K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set using Higher Order Neural Network and PCA

Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse

***Abstract*— Neural network techniques have been successfully applied for diagnosis of Type II diabetes. We propose a K-Fold cross validation method for classification of PIMA Indian diabetes data set. The classification accuracy is computed with PCA preprocessing and higher order neural network. The problem of missing data in the analysis and decision making process is handled through PCA. PCA also scales the data in the same range of values.**

***Index Terms*— Type II diabetes, Pima Indian data set, higher order neural networks, data pre-processing, K cross validation, PCA.**

I. INTRODUCTION

Diabetes is characterized by persistently high levels of blood glucose resulting from defects of insulin secretion, insulin action or both. In early stages it is challenging task to diagnose diabetes due to complex inter-dependence on various factors. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. Neural network techniques have been effectively used as a medical decision making tool. There has been a wide research in health care benefits involving the applications of artificial neural networks (ANN) to the clinical diagnosis, prognosis and survival analysis in medical domain [1]-[2]. 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 [3]. k-NN algorithms are among the most popular methods used in statistical pattern recognition [4]. 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 [5]. 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. The lack of a formal framework for choosing the size of the neighbourhood k is a problem. 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% [6]. 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 [7]. This paper is organised as follows: Section 2 proposes the methodology comprising of higher order neural network (HONN) model with PCA data processing for detection of diabetes on Pima Indian data set. Section 3 discusses the simulation and results and section 4 concludes the paper.

II. METHEDOLOGY

Higher Order Neural Network (HONN)

Traditional neural networks use linear aggregation, which are mathematically inconvenient. The exhaustive architecture of these networks can be reduced by incorporating non-linearity in aggregation, which generates higher order terms. One such method is to use multiplication instead of summation at the node.

An error back propagation (BP) based learning using a norm-squared error function is described as follows [8]. The aggregation function is considered as a product of linear functions in different dimensions of space. A bipolar sigmoidal activation function is used at each node. This kind of neuron itself looks complex in the first instance but when used to solve a complicated problem needs less number of parameters as compared to the existing conventional models. Figure 2 shows a feed forward higher order neural network (HONN).

Manuscript received on January, 2013.

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.

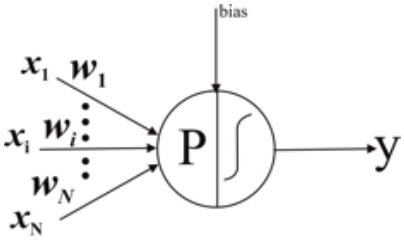

Fig. 1 Node Structure of MNN

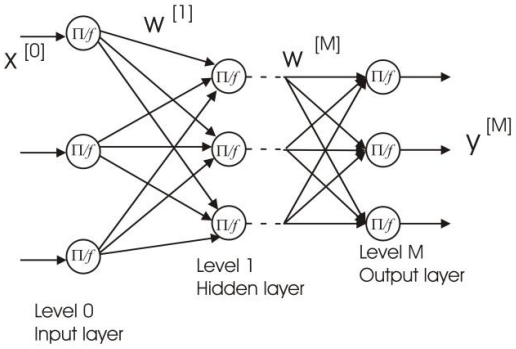

Fig. 2 Higher Order Neural Network

Here the operator P is a multiplicative operation and the aggregation u before applying activation function is given by:

$$u = \prod_{i=1}^{n} (w_i x_i + b_i) \tag{1}$$

The output at the node y is given by

$$y = f(u) = \frac{1-e^{-u}}{1+e^{-u}} \tag{2}$$

The mean square error is given by

$$E = \frac{1}{2N} \sum_{p=1}^{N} (y^p - y_d^p)^2 \tag{3}$$

Where, p is the number of input patterns.

The weight update equation for the split complex back propagation algorithm is given by

$$\Delta w_i = -\eta \frac{dE}{d w_i} = -\frac{1}{2}\eta (y-d)(1+y)(1-y) \frac{u}{(w_i x_i + b_i)} x_i \tag{4}$$

where, η is the learning rate and d is the desired signal. The bias is updated as

$$\Delta b_i = -\eta \frac{dE}{d b_i} = -\frac{1}{2}\eta (y-d)(1+y)(1-y) \frac{u}{(w_i x_i + b_i)} \tag{5}$$

$$w_i^{new} = w_i^{old} + \Delta w_i \tag{6}$$

$$b_i^{new} = b_i^{old} + \Delta b_i \tag{7}$$

The weights are updated after the entire training sequence has been presented to the network once. This is called learning by epoch. The algorithm is extended to train multi layer HONN feed forward neural network.

