ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, Vol 8(9), 797–803, May 2015
DOI: 10.17485/ijst/2015/v8i9/52930

Heart Disease Prediction Using Hybrid Genetic Fuzzy Model

T. Santhanam[1] and E. P. Ephzibah[2*]

[1]D.G. Vaishnav College, Arumbakkam, Chennai, Tamilnadu, India
[2]School of Information Technology and Engineering, VIT University, Vellore, Tamilnadu, India; ep.ephzibah@vit.ac.in

Abstract

The objective of the work is to diagnose heart disease using computing techniques like genetic algorithm and fuzzy logic. The system would help the doctors to automate heart disease diagnosis and to enhance the medical care. In this paper a hybrid genetic-fuzzy heart disease diagnosis system is designed. The genetic algorithm is used for a stochastic search that provides the optimal solution to the feature selection problem. The relevant features selected from the dataset help the diagnosing system to develop a classification model using fuzzy inference system. The rules for the fuzzy system are generated from the sample data. Among the entire rule set the important and relevant subset of rules are selected using genetic algorithm. The proposed work uses the benefits of genetic algorithms and fuzzy inference system for effective prediction of heart disease in patients. The selected features are sex, serum cholesterol (chol), maximum heart rate achieved (thalach), Exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), number of major vessels coloured (ca) and thal value. Fuzzification using Fuzzy Gaussian membership function and defuzzification using centroid method improves the performance of the system. The work has been evaluated using the performance metrics like accuracy, specificity, sensitivity, confusion matrix that help in proving the efficiency of the work. The obtained classification accuracy is 86% using the stratified k fold technique with the values for specificity and sensitivity as .90 and .80 respectively. The number of attributes has been reduced from 13 to 7 from heart disease dataset available in the UCI Machine learning repository. When compared with the existing system the accuracy of the proposed work has been increased by 1.54%. The proposed model is named as GAFL model called Genetic Algorithm Fuzzy Logic model for effective heart disease prediction. It is easy to build the model thereby providing an easy option to be used in hospitals and medical centers for the aid of the physicians.

Keywords: Feature Selection, Fuzzy Logic, Gaussian Membership Function, Genetic Algorithm, Heart Disease Prediction

1. Introduction

Medical mining involves computerized tools and techniques that help in providing the benefits to health systems. Especially artificial intelligence techniques are most commonly used for disease diagnosis[1–3]. The neural network classifier helps in diagnosing the diseases by developing a model using feed forward neural network, multi-layer perceptron neural network, and back propagation neural network. Genetic algorithms help in medical mining using their stochastic searching technique, the fitness function along with a set of genetic operators. The fuzzy logic is a tool for providing solution to the problems that deal with fuzzy input data[4]. The proposed work takes into account the Genetic Algorithm (GA) for feature selection, and fuzzy logic for classification. The dataset chosen is the heart disease dataset that contains records of patients with and without heart disease. The objective of this work is to design a model that can help in predicting whether an incoming patient has heart disease or not. GA is one of the most effective feature selection methods[5]. It is a stochastic searching technique that helps in producing optimal solution for optimization problem.

*Author for correspondence

GA is used for feature selection and thus reducing the number of attributes in the dataset which in-turn narrows down the search. A fitness function evaluates the fitness of a chromosome in a population. A chromosome is the characteristic of the data. The design of fitness function is fundamental for the genetic algorithm, as it decides the termination criterion[6]. In the proposed work the attributes in the dataset are selected using GA and the fuzzy inference system further performs classification and prediction.

