Data Pre-processing and Neural Network Algorithms for Diagnosis of Type II Diabetes: A Survey

Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse

Abstract*— Diagnosis of type II diabetes in early stages is very challenging task due to complex inter dependence on various factors. It requires the critical need to develop medical diagnostic support systems which can be helpful for the medical practitioners in the diagnostic process. Neural network techniques have been successfully applied to the diagnosis of many medical problems. In this survey we compare the various neural network techniques for the diagnosis of diabetes. The Pima Indian data set is used to study the classification accuracy of the neural network algorithms. The various data pre-processing techniques are surveyed to improve the predictive accuracy of the neural network algorithms.

Index Terms*— Type II diabetes, Pima Indian data set, neural networks, data pre-processing.

I. INTRODUCTION

Diabetes is a major growing health problem all over the world which causes death, especially in industrial and developing countries. This chronic disease may lead to long term complications and death. It can cause high risk of kidney failure, nervous system damage, blindness and heart diseases. In this disease the body does not produce or properly use insulin, the hormone that unlocks the cells of the body, allowing the glucose to enter and fuel them. In the absence of insulin the cells become starved of glucose energy despite the presence of abundant glucose in blood stream. Complications of diabetes are related to blood vessel diseases and generally classify into small vessel disease such as those involving the eyes, called as diabetic retinopathy. It occurs in the patients who have diabetes for at least five years with the leakage of protein in small blood vessels at the back of eyes and the blood in retina.

Disease in blood vessel also causes the formation of small aneurysms and new but brittle blood vessel can lead to retinal scarring and retinal detachments thus impairing vision. Kidney damage from diabetes is called diabetic nephropathy. Initially disease blood vessel in the kidney causes the leakage protein in the urine. Later on kidneys lose their ability to cleanse and filter blood. The accumulation of toxic waste product in blood leads to the need for dialysis. Nerve damage from diabetes is called diabetic neuropathy and it is also caused by disease of small blood vessels. The blood flow to the nerves is limited leaving the nerves without blood flow and they get damaged or die as a result. The symptoms of nerve damage are numbness burning, aching of feet and lower extremities [1]. There are two major types of diabetes type I and type II. Type I is basically diagnosed in children which is usually known as Juvenile diabetes and type II is most common form of diabetes. A patient with type II diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose level [2]. Diagnosis of type II diabetes in early stages is very challenging task due to complex inter dependence on various factors. It requires the critical need to develop medical diagnostic support systems which can be helpful for the medical practitioners in the diagnostic process. World Health Organisation (WHO) predicts that developing countries will bear the brunt of this epidemic in the 21st century. Recently, more then 70% people with diabetes live in low and middle income countries. India has world largest diabetes population, followed by China with 43.2 million. The age of 40-59 years is the largest group which may be affected from diabetes, which is expected to move to 196 million cases at the age of 60-79 years. The estimated global expenditure on diabetes will be at least 428 billion dollors in 2010 and 561 billion dollors in 2030 [3]. This survey paper is organised as follows: Section 2 consist of diabetes data set and methods for data pre-processing. Section 3 describes neural network algorithms and approach for Type II diabetes detection and the performance evaluation through accuracy. Sections 4 conclude the paper.

II. DIABETES DATA AND PRE-PROCESSING

Neural network training can be made more efficient by performing certain pre-processing steps on the network inputs and targets. If a training data set contains irrelevant attributes classification analysis may produce less accurate results. Data preprocessing is required to improve the predictive accuracy. The problem of missing data poses difficulty in the analysis and decision-making processes and the missing data is replaced before applying it to NN model. Without this pre-processing, training the neural networks would have been very slow. 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.

Manuscript Received on October, 2012.

Raj Anand, Department of Computer Science, Oriental College of Technology, Bhopal, India.

Vishnu Pratap Singh Kirar, Department of Electronics & Communication, Truba Institute of Engineering & Information Technology, Bhopal, India.

Dr. Kavita Burse, Department of Electronics & Communication, Oriental College of Technology, Bhopal, India.

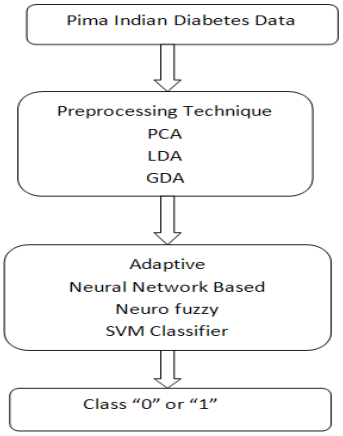

Fig. 1 Data Pre-processing of Pima Indian Data Set

Data pre-processing 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. This section consists of two sub sections: diabetes data set and data pre-processing. The neural network model for type II diabetes is developed using Pima Indian Dataset and the objective of the data-processing is to prepare the data set for neural network algorithms as given in Fig. 1.

