Enhanced Prediction of Heart Disease by Genetic Algorithm and RBF Network

A. Durga Devi

M.Phil scholar, Department of Computer Science, St. Xavier's College, Palayamkottai[1]

ABSTRACT

A huge amount of data is available in hospitals and clinics. These data become useful and meaningful, only when data mining techniques are applied on them to retrieve knowledge. Nowadays, most of the human deaths occur by heart diseases. It is very important to find a prediction system with more accuracy. This prediction system predicts heart disease risk with an enhanced accuracy. The enhanced accuracy is obtained by the attribute reduction. The data mining technique RBF Network is used to create the classifier. Genetic Algorithm is used for attribute selection or reduction. The result shows that the RBF Network performs better than the other data mining techniques Naïve Bayes and J48.

Keywords: RBF Network, Genetic Algorithm, attribute selection, Naïve Bayes, J48

1. INTRODUCTION

Data mining can be defined as extracting useful knowledge from a large amout of data. Among many application fields of data mining, medical field is very important. In medical field, data mining is mainly used for disease diagnosis. The disease diagnosis is a careful task of doctors. Data mining makes easier their task. There are many disease diagnosis systems today. More accurate systems will give more accurate results. Accuracy is a main factor in disease diagnosis.

1.1 Heart Disease

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart disease has emerged as the number one killer disease in India [7]. The rate of death caused by heart disease is increasing continuously. It is the most threatening disease today. About 25 per cent of deaths in the age group of 25- 69 years occur because of heart diseases. In urban areas, 32.8 per cent deaths occur because of heart ailments, while this percentage in rural areas is 22.9. According to the California-based CADI (Coronary Artery Disease among Asian Indians) Research Foundation, India will have 62 million heart patients by 2015.

1.2 Attribute Reduction or Selection

Attribute selection reduces the data set size by removing irrelevant or redundant attributes. The goal of attribute selection to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes [6]. There are many methods for the purpose of attribute selection. In this paper, genetic search method is used for attribute selection.

1.3 Classification

Classification is a data mining technique. It is the process of finding a classifier or model to predict the class of objects whose class label is unknown [6]. The machine learning involved in classification is supervised. That is, class label of each record in training data set is known. Classification contrasts with another data mining technique clustering where machine learning is unsupervised. That is, class labels are unknown. There are many algorithms or methods for the purpose of classification. Naïve Bayes, J48 and RBF Network are some of important classification methods. In this paper, RBF Network is used for the purpose of classification

RBF Network

Radial basis function (RBF) networks are feed-forward networks trained using a supervised training algorithm [11]. They are typically configured with a single hidden layer of units whose activation function is selected from a class of functions called basis functions. While similar to back propagation in many respects, radial basis function networks have several advantages. They usually train much faster than back propagation networks. The major difference between RBF networks and back propagation networks is the behavior of the single hidden layer. Rather than using the sigmoidal or S-shaped activation function as in back propagation, the hidden units in RBF networks use a Gaussian or some other basis kernel function.

2. RELATED WORKS

In [1], a comparative analysis of data mining classification techniques was done by Dr. S. Vijarani et al for heart disease prediction. They used the Cleveland heart disease dataset for data analysis [5]. In this paper, three classification

function techniques in data mining are compared for predicting heart disease. They are function based Logistic, Multilayer perceptron and Sequential Minimal Optimization algorithm. The authors concluded that the logistic classification function technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least error rate.

A heart disease prediction system was proposed by R. R. Ade et al in [2]. The system predicts the heart disease by the techniques Support Vector Machine (SVM) and Naive Bayes. The authors used the Cleveland heart disease dataset [5] for data analysis. In this system, the authors categorized the medical data into five categories namely no, low, average, high and very high. If unknown sample comes then the system will predict the class label of that sample. The authors divided the system into training phase and testing phase. In training phase, a model (classifier) is constructed from the labeled dataset (training dataset) using one of the two techniques SVM and Naive Bayes. In testing phase, prediction is done for unlabeled dataset using the classifier constructed in training phase.

