Study of Diabetes Prediction using Feature Selection and Classification

[1.] Khyati K. Gandhi, [2.] Prof. Nilesh B. Prajapati
[1.] PG Student,Computer department,Birla Vishvakarma Mahavidhylaya, Vallabh Vidyanagar,India
[2.] IT Department, Birla Vishvakarma Mahavidhylaya, Vallabh Vidyanagar,India

***Abstract*— Diabetes mellitus is one of the most serious health challenges in both developing and developed countries. It has become leading cause of death. Detection of diabetes with optimal cost and better performance is the need of the age. Medical data are multidimensional; hence data pre-processing step is applied to high dimensional data. Feature selection is a pre-processing step that is applied to high dimensional dataset to reduce number of dimensions by selecting the most informative features that influence the diagnosis of the disease. The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard dataset for testing data mining algorithms to see their prediction accuracy in diabetes data classification. The F-score feature selection method and k-means clustering select the optimal feature subsets of the medical datasets that enhances the performance of the Support Vector Machine classifier. The performance of the SVM classifier is empirically evaluated on the reduced feature subset of Diabetes dataset. Then performance is validated using four parameters namely the Accuracy of the classifier, Area Under ROC (Receiver operating characteristics) Curve, Sensitivity and Specificity.**

Keywords*— Feature selection, data mining, F-score, SVM classifier, K-means clustering, area under ROC curve.

I. INTRODUCTION

Diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. So there is need for diabetes detection increasing today.

Medical data mining is process of finding useful pattern that helps in medical diagnosis. Data mining is the process of extraction of knowledge. So, the predictability of disease will become more effective and early detection of disease will help in patient care. Data classification is one of the tasks in data mining. We try to design a classifier that detects diabetes with optimal cost and better performance. Data pre-processing is required to prepare the data for data mining and machine learning to increase the predictive accuracy. Data pre-processing procedures could reduce the amount of data to be analyzed without losing any critical information, improve the quality of the data, enhance the performance of the actual data mining algorithms and reduce the execution time of mining algorithms [1]. Finally Based on reduced dataset classifier detects diabetes disease.

A. Pima Indian Diabetes Dataset

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

The binary response variable (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 instances are there in data set. There are 268 instances are diabetes positive and 500 instances are diabetes negative. We have to populate the database by the above data set for diabetes predication.

B. Various methods of Diabetes prediction

- Feature selection methods

Feature selection is a data pre-processing step applied to diabetes dataset. It selects subset of features from whole feature set based on some statistical score and removes redundant features that do not contribute to performance.

There are three main approaches in feature selection: filter, wrapper, and embedded methods [20]. Filter methods select high ranked features based on a statistical score as a pre-processing step. Wrapper and embedded approach require considering the design of a classifier to select subset of features. Various filtering methods like F-score [1], reliefF [3] are available for feature selection.

We apply data mining techniques to extract knowledge from medical data. Generally medical data are high dimensional. If diabetes dataset contains irrelevant and redundant features then classification result is less accurate. To improve classification accuracy we apply feature selection to dataset. Here we apply Fast correlation based filter that try

to select small no of features that is also optimal feature subset.

F-score is one of the filtering methods. The F-score method uses the F-score values to measure the discriminating power of individual features in the database in respect to class labels [1]. F-score method calculates F-score of all the features of dataset. Based on the estimated values all features are arranged in descending order of importance. The features with high rank are known as most informative features. In feature selection we select most informative features. Then one feature is removed at a time and K-means clustering is applied. Clustering error is used as performance indicator to select optimal feature subset from different subsets. Plasma glucose concentration, body mass index and age form optimal feature subset as a result of feature selection. So feature selection methods reduce no of dimensions of dataset.

Classification methods make use of attributes in the process of classification. Wrapper methods uses classification algorithm to select the optimal attributes. New hybrid approach comprising of two conventional machine learning algorithms has been proposed to carry out attribute selection [4]. GA is evolutionary algorithm searches for the most promising attributes. Then SVM determine fitness value of each attribute. The fitness value of attributes is used to select optimal feature subset.

In short filtering method removes undesirable features before classification begins while the wrapper method apply classification algorithm to select optimal features. Wrapper method gives higher classification accuracy. The only drawback of the wrapper approach would be a longer runtime because the ML algorithm has to run iteratively in the search for the attribute subsets [4].

Recursive feature elimination with support vector machine is based on the embedded method. It is iterative process. It removed redundant features during iteration. However, this algorithm is limited theoretically to the linear kernel in SVM, because it is difficult to calculate the weight vector for a non-linear kernel because of the kernel characteristics of the implicit mapping [3].

A medical diagnosis is a classification process. Classification methods aim to predict diabetes. It tries to assign a class to previously unseen record as accurately as possible.

