International Journal of Engineering and Applied Sciences (IJEAS)
ISSN: 2394-3661, Volume-2, Issue-5, May 2015

Classification of Diabetes Mellitus Using Machine Learning Techniques

Amit kumar Dewangan, Pragati Agrawal

Abstract— **Diabetes-Mellitus refers to the metabolic disorder that happens from misfunction in insulin secretion and action. It is characterized by hyperglycemia. The persistent hyperglycemia of diabetes leads to damage, malfunction and failure of different organs such as kidneys, eyes, nerves, blood vessels and heart. In the past decades several techniques have been implemented for the detection of diabetes. The diagnosis of diabetes is very important now a days using various types of techniques. Here, there are various techniques, their classification and implementation using various types of software tools and techniques. The diagnosis of diabetes can be done using Artificial Neural Network, K-fold cross validation and classification, Vector support machine, K-nearest neighbor method, Data Mining Algorithm, etc. Using these techniques, we attempt to make an ensemble model by combining two techniques: Bayesian classification and Multilayer Perceptron for the accuracy, sensitivity and specificity measures of diagnosis of diabetes-mellitus.**

Index Terms— **Diabetes, Classification, Accuracy, Sensitivity, Specificity, Hybrid Model, Ensemble Model.**

I. INTRODUCTION

In medical science, diagnosis of health condition is a very challenging task. Diabetes Mellitus is one of the most important serious challenges in both developed and developing countries. Medical history data comprises of a number of tests essential to diagnose a particular disease and the diagnosis are based on the experience of the physician; a less experience physician can diagnose a problem incorrectly. Data mining applications can greatly benefit all parties involved in the health care industry. In health care, there is a vast data, and this data has no supervisory value until converted into information and knowledge, which can help limit costs, enhance profits, and maintain high quality of patient care. The main aim of this work is the detection of Diabetes Mellitus using an hybrid model classification comprised of Bayesian classification and Multilayer Perceptron and classify the data as diabetic and non diabetic. Generally, Most of the cases of diabetes are categorized into two main etiopathogenic categories:

- Type 1
- Type 2

A. *TYPE-1 DIABETES*

It occurs due to complete lacking of insulin secretion. It is detected by serological corroboration of an autoimmune pathology process occurring in islets of pancreas and by genetic markers.

In Type-1 diabetes:

The stages are:

I. Normoglycemia
(normal glucose regulation)

II. Hyperglycemia leads to :
   i. Impaired Glucose Tolerance
   (Prediabetes)
   ⇩
   ii. Diabetes Mellitus
      a. Not insulin requiring.
      b. Insulin requiring for control.
      c. Insulin requiring for survival.

B. *TYPE-2 DIABETES*

It occurs due to combination of resistance to insulin action and deficient compensatory insulin secretory response.

In Type-2 diabetes:
The stages are:

I. Normoglycemia
(normal glucose regulation)

II. Hyperglycemia leads to :
   i. Impaired Glucose Tolerance
   (Prediabetes)
   ⇩
   ii. Diabetes Mellitus
   (Not insulin requiring)

These are the types of diabetes [1].

Classification is one of the most important decision making techniques in many real world problem. In this paper, the main objective is to classify the data as diabetic or non diabetic and improve the classification accuracy. For many

Amit kumar Dewangan, Assistant Professor, Deptt. of Computer Science & Information technology Dr. C V Raman University, Kota, Bilaspur (C.G.), India
Pragati Agrawal, Research Scholar, Deptt. of Computer Science & Information technology, Dr. C V Raman University, Kota, Bilaspur (C.G.), India

classification problem, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data set for testing.

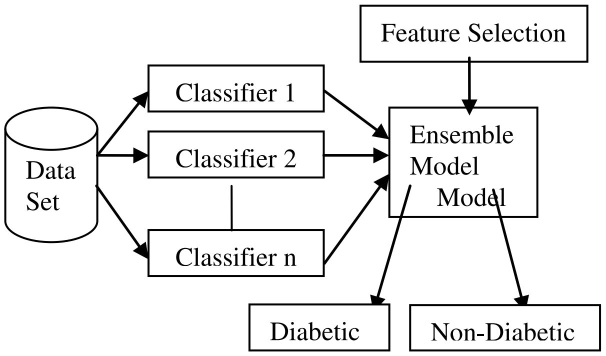

Figure 1.1: Proposed model

The main task of this research work is partitioned into three stages:

First, Classification accuracy achieved with individual model.

Second, Ensemble model used to achieve high accuracy compare to its individual model.

Third, Feature Selection technique applied on best ensemble model in order to achieve high accuracy which is computationally efficient.

In this work, we have used various classification techniques for classification of diabetic and non diabetic data. This data set is binary class problem data set either is diabetic or non diabetic.

