V. Anuja Kumari, R.Chitra / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 2, March -April 2013, pp.1797-1801

Classification Of Diabetes Disease Using Support Vector Machine

V. Anuja Kumari[1], R.Chitra[2]

[1]PG Student, Department of Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, Kanyakumari District, India
[2] Associate Professor, Department of Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, Kanyakumari District, India

Abstract

Diabetes mellitus is one of the most serious health challenges in both developing and developed countries. According to the International Diabetes Federation, there are 285 million diabetic people worldwide. This total is expected to rise to 380 million within 20 years. Due to its importance, a design of classifier for the detection of Diabetes disease with optimal cost and better performance is the need of the age. The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard for testing data mining algorithms to see their prediction accuracy in diabetes data classification. The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease.

I. INTRODUCTION

Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. Diabetes is a condition in which your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. There are two general reasons for diabetes: (1) the pancreas does not make enough insulin or the body does not produce enough insulin. Only 5-10 % of people with diabetes have this form of the disease (Type-1). (2) Cells do not respond to the insulin that is produced (Type-2). Insulin is the principle hormone that regulates uptake of glucose from the blood into most cells (muscle and fat cells). If the amount of insulin available is insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. Diabetes mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem [5]. A classifier is required and to be designed that is cost efficient, convenient and accurate. Artificial intelligence and Soft Computing Techniques provide a great deal of human ideologies and are involved in human related fields of application. These systems find a place in the medical diagnosis.

A medical diagnosis is a classification process. A physician has to analyze lot of factors before diagnosing the diabetes which makes physician's job difficult. In recent times, machine learning and data mining techniques have been considered to design automatic diagnosis system for diabetes [9]. Recently, there are many methods and algorithms used to mine biomedical datasets for hidden information including Neural networks (NNs), Decision Trees (DT), Fuzzy Logic Systems, Naive Bayes, SVM, cauterization, logistic regression and so on [1,3,7]. These algorithms decrease the time spent for processing symptoms and producing diagnoses, making them more precise at the same time. There is a great variety of methods related to diagnosis and classification of diabetes disease in the literature. Polat et al. [2] used principal component analysis and neuro fuzzy inference for diabetes data classification. Deng and kasabov [6] obtained 78.4% classification accuracy with 10-fold cross-validation (FC) using ESOM. Yu et al. [10] combined Quantum Particle Swarm Optimization (QPSO) and Weighted Least Square (WLS) Support Vector Machine to diagnose Type-2 diabetes. Smith et al. [12] proposed a neural network ADAP algorithm to build associative models. 576 randomly selected data are used for training and 192 test cases showed an accuracy of 76%. Quinlan [13] applied c4.5 algorithm and the classification accuracy was 71.1%. Sahan et al. [14] used Attribute Weighted Artificial Immune System with 10- fold cross validation method and obtained a classification accuracy of 75.87%. There have been many other methods used for the classification of diabetes dataset with accuracy between 59% and 77.5%. SVMs have shown remarkable success in the area of employing Computer Aided Diagnostic systems (CAD) to improve diagnostic decisions [4]. The Support Vector Machine (SVM) is a novel learning machine introduced first by Vapnik and has been applied in several financial applications recently, mainly in the area of time series prediction and classification.

This paper is organized as follows: Section II briefly review some basic concepts of SVM and the kernel function selection. Section III describes the classification process using SVM classifier. The experimental results are given in Section IV. Finally, Section V concludes the paper.

II. SUPPORT VECTOR MACHINE

A. SVM Model Generation

SVM is a set of related supervised learning method used in medical diagnosis for classification and regression [1,16]. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory i.e. the so called structural risk minimization principle. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [1, 8]. For example, given a set of points belonging to either one of the two classes, an SVM finds a hyperplane having the largest possible fraction of points of the same class on the same plane. This separating hyperplane is called the optimal separating hyperplane (OSH) that maximizes the distance between the two parallel hyper planes and can minimize the risk of misclassifying examples of the test dataset.

