A NOVEL PROBABILISTIC ARTIFICIAL NEURAL NETWORKS APPROACH FOR DIAGNOSING HEART DISEASE

Sadri Sa'di[1], Ramin Hashemi[2], Arman Abdollapour[3], Kamal Chalabi[1] and Mohammad Amin Salamat[1]

[1]Department of Computer Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran
[2] Executive Master of Business Administration (EMBA) Student, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
[3]Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

ABSTRACT

In today's world one of the most common diseases are heart disease which its mortality and disability is high. Therefore, heart disease is one of the biggest health problems in the world. Since the diagnosis of heart disease in people is very important, a method should be used in the right diagnosis of heart diseases that have the least errors in heart disease diagnosis. For this reason, in this paper, Probabilistic Neural Networks (PNNs) for the diagnosis of heart disease from a dataset that includes 303 samples from different patients is used. In this paper, we have implemented PNN in the MATLAB environment. As well as, the efficiency criteria in this paper is to maximize accuracy of heart disease diagnosis in the process of training and testing. According to the Cleveland dataset which contains 303 samples, we found that the accuracy of training and test accuracy are 87% and 75% respectively.

KEYWORDS

Artificial Neural Networks (ANNs), Probabilistic Neural Network (PNN), training accuracy, test accuracy, heart disease.

1. INTRODUCTION

Heart disease is the leading cause of death in both men and women. Moreover, heart attack arises from the death of heart muscle cells caused by the reduction or stoppage of blood flow in the heart arteries. Although it often occurs in people over the age of 40 years, but can occur in any age groups [1]. According to the report of World Health Organization, Heart disease among other diseases is the most common cause of death. Diagnosis of the disease can be difficult due to non-specific symptoms or other common symptoms with the other illnesses. Human life depends on effective work of heart [1, 2]. The important factors that increase the risk of heart disease include smoking, high blood pressure, high blood cholesterol, family health history, physical inactivity, obesity, etc. [1, 2]. Signs and symptoms of heart disease are the most important diagnostic tools for medics. Therefore, the diagnosis of heart disease needs high experience and skills. However, early diagnosis and specific medical care of heart patients can greatly prevent sudden death in these patients and reduce the high costs of surgery and other treatment courses [1, 2, 3, 4]. Up to now, various methods have been proposed to solve this problem. One of these methods is ANNs. In the recent decades the cooperation between engineers and medics equipped medical science

DOI:10.5121/ijfcst.2015.5605

with the latest intelligent tools which with using of them better diagnosis and treatments is provide to patients. This automatically reduces medical errors, costs, and life damage [1, 2]. ANNs have an important role in the medical field to solve health problems and correct diagnosis of diseases. ANNs have different models which with using these models complex medical processes can be designed as software and be implemented. Then, these software systems can be used to increase the accuracy of disease diagnosis [1, 2]. That is why up to now ANNs has been used in diagnosis of various diseases such as heart disease [5, 6], diabetes disease [7], thyroid disease [8], and etc. ANNs are able to learn like humans. An ANN is set for specific tasks such as pattern recognition and classification of information during a learning process [1, 2, 9, 10]. The learning through experience and generalizability ability to solve new problems will lead this approach over other strategies [1, 2, 9, 10]. There are different models of ANNs, including PNN [11], Radial bases function (RBF) [10, 11], Generalized regression neural network (GRNN) [2, 11], Multi-layer perceptron (MLP) [9, 11] and any of these models can be used to diagnose heart disease.

PNN was the model used in this article to diagnose heart disease. In this article we have tried to raise the accuracy of heart disease diagnosis in a dataset via using PNN to achieve a maximum optimized response. The important characteristic in using ANNs is applying standard data for diagnosis. In this article, a data set named Cleveland containing 303 data sample was used. This data was used for training PNN as well as in this article PNN model is implemented in MATLAB environment.

In the second part, the previous works in the field of diagnosis of heart disease with the models of ANNs will be highlighted. In the third part of the article, we will talk about Cleveland data set, PNN, the proposed method, the result of training and testing of data on the Cleveland data set which was used for the diagnosis of disease. Conclusions and future work will be discussed in Part Four.

