www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 3 March 2013 Page No. 781-788

Predictive Model Of Stroke Disease Using Hybrid Neuro-Genetic Approach

[1]K.Priya, [2] T.Manju,[3]R.Chitra
[1, 2] PG Student, Computer Science Department,
[1, 2]Noorul Islam Centre For Higher Education, Tamil Nadu, India
[3]Associate professor, Computer Science Department,
[3]Noorul Islam Centre For Higher Education, Tamil Nadu, India
E-mail: priyakpr.k@gmail.com, ermanjucs@gmail.com, jesi_chit@yahoo.co.in

Abstract- Stroke is a major life threatening disease to cause of death and it has a serious long term disability. The time taken to recover from stroke disease depends on patient's severity. Number of work has been carried out for predicting various diseases by comparing the performance of predictive data mining. In the proposed work MLFFNN with back propagation algorithm is used. The number of hidden neurons is optimized by genetic algorithm. This work demonstrates about ANN based prediction of stroke disease by improving the accuracy with higher consistent rat using optimized hidden neurons. In this algorithm determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease. The data is collected from 300 patients. Among that 180 patients having disease .In the proposed work 196 data are used for training and 104 data is used to test the performance of the system.

Keyword: Artificial Neural Network, Genetic Algorithm, Multi Layer Perception, Hybrid Neuro Genetic, Stroke disease.

I INTRODUCTION

In today's world data mining plays a vital role for prediction of diseases in medical industry. In medical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of same category. The physicians may not able to diagnose it correctly. So it is necessary to identify the important diagnostic features of a disease and this may facilitate the physicians to diagnosis the disease early and correctly.

Stroke is a life threatening disease that has been ranked third leading cause of death in states and in developing countries [12]. Stroke patients are often cared for by neurologists, because of the complex nature of the symptoms caused by damage to the brain. However, strokes are very closely related to heart disease.

The following signs may mean that had a stroke: sudden weakness or numbness of the face, arm, or leg on one side of the body; sudden confusion, trouble talking, or trouble understanding; sudden dizziness, loss of balance, or trouble walking; sudden trouble seeing out of one or both eyes or sudden double vision; sudden severe headache[11].

K.Srinivas et al. [4] to examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naive Bayes and Artificial Neural Network to massive volume of healthcare data. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. Data mining is an essential step of knowledge discovery.

A.Sudha et al. [1] to propose the classification algorithm like Naive bayes, Decision tree and Neural Network for predicting the stroke diseases. In this existing system a predictive model for cerebrovascular disease using data mining utilized the data mining techniques. It adopted the classification algorithm like decision trees, Bayesian classifier and back propagation neural network. In this work it consists of attributes containing patient's medical history and symptoms. The records with irrelevant data were removed from data warehouse before mining process occurs.

D.Shanthi, et al. [7] proposed to functional model of ANN to aid existing diagnosis methods. The Back propagation algorithm was used to train the ANN architecture and the same had been tested for the various categories of stroke disease. The data for this study had been collected from 50 patients who had symptoms of stroke disease. The data had standardized so as to be error free in nature. All the fifty cases were analyzed after careful scrutiny with the help of the Physicians. Data were analyzed in the dataset to define column parameters and data anomalies

P.K. Anooj [3] to predict the heart disease using Weighted Fuzzy Rules were proposed. Clinical decision support system which uses knowledge from medical experts and transfers this knowledge into computer algorithms manually. To handle this problem, machine learning techniques had been developed to gain knowledge automatically from examples or raw data. Here, a weighted fuzzy rule-based clinical decision support system (CDSS) was presented for the diagnosis of heart disease, automatically obtaining knowledge from the patient's clinical data. The proposed clinical decision support system for the risk prediction of heart patients consists of two phases: (1) automated approach for the generation of weighted fuzzy rules and (2) developing a fuzzy rule-based decision support system.