In recent times, machine learning and data mining techniques have been considered to design automatic diagnosis system for diabetes. 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].

Support Vector machine is one of the supervised learning used in classification problem. 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). It maximizes the distance between the two parallel hyper planes. Minimize the risk of misclassification. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [5].

When the training samples are linearly separable, SVM yields the optimal hyper plane that separates two classes with no training error, for linearly non separable cases, there is no such a hyper plane that is able to classify every training point correctly. However the optimization idea can be generalized by introducing the concept of soft margin [7]. RBF kernel is used for non linear classification. Use cross-validation to find the best parameter C and y. Use the best C and y to train the whole training set.

Other forms of algorithms like neural networks create very complex models to solve problems but are difficult to analyse theoretically [4]. There has been wide spectrum of work on developing ANN based classification models consisting of many hidden layers and large number of neurons in the hidden layers [15]. More no of hidden layers and more neurons give good result but increase cost of computation.

- Classification methods

C. Comparison of various diabetes prediction methods

Table 1.3.1 Comparison of various methods and performance measure of diabetes prediction

Authors	Feature selection	Classification method	Accuracy	Sensitivity	Specificity	AUC
B. Sarojini, N. Ramaraj	f-SCORE	SVM	98.9247	99.33	98.73	0.997
B. Sarojini, N. Ramaraj	FCBF	SVM	77.474	-	-	0.8344
Baek Hwan Cho et al.	ReliefF, SVM-RFE, sensitivity analysis with SVM	SVM	88.4	50.20	99.6	0.784
K.C. Tan et al.	GA	SVM	78.26	-	-	-
Manaswini Pradhan et al.	GA	ANN	73.83	-	-	-

II. CONCLUSION

We give emphasis on feature selection in diabetes prediction. Feature selection reduces the no of dimensions by selecting most informative features based on some statistical score. Then performance of classification is evaluated on the reduced diabetes dataset. We can say that embedding feature selection method in diabetes prediction improved performance of classification. F-score gives better performance of classification than other feature selection methods.

ACKNOWLEDGEMENT

We would like to thank reference authors and also like to thank the anonymous reviewers, whose comments and suggestions have helped them to improve the quality of the original manuscript.

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] K. Rajesh, V. Sangeetha, Application of Data Mining Methods and Techniques for Diabetes Diagnosis, IJEIT Volume 2, Issue 3, September 2012.
[7] Yi Liu, Yuan F. Zheng, FS_SFS: A novel feature selection method for support vector machines, Published by Elsevier Ltd., October 2005.
[8] Sunita Beniwal, Jitender Arora, Classification and Feature Selection Techniques in Data Mining, Vol. 1 Issue 6, August – 2012.
[9] M. A. Pradhan, Abdul Rahman, PushkarAcharya, RavindraGawade, AshishPateria, Design of Classifier for Detection of Diabetes using Genetic Programming, International Conference on Computer Science and Information Technology Pattaya Dec. 2011.
[10] 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.
[11] F.C. Li, F.L. Chen, G.E. Wang, Comparison of Feature Selection Approaches based on the SVM Classification, IEEE 2008.
[12] Fawzi Elias Bekri, A. Govardhan, EMA-QPSO based Feature Selection and Weighted Classification by LS-SVM for Diabetes Diagnosis, International Journal of Engineering and Advanced Technology (IJEAT), 2012.
[13] UCI repository of machine learning Databases, Pima Indian Diabetes Dataset. [Online] Available at: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.
[14] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian , Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search.
[15] Manaswini Pradhan , Dr. Ranjit Kumar Sahu, Predict the onset of diabetes disease using Artificial Neural Network (ANN), International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004) Volume 2, Issue 2, April 2011.
[16] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proc 12th Int Conf on Machine Learning (ICML-03), Washington, D.C., pages 856–863, San Francisco, CA, 2003. Morgan Kaufmann.
[17] Ayush Sood, Steven Diamond, Shizhi Wang, Type 2 Diabetes Mellitus Classification, December 14, 2012.
[18] Hongliang Fei, Brian Quanz and Jun Huan, GLSVM: Integrating Structured Feature Selection and Large Margin Classification.
[19] Murali S. Sanker, Using neural network to predict onset of diabetes mellitus, journal of chemical and information and computer sciences.
[20] S.Vijiyarani, S.Sudha, Disease Prediction in Data Mining Technique – A Survey, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013.
[21] Guyon I, Elisseeff A. An introduction to variable and feature selection, J Mach Learn Res 2003; 3: 1157—82.
[22] Ashwinkumar.U.M and Dr Anandakumar.K.R, Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques, 2 2nd International Conference on Computer Design and Engineering, IPCSIT vol. 49 (2012).