Heart Disease Classification By Genetic Algorithm

Zainab Falah Hasan **Asraa Abdalluh Hussein**
University of Babylon, science collage for women / Computer science
zainab_ga@yahoo.com esraa_zd@yahoo.com

Abstract

Classification is predicting of a correct output for related inputs. This is done by using many techniques, some of them depends on a training process that uses set of features and its related output. In the training process, the algorithm finds the relationships between the features and its related output. In this work, classification heart disease is done using genetic algorithms. Genetic algorithm is used to get best successful system of classification. The chromosome consists of the states using for classification, this means every gene acts state from database. Database is clustered to chromosome to find fitness of chromosome(accuracy of classification).

Key words : genetic algorithms, clustering, patterns recognition.

الخلاصة

التصنيف هو التنبؤ بالاخراج الصحيح للمدخلات المرتبطة.هذا يتم باستخدام عدة تقنيات , بعض منها يعتمد على عملية التدريب التي تستخدم مجموعة من الخصائص و الاخراج المرتبط بها. في عملية التدريب, الخوارزمية تجد العلاقة بين الخصائص والاخراج المرتبط بها. في هذا العمل, تصنيف مرض القلب يتم باستخدام الخوارزمية الجينية. الخوارزمية الجينية تستخدم للحصول على افضل نظام ناجح للتصنيف. الكروموسوم يتكون من الحالات المستخدمة للتصنيف, هذا يعني انه كل جين يمثل حالة من قاعدة البيانات. قاعدة البيانات يتم عنقدتها الى الكروموسوم لايجاد صلاحية الكروموسوم(دقة التصنيف).

الكلمات المفتاحية: الخوارزميات الجينية, العنقدة , تمييز الانماط.

1- Introduction

The classification of data is based on the set of data features used. Therefore, feature selection and extraction are main in optimizing performance, and strongly affect classifier design. Defining suitable features often requires interaction with experts in the application area. Genetic algorithms are good candidates for this task since GAs are most useful in multiclass.

Genetic algorithm using to mine a classification rules in large datasets proposed by Vivekanandan and Nedunchezhi(2010). Building a rule based classification model for these huge data sets using Genetic Algorithm becomes an extremely complex process. They build a model incrementally. An incremental Genetic Algorithm was evolved small components by evolution of the data set which reduce the cost of learning and making it suitable for large data set to build the rule based classification model in a fine granular method .

Parallel Genetic Algorithm depended on Clustering have been proposed by Kanungo *et al.*,(2007) was used to classify the background and objects. They used a method uses the histogram of the original image where Parallel Genetic Algorithm depended on clustering notion was used the discrete nature of the histogram distribution to determine the optimal threshold.

Fukunaga (1990) divided the space into the parts of classes and defined a problem of evaluating density functions in a high-dimensional space as pattern recognition.

Robert *et al.,* (2002) described Pattern recognition for metal defect detection, they described how to classify a data set including features extracted from metal strips using pattern recognition algorithms.

Mehdi.Neshat and Ali.Adeli (2010) used Mamdani inference expert system for diagnosis of heart disease. Their system consists of one output to indicate the state of patient and 13 inputs.

Hitoshi *et al.,* (2001) used to classify Tissue based on selecting Informative Genes with Parallel Genetic Algorithms, they filtered out the informative genes related to classification.

2- Genetic Algorithms

Genetic algorithms are premised on natural selection and genetic for the Darwin's evolutionary ideas and they are search about solutions in heuristic direction (Pratibha and Manoj, 2010). GA has three main applications, namely, intelligent search, optimization and robot learning.

In every generation, many of solutions generated by genetic algorithm represent individuals of population. Initial population created by randomly generating of many solutions. The number of generated solutions called population size (Pratibha and Manoj, 2010). A suitable encoding is used for the solution to problem. Many encoding techniques used to provide representation of genetic algorithms. Some of the encoding methods:

1. Binary encoding: in this method, the value of genes of chromosome is either 0 or 1.
2. Real-number encoding: in this method, the value of genes of chromosome is real number.
3. Integer encoding: in this method, the value of genes of chromosome is integer number.

The fitness of each individual is computed to measure its efficiency.

Selection operation is a main operation in genetic algorithm used to choose the best individuals in the population to obtain the new individuals. New individuals will participate to create the next generation of population. The next generation of population is created with a hope to reach the optimal solution (Rakesh and Jyotishree, 2012). The most common types of selection methods are :

1. Roulette wheel selection: in this selection, selection of individuals for next generation based on the fitness level (Noraini Mohd Razali, 2011).
2. Rank selection: in ranking selection, every individual in the population has its rank by sorting the population using fitness from best to worst and selection is based on rank (Noraini Mohd Razali, 2011).
3. Tournament selection: this method select the best individual among a random set of individuals. Tournament size represents the number of individuals in the set (Noraini Mohd Razali, 2011).