The Experimental data had been a backbone in many research works. In[7] the authors have used a learning method that deals with uncertain data. The training dataset helps in constructing the fuzzy decision tree. The attributes are ranked based on the measure of discrimination like entropy and ambiguity. They have devised an algorithm for incremental dynamic development of the decision tree fuzzy classification. The authors of the paper[8] have used the three types of decision trees like non ordered, ordered and stable decision trees for medical decision making system for breast cancer diagnosis. Out of the three types, the method using non ordered decision tree performed better with a minimum error of 0.1040. Mohamed et al., in their paper[9] have proposed an algorithm to generate fuzzy rough decision tree using fuzzy logic and rough sets. They have taken into consideration the medical datasets like Wisconsin breast cancer and pima Indian diabetes and obtained an accuracy of 96.1 % and 86.4 % respectively. Nael and Robert[10] have provided a software tool for the fuzzy decision tree generation using the fuzzy ID3 algorithm as its base. The experimental results have proved that the rule-set reduction method can improve the accuracy and reduce the number of rules required for classification. The measures like information gain, classification ambiguity value and the gini index have been used for the RFDT (Reduced set Fuzzy Decision Tree) construction and produced the experimental results as 99.29% and 82.22% accuracy for breast cancer and heart disease datasets respectively using information gain, 97.89% for breast cancer and 82.96% for heart dataset using classification ambiguity, 99.29% and 82.22% for breast cancer and heart disease dataset respectively using Gini index. It is to be noted that the accuracy for the heart disease diagnosis using this method is lesser when compared to our proposed approach. In their paper on “A threshold fuzzy entropy based feature selection for medical database classification” the authors have used a fuzzy entropy measure to identify the feature relevance and have reduced the number of features from 13 to 3 features 11. The reduced features have been tested with a Radial Basis Function (RBF) network classifier and the accuracy was found to be 84.46% for the heart disease dataset. The proposed approach reduces the number of attributes from 13 to 7 and improves the accuracy by 1.54% when compared to the previous approach.

1.1 Data Set

The heart disease dataset is taken from the UCI machine learning repository. This data set has been used by many researchers especially for classification using fuzzy logic[12,13]. Manually removing 6 records that contain missing vales, the dataset resulted in 297 samples. There are[14] attributes including the class label. The information contained in the dataset is effective and helps in identifying the hidden pattern. Not all the attributes are effective but only a few are relevant for classification and prediction of the disease. Identifying the important attribute set is an important task that helps in data cleaning, eliminating irrelevant attributes, removing noise form the data etc. The attributes in the dataset are as follows (Table 1): age of the patient, gender, the chest pain type, the resting blood pressure in mmHg, serum cholesterol in mg/dl, the fasting blood sugar in mg/dl, resting electro cardio graphic results, thalach (maximum heart rate achieved), exercise induced angina, oldpeak, slope, number of major vessels colored by fluoroscopy and finally thal with values as normal, fixed defect and reversible defect. The experts can easily identify the important attributes thereby reducing the number of features. On the other hand the computing techniques can also be used for selecting the important attributes. In this work we have identified the important features using the genetic algorithm a stochastic search technique.

2. Methods

2.1 Genetic Algorithms

GA[14] helps in solving many real time problems using the process of evolution of species. The input to the algorithm is called as chromosome that contains the parameters that have unique characteristics. Each chromosome consists of a collection of genes. A gene expresses the characteristic of the input. A collection of such chromosomes form a population. A chromosome in the population provides the solution to the problem after a series of iterations

Table 1. The attributes and their numeric representations

Attribute No	Attribute names
1	age
2	sex
3	cp
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal

called generations. All genetic information gets stored in the chromosomes[15]. Every generation is better than its previous one as the possibility of obtaining the solution gets better. The solution is obtained with the inclusion of fitness function, the genetic operators like selection, crossover and mutation in every generation.

2.1.1 Chromosome Representation

Designing the fitness function is the fundamental part of the GA that helps to get the optimal solution[16]. The fitness functions available in the literature are most commonly problem dependent. Functions that work well for some datasets may not produce better results for other data sets. The fitness function value is the measurement that helps to check the nearness of the optimal solution. The three genetic operators are briefly outlined below:

Selection is the process of selecting the parents among the population so that they can be used for crossover and mutation operations. Selection method represents the mechanism that determines the number and the type of parent chromosome to be selected. The selection method for this work is the roulette wheel selection. It is one of the traditional selection techniques. The chromosome is selected based on the probability proportional to its fitness value. Fitter chromosomes have a better chance to be selected in this approach.

