Analysis and Prediction of Diabetes Mellitus Using PCA, REP and SVM

Tarun Jhaldiyal, Pawan Kumar Mishra

***Abstract*— Disease diagnosis is a major problem area for researchers for a long time. To accurately diagnosis a disease is of prime concern for a doctor. To help the medical personnel with the diagnosis tool, many engineering techniques have evolved in the past. There are various conventional methods of disease diagnosis, but application of soft computing technique with information technology has given a new dimension to this area. In this particular work two different approaches have been proposed for the classification of subjects into two classes namely: Diabetic & Non-diabetic. The techniques undertaken are PCA + REP & PCA + SVM. The results obtained are very interesting and show improvement from the previous works. There is enough scope for improvement in this field and with the advent of faster and more accurate learning techniques the results can surely be improved considerably.**

***Index Terms*— Principal Component Analysis (PCA), Reduced Error Pruning Tree (REP Tree), Support Vector Machine (SVM).**

I. INTRODUCTION

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. But, expert systems and different artificial intelligence techniques for classification also help experts in a great deal . Most of the work related to machine learning in the domain of diabetes diagnosis is concentrated on the study of the Pima Indian Diabetes dataset in the UCI repository. In this paper, two classifier techniques with Principal component analysis are implemented for the forecasting of Diabetes and concluded with best forecasting techniques which has a maximum accuracy. Implemented techniques are listed below.

1. Principal Component Analysis (PCA) with REP Tree
2. PCA with SVM (Support Vector Machine)

II. THE DATA

The dataset which we use in our work is Pima Indians Diabetic database from UCI Repository of Machine Learning Databases. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. The binary response variable takes the values '0' or '1,' where '1' means a positive test for diabetes and '0' is a negative test for diabetes. There are 268 (34.9%) cases in class '1' and 500 (65.1%) cases in class '0.'

III. PRINCIPAL COMPONENT ANALYSIS

Principal Components Analysis (PCA) is used to compress data in such a way that the least information is lost. It does so by truncating data and thereby leaving out the data which is of the least importance to the information stored in the data. This PCA process is called dimensionality reduction, because a vector x which contains the original data and is N-dimensional is reduced to a compressed vector c which is M-dimensional, where M<N.

IV. BLOCK DIAGRAM

The overall system block diagram is shown in figure.1. In the first part the original diabetic database consisting of the entire 8 feature is shown. This dataset will be utilized for all the classification tasks throughout the study. The entire experimental work can be divided into two major sections:

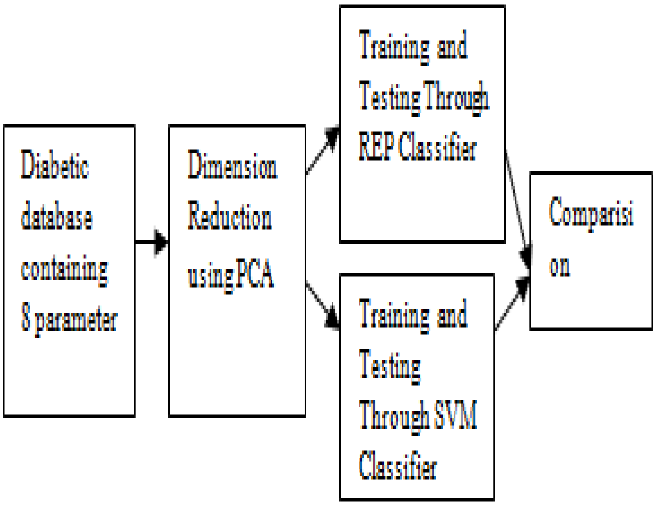

Diabetic database containing 8 parameter
Dimension Reduction using PCA
Training and Testing Through REP Classifier
Training and Testing Through SVM Classifier
Comparisi on

Figure 1. System Block Diagram

.

A. *PCA WITH REP*

To improve the accuracy of classification some kind of modification was required to be done in the classification technique. For that purpose Principal Component Analysis

Manuscript received August 23, 2014.

Tarun Jhaldiyal, Department of Computer Science & Engineering, Uttarakhand Technical University, Dehradun, India, 9997311303.

Pawan Kumar Mishra, Department of Computer Science & Engineering, Uttarakhand Technical University, Dehradun, India, 9411413650.

