A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks

T.Jayalakshmi
Computer Science Department
CMS College of Science and Commerce
Coimbatore, INDIA
jayas20@rediffmail.com

Dr.A.Santhakumaran
Statistics Department
Salem Sowdeswari College
Salem, INDIA
ask.stat@yahoo.com

***Abstract*—Many real world problems can be solved with Artificial Neural Networks in the areas of pattern recognition, signal processing and medical diagnosis. Most of the medical data set is seldom complete. Artificial Neural Networks require complete set of data for an accurate classification. This paper dwells on the various missing value techniques to improve the classification accuracy. The proposed system also investigates the impact on preprocessing during the classification. A classifier was applied to Pima Indian Diabetes Dataset and the results were improved tremendously when using certain combination of preprocessing techniques. The experimental system achieves an excellent classification accuracy of 99% which is best than before.**

Keywords- Artificial Neural Networks; Diabetes Mellitus; Missing Value Analysis; Pre-Processing Methods.

I. INTRODUCTION

Medical information systems in modern hospitals and medical institutions become larger and larger; it causes great difficulties in extracting useful information for decision support. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are essential. It has been proven that the benefits of introducing machine learning into medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources. Artificial Neural Networks (ANN) is currently the next promising area of interest. Already it could successfully apply to various areas of medicine such as diagnostic systems, bio chemical analysis, image analysis and drug development. The benefit of using ANN is that they are not affected by factors such as fatigue, working conditions and emotional state. Diabetes mellitus is a lifelong disease resulted from the underproduction or reduced action of hormone insulin. This dysfunction of insulin results in blood glucose levels out of normal range, leading to many short and long-term complications [13]. Depending on the cause of this insulin insufficiency, two different types of the disease are distinguished. Type I diabetes mellitus and Type II diabetes mellitus. This paper deals about the classification of Type II diabetes.

Varieties of techniques have been applied to deal with the classification problems. Many previous research works shows that neural network classifiers have a better performance, lower classification error rate, and more robust to noise than other methods [1]. One type of neural network commonly used for classification is a Multi Layer Perceptron (MLP) a feed forward net with one or more layers of nodes and with back propagation for training [12]. Increasing both the number of hidden layers and neurons will make the network more flexible to the mapping to be implemented [6]. Moreover, increasing the number of hidden neurons increases the risk of over fitting.

To overcome these problems, this paper present two different approaches. The first approach attempts to construct the data sets with reconstructed missing values. It can achieve higher correct classification rates than the standard back propagation method. Reconstruction of missing values may shift input patterns, and the network may have to settle on a complicated solution to satisfy all reconstructed patterns. In contrast, the second approach attempts to reduce the actual learning time by using data pre-processing. Pre-processing can speed up the training time by starting the training process for each feature within the same scale. It is especially useful for modeling applications where the inputs are generally on widely different scales. In this study, various missing value analysis and pre-processing methods are analyzed. This paper also investigates the reconstruction values with pre-processing and observes the results. This shows that the results are tremendously improved when applying these concepts.

This paper organized as follows: Section II briefs about Artificial Neural Networks, Section III gives the description of the Diabetes mellitus. Section IV provides the details about back propagation method, Section V proposes the methodology and Section VI concludes the paper

II. ARTIFICIAL NEURAL NETWORKS

A Neural Network is a massively parallel-distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It is very sophisticated modeling technique capable of modeling extremely complex functions.

978-0-7695-3958-4/10 $26.00 © 2010 IEEE
DOI 10.1109/DSDE.2010.58

ANNs attempt to create machines that work in a similar way to the human brain by building them using components that behave like biological neurons. However the operation of artificial neural networks and artificial neuron is far more simplified that the operation of the human brain. The brain consists of millions of these neurons, which may be specialized in some task or not. The behaviour of the brain inspired to devise an artificial neuron called perceptron, which is the basis of all neural network models. It resembles the brain in two respects a) Knowledge is acquired by the networks from the environment through a learning process. b) Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge. Neural networks have advantages over classical statistical approaches especially when the training set size is small compared with the dimensionality of the problem to be solved and the underlying data distribution is unknown. Learning is essential to most of the neural network models. Learning can be supervised, when the network is provided with the correct answer for the output during training, or unsupervised, when no external teacher is present. Error is calculated from the actual output to the calculated output. The goal of the neural network learning is to iteratively adjust weights, in order to globally minimize a measure of the difference between the actual output of the network and the desired output as specified by the teacher.