Obi J.C. and Imainvan A.A [2] Neuro Fuzzy Logic procedure for the medical diagnosis of Alzheimer employed by physician was proposed. Alzheimer Disease (AD) is a form of dementia; it is a progressive, degenerative disease. Alzheimer is a brain disease that causes problems with memory, thinking and behavior. It is severe enough to interfere with daily activities. The proposed system was a useful decision support approach for the diagnosis of Alzheimer. Neural network (NN) consists of an interconnected group of neurons. Artificial Neural Network (ANN) was made up of interconnecting artificial neurons.

D.Shanthi, et al. [7] proposed to functional model of ANN to aid existing diagnosis methods. The Back propagation algorithm was used to train the ANN architecture and the same had been tested for the various categories of stroke disease. The data for this study had been collected from 50 patients who had symptoms of stroke disease. The data had standardized so as to be error free in nature. All the fifty cases were analyzed after careful scrutiny with the help of the Physicians. Data were analyzed in the dataset to define column parameters and data anomalies. Data analysis information needed for correct data preprocessing.

II. DISEASE DATA SOURCE

The data for this work have been collected from patients who have symptoms of stroke disease. The stroke data set is collected from UCI repository. Each data set consists of 14 attributes, in the data set Cleveland taken the stroke disease. This is publicly available dataset in the Internet. All the symptoms are analyzed carefully for the prediction of stroke. In those attributes totally consist of 14 attributes such as age, sex, hypertension, iabetes Mellitus, Smoking, High blood cholesterol, Alcohol abuse, Headache, Vomiting, Loss of Consciousness, Transient ischemic attack, Atrial fibrillation, etc

Sl No	Features	Description
1.	age	Age in Years

2.	Sex	Male=1, Female=0
3.	Hypertension	Yes=1,no=0
4.	High cholesterol	Severe=1, normal=0
5.	Smoking	Yes=1,no=0
6.	High blood pressure	Severe=1, normal=0
7.	Alcohol abuse	Yes=1, no=0
8.	Headache	Yes=1,no=0
9.	Vomiting	Yes=1,no=0
10.	Loss of Consciousness	Coma=1, alert=0
11.	Diabetes Mellitus	Value1:>120 mg/dl; value0<120 mg/dl
12.	Transient ischemic attack,	present=1, Absent=0
13.	Atrial fibrillation	Present=1,absent=0
14.	Hand / Leg numbness	Present=1,absent=0

Table1: Input Parameter

III. Multi layer Feed Forward Network

The GA-based MLFNN was mainly developed to search for the optimal training parameters, i.e. the number of neurons in the hidden layer, the learning rate, the momentum rate, the transfer function in the hidden layer and the learning algorithm. In this project it takes the 14 input layers, 6 hidden layers and two outputs.

3.1 Back Propagation Network

BP algorithm is a supervised learning method, which it is the most widely used algorithm for training MLP neural network. The idea of the BP is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the learning error will be at minimal. ANN nodes in BP algorithm are organized in layers, send their signals forward and then the learning error (difference between actual and expected results) is calculated and propagated backwards until met satisfactory learning error.

Figure 1 show the interconnection between nodes which is usually referred to as a fully connected network or multilayer perceptron (MLP). Multilayer architecture means that the network has several layers or nodes usually referred to as input layer, hidden layer and output layer. MLP network can be used with great success to solve both classification and function approximation problems.

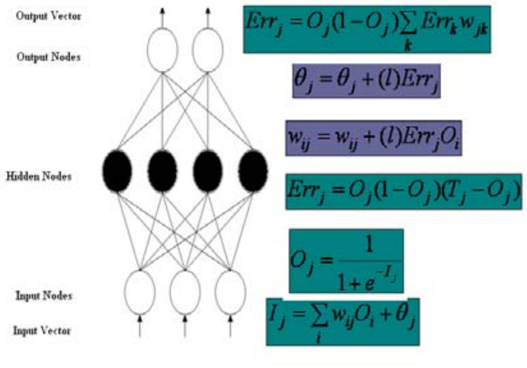
Output Vector
Output Nodes
Hidden Nodes
Input Nodes
Input Vector
$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$
$\theta_j = \theta_j + (l)Err_j$
$w_{ij} = w_{ij} + (l)Err_j O_i$
$Err_j = O_j(1-O_j)(T_j - O_j)$
$O_j = \frac{1}{1+e^{-I_j}}$
$I_j = \sum_i w_{ij} O_i + \theta_j$

