# Genetic Algorithm for Classification in Data Mining

Anuradha Purohit, Jai Kumar Rai, Pooja Sharma, Shashank Soni, Vaishali Gupta

*Computer Technology and Applications Department, SGSITS, Indore(M.P)*

anuradhapurohit@rediffmail.com

jai.r09@yahoo.co.in

pooja.sgsits@hotmail.com

shashank.cool1986@gmail.com

guptavaishali10@yahoo.in

***Abstract*— This paper focuses on performing rule based classification in Data Mining using Genetic Algorithm (GA). The objective functions namely comprehensibility, predictive accuracy, interestingness, coverage are used to evaluate the fitness value of each classification rule obtained using GA. For presenting our approach a real dataset is used. Experimental results shows that the rule based classification designed using GA efficiently identify the pattern using multiple objective functions.**

**Keywords—** Genetic Algorithm, data mining, classification, fitness function, crossover, mutation, reproduction.

## I. Introduction

Genetic Algorithms are stochastic search algorithms that borrow some concepts from nature [1]. It is one of the evolutionary computational methods. It works on a population of possible solution and it is probabilistic. It is different from most of the traditional optimization methods.

GA is a method for moving from one population of "chromosomes" (e.g. strings of zero and ones or "bits") to a new population by a kind of "natural selection" together with genetics inspired operators of crossover, mutation and reproduction. Each chromosome consists of "genes"(e.g. "bits"), each gene being an instance of particular "allele". The selection operators chooses those chromosomes in the population that will be allowed to reproduce, and on average the fitter chromosomes produce more off springs then the less fit ones[2]. The crossover operator oversees the mating process of two chromosomes. The crossover operator decides what genetic material from each parent is passed on to the child chromosome. The mutation operator takes each chromosome in the offspring pool and randomly changes part of its genetic make-up.

The process of reproduction, cross-over, mutation and formation of new population completes one generation cycle. GA was initially used for machine learning system, but it was soon realized that GA has great potential in function optimization. As GA is a search optimization technique hence it can be used in data mining. GA is a powerful tool for solving problems and for simulating natural systems in a wide variety of scientific fields.

## II. Basic concepts

Knowledge Discovery in Databases (KDD) is the process of automatic discovery of previously unknown patterns, rules and other regular contents implicitly present in large volume of data [3]. The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact, more abstract or more useful. The notions of finding useful patterns in data have been given a variety of names, including data mining, knowledge extraction, information discovery and pattern processing.

The rules are discovered/mined using four objectives: namely, predictive accuracy, comprehensibility, interestingness and coverage with the respect of classification [4].

Predictive accuracy: The basic idea is to predict the value that some attributes will take in future based on previously observed data. We want the discovered knowledge to have a high predictive accuracy.

Comprehensibility: The discovered knowledge should be comprehensible for the user. This is necessary whenever discovered knowledge is to be used for supporting a decision to be made by a human being. Knowledge comprehensibility can be achieved by using high-level knowledge representations [5].

Interestingness: Interestingness measures play an important role in data mining, regardless of the pattern being mined. These measures are intended are intended for selecting patterns according to the interest to the user [6].

Coverage: The quality of a classification rule is evaluated using coverage [7].

## A. Data Mining

Data Mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown [8]. The major reason that data mining has attracted a great deal of attention in information industry is due to the wide availability of huge amount of data and need for turning the data into useful information and knowledge. The common algorithm in current data mining practice includes classification.

Classification: Classification is the process of finding a set of functions which describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the classes and objects whose class label is unknown [9].

GA has been shown to be an effective tool to use in data mining. It can be used to improve results of already designed algorithm as an optimization rule or can be used for searching. The problem that may be solved by genetic algorithms vary from the optimizing a variety of data mining techniques [10]. In this paper we will discuss how GA can be used for classification.

# III. Proposed Approach for Classification

Genetic algorithms are probabilistic search algorithms which are based on natural selection and genetics. They combine the concept of survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm [11]. The GA starts with the generation of initial population that are stored as binary string in the computer. Selection operator improves the quality of the individual and therefore focuses on the search of promising regions in the search space.

## *A. Genetic Representations*

Each individual in the population represents a rule which is in the form of "IF antecedent THEN consequent ". The antecedent part of rule is conjunction of n-1 attributes. Here n is the total number of attributes which is going to be mined. Each condition in antecedent contains attributes and its value. Consequent consist of single attribute which represents class attributes.