Crossover is a recombination operator that selects the parents from the pool of population and interchanges the position of the values based on the crossover point fixed. The values before the fixed point from one chromosome is transferred to the first part of the new chromosome and the values that are in the second chromosome are transferred to the second section of the new one, thus inheriting the features of both the parents. Among the various crossover techniques the intermediate crossover has been chosen in this work. This technique is applicable for real valued chromosomes as they produce the offspring using the following formula[17.]

$$\text{Offspring} = \text{parent 1} + \text{Alpha (parent2-parent1)} \quad \text{------(1)}$$

Alpha is a scaling factor chosen uniformly at random.

Mutation is a process of flipping or changing the gene values based on its given probability (mutation probability) value in binary or real valued representation respectively. Mutation operator always accelerates and explores the search space. It helps to escape from local minima and an appropriate value for this operation can lead to the optimal solution as it maintains the diversity in the population. Mutation probability decides the frequency of performing mutation in every generation. This study uses the Gaussian mutation with the values for scale and shrinking to be 0.05 and 1 respectively. This method adds a random value to the selected chromosome taken from the Gaussian distribution centered on zero. The scale value determines the standard deviation at the initial generation and the shrink parameter is a controlling parameter that controls the standard deviation value throughout the generations. The stochastic search of the genetic algorithm stops based on the convergence criteria. The various stopping conditions are: when the process reaches the maximum number of generations, the maximum time limit fixed by the user (elapsed time), a state when it finds no improvement if the fitness values of the individuals for a pre-mentioned number of generations, a state when there is no significant improvement in the objective function value (called stall generations and stall time limit). Genetic algorithms combine the high performance notions to achieve better performance for getting optimal solution.

2.2 Fuzzy Logic

The primary task of our work is to perform classification using fuzzy logic. In 1965 Lotfi A. Zadeh proposed a fuzzy set theory that is more applicable to artificial intelligence, especially, for the problems that have uncertain input values[18]. Fuzzy logic is a form of uncertain or many – valued logic. This logic provides approximate solutions rather than accurate as it handles the concept of partial truth where the truth value can be in the range between completely true and completely false.

2.2.1 Membership Functions

Fuzzy membership functions are devised based on the problem to be solved and the fuzzy set chosen for the same. Membership function represents the fuzzy set and also provides a measure of the degree of similarity of an entity to a fuzzy set. Most common shapes for design of membership functions are triangular, trapezoidal, linear, Gaussian, bell-shaped etc. In the proposed work the Gaussian membership function is chosen as it is comprehensible and appropriate to the problem.

A membership function for a fuzzy set A on the universe of discourse X is defined as μ_A: X → [0, 1], where each element of X is mapped to a value between 0 and 1. This value, called membership value or degree of membership, quantifies the grade of membership of the element in X to the fuzzy set A. Membership functions allow us to graphically represent a fuzzy set. The x axis represents the universe of discourse, whereas the y axis represents the degrees of membership in the [0, 1] interval. The Figure 1 depicts the Gaussian membership function used in the proposed work.

$$\mu_A(x, c, s, m) = \exp\left[-\frac{1}{2}\left|\frac{x-c}{s}\right|^m\right]$$

– c: centre

– s: width

– m: fuzzification factor (e.g., m=2)

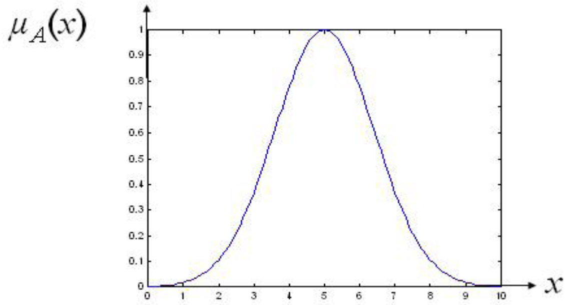

Figure 1. Gaussian membership function.