2. RELATED WORKS

Significant researches have been carried out in the field of heart disease diagnosis with ANN models.

Soleimanian and colleagues [6] performed diagnosis of heart disease with a dataset, which included 40 patients with 6 features of gender, name, age, blood pressure, smoking, and heart failure. The ANN model used is MLP. MLP has three input, hidden and output layers which the layers are connected through neurons and connection lines between them. The used neurons in input layer are equal to the above mentioned 6 features and the hidden layer neurons are equal to 4 features. These 4 neurons are obtained through experience and test to come to a better and accurate predict. The neurons in output layer are equal to the number of classifications which digit 2 is selected because the object vector has two members. It means that in data separation “yes” is for those who has heart failure and “no” is for those who don’t have any heart disease. Finally, the prediction accuracy under the mentioned conditions is 95% during training phase and 85% in the testing phase.

Researchers in [3] has used techniques such as MLP of ANNs, Decision Trees and Support Vector Machine (SVM) to predict coronary heart disease in patients. Dataset used for these techniques includes 12 features, such as high blood pressure, smoking, diabetes, gender, age and etc. The Dataset includes 1000 sample. The accuracy of coronary heart disease prediction in SVM, MLP and Decision Trees are 92.1%, 91% and 89.6%, respectively. Which among them SVM has more accuracy to predict the disease.

In another research, Zebardast and colleagues [2] used GRNN model for heart disease diagnosis in which 4 Datasets were used for diagnosis with identical characteristics and different numbers of samples. These dataset include 14 features such as age, gender, blood pressure, cholesterol, diabetes, the number of smoking in a day, and etc. After normalizing data, four dataset used for training and testing was grouped that 90% of them is used for training and 10% of them were used for the experiment. The result of this training and testing of dataset showed that the accuracy in the training phase of all dataset is between 97% to 100%, and accuracy in the testing phase of each dataset were between 75% to 96%.

In Reference [12] techniques such as Decision Trees, Naive Bayes and MLP of ANNs is used to predict heart disease in people. In this study, two dataset including 303 and 270 samples with 15 attributes for each dataset is used. The performance of each of the used methods is compared based on the prediction accuracy measure. Accordingly, the accuracy of Naive Bayes, Decision Trees and MLP is equal to 90.74%, 100% and 99.62%, respectively. As a result it was found that MLP has the perfect accuracy rate to predict disease.

Researchers in [13] have used the GRNN and RBF of ANNs models for heart disease diagnosis. In this study, 300 samples have been used for training and after the training it has been find that RBF has a better performance for heart disease diagnosis in comparison to GRNN.

3. THE PROPOSED MODEL FOR HEART DISEASE DIAGNOSIS

ANNs are used for parallel processing of information. ANNs are formed from neurons and connection lines between neurons. A neuron is the smallest unit of information processing that forms the basis of ANNs performance. In the mathematical model proposed for a neuron it is tried to consider the main features of a normal neuron. Neurons are interconnected processing elements that work together to solve a problem [1, 2, 8, 9, 10, 11].Today, like people, learning in ANNs has lead to the application of ANNs in every industry. In order to learn, one of the most important parts related to the creation and development of ANNs the data selection [1, 2, 8, 9, 10, 11]. In this article, the data selected for heart disease diagnosis were data collected from Cleveland hospital [14] which includes 303 samples and each sample has 14 features. In this article the data is used for training and testing of PNN which the 14 features in this article have shown in table 1.

