Diabetes prediction using feature selection and classification

Khyati K. Gandhi[1], Prof. Nilesh B.Prajapati[2]

[1]*PG Student, CE Department, BVM Engg. College, Vallabh Vidhyanagar, kvmehta108@gmail.com*
[2]*IT Department, BVM Engg. College, Vallabh Vidhyanagar, nilesh.prajapati@bvmengineering.ac.in*

Abstract —Medical data mining is becoming increasingly important in healthcare. The diversity of medical data collected/stored for diagnosis and prognosis and the availability of widespread data mining techniques to process these data place medical data mining in a unique position to truly impact patient care using these stored data. Medical data are high dimensional in nature. It contains irrelevant and redundant features that reduce prediction accuracy so data pre-processing is required to prepare data for mining task. Feature selection has been an active and fruitful field of research and development for decades in statistical machine learning, data mining. It is effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results. Feature selection is the pre-processing technique that selects optimal feature subset from whole features. F-score method and K-means clustering is used for feature selection. The performance of the SVM classifier is empirically evaluated on the reduced feature subset of Pima Indian diabetes dataset is one of the standard dataset available at UCI machine learning laboratory used for testing data mining algorithms to see their prediction accuracy in diabetes data classification.

Keywords—Data mining, Feature selection, F-score, SVM classifier, K-means clustering.

I. INTRODUCTION

Diabetes is often called a modern-society disease because widespread lack of regular exercise and rising obesity rates are some of the main contributing factors for it. Diabetes is a very serious disease that, if not treated properly and on time, can lead to very serious complications, including death [11]. 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. Data classification problem is studied by statisticians and machine learning researchers. Data classification is widely used in variety of engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence. The goal of the data classification is to classify objects into a number of categories or classes. For a given dataset the task of classification is to assign a class to the data object.

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 [5].

1.1 Pima Indian Diabetes Dataset

The Pima Indian Diabetes data set [8] was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases [1, 2, 6, 7].

There are eight clinical findings (features):

1. Number of times pregnant
2. Plasma glucose concentration a 2 h in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. Two hour serum insulin (mu U/ ml)
6. Body mass index
7. Diabetes pedigree function
8. Age (years).
9. Class variable 0 or 1

Class variable takes the values 0 or 1, where 1 means tested positive for diabetes and 0 means tested negative for diabetes. The Pima Indian diabetes dataset is widely used for testing classification algorithm.

There are total 768 samples are there in Pima Indian diabetes dataset, 268 samples are diabetes positive and 500 samples are diabetes negative.

II. PROPOSED SYSTEM

The proposed system is based on Feature selection and SVM classifier. It is depicted in figure 1.

Feature selection is one of the important and frequently used data pre-processing techniques for data mining applications in medicine. Data pre-processing is required to prepare the data for data mining tasks to increase the predictive accuracy. The first motivation is quite evident, since fewer features require less run time to train and to apply the classifier [9].

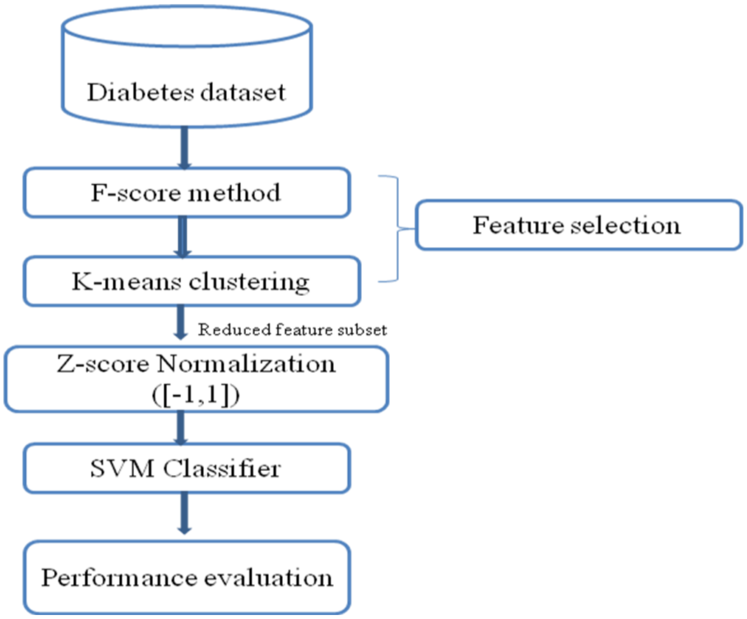

Figure 1 Task flow of proposed system

2.1 F-score method [1]

F-score is a one of the filtering method used for feature selection. The F-score values of each feature in dataset are computed and the features with relatively high F-scores are considered as “informative”.