2.2.2 Fuzzy Inference System

A fuzzy inference system helps in mapping the inputs to the corresponding output using predefined fuzzy rules available in the knowledge base. The knowledge base consists of if-then rules that specify the relationship between the input and output fuzzy sets[19]. As it requires the input in fuzzy values, the input is fuzzified and for the user to better understand the output, the output from the inference system is defuzzified. The inference system is developed with a series of activities like[20],

1. Developing the fuzzy rules,
2. Fuzzifying the input values based on the membership function,
3. Combining the fuzzy input and the fuzzy rules to generate the rule strength.
4. The rule strength consequence is again combined with the output membership function to generate the output distribution and
5. Finally the output is defuzzified to give the output in crisp value.

2.2.3 Fuzzy Rule Generation

The fuzzy rule generation is a vital task that helps in mapping the input to its corresponding output. Rules can be framed using any method that provides an antecedent and consequent form as given below:

If antecedent then consequent;

If a_1, a_2...a_n are the attributes and c_1, c_2...c_m are the class labels then a fuzzy rule can be framed based on the linguistic values like high, medium, low. The values n and m are the number of attributes and number of classes respectively. These linguistic values are also called as the labels that can vary in number based on the type of problem to be solved. Therefore the fuzzy rule can be framed as follows:

If a_1 is high and a_2 is low and a_3 is medium then class is c_2;

The data has been normalized and split into three distinct regions namely ‘1’, ‘2’ and ‘3’ for the attributes that have a wide range of values. The value 1 represents ‘low’, 2, ‘medium’ and 3, ‘high’. For attributes like age, trestbps, cholesterol, thalach, oldpeak the following (Figure 2) implies the Gaussian fuzzy membership that is used:

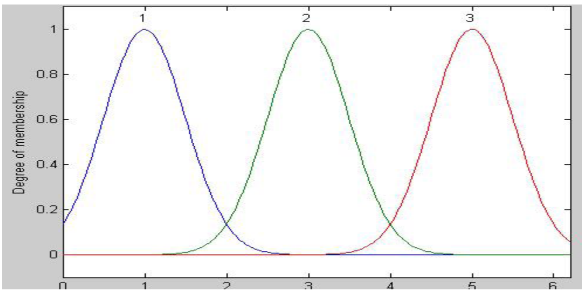

Figure 2. Representation of the membership functions.

For the other attributes the membership values are based on their actual values taken from the dataset.

2.2.4 Fuzzy Classifier

Fuzzy classification is a supervised learning method where the fuzzy model understands the data with its rules and class label of the training data and predicts the target value for the set of test data. The proposed work uses the stratified 10 fold technique which is a popular choice for estimating the test error on classification algorithms. It randomly divides the training set into 10 disjoint subsets. Each subset has roughly equal size and also the same class proportions as in the training set. The steps involved in the process are to identify a subset for testing with all the other subsets as training subsets. Using the trained model the testing subset is classified. This process continues for ten times with different training and testing data. Finally the average accuracy if calculated.

The proposed work contains different labels based on the values of the attributes available in the dataset. The attributes and their corresponding labels are as given in Table 1. The actual labels are provided based on the data in the dataset whereas the modified labels are framed based on the fuzzy classifier. The Table 2 provides the details about the attributes and their modified labels. The fuzzy rules are also based on these labels. Some of the sample rules are given below:

Rule 1:if a_1 is 3 and a_2 is 1and a_3 is 1 and a_4 is 2 and a_5 is 2 and a_6 is 2 and a_7 is 2 and a_8 is 2 and a_9 is 2 and a_{10} is 2 and a_{11} is 3 and a_{12} is 4 and a_{13} is 2 then result is "healthy"

Rule 2:if a_1 is 3 and a_2 is 1and a_3 is 4 and a_4 is 2 and a_5 is 3 and a_6 is 2 and a_7 is 2 and a_8 is 1 and a_9 is 1 and a_{10} is 1 and a_{11} is 2 and a_{12} is 3 and a_{13} is 1 then result is "Sick"

Rule 3:if a_1 is 3 and a_2 is 1and a_3 is 4 and a_4 is 1 and a_5 is 2 and a_6 is 2 and a_7 is 2 and a_8 is 2 and a_9 is 1 and a_{10} is 2 and a_{11} is 2 and a_{12} is 2 and a_{13} is 3 then result is "Sick"