Figure 1: Multi Layer Perceptron Structure

There are two phases in BP learning algorithm which are feed forward phase and backward phase. In feed forward process, the dataset is presented to the input layer and the network propagates the input pattern from layer to layer until the output pattern is generated. The output is obtained from a summation of the weighted input of a node and maps to the network activation function. Equation (a) and (b) show the calculation formula from input layer (*i*) to hidden layer (*j*) and equation (c) and (d) show formula for hidden layer (*j*) to output layer (*k*). The network activation function as in equations (a) and (c)is Sigmoid Activation Function.

Between input (*i*) and hidden (*j*)

$$O_j = f(net_j) = \frac{1}{1+e^{-net_j}} \qquad \textbf{(a)}$$

$$net_j = \sum_i w_{ij} O_i + \theta_j \qquad \textbf{(b)}$$

where

O_j is the output of node *j*

O_i is the output of node *i*

w_{ij} is the weight connected between node *i* and *j*

θ_j is the bias of node *j*

Between hidden (*j*) and output (*k*)

$$\boldsymbol{O_k = f(net_k) = \frac{1}{1+e^{-net_k}}} \qquad \textbf{(c)}$$

$$\boldsymbol{net_k} = \sum_k w_{jk} O_j + \theta_k \qquad \textbf{(d)}$$

Where:

O_k is the output of node *k*

O_j is the output of node *j*

w_{jk} is the weight connected between node *j* and *k*

θ_k is the bias of node *k*

Error is calculated using equation (e) to measure the differences between desired output and actual output that has been produced in feed forward phase. Error than propagated backward through the network from output layer to input layer as represented below. The weights are modified to reduce the error as the error is propagated.

$$\textbf{Error} = \frac{1}{2}\left(\boldsymbol{Output_{desired} - Output_{actual}}\right)^2$$

(e)

Based on the error calculated, back propagation is applied from output (*k*) to hidden (*j*) as shown by equation (f) and (g).

$$\boldsymbol{w_{ji}(t+1) = w_{ji}(t) + \Delta w_{ji}(t+1)} \ \textbf{(f)}$$

$$\boldsymbol{\Delta w_{ji}(t+1) = \eta \delta_k O_j + \alpha \Delta w_{ji}(t)} \ \textbf{(g)}$$

With

$$\boldsymbol{\delta_k = O_k(1 - O_k)(t_k - O_k)} \qquad \textbf{(h)}$$

Where

w_{ji}(*t*) is the weight from node *j* to node *i* at time *t*

Δw_{ji} is the weight adjustment

η is the learning rate

α is the momentum rate

δ_j is error at node *j*

δ_k is error at node *k*

O_i is the actual network output at node *i*

O_j is the actual network output at node *j*

O_k is the actual network output at node *k*

w_{kj} is the weight connected between node *j* and *k*

θ_k is the bias of node *k*

This process will be repeated iteratively until convergence is achieved (targeted learning error or maximum number of iteration).

3.2 Genetic Optimized Neural Network

GA is an iterative procedure that consists of a constant-size population of individuals called chromosomes, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The GA can be employed to improve the performance of BPN in different ways. GA is a stochastic general search method, capable of effectively exploring large search spaces, which has been used with BPN for determining the number of hidden neurons.

The hybrid GA-ANN has been used in the diverse applications. GA has been used to search for optimal hidden-layer architectures, connectivity, and training parameters (learning rate and momentum parameters) for ANN for predicting community-acquired pneumonia among patients with respiratory complaints [10]. GA has been used to initialize and optimize the connection weight of ANN to improve the performance ANN and is applied in a medical problem for predicting stroke disease [7]. GA has been used to optimize the ANN parameters namely: learning rate, momentum coefficient, Activation function, Number of hidden layers and number of nodes.