Given labeled training data as data points of the form

$$M = \{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)\}$$

where y_n =1/-1, a constant that denotes the class to which that point x_n belongs. n=number of data sample. Each x_n is a p-dimensional real vector. The SVM classifier first maps the input vectors into a decision value, and then performs the classification using an appropriate threshold value. To view the training data, we divide (or separate) the hyperplane, which can be described as:

Mapping: $w^T.x + b = 0$ (1)

where w is a p-dimensional weight vector and b is a scalar. The vector w points perpendicular to the separating hyperplane. The offset parameter b allows to increase the margin. When the training data are linearly separable, we select these hyperplanes so that there are no points between them and then try on maximizing the distance between the hyperplane. We have found out the distance between the hyperplane as 2/ |w|. To minimize |w|, we need to ensure that for all i either

$$w.x_i - b \geq 1 \; or \; w.x_i - b \leq -1 \quad (2)$$

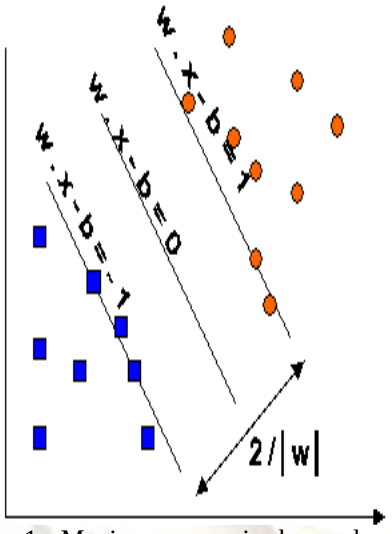

Fig. 1. Maximum margin hyperplanes for SVM trained with samples from two classes

B. Radial Basis kernel function

The Radial Basis Function (RBF) kernel of SVM is used as the Classifier, as RBF kernel function can analyse higher-dimensional data [1,17]. The output of the kernel is dependent on the Euclidean distance of x_j from x_i (one of these will be the support vector and the other will be the testing data point). The support vector will be the centre of the RBF and γ will determine the area of influence this support vector has over the data space.

RBF Kernel function can be defined as

$$k(x_i, x_j) = \exp\left(-\gamma \left||x_i - x_j|\right|^2\right), \quad \gamma > 0 \quad (3)$$

where γ is a kernel parameter and x_i is the training vector. A larger value of γ will give a smoother decision surface and more regular decision boundary. This is because an RBF with large γ will allow a support vector to have a strong influence over a larger area. The best parameter set is applied to the training dataset and the classifier is obtained. The designed classifier is used to classify the testing dataset to get the generalization accuracy.

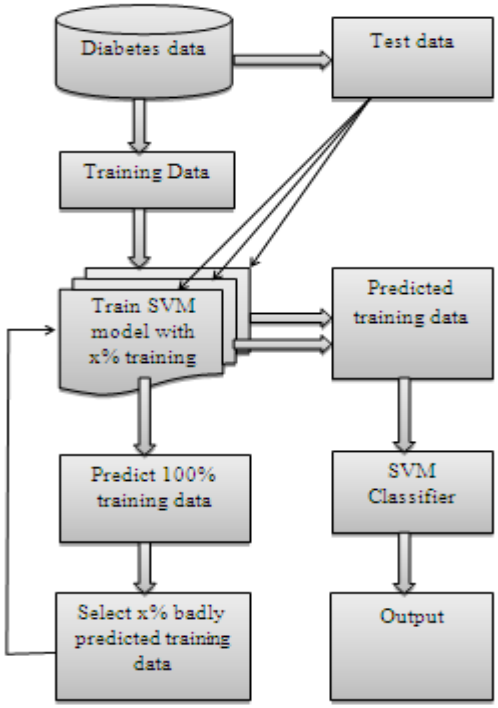

Fig.2. Architecture of the Proposed System

III. DISEASE CLASSIFICATION USING SVM

A. Experimental Setup

The SVM models for classification have been developed for the classification of diabetes dataset. The experiments are conducted on Matlab R2010a. The datasets are stored in MS Excel documents and read directly from Matlab. The diagnostic performance of the developed models is evaluated using Receiver Operating Characteristic (ROC) curve. In ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold.

B. Diabetes Disease Dataset

The Pima Indian diabetes dataset, donated by Vincent Sigillito, is a collection of medical diagnostic reports from 768 records of female patients at least 21 years old of Pima Indian heritage, a population living near Phoenix, Arizona, USA [19]. The binary target variable takes the values '0' or '1' while '1' means a positive test for diabetes, '0' means a negative test. There are 268 cases in class '1' and 500 cases in class '0'. The significance of the automatically selected set of variables was further manually evaluated by fine tuning parameters. The variables included in the final selection were those with the best discriminative performance.