III. DIABETES MELLITUS

Diabetes mellitus is the most common endocrine disease. The disease is characterized by metabolic abnormalities and by long-term complications involving the eyes, kidneys, nerves, and blood vessels. The diagnosis of symptomatic diabetes is not difficult. When a patient presents with signs and symptoms attributable to an osmotic diuresis and is found to have hyperglycemia essentially all physicians agree that diabetes is present. The two major types of diabetes are Type I diabetes and Type II diabetes. Type I diabetes usually diagnosed in children and young adults, and was previously known as juvenile diabetes [10] [14]. Type I diabetes mellitus (IDDM) patients do not produce insulin due to the destruction of the beta cells of the pancreas. Therefore, therapy consists of the exogenous administration of insulin. Type II diabetes is the most common form of diabetes. Type II diabetes mellitus (NIDDM) patients do produce insulin endogenously but the effect and secretion of this insulin are reduced compared to healthy subjects [3]. Currently cure does not exist for the diabetes, then only option is to take care of the health of people affected, maintained their glucose levels in the blood to the nearest possible normal values [5].

IV. BACK PROPAGATION ALGORITHM

Back propagation algorithm is a classical domain dependent technique for supervised training. It works by measuring the output error calculating the gradient of this error and adjusting the ANN weights (and biases) in the descending gradient direction. Hence, Back propagation is a gradient-descent local search procedure [9]. A study shown that approximately 95% of the reported neural network applications utilize multi layer feed forward neural network with back propagation algorithm. The training process is an incremental adaptation of connection weights that propagate information between simple processing units called neurons. The neurons are arranged in layers and connections between the neurons of one layer to those of the next exist. The architecture of multilayer feed forward neural network is shown in Figure.1.

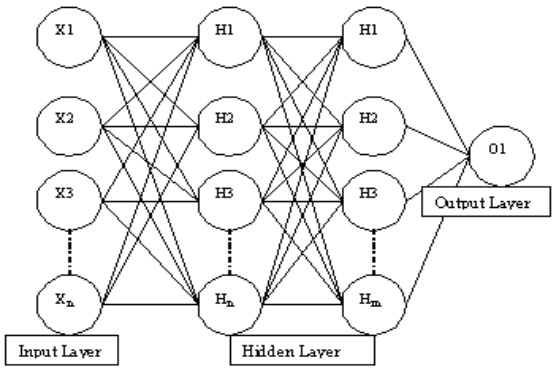


Figure 1. Multi Layer Feed Forward Network

Back propagation is the generalization of the Widrow-Hoff learning rule to multilayer networks and non-linear differentiable transfer functions [15].

V. METHODOLOGY

The performance of the proposed method is demonstrated by comparing the results of various missing value analysis techniques using back propagation Levenberg Marquardt method. The performance criteria used in this research focuses on the speed of convergence and can be measured in terms of number of epochs and classification accuracy.

A. Classification data

The problem that has been chosen for this research is to classify the Type II diabetic data using LM classification algorithm. To conduct experiments Pima Indian data set is used. The data set is difficult to classify [8]. It consists of 768 data samples. The set consists of eight input parameters such as number of times pregnant, plasma glucose concentration, blood pressure, triceps skin fold thickness, serum insulin, body mass index, diabetic pedigree function and age. The network topology used for this study is 8-8-8-1.

B. Missing Data analysis

Back propagation neural networks have been applied for classification problems in real world situations. A drawback of this type of neural network is that it requires a complete set of input data, where as real world data is seldom complete [11]. The problem of databases containing missing values is a common one in the medical environment. It may be that the medical procedures was not needed or that the

physiological measurements were taken but not recorded by the person perhaps due to time constraints [2]. ANNs cannot interpret missing values, and when a database is highly skewed, ANNs have difficulty in identifying the factors leading to a rare outcome. It poses difficulty in the analysis and decision-making processes which depend on these data, requiring methods of estimation that are accurate and efficient. Various techniques exist as a solution to this problem, ranging from deletion to methods employing statistical and Artificial Intelligent techniques to impute for missing values. The following can deal with this issue. The first and easiest way to deal with missing values is simply delete all the cases with missing values for the variable under consideration. This technique however may lead to the loss of potentially valuable information about patients whose values are missing. The second approach is to replace all missing values with the mean. The method of replacing by average is to replace all missing values of an attribute by the average of all available values of the same attribute in the training set. Replacing the missing values with the means might bias the databases towards the sicker ones. The third technique is to replace all the missing values with zeros. The method of replacing by zero is simply to replace all missing values by zero. If the values are important for clinical management the assessment of missing values leads to poor classification. The fourth technique K-nearest neighbour (KNN) method replaces missing values in data with the corresponding value from the nearest-neighbour column. The nearest-neighbour column is the closest column in Euclidean distance. If the corresponding value from the nearest-neighbour column is also missing means, the next nearest column is used.

