Cardiovascular disease prediction system using genetic algorithm and neural network

Conference Paper · February 2012

DOI: 10.1109/ICCCA.2012.6179185

Dr Bhuvaneswari Amma Ng
VIT University Chennai

Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network

Bhuvaneswari Amma N.G.
Department of CSE, Sudharsan Engineering College, Sathiyamangalam
Pudukkottai, Tamilnadu, India. shreena83@yahoo.com

Abstract— **Medical Diagnosis Systems play a vital role in medical practice and are used by medical practitioners for diagnosis and treatment. In this paper, a medical diagnosis system is presented for predicting the risk of cardiovascular disease. This system is built by combining the relative advantages of genetic algorithm and neural network. Multilayered feed forward neural networks are particularly suited to complex classification problems. The weights of the neural network are determined using genetic algorithm because it finds acceptably good set of weights in less number of iterations. The dataset provided by University of California, Irvine (UCI) machine learning repository is used for training and testing. It consists of 303 instances of heart disease data each having 14 attributes including the class label. First, the dataset is preprocessed in order to make them suitable for training. Genetic based neural network is used for training the system. The final weights of the neural network are stored in the weight base and are used for predicting the risk of cardiovascular disease. The classification accuracy obtained using this approach is 94.17%.**

Keywords- Genetic Algorithm; Neural Network, Backpropagation Algorithm; Cardiovascular Disease; Prediction Engine

I. INTRODUCTION

Cardiovascular disease refers to the class of diseases that involve the heart or blood vessels. Cardiovascular disease technically refers to any disease that affects the cardiovascular system; it is usually used to refer to those related to atherosclerosis. Cardiovascular diseases include coronary heart disease, cerebrovascular disease, raised blood pressure, peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. In practice, cardiovascular disease is treated by cardiologists, thoracic surgeons, vascular surgeons, neurologists, and interventional radiologists, depending on the organ system that is being treated. The heart is the organ that pumps blood to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidney suffer and if the heart stops working, death occurs within minutes. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to heart disease [2].

Medical diagnosis is an important yet complicated task that needs to be done accurately and efficiently. The automation of this system is very much needed to help the physicians to do better diagnosis and treatment. The representation of medical knowledge, decision making, choice and adaptation of a suitable model are some issues that a medical system should take into consideration. Medical progress is always supported by data analysis which improves the skill of medical experts and establishes the treatment technique for diseases. The purpose of medical diagnosis system is to assist physicians in defining the risk level of an individual patient. The heart disease dataset found in University of California, Irvine Machine Learning Repository is used for training and testing the system [10]. The purpose of using this dataset is to provide a complex, real world data example where the relationships between the features are not easily discovered by casual inspection.

In this proposed system, the advantages of genetic algorithm and neural network are combined to predict the risk of cardiovascular disease. Genetic algorithm is an optimization algorithm that mimics the principles of natural genetics. It finds acceptably good solutions to problems acceptably quickly. In many applications, knowledge that describes desired system behavior is contained in datasets. When datasets contain knowledge about the system to be designed, a neural network promises a solution because it can train itself from the datasets. Neural networks are adaptive models for data analysis particularly suitable for handling nonlinear functions. By combining the optimization technique of genetic algorithm with the learning power of neural network, a model with better predictive accuracy can be derived.

II. RELATED WORK

A number of approaches have been used to predict the risk of cardiovascular diseases. Hai, Hussain, and Xin (2008), in their work proposed neural based learning classifier system for classifying data mining tasks. They conducted experiments on 13 different datasets from the University of California, Irvine repository and one artificial dataset. They showed that neural based learning classifier system performs equivalently to supervise learning classifier system on five datasets, significantly good performance on six datasets and significantly poor performance on three datasets [3].

Shantakumar and Kumaraswamy (2009), in their work proposed an intelligent and effective heart attack prediction system using data mining and artificial neural network. They also proposed extracting significant patterns for heart disease prediction. They used K-means clustering to extract the data appropriate to heart attack from the warehouse.They used MAFIA algorithm to mine the frequent patterns[6][7].