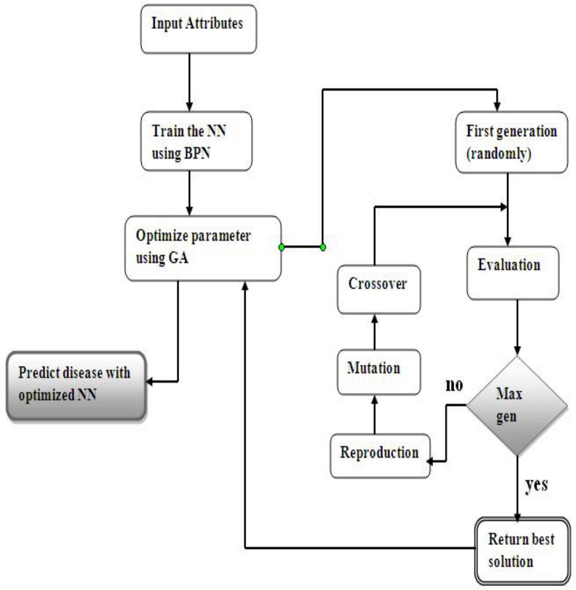

Figure 2: Flow chart for Hybrid Neuro GA

Determine the input feature with the help of expert. Initialize the input features that are randomly selected. Design the suitable network then training it and compute the fitness variable. Then perform the genetic algorithm from this generate then new input. Store the best feature for diagnosis of disease. Test the data to diagnose the optimum value.

IV GENETIC ALGORITHM

GAs search by simulating evolution, starting from an initial set of solutions or hypotheses, and generating successive generations of solutions [6]. Genetic algorithm is a ways of solving the problems and used evolve a solution to a problem. The operation of GAs is selection, crossover, mutation and accepting. There are several terms in GA [8]. *Fitness* is a measure of the goodness of a data where it measures how well the data fits the search space or solves the problem. *Selection* is a process for choosing a pair of organisms to reproduce while *Crossover* is a process of exchanging the genes between the two individuals that are reproducing. *Mutation* is the process of randomly altering the data.

Pseudo code for GN:

```
BEGIN
INITIALIZE population with random candidate solution.
EVALUATE each candidate;
REPEAT UNTIL (termination condition) is satisfied DO
    1. SELECT parents;
    2. RECOMBINE pairs of parents;
    3. MUTATE the resulting offspring;
    4. SELECT individuals or the next generation
END
```

V PREDICTION OF STROKE DISEASE

In this project, Hybrid Neuro Genetic algorithm is proposed to select input features for the diagnosis of stroke diseases. Which is based on the Application of Artificial Neural Network (ANN) can be time-consuming due to the selection of input features for the Multi Layer Perceptron (MLP).The main goal of this work is to predict the stroke diseases, which is done by some attributes or clinical variables.

In this algorithm determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not. To retrieve the significant data from the database by using data mining techniques. To predict the stroke disease from the extracted data using Hybrid Neuro Genetic algorithm.

The proposed model uses predictive Hybrid Neuro Genetic Algorithm for predicting the presence of stroke disease for dimensionality reduction. So the reduced subset of the attributes could be used as inputs. Once pre-processing is over, the dataset containing more than 1000 records is used for predicting stroke diseases. By using these records it is very difficult and time consuming task to identify the diseases. It deals with huge amount of dataset and reduces it to a lower dimension. Then normalization used to remove the negative datasets. The attribute values need to be normalized, in order to avoid the high value attributes that may confuse or underestimate the low value attributes. The Figure3 represents the Prediction of Stroke Disease.

