Classification Of Diabetes Patients Using Kernel Based Support Vector Machines

G. A . Pethunachiyar

The Tamilnadu Dr.Ambedkar Law University,
Chennai,India
E-mail: apethu@gmail.com

Abstract : **Diabetes mellitus (DM) is a collection of metabolic diseases that influence the human pressure significantly worldwide. Detection of patient with diabetes at early stage is the most crucial task and helps to avoid the risk of the people from the diseases that lead to cause death. In diabetes research, the machine learning plays the important role in detecting the diseases at an early stage. There are more machine learning algorithms used for the research. Support Vector Machines (SVM) is the most successful and widely used algorithm. In this paper, SVM with different kernel functions are applied. SVM with linear kernel showed the highest accuracy value for the classification of diabetes.**

Keywords - Diabetes, Kernel, Linear, Machine Learning, Support Vector Machines

I. INTRODUCTION

Diabetes, a chronic metabolic non communicable disease (NCD), has reached 415 million worldwide and in 2040, it is expected to increase by 642 million. India is the second highest number of people with diabetes in the world. Diabetes caused if pancreas fail to produce enough insulin or responding of the human body cells not proper for the insulin produced. The people with diabetes are affected personally and economically. If patients with diabetes are left untreated then it leads to the cause of visual impairments, heart problems and skin problems. For the past five years, remarkable progress is placed by machine learning in health-care and medical field. Because of efficiency produced by machine learning, it is used in many real time applications. So, it is mandatory for the physicians to identify and diagnose at an early stage for the peaceful life of human being. In entire world, the medical field contains the large volume of data with different formats. Processing and mining the knowledge from those data is a challenging task. Machine learning algorithms are the appropriate choice for detecting the diabetes at an early stage with significant level of confidence to provide the individuals with good quality life. In this study, we applied the SVM based classification techniques to identify the patients with diabetes. This paper is arranged as follows:

Section 2 provides the related work
Section 3 provides the proposed methodology
Section 4 presents a results and discussion, and
Section 5 providees conclusions.

II. RELATED WORK

A literature survey shows that lot of results on diabetes all over the world. Diabetic data set from CPCSSN database taken for analysis [1]. The researcher used the bagging ensemble techniques using J48 for classifying different age groups patients with diabetes mellitus. Regression based data mining technique used in [2] for predictive analytic in diabetic treatment. Here, Oracle Data Miner tool was used for prediction analysis. The study was conducted in 2012 [3] for predicting diabetes by using the common risk factors. For the performance analysis, different classification techniques such as decision tree, Neural Networks and logistic regression were considered. The logistic model outperformed the other two in accuracy rate. Common attributes taken for study are family history, characteristics and lifestyle risk. In [4], Investigation performed among participants who registered in an adapted Diabetes Prevention Program (DPP) for weigh reduction. The findings in the above support the participants for dietary monitoring by themselves and motivate the participants to increase the levels in exercise. The investigation also performed to analyze the treatment for hypertension based on regression technique. Study conducted in [5] shows the importance of machine learning methods in medical field and it gradually increasing over the years. The survey proves that the large amount of data was generated from the wide research carried out in all aspects of diabetes and machine learning algorithms had an important role in most of the applications. It also proves that Support vector machines (SVM) is the mostly used algorithm. Even though, DM can be categorized into many types, the major types are type 1 diabetes and

type 2 diabetes. Type-2 diabetes is mainly caused by the resistance of insulin in our body and it is the general type in ninety percent of the diabetes patients. Ten percentages of the diabetic patients are affected by Type-1 diabetes. From the study [6], it is observed that Diagnosis of diabetes patients depends on blood glucose levels.

