International Journal of Emerging Technology and Innovative Engineering
Volume 1, Issue 8, August 2015 (ISSN: 2394 – 6598)

PREDICTION OF DIABETES USING BACK PROPAGATION ALGORITHM

M.Durairaj
Assistant Professor
School of Com Sci., Eng. & Application,
Bharathidasan University,Trichy,TN,India
durairaj.bdu@gmail.com

G.Kalaiselvi
Research Scholar
School of Com Sci., Eng. & Application,
Bharathidasan University,Trichy,TN,India
gkalaiselvik@gmail.com

ABSTRACT

Diabetes affected more than 246 million of people in a worldwide, among a majority of them being women. According to the WHO report, with 2025 this number is expected to grow near or more than 380 million. Diabetes occurs when a body is unable to generate or react correctly to insulin which is wanted to regulate glucose. Besides leading to heart disease, diabetes also increase the risk of rising kidney disease, blindness, nerve damage, in addition to blood vessel damage. Diabetes disease diagnosis needs proper interpretation of the diabetes data which in turn makes it as remarkable research issue to be focused. This study has taken the diabetes disease diagnosis data set from, Pima Indians. For this purpose, a back propagation Network which was trained by Levenberg–Marquardt (LM) algorithm is used. In this work the UCI machine learning database focusing on diabetes disease diagnosis is used and the results of the work a compared with the results of the previous studies reported.

Keywords

Diabetes disease diagnosis, Back Propagation Network, Leven berg–Marquardt algorithm.

1. INTRODUCTION

We In general, a person is considering to be suffering as of diabetes, when blood sugar levels are above normal (4.4 to 6.1mmol/L). Pancreas in the human body produces insulin, a hormone that is responsible to help glucose appear at each cell of the body. A diabetic patient basically has low production of insulin or their body is not able to use the insulin well. There are three main types of diabetes, viz. Type 1, Type 2 and Gestational. Type 1 is an auto immune disease going on at a very young age of below 20 years. In this type of diabetes, the pancreatic cells that produce insulin have been destroyed. Type 2 refers to the situation when the various organs of the body become insulin resistant, and pancreas doesn't build the required amount of insulin. Gestational diabetes tends to occur in pregnant women, as the pancreas don't build sufficient amount of insulin. All these types of diabetes need treatment and if they are detected at an early state, one can avoid the complications related with them.

The Back Propagation Networks (BPNs) have been successfully used in replacing conventional pattern recognition methods for the disease diagnosis systems.But, since it applies the steepest descent method to update the weights, it suffers from a slow convergence rate and often yields sub optimal solutions. A variety of related algorithms have been introduced to address that problem. A number of researchers have carried out comparative studies of BPN training algorithms. Leven berg–Marquardt (LM) algorithm has been used in this study, which provides generally faster convergence and better estimation results than other training algorithms.

The pervious diagnosis result of LM algorithm reported on Pima Indian diabetes disease dataset was not better than the diagnosis results produced by other training algorithms. This can be because of that, LM algorithm converges very fast but it can cause the memorization effect when the overtraining occurs. If a Probability Neural Network(PNN) starts to memorize the training set, its generalization starts to decrease and its performance may not be improved for untrained test sets. The proposed work in memorization of the training set can be because of the overtraining. So, before the starting of memorization, we must determine the optimum trained neural network using the maximum accuracy value of the test data. Using the optimum trained neural network, Pima Indian diabetes disease diagnosis can be made with better accuracy.

This work aims to use Back Propagation Network(BPN) with LM training algorithm for the prediction and classification of diabetes on Pima Indian Dataset repository. Other network type will be compared with the LM algorithm to answers the prediction accuracy on diabetes dataset. For the enhancement of accuracy, the selection of training,

validation and testing dataset will be selected suitably. The data cleaning is essential to improve the accuracy and correct classification.

This paper organized as follows: Section 1 discusses various training algorithm applied on the prediction and analysis. Section 2 presents the different approaches include in neural network techniques. Section 3 provides the methodology applied in Back Propagation Nework algorithm. Section 4 discusses the main aim of the Pima Indian Dataset description. Section 5 presents the exprimetation of training process. Section 6 compares these techniques and discussed. Section 7 concludes with our findings.

2. LITERATURE REVIEW

Data mining is the extraction of useful information from the large volume of data [13]. Data mining has been applied in various fields like medicine, marketing, banking, etc. In medicine, predictive data mining is used to diagnose the disease at the earlier stages itself and helps the physicians in treatment planning procedure.