A. Diabetes Data

The source of Pima Indian diabetes data set is the UCI machine learning repository [4]. The data source uses 768 samples with two class problems to test whether the patient would test positive or negative for diabetes. All the patients in this database are Pima Indian women at least 21 years old and living near Phoenix Arizona, USA. This data set is most commonly used for comparison of diabetes diagnosis algorithms [5]. The dataset consists of 9 attributes as shown in Table 1. [6]-[7].

Table 1. Attributes of Diabetes Data Set

No	Attribute	Description	Missing Value
1.	Pregnant	A record of the number of times the woman pregnant	110
2	Plasma glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)	5
3.	Diastolic BP	Diastolic blood pressure	35
4.	Triceps SFT	Triceps skin fold thickness (mm)	227
5.	Serum-Insulin	Two hours serum insulin (muU/ml).	374
6.	BMI	Body mass index (weight Kg/height in (mm) ²)	11
7.	DPF	Diabetes pedigree function	0
8.	Age	Age of patient(year)	0
9.	Class	Diabetes on set within five year	0

The class distribution is mentioned in Table 2.

Table 2. Class Distribution

Class Value	Number of instances
0	500
1	268

The total number of instances in the database is 768. The class value '0' is for 500 instances i.e., non diabetic and the class value '1' is for 268 instances i.e., diabetic. The statistical analysis is in Table 3.

Table 3. Statistical Analysis

Attribute No.	Mean	Standard Deviation
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

B. Data Pre-processing Techniques

There are many techniques for data pre processing like k-nearest neighbour (k-NN) method for missing diabetes data, principle component analysis (PCA), linear discriminant analysis (LDA) and fuzzy neural network. k-NN method for classification is a direct approach for classifying the object which is represented as points defined in a feature space [8]. k-NN algorithms are among the most popular methods used in statistical pattern recognition [9]. The technique k-NN method replaces missing values in data with the corresponding values from the neighbouring column in Euclidean distance. If the corresponding value from the nearest neighbour column is also missing the value from the next nearest column is used [10].The models are conceptually simple and empirical studies have shown that their performance is highly competitive against other techniques. The main shortcoming of k-NN is the lack of any probabilistic semantics which would allow posterior predictive probabilities to be employed in for example, assigning variable losses in a consistent manner. In addition the selection of the value of k, the number of nearest neighbours, is not straightforward without resorting to cross-validation (CV) [11]. The lack of a formal framework for choosing the size of the neighbourhood k is a problem. Furthermore, the method can only make discrete predictions by reporting the relative frequency of the classes in the neighbourhood of the prediction point. In [8] a probabilistic framework for the k-NN method is presented that largely overcomes these difficulties. Uncertainty is accommodated via a prior distribution on k as well as in the strength of the interaction between neighbours. These prior distributions propagate uncertainty through to proper probabilistic predictions that have continuous support on (0, 1). The method makes no assumptions about the distribution of the predictor variables. The method is also fully automatic with no user-set parameters and empirically it proves to be highly accurate on many bench-mark data sets [12].

PCA is a very popular pre-processing method. 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. This means that for a given set of data, the features that exhibit the most variance are the most descriptive for determining differences between sets of data. This is accomplished by using eigenvector analysis on either the covariance matrix or correlation matrix for a set of data. Polat and Gunes have used neuro-fuzzy interface system and PCA for diabetes diagnosis. The proposed system has two stages. In the first stage, dimension of diabetes disease dataset that has 8 features is reduced to 4 features using principal component analysis. In the second stage, diagnosis of diabetes disease is conducted via adaptive neuro-fuzzy inference system classifier. The obtained classification accuracy of the system is 89.47% [13]. Linear Discriminant Analysis (LDA) is a widely used technique for pattern classification. It seeks the linear projection of the data to a low dimensional subspace where the data features can be modeled with maximal discriminative power. The main computation involved in LDA is the dot product between LDA base vector and the data, which is costly element-wise floating point multiplications [14].

III. NEURAL NETWORK ALGORITHMS AND ITS APPROACH

There has been a wide research in health care benefits involving the applications of artificial neural networks (NN) to the clinical diagnosis, prognosis and survival analysis in medical domain [15]-[16]. NN classification systems enable medical data analysis in shorter time and remain unaffected by human errors caused by inexperience or fatigue. We have reviewed research publications on diabetes diagnosis using UCI machine learning database. The feed forward multilayer perceptron (MLP) network has been successfully used in medical applications [17].