III. MATERIALS AND METHODS

In Fig 1, the simulation model of a proposed algorithm is depicted.
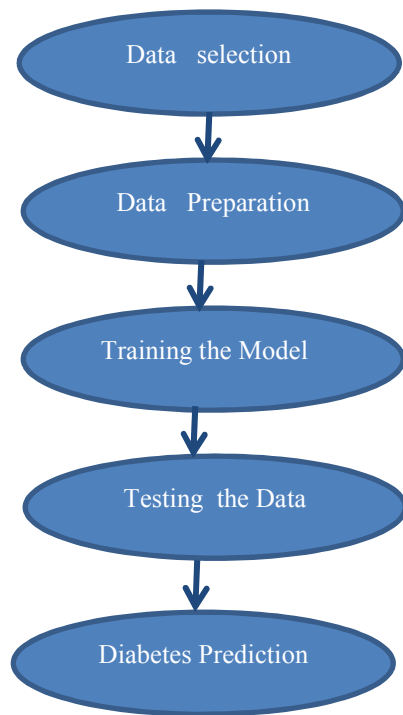
Fig. 1: Proposed Model

3.1 Data Collection

Data sets are taken from UCI Machine Learning Repository that is available in the web link titled Data sets. There are many data sets available for diabetes. Data set with 332 records were considered for study. There are 7 input variables and one output variables. The output variable value is 1 and 2. The value 1 indicates the patients without diabetes and 2 for patients with diabetes.

3.2 Tools and Techniques

There are many types of machine learning algorithm are available and each has its own advantages and disadvantages. Here, a predictive analysis technique is used to predict the patients with diabetes in more accuracy rate. The process of detecting the diabetes patients using machine learning algorithm involves the following five steps.

A) **Data Selectio**n: The first step in this study is selecting the data from UCI repository by rectifying the errors like missing values, inconsistent in data and wrong information in the data set.

B) **Data Preparation**: Databases stored in excel and text format are divided into training and testing data. 70% and 30% of data are considered for training and testing respectively.

C) **Machine Learning**: This is the crucial stage for obtaining the research objectives. Machine learning algorithms like Naïve Bayes, Logistic Regression, Support Vector Machines and Neural Networks etc., are available for prediction.

Support Vector Machines (SVM) is selected for efficient prediction.

D) **Building Model**: Model is built for the training data set using the SVM.

E) **Prediction and Deployment**: Test data is applied to the built model to make predictions on the resultant value produced.

3.3 Proposed Algorithm

Support Vector Machines (Kernels)

In this study, Kernel based SVM is implemented. There values for the kernel-type parameter are considered for investigation. They are linear, radial and polynomial.

A) Linear Kernel SVM

Let x be the input and support vector for each x be x1. The kernel can be determined as

$$K(x, x1) = sum(x * x1).$$

Measuring distance between the data and the support vectors is called as kernel. Grouping the Inputs in linear is known as Linear Kernel. The kernel types like Polynomial and Radial changes the input into its higher dimensions.

B) Polynomial Kernel SVM

Polynomial kernel can be written as

$$K(x, x1) = 1 + sum(x * x1) \char`^d$$

It allows the curved lines for input space.

C) Radial Kernel SVM

The radial kernel can be written as follows.

$$K(x, x1) = exp (-gamma * sum ((x - x1^2))$$

The default value taken for gamma is 0.1 but the gamma value lies between 0 and 1.It can be used to create complex regions in two dimensional spaces.

IV. RESULTS AND DISCUSSION

All the parameters in the data set are taken for study. The accuracy of prediction is calculated using confusion matrix. Confusion matrix of all the methods is displayed in Table 1. Confusion matrix is produced for the testing data set. It has 100 records. Accuracy of different methods of a kernel is produced in Table 2. It contains specificity, sensitivity and accuracy. From Table 2, we can conclude that the linear based kernel produces 100% accuracy in predicting the diabetes patients.ROC curve is used to evaluate the different kernel functions in SVM. Here, Sensitivity (True Positive Rate) and Specificity (False Positive Rate) for different parameters is plotted using R tool is specified in Fig 2.

Table 1: Performance measures of different kernel functions

Confusion Matrix								
SVM Kernel="Linear"			SVM Kernel="Radial"			SVM Kernel="Polynomial"		
	F	T		F	T		F	T
F	65	0	F	65	0	F	65	0
T	0	35	T	1	34	T	10	25

Table 2: Accuracy Rate produced by different kernel functions

Kernel Type	Correctly Identified Instances	Specificity	Sensitivity	Accuracy
Linear	100	1.0	1.0	100%
Radial	99	1.0	0.98	99%
Polynomial	89	0.87	1.0	90%