Asha Gowda Karegowda, et.al. [1] provided the application of hybrid GA(Genetic Algorithm) and BPN(Back Propagation Network). They experimented for classification of PIMA dataset. They concluded that the Back Propagation learns by making modifications in weight values by using gradient method starting at the output layer then, moving backward through the hidden layers of the neural network and hence is prone to lead to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure.

Ravi Sanakal, et.al. [2] presented a diagnostic Fuzzy cluster FCM as well as SVM using SMO and decided which technique helps in diagnosis of Diabetes disease. The best result is obtained in a FCM with an accuracy of 94.3% and the positive predictive value is 88.57%. SVM has an accuracy of 59.5% which is quite low. These results are quite satisfactory only, due to the fact that detecting the Diabetes is a very complex problem.

Rajesh, et.al.[3] presented the C4.5 algorithm for classification and the classification rate obtained was 91%. Future enhancement of this work includes improvisation of the C4.5 algorithms in order to improve the classification rate with greater accuracy.

Radha, et.al. [4] demsonstrated the application of five classification techniques namely C4.5, SVM, K-NN, PLR, and BLR to predict the diabetes disease in patients. They pointed out that necessary to intend an automatic classification tool. In this study, these five techniques were chosen based on the computing time, in which BLR has the lowest computing time with 75% accuracy and error rate of 0.27. The second one with more accuracy rate is SVM while comparing with other techniques. The accuracy of BLR is 75% from the results obtained. The BLR algorithm plays a vital role in data mining techniques.

RajAnand, et.al. [5] presented a novel approach to Pima Indian diabetes data diagnosis using PCA (Principle Component Analysis) and HONN (Higher Order Neural Network). The HONN can perform diabetes classification with parsimonious representation of node architecture due to its generation of higher order terms. A lower mean square error and faster convergence is attained with PCA preprocessing.

Veena Vijayan, et.al.[6] suggested for the expectation of maximization of algorithms, among K Nearest Neighbor algorithm, K-means algorithm, Amalgam KNN algorithm and Adaptive Neuro Fuzzy Inference System algorithm. From the observation EM (Expectation Maximization) possess the least classification accuracy. Amalgam KNN and ANFIS (Adaptive Neuro Fuzzy Inference system) provide the better classification accuracy results. Amalgam KNN comprises both the feature of KNN and K means. ANFIS incorporates both the features.

Priya, et.al. [7] proposed a method of applying Neural Networks for classification. The result produced by this model is higher than the other models, since it performs classification using Neural networks in the Rapid miner tool. This has produced an improvement in the accuracy when compared to the other techniques.

Blanca S. Leon, et.al.[8] demonstrated the application of Recurrent Neural Networks (RNN) for modeling and control of glucose–insulin dynamics in T1DM (type 1 diabetes mellitus) patients. The proposed RNN, used in these experiments, captures very well the complexity associated with blood glucose level for type 1 diabetes mellitus patients.

Paul S. Heckerling, et.al. [9] presented the predictor of variables derived from a neural network genetic algorithm which are accurately discriminated the urinary tract infection from non infection in women with urinary complaints. Clinical variables are important in predicting infection differed depending on the uropathogen colony count used to define urinary infection in their work. In addition, some variables predicted urine infection in unexpected ways, and interacted with other variables in making those predictions.

Sebastian Polak, et.al. [10] presented a study in which ANNs as well as other information technology tools are able to identify and analyze relationships in data even when some of the inputs are very complex and difficult to be defined. Therefore, the research carried-out in virtual space is gaining importance in medical applications as well. Superiority of ANNs is pronounced in their ability of automatic identification of complicated relationships.

3. METHODOLOGY

3.1 Back Propagation Network

The backpropagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The backpropagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the backpropagation algorithm 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 error will be minimal.

Back Propagation Algorithm

1. The network is first initialized by setting up all its weights to be small random numbers say between 0 and 1.
2. The input pattern is applied and the output calculated (this is called the forward pass). The calculation gives an output which is completely different to what actually needed (the Target), since all the weights are random.
3. We then calculate the Error of each neuron, which is essentially: Target – Actual Output. This error is then used mathematically to change the weights in such a way that the error will get smaller.
4. The Output of each neuron will try to get closer to its Target (this part is called the reverse pass).
5. This process is repeated again and again until the error is minimal.