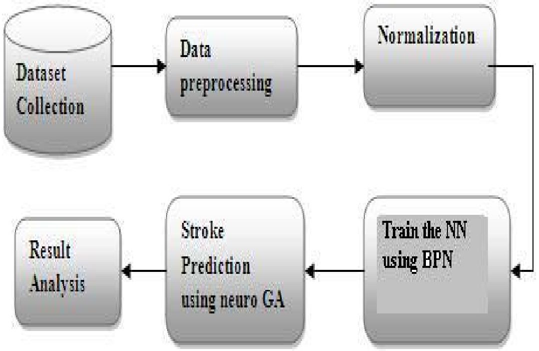

Figure 3: Architecture diagram for prediction of stroke diseases

In the first step the stroke consist of dataset which is collected from the medical institute. The dataset consists of patient information, patient history, Gene diagnosis disease database which contains the symptoms of stroke disease. Then it performs the preprocessing technique. So that noisy data, duplicate records, missing data, inconsistent data are removed. Then perform the normalization, in this work it perform the min-max normalization. So that negative values are removed. Then train the NN by using the BPN. Then predict the stroke disease using the neuro genetic algorithm. Then diagnosis the disease whether the patient suffer from stroke or not.

VI RESULT ANALYSIS & DISCUSSION

The experimental results of the heart attack disease system for prediction are explained in this section. Here, the performance of the proposed system is compared with the neural network-based system to evaluate the sensitivity, specificity and accuracy. Experimental environment and evaluation the proposed Multi Layer Feed Forward Neural network system has been implemented using MATLAB.

The dataset for the project have been collected from 294 patients who have symptoms of stroke disease. Stroke datasets are divided into two parts for training 196 and for testing 98 data are used. It has trained and tested for 1000 iteration with random weight of hidden layers. The weights in ANN are encoded in such away weight is between -1.0 to +1.0. The objective function is minimization of the Mean Square Error (MSE). The fitness function considered is the minimum MSE and computed by recalling the network. Performance of the GA-NN with 14 attributes is for 14-6-1 topology that has been shown in figure 4.

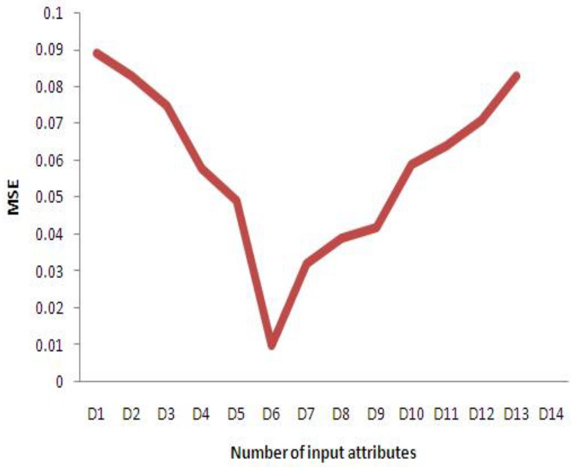

Figure 4: Minimum MSE

In the proposed system to find the risk factors of stroke patients and obtained results are evaluated with namely sensitivity, specificity, and accuracy, Sensitivity, specificity and accuracy are the commonly used statistical measures to illustrate the medical diagnostic test and especially, used to enumerate how the test was good and consistent The accuracy is measured based on sensitivity and specificity. Based on patient risk levels with diagnosis result = normal and diagnosis result = Abnormal, number of attributes used with patient suffering from diseases is measured. Figure 4 shows the result analysis and accuracy for predicting stroke diseases. In order to find these metrics, we first compute some of the terms like, True positive, True negative, false negative and false positive.

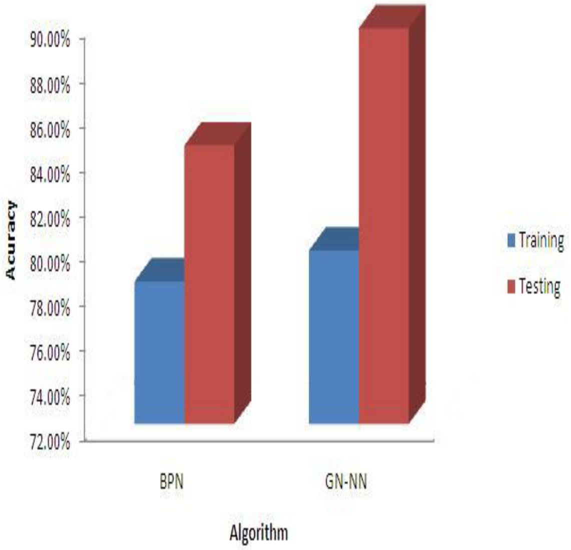

Figure 5: performance of algorithm

Sensitivity = TP/TP+FN
Sensitivity = TN/TN+FP
Accuracy = $\left[\frac{TN+TP}{TN+TP+FN+FP}\right]$

The table 2 shows the performance study of the algorithm. According to these values the accuracy is calculated and analyzed. Out of 300 input features 196 for training and 104 for testing. The accuracy of the training dataset is 79.7% and testing accuracy is 89.67%. Performance can be determined based on the evaluation time of calculation and the error rates.