Shanthi, Sahoo, and Saravanan (2009), in their work proposed a decision support system using evolving connection weights of artificial neural networks with genetic algorithms with the application to predict stroke disease. The data for their work have been collected from 150 patients who have the symptoms of stroke disease. The result shows that this hybrid approach is better compared to traditional artificial neural network and feature selection using genetic algorithm [8].

Niti, Anil, and Navin (2007), in their work proposed a decision support system for heart disease diagnosis using neural network. They trained their system with 78 patient records and the errors made by humans are avoided in this system [5].

Anbarasi, Anupriya, and Iyengar (2010), in their work proposed an enhanced prediction of heart disease with feature subset selection using genetic algorithm. They predicted more accurately the presence of heart disease with reduced number of attributes. They used Naïve Bayes, Clustering, and Decision Tree methods to predict the diagnosis of patients with the same accuracy as obtained before the reduction of attributes. They concluded that the decision tree method outperforms the other two methods [1].

Latha and Subramanian (2007), in their work proposed an intelligent heart disease prediction system using CANFIS and genetic algorithm. They simulated their work using Neurosolution software. Cleveland heart disease dataset is used for analysis. The mean square error obtained is only 0.000842 [4].

Yung-Keun Kwon and Byungo-Ro Moon (2007), in their work proposed a hybrid Neuro-Genetic system for stock trading. They tested the system with 36 companies in NYSE and NASDAQ for 13 years from 1992 to 2004. They used a recurrent neural network as the prediction model. Genetic algorithm was used to optimize the weights of the neural network. The hybrid system showed notable improvement on the average over the buy and hold strategy [9].

Comparing to the works discussed above, the work discussed in this paper is different by using genetic algorithm and neural network for predicting the risk of cardiovascular disease. Genetic based neural network is used to train the system. The weights of the neural network are optimized using genetic algorithm.

III. SYSTEM ARCHITECTURE

The architecture of the proposed system is illustrated in Figure 1.The major components of this system are Cardiac Database, Preprocessing Engine, Weight Optimization Engine, Training Engine, Weight Base, and Prediction Engine.

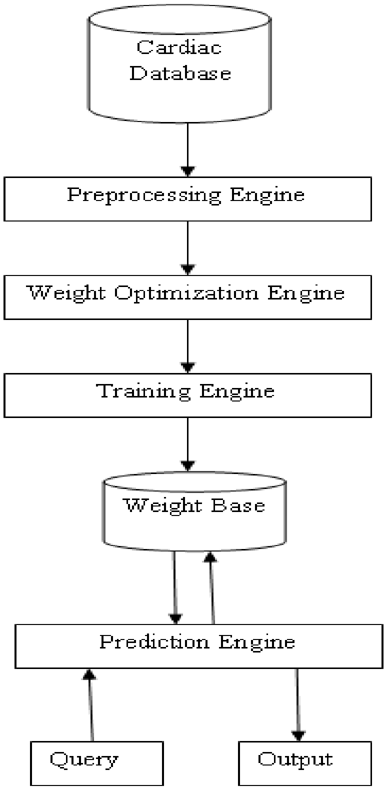

Figure 1 . System Architecture

A.Cardiac Database

The Cleveland heart disease data provided by the University of California, Irvine Machine Learning Repository [10] is used for analysis of this work. The dataset has 13 numeric input attributes namely age, sex, chest pain type, cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, old peak, slope, number of vessels colored and thal. It also has the predicted attribute ie) the class label. The description of the dataset is tabulated in Table 1.

B. Preprocessing Engine

Preprocessing is an important step in the knowledge discovery process, as real world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. In this proposed work, the most probable value is used to fill in the missing values. Data transformation routines convert the data into appropriate forms for mining. Normalization is useful for classification purpose. By normalizing the input values for each attribute measured in the training tuples will speed up the learning process. In this work, the normalization technique used is min-max normalization. The min-max normalization given in equation (1) discussed in [12] is defined as follows:

$$v^1 = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) \quad \textbf{(1)}$$

In this work, the following attributes are normalized: age, trestbps, chol, thalach, and oldpeak.

