Heart disease prediction using risk factors in genetic neural network based data mining

[1] V.Poornima, [2] Dr. D. Gladis
[1] Research Scholar, [2] Principal
[1, 2] Department Of Computer Science
[1] Bharathiar University, Coimbatore, Tamil Nadu, India.
[2] Bharathi Womens College

Abstract : Data mining techniques have a wide range application in clinical solution support systems for prediction and analysis of various diseases with high accuracy. These practices have an ability to determine the various unseen patterns and deals in medical data and also used in designing clinical support systems. Thus the most important applications of these system are in analyses of heart diseases. Over the world, heart diseases is a major one of the leading causes to death. In every system that forecast the heart diseases uses the clinical dataset and it has various limitations and inputs of many complex tests results. There is no much system which predicts the heart diseases based on Risk Factors (RF) such as family history, age, diabetes, hypertension, high cholesterol, alcohol intake, tobacco smoking, physical inactivity or obesity etc. These visible RFs are commonly observed in all heart disease patients. The system based on the RFs will not only help the medical professionals but it would act as a warning for the presence of heart diseases before he/she goes for costly medical checkups or visits the hospital. Therefore this paper presents a technique for forecasting the heart disease by the use of major RFs. This involves data mining tools, neural networks and Genetic Algorithms (GA) and the system uses the global optimization benefit of GA for initialization of neural network weights. The learning is fast, more stable and accurate as compared to back propagation. The system was implemented in Matlab and predicts the risk of heart disease with an accuracy of 89%

***IndexTerms* - Data mining, Matlab, genetic algorithms, neural networks.**

I. Introduction

Every day many people die because of various heart diseases. It is considered as one of the major preventable diseases that causes more than 12 million deaths every year [1]. RFs which increases the likelihood of heart diseases are categorized as modifiable and other as non-modifiable RFs. Modifiable factors like abnormal blood lipids, pressure, diabetes, obesity etc. are preventable and non-modifiable factors like sex, age, eating habits, family history etc. are uncontrollable [2,3]. Using the corresponding values of those RFs there are numerous algorithms used in predicting the risk of occurring heart diseases.

Even though there are large number of RFs when predicting the occurrence of heart diseases use only few of them. Because when increasing the number of features it may reduce the accuracy and the performance of classifier that uses for prediction. Data mining and neural network technologies are commonly used in extracting important and knowledgeable information from medical data [4,5]. In the medical domain data mining has a large prospective for revealing the unseen patterns in data sets. With the help of these patterns clinical diagnosis can be carried out. Yet the existing raw medical data are broadly circulated, heterogeneous in nature and voluminous. The data are to be collected in a systematic form and these collected data creates to form a hospital information system. Data mining technology provides a method to novel and hidden patterns in the data. Both Data mining and statistics attempt towards determining the various patterns and structures in data. Data mining has its applications towards various heterogeneous fields but statistics deals with heterogeneous numbers only.

The most active model to forecast patients with heart disease seems to be Naïve Bayes followed by Neural Network and Decision Trees [5]. Continuous data can also be used instead of just categorical data. The other field is to make use of Text Mining to mine the various unstructured data which are available in healthcare databases. Another challenge would be to integrate data mining and text mining [6]. A generalized definition of data mining is provided as “Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data” [7]. Cluster analysis is one area of machine learning of particular interest to data mining. It provides for the organization for a collection of represented as a vector in a multidimensional space, patterns, into clusters based on the similarity of these patterns [8].

Therefore this work provides a technique for prediction of heart disease using the major RFs. The technique involves data mining tools, GA and neural networks. The hybrid system applied uses the global optimization benefits of GA for initialization of neural network weights. The learning is fast, more stable and accurate as compared to back propagation. The system predicts the risk of heart disease with an accuracy of 89%, this is achieved through the implementation carried in Matlab.

