Prediction of Heart Disease using Soft Computing and Data Mining

[1]K.Sudhakar, [2]Dr. M. Manimekalai

[1]Research Scholar, [2]Director, Department of MCA,
[1]Department of Computer Science,
[1]Shrimati Indira Gandhi College, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

***Abstract* - Cardiovascular disease (CVD) has become the primary killer worldwide and is expected to cause more deaths in the future. Prediction and prevention of CVD have therefore become important social problems. Many groups have developed prediction models for asymptomatic CVD by classifying its risk based on established risk factors (e.g., age, sex, etc.). The growth of medical databases is very high. This rapid growth is the main motivation for researchers to mine useful information from these medical databases. As the volume of stored data increases, data mining techniques play an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. In the proposed system, the genetic algorithm was first used to reduce the number of attributes that are needed for predicting the cardio vascular disease and then the neural network was employed for prediction of disease from the reduced set of attributes.**

***Keywords* - Cardiovascular Disease, Data Mining, Genetic Algorithm, Neural Networks**

I. Introduction

Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of such systems would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Models developed from these techniques will be useful for medical practitioners to take effective decision.

The World Health Organization has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths in the United States and other developed countries occur due to cardio vascular diseases. It is also the chief reason of deaths in numerous developing countries. On the whole, it is regarded as the primary reason behind deaths in adults. The term Heart disease encompasses the diverse diseases that affect the heart.

The heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes. Life itself is completely dependent on the efficient operation of the heart.

Cardiovascular disease (CVD) refers to any condition that affects the heart. Many CVD patients have symptoms such as chest pain (angina) and fatigue, which occur when the heart isn’t receiving adequate oxygen. As per a survey nearly 50 percent of patients, however, have no symptoms until a heart attack occurs. A number of factors have been shown to increase the risk of developing CVD. Some of these are [1]:

- Family history of cardiovascular disease
- High levels of LDL (bad) cholesterol
- Low level of HDL (good) cholesterol
- Hypertension
- High fat diet
- Lack of regular exercise
- Obesity

With so many factors to analyze for a diagnosis of cardiovascular disease, physicians generally make a diagnosis by evaluating a patient’s current test results. Previous diagnoses made on other patients with the same results are also examined by physicians. These complex procedures are not easy. Therefore, a physician must be experienced and highly skilled to diagnose cardiovascular disease in a patient. Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden pattern in data and provide better patient care and effective diagnostic capabilities.

Data Mining

Data mining is the process of automating information (knowledge) discovery. Knowledge Discovery in Databases (KDD) is the process of getting high-level knowledge from low-level data. Data mining plays an important role in the KDD. Data mining is an interdisciplinary field. Its main aim is to uncover relationships in data and to predict outcomes. Researchers are trying to find satisfactory solutions in a reasonable time through search techniques as many problems are difficult to be solved in a feasible time by analytically. Hence data mining got its importance.

Data mining helps to extract patterns in the process of knowledge discovery in databases in which intelligent methods are applied. The emerging field of data mining promises to provide new techniques and intelligent tools which help the human to analyze and understand large bodies of data remains on difficult and unsolved problem.

Data mining techniques have been widely used in diagnostic and health care applications because of their predictive power. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion.

The common functions in current data mining practice include Classification, Regression, Clustering, Rule generation, Discovering association rules, summarization, dependency modeling, and sequence analysis. Classification is one of the important techniques of data mining. The input to the classification problem is a data-set called the training set having a number of attributes. The attributes are either continuous or categorical. One of the categorical attributes is class label or the classifying attribute. The objective is to use the training set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data-set.

Various data mining problems can be handled effectively by soft computing techniques. These techniques are fuzzy logic, neural networks, genetic algorithms and rough sets, which will lead to an intelligent, interpretable, low cost solution than traditional techniques.

Artificial Neural Network (ANN) is one of the most used data mining method to extract patterns in an intelligent and reliable way and has been greatly used to find models that describe data relationship [3]. In view of the above said significant characteristics of ANN, Neural Network technique is adopted in this study for data classification. To perform classification task of medical data, the neural network is trained using Back propagation algorithm. As the structure of neural network is convenient for parallel processing, the output at each neuron in different layers is calculated in parallel. The performance of the network is analyzed with various types of test data.