(PCA) was used along with REP. PCA is a simple, non-parametric method for extracting relevant information from confusing data sets. Here, first the dataset is reduced to a lower dimension using Principal Component Analysis. After this the reduced dataset is trained using the REP classifier and then the testing of the REP model is carried out. Here 8 parameters are used for determining whether a sample is diabetic or non-diabetic, it might be possible that some of the features may play a more significant role in diagnosis than others and hence if the pattern pertaining to the performance of each of the features can be studied, the dimension of the dataset can be reduced by removing the less relevant features from the complete set. Now after performing PCA on this dataset we get a reduced dataset with lesser number of features. The number of features to be selected can be set as per our requirement as in PCA the features which contribute to the maximum amount of variation of the dataset are ordered first and those with least variation are set to the end. Here we have selected the features which are responsible for maximum variance of the total variation of the dataset and the remaining features which contribute less variation are discarded. These criteria yielded us 3 parameters i.e. there are 3 parameters which contribute maximum variation of the whole dataset while the other five contribute less variation so they can be neglected without much loss of information.

Now PCA reduce the number of features to 3 from a total of 8. Now after the dimension of the dataset has been reduced, this reduced dataset will be used for classification using REP.

The features selected are:

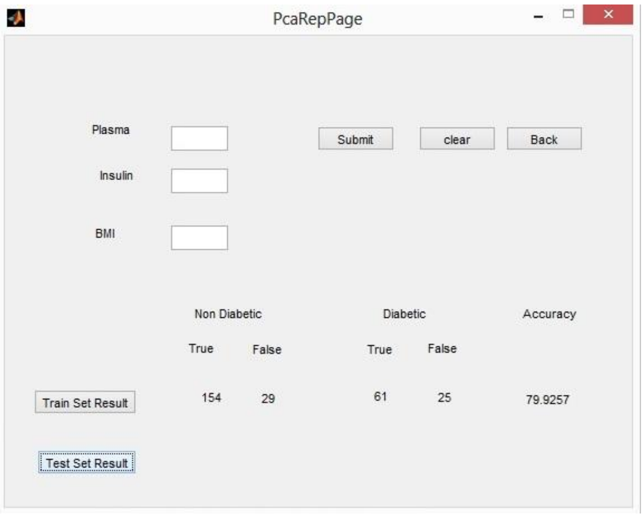
1. Plasma glucose concentration a 2 hrs in an oral glucose tolerance test (GTT).
2. Insulin (Insu).
3. Body Mass Index (BMI).

B. PCA WITH SVM

The modifications which were done in case of REP classifier using PCA were also implemented for SVM so that even comparison can be made between the classifiers. Similar, the dataset was first reduced to a lower dimension and the reduced dataset is used for training through SVM. The results of the two classifier techniques are tabulated and compared at the end in order to find the best classifier technique out of the two proposed ones.

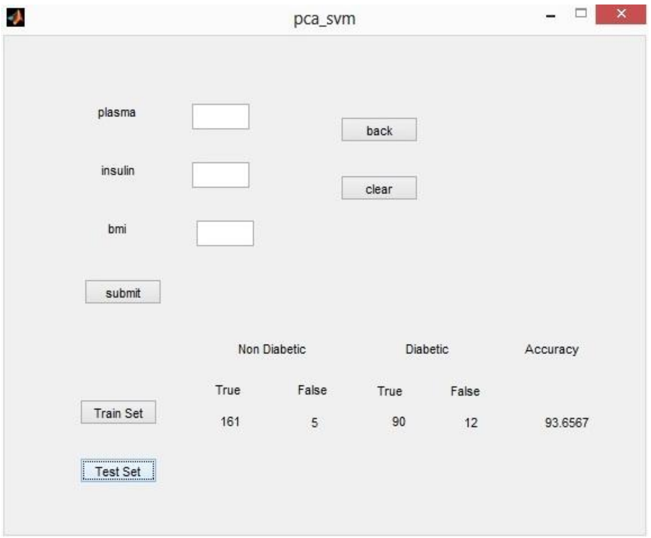
C. IMPLEMENTATION OF BOTH CLASSIFIER TECHNIQUES IN MATLAB GUI

A graphical user interface (GUI) is a graphical display in one or more windows containing controls, called components that enable a user to perform interactive tasks. The user of the GUI does not have to create a script or type commands at the command line to accomplish the tasks. Unlike coding programs to accomplish tasks, the user of a GUI need not understand the details of how the tasks are performed



	Non Diabetic		Diabetic		Accuracy
	True	False	True	False	
Train Set Result	154	29	61	25	79.9257

Figure 2. PCA with REP, Matlab GUI



	Non Diabetic		Diabetic		Accuracy
	True	False	True	False	
	161	5	90	12	93.6567

Figure 3. PCA with SVM, Matlab GUI

Table 1- Comparision Of Different Classifier Techniques.