II. Data Mining

Data Mining (DM) is major anxious with the study of data and DM tools and techniques are used for discovery patterns from the data set. The most significant aim of DM is to find patterns mechanically with low input and efforts. DM is a dominant tool able to use decision building and for forecasting expectations trends of market. DM tools and techniques can be effectively functional in different fields in different forms. Now a days for data analysis too many Organizations begin to use DMs a tool.

By using Mining tools and techniques, different fields of business get advantage by simply assess various trends and pattern of market and to make rapid and efficient market trend analysis and it is also helpful tool for the diagnosis of diseases.

III. Techniques Used in Data Mining

Classification

Classification is a classic data mining technique based on machine learning. Mainly classification referees in classifying the item in collective set of data in one predefined set of groups. Classification technique uses mathematical techniques such as decision trees, linear programming, neural network and statistics.

Clustering

Clustering is a data mining technique that makes significant or helpful cluster of substance that have similar feature using mechanical technique. Dissimilar from classification, clustering technique also defines the classes and put objects in them, as in classification objects are assigned into predefined classes. For example in prediction of heart disease by using clustering obtain cluster or state that list of patients which have same RF. Funds this makes the split list of patients with high blood sugar and related RF so on.

Association

Association is one of the best known data mining technique. In association, a pattern is exposed based on a relationship of a particular item on other items in the same operation. For example, the association technique is used in heart disease prediction as it say to us the relationship of dissimilar attributes used for analysis and sort out the patient with all the RF which are necessary for prediction of disease.

Prediction

The prediction as it name indirect is one of a data mining techniques that discovers relationship between independent variables and relationship among dependent and independent variables. For example, prediction analysis technique can be used in sale to predict profit for the future if consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data and can draw a fixed regression curve that is used for profit prediction.

Data Analysis and Encoding

The problem with RFs related to heart disease is that there are many RFs involved like age, usage of cigarette, blood cholesterol, person's fitness, blood pressure, stress and etc. and understanding and categorizing each one according to its importance is a difficult task. Also a heart disease is often detected when a patient reaches advanced stage of the disease [9]. Hence the RFs were analyzed from various sources [10]-[11]. The dataset was composed of 12 important RFs which were sex, age, family history blood pressure, Smoking Habit, alcohol consumption, physical inactivity, diabetes, blood cholesterol, poor diet, obesity .The system indicated whether the patient had risk of heart disease or not. The data for 50 people was collected from surveys done by the American Heart Association [11]. Most of the heart disease patients had many similarities in the RFs [12]. The TABLE I below shows the identified important RFs and the corresponding values and their encoded values in brackets, which were used as input to the system.

Table I : RISK FACTORS VALUES AND THEIR ENCODINGS

	Risk Factors	Values
1.	Sex	Male (1), Female (0)
2.	Age (years)	20-34 (-2), 35-50 (-1), 51-60 (0), 61-79 (1) , >79 (2)
3.	Blood Cholesterol	Below 200 mg/dL - Low (-1) 200-239 mg/dL - Normal (0) 240 mg/dL and above - High (1)
4.	Blood Pressure	Below 120 mm Hg- Low (-1) 120 to 139 mm Hg- Normal (0) Above 139 mm Hg- High (-1)
5.	Hereditary	Family Member diagnosed with HD -Yes (1) Otherwise –No (0)
6.	Smoking	Yes (1) or No (0)
7.	Alcohol Intake	Yes (1) or No (0)
8.	Physical Activity	Low (-1) , Normal (0) or High (-1)
9.	Diabetes	Yes (1) or No (0)
10.	Diet	Poor (-1), Normal (0) or Good (1)
11.	Obesity	Yes (1) or No (0)
12.	Stress	Yes (1) or No (0)
Output	Heart Disease	Yes (1) or No (0)

Data analysis has been carried out in order to transform data into useful form, for this the values were encoded mostly between a range [-1, 1]. Data analysis also removed the inconsistency and anomalies in the data. This was needed. Data analysis was needed for correct data preprocessing. The removal of missing and incorrect inputs will help the neural network to generalize well.

