A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier

Mani Butwall
M. Tech. Scholar
Computer Science Department
Sushila Devi Bansal College of Technology, Indore
(India)

Shraddha Kumar
Asst. Professor
Computer Science Department
Sushila Devi Bansal College of Technology, Indore
(India)

ABSTRACT

Diabetes mellitus is an interminable disease that forces excessively high human, social and financial expenses for a nation. Additionally, minimizing its commonness rate and in addition its excessive and risky confusions requires viable administration. Diabetes administration depends on close participation between the patient and health awareness experts.

Data mining gives a diversity of methods to investigate large data keeping in mind the end goal to find hidden knowledge. This study is an effort to plan and execute a descriptive data mining approach and to devise association standards to envisage diabetes behaviour in arrangement with particular life style parameters, including physical activity and emotional states, especially in elderly diabetics. Proposed methodology is based on Random Forest Classifier.

General Terms

Medical Data mining, Classification using Random Forest Classifier.

Keywords

Data mining, Diabetes mellitus, Random Forest Classifier. Pima Indian Diabetic Database

1. INTRODUCTION

Data mining with greatly concentrated and far reaching applications in numerous associations is getting to be progressively famous and even crucial in medical services. Data mining applications can extraordinarily advantage all gatherings included in the medical service industry. It can be utilized via; medical insurers to distinguish fraud and misuse, insurance associations to settle on client relationship administration decisions, doctors to recognize powerful medications and patients to get better and more reasonable medical services. Data mining gives philosophy and innovation to process and investigate immense measures of data into helpful information for decision making that is an essential piece of medical service management.

Medicinal services frameworks are rich in information yet poor in knowledge so there is tremendous need of having methods and apparatuses to extract the information from the huge data set so that medical diagnosis could be possible. Medicinal diagnosis is viewed as an imperative yet confounded undertaking that needs to be executed precisely and effectively.

2. RELATED WORK ON DIABETIC DATASET CLASSIFICATION

2.1 Diabetes Mellitus

Diabetes is a standout amongst the most well-known non-transmittable diseases in the world. It is assessed to be the fourth or fifth cause for death in most of the nation [1].

Diabetes is an unending malady that happens when the body cannot sufficiently deliver insulin or cannot utilize it adequately which brings about BG (Blood Glucose) not to be ingested appropriately by the body cells and stays circulation in the blood.

Categories of diabetes:

- Type-1 diabetes
- Type-2 diabetes
- Gestational diabetes

Type 1 diabetes incorporates diabetes that is basically an aftereffect of pancreatic beta cell devastation that prompts insufficient creation of insulin. This type can influence any age yet typically happens in youngsters and youth. These diabetics can lead a typical life through mix of a day by day insulin treatment, solid eating regimen, close checking and normal physical activity [2].

Type 2 diabetes is the most well-known, one that generally happens in adult peoples yet is progressively seen in kids and young people, as well. This type is otherwise called an insulin resistance in light of the fact that, in this type the body can create insulin yet possibly it is not adequate or the body can't react to its impact leading glucose remains flowing in the blood. This type likewise incorporates LADA (Latent Autoimmune Diabetes in Adults), depicting a lessened number of individuals with diabetes type 2 who seem to have an insusceptible intervened loss of pancreatic beta cells [3]. Numerous type 2 diabetics can control their BG level through a healthy eating routine and an expanded physical movement.

Gestational diabetes mellitus alludes to a glucose intolerance with onset or first acknowledgment amid pregnancy because of ineffectively oversaw BG. This collection must be nearly checked to control their BG level and minimize the danger for the child. This could be possible by healthy eating regimen, moderate physical activity and at times insulin treatment or oral drug [2].

Proposed framework helps the patients from experiencing different tests like blood tests, checking diastolic and systolic blood pressure and so on. In the wake of having this frameworks, the patent itself is capable to analyze whether he/she is experiencing Diabetes Mellitus or not. These estimations are in view of the indications' which happens in right on time phases of Diabetes Mellitus and on some physical conditions.

