ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Application of Data Mining Methods and Techniques for Diabetes Diagnosis

K. Rajesh, V. Sangeetha

Abstract*-- Medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. This project aims for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.

Index Terms — **Data Mining, Healthcare, Diabetes Research, Clinical Data, Classification, Diabetes Dataset.**

I. INTRODUCTION

Diabetes mellitus, or simply diabetes, is a set of related diseases in which the body cannot regulate the amount of sugar in the blood [1]. It is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. This high blood sugar produces the classical symptoms of polyuria, polydipsia and polyphagia [2]. There are three main types of diabetes mellitus (DM). Type 1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as non insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes". The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. As of 2000 it was estimated that 171 million people globally suffered from diabetes or 2.8% of the population. Type-2 diabetes is the most common type worldwide [3]. Figures for the year 2007 show that the 5 countries with the largest amount of people diagnosed with diabetes were India (40.9 million), China (38.9 million), US (19.2 million), Russia (9.6 million), and Germany (7.4 million) [3]. Data Mining [4] refers to extracting or mining knowledge from large amounts of data. The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain. Classification [5] maps data into predefined groups. It is often referred to as supervised learning as the classes are determined prior to examining the data. Classification Algorithms usually require that the classes be defined based on the data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to class. Pattern Recognition is a type of classification where an input pattern is classified into one of the several classes based on its similarity to these predefined classes. Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns in data which involves Selection, Pre-processing, Transformation, Data Mining and Evaluation.

II. RELATED WORK

Santi Wulan Purnami et al. [6], in their research work used support vector machine for feature selection and classification of breast cancer and also emphasizes how 1-norm SVM can be used in feature selection and smooth SVM (SSVM) for classification. Two problems addressed here are, the first is to identify the importance of the parameters on the breast cancer. The second research problem is to diagnose breast cancer based on nine attributes of Wisconsin breast cancer dataset. To identify the importance of the parameters, the 1-norm SVM of the original data was done. The stronger parameters are as follows: parameter 1 (Clump thickness), parameter 3(Uniformity Of Cell shape), parameter 6 (Bare Nuclei), parameter 7 (Bland Chromatin), and parameter 9(Mitoses), while parameter 2 (Uniformity Of Belsize), parameter 4 (Marginal Adhesion), parameter 5(Single Epithelial Cell Size) and parameter 8 (Normal Nucleoli) are weaker. The obtained training and testing classification accuracy using 10 fold cross validation were 97.52% and 97.01% respectively. When one of the weak parameters was removed both training and testing shows a little decrease in accuracy. Pardha Repalli [7], In their research work predicted how likely the people with different age groups are affected by diabetes based on their life style activities. They also found out factors responsible for the individual to be diabetic. 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 modeling has 50784 records with 37 variables. They computed a new variable age_new as nominal variable, dividing in to three group's young age, middle age and old age and the target variable diabetes_diag_binary is a binary variable. They found 34% of the population whose age was below 20 years was not affected by diabetes. 33.9% of the population whose

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

age was above 20 and below 45 years was not affected by diabetes. 26.8% of the population whose age was above 45 years was not diabetic. Joseph L. Breault [8], In his research work used the publicly available Pima Indian diabetic database (PIDD) at the UCIrvine Machine Learning Lab. They tested data mining algorithms to predict their accuracy in predicting 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 test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI .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%. Using a group of 10 random samples the mean accuracy was 73.2%. G. Parthiban et al. [9] The main objective of their research paper is to predict the chances of diabetic patient getting heart disease. In this study, we are applying Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. 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. It is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. The data set used in their work was clinical data set collected from one of the leading diabetic research institute in Chennai and contain records of about 500 patients. The clinical data set specification provides concise, unambiguous definition for items related to diabetes. The WEKA tool was used for Data mining. They used 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. Padmaja et al. [10] In their research aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of women suffering from diabetes. They used Data mining functionalities like clustering and attribute oriented induction techniques to track the characteristics of the women suffering from diabetes. Information related to the study was obtained from National Institute of Diabetes, Digestive and Kidney Diseases. The results were presented in the form of clusters. Those clusters denote the concentrations of the various attributes and the percentage of women suffering from diabetes. The results were evaluated in five different clusters 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. The study predicts the state of diabetes i.e., whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering from diabetes with specific characteristics. This is used to effectively in diagnosis and treatment.

