Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Genetic Algorithm in Data Mining

Ranno Agarwal

Scholar M.Tech (CSE)

U.P. Technical University, Lucknow (UP) India

Abstract- GAs are used in various fields of Data mining to get the optimized solutions for the better performance of the data that are required in decision making and process the accurate result. genetic algorithm for classification rule mining techniques that discover comprehensible if then rules using a generalized uniform population method and a uniform operator inspired from the uniform population method. Initial population is generated by methodically eliminating the randomness by generalized uniform population method . In the subsequence generations, genetic diversity is ensured and premature convergence is prevented by the uniform operator

Keywords- Genetic algorithm, Operator for Genetic algorithm, Encoding, Uniform operator, The fitness function, Generalized uniform population, Genetic algorithm application

I. INTRODUCTION

Genetic algorithm are stochastic search methods which have been inspired by the process of biological evolution. Because of gas robustness and their uniform approaches to large number of different classes of problems, they have been used in many applications. Data mining is also one of the important application fields of Genetic algorithm .

In data mining a Genetic algorithm can be used either to optimize parameters for other kind of data mining algorithms or to discover knowledge by itself. The advantage of Genetic algorithm become more obvious when the search space of a task is large. Genetic algorithms are a probabilistic search and evolutionary optimization approach Which is inspired by Darwin's theory about evolution. And this technique used in computing to find exact or approximate solution to optimization and search problems.

II. GENETIC ALGORITHM AS DATA MINING TECHNIQUES

Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. 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.

In this we can take a set of population solution represented by chromosomes from one population are taken and that is used to create a new population and hope that the new generation is better then the existing previous one. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. The main steps are-

1. Start with a randomly generated population of n chromosomes
2. Calculate the fitness f(x) of each chromosomes x in the population.
3. Repeat the steps until n offspring have been created.
   - randomly select a pair of parent chromosomes from the current population
   - cross the pair at a randomly chosen point to form two offspring
   - randomly mutate the two offspring and add the resulting chromosomes to the population
   - calculate the fitness of the resulting chromosomes

4 Let the n finest chromosomes survive to next generation

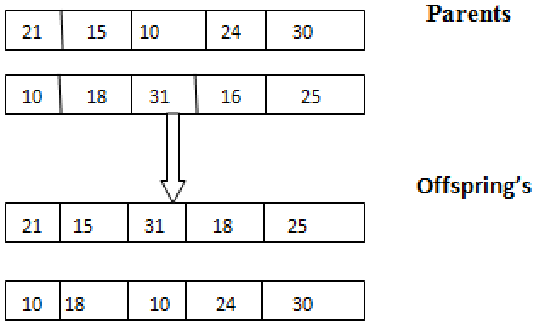
5 Go to step 3

III. OPERATORS FOR GENETIC ALGORITHM

A. Crossover

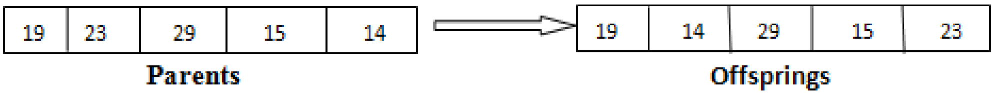
crossover probability crossover the parents to form a new offspring(children). If no crossover was performed, offspring is an exact copy of parents We use a Subset Size-Oriented Common Feature Crossover Operator (SSOCF) which keeps useful informative blocks and produces offspring's which have the same distribution than the parents. Offspring's are kept, only if they fit better than the least good individual of the population.

© 2015, IJARCSSE All Rights Reserved



B. Mutation

During the mutation stage, a chromosome has a probability *pmut* to mutate. If a chromosome is selected to mutate, we choose randomly a number *n* of bits to be flipped then *n* bits are chosen randomly and flipped. The mutation is an operator which allows diversity. With a mutation probability mutate new offspring at each locus (position in chromosome).



C. Selection

Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected

IV. THE PROPOSED GENETIC ALGORITHM:

A. Encoding:

let n be the no of predicted attributes In the data being mined, then a chromosomes is composed of n genes where each genes corresponds to a condition containing one attributes. Each ith gene is partitioned into 3 fields: Flag (fi), relational operator (RO i) and value (n i) as in fig:

A chromosome correspond to entire IF part of the rule and each gene is corresponds to one condition in the IF part. The chromosomes do not involve the class predicted by a rule . In a given run of genetic algorithm all chromosomes are searching for rules predicting the same class.

The genetic algorithm is run at least once for each class (value of the goal attributes)

Gene1			....	....	gene n			
F1	RO1	V1	....	....	Fn	Ron	Vn	

Fig: chromosomes representation

The flag field (fi) is a binary valued variable in the range. This variable indicated whether or not the corresponding attributes is involved in the rule 1 shows that the corresponding condition will be involved in the rule while 0 shows the condition will be removed from the rule. Although each chromosomes has a fixed length, the genes are interpreted in a such a way that the rule has a variable length , hence, different chromosomes corresponds to rule with different number of condition.