2.2 Literature Review of Classification of Diabetic Dataset

Statistics given by the Centers for Disease Control states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modelling has 50784 records with 37 variables. After computational calculation they found 34% of the population with age<20 years and 33.9% of the population with 20<age<45 was not affected by diabetes. 26.8% of the population with age>45 years was not diabetic [7].According to another research work, used the Pima Indian diabetic database (PIDD) at the UCI Machine Learning Lab has been tested data mining algorithms to predict their accuracy in diabetic status from the 8 variables given. Out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. Rough sets as a data mining predictive tool applied rough sets to PIDD using ROSETTA software. The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81% [8]. They proposed a system which predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. They used Naïve Bayes Classifier and data set collected from diabetic research institute in Chennai which contain records of about 500 patients. The WEKA tool was used for Data mining with 10 fold cross validation. They found most of the diabetic patients with high cholesterol values are in the age group of 45 – 55, have a body weight in the range of 60 – 71, have BP value of 148 or 230, have a Fasting value in the range of 102 – 135, have a PP value in the range of 88 – 107, and have a A1C value in the range of 7.7 – 9.6 [9]. One more research concluded that the women suffering from diabetes can be tracked by clustering and attribute oriented induction techniques and dataset was collected from National Institute of Diabetes, Digestive and Kidney Diseases. The results were evaluated in five different clusters denoting concentrations of the various attributes and the percentage of women suffering from diabetes and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3 [10]. The data mining approaches in Diabetes diagnosis yields a result of almost 91-92%.

Table 1: Comparative Results of Diabetes Diagnosis Methods

Methods	Accuracy %	Paper
Neural Network	92.8	[11]
Neural Network	92.5	[12]
Rough Sets	82.6	[13]
Logistic Regression	77.19	[14]
C 4.5	91	[15]

3. DATABASE EXPLANATION

Title: Pima Indians Diabetes Database

Sources:

Original owners: National Institute of Diabetes and Digestive and Kidney Diseases
Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Centre, RMI Group Leader, Applied Physics Laboratory, The Johns Hopkins University, Johns Hopkins Road, Laurel, MD 20707 (301) 953-6231(c)
Date received: 9 May 1990

Past Usage: **Smith et al,** [4] investigated the diagnostic, binary-valued variable whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cut-off of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

Relevant Information: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogy of perceptron-like devices. It is a unique algorithm; see the paper for details.

Number of Instances: 768
Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours 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 (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Missing Attribute Values: Yes

Class Distribution: (Class value 1 is interpreted as "tested positive for diabetes")

Table 2:Class Distribution

Class Value	Number of instances
0	500
1	268

Table 3:Brief Statistical Analysis

Attribute Number	Mean	Standard Deviation
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

60% of data used for training
40% of data for testing.

4. RANDOM FOREST CLASSIFIER

A random forest classifier is the assembly of tree-structured classifiers (Figure 1) [5]. This algorithm supplements the objects from array of input to every tree of forest. The elements of the unit vector are individually voted for classification by every single tree. The forest filters the most voted classifications out of the forest.

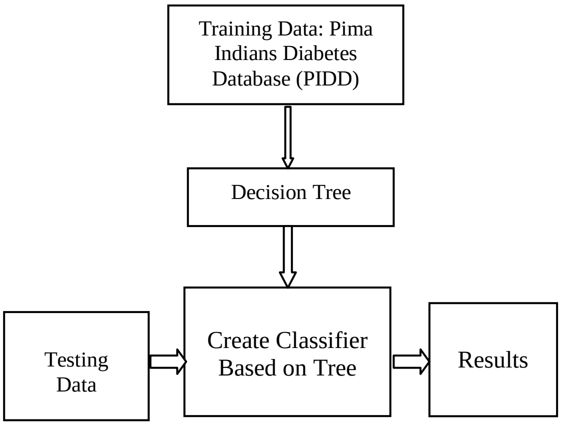
Training Data: Pima Indians Diabetes Database (PIDD)
Decision Tree
Testing Data
Create Classifier Based on Tree
Results

Figure 1: Algorithmic Sequence of Random Forest Classifier

The pseudo code for generation of tree is:

1. N is the total number of cases in training set, sampling of every case is executed randomly to replace from original data. The sampling set generated is the training set for tree formation.
2. m Predictors are anonymously selected at every node out of input variables of size M where ($m \ll M$). The nodes are split by the best split on m predictors. The value of m is kept stationary during forest grow.
3. By default, $m = \sqrt{M}$
4. M (to minimize the distance with optimal results)
5. The tree is allowed to extend the maximum achievable height without any shear.