Approach	Training	Testing
BPN	78.37%	84.44%
GA-NN	79.75%	89.67%

Table 2: Performance of algorithm

BPN is widely used in the learning algorithm in Neural Network for the many applications. However, BP learning depends on the several parameters in the MLFFNN such as learning rate and momentum rate number of neurons in the hidden layers. Due to this, GA has been used to obtain the optimal parameter value and weight for the BP learning. So that the performance of GA is increased better than the MLFFNN.

VII CONCLUSION

Thus the project has to be predicting the stroke disease, which is defined by some attributes or clinical variables. In order to predict the stroke disease it adopts hybrid Neuro genetic algorithm. Healthcare industry makes use of data mining techniques and generates huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. This work demonstrates about ANN based prediction of stroke disease by improving the accuracy with higher consistent rat using optimized hidden neurons. In this algorithm determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease. The data is collected from 300 patients. Among that 180 patients having disease .In the proposed work 196 data are used for training and 104 data is used to test the performance of the system. It could be reduce the medical error, enhance patient safety. By using this algorithm it will easy to diagnosis of patient with stroke diseases with more accuracy.

VIII REFERENCE

[1] Sudha.A, Gayathiri.P, Jaisankar.N, (2012) “Effective Analysis and Predictive Model of Stroke Disease using Classification Methods”, *International Journal of Computer Applications* (0975 – 8887) Volume 43– No.14, pp.0975 – 8887.

[2] Obi J.C. and Imainvan A.A, (2011) “Decision Support System for the Intelligient Identification of Alzheimer using Neuro Fuzzy logic” *International Journal on Soft Computing (IJSC),* Vol.2, No.2,pp.25-38.

[3] Anooj P.K, (2011), “Clinical decision support system: Risk level predictionof heart disease using weighted fuzzy rules”, *Journal of King Saud University Computer and Information Sciences* (2011)Vol(11) pp.309-314 .

[4] Srinivas.K Kavihta Rani.B Dr.Govrdhan.A, (2010), “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, *International Journal on Computer Science and Engineering* Vol. 02, No. 02, pp.250-255.

[5] **“**Global status report on no communicable diseases”, (2010), *World Health Organization*

[6] RC Chakraborty, 2010 “Fundamentals of Genetic Algorithm”

[7] Shanthi.D, Dr.Sahoo.G Dr.Saravanan.N, (2009), “Evolving Connection Weights of Artificial Neural Network Using Genetic Algorithm With Application to the Prediction Stroke Diseases”, *International Journal of Soft Computing* –Vol (2), pp.95-101.

[8] Shanthi.D, Dr.Sahoo.G Dr.Saravanan.N, (2008),“Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke”, *International Journals of Biometric and Bioinformatics (IJBB)*, Volume (3) : Issue (1), pp.250-255.

[9] Tom V Marhew ,“Genetic algorithm”

[10] H. Paul S., G. Ben S., T. Thomas G., W. Robert S.,(2004), ” Use of genetic algorithms for neural networks to predict community-acquired pneumonia”, Artificial Intelligence in Medicine, Vol. 30, Issue 1, pp.71-84.

[11] Dallas.K, (2003) “ Heart Disease and Stroke Statistics “, *American Heart Association*.

[12] Mohr J.P (2001) “Stroke Analysis”, 4th Edition, Oxford Press, New York.

[13] Lawrence M brass, M.D “Stroke” Chapter 18.

[14] “Pocket Guidelines for Assessment and Management of Cardiovascular Risk”, World Health Organization.