IV. Neural Network Weight Optimization by Genetic Algorithm

This system uses back propagation algorithm for learning and training the neural network, but there are two major disadvantages with back propagation algorithm. First is that the initialization of the NN weights is a blind process hence it is not possible to find out globally optimized initial weights and there is a danger that the network output would run towards local optima hence the overall tendency of the network to find out a global solution is greatly affected. The second problem is that back propagation algorithm is very slow in convergence and there is a possibility that network never converges [13]. This problem of local optimum solution can be solved by optimizing the initial weights of neural network. For this we use a GA which is specialized for global searching [14]. For this we first determine the number of inputs, layers and hidden neurons of the neural network and then we would use the back propagation algorithm to train the networks using the weights optimized by GA.

V. Neural Network Architecture

A multilayered feed-forward network is used having 12 input nodes 10 hidden nodes and 2 output nodes. The number of input is based on the final set of RFs for each patient which is given in TABLE I. number of hidden nodes must be decided for which the training is fast and the network gives the best output.

The first step is to initialize the weights of neural network using the 'configure' function available in MATLAB. Then these configured weights are passed to the GA for optimization according to the fitness function. Once the weights are optimized, the Levenberg-Marquardt back propagation algorithm is used for training and learning and 'trainlm' is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. The 'trainlm' is often the fastest back propagation algorithm in the toolbox, and is highly recommended as a first-choice supervised algorithm, although it does require more memory than other algorithms. Maximum number of epochs to train is set to a default value 100. The learning stops at a predefined minimum error after modifying network weights and adjusting them to an optimal quantity at which the classification is accurate. The predicted output would be presence or absence of a heart disease.

VI. PARAMETER SETTINGS

The system was developed using MATLAB R2012a. Global Optimization Toolbox and the Neural Network Toolbox were used for implementing the algorithm [15]. The data for RFs related to heart diseases collected from 50 people is provided in TABLE II. ANN is initialized with the 'configure' function, with each weigh being between -1.0 to +1.0. These weights are then passed to the GA which uses the mean square error as the fitness function. The interconnecting weights and thresholds of the trained neural network are passed to the genetic algorithm. The number of neurons in the three layer neural networks is 12, 10, and 2 respectively in input, hidden and output layer. Hence there are (12x10+10) + (10x2+2) = 152 total weights and biases. The weights in the ANN are encoded in such a way each weight is being between -1.0 to +1.0. After that weights are assigned to each link. Weights adjustment using GA is done with 'population size =20'.In this application, each string or chromosome in the population represents the weight and bias values of the network. Fitness function is calculated for each chromosome based on mean square error. The fitness function used is mean square error (mse) which is calculated as below:

$$\sum k = \frac{(Ok - Tk)2}{n}$$

After selection, crossover and mutation in GA, the chromosomes with lower adaptation are replaced with better ones, and the better and fitter chromosomes (optimized solutions) that correspond to the interconnecting weights and thresholds of neural network are generated. A small value, closes to zero, shows that the network has generalized well and is ready for the classification problem. In this method GA searches among several set of weight vectors simultaneously. The initial population is randomly generated. By selecting suitable parameters, like selection criteria, probability of cross-over, probability of mutation, initial population, etc., to the GA, high efficiency and performance can be achieved.

TABLE III

Some Parameters Used In GA

Search Method	Genetic Algorithm
Population Size	20
Generations	100
Crossover Fraction	0.800

Migration Interval	20
Migration Fraction	0.2000
Elite Count	2
TolFun	1.0000e-006

VII. RESULTS AND DISCUSSION

The input data consisted of RFs collected from 50 people through case studies provided at the website of the American Heart Association [16]. The data was encoded as shown in TABLE II. 70% of the data was used for training and 15% each for testing and validation. A confusion matrix is produced using Matlab and accuracy is determined (shown in TABLE IV) as Accuracy = (TP + TN**)** / **(**TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively. The accuracy of prediction of heart disease on the training data was calculated as 89% and accuracy on validation data was 96.2%. The least mean square error (MSE) achieved was 0.034683 after 12 epochs, as shown in Figure 1. Results show GA and neural network approach gives better average prediction accuracy than the traditional ANN.