Suppose that an individual consist of 17 attribute values, where attributes are namely hair, feathers, eggs, milk, predator, toothed, domestic, backbone, fins, legs, tail, cat size, airborne, aquatic, breathes, venomous and type and their values are in the form of 0 and 1. The string is generated as follows.

| 10010010100100113 |
|---|

Here for encoding the rules we use only those attributes which contains one with them and leave the remaining attributes. Hence here according to the string used attributes are hair, milk, domestic, fins, catsize, breathes and venomous. This string is then compared with the string of the dataset and the attributes which contains one are replaced by the original value of the attribute from the dataset. Thus, the part so formed is the antecedent part or IF part and the last attribute represent the class attribute. In this example the string belongs to the third class and hence, the interpreted rule is
(if hair=1^milk=1^fins=0^domestic=1^catsize=0^brathes=1 ^venomous=0) THEN (Type=3)

Here in the rule we are using both logical AND and logical OR operator which can be easily extended to represent the rule antecedent having several conditions.

## *B. Fitness Function*

The discovered rules should have predictive accuracy, Comprehensibility, Coverage and Interestingness. By using these four objective functions we calculate the fitness value of a string.

1. Comprehensibility metric: A standard way to calculate comprehensibility is to count the number of rules and number of conditions in these rules.
   If there is rule having at most M conditions, the Comprehensibility of a rule C(R) is given by:

   $$C(R) = 1-\frac{M}{R}$$

   Where, M is the at most conditions and
   R is the total number of attributes excluding class and Name.

2. Predictive accuracy: as described our rules are in the form of IF A THEN C. It can be calculated as

   $$PredicAcc = n\ \frac{(|A\&C|)}{|A|}$$

   Where, |A & C| is defined as the number of records satisfying both A and C.

3. Coverage: Coverage can be calculated as

   $$Cov = \frac{|A|}{\text{No. of Records}}$$

Where, |A| is the number of antecedent.

4. Interestingness: The interestingness is calculated using two terms, one for the antecedent of the rule and other for the consequent. The degree of interestingness of the rule antecedent is calculated by information theoretical measure [12].

$$RInt = 1 - \frac{\sum_{i=1}^{n-1} InfoGain(Ai)}{n-1}$$

$$\log_2(|dom(G)|)$$

InfoGain (Ai) =Info (G) – Info (G|Ai)

Info (G) = $- \sum_{i=1}^{mk}(P(g1)\log_2(P(g1)))$

Info(G/Ai)=∑((p(v) (∑(-p(g/v)log(g/v)))
from i=1 to n

Where, m=number of possible value of goal attributes G.
n =number of possible values of the attributes Ai and log is the base of 2.

The overall fitness is computed as

$$F(x) = \frac{w1.C(R) + w2.PredAcc + w3.RInt + w4.Cov}{w1+w2+w3+w4}$$

Where w1, w2, w3,w4 are the user defined weights.

## C. *Genetic Operators*

Various Genetic Operators are applied in order to perform the generations.

Crossover: It is a genetic operator that combines (mates) two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. Crossover occurs during evolution according to a user-definable crossover probability [13].

Here we are using one single point crossover with tournament selection.

A crossover operator that randomly selects a crossover point within a chromosome then interchanges the two parent chromosomes at this point to produce two new offspring.

Consider the following 2 strings which have been selected for crossover. The "|" symbol indicates the randomly chosen crossover point.
Parent1: 11001|010100011004
Parent2: 00100|111000011002

After interchanging the parent chromosomes at the crossover point, the following offspring are produced:

Offspring1: 11001|111000011002
Offspring2: 00100|010100011004
The crossover is performed over 80 percent of the population.

Mutation: The mutation operator randomly transforms the value of an attribute into another value belonging to the same domain of the attribute.

Here in our approach we exchange the bits of the string
Consider the two string

Parent1: 10110110011010114
Parent2: 11010010101011107
Now by applying mutation the 16 bits of the string are get replaced and the class attribute remain the same. That is on the place of one zero is there. Hence after mutation the string will be:

Offspring1: 01001001100101004
Offspring2: 00101101010100017
The mutation is performed over 10 percent of the population.

Reproduction: Reproduction is performed on the strings which are having the high fitness value and is as it is passed on to the next generation. Reproduction is performed over the 10 percent population.

# IV. Algorithm

1. Initialize the population, g=1.
2. Evaluate the objective function (Predictive accuracy, Coverage, Interestingness and Comprehensibility)
3. Calculate the fitness value of each individual.
4. While (g <=specified no of generation)
5. Apply single point crossover.
6. Apply mutation.
7. Apply reproduction
8. Evaluate the objective function.
9. Assign the fitness.
10. g=g+1;
11. End While
11. Decode the strings as an IF-THEN rule

# V. Experimental Results

The experiment to test the proposed approach is performed on Zoo dataset.

## A. Description of Dataset

This dataset contains 101 instances and 18 attributes. Each instance in the dataset corresponds to the animal. While processing the names of the attributes are removed. There are about seven classes in which all the instances are differentiated. The last attribute is class and the other attributes are namely hair, feathers, eggs, milk, predator, toothed, domestic, backbone, fins, legs, tail, and catsize, airborne, aquatic, breathes, venomous, type. The type attribute shows the class attribute.