TABLE 1. SUMMARY OF THE HEART DISEASE DATASET

Attribute	Description	Domain of value
Age	Age in years	29 to 77
Sex	Sex	Male (1) Female (0)
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)
Trestbps	Resting blood sugar	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)
Restecg	Resting ECG result	Normal (0) ST-T wave abnormality (1) LV hypertrophy (2)
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise induced angina	Yes (1) No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)
Ca	Number of major vessels coloured by fluoroscopy	0–3
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)
Num	Heart disease	0–4

C. Weight Optimization Engine

Weight determination technique discussed in [11] is used for optimizing the weights of the neural network. Genetic algorithm uses a direct analogy of natural behavior, work with a population of individual strings, each representing a possible solution to the problem. Each individual string is assigned a fitness value which is an assessment of how good a solution is, to a problem. The best fit individuals participate in reproduction by cross breeding with other individuals in the population. A whole new population of possible solutions to the problem is generated by selecting the best fit individuals from the current generation. This new generation contains characteristics which are better than their ancestors.

The following is the methodology for weight optimization used in this work as discussed in [11]:

Step 1: Get the input-output pairs (I_i, T_i), i=1, 2, ..,N where Ii = $(I_{1i}, I_{2i}, \ldots, I_{ni})$ and Ti=$(T_{1i}, T_{2i}, \ldots, T_{ni})$ of the neural network, Chromosome C_i, i=1,2,…,p belonging to the current population P_i whose size is p.

Step 2: Extract weights W_i from C_i using the actual weight w_k which is given in equation (2):

$$W_k = \begin{cases} + \dfrac{x_{kd+2}10^{d-2} + x_{kd+3}10^{d-3} + \ldots + x_{(k+1)d}}{10^{d-2}} & , \text{if } 5 = x_{kd+1} = 9 \\[2ex] - \dfrac{x_{kd+2}10^{d-2} + x_{kd+3}10^{d-3} + \ldots + x_{(k+1)d}}{10^{d-2}} & , \text{if } 0 = x_{kd+1} < 5 \end{cases} \quad \textbf{(2)}$$

where $x_1, x_2, \ldots, x_d, .., x_L$ represent a chromosome and $x_{kd+1}, x_{kd+2}, \ldots, x_{(k+1)d}$ represent the kth gene (k≥0) in the chromosome.

d represent the number of digits to be randomly generated for representing a weight value

Step 3: Keeping W_i as a fixed weight setting train the neural network for the N input instance. Calculate error E_i using equation (3):

$$\mathbf{E_i = \sum_{J} (T_{ji} - O_{ji})^2} \quad \textbf{(3)}$$

Where O_{ji} is the output calculated by the NN

Step 4: Calculate the root mean square E of the errors using equation (4):

$$\mathbf{E = \sqrt{\sum_{i} E_i / N}} \quad \textbf{(4)}$$

where E_i , i=1,2,…,N

Step 5: Calculate the fitness value F_i using equation (5) for each of the individual string of the population

$$\mathbf{F_i = 1/E} \quad \textbf{(5)}$$

Step 6: Select parent using roulette wheel parent selection. Apply single point crossover and mutate child chromosome to the parent chromosome.

Step 7: Check for benefiting of child chromosome with the objective function. Replace the old generation by the new generation and name it the best chromosome.

Repeat steps 2 to 7 till the stopping criterion is met.

D. Training Engine

Back propagation algorithm discussed in [11] is used for training the neural network. The reason for choosing this algorithm is, it can find a good set of weights in a reasonable amount of time. Backpropagation is a variation of gradient search. It uses a least-square optimality criterion. The key to backpropagation is a method for calculating the gradient of the error with respect to the weights for a given input by propagating error backwards through the network. The neural network has the capability to quickly classify a dataset. It is trained on a set of training data until it reaches a predefined threshold level. The Backpropagation algorithm can be outlined as follows:

Step 1: Get the number of input nodes, hidden nodes and output nodes.

