PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS FOR DIABETES CLASSIFICATION

Midhila. M* and S. Padmavathi**

* **Dept of Computer Science and Engineering, Amrita School of Engineering,Coimbatore
Amrita vishwa vidhyapeetham, Amrita University, India
cb.en.p2cvi15005@cb.students.amrita.edu
s_padmavathi@cb.amrita.edu

ABSTRACT: Diabetes during pregnancy is a major issue common among Indian women. Prediction of diabetes based on the test done during the pregnancy period plays a significant role in the treatment. Results of various clinical test conducted during the pregnancy can be considered as parameters for diabetes classification. In this paper, Classifiers such as Support Vector Machine, probablistic, Tree based and regression are used to predict diabetes based on the parameters. The SVM classifiers performed well on the dataset giving the highest accuracy of 78% while regression based classsifier scored a minimum of 65% accuracy.

KEYWORDS: Diabetes, SVM classifier, Regression, Probablistic classifier.

INTRODUCTION

Diabetes is a common disease found in India.It is characteized by high sugar level in blood. The blood sugar level is tested before and after food. If any one exceeds a predefined threshold,the person is said to be diabetic. Diabetes is related to insulin secretion of the body. There are two types of diabetes,Type-1 diabetes is caused when enough insulin is not secreted. This cannot be prevented and is usually treated with insulin injection. Type-2 diabetes is caused due to the insulin resistance occuring mainly due to obesity. This can be prevented by proper exercise. Many women get Type-1 diabetes during pregnancy. Among Indian women who are diabetic, a majority are affected during pregnancy. This paper considers the test parameters extracted from PIMA Indian women dataset and tries to predict if they are diabetic using various classifiers. The performance of Suppot Vector Machine classifier, probablistic,Treebased,regression based is analysed and results are summarized in this paper. The data set is referred from the UCI data repository.
Eight features such as,Plasma glucose concentration a 2 hours in an oral glucose tolerance test,Number of times pregnant,Diastolic blood pressure (mm Hg) Triceps skin fold thickness(mm),Diabetespedigree function 2-Hour serum insulin(mu U/ml),Body mass Age (years) and Body mass index (weight in kg/(height in m)^2) are considered as features for diabetes classification.Naive bayes classifier is a traditonal probablistic classifier which generally perfoms well for larger data set. Decision tree claasifier suites for a heirarchical selection of features. ZeroR classifier is a regression based classifier and binary SVM classifier are considered for analysis in this paper.

LITERATURE SURVEY

Najmeh [2] proposed that diabetes can be analysed by applying different classifiers such as Bayesian, Functional, Rule-base,and Decision Trees. Experimental results on Pima Indian Diabetes (PID) and concluded as Logistic core has better performance in comparison with other classifiers. David et.al [21] deals with a comparison of a popular SVM implementation (libsvm) to 16 classification methods and 9 regression methods on “abalone” and the “cpu-Small”. SVMs showed mostly good performances both on classification and regression tasks. Nahla et al. [10] proposes a system which uses SVM for the diagnosing purposes. It uses an additional explanation module, where the “black box” model of an SVM turned into an intelligible representation for diagnostic decision. The data from 4682 subjects of age 20 years and above are collected using a questionnaire. From the paper it signifies that SVM is a useful tool for predicting

diabetes. Robert et al. [21] deals with SVM to know whether it is qualitatively robust for any fixed regularization parameter λ. The results show that SVMs are the solutions of a well-posed mathematical problem in Hadamard's sense.

3 OVERVIEW OF CLASSIFIERS

The classifiers considered in this paper are SVM, Bayesian network, Decision trees, Naive Bayes and ZeroR classifiers. A brief overview of each classifier is presented below.

SVM Classifier

Linearly separable SVM produces a hyperplane that linearly divides the positive and negative classes during the training. The samples are considered to be in high dimensional feature space or in input space. This determines the principal hyperplane which separates the two samples The linear separator is designed in such a way that it is located at the maximum distance from the hyperplane to the nearby samples. The decision boundary for a linear classifier can be represented as follows (1):

$$wTx + b = 0 \qquad (1)$$

From the equation, w and b represents the weight vector and the bias factor respectively. SVM obtains a decision boundary, which is away from the sample points. The points used for identifying the position of the separators are known as support vectors.For a linearly separable data there will be a pair of (*w,b*) which satisfies equation (2)and(3)

$$wTx_i + b \geq 1, \text{if } y_i = \quad 1 \qquad (2)$$

$$wTx_i + b \leq -1, \text{if } y_i = -1 \qquad (3)$$

The classifier is defined in equation (4):

$$f(x) = \text{sign}(wTx + b) \qquad (4)$$

The functional margin can be scaled according to the ease of solving large SVM. The value of margin can be decided on the basis of input vector and the value is equivalent to be one. All the points present in the data sample can be represented using the following equation (5):

$$y_i\left(\vec{w}^T\vec{x}_i + b\right) \geq 1 \qquad (5)$$

Bayesian Network

A graphical model that encodes the probabilistic relations among the variables is known as a Bayesian network. They are statistical classifiers. Bayes theorem provides a method of calculating the probability of a hypothesis based on its prior probability.

