International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Diagnosis of Pima Indians Diabetes by LDA-SVM Approach: A Survey

Ankita Parashar
Mtech Scholar CSE/OCT,
Rgtu/ Bhopal/ India

Kavita Burse
Director OCT,
Rgtu / Bhopal/ India

Kavita Rawat
Assistant Prof. OCT,
Rgtu / Bhopal/ India

***Abstract*- Prediction of diabetes in late stages is a very complex task, as several environmental factors are responsible for it. There are certain medical diagnosis systems which are helpful for physicians to know whether a patient is diabetic or not. In our study the proposed system is LDA-SVM. In first stage Linear Discriminant Analysis is used for reducing the feature variables of the Pima Indians diabetes dataset. In next stage two techniques are used for the classification- Support Vector Machine and Feed Forward Neural Network. Comparatively SVM gives better classification accuracy than FFNN.**

Keywords*- *Linear Discriminant Analysis, Support Vector Machine, Feed Forward Neural Network.

I. INTRODUCTION

Diabetes is a major health concern affecting all age groups all over the world. Diabetes causes death, and may give rise to heart disease, blindness, kidney disease and other health related problems. It causes mainly due to genetics or certain environmental conditions like obesity, lack of physical exercise, eating habits, unhealthy life style etc. However it can be controlled by proper control over diet and regular exercise, yoga etc [1].

Food which contains Carbohydrates is turned into glucose. Insulin hormone, produced by pancreas, helps glucose to move from blood streams to body cells. In the absence of insulin or insufficient Insulin, blood glucose level increases than the normal, due to accumulation of glucose in blood [2]. There are two types of diabetes namely- type-1 and type-2. Type-1 diabetes occurs in children called Juvenile diabetes, which can be easily diagnosed as the cells (of pancreas) which produce insulin are destroyed, resulting in deficiency of insulin. Type-2 diabetes occurs in adults. In type-2 diabetes, sufficient insulin is not produced in pancreas, in order to regulate the glucose level [3].

This is a very common type of diabetes. In most cases, people are not aware of the disease. Due to such complexities it is hard to diagnose diabetes by the physicians. The physician usually takes decision by considering two factors- either by evaluating current test results of patients or by comparing the results with other patients and analyzing the previous decisions made under the same condition [4]. Therefore we need an intelligent diagnosis system in order to help physicians in diabetes diagnosis which is time saving and efficient.

Diabetes data

In our study Pima Indians diabetes dataset is taken from UCI machine learning repository. The dataset consists of 768 Samples; with classes to test the patients.Class1 is of normal patients with 500 samples, and Class2 contains diabetic patients with 268 samples. All patients in this dataset are Pima Indian women whose age is at least 21 years old and living near Phoenix, Arizona and USA [1]. The dataset consists of 9 attributes as shown in tables below [1] [5].

Table1. Attributes of Dataset

Features	Description
Attribute1	Number of times pregnant
Attribute2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Attribute3	Diastolic blood pressure(mm Hg)
Attribute4	Triceps skin fold thickness (mm)
Attribute5	2-hour serum insulin(mu U/ml)
Attribute6	Body mass index(weight in kg/(height in m)^2)
Attribute7	Diabetes pedigree function
Attribute8	Age(years)
Attribute9	Class(0 or 1)

Table2. Class distribution

Class value	Number of Samples
0	500
1	268

Table3. Brief 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

II. RELATED WORK

1. In this paper [5], Esin Dodantekin et al have proposed an intelligent system based on Linear Discriminant Analysis (LDA) and Adaptive Network Based Fuzzy Inference System (ANFIS) for diabetes diagnosis. In first stage LDA is used to separate feature variables between healthy and diabetic patient data. The features obtained in the first

stage are given as input to ANFIS for classification in the second stage. ANFIS is a structure with five layers, 256 rules and single output. The classification accuracy obtained of the proposed LDA-ANFIS diagnosis system was about 84.61%.