Two types of transformations used to create new population of solutions are: (Muhammad and Abido, 2009):-

1. Crossover, which forms new individuals by combining parts from two individuals. The crossover points of any two chromosomes are selected randomly (Nitasha Soni and Dr .Tapas Kumar , 2014). There are many types of crossover such as:

Single Point Crossover: in this method, two children created by exchange two parents their parts after crossover point that is randomly selected (Nitasha Soni and Dr .Tapas Kumar , 2014).

Two point crossover: two points crossover randomly selected from a chromosome to obtain two new children by exchange the parts of parents among selected crossover points (Nitasha Soni and Dr .Tapas Kumar , 2014).

2. Mutation, which forms new individuals by making changes in a single individual. Mutation is help to overcome on local optimum. There are many types of mutation such as:

Single Point Mutation: single gene is randomly selected to be mutated and its value is changed depending on the encoding type used (Young. and Ying, 2004).

Multi Point Mutation: multi genes are randomly selected to be mutated and their values are changed depending on the encoding type used (Young and Ying, 2004).

The fitness is computed for the new individuals to evaluation them. A next generation of offspring is created by selecting the more fit solutions from the parent population and applying the transformation operations (Muhammad and Abido 2009).

Genetic algorithm reaches to the optimal or suboptimal solution after many generations. The flowchart of a GA work is presented in the fig.1 (Amit, 2000) .

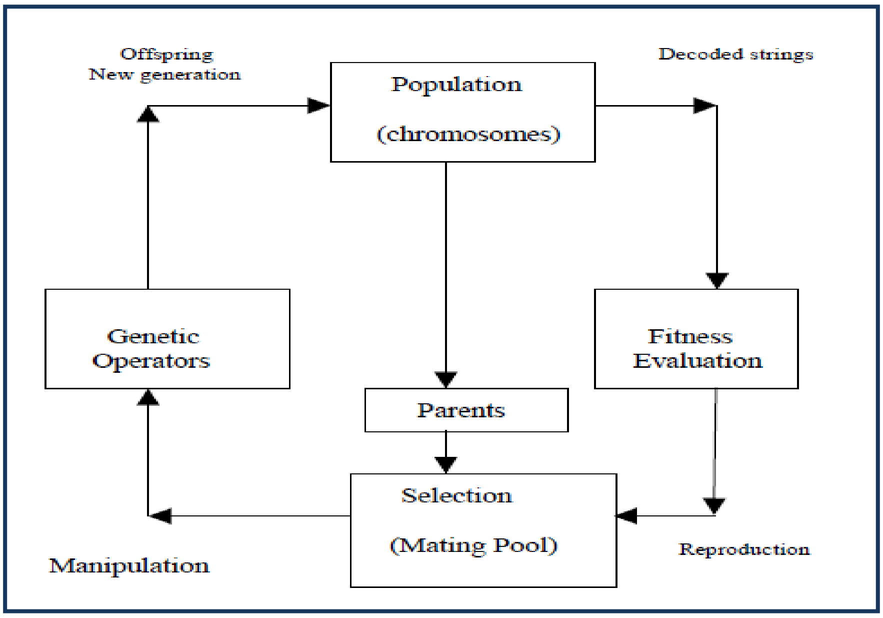


Fig. 1: The Block diagram of Genetic Algorithms Work.

3- Genetic Algorithm to Classification

The Classification of heart disease in which every state consists from 14 features where the class feature represents if person is sick or no using G.A explained by the following steps:

1-Choose random input data: initial population.

2-Use input data for classification .

3-Measure accuracy of classification for input data: evaluation of input data.

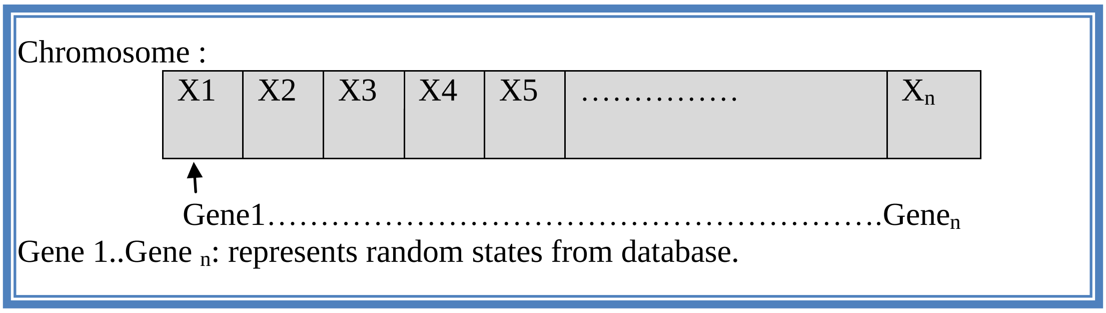
4-Explore other input data for better classification: generation new solutions.