There are 15 Boolean attribute and 2 numeric attribute and name.

**Classes:**

Class 1:Mamalia
Class 2:Aves
Class 3:Reptilia
Class 4:Actinopterygii
Class 5:Amphibia
Class 6:Insecta
Class 7:Bivalvia

## B. Results

Experiments have been performed using JAVA 6.0 on windows. For the following dataset the simple genetic algorithm had 200 individuals in the population and was run for 60 generations. The data set is divided into two parts training set and test set. On the training set the rules are discovered and are taken 60 percent of the data set. The testing is performed on the remaining 40 percent test set. We represent the predicted class to all individuals of the population which is never modified during the running of the algorithm. Hence, we get the corresponding rules for each class. Table I shows the mined rules of each class with its accuracy.

Table I
Discovered Rules

| Class | Mined rules | Accuracy |
|---|---|---|
| 1 | If (hair=1) ^ (eggs=0) ^ (domestic=0) ^(Venomous=1) Then (Type=1) | 0.8856 |
| 2 | If (hair=1) ^ (legs=2) ^ (domestic=0) ^ (Venomous=1) Then (Type=2) | 0.8475 |
| 3 | If (eggs=1) ^ (tail=1) ^ (domestic=0) ^ (backbone=0)^ (fins=0)Then (Type=3) | 0.7655 |
| 4 | If (toothed=0)^ (eggs=1) ^ (aquatic=1) ^(tail=1) Then (Type=41) | 0.8345 |
| 5 | If (airbone=0)^(catsize=1) ^ (breathe=0)^(toothed=1)^(legs=4) Then (Type=5) | 0.7965 |
| 6 | If (hair=0) ^ (legs=6) ^ (eggs=0) Then (Type=6) | 0.8355 |
| 7 | If (predator=1)^(eggs=0)^ (toothed=0) ^ (domestic=1) Then (Type=5) | 0.7125 |

## *VI Conclusion*

We have presented a Genetic Algorithm based approach for performing classification in data mining. We have used four objective functions for performing the rule based classification, and hence much more accurate fitness values are obtained. We have used Zoo dataset for the training and testing of the classifiers rules obtained for each class. Our approach is giving satisfactory results.

## *References*


[1] Lau Tung Leng, "Guided Genetic Algorithm", pp. 6-8 University of Essex ,United Kingdom.

[2] S.Rajsekaran, "Neural Networks, Fuzzy logic and Genetic Algorithms", Prentice Hall of India Publications, 2007.

[3] Vladan Devedzic, "Knowledge Discovery and Data Mining in Databases", Fon-School of Business Administration , University of Belgrade, Yugoslavia.

[4] Usama Fayyad, Geogry Piatetsky-Shapiro, Padhraic Smyth ,"From data mining to knowledge discovery from databases", pp. 37-40, AAAI97, Providence, Rhode Island July 27-31, 1997.

[5] S.Dehuri, R Mall, "Mining Predictive and Comprehensible Classification Rules using Multi Objective Genetic Algorithm", In Proceeding of ADCOM, pp. 99-104, India, 2004.

[6] Liqiang Geng and Howard J. Hamilton, "Interestingness measures for Data Mining: A Survey", volume 38 Issue in 3,2006, University of Regina, Saskatchewan, Canada.

[7] Vipin Kumar and Pang-Ning Tan, "Introduction to Data Mining",Michigan State University pp 208-214 India.

[8] Arabinda Nanda and Saroj Kumar,"Data Mining and Knowledge Discovery in Databases: An AI Perspective", Proceedings of national seminar on data mining, Bhubneshwar ,Orrisa.

[9] Jiawei Han and Micheline Kamber, "Data Mining: concepts and Techniques", pp. 10-24, 20, Morgan Kaufmann Publishers, san Francisco, 2000.

[10] Behrouz Minaci-Bidgoli, William F. Punch, ”Using Gentic Algorithm for Data Mining Optimization in an Educational–Web based System”, GARAGe, Michigan State University East Lansing.

[11] S.Bandyopadhyay, ”Pattern Classification using Genetic Algorithm” pp. 1171–1181, Volume 19 , Issue 13, India, 1998.

[12] A.A Freitas(1998),” On Objective Measures of Rule Surprisingness”, Proc. Of 2nd European Sympisiom on Principle of Data Mining and Knowledge Discovery(PKDD-98). Lecturer Notes in Artificial Intellegence,1510.pp1-9

[13] Gilbert Syswerda, ”Uniform Crossover in genetic algorithms”, Morgan Kaufmann publishers Inc. san Francisco, CA,USA, pp. 2-9.