

# Association Rule Optimize by Multi Objective Genetic Algorithm

Deep Hakani<sup>1</sup> Harshita Kanani<sup>2</sup>

<sup>1,2</sup>Dept. of Computer Engineering,  
<sup>1,2</sup>LDRP, KSV Gandhinagar, Gujarat

**Abstract**—Association analysis is the task of bringing out relationship among data. Association analysis is most popular analysis technique in data mining for classifying biological data. When dealing with Biological data large search spaces may arise. Genetic algorithms deal with large search space very effectively. Combining association rule mining and Genetic algorithms to classify biological data is a novel and extensive research area. The purpose of this thesis is to classify biological data with association rule and genetic algorithm. A Genetic Algorithm (GA) is an iterative search, optimization and adaptive machine learning technique premised on the principles of Natural Selection. A GA is a search method that functions analogously to an evolutionary process in a biological system as it mimics evolution and competition between individuals in natural selection. It generates a better solution from existing solutions. Neither programmer nor genetic algorithm has to know how to solve a given problem; solution is just bred. Gas is one of the most robust problem solving techniques. They can find solutions of NP-hard problems easily. For problems with a larger parameter space and where the problem itself can be easily specified, GA can be an appropriate solution. In this paper described all the method of association rule mining with genetic algorithm. “Genetic Algorithms are software procedures modeled after genetics and evolution” Index Terms Ensemble System, Association, Classification.

## I. INTRODUCTION

Data mining, which is also referred to as knowledge discovery in databases (KDD), means a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases. There are also many other terms, which carry a similar or slightly different meaning, such as data dredging, knowledge extraction, data archaeology. [1]

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 leads 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.

## II. ASSOCIATION RULE MINING

Association rule is a data mining technique which discovers the strong associations or correlation relationships among given data. The association rules problem is as follows:

Let  $I = \{i_1, i_2, \dots, i_n\}$  be a set of literals call items. Let  $D$  be a set of all transactions where each transaction  $T$  is a set of items such that  $T \subseteq I$ . Let  $X, Y$  be a set of items such that  $X, Y \subseteq I$ . An association rule is an implication in the form  $X \Rightarrow Y$ , where  $X \subset I, Y \subset I, X \cap Y = \emptyset$  [1].

Association rule mining is to find out association rules that satisfy the pre-defined minimum support and confidence from a given database [1]. The problem is usually decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database, those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is  $L_k, L_k = \{i_1, i_2, \dots, i_k\}$ , association rules with this itemsets are generated in the following way: the first rule is  $\{i_1, i_2, \dots, i_{k-1}\} \Rightarrow \{i_k\}$ , by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem.

The support of an association rule is the percentage of groups that contain all of the items listed in that association rule. The percentage value is calculated from among all the groups that were considered. This percentage value shows how often they joined rule body and rule head occur among all of the groups that were considered.[1]

$$\text{support} = \frac{(X \cup Y).count}{n}$$

The confidence of an association rule is a percentage value that shows how frequently the rule head occurs among all the groups containing the rule body. The confidence value indicates how reliable this rule is. The higher the value, the more often this set of items is associated together.[1]

$$\text{confidence} = \frac{(X \cup Y).count}{X.count}$$

The discovery of association rules has been known to be useful in selective marketing, decision analysis, and business management. An important application area of mining association rules is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased together. With

the increasing use of the record-based databases whose data is being continuously added, recent important applications have called for the need of incremental mining. In dynamic transaction databases, new transactions are appended and obsolete transactions are discarded as time advances.

### III. GENETIC ALGORITHM

Genetic algorithms are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved.

The three most important aspects of using genetic algorithms are: (1) definition of the objective function, (2) definition and implementation of the genetic representation, and (3) definition and implementation of the genetic operators.

Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce.

This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied.

Depending on the kinds of the data to be mind or on the given data mining applications, data mining systems may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis; signal processing, Computer graphics, Web technology, economics, Business, bioinformatics or psychology.