Step 2: Get the weights from the weight optimization subsystem.

Step 3: For the training data, present the set of inputs and outputs. By using the linear activation function, the out put of the input layer is evaluated as $\{O\}_I=\{I\}_I$, where $\{I\}_I$ is the training data set.

Step 4: Compute the inputs to the hidden layer by multiplying the corresponding weights of synapses using equation (6):

$$\{I\}_H= [V]^T\{O\}_I \qquad (6)$$

where $[V]^T$ is the weight matrix for input to hidden layer obtained from weight optimization subsystem.

Step 5: Let the hidden layer units evaluate the output using the sigmoidal function as given in equation (7):

$$\{O\}_H= 1/(1+e^{-I}{}_{Hi}) \qquad (7)$$

Step 6: Compute the inputs to the output layer by multiplying the corresponding weights of synapses as given in equation (8):

$$\{I\}_O= [W]^T\{O\}_H \qquad (8)$$

where $[W]^T$ is the weight matrix for hidden to output layer obtained from weight optimization subsystem.

Step 7: Let the hidden layer units evaluate the output using the sigmoidal function as given in equation (9):

$$\{O\}_O= 1/(1+e^{-I}{}_{Oj}) \qquad (9)$$

Step 8: Calculate the error and the difference between the network output and the desired output using equation (10):

$$E_i=\sum_{J} (T_{ji}-O_{ji})^2 \qquad (10)$$

Where T_{ji} is the desired output and O_{ji} is the output calculated by the neural network.

D. Weight Base

Weight base contains the final weights of the trained neural network. These weights are used by the prediction engine to predict the severity of cardiovascular disease.

E. Prediction Engine

Prediction engine predicts the severity of cardiovascular disease. When the user submits the query to the prediction engine, it gets the weights from the weight base and predicts the severity of the disease. The proposed prediction engine is shown in Figure 2. It consists of 13 nodes in the input layer, 7 nodes in the hidden layer and only one node in the output layer. It gets the weights from the weight base and works same as that of backpropagation network's initial iteration. But, no error is calculated.

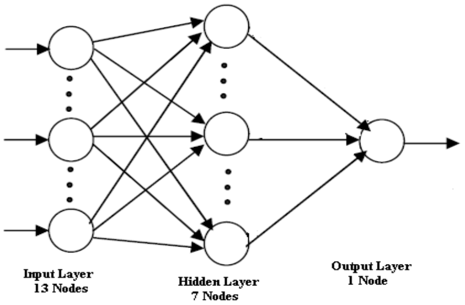

Figure 2. Proposed Prediction Network

IV. EXPERIMENTAL RESULTS

The Cleveland Heart Disease Dataset provided by the UCI Machine Learning Repository [8] is used for training and testing the medical diagnosis system. The distribution of dataset is given in Table 2. Among the 303 instances of data, 200 instances are used for training and 103 instances are used for testing.

TABLE 2. DISTRIBUTION OF DATA

Class	*0* *Absent*	*1* *Low*	*2* *Medium*	*3* *High*	*4* *Serious*
Training	109	38	20	23	10
Testing	55	17	16	12	3

The cardiac dataset is preprocessed in order to make it suitable for further processing. The initial population of weights is randomly generated. The population size is determined by the number of nodes in the neural network. The number of nodes in the input layer, hidden layer, and output layer are 13, 7 and 1 respectively. Therefore, the total number of weights needed for training is calculated using equation (11) which is discussed in [11]:

Weights = (Number of input nodes + Number of output nodes) x Number of hidden nodes (11)

In this work, the required number of weights is 98. Here each weight is considered as a gene and the gene length is assumed as 5. Then, the string length is calculated using equation (12) as discussed in [11]:

String length = (Number of input nodes+ Number of output nodes) x Number of hidden nodes x gene length (12)

