700	100	25
	200	31
	300	32
800	100	25
	200	32
	300	32
1000	100	26
	200	33
	300	33

Our proposed technique implemented on different data sets with satisfactory results. Now we present the results on one such data set having 40 attributes and 3000 records. The crossover and mutation probabilities were taken respectively as 0.8 and 0.02; 5 point crossover operator was used for our experiment and the population size was kept fixed as 40. We sampled size and numbers of rules generated are put in the following Table.

From the rule sets generated for different samples and for different number of generations it is observed that after 200 generations it ceases to generate more rules; in other words after that number of generations the GA converges. From the results given above it can be seen that only for the third sample, it gives an extra rule at the cost of 100 additional generations. Moreover, only a very few number of attributes got involved in the rules, which means that all the attributes are not equally important; and the rules are simple to understand (comprehensible). The generated rules were not that much interesting (interestingness value was order of 0.005). If the confidence of the rule is used as one measure, sometimes some rules with SUP (A) = 1, SUP(C) = 1, and SUP(AC)= 1 may be generated. That rule will have a confidence 100%. So there is a chance that the rule may be declared as a non-dominated rule. But the records satisfying that rule may be noise also. Current algorithms do not face this problem, because the user parameter called minimum support eliminates the probability of generation of such rules. Instead of the confidence, we used the support of the rule as one measure to evaluate the rule thereby overcome.

VI. Conclusion

Discovering association rules is at the heart of data mining. Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. This thesis uses an association rule based genetic algorithm to solve the multi-objective rule mining problem using three measures comprehensibility, interestingness and the predictive accuracy. We discus an approach to represent the rules as chromosomes, where each chromosome represents a separate rule. To improve the efficiency of this algorithm, some refinement may be required. For example, this algorithm works on a sample of the original database, and the sample may not truly reflect the actual database. In the present work, we use the random sampling method. A perfect sample improved the correctness of the rules generated by the algorithm. Moreover, we test the approach only with the numerical valued attributes. It must be tested with the categorical attributes also.

References

[1] F. H. AL-Zawaidah, Y. H. Jbara, and M. AL-Abed Abu-Zanona, "An improved algorithm for mining association rules in large databases," *World of Computer Science and Information Technology Journal*, vol. 1, no. 7, pp. 311-316, 2011.

[2] P. P. Wakabi-Waiswa and V. Baryamureeba, "Extraction of interesting association rules using genetic algorithms," *Advances in Systems Modelling and ICT Applications*, pp. 101-110, 2008.

[3] M. R. Kumar and K. Iyakutti, "Genetic algorithms for the prioritization of associatio rules," *IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches & Practical Applications AIT*, pp. 35-38, 2011.

[4] S. Dehuri, A. K. Jagadev, A. Ghosh, and R. Mall, "Multi-objective genetic algorithm for association rule mining using a homogeneous dedicated cluster of workstations," *American Journal of Applied Sciences*, vol. 3, no. 11, pp. 2086-2095, 2006.

[5] R. Haldulakar and J. Agrawal, "Optimization of association rule mining through genetic algorithm," *International Journal on Computer Science and Engineering*, vol. 3, no. 3, pp. 1252-1259, Mar. 2011.

[6] K. Indira and S. Kanmani, "Performance analysis of genetic algorithm for mining association rules," *International Journal of Computer Science Issues*, vol. 9, no. 2, pp. 368-376, March 2012.

[7] S. Das and B. Saha, "Data quality mining using genetic algorithm," *International Journal of Computer Science and Security*, vol. 3, no. 2, pp. 105-112, 2009.

[8] F. Q. Shi, S. Q. Sun, and J. Xu, "Association rule mining of Dansei knowledge based on rough set," *Computer Integrated Manufacturing Systems*, vol. 14. pp. 407-411, 2008.

[9] M. Saggar, A. K. Agarwal, and A. Lad, "Optimization of association rule mining using improved genetic algorithms," in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, 2004, vol. 4, pp. 3725-3729.

[10] A. Salleb-Aouissi, C. Vrain, and C. Nortet, "Quant miner: A genetic algorithm for mining quantitative association rules," *IJCAI*, pp. 1035-1040, 2007.

Mr. Amit Kumar Barai presently working at Dredging Corporation of India Ltd (Mini-Ratna PSU)., Visakhapatnam, India.. He has more than four and half years experience in various reputed organization in India. The degree of B.E. secured in Computer Science & Engineering from North Bengal University, Darjeeling in 2004, M.Tech.in Computer Science & Engineering from S.A.T.I., Vidisha, India in 2009. Research Interest includes Network Security, Watermarking, Data Mining Techniques. Image Processing, Image compression, and *Individual Risk Premium Modification* (*IRPM*). He published three Research papers in various International Journals/Conference.