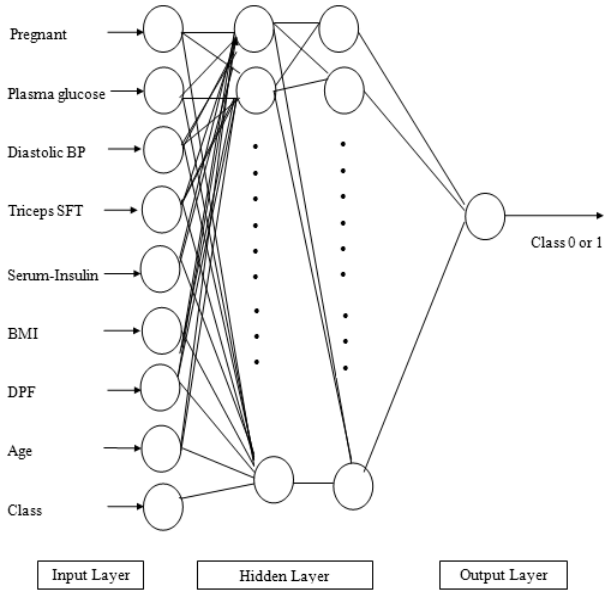

Fig. 2 Feed Forward NN model for Diabetes Diagnosis

The back propagation algorithm used to train the MLP suffers from slow convergence, sometimes getting trapped in local minima and it scales up poorly when the task becomes larger and more complex. The convergence of BP can be improved using Lavenberg-Marquardt (LM) algorithm. The drawback of LM algorithm is that it is expensive both in memory and number of operations required. The network may also get over trained due to memorization effect and its performance suffers on untrained data [18].

In [19] a new generalized feed forward neural network (GFNN) architecture for pattern classification and regression is proposed. The GFNN architecture uses a generalized shunting neuron (GSN) model as the basic computing unit, which includes the perceptron and the shunting inhibitory neuron. The average test error was generally between 20 and 22%, with best case results as low as 17.7% which is comparable to previously reported result of 20%. In [20] a new cascade learning system based on generalized discriminant analysis (GDA) and least square support vector machine (SVM) is proposed. The proposed system consists of two stages. The first stage uses GDA to separate feature variables between healthy and patient (diabetes) data as pre-processing process. The second stage uses LS-SVM for classification of diabetes dataset. The proposed system called GDA-LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation. The obtained classification accuracy is 82.05%. The classification accuracy is further improved by the LDA Morlet Wavelet Support Vector Machine Classifier: LDA–MWSVM of Calisir et al [21]. The feature extraction and feature reduction is done using the LDA method and the classification is achieved using MWSVM classifier. The classification accuracy achieved by the system is about 89.74%.

M Islam *et al* have proposed a back propagation chaotic learning algorithm (BPCL) [22]. Standard BP trains NNs with a constant value of learning rate. On the other hand, biological systems such as human brain involve chaos and the learning rate is modified using a chaotic time series. The size of NN for diabetes diagnosis is 8-4-2. It is shown that BPCL has good generalization ability and also fast convergence rate than BP. Wyk et al have proposed to replace the bounded activation function in feed forward neural network by particle swarm optimization (PSO) that uses adaptive sigmoid activation function and the algorithm is tested on diabetes diagnosis. PSO removes the need for scaling the inputs as required by a bounded activation function of fixed domain [23].

Table 4. Classification Accuracy of Different Methods from Literature

Method	Classification Accuracy	Author
BP	75.2	Statlog
SNB	75.4	Ster et al.
NB	75.5	Ster et al.
kNN	75.5	Grudzinski
MML	75.5	Zarndt
RBF	75.7	Statlog
LVQ	75.8	Ster et al.
Semi-Naïve Bayes (5xCV)	76	Friedman
MLP + BP	76.4	Ster et al.
FDA	76.5	Ster et al.

ASI	76.6	Ster et al.
SMART	76.8	Statlog
GTO DT (5xCV)	76.8	Bennet et al.
BFGS quasi Newton	77.08	Yildirim et al.
LM	77.08	Yildirim et al.
LDA	77.5	Statlog et al.
GD	77.6	Yildirim et al.
SVM (5xCV)	77.6	Bennet et al.
GDA-LS-SVM	79.16	Polat et al.
GRNN	80.21	Yildirim et al.
LDA-MWSVM	89.74	Calisir et al.
MI-MCS-SVM	93.58	Giveki et al.

The fuzzy models integrated with neural networks enhance the learning capabilities of diabetes diagnosis system. LDA-ANFIS intelligent diagnosis system for diabetes proposed in [24] consists of two phases: in the first phase, LDA is used to separate features variables between healthy and diabetes data. In second phase, the healthy and diabetes features obtained in first phase are given to inputs of ANFIS classifier. The LDA-ANFIS is a five layered structure with 7 inputs, 256 rules and 1 output. The input membership functions are the type bell shaped. The sum squared error obtained is 0.000001 in 1543 epochs. The classification accuracy of this LDA-ANFIS intelligent system was obtained about 84.61%.