About one-third of the cases are subsided or neglected in the training set of a tree when drawn by sampling with replacement. This Out-of-Bag data is employed for the test set that gets unbiased estimation of the classification errors. Henceforth the need for separate test and cross validation to get an unbiased estimate minimizes of the test set error. The out-of-bag estimates are unbiased [5].

5. SIMULATION AND RESULTS

The performance of proposed technique has been studied by means of MATLAB simulation.

Test Result for Diabetic

```
Number of times pregnant: 6
Plasma glucose concentration a 2 hours in an oral glucose tolerance test: 148
Diastolic blood pressure (mm Hg): 72
Triceps skin fold thickness (mm): 35
2-Hour serum insulin (mu U/ml): 0
Body mass index (weight in kg/(height in m)^2): 33.6000
Diabetes pedigree function: 0.6270
Age (years): 50

------------------------------------------------------------
Testing with Neural Network:

There are 100 % chances that person is diabetic.
So, Neural Network Considering data as diabetic.
------------------------------------------------------------
Testing with Random Forest:

There are 96.7 % chances that person is diabetic.
So, Random Forest Considering data as diabetic.
---------------------------------------------------------
```

Test Result for Non-diabetic

```
Number of times pregnant: 0
Plasma glucose concentration a 2 hours in an oral glucose tolerance test: 137
Diastolic blood pressure (mm Hg): 40
Triceps skin fold thickness (mm): 35
2-Hour serum insulin (mu U/ml): 168
Body mass index (weight in kg/(height in m)^2): 43.1000
Diabetes pedigree function: 2.2880
Age (years): 33

------------------------------------------------------------
Testing with Neural Network:

There are 0.195 % chances that person is diabetic.
So Neural Network Considering data as non-diabetic.
------------------------------------------------------------
Testing with Random Forest:

There are 20 % chances that person is diabetic.
So, Random Forest Considering data as non-diabetic.
------------------------------------------------------------
```

Result of Neural Network (Previous method [6])

Feed forward NN with;

- Input layer - 8 neuron
- 1 hidden layer - 8 neuron
- Output layer - 2 neuron

Training method - Scaled Conjugate Training

Performance criteria – MSE

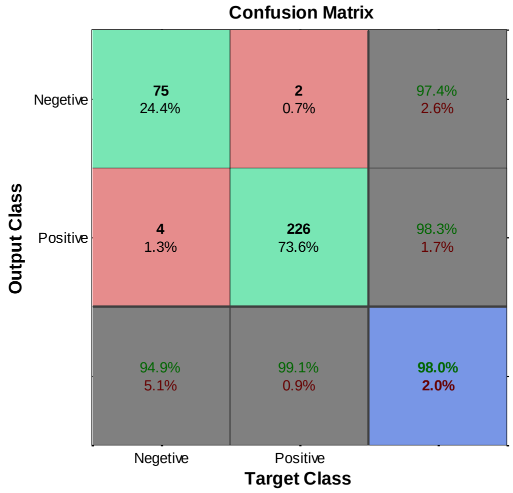

Figure 2: Confusion Matrix for Neural Network based previous approach [5]

Random Forest Classifier (Proposed Method)

Forest of 100 trees with 20 splits and depth of 9.