In this work, the string length is 490. Therefore, the initial population of weights is 490. Using the fitness function, the best fit and worst fit individuals are selected and then duplicate the best fit with the worst fit. Then, the reproduction operators are applied and the process is continued until the best solution is obtained. Backpropagation algorithm is used for training and the mean square error between the actual and desired output is reduced to a predetermined level. The final weights are stored in the weight base. This is used by the prediction engine to predict the risk of cardiovascular disease. The training and testing dataset classification by genetic based neural network is given in Table 3. and Table 4. respectively. The classification accuracy of training set 99%. The performance measures of the testing test is given in Table 5. The classification accuracy of testing set is 94.17%.

TABLE 3. CLASSIFICATION OF TRAINING DATA

Class	*0*	*1*	*2*	*3*	*4*
	Absent	*Low*	*Medium*	*High*	*Serious*
Yes	108	37	20	23	10
No	1	1	0	0	0

TABLE 4. CLASSIFICATION OF TESTING DATA

Class	*0*	*1*	*2*	*3*	*4*
	Absent	*Low*	*Medium*	*High*	*Serious*
Yes	52	16	15	12	2
No	3	1	1	0	1

TABLE 5. PERFORMANCE MEASURES

Measures	*Classifier*
True Positive	94.55%
False Positive	6.25%
True Negative	93.75%
False Negative	5.45%
Accuracy	94.17%

V. CONCLUSION AND FUTURE WORK

In this paper, a framework for Decision Support System is developed for the analysis of medical data. The Heart Disease dataset is taken and analyzed to predict the severity of the disease. A genetic based neural network approach is used to predict the severity of the disease. The data in the dataset is preprocessed to make it suitable for classification. The weights for the neural network are determined using genetic algorithm. The preprocessed data is classified into five classes based on the severity of the disease using Backpropagation Algorithm and the final weights of the neural network are stored in the weight base. These weights are used for predicting the risk of cardiovascular disease. All the attributes are taken into consideration to predict the risk of cardiovascular disease. The accuracy obtained is 94.17%.

There are many interesting aspects for future work. This system can be enhanced by using Genetic Algorithms and Principal Component Analysis to reduce the dimension of the dataset and is used to predict the risk of cardiovascular diseases.

REFERENCES

[1] M.Anbarasi, E.Anupriya, 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, No.10, pp.5370-5376, 2010.

[2] S.Goenka, D.Prabhakaran, V.S.Ajay, and K.S.Reddy, "Preventing Cardiovascular Disease in India – Translating Evidence to Action", Current Science, Vol.97, No.3, pp.367-377, 2009.

[3] Hai H.Dam, Hussain A.Abbass and Xin Yao, "Neural – Based Learning Classifier Systems", IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, pp.26-39, 2008.

[4] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences, Vol.3, No.3, pp.157-160, 2007.

[5] Niti Guru, Anil Dahiya and Navin Rajpal, " Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review,Vol.8, No.1, pp.99-101,2007.

[6] Shantakumar B.Patil and Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol.31, No.4, pp.642-656, 2009.

[7] Shantakumar B.Patil and Y.S.Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", International Journal of Computer Science and Network Security ,Vol.9, No.2, pp.228-235, 2009.

[8] D.Shanthi, G.Sahoo and N.Saravanan, "Evolving Connection Weights of Artificial Neural Networks using Genetic Algorithm with Application to the Prediction of Stroke Disease", International Journal of Soft Computing, Vol.4, No.2, pp.95-102, 2009.

[9] Yung-Keun Kwon and Byungo-Ro Moon, "A Hybrid Neuro-Genetic Approach for Stock Forecasting", IEEE Transactions on Neural Networks,Vol.18, No.3, pp.851-864, 2007.

[10] http://www.ics.edu, UCI Repository of Machine Learning Data bases, Cleveland Heart Disease Dataset.

[11] S.Rajasekaran and G.A.Vijayalakshmi Pai, "Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications", Prentice Hall of India, 2007.

[12] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishers, 2009.