Genetic Algorithm

In the field of artificial intelligence, a genetic algorithm (GA) [2] is a search heuristic that imitates the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate optimized solutions using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

The evolution usually starts from a population of randomly generated individuals, and is an iterative process, with the population in each iteration called a generation. In each generation, the fitness of every individual in the population is evaluated; the fitness is usually the value of the objective function in the optimization problem being solved. The more fit individuals are stochastically selected from the current population, and each individual's genome is modified (recombined and possibly randomly mutated) to form a new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. A typical genetic algorithm requires: a genetic representation of the solution domain and a fitness function to evaluate the solution domain.

A standard representation of the solution is as an array of bits. The main property that makes the genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. Variable length representations may also be used, but crossover implementation is more complex in this case. The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. Initially many individual solutions are (usually) randomly generated to form an initial population. During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. The next step is to generate a second generation population of solutions from those selected through genetic operators: crossover (also called recombination) and mutation.

Neural Networks

Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an expert in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer “what if” questions.

Neural Networks use a set of processing elements analogous to neurons in the brain. A Neural Network (NN) consists of many Processing Elements (PEs), loosely called “neurons” and weighted interconnections among the PEs. Each PE performs a very simple computation, such as calculating a weighted sum of its input connections, and computes an output signal that is sent to other PEs. The training (mining) phase of a NN consists of adjusting the weights (real valued numbers) of the interconnections, in order to produce the desired output [6].

The Artificial Neural Network (ANN) is a technique that is commonly applied to solve data mining applications. Neural Network is a set of processing units when assembled in a closely interconnected network, offers rich structure exhibiting some features of the biological neural network. The structure of neural network provides an opportunity to the user to implement parallel concept at each layer level. Another significant characteristic of ANN is fault tolerance. ANNs are well suited in situations where information is noisy and uncertain. ANN are an information processing methodology that differs drastically from conventional methodologies in that it employ training by examples to solve problem rather than a fixed algorithm [4,5].

They can be divided into two types based on the training method:

- Supervised training and
- Unsupervised training

Networks that are supervised require the actual desired output for each input where as unsupervised networks does not require the desired output for each input.

A key feature of neural networks is an iterative learning process in which data cases are presented to the network one at a time, and the weights associated with the input values are adjusted each time [6]. After all cases are presented, the process often starts over again. During this learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of input samples. Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights are chosen randomly. Then the training or learning begins.

The most popular neural network algorithm is back-propagation algorithm. Although many types of neural networks can be used for classification purposes [7], the focus is on the feedforward multilayer networks or multilayer perceptrons which are the most widely studied and used neural network classifiers. The feedforward back-propagation architecture is the most popular, effective, and easy-to-learn model for complex, multi-layered networks. Its greatest strength is in non-linear solutions to ill-defined problems. The typical back-propagation network has an input layer, an output layer, and at least one hidden layer. Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. The higher the weight the more important it is. There is no theoretical limit on the number of hidden layers but typically there are just one or two. Some work has been done which indicates that a maximum of five layers (one input layer, three hidden layers and an output layer) are required to solve problems of any complexity. Each layer is fully connected to the succeeding layer.

Training inputs are applied to the input layer of the network, and desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network, and the output of each element is computed layer by layer. The difference between the output of the final layer and the desired output is back-propagated to the previous layers, usually modified by the derivative of the transfer function, and the connection weights are normally adjusted. This process proceeds for the previous layers until the input layer is reached [8].

The advantages of Neural Networks for classification are:

- Neural Networks are more robust because of the weights
- The Neural Networks improves its performance by learning. This may continue even after the training set has been applied.
- The use of Neural Networks can be parallelized as specified above for better performance.
- There is a low error rate and thus a high degree of accuracy once the appropriate training has been performed.
- Neural Networks are more robust in noisy environment Artificial neural networks provide a powerful tool to help doctors analyze, model, and make sense of complex clinical data across a broad range of medical applications [9].
- Neural networks can be used to extract rules from a disease classification. From the rules system so discovered, we can predict if someone will have a particular stage of a particular disease.