The relational operator (RO i)field is a variable related to categorical or continuous range of the ith condition . it indicates the relational operator used in the ith condition..if the attributes is categorical this field can involve the operator = /=. If the attributes is continuous , this field can involve the operator < =. The value (v) field involves one of the values belonging to the domain of the attribute.

B. Uniform operator

It is an operator used to obtain high quality chromosomes in each generation of the genetic algorithm. Based on the problem in each generation four or more with high quality chromosomes from two different best chromosomes are generated.

The position of different bits in the chromosomes are saves for binary encoding and the array size of which is the no of different position bits are generated, then from the array four different array are generated similar to uniform population method.

C. fitness function

The fitness function combines two indicators commonly used in medical domain ,namely the sensitivity and the specificity. the term used .are:

TP=no of true positive instances,

FP= no of false positive instances,

© 2015, IJARCSSE All Rights Reserved

FN= no of false negative instances,
TN= no of true negative instances

Sensitivity is the probability of the test of the finding disease among those who have the diseases or the proportion of the people with diseases who have a positive test result

$$\text{Sensitivity}=TP= (TP+FN)$$

Specificity is the probability of the test finding no diseases among those who do not have the disease or the proportion of the people free of diseases who have a negative test

$$\text{Specificity}=TN= (TN+FP)$$

Finally the fitness function used for classification is defined as the product of these two indicators;

$$\text{Fitness} =\text{sensitivity} + \text{specificity}$$

***D. Generalized uniform population*:**

In this study creating initial population is performed in a systematic way to inspired by the uniform population method. In this method it is assumed that the range of genes in the chromosomes has been known. First two chromosomes are generated according to these ranges all genes of which is in the lower bound and the other of which is in the upper bound of this range.

Let these chromosomes be C0 , C1 respectively as in fig. A is represent set of genes with equal length , if possible.

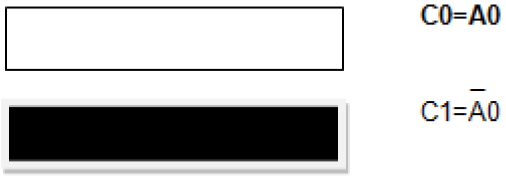


Fig: Initial chromosomes

C1 can be thought as the complement of C0 and the other chromosomes will be generated based on the complementation. The proposed method will be explained for binary encoding but this method can also be applied to other types of encoding.

In uniform population method, 2r-1 new chromosomes are generated from a chromosomes randomly. here ,if the population size is fixed 2r-2 new chromosomes will be generated in a systematic way .

By this method, initial population is distributed in the feasible resion uniformly. Chromosomes are not far away from the solution and are prevented to go and remain on stack to a local solution, this method can be applied to all types of encoding such as binary encoding, floating point based encoding, and string based encoding.

In real type encoding ,a random number can be generated in the half open interval and each gene complemented in the binary encoding are multiplied by this random umber.

V. APPLICATION OF GENETIC ALGORITHM

There is a greater scope of GA in data mining in future application to stimulate the data mining concepts. Genetic algorithms are widely applicable to classification by means of inductive learning. . GAs also provides a practical method for optimization of data preparation and data transformation steps Genetic Algorithms is an effective tool to use in data mining and pattern recognition. The main applications of Genetic algorithm are

- **Financial data analysis**
  Barclay’s global investors
  Pan agora asset management
  Fidelity funds
- **Engineering design**
  General electric
  Boeing
- **Operations and supply chain management**
  General motors
  Volvo
  Cemex

VI. CONCLUSION

Genetic algorithm is a search technique used in computing to find exact or approximate solution to optimization and search problems Genetic algorithms are inspired by Darwin's theory about evolution. Genetic algorithms are categories as global search heuristics Genetic algorithms are a probabilistic search and evolutionary optimization approach. The Genetic algorithm proposed as a search strategy to find accurate and comprehensible knowledge within large database that may be considered as search space .GA will evaluate each individual as a potential solution according to a predefined evaluation function .The evaluation function assigns a value of goodness to each individual based on how well the individual solves a given problem.

ACKNOWLEDGMENTS

Our thanks to the experts and authors of referenced journals who have contributed towards development of the paper and help us for making the concepts clear.

REFERENCES

[1] M. Pei, E.D. Goodman, and W.F. Punch. Feature extraction using genetic algorithms. Technical report, Michigan State University : GARAGe, June 1997.

[2] Luo , Qi . (2008). "Advancing Knowledge Discovery and Data Mining" Knowledge Discovery and Data Mining, 2008.WKDD 2008

[3] A Genetic Algorithm for Fe *Laetitia Jordan∗ Clarisse Dhaenens∗ El-Ghazali Talbi∗* LIFL, University of Lille, Bˆat M3 Cit´e Scientifique 59655 Villeneuve d'Ascq cedex Franceature S election in Data-Mining for Genetics

[4] Kamble, Atul (2010). "Incremental Clustering in Data Mining using Genetic Algorithm". International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010.