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. Preprocessing also scales 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. The source of Pima Indian diabetes data set is the UCI machine learning repository [9]. 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. The dataset consists of 9 attributes as shown in Table 1.

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

III. SIMULATION AND RESULTS

Divide the data into K roughly equal parts.

Testing	Training	Training	Training	Training

Training	**Testing**	Training	Training	Training

.
.
.

Training	Training	Training	Training	**Testing**

Fig. 3 K-Fold cross validation of data set

A K-fold partition of the data set is created. For each K experiments, K-1 folds are used for training and the remaining one for testing. Then the average error across all K trials is computed as follows:

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set K-1 times. The variance of the resulting estimate is reduced as K is increased [10]. The scatter plot of data after applying PCA is indicated by the red region of fig. 4. The MSE computed by K-fold cross validation method for training is 7.5257e-04 for normalization and 1.0386e-05 for PCA. The MSE computed by K-fold cross validation method for testing is 2.4101e-04 for normalization and 1.4219e-05 for PCA.

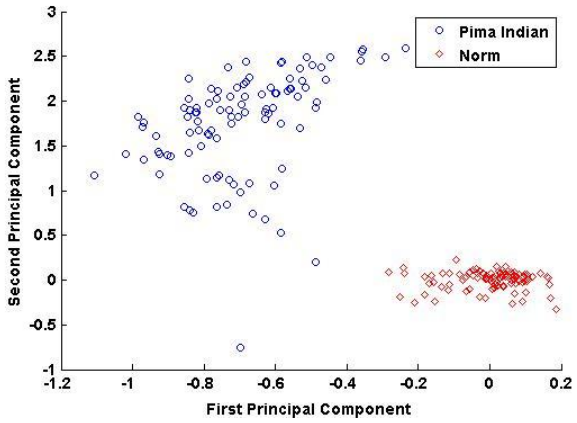

Fig. 4 Principal Component Analysis of data set

Table 2. Training Performance Table

K	MSE for Normalization	MSE for PCA
1	7.6103e-04	1.0374e-05
2	7.5042e-04	1.0333e-05
3	7.5720e-04	1.0277e-05
4	7.5247e-04	9.7367e-06
5	7.5332e-04	1.0619e-05
6	7.5212e-04	1.0629e-05
7	7.5217e-04	1.0789e-05
8	7.4180e-04	1.0331e-05

E(norm) = 7.5257e-04
E(pca) = 1.0386e-05

Table 3. Testing Performance Table

K	MSE for Normalization	MSE for PCA
1	6.5677e-04	2.1523e-05
2	5.7988e-04	2.4154e-05
3	5.4197e-04	1.2114e-05
4	7.5869e-05	3.6888e-05
5	2.1202e-05	1.4273e-06
6	8.1666e-06	4.5687e-07
7	1.9067e-05	7.5647e-06
8	3.2579e-05	9.6264e-06

E(norm) = 2.4101e-04
E(pca) = 1.4219e-05

IV. CONCLUSION

In this paper we proposed a novel approach to Pima Indian diabetes data diagnosis using PCA and HONN. 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.

REFERENCES

1. P.J.G. Lisboa, "A review of evidence of health benefit from artificial neutral networks in medical intervention, Neural Network," pp. 11-39, 2002.
2. M. Shanker, M.Y. Hu, and M.S. Hung, "Estimating probabilities of diabetes mellitus using neural network", "SAR and QSAR in Environment Research," vol. 2, pp. 133-147, 2000.
3. G. Guo , H. Wang , D.Bell ,Y. Bi and K. Greer," KNN model-based approach in classification", Springer-Verlag Berlin, 2888, 2003, 986-996.
4. 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.
5. 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.
6. 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.
7. F. Tang, and H.Tao, "Fast linear discriminant analysis using binary bases", "Proc. of the 18th International Conference on Pattern Recognition (ICPR"06)", (2006).
8. R. N. Yadav, P. K. Kalra and J. John," Time series prediction using single multiplicative neuron model," Applied Soft Computing, Vol 7, pp 1157-1163, 2007.
9. A. Frank and A. Asuncion, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science," 2010.
10. http://www.cs.cmu.edu/~schneide/tut5/node42.html