M. Anbarasi et al introduced a system to predict the presence of heart disease with reduced number of attributes [3]. Originally, there are thirteen attributes in the dataset involved in predicting the heart disease. In this work, Genetic algorithm is used to determine the attributes. Thirteen attributes are reduced to 6 attributes using genetic search. The three classifiers Naive Bayes, Classification by clustering and Decision Tree are used to predict the heart disease with the same accuracy as obtained before the reduction of number of attributes. The authors concluded that the Decision Tree data mining technique performs better than the other two techniques. The Decision Tree data mining technique shows more accuracy and low error rate than the other two data mining techniques. Classification via clustering performs poor compared to other two methods.

Nidhi Bhatla et al studied different data mining techniques that can be employed in automated heart disease prediction systems [4]. Various techniques and data mining classifiers are defined. The analysis shows that Neural Network with 15 attributes has shown the highest accuracy i.e. 100%. On the other hand, Decision Tree has also performed well with 99.62% accuracy by using 15 attributes. Moreover, in combination with Genetic Algorithm and 6 attributes, Decision Tree has shown 99.2% efficiency.

3. DATA SOURCE

The dataset used is the Cleaveland heart disease dataset obtained from UCI Repository [5]. The dataset has 14 attributes and 303 records. WEKA tool is used for data analysis of this data set. The last attribute **num** is the class attribute which is predicted by this prediction system.

No	Attributes	Description
1.	age	Age in years
2.	sex	Sex
3.	cp	Chest pain type
4.	trestbps	Resting blood pressure
5.	chol	Serum cholesterol in mg/dl
6.	fbs	Fasting blood sugar
7.	restecg	Resting electrocardiographic Results
8.	thalach	Maximum heart rate achieved
9.	exang	Exercise induced angina
10.	oldpeak	ST depression induced by Exercise relative to rest
11.	slope	The slope of the peak exercise ST segment
12.	ca	Number of major vessels (0-3) colored by fluoroscopy
13.	tal	Defect type
14.	num	Diagnosis of heart disease (angiographic disease status)

Table 1. Heart Dataset

4. METHODS USED IN THE PROPOSED SYSTEM

The two methods used in the proposed system are Genetic Algorithm and RBF Network. These two methods are explained in the sections 4.1 and 4.2.

4.1 Genetic Algorithm

This algorithm is used for attribute selection in the proposed system. Genetic Algorithms are algorithms that simulate the logic of Darwinian natural selection theory. Best attributes are selected from a set of attributes.

Terms of Genetics:

Selection: Selecting individuals for creating the next generation **Chromosome:** A string of genes

Genes: Blocks of DNA, responsible for a particular characteristic of the individual
Individual: Same as chromosome

Population: Number of individuals present with same length of chromosome **Fitness:** Fitness is the value assigned to an individual

Fitness function: Fitness function f(x) is a function which assigns fitness value to the individual **Crossover:** Genes from parents combine to form a new chromosome

Mutation: Changing randomly the gene in an individual.
Selection: Selecting individuals for creating the next generation

Crossover and mutation are two basic operators of genetic algorithm. Performance of genetic algorithm very depends on them[9]. The general genetic algorithm is given below.

START
1. Create initial population of n chromosomes
2. Assign fitness f(x) to all chromosomes
3. DO UNTIL best solution is found
    1. Select individuals from current generation
    2. Create new offspring with mutation and/or crossover
    3. Compute new fitness for all individuals
    4. Kill all the unfit individuals to give space to new offspring
    5. Check if best solution is found
END LOOP
END

Correlation feature selection

CFS (Correlation Based feature Selection) is used as a subset evaluating mechanism (Fitness Function) for Genetic Algorithm. CFS evaluates subsets of attributes on the basis of the following concept, "Good attribute subsets contain attributes highly correlated with the class, but uncorrelated to each other".

The following equation gives the merit of a feature subset *S* consisting of *k* features:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}.$$

Here, $\overline{r_{cf}}$ is the average value of all feature-class correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations.

4.2 RBF Network

RBF networks have three layers: input layer, hidden layer and output layer [10]. Each neuron in the input layer corresponds to each predictor variable. Each neuron in hidden layer consists of a radial basis function .The output layer has a weighted sum of outputs from the hidden layer to form the network outputs.

