Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors

Syed Umar Amin[1], Kavita Agarwal[2], Dr. Rizwan Beg[3],
[1,2,3]Department of Computer Science & Engineering
Integral University, Lucknow, India
syed.umar.amin@gmail.com[1], kavitalucknow@gmail.com[2], rizwanbeg@gmail.com[3]

***Abstract-* Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy. These techniques have been very effective in designing clinical support systems because of their ability to discover hidden patterns and relationships in medical data. One of the most important applications of such systems is in diagnosis of heart diseases because it is one of the leading causes of deaths all over the world. Almost all systems that predict heart diseases use clinical dataset having parameters and inputs from complex tests conducted in labs. None of the system predicts heart diseases based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, obesity or physical inactivity, etc. Heart disease patients have lot of these visible risk factors in common which can be used very effectively for diagnosis. System based on such risk factors would not only help medical professionals but it would give patients a warning about the probable presence of heart disease even before he visits a hospital or goes for costly medical checkups. Hence this paper presents a technique for prediction of heart disease using major risk factors. This technique involves two most successful data mining tools, neural networks and genetic algorithms. The hybrid system implemented uses the global optimization advantage of genetic algorithm 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%.**

Keywords- data mining, heart disease risk factors, prediction and diagnosis systems.

I. INTRODUCTION

Heart diseases are the number one cause of death globally: more people die annually from Heart diseases than from any other cause. An estimated 17.3 million people died from Heart diseases in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke [1]. Recent research in the field of medicine has been able to identify risk factors that may contribute toward the development of heart disease but more research is needed to use this knowledge in reducing the occurrence of heart diseases. Diabetes, hypertension, and high blood cholesterol have been established as the major risk factors of heart diseases. Life style risk factors which include eating habits, physical inactivity, smoking, alcohol intake, obesity are also associated with the major heart disease risk factors and heart disease [2,3].There are studies showing that reducing these risk factors for heart disease can actually help in preventing heart diseases [4]. There are many studies and researches on the prevention of heart disease risk. Data from studies of population has helped in prediction of heart diseases, based on blood pressure, smoking habit, cholesterol and blood pressure levels, diabetes. Researchers have used these prediction algorithms in adapted form of simplified score sheets that allow patients to calculate the risk of heart diseases [6]. The Framingham Risk Score (FRS) is a popular risk prediction criterion which is used in algorithms for heart disease prediction [7].

This study aimed at developing an intelligent data mining system based on genetic algorithm optimized neural networks for the prediction of heart disease based on risk factors' categories. The system was implemented using MATLAB R2012a.

II. DATA MINING TECHNIQUES

Data mining techniques are used to explore, analyze and extract medical data using complex algorithms in order to discover unknown patterns. Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease [8], diabetes [9], stroke [10] and cancer [11] and many data mining techniques have been used in the diagnosis of heart disease with good accuracy. Researchers have been applying different data mining techniques such as naïve bayes, neural network, decision tree, bagging, kernel density, and support vector machine for prediction and diagnosis of heart diseases [13]-[15]. One of the systems [16] uses neural based learning classifier for classifying data mining tasks showed that neural based learning classifier system performs equivalently to supervise learning classifier. IEHPS [17] intelligent and effective heart attack prediction system was built using data mining and neural networks and it proposed extracting significant patterns for heart disease prediction using K-means clustering and used MAFIA algorithm to mine the frequent patterns. Polatet al., developed system using hybrid fuzzy and k-nearest neighbour approach for the prediction of heart disease, which had 87% accuracy in diagnosis [18]. In another system [19] neural network ensemble was used in the diagnosis of heart disease with an accuracy of 89.01%. Latha and Subramanian (2007), proposed an intelligent heart disease prediction system using CANFIS and genetic algorithm which had a very low mean square error [20].Analyzing the different techniques discussed, this paper proposes a novel system using genetic algorithm and neural

978-1-4673-5758-6/13/$31.00 © 2013 IEEE

network for predicting the risk of heart diseases. Genetic algorithm is used to optimize neural network weights. What is even more different in this paper is that it is the first time that such a hybrid technique is applied on risk factors for the accurate prediction of heart disease. Hence the main objective is not only to use this system in clinical decision support but to also use this system as risk indicator so that it helps people reduce the risks of having any heart disease in future.