2. In this paper [6], Kelam Polat et al have proposed a learning system based on Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM). In first stage GDA is used to separate the feature variables between healthy and diabetic patient's data as pre-processing step. The second stage used LS-SVM for the classification of Diabetes dataset. The proposed system called GDA-LS-SVM which obtained 82.05% accuracy with 10 fold cross validation.
3. In this paper [7], Manjeevan Seera et al have proposed a hybrid intelligent system that consists of three techniques- Fuzzy Min-Max Neural network (FMM), Classification and Regression Tree (CART) and Random Forest Model (RF) to examine its efficacy over three medical benchmark datasets namely- Wisconsin Breast Cancer, Pima Indians Diabetes, and Liver Disorder. The experiment results reveal that FMM-CART-RF yields better performance in terms of accuracy, sensitivity and specificity as compared to FMM and FMM-CART.
4. In this paper [8], Mostafa Fathi Ganji et al have proposed called FCS-ANTMINER based on Ant Colony Optimization (ACO) which is used for the extraction of set of fuzzy rules for the diagnosis of diabetes disease. The method is applied in two stages training and testing stage. The obtained classification accuracy of the proposed method is 84.24%.
5. In this paper[9], Humar Kahramanli et al have proposed a hybrid neural network which include Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN) for the classification of two medical dataset- Pima Indians diabetes and Cleveland heart disease, which obtained classification accuracy of 84.24% and 86.8% respectively, using k-fold cross validation.
6. In this paper[10],Sean N. Ghazavi et al have proposed three methods based on fuzzy modeling namely- fuzzy k-nearest neighbor algorithm, fuzzy clustering- based modeling and Adaptive network-based fuzzy inference system(ANFIS). These techniques were employed over Wisconsin breast cancer dataset and Pima Indians diabetes dataset with classification accuracy of 97.17% and 77.65% respectively

III. PROPOSED WORK

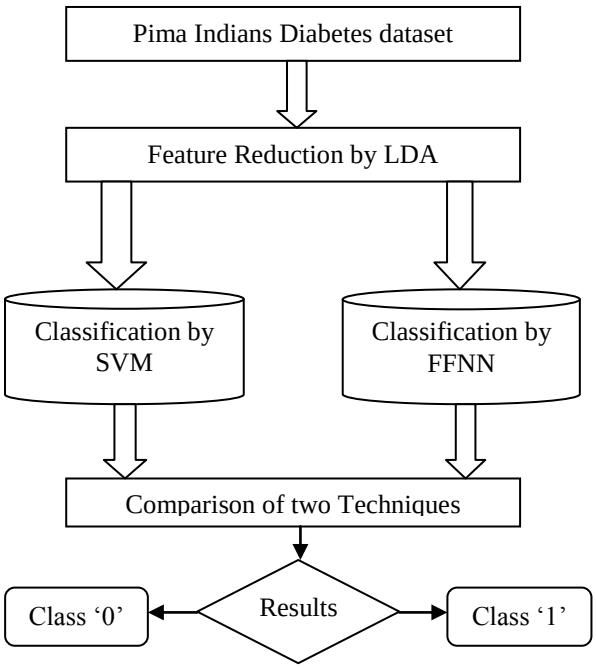

Fig.1

Linear Discriminant Analysis: LDA is used to separate feature variables between healthy and patients data. This is a technique of reducing feature variables, and benefits the supervised learning approach to find a set of base vectors [5].These base vectors are a ratio between the class scatters of the set of training samples. Base vectors can be attained by representation in subspace of LDA using a simple linear projection [4].The main evaluation included in LDA is the dot product between base vector and the data.

Support Vector Machine: SVM is used in supervised learning process. It constructs hyperplane surfaces that classify examples with a largest margin, and predicts whether the examples fall into one class or other separated by a margin. If the data points are non-linearly separable, they are transposed to higher dimensional feature space, which are introduced by kernels [4]. In our study we use default linear kernel and the classifier is called Linear-SVM, which classifies the data with greater efficiency.