C. *Preprocessing of Input Data*

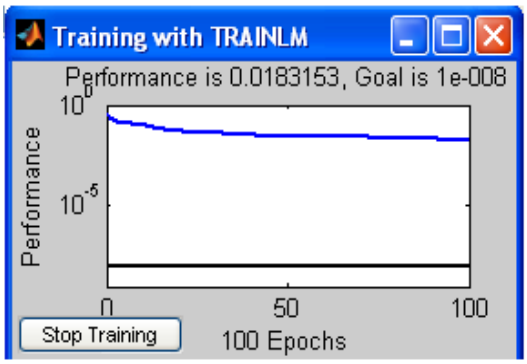
Neural network training could be made more efficient by performing certain pre-processing steps on the network inputs and targets. Network input processing functions transforms inputs into better form for the network use. The normalization process for the raw inputs has great effect on preparing the data to be suitable for the training. It can be used to scale the data in the same range of values for each input feature in order to minimize bias within the neural network for one feature to another. Data normalization can also speed up training time by starting the training process for each feature within the same scale. It is especially useful for modeling application where the inputs are generally on widely different scales. Principle Component Analysis (PCA) is one of the most powerful pre-processing techniques. Principal Component's normalization is based on the premise that the salient information in a given set of features lies in those features that have the largest variance. The PCA algorithm normalizes the components so that they have zero mean and unity variance [7]. This is accomplished by using eigenvector analysis on either the covariance matrix or correlation matrix for a set of data.

D. *Classification structure*

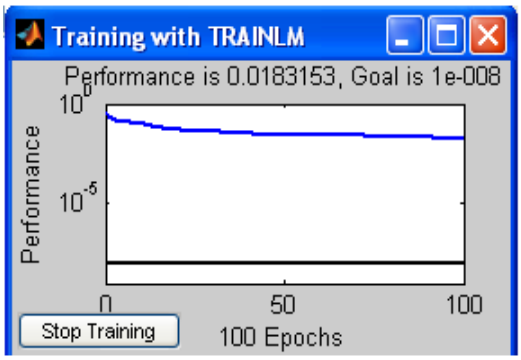
Classification structure used in the proposed method is four layers feed forward networks i.e. one input layer, two hidden layers and one output layer. The ANN has eight input nodes, eight hidden nodes and one output node. The network was trained by Levenberg Marquardt back propagation algorithm with tansigmoid activation function. The network parameters such as learning rate, momentum constant, training error and number of epochs can be considered as 0.9, 0.9, 1e-008 and 100 respectively. Before training, the weights are initialized to random values. The reason to initialize weights with small values is to prevent saturation. To evaluate the performance of the network the entire sample was randomly divided into training and test sample. The model is tested using the standard rule of 80/20, where 80% of the samples are used for training and 20% is used for testing. In this classification method, training process is considered to be successful when the Mean Square Error (MSE) reaches the value 1e-008. On the other hand the training process fails to converge when it reaches the maximum training time before reaching the desired MSE. The training time of an algorithm is defined as the number of epochs required to meet the stopping criterion.

E. *Experimental results*

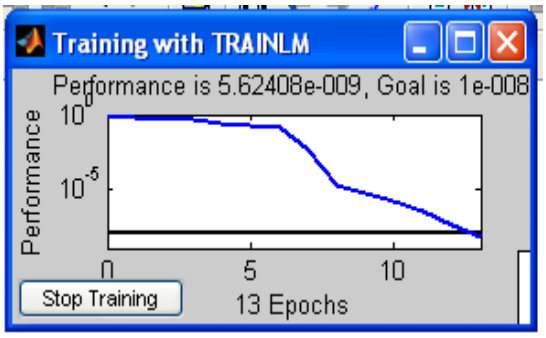
A computer simulation has been developed to study the impact of pre-processing and missing value techniques. The simulations have been carried out using MATLAB. Various networks were developed and tested with random initial weights. The network is trained ten times and the performance goal is achieved at different epochs which are shown in Figure.2



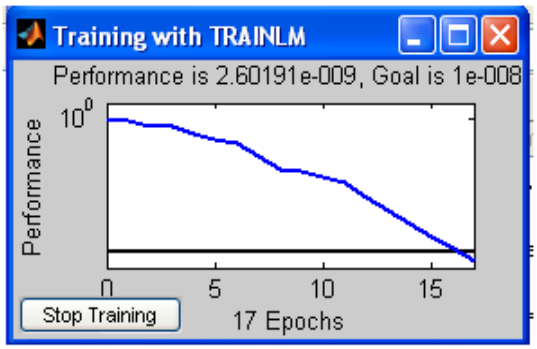
(a) Omit the samples



(b) Replace with Zero



(c) Replace with Mean



(d) Replace with KNN

Figure 2. Simulation Results

The impact of the missing values can be assessed by taking the average of ten runs and measured in terms of classification accuracy (Table I). It shows that the accuracy was tremendously improved when using K-nearest neighbour and mean with PCA pre-processing method. The confusion matrix shows the correct classification rate against incorrect classification (Table II).