There are eight numeric variables: (1) Number of times pregnant, (2) Plasma glucose concentration a 2h in an oral glucose tolerance test (3) Diastolic blood pressure (mm Hg) (4) Triceps skin fold thickness (mm) (5) 2-hour serum insulin (mu U/ml) (6) Body mass index (7) Diabetes pedigree function (8) Age (years). Although the dataset is labeled as there are no missing values, there were some liberally added zeros as missing values. Five patients had a glucose of 0, 28 had a diastolic blood pressure of 0, 11 more had a body mass index of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After the deletion there were 460 cases with no missing values.

C. Training and test dataset evaluation

To evaluate the robustness of the SVM models, a 10-fold cross-validation was performed in the training data set. The training data set is first partitioned into 10 equal-sized subsets. Each subset was used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the remaining nine datasets. This cross-validation process was repeated 10 times, and each subset serve once as the test data set. Test data sets assess the performance of the models.

IV RESULT ANALYSIS

To analyze the performance of classification, the accuracy and AUC measures are adopted. Four cases are considered as the result of classifier.

TP (True Positive) : the number of examples correctly classified to that class.

TN (True Negative): the number of examples correctly rejected from that class.

FP (False Positive): the number of examples incorrectly rejected from that class.

FN (False Negative): the number of examples incorrectly classified to that class.

The classification experiments are conducted on the Diabetes dataset. The SVM classifier with RBF kernel is used for classification. The diabetes dataset contains 460 data, 200 data are used for training and 260 data for testing. The results of SVM classification for Diabetes dataset are analysed. Table I shows the number of data used for training, testing, number of attributes used and the accuracy of the classifier for the diabetes data set.

TABLE I
CLASSIFICATION ACCURACY USING SVM

Data set	Samples	Training data	Testing data	Attributes	No. of Classes	Using SVM (with RBF kernel)
Diabetes	460	200	260	8	2	0.755

The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible values of the variables being classified. From the results obtained the following equations are used to measure the Accuracy, Sensitivity, and Specificity.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Receiver Operating Characteristic (ROC) curves were generated based on the predicted outcome. ROC curve is a technique for summarizing a classifier's performance over a range by considering the tradeoffs between TP rate and FP rate. The TP rate and FP rate are calculated as:

$$Sensitivity(TP\ rate) = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity\ (FP\ rate) = \frac{FP}{FP+TN} \quad (6)$$

Sensitivity and Specificity are statistical measures that describe how well the classifier discriminates between a case with positive and with negative class. Sensitivity is the detection of disease rate that needs to be maximized and Specificity is the false alarm rate that is to be minimized for accurate diagnosis. The results in Table II show the performance of SVM classification.

TABLE II
PERFORMANCE OF SVM CLASSIFIER

Dataset	Accuracy	Sensitivity	Specificity
Diabetes	78%	80%	76.5%

The training set accuracy of Diabetes data set is 65.8 and the testing accuracy is 78.2 for the SVM classifier. From the cross validation accuracy it is noticed that there is significant improvement in the accuracy if the number of training samples increases. Figure 3 shows the training set accuracy of the diabetes data set. Receiver Operating Characteristic curve is plotted between false positive rate and true positive rate and it describes the measure of positively predicting the disease. Receiver Operating Characteristic curve for diabetes data is shown in figure 4.