III. PROPOSED SYSTEM

We have applied data mining techniques to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes or not. The training dataset used for data mining classification was the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases from UCI Machine Learning Repository [11]. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise, as the data is complete with no missing values. We applied different classification techniques to Pima Indians Diabetes Database and the error results obtained is tabulated in table

IV. PROPOSED SYSTEM DESIGN

A. *Dataset Used*

The training dataset used for data mining classification was the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise, as the data is complete. The diagrammatic representation of the proposed system design is given in Figure.

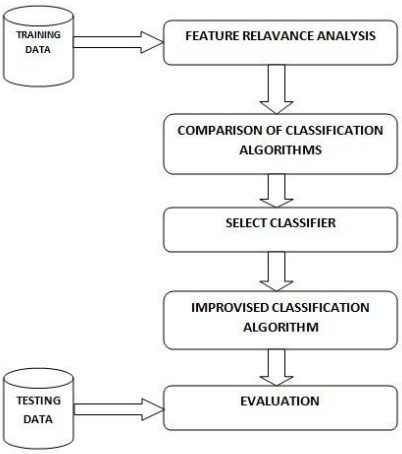

Fig 1. Proposed Architecture

The attributes in the dataset are given in Table I.

Feature selection [12] is the technique that is applied to the dataset to obtain a reduced subset of key attributes to be used in the classification exercise. Feature Relevance Analysis was performed on the given dataset to rank the

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

features in accordance with the relevance to the class label. There are many feature different techniques available for use. As the dataset consists of continuous attributes, filtering techniques which would be effective for such type of data has been selected and applied. The filtering techniques and the results obtained are given in Table II.

Table I: Diabetes Dataset Attributes

S. No	Attributes	Type
1	Number of times pregnant	Continuous
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (kg/m)^2)	Continuous
7	Diabetes pedigree function	Continuous
8	Age (years)	Continuous
9	Class variable (0 or 1)	Discrete

Table II: Filtering Results

Filtering Technique	No. of Attributes	
	Before filtering	*After filtering*
Fisher	8	6
Runs	8	2
ReliefF	8	3
Step Disc	8	5

B. ***Comparison of Classification Algorithms***

We applied different classification techniques to Diabetes dataset and the error results obtained is tabulated in table given below.

Table III: Comparison of Classification Algorithms

S.No	Technique	Error Rate
1	C-RT	0.2148
2	CS-RT	0.2148
3	C 4.5	0.0938
4	ID3	0.2279
5	K-NN	0.1966
6	LDA	0.2161
7	NAÏVE BAYES	0.2461
8	PLS-DA	0.2253
9	SVM	0.2253
10	RND TREE	0.0

In the above ten Classification Algorithms RND TREE gives 100% accuracy but the rule set is huge and this algorithms is suffering from over fitting of data. C4.5 gives ~91% classification. Since C4.5 Algorithm is mainly used for most of the medical application we use C4.5 for the classification.

C. ***C4.5 Classification Algorithm***

C4.5 is a well known decision tree induction learning technique that has been successfully and extensively applied for medical data. C4.5 [13][14] is a software extension of the basic ID3 algorithm designed by Quinlan. The number of attributes and error rates obtained in classification using C4.5 is given in Table IV.

Table IV: Feature Relevance Analysis Results

Filtering Technique	No. of Attributes		Error Rate in Classification	
	Before filtering	*After filtering*	*Before filtering*	*After filtering*
Fisher	8	6	0.0938	0.1224
Runs	8	2	0.0938	0.1875
Relief F	8	3	0.0938	0.1576
Step Disc	8	5	0.0938	0.1237

It can be observed that C4.5 algorithm gives a classification rate of ~ 91% without feature relevance.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

However, when feature relevance technique is applied, the classification rate decreases to lesser than 88% . The classification rules obtained by applying C4.5 algorithm is given below.