II. Proposed System

The dataset under consideration has been taken from University of California Irvin (UCI). Originally 13 attributes were involved in predicting the heart disease. The thirteen attributes have been shown in Table 1. Feature extraction method has been used to find the minimal subset of attributes that is equivalent to original set of attributes. The number of attributes has been reduced to 6 using Genetic Search. The genetic search starts with zero attribute and an initial population with randomly generated rules. Based on the concept of survival of fittest new population is constructed to obey the fittest rule in the current population, as well as offspring of these rules. The process of generation continues until it evolves a population P where every rule in P satisfies the fitness threshold. With initial population of 20 instances, generation continued till the twentieth generation with cross over probability of 0.6 and mutation probability of 0.033. In this way, genetic search lead to the selection of 6 attributes out of the 13 mentioned (Refer Table 2).

The reduced set of attributes has been sent as input to the feed forward neural network with 6 neurons as input and four output classes predicting the chances of getting the cardiovascular disease.

Table 1 UCI Heart Dataset

Predictable Attribute Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))
Key Attribute Patient ID – Patient’s identification number

Input Attributes

1. Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
8. CA – number of major vessels colored by floursopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholestrol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Oldpeak – ST depression induced by exercise

Table 2 List of Reduced Attributes

Predictable Attribute

Diagnosis

Value 0: No heart disease Value 1: Has heart disease

Reduced Input Attributes

Type :Chest Pain Type
Rbp :Reduced blood pressure
Eia :Exercise Induced Angina
Oldpk :Old Peak
Vsal :No. of vessels coloured
Thal :Maximum Heart Rate achieved

Training the Neural Network

In the proposed system, the neural network is trained with Heart Diseases database by using feed forward neural network model and backpropagation learning algorithm with momentum and variable learning rate. The input layer of the network consists of 6 neurons to represent each reduced set of attributes.

The number of classes are four: 0 - normal person, 1- first stroke, 2- second stroke and 3- end of life. The output layer consists of two neurons to represent these four classes. The backpropagation algorithm is used to train the neural network during the training process.

Testing the Network

For testing the performance of the net, various samples are collected as test data. The test data is given as the input to the trained network and the output of the net is calculated with the adjusted weights. The output of the net is compared with the target output to study the learning ability of the network for classifying the heart disease data.

III. Conclusion

Classification is an important problem in the rapidly emerging field of data mining. Owing to the wide range of applicability of artificial intelligence and their ability to learn complex and nonlinear relationships, including noisy or less precise information, soft computing techniques are well suited to solve problems in biomedical engineering. In this work, genetic algorithm and neural network are employed for the classification of dataset. The genetic algorithm uses the feature extraction method for reducing the number of attributes and the neural network uses the feedforward backpropogation algorithm with the reduced set of attributes.

References

[1] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

[2] M. Anbarasi and E. Anupriya, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", vol.2 no.10, pp. 5370 – 5376, 2010.

[3] Sunghwan Sohn and Cihan H. Dagli, (2004) "Ensemble of Evolving Neural Networks in classification", Neural Processing Letters 19: 191-203, Kulwer Publishers.

[4] K. Anil Jain, Jianchang Mao and K.M. Mohiuddi, (1996) "Artificial Neural Networks: A Tutorial", IEEE Computers, pp.31-44.

[5] George Cybenk,, (1996)"Neural Networks in Computational Science and Engineerin", IEEE Computational Science and Engineering, pp.36-42

[6] R. Rojas, (1996) "Neural Networks: a systematic introduction", Springer-Verlag.
[7] R.P.Lippmann,"Pattern classification using neural networks, (1989)" IEEE Commun. Mag., pp. 47–64.
[8] Simon Haykin, (2001) "Neural Networks – A Comprehensive Foundation", Pearson Education.
[9] W. G. Baxt, (1990) "Use of an artificial neural network for data analysis in clinical decision making: The diagnosis of acutecoronary occlusion," *Neural Comput.*, vol. 2, pp. 480–489.