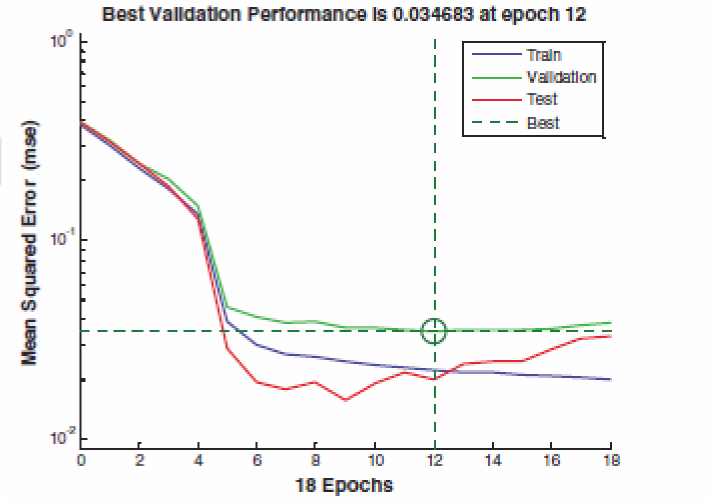

Figure 1: Performance Graph

TABLE IV

Data Sets

Data Set	Number of Data	Accuracy (%)
Training Set	34	96.2%
Test Set	8	92%
Validation Set	8	89%
Total instances	50	-

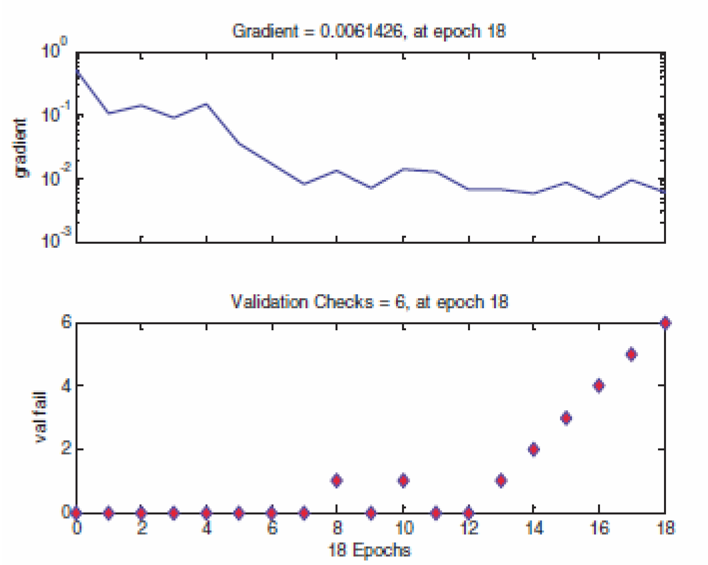

Figure2: Training State Graph

VIII. CONCLUSION

Data mining techniques and methods applied in patient medical dataset has resulted in innovations, standards and decision support system that have significant success in improving the health of patients and the overall quality of medical services. But we still need systems which could predict heart diseases in early stages. In this study, a new hybrid model of Neural Networks and GA to optimize the connection weights of ANN so as to improve the performance of the Artificial Neural Network. The system uses identified important RFs for the prediction of heart disease and it does not require costly medical tests. RFs data of 50 patients was collected and the results obtained showed training accuracy of 96.2% and a validation accuracy of 89% as specified in TABLE IV. With using hybrid data mining techniques we could design more accurate clinical decision support systems for diagnosis of diseases. We can build an intelligent system which could predict the disease using RFs hence saving cost and time to undergo medical tests and checkups and ensuring that the patient can monitor his health on his own and plan preventive measures and treatment at the early stages of the diseases.