The F-Score values are estimated using the following equation.

$$F(i)= \frac{(X_{pavg}(i)-X_{avg}(i))^2 +(Xnavg(i)- X_{avg}(i))^2}{(1/Npos) \sum_{k=1}^{Npos} (Xpos(k,i)- X_{pavg}(i))^2 + (1/Nneg) \sum_{k=1}^{Nneg} (Xneg(k,i)- X_{navg}(i))^2}$$

$X_{avg}(i)$, $X_{pavg}(i)$, Xnavg(i) are the average of the ith feature of the whole, positive and Negative datasets respectively. Xpos(k,i) is ith feature of the kth positive instance. Xneg(k,i) is ith feature of the kth negative instance. Npos, Nneg are number of positive and negative instances respectively.

Table 2.1 F-score values of features of Pima Indian diabetes dataset

Features	**F-score Value**
f2:Glucose tolerance test	0.31334
f6:Body mass index	0.114161
f8:Age	0.073494
f1: No. of times pregnant	0.058048
f7: Diabetes pedigree function	0.034887
f3: Diastolic blood pressure	0.004827

The estimated F-score values of the features are arranged in decreasing order of importance. The least rank features are eliminated one at a time from backwards and the performance of the K-means clustering algorithm is observed.

2.2 K-means clustering

K-means clustering is one of the popular partitioning algorithms which use Euclidean distance as the dissimilarity method. The features with low F-scores are removed one at a time and the clustering error is used as the performance indicator to determine the optimal feature subset. The reduced feature subset that gives the minimal clustering error is considered to be the optimal feature subset.

Optimal subset selected after feature selection: f2, f6, f8.

2.3 Z-score normalization

Z-score Normalization is one of the data pre-processing steps of data mining. It scales data in [-1, 1] range.

2.4 SVM classifier

SVM is a set of related supervised learning method used in medical diagnosis for classification and regression.SVM is represented with the help of hyper plane. For example, given a set of points belonging to either one of the two classes, an SVM finds a hyper plane having the largest possible fraction of points of the same class on the same plane. This separating hyper plane is called the optimal separating hyper plane (OSH) that maximizes the distance between the two parallel hyper planes and can minimize the risk of misclassifying

Examples of the test dataset [5]. It is shown in following figure 2.

Given some training data D, a set of n points of the form:

$$D= \{(X_1,Y_1), (X_2,Y_2)....(X_n,Y_n)\}$$

Here Y_i= -1/1 that denotes the class to which data point X_n belongs. Any hyper plane can be written as the set of points satisfying

$$W^T.X + b=0$$

Here W is normal to hyper plane, b is a constant.

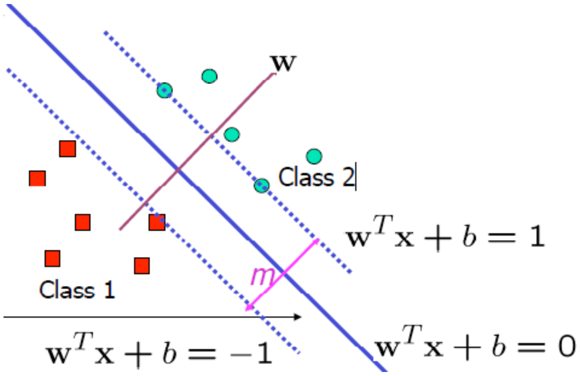

Figure 2 SVM training with two classes

We add the following constraint:

W^T. X_i + b<=-1 for x_i of first class and W^T. X_i + b>=1 for x_i of second class

This can be written as, we try to minimize ||w|| Subject to Y_i (W^T. xi + b)>=1, for all i.

- **RBF kernel**

Proper parameters setting can improve the SVM classification accuracy [10]. RBF kernel function is used as classifier. It is able to analyse high dimensional data. The output of the kernel is dependent on the Euclidean distance. RBF Kernel function can be defined as:

$$K\ (X_i,\ X_j) = EXP\ (-y||X_i-X_j\ ||^2\ )$$

Here, y is Kernel parameter and $X_{i,}$ X_j are Support vector and testing data point.