Plasma glucose concentration < 127.5000
Body mass index < 26.4500
then Class variable = **Tested Negative**
Body mass index >= 26.4500
Age < 28.5000
Body mass index < 30.9500
Then **Class variable = Tested Negative**
Body mass index >= 30.9500
2-Hour serum insulin < 168.5000
Triceps skin fold thickness < 44.5000
Triceps skin fold thickness < 40.5000
Diastolic blood pressure < 53.0000
then **Class variable = Tested Positive**
Diastolic blood pressure >= 53.0000
Diastolic blood pressure < 79.0000
Plasma glucose concentration < 92.5000
then **Class variable = Tested Negative**
Plasma glucose concentration >= 92.5000
Body mass index < 33.7000
2-Hour serum insulin < 65.0000 then Class variable = **Tested Positive**
2-Hour serum insulin >= 65.0000 then Class variable = **Tested Negative**
Body mass index >= 33.7000 then Class variable = **Tested Negative**
Diastolic blood pressure >= 79.0000
Plasma glucose concentration < 93.0000 then Class variable = **Tested Negative**
Plasma glucose concentration >= 93.0000
Body mass index < 36.5500 then Class variable = **Tested Negative**
Body mass index >= 36.5500 then Class variable = **Tested Positive**
Triceps skin fold thickness >= 40.5000 then Class variable = **Tested Positive**
Triceps skin fold thickness >= 44.5000 then Class variable = **Tested Negative**
2-Hour serum insulin >= 168.5000 then Class variable = **Tested Negative**
Age >= 28.5000
Plasma glucose concentration < 99.5000
2-Hour serum insulin < 88.0000
2-Hour serum insulin < 21.0000
Number of times pregnant < 3.5000 then Class variable = **Tested Negative**
Number of times pregnant >= 3.5000
Triceps skin fold thickness < 20.5000 then Class variable = **Tested Negative**
Triceps skin fold thickness >= 20.5000
Diabetes pedigree function < 0.2885 then Class variable = **Tested Negative**
Diabetes pedigree function >= 0.2885 then Class variable = **Tested Positive**
2-Hour serum insulin >= 21.0000 then Class variable = **Tested Negative**
2-Hour serum insulin >= 88.0000 then Class variable = **Tested Positive**
Plasma glucose concentration >= 99.5000
Diastolic blood pressure < 91.0000
Diabetes pedigree function < 0.5610
Age < 54.5000
Triceps skin fold thickness < 28.0000
Body mass index < 27.9500 then Class variable = **Tested Positive**
Body mass index >= 27.9500
Age < 29.5000 then Class variable = **Tested Negative**
Age >= 29.5000
Body mass index < 29.6500 then Class variable = **Tested Negative**
Body mass index >= 29.6500 then Class variable = **Tested Positive**
Triceps skin fold thickness >= 28.0000
Age < 41.0000
Plasma glucose concentration < 122.5000
Plasma glucose concentration < 111.5000 then Class variable = **Tested Negative**
Plasma glucose concentration >= 111.5000
Body mass index < 37.0000 then Class variable = **Tested Positive**
Body mass index >= 37.0000 then Class variable = **Tested Negative**
Plasma glucose concentration >= 122.5000 then Class variable = **Tested Negative**
Age >= 41.0000 then Class variable = **Tested Negative**
Age >= 54.5000 then Class variable = **Tested Negative**
Diabetes pedigree function >= 0.5610
Number of times pregnant < 6.5000
2-Hour serum insulin < 120.5000
Age < 34.5000 then Class variable = **Tested Negative**
Age >= 34.5000 then Class variable = **Tested Positive**
2-Hour serum insulin >= 120.5000 then Class variable = **Tested Positive**
Number of times pregnant >= 6.5000 then Class variable = **Tested Positive**
Diastolic blood pressure >= 91.0000 then Class variable = **Tested Negative**
Plasma glucose concentration >= 127.5000
Body mass index < 29.9500
Body mass index < 23.2000 then Class variable = **Tested Negative**
Body mass index >= 23.2000
Age < 60.5000
Plasma glucose concentration < 160.0000
Age < 21.5000 then Class variable = **Tested Negative**
Age >= 21.5000
2-Hour serum insulin < 132.5000
Triceps skin fold thickness < 28.0000
Number of times pregnant < 1.5000 then Class variable = **Tested Negative**
Number of times pregnant >= 1.5000