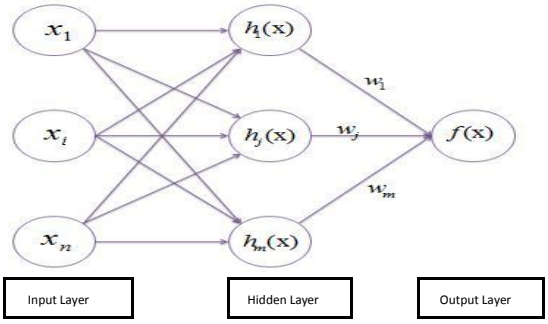

Figure 1. Architecture of RBF Network

$$h(x) = \exp\left(-\frac{(x-c)^2}{r^2}\right)$$

$$f(\mathrm{x}) = \sum_{j=1}^{m} w_j h_j(\mathrm{x})$$

h(x) is the Gaussian activation function with the parameters *r* (the radius or standard deviation) and *c* (the center or average taken from the input space) defined separately at each RBF unit. f(x) is the function for finding output.

Learning Process

The learning process is based on adjusting the parameters of the network to reproduce a set of input-output patterns. There are three types of parameters; the weight *w* between the hidden nodes and the output nodes, the center *c* of each neutron of the hidden layer and the unit width *r*.

Any clustering algorithm can be used to determine the RBF unit centers (e.g., K-means clustering). A set of clusters each with *r*-dimensional centers is determined by the number of input variables or nodes of the input layer. The cluster centers become the centers of the RBF units. The number of clusters, *H*, is a design parameter and determines the number of nodes in the hidden layer. The K-means clustering algorithm proceeds as follows:

1. Initialize the center of each cluster to a different randomly selected training pattern.
2. Assign each training pattern to the nearest cluster. This can be accomplished by calculating the Euclidean distances between the training patterns and the cluster centers.
3. When all training patterns are assigned, calculate the average position for each cluster center. Then, they will become new cluster centers.
4. Repeat steps 2 and 3, until the cluster centers do not change during the subsequent iterations.

Unit width (*r*)

When the RBF centers have been established, the width of each RBF unit can be calculated using the K-nearest eighbors algorithm. A number K is chosen, and for each center, the K nearest centers is found. The root-mean squared distance between the current cluster center and its K nearest neighbors is calculated, and this is the value chosen for the unit width (*r*). So, if the current cluster center is c_j, the *r* value is:

$$r_j = \sqrt{\frac{\sum_{i=1}^{k}(c_j - c_i)^2}{k}}$$

A typical value for K is 2, in which case s is set to be the average distance from the two nearest neighboring cluster centers.

Weights (*w*)

Using the linear mapping, **w** vector is calculated using the output vector (**y**) and the design matrix **H**

$$\mathrm{y} = \mathrm{wH}$$
$$\mathrm{w} = (\mathrm{H'H})\,\mathrm{H'y}$$

The two layers of the RBF network are trained separately, First, RBF centers c and the scaling parameters are determined, and subsequently the output layer is adjusted.

5. PROPOSED SYSTEM

As shown in the Figure 2, the proposed system has the following steps,

Step 1: First, the heart disease data is loaded in Weka and preprocessed.
Step 2: The preprocessed data is subjected attribute selection process
Step 3: A classifier is created from the reduced data
Step 4: The created classifier is used to predict heart disease of an unlabeled data.

5.1 Preprocessing

The first step of the proposed system is preprocessing. The heart disease dataset is loaded in Weka. In the data set, 7 records have missing values. These values have to be cleaned. In WEKA the data cleaning filter

ReplaceMissingValues is used to clean the missing values. This filter replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

5.2 Attribute Selection

After preprocessing the dataset is subjected to attribute selection. Attribute selection is done by genetic algorithm. Through Genetic Algorithm best features or attributes are selected.

Algorithm for Attribute selection by Genetic Algorithm

1. Initial population is created from the original attribute set.
2. Fitness value for each individual in the population is calculated using CFS.
3. Crossover and mutation operations occur on population.
4. After crossover and mutation, new population is created.
5. Best individuals are selected from newly created population based on the fitness values.
6. The steps 3 – 5 are repeated until generation count reached.