Missing Value Techniques	Omit the samples	Replace with Zero	Replace with Mean	Replace With Knn
Accuracy(%)	68.56	66.67	99.93	99.80

TABLE I. CLASSIFICATION ACCURACY

method	Omit the samples		Replace with Zero		Replace with Mean		Replace With Knn	
class	D	ND	D	ND	D	ND	D	ND
D	76	27	70	31	101	0	100	0
ND	21	29	20	32	0	52	1	52

TABLE II. CONFUSION MATRIX (A, B, C, D)

VI. CONCLUSION

This paper demonstrates the impact of pre-processing and missing values. An experimental result shows that maximum accuracy is achieved with minimum training time. It proves that some combination of missing values and pre-processing the accuracy was tremendously improved. The novel classification method can be applied for different training method that provides high accuracy.

REFERENCES

[1] Arit Thammano and Asavin Meengen, "A New Evolutionary Neural Network Classifier", T.B.Ho, D.Cheung, and H.Liu (Eds.): PAKDD 2005, LNAI 3518, Springer-Verlag Berlin, pp. 249-255. 2005.

[2] Colleen M Ennett, Monique Frize, C.Robin Walker, "Influence of Missing Values on Aritificial Neural Network Performance", Proceedings of Medinfo, 2001, pp.449-453.

[3] Edgar Teufel1, Marco Kletting1, Werner G.Teich2, Hans-Jorg Pfleiderer1, and Cristina Tarin-Sauer3, "Modelling the Glucose Metabolism with Backpropagation Through Time Trained Elman Nets", IEEE 13th Workshop on Neural Networks for Signal Processing, NNSP'03, 17-19 Sept. 2003, pp.789 - 798

[4] Fuluf helo V Nelwamondo, Shakir Mohammed and Tshilidzi Mawala, "Missing Data: A comparison of neural network and expectation maximization techniques", Current Science, Vol 93, No 11, 2007.

[5] Humberto M.Fonseca; Victor H.Ortiz, Agustin LCabrera., "Stochastic Neural Networks Applied to Dynamic Glucose Model for Diabetic Patients", *Ist ICEEE,* 2004, pp.522 - 525

[6] Igor Aizenberg, Claudio Moraga, "Multilayer Feedforward Neural Network Based on Multi-Valued neurons (MLMVN) and a back propagation learning algorithm", Soft Computing, 20 April 2006, pp.169-183.

[7] Junita Mohammad-saleh, Brain S.Hoyle, "Improved Neural Network performance using Principle Component Analysis on Matlab", Int. journal of the computer, the internet and management, Vol.16, No 2, 2008, pp.1-8.

[8] Md.Shahjahan1,M.A.H.Akhand2, and K.Murase1, "A Pruning Algorithm for Training Neural Network Ensembles", SICE 2003 Annual Conference ,Volume 1, 2003, pp.628 – 633.

[9] M.Nawi1, R.S. Ransing and M.R. Ransing, "An Improved Conjugate Gradient Based learning Algorithm for Back Propagation Neural Networks" International journal of Computational Intelligence, 2008.

[10] Rajeeb Dey and Vaibhav Bajpai, Gagan Gandhi and Barnali Dey, "Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus", IEEE Region 10 Colloquium and the 3rd ICIIS, Dec 2008, PID 155.

[11] P.K.Sharpe and R.J.Solly, "Dealing with Missing Values in Neural Network based Diagnostic Systems", Neural Computing and Applications, Springer-Verlag London Ltd, 1995, pp.73-77.

[12] Roelof K Brouwer PEng, PhD, "An Integer Recurrent Artificial Neural Network for Classifying Feature Vectors". ESSANN, Proceedings – Eur. Symp on Artificial Neural Networks, April 1999, D-Facto public, pp. 307-312.

[13] S.G.Mougiakakou,K.Prountzou,K.S.Nikita , "A Real Time Simulation Model of Glucose-Insulin Metabolism for Type 1 Diabetes Patients" , 27th Annual Int. Conf. of the Engineering in Medicine and Biology Society, IEEE-EMBS 17-18 Jan. 2006 pp.298 - 301.

[14] Siti Farhanah, Bt Jafan and Darmawaty Mohd Ali, "Diabetes Mellitus Forecast using Artificial Neural Networks (ANN)", Asian Conference on sensors and the international conference on new techniques in pharamaceutical and medical research proceedings (IEEE), Sep 2005, pp. 135-138.

[15] Syed Muhammad Aqil Burney, Tahseen Ahmed Jilani, Cemal Ardil, "Levenberg-Marquardt Algorithm for Karachi Stock Exchange Share Rates Forecasting", Proc of world Academy of Science, Eng And Tech, Vol 3, Janry 2005.