Genetic algorithms were formally introduced in the United States in the 1970s by John Holland at University of Michigan. The continuing price/performance improvements of computational systems have made them attractive for some types of optimization. In particular, genetic algorithms work very well on mixed (continuous and discrete), combinatorial problems. They are less susceptible to getting 'stuck' at local optima than gradient search methods. But they tend to be computationally expensive. [13]

To use a genetic algorithm, you must represent a solution to your problem as a genome (or chromosome). The genetic algorithm then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one(s).

This presentation outlines some of the basics of genetic algorithms. The three most important aspects of using genetic algorithms are: (1) definition of the objective function, (2) definition and implementation of the genetic representation, and (3) definition and implementation of the genetic operators. Once these three have been defined, the generic genetic algorithm should work fairly well. Beyond that you can try many different variations to improve performance, find multiple optima (species - if they exist), or parallelize the algorithms.

After an initial population is randomly generated, the algorithm evolves the through three operators: selection which equates to survival of the fittest; crossover which represents mating between individuals; mutation which introduces random modifications.

#### A. Multi-Objectives Genetic Algorithm [12]

Being a population based approach, GA are well suited to solve multi-objective optimization problems. A generic single-objective GA can be easily modified to find a set of multiple non-dominated solutions in a single run. The ability of GA to simultaneously search different regions of a solution space makes it possible to find a diverse set of solutions for difficult problems with non-convex, discontinuous, and multi-modal solutions spaces. The crossover operator of GA may exploit structures of good solutions with respect to different objectives to create new non-dominated solutions in unexplored parts of the Pareto front. In addition, most multi-objective GA do not require the user to prioritize, scale, or weight objectives. Therefore, GA has been the most popular heuristic approach to multi-objective design and optimization problems. Jones et al. reported that 90% of the approaches to multiobjective optimization aimed to approximate the true Pareto front for the underlying problem. A majority of these used a meta-heuristic technique, and 70% of all meta-heuristics approaches were based on evolutionary approaches.

The first multi-objective GA, called Vector Evaluated Genetic Algorithms (or VEGA), was proposed by Schaffer. Afterward, several major multi-objective evolutionary algorithms were developed such as Multi-objective Genetic Algorithm (MOGA), Niche Pareto .

Applying Genetic Algorithms (GA) to solve multiple objectives optimization problems has to deal with the duplicate issues of searching large and complex solution spaces and dealing with multiple, potentially conflicting objectives. Selection of a solution from a set of possible ones on the basis of several criteria is considered a difficult problem. Due to this difficulty, most of researchers reduce the problem to a monocriterion one. Mathematical programming techniques and the popular weighted sum approach have been developed. On the meta-heuristic side, Schaffer was one of the first to recognize the possibility of exploiting Evolutionary Algorithm's to treat multiple-objectives problems.

Classical GA's use fitness-based selection, and thus require scalar fitness information. So the objectives are often artificially combined into a scalar function.

Other GA's use ranking methods to grade the population, without using the Decision Maker's (DM) preferences. Since the best solution may not necessarily belong to the pareto optimal set, the classical methods are a kind of local optimality search rather than a global one.

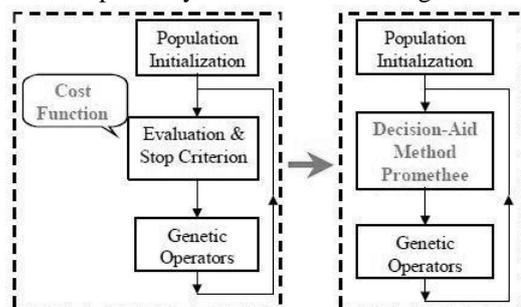


Fig. 1: Classic and New Multi-Objectives GA Structure

To come out of this kind of problems, we use the multi-criteria decision-aid (MCDA) method called Promethee II [2]. It computes a 'net flow' (f) associated with

each solution. This flow gives us a ranking, called the Promethee II complete ranking, between the different solutions in the population. The weights (associated with each criterion) are involved in the computation of the  $f$  number and represent the relative influence of each criterion. Thus the solutions are not compared according to a cost function yielding an absolute fitness of the individuals as in a classical GA, but are compared to each other thanks to flows, depending on the current population. In order to avoid a drift towards locally optimal solutions, elitism is used, i.e., the best-ever solution takes part in the evaluation of the F flows.