TABLE II

No.	Sex	Age	Blood Cholestrol	Blood Pressure	Hereditary	Smoking	Acohol Intake	Physical Activity	Diabetes	Diet	Obesity	Stress	Heart Disease
1	Female	35	High	Normal	No	No	Yes	Low	Yes	Poor	Yes	Yes	Yes
2	Male	70	Low	Low	No	No	Yes	High	Yes	Normal	No	No	No
3	Female	60	High	High	No	No	No	Normal	Yes	Poor	Yes	Yes	Yes
4	Female	36	Low	Normal	No	No	No	Normal	No	Good	No	No	No
5	Male	30	Low	Normal	No	No	Yes	High	No	Normal	No	No	No
6	Female	39	Low	Normal	Yes	No	Yes	High	Yes	Normal	No	Yes	No
7	Female	41	High	Normal	No	No	No	Low	No	Poor	Yes	No	No
8	Male	70	High	Normal	No	No	Yes	Low	No	Poor	Yes	No	Yes
9	Male	65	Normal	High	Yes	Yes	Yes	Normal	Yes	Poor	Yes	No	Yes
10	Male	30	Normal	High	No	Yes	No	Normal	No	Good	No	Yes	No
11	Female	31	Low	Normal	No	No	No	High	No	Normal	No	No	No
12	Female	29	Low	Normal	No	No	Yes	High	No	Good	No	No	No
13	Male	30	Low	Normal	No	No	Yes	Normal	No	Normal	No	No	No
14	Female	45	Normal	High	Yes	Yes	No	Normal	Yes	Normal	Yes	Yes	No
15	Male	25	High	Normal	Yes	Yes	Yes	Low	Yes	Normal	No	No	Yes
16	Female	37	Normal	Normal	No	No	No	Normal	Yes	Poor	No	Yes	No
17	Female	37	Normal	High	No	Yes	Yes	High	No	Poor	No	Yes	No
18	Male	53	High	Low	No	Yes	No	Normal	Yes	Normal	No	Yes	No
19	Male	57	High	Normal	No	Yes	No	Low	No	Poor	Yes	Yes	Yes
20	Male	52	High	Low	No	No	No	Normal	Yes	Poor	Yes	No	No
21	Male	48	Normal	Normal	Yes	Yes	Yes	Normal	No	Normal	No	No	Yes
22	Male	62	High	High	No	Yes	Yes	Normal	Yes	Normal	No	No	Yes
23	Male	56	Normal	High	No	Yes	Yes	Low	No	Poor	Yes	No	Yes
24	Female	27	Low	Normal	No	No	No	High	No	Good	No	Yes	No
25	Male	33	Normal	Normal	No	No	No	Normal	Yes	Good	No	No	No