Decision tree

This is a hierarchical tree based classifier and indicates a course of action to be taken for each value or combination of values of one or more variables or parameters. The decision attributes allows us to partition, the entire universe into blocks determined by possible decisions. The block used here is known as classes.

Naïve bayes: It is a commonly used probabilistic classifier that is based on Bayes'rule or Bayes'theorem. Naïve bayes classifier can predict the probabilities of a given sample belonging to a particular class. Here the attribute value on a given class is assumed to be independent to the values of other attributes. Naïve bayes is calculated from independent probability and is also called as class conditional independence. The probability value can be calculated from the equation (6). This is more accurate for natural datasets where the classes are clearly defined.

$$\mathbf{P(C = c_k|X = x) = P(C = c_k) \times \frac{P(X=x|C=c_k)}{P(x)}} \qquad (6)$$

Here all possible events fall into exactly one class. C is the class with values ranging from (C1, C2…..Ck).

ZeroR classifier: The simplest regression based classification technique which relies on the target and ignores all other predictors. It predicts the majority category as a class and can be used as a benchmark for other classification methods.

EXPERIMENTAL RESULTS

Diabetes patient reports which are collected from various sources available in the UCI repository. One of the standard machine learning dataset from the repository is the PID dataset, which holds 768 samples. The dataset contains the details of PIMA Indian women at least 21years old and living near Phoenix, USA. Each sample has 8 features obtained from clinical test as mentioned in the introduction. The classifier considers these features and classifies into diabetic class (1) or non diabetic class (0). From the dataset 268 samples belongs to diabetic and 500 samples belongs to non-diabetic.

The performance of a classifier can be analysed by TP rate,FP rate,Precisions,Recall,Classification rate and F-measure as specified in equations from (7) to (12). These are calculated based on the values from confusion matrix as given in Table 1.

$$\text{TP rate} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100 \quad (7)$$

$$\text{FP rate} = \frac{\text{FP}}{(\text{FP}+\text{TN})} \times 100 \quad (8)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{TN})} \times 100 \quad (9)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100 \quad (10)$$

$$\text{F} - \text{measure} = 2 \times \text{Precision} \times \frac{\text{Recall}}{\text{Precision}+\text{Recall}} \quad (11)$$

$$\text{Classification} = \frac{\text{TN}+\text{TP}}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \times 100 \quad (12)$$

Table 1: Confusion Matrix

Predicted class	Actual class	
	Diabetic	Non-diabetic
Diabetic	TP	FP
Non-diabetic	FN	TN

True Positive (TP): Diabetic People correctly detected as diabetic.
False Positive (FP): Healthy people wrongly detected as diabetic.
True Negative (TN)**:** HealthyPeople properly detected as healthy
False Negative (FN): Diabetic people wronglydetected as healthy.
The significant fractions of retrieved instances are known as precision. The fraction of appropriate instances that are retrieved is the recall and represented by sensitivity.
This paper uses WEKA which is a most widely used machine learning software in Java. The results of classifiers such as SVM, BayesNet,Decision Trees,Naive bayes and ZeroR on PIMA dataset is given in table (2) and (3) . Table 2 gives the results of classifiers on non diabetic dataset and Table 3 gives the results on diabetic dataset. A 10-Fold cross validations are performed and it divides the dataset into 10 partitions.The classifier runs for 10 times and uses different partitions as test set.

Table 2: Classification of Tested negative class

Classifier	TP	FP	PRECISION	RECALL	F-MEASURE	TN	FP
SVM	89	45.9	78.5	89.8	.838	449	51
BayesNet	816	39.2	79.5	81.6	.806	408	92
DecisonTree	81.6	39.2	79.5	81.6	.806	405	95
NaiveBayes	84.4	38.8	80.2	84.4	.823	422	78
ZeroR	1	1	65.1	1	.78	500	0

Table 3: Classification of Tested positive class

Classifier	TP	FP	PRECISION	RECALL	F-MEASURE	TP	FN
SVM	54.1	10.2	74	54.1	.625	123	145
BayesNet	60.8	18.4	63.9	60.8	.623	105	163
DecisonTree	60.8	18.4	63.9	60.8	.623	126	142
NaiveBayes	61.2	15.6	67.8	61.2	.643	104	164
ZeroR	0	0	0	0	0	268	0

The overall classification accuracy of SVM, Bayes net, Descision tree, naïvebayes ,zero-R classsifiers are 78%, 74.3%, 71%, 76% and 65% respectively are plotted in fig(1).

