ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 11, November 2013

Classifying biological Data based on Association rule using Multi-Objective Genetic Algorithm

Deep Hakani, Dept. of Computer Engineering, LDRP, KSV, Gandhinagar, Gujarat,ph.no- 9724082291.
Harshita Kanani, Assistant Professor, Dept. of Computer Engineering, LDRP, KSV, Gandhinagar, Gujarat, ph.no-9924505607.

Abstract— **In data mining many techniques are available for predicting of frequent pattern. One technique is association rule mining algorithm which can solve critical problem in biological field but this algorithm has limitation of space, time complexity and accuracy. Classification rules are one kind of conditional rules which can be used to discover data from large data sets. The model that is built is called a Classifier or Predictor depending upon whether the model finds the unknown data class or data value.**

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.

"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]

- **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 = {i1, i2, …, in} 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 ⊆ I. Let X, Y be a set of items such that X, Y ⊆ I. An association rule is an implication in the form X ⇒ Y, where X ⊂ I, Y ⊂ I, X ∩ Y = ∅ [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 Lk, Lk= fI1, I2,….., Ik ¡1,Ikg, association rules with this itemsets are generated in the following way: the ¯rst rule is fI1, I2,…, Ik ¡1 g) fIkg, 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]

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

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]

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

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.

- **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. [11]

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:

1. **selection** which equates to survival of the fittest;
2. **crossover** which represents mating between individuals;
3. **mutation** which introduces random modifications.

Multi-Objectives Genetic Algorithm[10]

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 rue 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) , Niched 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.

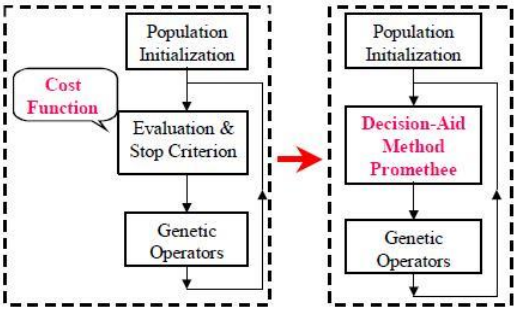

Figure1.1 : Classic and New Multi-Objectives GA Structure

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. The basic steps of the classical GA are shown below.

Generate an initial population;
Evaluate fitness of individuals in the population;
repeat
  Select parents from the population;
  Recombine parents to produce children;
  Mutate children;
  Evaluate fitness of the children;
  Replace some or all of the population by children;
until a satisfactory solution has been found;
The new MOGA steps are the following:
Generate an initial population;
Order individuals in the population using Promethee II;
repeat
  Select parents
  Recombine best parents from the population;
  Mutate children;
  Use Promethee II to order the new population;
  Replace some or all of the population by children;
until a satisfactory solution has been found;

The method is integrated in the grouping genetic Algorithm (GGA), and uses a group-oriented encoding. We apply it to the design of hybrid assembly lines, dealing with many objectives (cost, balance, reliability, congestion,..)

II. EXISTING METHODS

Table 1 : Comparison of all existing methods

Techniques	Description	Advantages/ Merits	Disadvantages
Association Rule, GA	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}	The proposed optimize association rule mining using new fitness function. In which fitness function divide into two classes" c1 and c2 one class for discrete rule and another class for continuous rule. Through this direction it got a better result.	The genetic algorithm does not sufficient effective and it can" t incorporate with other techniques, so it will need to improve in future work [5]

	C1 = {those value or Data minsupport less than 0.5} C2 = {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 is the probably of selection of individual whose fitness value is greater than one and f(α) is a those value whose fitness is less than one but near to the value of 1. Now pi Where α is an adjustment factor. $\text{pi} = \frac{f(xi).e^{-\alpha f(\alpha)}}{\sum f(xj)}$ **Genetic Operation:** ***Select operation****:* In this algorithm it restores each chromosome in the population to the corresponding rule, and then calculate selection probability **pi** for each rule based on formula. Then apply ***Crossover operation and Mutation Operation.*** ***Rules extraction:*** Extraction criteria are: output the rule which meets the minimum confidence given by users, otherwise abandon it.		
Data Mining, GAs, Association Rule Mining	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.	They 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.	This technique needs major modifications to improve the complexity reduction of Association rule mining and Genetic Algorithms by using distributed computing [6]
GA, Association Rules, Support,	the genetic algorithm is applied over the rules fetched from Apriori association rule mining. The proposed method for generating association rule by genetic algorithm is as follows:	The genetic algorithm has been applied on the generated frequent itemsets to generate the	This technique need to minimize the complexity of the genetic algorithm and scanning of database