- **Soft margin**

If there exists no hyper plane that can split the "yes" and "no" examples, the *Soft Margin* method will choose a hyper plane that splits the examples as cleanly as possible. The method introduces non-negative slack variables, ξ_i, which measure the degree of misclassification of the data x_i

$$y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1 - \xi_i \quad 1 \leq i \leq n.$$

The objective function is then increased by a function which penalizes non-zero ξ, and the optimization becomes a trade off between a large margin and a small error penalty. If the penalty function is linear, the optimization problem becomes:

$$\text{Minimize } ½\ ||w||^2\ +C \sum_{i=1}^{n} \xi_i$$

- **Training and testing a model**

We applied type of cross validation to the training. First, we divide randomly whole training examples into equal-sized subsets. One subset is used for testing and rest subsets are used for training SVM. The trained SVM classifier is then tested using the one subset, and its classification error is recorded. This is repeated m times; m is equal to number of partitions. Third, the classification errors are averaged to obtain an estimate of the generalization error of the SVM classifier. In the end, the model with the smallest generalization error will be adopted.

2.5 Performance evaluation parameters

The proposed model is validated using four parameters namely the Accuracy of the classifier, Area Under ROC Curve, Sensitivity and Specificity [1, 5].

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.

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total no of instances}}$$

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True Positive + False Negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative+ False Positive}}$$

ROC describes the tradeoff between Sensitivity and Specificity, as well as the performance of the classifier, can be visualized and studied using the Receiver Operating Characteristic (ROC) curve.

Table 2.5.1 Classification result of SVM classifier

Dataset	**Accuracy %**	**Sensitivity %**	**Specificity %**	**AUC**
Pima Indian diabetes dataset	98	97.77	97.79	0.9

III. CONCLUSION

The most significant aspect of this study is to apply data mining technology for predicting diabetes. We performed a pre-processing step to deal with dataset like feature selection method, normalization and assessed machine-learning technique such as SVM. Feature selection reduces the no of dimensions by selecting most informative features based on some statistical score. F-score gives better performance of classification than other feature selection methods like relief and reliefF filtering methods.Then performance of SVM classifier is evaluated in term of accuracy, sensitivity, specificity and AUC. Embedding feature selection and data normalization performance of SVM classifier is improved. We applied different functions of SVM on various datasets having same no. of features but with different values then we get variation in accuracy in diabetes classification.

IV. REFERENCES

[1] Dr. B. Sarojini,Dr. N. Ramaraj , Enhancing Medical Prediction using Feature Selection (IJAE), Volume (1): Issue (3), 2011.

[2] Sarojini Balakrishnan, Ramaraj Narayanaswamy, Feature Selection Using FCBF in Type II Diabetes Databases International Conference on IT , March 2009, Thailand.

[3] Baek Hwan Cho, Hwanjo Yu, Kwang-Won Kim,Tae Hyun Kim, In Young Kim, Sun I. Kim ,predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine (2008) 42, 37—53.

[4] K.C. Tan, E.J. Teoh, Q. Yua, K.C. Goh, A hybrid evolutionary algorithm for attribute selection in data mining,2008 Published by Elsevier Ltd..

[5]V. Anuja Kumari, R.Chitra, Classification of Diabetes Disease Using Support Vector Machine. Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[6]Sunita Beniwal, Jitender Arora,Classification and Feature Selection Techniques in Data Mining, Vol. 1 Issue 6, August – 2012.

[7] Mrs. Madhavi Pradha, Ketki Kohale, Parag Naikade, Ajinkya Pachore, Eknath Palwe, Design of Classifier for Detection of Diabetes using Neural Network and Fuzzy k-Nearest Neighbor Algorithm, IJCER, Vol. 2 Issue. 5, September-2012.

[8] UCI repository of machine learning Databases, Pima Indian Diabetes Dataset. [Online] Available at: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[9] Yi Liu, Yuan F. Zheng, FS_SFS:Anovel feature selection method for support vector machines, Published by Elsevier Ltd. ,October 2005.

[10] Comparison of Feature Selection Approaches based on the SVM Classification, F.C. Li, F.L. Chen, G.E. Wang, IEEE 2008

[11] Manaswini Pradhan, Dr. Ranjit Kumar Sahu, Predict the onset of diabetes disease using Artificial Neural Network, International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004) 303 Volume 2, Issue 2, April 2011.