Table 1. Features used for heart disease diagnosis with their values [2, 14]

No. of Feature	Feature	Descriptions and Feature values
1	Age	Numerical values
2	Sex	Male=1 Female=0
3	Chest pain type	Typical angina=1 Atypical angina=2 Non-angina pain=3 Asymptomatic=4
4	Resting blood pressure	Numerical values in mm hg
5	Serum cholesterol	Numerical values in mm/dl
6	Fasting blood sugar	Fasting blood sugar > 120 mg/dl (True=1; False=0)
7	Resting electrographic results	Normal=0 Having ST-T wave abnormality=1

		Left ventricular hypertrophy=2
8	Maximum heart rate achieved	Numerical values
9	Exercise induced angina	Yes=1 No=0
10	ST depression induced by exercise relative to rest	Numerical values
11	Slope of the peak exercise ST segment	Up sloping=1 Flat=2 Down sloping=3
12	Number of major vessels colored by fluoroscopy	Value = 0-3
13	Defect type	Normal=3 Fixed defect=6 Reversible defect=7
14	Diagnosis of heart disease	Abnormal=1 Normal=0

According to Table 1, the first 13 features are used as PNN inputs and the 14th feature which is used to determine whether the individual is healthy or sick is used as the output features. The values of 14th feature as represented in Table 1 are in the form of 0 and 1. A zero means that the individual is healthy and 1 means that the individual is sick.

As it was stated, the proposed model in this paper for heart disease diagnosis in PNN of ANNs. PNN is used for classification issues [11]. In fact, these types of ANNs are also another form of RBF. PNN has three layers including an input layer, a layer called the radial basis layer (a hidden layer) and a competitive layer (an output layer). The radial basis layer used the Gaussian transfer function [11]. The number of neurons in this layer is equal to the number of training data set. This layer calculates the distance of input vector from training vector. Accordingly, the operation provides a vector that its elements determine the distance between the input and training input. Also, the hidden layer produces a possibility vector as the output of network. Finally, this layer selects the maximum possible value of probability from the probability vector and generates 1 for that and 0 for the rest of the probabilities [11]. Formulas (1, 2) show the way of Gaussian function [11] calculation for each hidden layer neuron.

$$F_{k,i}(X) = \frac{1}{(2\pi\sigma^2)} h_i \qquad (1)$$

$$h_i = \exp\left(-\frac{\left\| X - X_{k,i} \right\|^2}{2\sigma^2} \right) \qquad (2)$$

According to formulas (1, 2), h_i is Gaussian function and X is input vector (x_1, x_2, ..., x_n) which is composed of n variables. Hidden layer neurons are divided in to k groups, each k group includes a class. i is the available neuron in group k which is calculated by Gaussian function. σ is a fixed amount that increases the accuracy of the training and testing and is used for hidden layer neurons [11]. Sum of each of neurons in one of group k is calculated by the formula (3) [11].

$$G_k(X) = \sum_{i=1}^{M_k} W_{ki} F_{k,i}(x), \quad K \in \{1,...,K\}, \qquad (3)$$

According to the formula (3), M_K is the number of neurons in the pattern of a class. W_{ki} is weights coefficients and that G_k (X) generates 1 output for 't' the maximum possible value of probabilities [11]. The data normalization value is often used instead of data values. Employing methods to put input data in a limited range is called normal values which are one of the pre-processing methods. Other advantage of putting inputs in a limited range is to inhibit the excessive growth of the weight. This limitation reduces the convergence time in ANN and minimizes achievable error. In other words, entering data in raw form reduces the speed and accuracy of the ANN, to prevent this, the data must be normalized before training. Different methods were tested to normalize the data, finally using formula (4) the data were normalized in the range [-1, 0] to increase the training performance [2, 15, 16, 17].

$$Input = \frac{Input - Max_{Input}}{Max_{Input} - Min_{Input}} \qquad (4)$$

In this paper, according to the Cleveland data set that contains 303 data samples, the number of data which were selected for training PNN is 80% i.e. 242 samples and the number of data which were selected for testing PNN is 20% i.e. 61 samples. If PNN answers correctly during testing the training work is completed. Otherwise, the training of network begins again. Finally, when the training phase of program is finished and generated the right outputs on all the inputs in this case it is determined that PNN weights are adjusted correctly. So after this, these values will be used for the diagnosis.

According to the proposed model, Figure 1 shows the used architecture in PNN of ANNs for heart disease diagnosis.