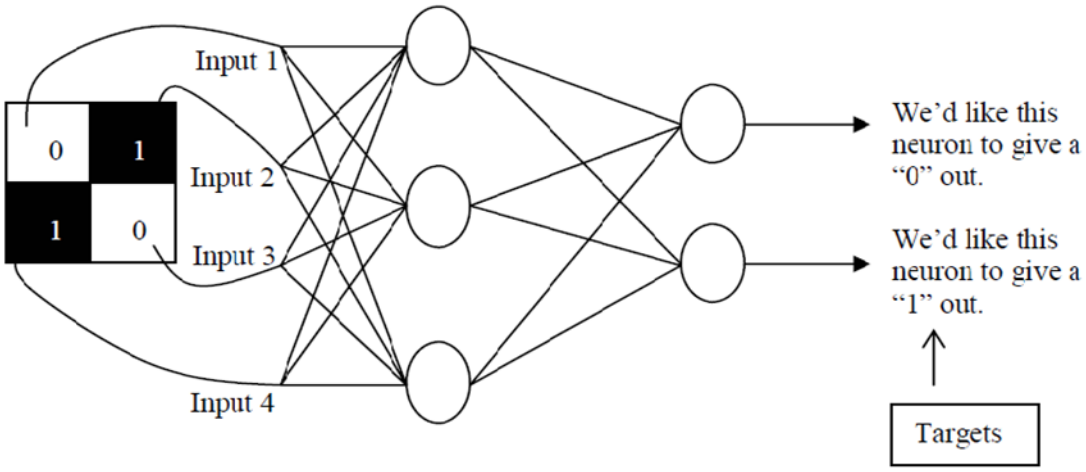

Fig 1: Represents the Back Propagation Network

4. DATA SET

The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases. The patients under consideration are the Pima Indian population living in Arizona, USA. More than 50% Pima Indian Population is suffering from diabetes and 95% of them are due to the overweight. Number of research has been done on these populations proved that obesity is the main cause for the diabetes. The data set mainly contain 9 attributes and 767 number of instances.The list of attributes from PIMA INDIAN Dataset used for experimentation is listed in Table1.

Table1. List of attributes from data sets for simulation tests

Attribute no	Attribute to be test	Symbols
1	Number of time pregnant	Preg

2	Glucose tolerance test to find the plasma glucose level concentration in saliva	Plas
3	Diastolic blood pressure measured in mmHg	Pre
4	Skin rashes and thickness fold in mm(triceps)	Skin
5	2-hour serum insulin in mu U/ml(INSULIN)	Insu
6	Body Mass Index	Mass
7	Diabetes pedigree function	Pedi
8	Age	Years
9	Diabetes class variable	Binvar

5. EXPERIMETATION

The proposed work is implemented using NNTool of MATLAB 7.12.0; The BPN structures employed in the study utilized the new function implemented in MATLAB. Detailed information about the realisation of the BPN structures can be found in the neural network toolboxes. Back Propagation algorithm is chosen for classifying the network. Classification accuracy has been used for the comparison of studies reported in literature focusing on diabetes disease diagnosis and using same database.

We checked the accuracy values of the test data and the training data and we determined the optimum trained neural network using the maximum accuracy value of the test data. Aother one method of performed function Mean Squared Error(mse) increase the all training data values.From this, the BPN with LM converge very fast and started to learn the training set after eight epoch.Additionally, the classification accuracy of BPN with LM obtain by this study using correct training was better than those obtained by other studies for the conventional validation method.

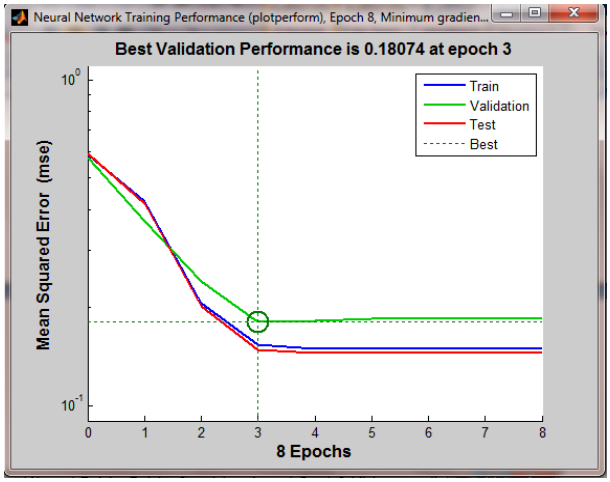

Fig2: Represents the Training Performance

6. RESULT AND DISCUSSION

The classification accuracies obtained by this research work and the best values of other studies for pima-diabetes disease dataset were presented table1.