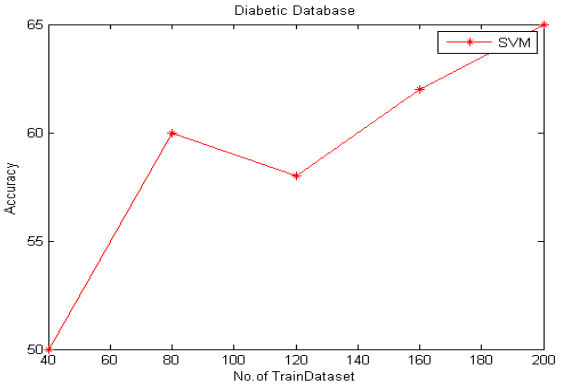

Fig. 3. Training set Accuracy of Diabetes data set

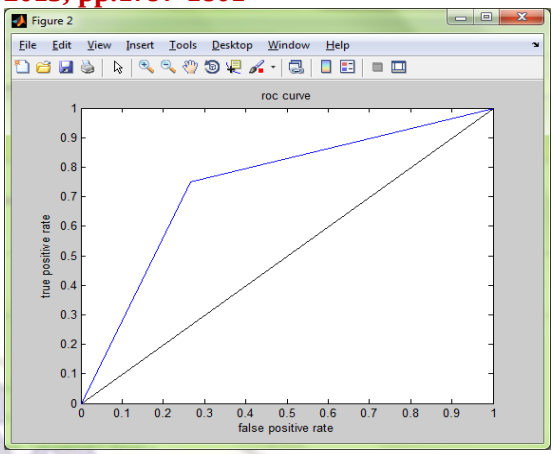

Fig. 4.ROC of SVM using Diabetes data set

V.CONCLUSION

In this paper, we have used datasets for diabetes disease from the machine learning laboratory at University of California, Irvine. All the patients' data are trained by using SVM. The choice of best value of parameters for particular kernel is critical for a given amount of data SVM approach can be successfully used to detect a common disease with simple clinical measurements, without laboratory tests. In the proposed work, SVM with Radial basis function kernel is used for classification. The performance parameters such as the classification accuracy, sensitivity, and specificity of the SVM and RBF have found to be high thus making it a good option for the classification process. In future the performance of SVM classifier can be improved by feature subset selection process.

REFERENCES

[1] Cortes, C., Vapnik, V., "Support-vector networks", *Machine Learning*, 20(2),pp. 273-297, 1995.

[2] Polat, Kemal and Salih Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Expert system with Applications*, pp. 702-710, Elseivier, 2007.

[3] Herron P., "Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms", *INLS 110, Data Mining*, 2004.

[4] N.Lavrac, E. Keravnou, and B. Zupan, "Intelligent Data Analysis in Medicine," in *Encyclopedia of Computer Science and Technology*, vol.42, New York:Dekker, 2000.

[5] Barakat,et al. " Intelligible Support Vector Machines for diagnosis of Diabetes Mellitus." *IEEE Transactions on*

Information Technology in Biomedicine, 2009.

[6] D. Deng and N. Kasabov, " On-line pattern analysis by evolving self- organizing maps", *In Proceedings of the fifth biannual conference on artificial neural networks and expert systems (ANNES)*, 2001, pp. 46-51.

[7] Balakrishnan Sarojini, Narayanasamy Ramaraj and Savarimuthu Nickolas, "Enhancing the Performance of LibSVM Classifier by kernel F-Score Feature Selection", *Contemporary Computing*, 2009, Volume 40, Part 10, pp. 533-543.

[8] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery", *Springer, 2(2),* pp.121-167, 1998.

[9] T.Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[10] Yue, et al. " An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLSSVM," *International Symposium on Intelligent Information Technology Application Workshops*, IEEE Computer Society, 2008.

[11] Van Gerven M. A.J., Jurgelenaite R., Taal B. G., Heskes., Lucas P. J.F., "Predicting carcinoid heart disease with the noisy-threshold classifier". *Artificial Intelligence in Medicine*, 2007, vol.40, 45-55.

[12] Smith, J.W., J. E. Everhart, et al.- "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", *Proceedings of the Symposium on Computer Applications and Medical Care (Washington, DC). R.A. Greenes*. Los Angeles, CA, IEEE Computer Society Press, 1988, pp. 261-265.

[13] Quinlan, J.R. "C4.5: programs for machine learning", *San Mateo, Calif., Morgan Kaufmann Publishers*, 1993.

[14] S.Sahan, K.Polat, H. Kodaz, and S. Gunes, " The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases", in *ICARIS*, 2005,p. 456-468.

[15] I.Tsoulos, D. Gavrilis, E. Glavas,- "Neural network construction and training using grammatical evolution", *Science Direct Neurocomputing Journal*, Vol.72, Issues 1-3, December 2008,pp. 269-277.

[16] V. Vapnik, "The Nature of Statistical Learning Theory." NY: *Springer- Verlag*. 1995.

[17] Park, J. and Sandberg, I. W., "Universal approximation using radial basis function networks", *Neural Computation*, vol.3, pp.246-257, 1991.

[18] P. Venkatesan and S. Anitha, "Application of a radial basis function neural network for diagnosis of diabetes mellitus" , *Current Science*, vol. 91, no.9,pp.1195-1199, 10 November 2006.

[19] UCI repository of bioinformatics Databases, Website: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[20] Hung, et al. "Feature selection and classification model construction on type 2 diabetic patients' data", *Journal of Artificial Intelligence in Medicine*, pp 251-262, Elsevier, 2008.