III. Materials And Methods

A. Data Analysis and Encoding

The problem with risk factors related to heart disease is that there are many risk factors 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 [21]. Hence the risk factors were analyzed from various sources [22]-[23]. The dataset was composed of 12 important risk factors 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 [23]. Most of the heart disease patients had many similarities in the risk factors [24]. The TABLE I below shows the identified important risk factors 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.

B. Neural Network Weight Optimization by Genetic Algorithm

This system uses backpropagation algorithm for learning and training the neural network, but there are two major disadvantages with backpropagation 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 backpropagation algorithm is very slow in convergence and there is a possibility that network never converges [25]. This problem of local optimum solution can be solved by optimizing the initial weights of neural network. For this we use a genetic algorithm which is specialized for global searching [26]. For this we first determine the number of inputs, layers and hidden neurons of the neural network and then we would use the backpropagation algorithm to train the networks using the weights optimized by GA.

C. Neural Network Architecture

A multilayered feed-forward network is used having 12 input nodes 10 hidden nodes and 2 output nodes. The number of inputs is based on the final set of risk factors 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 genetic algorithm for optimization according to the fitness function. Once the weights are optimized, the Levenberg-Marquardt backpropagation 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 backpropagation 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.

TABLE II
PATIENT'S CASE STUDY DATA IN ENCODED FORM

No	Sex	Age	Blood Cholesterol	Blood Pressure	Hereditary	Smoking	Alcohol 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	Yes	Yes
24	Female	27	Low	Normal	No	No	No	High	No	Good	No	No	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	Yes	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	No	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	Yes	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	Yes	No	Yes	Low	No	Poor	No	Yes	No
37	Female	25	Low	Normal	No	No	No	High	No	Poor	No	No	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	No	Yes

40	Male	28	Low	Normal	No	No	No	Normal	No	Poor	No	No	No
41	Female	45	High	Normal	No	No	Yes	Low	Yes	Poor	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	Normal	No	No	No	Normal	No	Normal	No	No	No
45	Male	35	Low	Normal	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	High	No	Yes	Yes	Normal	Yes	Normal	Yes	Yes	Yes
50	Male	36	Normal	High	No	Yes	No	Normal	No	Poor	No	No	No

IV. Parameter Settings

The system was developed using MATLAB R2012a. Global Optimization Toolbox and the Neural Network Toolbox were used for implementing the algorithm [27]. The data for risk factors 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 genetic algorithm 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:

$$mse = \sum_k (O_k - T_k)^2 / n$$

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
PopulationSize	20
Generations	100
CrossoverFraction	0.8000
MigrationInterval	20
MigrationFraction:	0.2000
EliteCount	2
TolFun	1.0000e-006

V. Results And Discussion

The input data consisted of risk factors collected from 50 people through case studies provided at the website of the American Heart Association [23]. 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 genetic algorithm and neural network approach gives better average prediction accuracy than the traditional ANN.