II. RELATED WORK :

Anand R. et al. (2013) have suggested Higher Order Neural Network and PCA for classification of pima Indian diabetes data set. Here we apply Error-Back Propagation Based Learning using Norm-Square Based Error. These type of functions are basically used in order to solve complete problems and it needs low number of parameters in comparision with known typical models. The proposed model gives lower mean square error and faster convergence is attained with PCA preprocessing [2].

Christobel Y. A. et al. (2013) have proposed a new class-wise K-Nearest Neighbor (CKNN) classification algorithm for classification of diabetes data. They have used diabetes data set for testing the CKNN algorithm and compared the various performance like accuracy, sensitivity and specificity with simple KNN. The proposed CKNN model gives better classification accuracy as 78.16% compared to simple KNN. Other performance measures are also better than simple KNN [3].

Kumari V. Anuja (2013) have proposed support vector machine (SVM) with Radial basis kernel function for classification of diabetes data. They have used Pima Indian diabetes data set which is collected from UCI repository and trained and tested on SVM as classifier. The proposed model achieved 78% accuracy which can be successfully used for diagnosing diabetes disease [4].

Parashar A. et al. (2014) have proposed Linear Discriminant Analysis and Support Vector Machine for the diagnosis of Pima Indians Diabetes dataset, where LDA reduces feature subsets and SVM is responsible to classify the data. They have also compared SVM with feed forward neural network (FFNN) but our proposed SVM+LDA gives better classification accuracy as 77.60% with 2 features [5].

Farahmandian M. et al. (2015) have applied diabetes data set on various classification algorithms like SVM, KNN, Naïve bayes,ID3, CART and C5.0 to classify the diabetes data. They have compared the classification accuracy of these models. SVM gives best classification accuracy as 81.77% compare to others [6].

III. METHODOLOGY :

In this research work, we will use data mining techniques like Multi layer Perceptron, and the Bayesian Net classification techniques and ensemble them for classification of diabetes data: MLP is a development from the simple perceptron in which extra hidden layers (additional to the input and output layers, not connected externally) are added. In this process, More than one hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and only possible) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes a trial-and-error process, determined by the nature of the problem at hand. The transfer function is generally a sigmoidal function. Multilayer Perceptron is a neural network that trains using back propagation learning [7].

Bayesian Net is a statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class or not. Let, X is a data sample whose class label is unknown. Let, H be some hypothesis, such that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds the given observed data sample X. P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X. Two or more models together form a new model is called an ensemble model. An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting are two techniques that is used in a combination of models. Each combines a series of k learned models (classifiers), $M1$, $M2$,…..Mk, with the aim of creating an improved composite model, M. Both bagging and boosting can be used for classification [8].

Feature subset selection is an important problem in knowledge discovery, not only for the insight gained for determining relevant modeling variables, but also for the improved International Journal of Decision Science & Information Technology, understandability, scalability, and, possible accuracy of the resulting models. In the Feature selection the main goal is to find a feature subset that produces higher classification accuracy. In this research work , we have used Information gain feature selection technique.

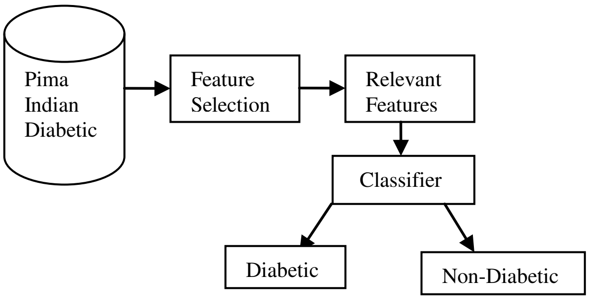

Figure 3.1: Feature Selection Model

A. DATASET AND MEASURES :

This section contains description of data set and performance measures of model shown below:

4.1 Pima Indian Diabetes Data Set

The Pima Indian diabetes data set is collected from UCI repository which is classified under two method diabetic and non diabetic. This data set consists of 8 attributes and 1 class. This data set also consists 768 instances [9].

Feature _Id	**Feature _Name**
F1	Pregnant
F2	Plasma glucose
F3	Diastolic Blood Pressure
F4	Triceps Skin Fold Thickness
F5	Serum-Insulin
F6	Body Mass Index
F7	Diabetes Pedigree Function
F8	Age
Class	Diabetic or Non- Diabetic

Table 4.1: Pima Indian Diabetes Data Set

1.2 Performance Measures

Performance of model can be evaluated various performance measures: classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [10].

Actual Vs. Predicted	**Positive**	**Negative**
Positive	**TP**	**FN**
Negative	**FP**	**TN**

Table 4.2: Confusion matrix

Various performance measures like sensitivity, specificity and accuracy are calculated using this matrix :

Accuracy	(TP+TN)/(TP+FP+TN+FN)
Sensitivity	TP/ (TP+FN)
Specificity	TN/ (TN +FP)

Table 4.3: Performance measures

The ranking of features from less important to high important as shown F3, F7, F1, F4, F5, F8, F6, F2. In this experiment we have eliminated the less important feature one by one and give to the best model (MLP+Bayesian Net). We have achieved same accuracy with 6 feature subset as we have achieved in full feature subset. There are 8 feature in the data set after removing F3 and F7 features from data set , we have got 81.89% accuracy of model which shows that our proposed model gives better performance with less number of features. Table 4.4 shows various performance measures of best model and Table 4.5 shows that accuracy of model with different feature subsets.