As seen in the table 2, pervious diagnosis result of the PNN with LM algorithm reported on Pima Indian diabetes disease dataset is very far from the result on 80%.From the results, it can be stated that the classification accuracy. This research work obtained 91% classification accuracy using the same method. This can be because of that, LM algorithm converges very fast but it can cause the memorization effect when the over training occurs. So, to prevent the memorization effect, the accuracy values of the test were checked during the training process determined the optimum trained neural network using the maximum accuracy value of the test data. Fig 2: shows the graphical representation of prediction accuracy.

Table2.Performance of Training Function

Studies	**Network used**	**Training function**	**Classification Accuracy (%)**
This study	Back Propagation Network(BPN)	Levenberg-marquart (LM)	91%
Previous study	Probability Neural Network(PNN)	Levenberg-marquart (LM)	80%

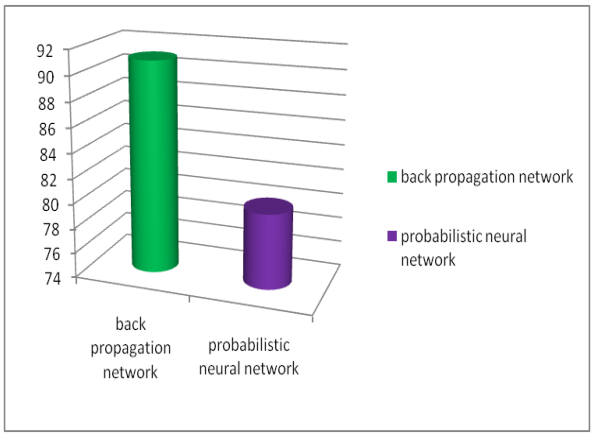

Fig3. Prediction of training function

7. CONCLUSION

This paper presents a methodology of diabetes disease diagnostic using Back Propagation Network with LM training algorithm on Pima Indian Dataset.The results were compared with the results of the previous studies reported. It was observed that Neural Network structures could be effectively used to help finding of diabetes disease. The classification accuracy of BPN with LM obtained by this study was better than those get by other studies for the predictable validation model.

REFERENCES

[1] Asha Gowda Karegowda ,A.S. Manjunath , M.A. Jayaram,Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes,International Journal on Soft Computing (IJSC), Vol.2, No.2, May 2011.

[2] Ravi Sanakal, Smt. T Jayakumari, ―Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine,International Journal of Computer Trends and Technology (IJCTT) – volume 11 number 2 May 2014.

[3] K.Rajesh,V.Sangeetha,Application of Data Mining Methods and Techniques for Diabetes Diagnosis,International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3 September 2012.

[4] P. Radha, Dr. B. Srinivasan, ―Predicting Diabetes by cosequencing the various Data Mining Classification Techniques,IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[5] Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse,K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[6] Veena Vijayan V. Aswathy Ravikumar, ―Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus,International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014 .

[7] S.Priya R.R.Rajalaxmi,An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network,International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, 2012.

[8] BlancaS.Leona,AlmaY.Alanisb,n,EdgarN.Sanchea Fernando Ornelas-Tellezc, EduardoRuiz-Velazquezb, ―Inverse optimal neural control of blood glucose level for type1diabetes mellitus patients,Journal of the Franklin Institute 349 (2012) 1851–1870.

[9] Paul S. Heckerling, Gay J. Canaris, Stephen, Flach, Thomas G. Tape,Robert S. Wigton, Ben S. Gerber, ―Predictors of urinary tract infection based on artificial neural networks and genetic algorithms,international journal of medical informatics 7 6, 2007.

[10] Sebastian Polak Aleksander Mendyk, ―Artificial neural networks based Internet hypertension prediction tool development and validation,Applied Soft Computing 8 (2008) 734–739.

[11] Pankaj Srivastava, Neeraj Sharma,Richa Singh, Soft Computing Diagnostic System for Diabetes, International Journal of Computer Applications (0975 – 888)Volume 47– No.18, June 2012.

[12] V.Karthikeyani,I.ParvinBegum,K.Tajudin,I.Shaha Begam, Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction,International Journal on Computer Science and Engineering (IJCSE), December 2012.

[13] M.Durairaj,V.Ranjani, ―Data Mining Applications In Healthcare Sector: A Study, in international journal of scientific & technology research volume 2, issue 10, October 2013.

[14] Hasan Temurtas a, Nejat Yumusak b, Feyzullah Temurtas c,d,* A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications 36 (2009) 8610–8615.

[15] M. Durairaj, G. Kalaiselvi - ¬Prediction of Diabetes Using Soft Computing Techniques- A Survey. International Journal Of Scientific & Technology Research Volume 4, Issue 03, March (2015) 2277-8616.