The choice of one solution over the others requires problem knowledge. It is the DM's task to adjust the weights to help the algorithm to find good solutions. Optimizing a combination of the objectives has the advantage of producing a single solution, requiring no further interaction with the DM. If this 'optimal' solution cannot be accepted, due to inappropriate settings of the weights, new runs may be required to adjust them until a suitable solution is found.

#### IV. EXISTING METHODS

##### A. Optimization of Association Rule Mining through Genetic Algorithm [5]

Apriori algorithm is used to generate the rules after that they used the optimization techniques. Gas use to optimize rules. Genetic Algorithm directly not work on the raw data then whole data had encoded in the form of Binary representation technique (0 and 1). The most important part of Genetic Algorithm is a design of Fitness Function:

$$f(x) = \frac{\text{Support}(x)}{\text{Minsupport}}$$

The value of  $q$  class is divided into two parts C1 and C2.  
 $q = \{C1, C2\}$

$C1 = \{\text{those value or Data minsupport less than } 0.5\}$

$C2 = \{\text{those value or Data minSupport greater than } 0.5\}$

$$f(q) = \frac{\text{Support}(C2)}{\text{Minsupport}}$$

The selection strategy based on the basis of individual fitness and concentration  $\pi_i$  is the probably of selection of individual whose fitness value is greater than one and  $f(\alpha)$  is a those value whose fitness is less than one but near to the value of 1. Now  $\pi_i$  Where  $\alpha$  is an adjustment factor.

$$\pi_i = \frac{f(x_i) \cdot e^{-\alpha f(\alpha)}}{\sum f(x_j)}$$

##### 1) Genetic Operation:

Select operation: In this algorithm it restores each chromosome in the population to the corresponding rule, and then calculate selection probability  $\pi_i$  for each rule based on formula.

Then apply Crossover operation and Mutation Operation.

##### 2) Rules extraction:

Extraction criteria are: output the rule which meets the minimum confidence given by users, otherwise abandon it.

##### B. Optimization of Association Rule Mining using Improved Genetic Algorithms [6]

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 genetic algorithms are applied over the rules fetched from association rule mining. Now, the authors firstly implemented Association Rule mining (using a-priori technique) by the help of their toolkit. And then the GAs are applied to evolve the rules which contains the negations in attributes and are of richer quality. Using this method rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining.

##### C. Association Rules Mining using Modified Genetic Algorithm [7]

In generate rules using Association rule and then apply Genetic algorithm to optimize rule is time taking process. So here to optimize rule apply modify genetic algorithm which time complexity is lesser. First apply Apriori for rules generations then apply genetic algorithm use support and confidence for fitness function. Below shown algorithm step for this algorithm.

- 1) Load the given dataset and convert the data into numeric values.
- 2) Apply Apriori algorithm to generate the association rules.
- 3) Define the fitness function for specific support & confidence values.
- 4) Generate N chromosomes randomly depending upon dataset length.
- 5) Iterate it & select the rule which satisfies the fitness conditions (according to genetic algorithm).

##### D. Optimized association rule mining using genetic algorithm [8]

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. In this paper the genetic algorithm is applied over the rules fetched from Apriori association rule mining.

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. This technique need to minimize the complexity of the genetic algorithm and scanning of database by applying theorem on the generated rule [3].

#### E. An Improved Algorithm for Mining Association Rules in Large Databases [9]

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.

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. The proposed technique need to improve in the mining multidimensional association rules from relational databases and data warehouses and also in mining multilevel association rules from transaction databases Use for transaction databases

#### F. Extraction of Interesting Association Rules Using Genetic Algorithms [10]

They proposed to use multi-point crossover operator. There were some difficulties to use the standard multi-objective GAs for association rule mining problems.