Rule 4:if a_1 is 1 and a_2 is 1and a_3 is 3 and a_4 is 1 and a_5 is 2 and a_6 is 2 and a_7 is 3 and a_8 is 3 and a_9 is 2 and a_{10} is 2 and a_{11} is 3 and a_{12} is 4 and a_{13} is 1 then result is "healthy"

Defuzzification methods are used to convert the fuzzy values to their corresponding crisp and understandable values. There are five defuzzification methods available. They are centroid, bisector, Smallest Of Maximum (SOM), Middle Of Maximum (MOM) and Largest Of Maximum (LOM). The defuzzification method used for the proposed work is centroid method as it is depicted in Figure 3. In this type it returns the center of area under the curve.

Table 2. The modified labels of the attributes for fuzzy classier

Attribute No	Attribute names	Actual labels	Modified Labels
1	age	Integer	1,2,3
2	sex	0,1	1,2
3	cp	1,2,3,4	1,2,3,4
4	trestbps	Integer	1,2,3
5	chol	Integer	1,2,3
6	fbs	0,1	1,2
7	restecg	0,1,2	1,2,3
8	thalach	Interger	1,2,3
9	exang	0,1	1,2
10	oldpeak	Real number	1,2,3
11	slope	1,2,3	1,2,3
12	ca	0,1,2,3	1,2,3,4
13	thal	3,6,7	1,2,3

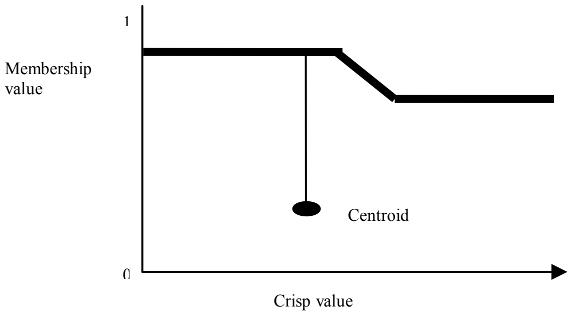

Figure 3. Defuzzification method used in the proposed work-centroid.

3. Experimental Results

The experimental results in Table 3 show that the selected set of attributes and the appropriate training set give a better performance when compared to the existing method in 11. The performance of the proposed model can be measured using some of the metrics like specificity, sensitivity and accuracy.

The Tables (Table 4 and 5) give the specificity, sensitivity and the confusion matrix for the proposed method with heart disease dataset. Specificity is the measure of finding the individuals correctly classified as "healthy" individuals where as sensitivity is a measure that finds individuals correctly classified as "sick"[21].

Table 3. Accuracy comparison of the proposed work with NNTS method

S No	Method	Authors	Features	Accuracy in % (stratified 10 fold technique)
1	Fuzzy entropy based method(NNTS)*	P. Jaganathan, R. Kuppuchamy[11]	3,12,13	84.46
2	Proposed method (GAFL System)**		2 5 8 9 10 12 13	**86**

*NNTS: Neural Network for Threshold Selection
**GAFL: Genetic Algorithm Fuzzy Logic System.

Specificity=TP/(TP+FN) -------------------------------(3)

Sensitivity=TN/(FP+TN) -------------------------------(4)
In the equations (3) and (4) TP means True Positive, FP means False Positive, TN means True Negative and FN means False Negative. Using the proposed method the specificity and sensitivity calculated are given in the Table 4. Finally, Table 5 provides the confusion matrix that explains the prediction accuracy of the proposed work.

Table 4. Specificity and Sensitivity measures

Specificity	.90
Sensitivity	.80

Table 5. Confusion matrix

	Sick	Healthy
Test Positive	38	5
Test Negative	9	48

The accuracy is the overall performance which is calculated using the equation (5).
Accuracy= (TP+TN)/(TP+TN+FP+FN) (5)
The calculated accuracy for the proposed approach is 86%.