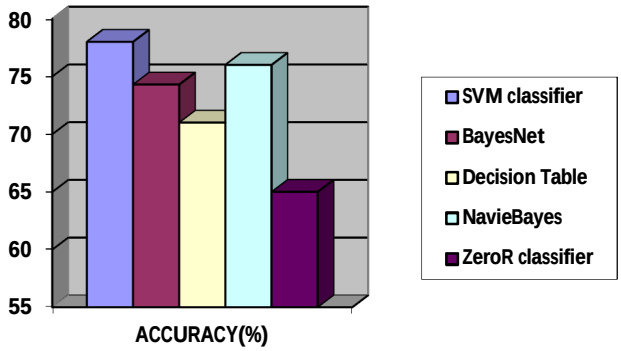

Figure 1. Accuracy of the classifier

CONCLUSION

The cause for Type-1 Diabetes is unknown. From the survey it couId be concluded that Indian women accquire Type-1 diabetes majorly during their pregnancy. The detection at early stage is a major health concern. This paper uses various machine learning algorithms to predict the diabetes occurring in pregnant women based on their clinical test results. The data set accquired from UCI repository.

From the results of various classifiers, SVM performs well with an overall classification rate of 78%. For non diabetic data set SVM has a highest Sensitivity rate of 89.8% and naïve bayes has a highest precision of 78.5 %.For diabetic dataset the sensitivity rate and precision rate is higher for SVM with 74% and 61.2% respectively. ZeroR classifier gave the minimum value for sensitiviity and precision in both test cases. Naive bayes gives a high precision value because of large number of samples in that category.

REFERENCES

[1] V. Vapnik, “The natural of statistical learning theory,” Springer,New York, 1995.
[2] NajmehHosseinpour,”Diabetes Diagnosis by Using Computational Intelligence Algorithms”, Volume 2, Issue 12, December 2012.
[3] S.Abe,”Support vector machines for pattern classification,” Springer,London-verlag,2006.
[4] A. Suarez and J.F. Lutsko, “Globally optimal fuzzy decision trees for classification and regression,” IEEE Transactions on Pattern Analysis and Machine Intelligence,, vol. 21, no. 12, 1999, pp. 1297-1311.
[5] UCI Machine learning Repository,"Pima Indians Diabetes Data Set”,http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.
[6] Raj Anand, Vishnu Pratap Singh Kirar and 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”, 2231-2307, 2 ,6, 2013.
[7] R.R. Bouckaert, et al., “WEKA---Experiences with a Java Open-Source Project,” The Journal of Machine Learning Research, vol. 11, 2010, pp. 2533-2541.

[8] K. D. Adeena and R. Remya, “Extraction of relevant dataset for support vector machine training: A comparison”, in 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015, 2015, pp. 222-227.
[9] Abinav, Anil kumar, Naveena Karthika, Pratibha, Ronsen, R. Gandhiraj, and Soman K.P., “SVM based Classification of Digitally Modulated Signals for Software Defined Radio”, in International Conference on Embedded Systems 2010, Coimbatore Institute of Technology, Coimbatore, 2010.
[10] Nahla Barakat , Department of Applied Information Technology, German University of Technology in Oman, Muscat, Oman , Andrew P. Bradley , Mohamed Nabil H. Barakat,”Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus”,IEEE Transaction Volume:14 Issue:4.
[11] Y.Ng Andrew and Michael I.Jordan,”OnDisriminativevs.Generativeclassifiers:A comparison of logistic regression and Navie Bayes”.
[12] Liu Jie. Support Vector Machine-based Network Intrusion Detection System, *Central South University Thesis*, 2008.
[13] Zhang Shuya, Zhao Yiming, Image classification algorithm and implementation basing on SVM. Computer engineering and application, Vol. 43, No. 25, pp. 125-127, 2007 .
[14] C. W. Hsu, and C. J. Lin, “A comparison of methods for multiclasssupport vector machines,” IEEE Transactions on NeuralNetworks, Vol. 13, 2002, pp. 415–425.
[15] Gloria L.A. Beckles and Patricia E. Thompson-Reidy the authors of “ Diabetes and Women’s Health Across the Life Stages”.
[16] K. Soman and V.A. Loganathan. R, “Machine Learning with SVM and other Kernel Methods”, Prentice Hall of India, 2009.
[17] Herron P., “Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms”, INLS 110, Data Mining, 2004.
[18] 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.
[19] David Meyera ,FriedrichLeischa , Kurt Hornikb, “The support vector machine under test”,Austria 2008.
[20] Robert Hable , Andreas Christmann, Department of Mathematics, “On qualitative robustness of support vector machines”, University of Bayreuth, Germany.