IV. CONCLUSION

This paper compares NN algorithms for diabetes diagnosis using Pima Indian database taken from UCI machine learning repository. Feed forward NN algorithms like MLP, MLP-LM, GFNN, LS-SVM, MW-SVM and ANFIS are successfully tested for diabetes diagnosis. More effective diagnosis can be attained by handling the missing values of data set by pre-processing methods like k-NN, PCA, LDA, GDA etc. We conclude that data pre-processing and feature extraction coupled with NN models improves the classification accuracy.

REFERENCES

1. http://www.medicinenet.com
2. J.C. Pickup, Williams G. (Eds.,), Textbook of diabetes, Blackwell Science, Oxford.
3. http://www.worlddiabetesfoundation.org
4. A. Frank and A. Asuncion, "UCI Machine Learning Repository
5. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science", (2010)
6. M. Giardina, Y. Huo, F. Auaje, P. McCullagh, "A missing data estimation analysis in type II diabetes databases", Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, 2005, pp. 347-352.
7. H. Temurtas, N. Yumusak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural network: Expert System with Applications", 36, 2009, 8610-8615.
8. A. J. Seibel, "Diabetes Guide WebMD", http://diabetese.webmd.com/guide/oral-glucose-tolerance-test, 2007
9. G. Guo , H. Wang , D.Bell ,Y. Bi and K. Greer," KNN model-based approach in classification", Springer-Verlag Berlin, 2888, 2003, 986-996.
10. M. Lee , T M Gatton and Lee K-K. "A monitoring and advisory system for diabetes patient management using rule based method and KNN, Sensors", 10, 2010, 3934-3953.
11. T Jayalakshmi and A Sahthakumaran", "Improved gradient descent back propagation neural network for diagnosis of type II diabetes mellitus", "Global journal of Computer Science and Technology", 9, 5 (Ver. 2.0), 2010, 94-97.
12. S. Manocha and M.A. Girolami, "An empirical analysis of the probabilistic K-nearest neighbour classifier", pattern recognition letters, 28,2007, 1818-1824.
13. C.C.Holmes and N.M. Adams, "A probabilistic nearest neighbour method for statistical pattern recognition", "Journal of the Royal Statistical Society - Series B: Statistical Methodology", 64, 2, 2002, 295-306.
14. K.Polat and S.Güneş", "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", "Digital image processing", 17, 4, 2007, 702-710.
15. F. Tang, and H.Tao, "Fast linear discriminant analysis using binary bases", "Proc. of the 18th International Conference on Pattern Recognition (ICPR'06)", (2006).
16. P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention, Neural Networks", 15, 2002, 11-39.
17. M. Shanker , M.Y. Hu, and M.S. Hung, "Estimating probabilities of diabetes mellitus using neural networks","SAR and QSAR in Environmental Research", 11, 2000, 2, 133-147.
18. F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks", "Expert system with applications, 36, 2009, 8610-8615.
19. K. Kayer and T. Yildirim, "Medical diagnosis on Pima Indian diabetes using general regression neural neteorks, Proc. International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ ICONIP)", 2003, pp.181-184.
20. G. Arulampalam and A. Bouzerdoum, "A generalised feed forward neural network architecture for classification and regression", "Neural Networks", 16, 2003,561-568.
21. K. POLAT, S.GUNES AND A. ARSLAN , "A CASCADE LEARNING SYSTEM FOR CLASSIFICATION OF DIABETES DISEASE: GENERALIZED DISCRIMINANT ANALYSIS AND LEAST SQUARE SUPPORT VECTOR MACHINE", "EXPERT SYSTEMS WITH APPLICATIONS", 34, 1,2008, 482-487.
22. D.Çalişir and E. Doğanteki., "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications", 38, 2011, 8311-8315.
23. M. Islam, Md R Rana, S.U. Ahmed, A. N. M. Enamul Kabir and Md. Shahjahan, "Training neural network with chaotic learning rate", "Proc. of IEEE International Confererence on Emerging trends in Electrical and Computer Technology (ICETECT 2011)",2011, 781-785.
24. A. B. Van Wyk , and A. P. Engelbrecht, "Lambda-Gamma learning with feed forward neural networks using particle swarm optimization", "Proc. IEEE symposium on swarm intelligence", Paris 2011, 1-8.
25. E. Dogantekin, A. Dogantekin, D. Avci and L.L. Avci, "An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS", "Digital Signal Processing", 20,2010, 1248-1255.