4. Conclusion and Future Work

A system that aids the physicians for accurate prediction of heart disease in patients has been devised using the computing techniques like genetic algorithms and fuzzy logic. Amidst various classification and prediction models this model is evaluated to be better providing an accuracy of 86%. The genetic algorithm has been applied to perform a stochastic search on the dataset to reduce the number of features from 13 to 7. The fuzzy inference system predicts the test data with the help of fuzzy Gaussian membership function and centroid defuzzification method. The future work includes the implementation of the proposed model for disease diagnosis using the other health related datasets. As the fuzzy classifier is suitable for uncertain data the proposed work can be extended for any data with uncertainty. The time and space complexities can be taken into consideration for overall performance of the proposed method.

5. References

1. Overiu M, Simon D. Biogeography-based optimization of neuro-fuzzy system parameters for diagnosis of cardiac disease. Proceedings of the 12th annual conference on Genetic and evolutionary computation; 2010. p. 1235–42.
2. Rajeswari K, Vaithiyanathan V, Amirtharaj P. Prediction of risk score for heart disease in India using machine intelligence. International Conference on Information and Network Technology; Singapore: IACSIT Press; 2011. p. 19–22.
3. Srimani PK, Koti MS. Knowledge discovery in medical data by using rough set rule induction algorithms. Indian Journal of Science and Technology. 2014 Jul; 7(7):905–15.
4. Aramideh J, Jelodar H. Application of fuzzy logic for presentation of an expert fuzzy system to diagnose anemia. Indian Journal of Science and Technology. 2014 Jul; 7(7):933–8.
5. Lahsasna A, Ainon RN, Zainuddin R, Bulgiba A. Design of a fuzzy-based decision support system for coronary heart disease diagnosis. J Med Syst. 2012; 26(5):3293–306.
6. Fida B, Nazir M, Naveed N, Akram S. Heart disease classification ensemble optimization using genetic algorithm. 2011 IEEE 14th International Multitopic Conference (INMIC); Karachi: IEEE; 2011. p. 19–24.
7. Marsala C. Fuzzy decision trees for dynamic data. IEEE Symposium on Evolving and Adaptive Intelligent Systems, EAIS; 2013. p. 17–24.

8. Levashenko V, Zaitseva E. Fuzzy decision trees in medical decision making support system. Proceedings of the Federated Conference on Computer Science and Information Systems; IEEE; 2012. p. 213–9.
9. Elashiri MA, Hefny HA, Abd Elwhab AH. Construct fuzzy decision trees based on roughness measures. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2012; 108:199–207.
10. Abu-halaweh NM, Harrison RW. An improved fuzzy decision tree induction tool. IEEE; 2010.
11. Jaganathan P, Kuppuchamy R. A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med. 2013; 43(12):2222–9.
12. Cheng Y, Miao D, Feng Q. Positive approximation and converse approximation in interval-valued fuzzy rough sets. Information Sciences: an International Journal. 2011; 181(11):2086–110.
13. Fan C-Y, Chang P-C, Lin J-J, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing. 2011; 11(1):632–44.
14. Holland JH. Adaptation in natural and artificial systems. Ann Arbor MI: The University of Michigan Press; 1975.
15. Sivanandam SN, Deepa SN. Introduction to genetic algorithms. Heidelberg: Springer-Verlag Berlin; 2008.
16. Fan Weiguo, Fox EA, Pathak P, Wu H. The effects of fitness functions on genetic programming-based ranking discovery for web search. J Am Soc Inform Sci Tech. 2004; 55(7):628–36.
17. Hong T. MCMC algorithm, integrated four-dimensional seismic reservoir characterization and uncertainty analysis in a bayesian framework. ProQuest LLC; 2008. p. 31.
18. Zadeh LA. Fuzzy sets. Inform Contr. 1965; 8:338–53.
19. Maarit. Fuzzy logic fundamentals, Chapter 3, Scribd. 2011. p. 70.
20. Sivasankar E, Rajesh RS. Knowledge discovery in medical datasets using a fuzzy logic rule based classifier. 2nd International conference on electronic computer technology; 2010. p. 208–13.
21. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008; 56(1):45–50.