If they 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. These rules should be kept. For this task, an external population is used. In this population no genetic operation is performed. It will simply contain only the non-dominated chromosomes of the previous generation. 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.

#### G. Genetic algorithms for the prioritization of Association Rules [11]

M. Ramesh Kumar and Dr. K. Iyakutti et. al. proposed a novel genetic algorithm based association rule mining algorithm is discussed in this paper. Prioritization of the rules has been discussed with the help of genetic 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 is been tested for the four data sets like Adult, Chess, Wine, Zoo. They presented a novel algorithm for the rule prioritizing, generated by the apriori algorithm through the application of genetic algorithm. The fitness function is designed based on the user's interesting measure and M is the threshold value of the interesting measure considered. 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 cross over rate is 6% and the mutation rate is fixed as 1%. The proposed genetic algorithm based association rule mining algorithm for the prioritization of

the rules. This approach significantly reduces the number of rules generated in the data sets. The fitness function is designed in such a way that to prioritize the rules based on the user preference. The technique can be extended by the incorporation of the other interesting measures in the literature to future work [6].

## V. PROPOSED METHOD

CBAMOGA, i.e. Classifying Biological Data using Association Rule Mining And Multiple Object Genetic Algorithm. This method mining biological data efficiently. It applies association rule analysis to find frequent itemset. Then apply multiple objective genetic algorithms. Select set of rule from generated rules. then evaluated it using three fitness function. This completes the first iteration then for the next iteration for selecting subset of rules apple crossover and mutation and generates new population. Old population replace it with new one. each time apply three fitness function to evaluate rules

### A. Algorithm

Step. 1 : Start

Step. 2 : Load a sample of records from the database that fits in the memory.

Step. 3 : Apply Association rule mining algorithm to find the frequent itemsets with the minimum support.

Step. 4 : Store frequent itemsets in A, where Suppose A is set of the frequent item set generated by Association rule mining algorithm.

Step. 5 : Set B=  $\Phi$  where B is the output set, which contains the association rule.

Step. 6 : Pre-process dataset in to form of chromosomes.

Step. 7 : Input the termination condition of Multi objective genetic algorithm.

Step. 8 : Select the two members from the frequent itemset using sampling method.

Step. 9 : Apply the crossover and mutation on the selected members to generate the association rules.

Step. 10 : Find three fitness functions( Fco(x), Fj(x), Fc(x) ) for each rule  $x \rightarrow y$  and check the following condition.

Step. 11 : Fco(x)=Comprehensibility =  $\log(1+ |Y|) + \log(1+ |XUY|)$

Step. 12 : Fj(x)=  $JM=f(y|x). \log_2 \frac{f(y|x)}{f(y)} + f(y|x). \log_2 \frac{1-f(y|x)}{1-f(y)}$  (1-

Step. 13 : Fc(x)=Confidence =  $(XUY) / (X)$

Step. 14 : if (Fco(x) > Comprehensibility, Fj(x) > J-measure, Fc(x) > min confidence)

Step. 15 : set B = B  $\cup \{x \rightarrow y\}$

Step. 16 : If the desired number of generations is not completed,

Step. 17 : then go to Step 8.

Step. 18 : classify data using Rule set B

Step. 19 : Stop

### B. Description

In first step load DATA set. Then Apply Association rule mining algorithm to find the frequent itemsets with the minimum support. Use Fp-growth algorithm for association rule mining. If rules is not satisfying minimum threshold

then remove it.

Store this frequent itemset set in to set A. No of rules generated by the fp-growth is very large so to reduce it and take more important rules among them. First take new empty set B in which store selected rule set. Now to apply CBAMOGA this algorithm first convert dataset into chromosome. So, apply pre-process2. It converts data set in to chromosome.

Then initialize 1st population. Then rules will be evaluating using multi object fitness function, there are three fitness function  $F_{co}(x)$ ,  $F_j(x)$ ,  $F_c(x)$ . Equations of this fitness function are shown below.

$F_{co}(x) = \text{Comprehensibility} = \log(1 + |Y|) + \log(1 + |XUY|)$

Here,  $|Y|$  and  $|XUY|$  are the number of attributes involved in the consequent part and the total rule, respectively.