Confidence	• Load a sample of records from the database that fits in the memory. • Apply Apriori algorithm to find the frequent itemsets with the minimum support. Suppose A is set of the frequent item set generated by Apriori algorithm. • The output set, which contains the association rule. • Input the termination condition of genetic algorithm. • Represent each frequent item set of A as binary string. • Select the two members from the frequent item set using Roulette Wheel sampling method. Apply the crossover and mutation on the selected members to generate the association rules. • Find the fitness function for each rule • if (fitness function > min confidence) • If the desired number of generations is not completed, Apply Apriori algorithm to find the frequent itemsets with the minimum support again.	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.	by applying theorem on the generated rule [7]
Association rules; Frequent Patterns; Apriori	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	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.	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 [8]
Association Rule Mining, GA	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	The proposed approach dealt with a challenging NP-Hard association rule mining problem of finding interesting association rules. The results of this paper were good since the discovered rules are of a high predictive accuracy and of a high interesting value.	This technique does not sufficient reliable for a large dataset, it need to improve for the application in large data set [9]

	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.		
Association Rule, Apriori, GA	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 fitness function is designed based on the users 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. 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 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 [10]

III. PROPOSE METHOD

There are two proposals

1)Preprocessing:

In this approach, first apply Genetic Algorithm for pre-processing DATA base. After apply the Genetic Algorithm reduce no of attribute for classification. Now apply Association rule base classification on data base pre-processed by genetic algorithm.

2)Classification Using Genetic Algorithm:

In this, Apply Association rule mining algorithm to find the frequent itemsets with the minimum support. The output set, which contains the association rule. After genetic algorithm is applied over the rules fetched from association rule mining.

Here, use second Approach first apply association rule mining then apply Genetic Algorithm to find best rule generated by the association rule mining algorithm, And classified data using that algorithm. Because in first approach apply genetic algorithm for pre-processing of data it reduce numbers of attributes but it may possible that some important attribute may also be reduce. So, it is better to use second approach because it first apply association rule on all attributes and then genetic algorithm to find best rule among them to classify data correctly.

CBAMOGA(Classifying Biological Data using Association rule mining Multi Objective Genetic Algorithm)

CBMOGA, i.e. Classifying Biological Data using Multiple Object Genetic Algorithm. This method mining biological data efficiently. It apply association rule analysis to find frequent itemset. Then apply multiple objective genetic algorithm. 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 generate new population. Old population replace it with new one. each time apply three fitness function to evaluate rules.

ALGORITHM

1. Start
2. Load a sample of records from the database that fits in the memory.
3. Apply Association rule mining algorithm to find the frequent itemsets with the minimum support.
4. Store frequent itemsets in A, where Suppose A is set of the frequent item set generated by Association rule mining algorithm.
5. Set B= Φ where B is the output set, which contains the association rule.
6. Pre-process dataset in to form of chromosomes.
7. Input the termination condition of Multi objective genetic algorithm.
8. Select the two members from the frequent item set using sampling method.

9. Apply the crossover and mutation on the selected members to generate the association rules.
10. Find Three fitness function (Fco(x), Fj(x), Fc(x)) for each rule x→y and check the following condition.
11. if (Fco(x) > Comprehensibility, Fj(x) > J-measure, Fc(x) > min confidence)
12. set B = B U {x →y}
13. If the desired number of generations is not completed,
    then go to Step 8.
14. classify data using Rule set B
15. Stop

Above see the algorithm for this meathead first apple fp-groth to find out rule then apply Multi Objective Genetic Algorithm to classify rule.

IV. CONCLUSION

Association rule mining can't deal with large database as well as can't deal with dense database so, to classify this type of database we use multi objective genetic algorithm. We propose a method CBAMOGA(Classifying Biological Data using Association rule mining Multi Objective Genetic Algorithm) is for the prioritization of the rules. Here in all existing methods to classify use one and only one fitness function or weighted fitness function but in our proposed method we use three different fitness function so, we get better results than existing methods and it also deal with large search space. By using this method we classify biological data correctly.

V.ACKNOWLEDGEMENT

The best way to have a good idea is to have a lot of ideas to choose and evolve from. I sincerely feel that the credit of a major research work could not be narrowed down to only one individual. This dissertation also involves many valuable contributions and is a combined effort of many people who have helped me for last one year throughout my work since its commencement. I could achieve successful completion of this thesis through co-operation and effective guidance of all these people whom I would like to thank for their patience and contributions

.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 (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.

[6] Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms"IEEE 2004.

[7] 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.

[8] 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.

[9] 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. [6] 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

[10] 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

[11] http://en.wikipedia.org/wiki/Weka_%28machine_learning%2