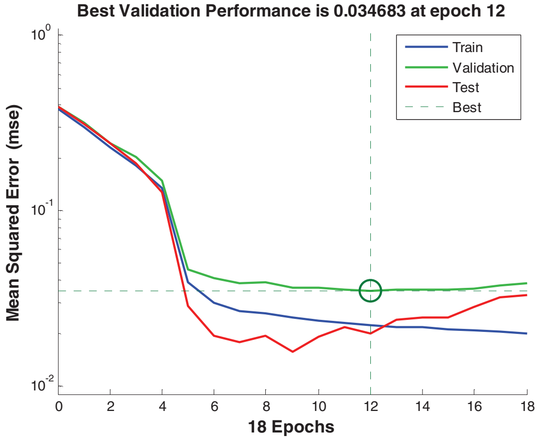

Figure 1: Performance Graph

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	-

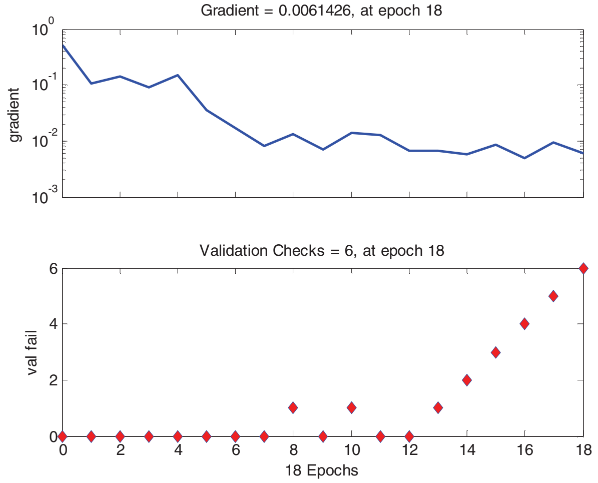

Figure 2: Training State Graph

VI. 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 Genetic Algorithm to optimize the connection weights of ANN so as to improve the performance of the Artificial Neural Network. The system uses identified important risk factors for the prediction of heart disease and it does not require costly medical tests. Risk factors 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 risk factors 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.

REFERENCES

[1] "Global atlas on cardiovascular disease prevention and control", WHO, 2011.

[2] Mozaffarian D, Wilson PW, Kannel WB, Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. Circulation 117: 3031–3038, 2008.

[3] Poirier P, Healthy lifestyle: even if you are doing everything right, extra weight carries an excess risk of acute coronary events. Circulation 117: 3057–3059, 2008.

[4] Wood D, De Backer, Prevention of coronary heart disease in clinical practice: recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention. Atherosclerosis 140: 199–270, 1998.

[5] Anderson KM, Odell PM, Cardiovascular disease risk profiles. Am Heart J 121: 293–298. 1991.

[6] Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J., 121: 293–298, 1991.

[7] Kannel WB, An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 110: 281–290, 1979.

[8] R. Das, I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications, Elsevier, pp. 7675–7680, 2009.

[9] T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach," Americas Conference on Information Systems, 2009.

[10] S. Panzarasa et al, "Data mining techniques for analyzing stroke care processes," in Proc. of the 13th World Congress on Medical Informatics, 2010.

[11] L. Li, T. H., Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, Elsevier, 2004.

[12] V. A. Sitar-Taut et al., "Using machine learning algorithms in cardiovascular disease risk evaluation," Journal of Applied Computer Science and Mathematics, 2009.

[13] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 250-255, 2010.

[14] H. Yan, et al., "Development of a decision support system for heart disease diagnosis using multilayer perceptron," in Proc. of the 2003 International Symposium on, vol. 5, pp. V-709- V-712.

[15] M. C. Tu, D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," Biomedical Engineering and Informatics, IEEE, 2009.

[16] 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.

[17] 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.

[18] Polat , K., S. Sahan, and S. Gunes, Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour based weighting preprocessing. Expert Systems with Applications 2007. 32 p. 625–631.

[19] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.

[20] 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.

[21] 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.

[22] Centre for Disease Control and Prevention, http://www.cdc.gov/heartdisease/risk_factors.htm.

[23] American Heart Association, http://www.heart.org/HEARTORG/Conditions

[24] 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.

[25] J. G. Yang, S. Y. Weng,, *Applied Textbook of Artificial Neural Network*, Zhejiang University Press, Hangzhou, 2001.

[26] Y. J. Lei, and X. W. Zhang, *Genetic Algorithm Toolbox of MatLab and its Application*, Xian University of Electronic Science and Technology Press, 2005.

[27] J. Guo, and W. J. Sun, *Theory of Neural Network and its Implementation with MatLab*, Electronic Industry Press, Beijing, 2005.