Classification Technique	Test Results				Accuracy (%)
	Non Diabetic (0's)		Diabetic (1's)		
	True	False	True	False	
PCA with REP	154	29	61	25	79.93
PCA with SVM	161	5	90	12	93.66

From the simulation results it clear that PCA with SVM approach is better than PCA with REP approach.

V. CONCLUSION

We have analyzed the Diabetic Patient data on the basis of data mining techniques. From the above study it has been observed that PCA with SVM perform well for diabtes mellitus prediction. Also the accuracy for PCA with REP classifier is good but in terms of accuracy PCA with SVM performs better than other classifier.

In future we can use other classifier techniques with PCA to improve the accuracy and testing time of result. Or we can use combination of other classifier techniques with other dimension reduction techniques to improve the accuracy and testing time of result.

REFERENCES

[1] Ashis Pradhan, “Support Vector Machine-A Survey”, International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 8, August 2012).

[2] Aqueel Ahmed and Shaikh Abdul Hannan, “Data Mining Techniques to Find Out Heart Diseases: An Overview”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[3] Mohamed,W.N.H.W, Salleh, M.N.M. ; Omar, A.H., “A comparative study of Reduced Error Pruning method in decision tree algorithms ”, International Conference on Control System, Computing and Engineering (ICCSCE), 2012 IEEE , 23-25 Nov. 2012, pages(392–397).

[4]Hanaa I. Elshazly, Neveen I. Ghali, Abir M. El Korany, Aboul Ella Hassanien, ‘’ Rough Sets and Genetic Algorithms: A hybrid approach to breast cancer classification ’’, 2012 World Congress on Information and Communication Technologies, 2012 IEEE.

[5] Cheng W., Leu S., Y Cheng, T. Wu,C.Lin, “Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan’s construction industry”, Accident Analysis and prevention, Vol,48,pp.214-222, 2012.

[6] 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

[7] Olaiya Folorunsho, “ Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database ” , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[8] V. Anuja Kumari, R.Chitra, “ Classification Of Diabetes Disease Using Support Vector Machine ”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, Issue 2, March -April 2013, pp.1797-1801

[9] Metin Zontul, Fatih Aydin, Gokhan Dogan, Selcuk Sener, Oguz Kaynar, “ Wind Speed Forecasting Using REP Tree And Bagging Methods In Kirklareli-Turkey ”, Journal of Theoretical and Applied Information Technology 10th October 2013. Vol. 56 No.1

[10] C. M. Velu, K. R. Kashwan “Visual Data Mining Techniques for Classification of Diabetic Patients”, 2013 3rd IEEE International Advance Computing Conference (IACC).

[11] Makhil Jabbar, Dr B.L Deekshatulu, Dr Priti Chandra, “Heart Disease Prediction using Lazy Associative Classification “, IEEE 2013.

[12] Syed Umar Amin, Kavita Agarwal and Dr. Rizwan Beg, “ Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors ”, Proceedings of IEEE Conference on Information and Communication Technologies (ICT 2013).

[13] Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, “Diabetes Mellitus Forecast Using Different Data Mining Techniques ”, 2013 4th International Conference on Computer and Communication Technology (ICCCT), 2013 IEEE.

First Author :

Mr. Tarun Jhaldiyal, Tarun Jhaldiyal is a Graduate in Computer Science Engineering from Uttarakhand Technical University, Dehradun. Presently he is pursuing M.Tech. in Computer Science Engineering from Uttarakhand Technical University, Dehradun. His area of interest include Autamata, Compiler, Computer Networks, Data Structure & Data Mining.

Second Author :

Assistant Professor Pawan kumar Mishra, Pawan Kumar Mishra is a Graduate in Computer Science Engineering from Dr. B.R Ambedkar University, Agra, ana Post graduate in Computer Science Engineering from, Uttarakhand Technical University, Dehradun. Presently he is pursuing PhD in Computer Science Engineering from, Uttarakhand Technical University, Dehradun. His area of interest include Data Structure, Object Oriented System, D.B.M.S, Mobile Computing, Software Engineering.