It is important that we extract only those rules that have a comparatively less occurrence in the entire database. Such a surprising rule may be more interesting to the users; which again is difficult to quantify. It is important that we extract only those rules that have a comparatively less occurrence in the entire database. Such a surprising rule may be more interesting to the users; which again is difficult to quantify.

$$F_j(x) = JM = f(y|x) \cdot \log_2 \frac{f(y|x)}{f(y)} + (1 - f(y|x)) \cdot \log_2 \frac{1 - f(y|x)}{1 - f(y)}$$

The J-Measure is a good indicator of the information content of the generated rules. In rule inference we are interested in the distribution of the rule "implication" variable Y, and especially its two events y and complement y. The purpose is to measure the difference between the priori distribution  $f(y)$ , i.e.  $f(Y = y)$  and  $f(Y \neq y)$ , and the posteriori distribution  $f(Y | X)$ . The J-Measure gives the average mutual information between the events y and  $f(Y = y)$ . The J-Measure shows how dissimilar our a priori and posteriori beliefs are about Y meaning that useful rules imply a high degree of dissimilarity.

$$F_c(x) = \text{Confidence} = (XUY) / (X)$$

It is same as confidence in apriori. Use of confidence is weather the rules are satisfying minimum confidence then we can take that rule for evaluation. [15]

If rule satisfy all fitness then select it put it into Set B. if criteria not match, create new population. To create new population use crossover operation then applies mutation operation. Then select some set of population which is not more useful for mining, and replace it with new generated population again evaluate it. This looping process is continue until N no of rule are not generated. After N iteration set B contain most important rules. Using that rules classify data. It gives more aureate classification because of three different fitness functions.

## VI. 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.

## REFERENCES

- [1] Agrawal R., Imielinski T., and Swami A., Mining association rules between set of items in large databases. In Proceedings of ACM SIGMOD, pages 207-216, May 1993
- [2] <http://www.obitko.com/tutorials/genetic-algorithms/ga-basic-description.php>
- [3] From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"
- [4] <http://www.cs.uic.edu/~liub/teach/cs583-fall-05/CS583-intro.ppt>
- [5] Rupali Haldulakar and Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm", International Journal on Computer Science and Engineering (IJCSSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.
- [6] Manish Saggarr, Ashish Kumar Agarwal and Abhimunya Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms" IEEE 2004.
- [7] Ms. Ruchi Saxena, Prof. Shailendra Shrivastava, Prof. Abhishek Mathur "Association Rules Mining using Modified Genetic Algorithm" International Journal of Scientific Engineering and Technology Volume No.1, Issue No.4, pg : 35-38
- [8] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., "Optimized association rule mining using genetic algorithm", Advances in Information Mining, ISSN: 0975-3265, Volume 1, Issue 2, 2009, pp-01-04.
- [9] Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona, "An Improved Algorithm for Mining Association Rules in Large Databases", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 2011, pp. 311-316.
- [10] Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba, "Extraction of Interesting Association Rules Using Genetic Algorithms", Advances in Systems Modelling and ICT Applications, pp. 101-110
- [11] M. Ramesh Kumar and Dr. K. Iyakutti, "Genetic algorithms for the prioritization of Association Rules", IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" AIT, 2011, pp. 35-38. An Introduction to the WEKA Data Mining System Zdravko Markov Central Connecticut State University
- [12] M. Ramesh Kumar Dr. K. Iyakutti "Application of Genetic algorithms for the prioritization of Association Rules" IJCA Journal Number 3 - Article 7
- [13] Multiple-Objectives Genetic Algorithm Brahim Rekiek Department of Applied Mechanics, CAD Unit, Brussels University (ULB) Av. F. D. Roosevelt 50, CP 165/41, B-1050 Brussels, Belgium
- [14] [http://en.wikipedia.org/wiki/Weka\\_%28machine\\_learning%2](http://en.wikipedia.org/wiki/Weka_%28machine_learning%2)