Feed Forward Neural Network: Feed Forward neural network is a widely used neural network applied in the fields of medical diagnosis, speech recognition and other information processing applications. FFNN is a layered structure with an input, hidden and an output layer. In our study it is implemented for the detection of diabetes, using log sigmoidal function whose accuracy is slightly lower than Support Vector Machine.

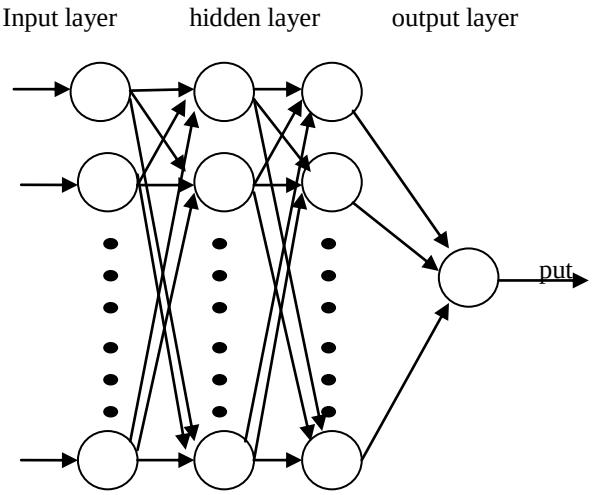

Fig.2

Attributes of Pima Indians diabetes dataset are given as input to the input layer. And the output layer shows the results and the class is represented by either "0" or "1" i.e. healthy(tested as negative) or diabetic patients(tested as positive) respectively.

As we see in Fig.1, the attributes of Pima Indians diabetes dataset are reduced in the first stage using Linear Discriminat Analysis (LDA). The remaining attributes after feature reduction are given as input to the SVM Classifier and Feed Forward Neural Network. The two techniques are then compared on the basis of Classification accuracy. The Classifier evaluating better performance shows the results by dividing the instances of diabetes into two classes tested positive(indicated by 1) and negative(indicated by 0).

IV. CONCLUSION

Diabetes disease gives rise to our related diseases in human body, which is hazardous to our health. It is necessary to detect such symptoms. To make diabetes diagnosis easier for Physicians, there are several methods. To attain greater performance we have reduced attributes of Pima Indians diabetes dataset using LDA. Then we have implemented and compared two different techniques Support Vector Machine using Linear kernel function and Feed Forward Neural Network. On comparison of these two methods, we conclude that SVM proved much better than FFNN.

REFERENCES

[1] Hasan Temurtas, Nejat Yumusak and Feyzullah Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", "Expert Systems with Applications", 36,2009,8610-8615.

[2] Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification", "Expert Systems with Applications",40,2013,5402-5412.

[3] B.M Patil, R.C Joshi and Durga Toshniwal, "Hybrid prediction model for Type-2 diabetic patients", "Expert Systems with Applications", 37 (2010) 8102-8108.

[4] Duygu Calisir and Esin Dogantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", "Expert Systems with Applications", 38(2011) 8311-8315.

[5] Esin Dogantekin, Akif Dogantekin and Derya Avci and Levent Avci, "An intelligent diagnosis system for diabetes on Linear Discriminat Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS", "Digital Signal Processing", 20(2010) 1248-1255.

[6] Kemal Polat, Salih Gunes and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine", "Expert Systems with Applications", 34(2008) 482-487.

[7] Manjeevan Seera and Chee Peng Lim, "A hybrid intelligent system for medical data classification", "Expert Systems with Applications", 41(2014)2239-2249.

[8] Mostafa Fathi Ganji and Mohammad Saniee Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis", "Expert Systems with Applications",38(2011)14650-14659.

[9] Humar Kahramanli and Novruz Allahverdi, "Design of hybrid system for diabetes and heart diseases", "Expert Systems with Applications", 35(2008)82-89.

[10] Sean N. Ghazavi and Thunshun W.Liao, "Medical data mining by fuzzy modeling with selected features", "Artificial Intelligence in Medicine", (2008) 43, 195-206.