Actual Vs. Predicted	Diabetic	Non Diabetic
Diabetic	17	11
Non Diabetic	10	78

Table 4.4: Confusion matrix of best model with 6 features

Number of features	Remove feature	Accuracy
8	Full Features	81.89
7	F3	81.03
6	F3,F7	**81.89**
5	F1, F3,F7	76.72
4	F4, F1, F3,F7	80.17
3	F5, F4, F1, F3,F7	81.03
2	F8, F5, F4, F1, F3,F7	77.58
1	F6, F8, F5, F4, F1, F3,F7	70.68

Table 4.5: Feature Selection on best model

Finally we can say that our proposed model is acceptable for classification of diabetes.

IV. EXPERIMENTAL WORK

This experiment have done with the help of open source data mining tools in window environment using net beans software. In this experiment, we have used java code and libraries which are available in WEKA.

In this experiment, we have used various individuals and hybrid classification models for classification of diabetes data. The analysis of models are done in two steps: first model is trained and tested. Various data mining techniques like C4.5, random forest (RF), Bayes Net and Multi Layer Perceptron (MLP) are trained using randomly training data set and after that the testing of the trained models is done using randomly tested data set. Partitions of data plays very important role in accuracy of models. Accuracy is varying from partition to partition. For example, when we see the accuracy of C4.5 model, 77.08% in case of 75-25%

training-testing partitions, 76.72% in case of 85-15% training-testing partitions and 75.32% in case of 90-10% training-testing partitions. Similarly accuracy is varying for others model in different partitions.

A hybrid model is suitable when its accuracy is high compare to its individual model. In this experiment, we have hybrid C4.5 and RF, similarly hybrid MLP and Bayes Net classification model to develop a robust model for classification of diabetes data. It is not necessary to improve classification accuracy in each partition. In case of C4.5+RF gives 79.31% of accuracy which is higher than its individuals model. Similarly, MLP+BayesNet, gives 81.89% of accuracy which is higher than its individuals model. In case of both ensemble model 85-15% training-testing partitions play important role for classification of diabetes data. Our proposed MLP + Bayes Net gives 81.89% of accuracy which is robust model for classification of data.

V. RESULT AND CONCLUSION:

In this study, we have taken various classification methods and ensemble them to give the new hybrid model in the search of finding the better result in terms of Accuracy, Specificity and Sensitivity. According to Table 4.5, we came to the conclusion that our model has achieved the highest Accuracy of 81.89% with 6 features and with the help of Table 4.3, it achieve the highest Sensitivity of 64.10% and achieve the highest Specificity of 90.90%.

Accuracy	81.89%
Sensitivity	64.10%
Specificity	90.90%

REFERENCES

[1] “Diagnosis & Classification of Diabetes Mellitus”, Diabetes Care, Volume 37, Supplement 1, 2014, pp. S81-S90.

[2] Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse, “ K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA ”, IJSCE, Volume-2, Issue-6, January 2013, pp. 436-438, ISSN: 2231-2307.

[3] Y. Angeline Christobel, P.Sivaprakasam, “ A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset”, IJEAT, Volume-2, Issue-3, February 2013, pp. 396-400, ISSN: 2249 – 8958.

[4] Kumari V. Anuja, Chitra R. (2013). Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. Vol. 3, pp. 1797-1801, ISSN: 2248-9622.

[5] Parashar A., Burse K., Rawat K. (2014). A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, pp. 378-383, ISSN: 2277 128X.

[6] Farahmandian M., Lotfi Y., Maleki I. (2015). Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study. MAGNT Research Report. Vol. 3, PP. 989-997, ISSN. 1444-8939.

[7] Pujari A. K. et al. (2012). Improving Classification Accuracy by Using Feature Selection and Ensemble Model. International Journal of Soft Computing and Engineering (IJSCE). International Journal of Soft Computing and Engineering (IJSCE)Vol. 2,pp. 380-386.

[8] Han, J.,& Micheline, K. (2006). Data mining: Concepts and Techniques, Morgan Kaufmann .Publisher.

[9] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: http://www.ics.uci.edu/~mlearn/databases/thyroid -disease/newthyroid.data(Accessed: 12 Jan 2015).

[10] H. S. Hota, Akhilesh Kumar Shrivas, S. K. Singhai (2011).An Ensemble Classification Model for Intrusion Detection System with Feature Selection, International Journal of Decision Science of Information Technology, Vol. 3, No. 1, 2011, pp.13-24.