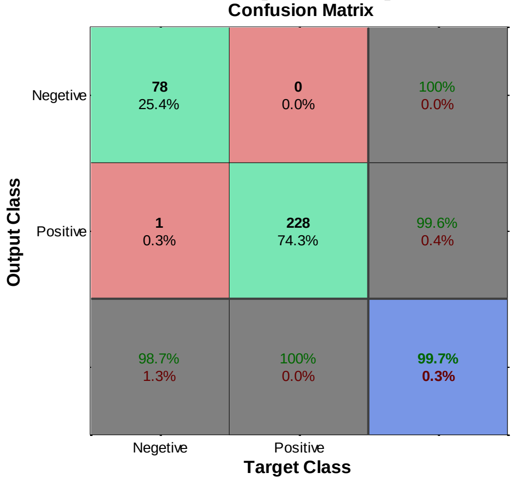

Figure 3: Confusion Matrix for proposed Random forest classifier approach

6. CONCLUSION AND FUTURE WORK

Data mining and machine learning algorithms in the medical field extracts distinctive concealed patterns from the medical data. They can be utilized for the examination of vital clinical parameters, expectation of different diseases, estimating assignments in pharmaceutical, extraction of medical knowledge, treatment planning support and patient administration. Various algorithms are present in literature for the prediction and finding of diabetes. These techniques give more precision than the accessible conventional frameworks. In this research work, Random Forest classifier has been used with different test parameters and it was found that it is effective in diagnosis of Diabetes mellitus when the person provide the required attributes value. Figure 2 and Figure 3 shows the confusion matrix for Neural Network approach [6] and the proposed Random Forest Classifier based approach respectively. After analyzing the confusion matrix we can say that our Random Forest Classifier based approach outperforms better with the accuracy of 99.7%. Future work can be directed on the use of hybrid classification algorithms to enhance the overall performance of the proposed system

7. REFERENCE

[1] IDF Diabetes Atlas, 6th edition, “International Diabetes Federation”, 2013, Online available at: http://www.idf.org/diabetesatlas

[2] “Definition, classification and diagnosis of diabetes, Prediabetes and metabolic syndrome”, Canadian Journal of Diabetes, Vol. 37, Canadian Diabetes Association. 2013. Online available at: www.canadianjournalofdiabetes.com.

[3] “Diagnosis and classification of diabetes mellitus”, American Diabetes Association, Diabetes Care, vol. 35, Supplement 1, 2012.

[4] Jack W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes, “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus”, IEEE Symposium on Computer Applications and Medical Care, pp. 261-265, 1988.

[5] Brieman L, Random Forests Statistics, Department University of California Berkeley, 45, 5-32, 2001.

[6] Sonu Kumari, Archana Singh, “A Data Mining Approach for the Diagnosis of Diabetes Mellitus”, IEEE 7th International Conference on Intelligent Systems and Control (ISCO), pp 373 – 375, 2013.

[7] Eurekalert, "Insufficient sleep may be linked to increased diabetes risk," July 11, 2010, <http://lifesciencelog.com/cluster53092620/.

[8] Bhatt K., Dalal P., Panwar A., “A Cluster Centres Initialization Method for Clustering Categorical Data Using Genetic Algorithm” International Journal of Digital Application & Contemporary research, 2013, Volume-2 Issue-1

[9] Huang, Zhexue, and Michael K. Ng. "A fuzzy k-modes algorithm for clustering categorical data." *Fuzzy Systems, IEEE Transactions on* 7.4 (1999): 446-452.

[10] Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." *DMKD*. 1997.

[11] Kumari, Sonu, and Archana Singh "A data mining approach for the diagnosis of diabetes mellitus" Intelligent Systems and Control (ISCO), 2013 7th International Conference on IEEE, 2013.

[12] Rajeeb Dey and Vaibhav Bajpai and Gagan Gandhi and Barnali Dey, "Application of artificial neural network technique for diagnosing diabetes mellitus", 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10

[13] Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse" Artificial Intelligence in Medicine 26.1 (2002): 37-54.

[14] Sa-ngasoongsong, Akkarapol, and Jongsawas Chongwatpol "An Analysis of Diabetes Risk Factors Using Data Mining Approach" *Oklahoma state university, USA* (2012)

[15] Rajesh, K., and V. Sangeetha. "Application of Data Mining Methods and Techniques for Diabetes Diagnosis." *age* 2.3 (2012)