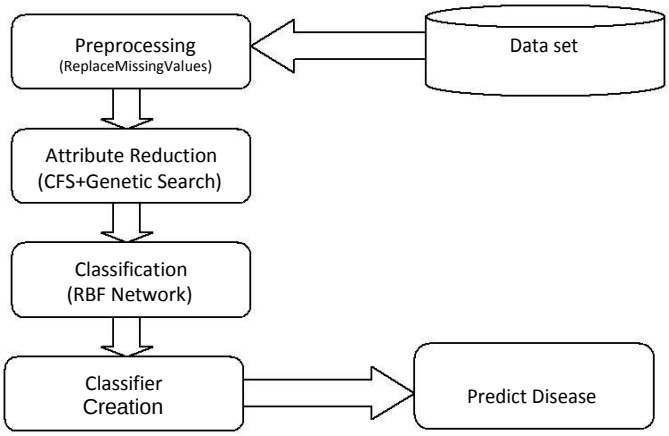

Figure 2. Proposed Framework

5.3 Classifier Creation by RBF Network

The classification technique used in the proposed system is RBF Network. It is available in weka.classifiers.functions. A classifier is created using RBF Network from the known heart data. The created classifier is used for prediction of unknown heart data. The classifier is created with cross validation of folds 10.

5.4 Heart Disease Prediction of a Single Unlabeled Instance

For this purpose, the tool NetBeans is used. A form is created using this tool. This form consists of textboxes, combo boxes and buttons. A new unlabeled instance values are entered using the combo boxes and text boxes. The created classifier is used for the prediction of the unknown data. Prediction is done when a button of the form is clicked and the result is shown in a message box.

6. RESULTS AND DISCUSSION

6.1 Reduction of Attributes

By Genetic Algorithm the 14 attributes are reduced to 9 attributes. Table 2 shows the list of reduced attributes.

No	Attributes
1.	Sex
2.	Cp
3.	Thalach
4.	Exang
5.	Oldpeak
6.	Slope
7.	Ca
8.	Thal
9.	No

Table 2. List of reduced attributes

6.2 Performance of the proposed System

The Table 5.2 shows the performance of RBF Network before and after attribute selection. The RBF Network shows 83.83 % of accuracy with 14 attributes. The prediction accuracy is increased to 85.48 % with 9 attributes. Performance of RBF Network is better after attribute reduction.

No of Attributes	Accuracy	Time taken to build model	Mean absolute Error
14 Attributes	83.83 %	0.33 sec	0.0915
9 Attributes	85.48 %	0.06 sec	0.0907

Table 3. Performance of Proposed System

6.3 Comparison with other classification techniques

The other two classification techniques taken are Naïve Bayes and J48. These two classification techniques have been used in many existing systems for heart disease prediction. Naïve Bayes gives the classification accuracy of 84.16% after attribute reduction. J48 gives the accuracy of 77.56%. RBF Network performs better than the two algorithms. The accuracy of RBF Network is 85.48%. The Table 4 shows performance of the three classification techniques.

Classification Techniques	Accuracy with	
	14 attributes	9 Attributes
Proposed System	83.83 %	85.48 %
Naïve Bayes	83.50 %	84.16 %
J48	77.56 %	77.56 %

Table 4. Comparison with other classification techniques

A graphical view for the performance of the three algorithms is shown in the Figure 3. It shows the classification accuracy of each algorithm. RBF Network shows better performance.

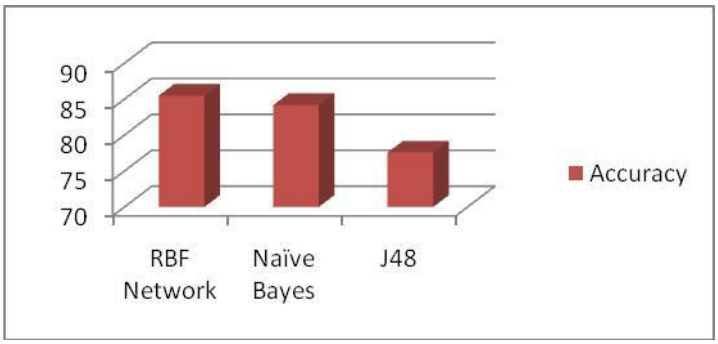

Figure 3. Comparison with other algorithms

From the Table 3, it is shown that accuracy is increased by attribute reduction. From the Table 4, it is clear that RBF Network is better than the other two algorithms**.**

6.4 Prediction of unlabeled data

The form shown in Figure 4 is created in NetBeans. It is used to load data, build classifier and to predict a single unlabeled heart data.