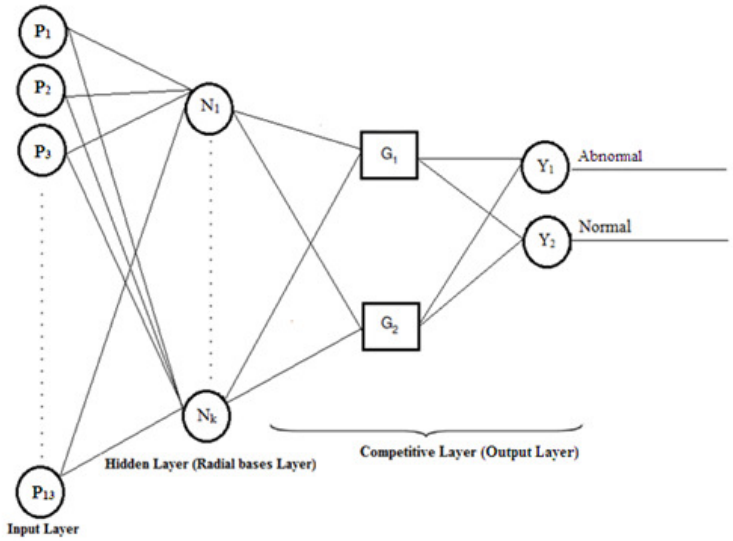

Figure 1. The used architecture in PNN for heart disease diagnosis

According to Figure 1, the PNN has an input layer, a hidden layer (radial basis layer) and a competitive layer (an output layer). The number of neurons in the input layer features is equal to 13 of above mentioned features (P_1 to P_{13}), the number of neurons in the hidden layer is equal to

NK (N_1- N_K) which equals to the number of training data set. The number of neurons in the output layer equals to the defined classes that is 2 neurons i.e. 1 for existing of a disease and 0 for the lack of disease.

The implemented structure is based on the newpnn function which is used to build ANNs. Also in this network we have determined the amount of σ as 0.3. Of course, the determination of this amount is quite experimental which in terms of this amount the accuracy of training and testing will be increased. According to the proposed model of PNN, the amount of accuracy for training and testing of Cleveland data set that includes 303 samples is shown. Figure 2 shows the accuracy of training and testing for heart disease diagnosis by PNN.

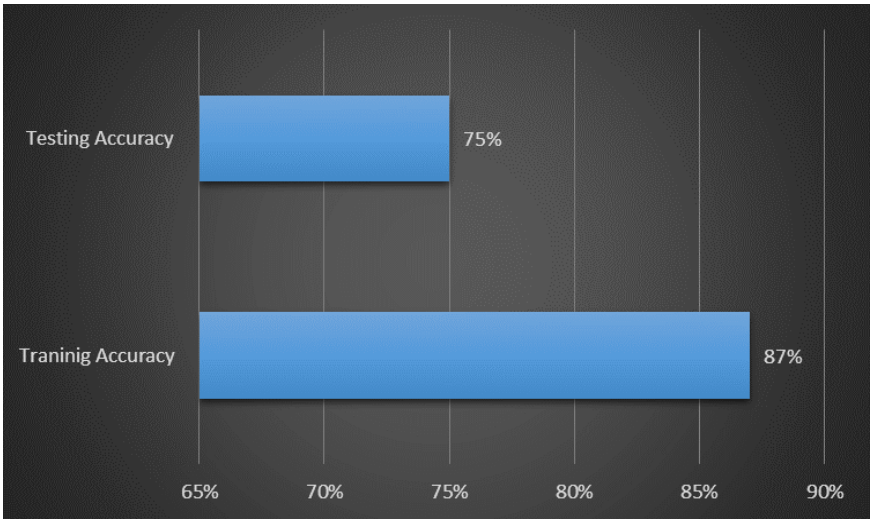

Figure 2. The accuracy rate of training and testing of Cleveland data sets

As Figure 2 makes clear the accuracy rate of training for 80% of data (242 samples) from 303 samples of Cleveland data set is 87% and the accuracy rate of testing for 20% (61 samples) from 303 samples of Cleveland dataset is 75%, respectively. Therefore, the training data is more accurate than the testing data. In heart disease diagnosis, both training and testing of Cleveland data set via using PNN have acted with a good accuracy.