5-Measure accuracy of classification for new solutions.

6-Loop the steps 4-5 until the best classification has been gotten.

3-1 Choose Random Input Data

Given input data from heart disease database, random number of these data is chosen. These data used with their features for classification. Random input data act as chromosomes and the gene represents one state from database. Initial population is generated in this step. The chromosome is shown as follows:



Gene 1..Gene $_n$: represents random states from database.

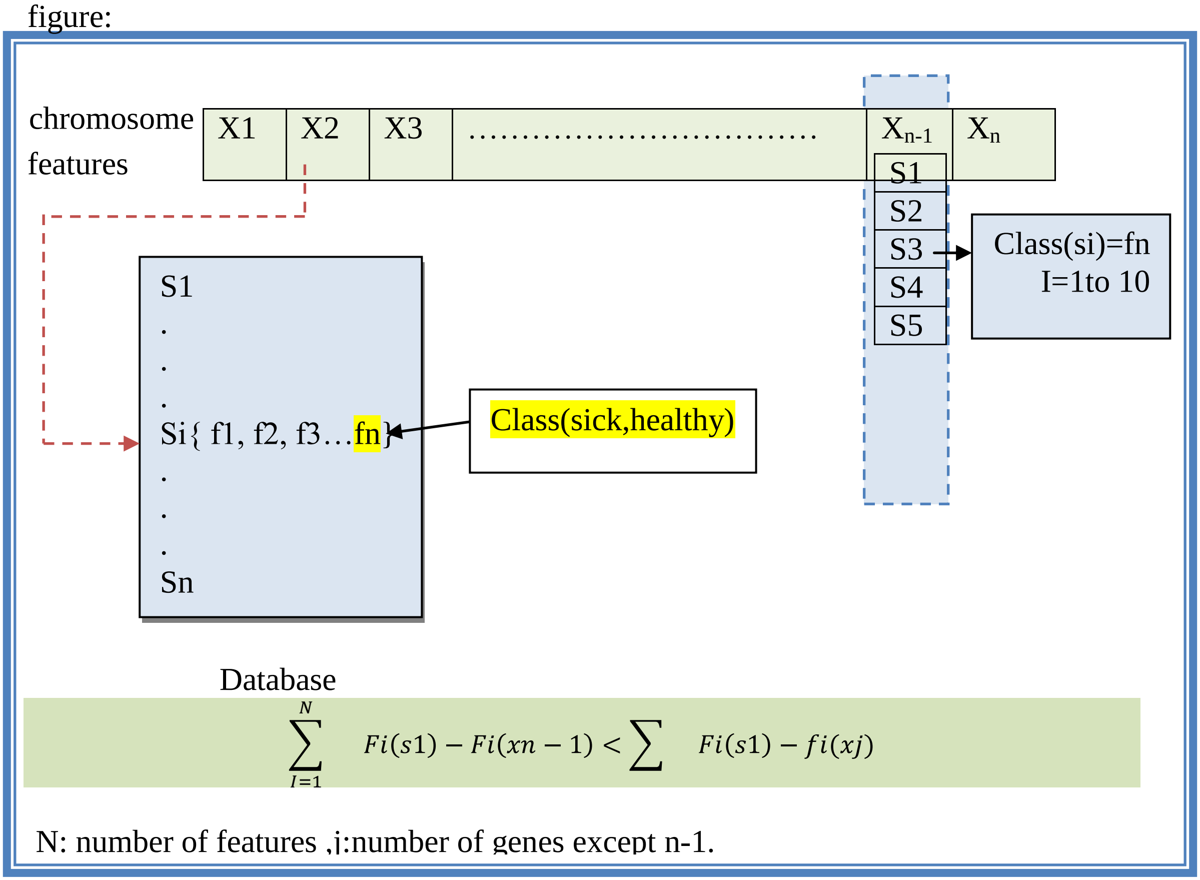
Fig. 2: Chromosome Representation.

Every gene represents state from database which has features used to classification.

3-2 Use Input Data for Classification

In this step, the initial population of chromosomes (input data) is used to classification the database to evaluate the chromosomes and compute fitness of each individual. Every gene represents state and has many features , the last feature acts the state of patient if he is sick or no. classification is done by collecting the similar states from database to state (gene) and the collected data take the class of state.