26	Female	33	Normal	Normal	No	No	Yes	Low	Yes	Poor	No	No	No
27	Male	37	High	Normal	No	No	Yes	Normal	No	Normal	No	Yes	No
28	Male	43	Normal	High	No	No	No	Normal	Yes	Poor	Yes	Yes	Yes
29	Male	46	Low	Normal	No	No	No	Normal	Yes	Poor	Yes	Yes	No
30	Female	36	Low	Normal	No	No	No	Normal	No	Normal	No	No	No
31	Female	29	Low	Normal	No	No	No	Normal	No	Good	No	Yes	No
32	Female	47	Normal	Normal	No	No	Yes	High	Yes	Normal	No	Yes	No
33	Male	58	High	High	No	Yes	Yes	Normal	Yes	Normal	No	Yes	Yes
34	Male	44	High	Normal	Yes	No	Yes	Normal	No	Normal	Yes	Yes	Yes
35	Female	36	Normal	High	No	No	No	Normal	No	Good	Yes	No	Yes
36	Male	42	Low	Normal	No	No	Yes	Low	No	Poor	No	No	No
37	Female	25	Low	Normal	No	No	No	High	No	Poor	No	Yes	No
38	Female	28	Low	Normal	No	No	Yes	High	No	Normal	No	No	No
39	Female	26	Low	Normal	Yes	No	No	Normal	No	Normal	Yes	Yes	Yes
40	Male	28	Low	Low	No	No	No	Normal	No	Normal	No	No	No
41	Female	45	High	High	No	No	Yes	Low	Yes	Normal	Yes	Yes	Yes
42	Male	63	Low	Low	No	No	Yes	High	Yes	Good	No	No	No
43	Female	55	High	High	No	No	No	Normal	Yes	Normal	Yes	Yes	Yes
44	Female	44	Low	Low	No	No	No	Normal	No	Normal	No	No	No
45	Male	35	Low	Low	No	No	Yes	High	No	Normal	No	No	No
46	Female	42	Normal	Normal	No	No	Yes	High	Yes	Good	No	No	No
47	Female	43	Normal	Normal	No	No	No	Low	No	Poor	Yes	No	No
48	Male	65	Normal	Normal	No	No	Yes	Low	No	Normal	Yes	Yes	Yes
49	Male	74	Normal	Normal	No	No	Yes	Normal	Yes	Normal	Yes	Yes	Yes
50	Male	36	Normal	Normal	No	No	No	Normal	No	Poor	No	No	No

.

REFERENCES

1. W.H. Organization, The top 10 causes of death, May 2014 (Online). Available: http://www.who.int/.
2. S.Amin, K. Agarwal and R. Beg, Genetic neural network based data mining in prediction of heart diseases using risk factors, in IEEE Conference on Information and Communication Technologies (ICT) 2013.
3. J. Thomas and T. Princy, Human Heart Disease Prediction System using Data Mining Techniques, in International Conference on Circuit, Power and Computing Technologies, 2016.
4. M. Jabbar, B. Deekshatulua and P. Chandra, Classification of heart disease using K-nearest neighbor and, in International Conference on Computational Intelligence: Modeling Techniques and Applications, 2013.
5. Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.
6. Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). TAPPING THE. Communications of the ACM, 49(9), 77.
7. ShantakumarB.Patil, Y.S.Kumaraswamy,” Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network”, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656© Euro Journals Publishing, Inc. 2009.
8. Thuy Nguyen Thi Thu, Darryl.N. Davis,”A Clustering Algorithm for Predicting cardio vascular risk”, Computer Science Department, Hull University, Cottingham Road, Hull, UK
9. N. Elfadil and A. Hossen, "Identification of Patients With Congestive Heart Failure Using Different Neural Networks Approaches", Journal Technology and Health Core, vol. 17 Issue 4, December 2009.
10. Centre for Disease Control and Prevention, http://www.cdc.gov/heartdisease/risk_factors.htm.
11. American Heart Association, http://www.heart.org/HEARTORG/Conditions
12. D. Isern, D. Sanchez, and A. Moreno, "Agents Applied in Health Care: A Review", International Journal of Medical Informatics, 79(3), pp.146-166, doi:10.1016/j.ijmedinf201O.01.003, 2010.
13. J. G. Yang, S. Y. Weng,, *Applied Textbook of Artificial Neural Network*, Zhejiang University Press, Hangzhou, 2001.
14. Y. J. Lei, and X. W. Zhang, *Genetic Algorithm Toolbox of MatLab and its Application*, Xian University of Electronic Science and Technology Press, 2005.
15. J. Guo, and W. J. Sun, *Theory of Neural Network and its Implementation with MatLab*, Electronic Industry Press, Beijing, 2005.
16. American Heart Association, *http://www.heart.org/HEARTORG/Conditions*