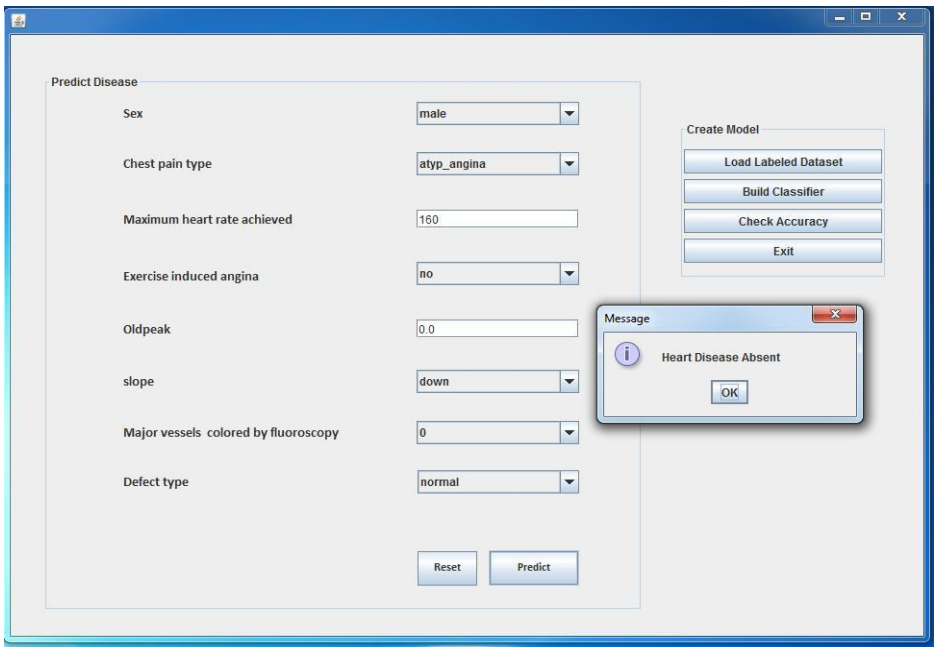

Figure 4. Prediction of Heart Disease of a Single Unlabeled Data

7. CONCLUSION AND FUTURE WORK

In this paper, a heart disease prediction system with better accuracy has been presented. Genetic algorithm has been used for attribute reduction and RBF Network for classification. Classification accuracy has been enhanced by reducing the number of attributes. From the experimental results, it is known that accuracy can be increased by reducing attributes. Attribute reduction has another benefit also. It reduces expense of patients by reducing the number of tests to be taken. The proposed system has given good results than the other two algorithms Naïve Bayes and J48.

RBF Network and Cleveland heart dataset have been used to create classifier in the proposed system. As future work, the following two changes can be done in the proposed system. Instead of RBF Network another classification technique can be applied. Another change is to use other heart dataset instead of Cleaveland heart dataset.

REFERENCES

[1] S. Vijayarani, S. Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013.

[2] R.Ade, Dhanashree, S. Medhekar, Mayur P. Bote, "Prediction using SVM and Naïve bayes", International Journal of Engineering Sciences and Research Technology, May 2013.

[3] M.Anbarasi, N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease With Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[4] Nidhi Bhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October – 2012.

[5] "Cleveland heart disease dataset",http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[6] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaumann Publisher, 2006

[7] "Indiatoday", http://indiatoday.intoday.in/story/India%27s+ no.1+killer:+Heart+disease/1/92422.html

[8] "Genetic Algorithms Understanding Using VB", http://paraschopra.com/tutorials/ga/

[9] "Introduction to Genetic Algorithms", http://www.obitko.com/tutorials/genetic-algorithms/crossover-mutation.php

[10]"Radial Basis Function Networks", http://www.saedsayad.com/artificial_neural_network_rbf.htm

[11]"Radial Basis Function Networks", www.eee.metu.edu.tr/~halici/courses/543LectureNotes/