4. CONCLUSION AND FUTURE WORKS

In recent decades, ANNs have played an important role in medical science whose main role was diagnosis. Therefore, in this article, the PNN is used for heart disease diagnosis. In order to diagnose this disease, Cleveland data set which includes 303 samples is used. Each sample includes 14 features. For training, 242 samples and for testing 61 samples were selected from among these 303 samples. The accuracy of heart disease diagnosis in training was 87% and the accuracy of testing was 75%. As is apparent, the accuracy of diagnosis, especially at the stage of training was high. The procedure was that the first 13 features of Table 1 were used as input of PNN and 14th features was used as the output which contains Zero (healthy) and one (sick). The future researches can be include the diagnosis of heart disease with other models of ANNs or with a combination of ANNs and fuzzy method for heart disease diagnosis in order to determine which methods is the best way to diagnose this disease.

REFERENCES

[1] B. Zebardast, R. Rashidi, T. Hasanpour, F. S. Gharehchopogh, "Artificial neural network models for diagnosing heart disease: a brief review", International Journal of Academic Research, Vol.6, Issue 3, pp.73-78, 2014.

[2] B. Zebardast, A. Ghaffari, M. Masdari, "A New Generalized Regression Artificial Neural Networks Approach for Diagnosing Heart Disease", International Journal of Innovation and Applied Studies, Vol. 4, No. 4, pp. 679-689, 2013.

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

[4] J.L. Patel, R.K. Goyal, "Applications of artificial neural networks in medical science", Curr. Clin. Pharmacol., 2(3), pp. 217-226, 2007.

[5] N. Ajam, "Heart Diseases Diagnoses using Artificial Neural Network", Network and Complex Systems,Vol.5, No.4, pp.7-10, 2015.

[6] F.S. Gharehchopogh, Z.A. Khalifelu, "Neural Network Application in Diagnosis of Patient: A Case Study", IEEE, International Conference on Computer Networks and Information Technology(ICCNIT 2011), Abbottabad, Pakistan, 11-13 July 2011, pp. 245-249, 2011.

[7] M. Pradhan, R. K. Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies, Vol. 2, Issue 2, April 2011.

[8] F. S. Gharehchopogh, M. Molany, F. D. Mokri, "Using artificial neural network in diagnosis of thyroid disease: a case study", International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.4, pp. 49-61, August 2013.

[9] B. Zebardast, I. Maleki, A. Maroufi, "A Novel Multilayer Perceptron Artificial Neural Network based Recognition for Kurdish Manuscript", Indian Journal of Science and Technology, Vol 7(3), pp.343– 351, March 2014.

[10] B. Zebardast, I. Maleki, "A New Radial Basis Function Artificial Neural Network based Recognition for Kurdish Manuscript", International Journal of Applied Evolutionary Computation, Vol. 4(4), pp.72-87, 2013.

[11] H. Demuth, M. Beale, "Neural Network Toolbox for Use with MATLAB", User's Guide, Version 4, The MathWorks, Inc. 3 Apple Hill Drive Natick, MA 01760-2098,840 pages, 2002.

[12] C.S. Dangare, S.S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, Volume 47– No.10, pp. 44-48, 2012.

[13] S. A. Hannan, R. R. Manza, R.J. Ramteke, "Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis", International Journal of Computer Applications, Vol. 7, No.13, pp. 7-13, 2010.

[14] Cleveland Clinic Foundation, Available: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease. [Last Accessed 20 September 2015].

[15] X. Tao, H. E. Michel, "Novel artificial neural networks for remote-sensing data classification", Proc. SPIE 5781, Optics and Photonics in Global Homeland Security, Vol. 5781, pp. 127—138, 2005.

[16] Q. Song, N. Kasabov, "TWRBF - Transductive RBF Neural Network with Weighted Data Normalization", 11th International Conference Neural Information Processing (ICONIP 2004), Calcutta, India, November 22-25, pp. 633-640, 2004.

[17] T. Jayalakshmi, A. Santhakumaran, "Statistical Normalization and Back Propagation for Classification", International Journal of Computer Theory and Engineering, Vol.3, No.1, February, pp. 89-93, 2011.