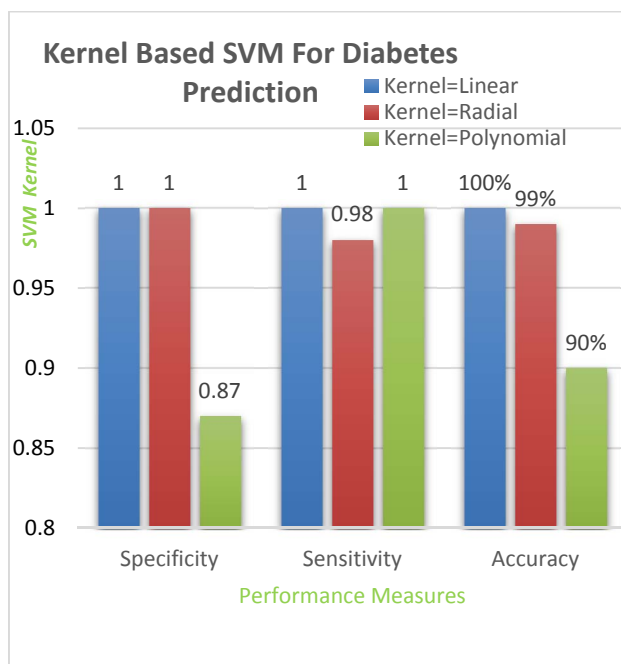
Fig 2: Performance Curve for all the Kernel Functions

IV. CONCLUSION

Support Vector Machine is the widely used techniques in prediction of diabetes patients. It evaluates the performance based on accuracy level for various kernel functions of SVM. A result shows that SVM with Linear Kernel outperforms the ther two kernel functions. The SVM with Linear Kernel function produced 100%, SVM with Radial Kernel Produced 99% and SVM with Polynomial kernel produced 90% for the chosen data set.

REFERENCES

1. Sajida Perveena,Muhammad Shahbaza,, Aziz Guergachib and Karim Keshavjeec ,Performance Analysis of Data Mining Classification Techniques to Predict Diabetes Procedia Computer Science.,82 (2016) 115 – 121.

2. Abdullah A. Aljumah , Mohammed Gulam Ahamad , Mohammad Khubeb Siddiqui Application of data mining: Diabetes health care in young and old patients , Journal of King Saud University - Computer and Information Sciences ,25(2013),pp.127-136

3. Xue-Hui Meng,, Yi-Xiang Huang,, Dong-Ping Raon, Qiu Zhang, , Qing Liu 2012, Comparison of three data mining models for predicting diabetes or prediabetes by risk factors , Kaohsiung Journal of Medical Sciences,29(2013), pp. 93-99

4. A.S. Almazyad, M.G. Ahamad, M.K. Siddiqui, A.S. Almazyad, Effective hypertensive treatment using data mining in Saudi Arabia Journal of Clinical Monitoring and Computing, 24 (6) (2010), pp. 391–401

5. Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras,Ioannis VlahavasIoanna Chouvarda ,Machine Learning and Data Mining Methods in Diabetes Research , 15(2017), pp.104–116

6. E.M. Cox, D. Elelman , Test for screening and diagnosis of type 2 diabetes Clin Diabetes, 4 (27) (2009), pp. 132–138

7. Kumar P.S., Umatejaswi V.Diagnosing Diabetes using Data Mining Techniques International Journal of Scientific and Research Publications, 7 (2017), pp. 705-709

8. Perveen S., Shahbaz M., Guergachi A., Keshavjee K.Performance Analysis of Data Mining Classification Techniques to Predict Diabetes Procedia Computer Science, 82 (2016), pp. 115-121

9. Vijayan V.V., Anjali C.Prediction and diagnosis of diabetes mellitus A machine learning approach 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (2015), pp. 122-127doi:10.1109/RAICS.2015.7488400.

10. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings-2010 International Conference on Computational

Intelligence and Communication Networks, CICN 2010, 554-559doi:10.1109/CICN.2010.109.

11. Nello Cristianini and John Shawe-Taylor, ―An Introduction to Support Vector Machines and Other Kernel-based Learning Methods‖, Cambridge University Press, 2000.

12. R.priya and P Aruna. Article: SVM and Neural Network based Diagnosis of Diabetic Retinopathy. International Journal of Computer Applications 41(1):6-12, March 2012