Similarity is done by computing distance between features of database for every input and features of state (gene). Classification is shown by the following figure:



$$\sum_{I=1}^{N} Fi(s1) - Fi(xn-1) < \sum Fi(s1) - fi(xj)$$

N: number of features ,j:number of genes except n-1.

Fig. 3: Classification system of Database of Heart Disease using Chromosome.

3-3 Measure Accuracy of Classification for Input Data

After classification of database using chromosomes , measuring the accuracy of classification of database to compute fitness of chromosomes is done.

This done as follows:

If class (si)= class(si)in data base then

Number of states have correct class =number of states have correct class+1

Accuracy of classification=(N /M)*100%.

N: number of states have correct classification.

M: number of all training states.

Fitness of chromosome= Accuracy of classification.

Class (s_i) acts feature no. 14 in database.

3-4 Explore Other Input Data for Better Classification

Crossover and mutation are applied on the population to generate new better chromosomes for classification . One-x crossover used for generating new chromosomes and 2m mutation is applied on new chromosomes. This done as follows.

1-Selection two individuals for crossover. Tournament is used to select random individuals to apply combination operations and get new solutions.

2-Apply 1x crossover on the selected chromosomes.

3-Apply 2m mutation on the children.

3-5 Measure Accuracy of Classification for New Solutions

This step similar in 3-3 , the fitness is computed to the new population . After the best classification is gotten, generation of new population stop. The best chromosome used to classification of database.

4 Results

After generation many population(solutions) to classification database of heart disease, the best classification is used with different numbers of data.

The results is gotten when training some of data, containing 14 features and the last feature acts class of state, by genetic algorithm are shown in the following table:

Table 1: Accuracy of Classification and used States.

No. of training data for classification	Accuracy of classification
66%	81%
50%	80%

5 Conclusions

Genetic algorithm resulting efficient search about the best classification for database based on features of states to access into perfect states used to build classification system and it is improved the solution by applying mutation on it.

Genetic algorithm efficiently classified database and succeeded to find some states able to classify all the database. GA also overcomes the problems of other techniques such as k.means to find optimal centers for classification.

6 References

Ali. Adeli and Mehdi.Neshat, 2010, A Fuzzy Expert System for Heart Disease Diagnosis, Proc. international multiconference of engineering and computer scientists (pp. 134–139). Vol. 2

Amit K. , 2000, Artificial Intelligence and Soft Computing, CRC Press LLC, Vol. 2.

Fukunaga. K. 1990, Introduction to statistical pattern recognition (2nd ed). Academic Press, Boston.

Juan Liu, Hitoshi Iba and Mitsuru Ishizuka, 2001, Selecting Informative Genes with Parallel Genetic Algorithms in Tissue Classification, Genome Informatics 12: 14–23.

Kanungo, P. ; P. K. Nanda and A. Ghosh, 2007, Classification of Objects and Background Using Parallel Genetic Algorithm Based Clustering, Electronic Letters on Computer Vision and Image Analysis 6(3):42-53.

Muhammad T. and M. A. Abido , 2009, Assessment of Genetic Algorithm Selection, Crossover and Mutation Techniques in Reactive Power Optimization, IEEE.

Nitasha Soni and Dr .Tapas Kumar , 2014**,** Study of Various Crossover Operators in Genetic Algorithms, International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 7235-7238.

Noraini Mohd Razaliand and John Geraghty, 2011, Genetic Algorithm Performance with Different Selection Strategies in Solving TSP, International Conference of Computational Intelligence and Intelligent Systems, Vol. 2, ISSN: 2078-0958.

Pratibha B. and Manoj K. ,2010, Genetic Algorithm – an Approach to Solve Global Optimization Problems, Indian Journal of Computer Science and Engineering, Vol. 1 No. 3.

Rakesh K. and Jyotishree, 2012, Effect Of Polygamy With Selection In Genetic Algorithms, International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, ISSN: 2231-2307.

Robert P.W. Duin, P.P. Jonker and D. de Ridder, 2003 , Structural, Syntactic, and Statistical and Pattern Recogition Steel Grips, vol. 1, no. 1, 20-23..

Vivekanandan P.; and R. Nedunchezhi, 2010, A Fast Genetic Algorithm For Mining Classificatin, International Journal on Soft Computing (IJSC), Vol.1, No.1.

Young-C. and Ying-H. C. , 2004 , A New Efficient Encoding Mode of Genetic Algorithms for the Generalized Plant Allocation Problem, Journal of Information Science and Engineering, Vol. 20.