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Number of times pregnant < 3.5000 then Class variable = **Tested Positive**
Number of times pregnant >= 3.5000 then Class variable = **Tested Negative**
Triceps skin fold thickness >= 28.0000 then Class variable = **Tested Positive**
2-Hour serum insulin >= 132.5000 then Class variable = **Tested Negative**
Plasma glucose concentration >= 160.0000 then Class variable = **Tested Positive**
Age >= 60.5000 then Class variable = **Tested Negative**
Body mass index >= 29.9500
Diastolic blood pressure < 61.0000 then Class variable = **Tested Positive**
Diastolic blood pressure >= 61.0000
Diastolic blood pressure < 96.5000
Plasma glucose concentration < 157.5000
Age < 30.5000
2-Hour serum insulin < 260.0000
Diabetes pedigree function < 0.3315 then Class variable = **Tested Negative**
Diabetes pedigree function >= 0.3315
Diabetes pedigree function < 0.3730 then Class variable = **Tested Positive**
Diabetes pedigree function >= 0.3730
Triceps skin fold thickness < 28.5000
Diastolic blood pressure < 73.0000 then Class variable = **Tested Positive**
Diastolic blood pressure >= 73.0000 then Class variable = **Tested Negative**
Triceps skin fold thickness >= 28.5000 then Class variable = **Tested Negative**
2-Hour serum insulin >= 260.0000 then Class variable = **Tested Negative**
Age >= 30.5000
Triceps skin fold thickness < 45.0000
Diabetes pedigree function < 0.4305
Triceps skin fold thickness < 31.0000
2-Hour serum insulin < 50.0000
Diabetes pedigree function < 0.2265 then Class variable = **Tested Negative**
Diabetes pedigree function >= 0.2265 then Class variable = **Tested Positive**
2-Hour serum insulin >= 50.0000 then Class variable = **Tested Negative**
Triceps skin fold thickness >= 31.0000 then Class variable = **Tested Positive**
Diabetes pedigree function >= 0.4305
Age < 44.5000
Plasma glucose concentration < 132.0000 then Class variable = **Tested Positive**
Plasma glucose concentration >= 132.0000
Number of times pregnant < 7.5000 then Class variable = **Tested Negative**
Number of times pregnant >= 7.5000 then Class variable = **Tested Positive**
Age >= 44.5000 then Class variable = **Tested Positive**
Triceps skin fold thickness >= 45.0000 then Class variable = **Tested Positive**
Plasma glucose concentration >= 157.5000
Body mass index < 46.1000
Body mass index < 40.8500
Triceps skin fold thickness < 26.5000
Diastolic blood pressure < 69.0000 then Class variable = **Tested Negative**
Diastolic blood pressure >= 69.0000 then Class variable = **Tested Positive**
Triceps skin fold thickness >= 26.5000 then Class variable = **Tested Positive**
Body mass index >= 40.8500 then Class variable = **Tested Positive**
Body mass index >= 46.1000 then Class variable = **Tested Positive**
Diastolic blood pressure >= 96.5000 then Class variable = **Tested Positive**

V. EVALUATION

The classification algorithm predicts the class label. The final output will be patterns which are used to find out whether the person is affected with Diabetes or not. The accuracy [4] of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Some of the performance measures are given below. A confusion matrix is a useful tool for analyzing classifier accuracy. Structure of confusion matrix is given below.

Table V: confusion matrix

	C1	C2
C1	True positives	False negatives
C2	False positives	True negatives

True Positive (TP) refers to positive tuples that were correctly labeled by the classifier. True Negative (TN) refers to negatives tuples that were correctly labeled by the classifier. False Positive (FP) refers to negatives tuples that were incorrectly labeled by the classifier. False Negative (FN) refers to positive tuples that were incorrectly labeled by the classifier.

Accuracy: Accuracy is the percentage of tuples that are correctly classified by the classifier

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Recall: Recall is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured.

$$\text{Recall} = TP / (TP + FN)$$

Precision: Precision is the proportion of the examples which truly have class x among all those which were classified as class x.

$$\text{Precision} = TP / (TP + FP)$$

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

VI. CONCLUSION AND FUTURE WORK

We have applied many classification algorithms on Diabetes dataset and the performance of those algorithms have been analysed. A classification rate of 91% was obtained for C4.5 algorithm. Future enhancement of this work includes improvisation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy in classification.

REFERENCES

[1] http://www.emedicinehealth.com/diabetes.

[2] http://en.wikipedia.org/wiki/Diabetes_mellitus.

[3] http://diabetes.co.in.

[4] Han, J., Kamber, M.: Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers (2000).

[5] Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.

[6] Santi Wulan Purnami, S.P. Rahayu and AbdullahEmbong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

[7] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach".

[8] Joseph L. Breault., "Data Mining Diabetic Databases:Are Rough Sets a Useful Addition? " .

[9] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[10] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[11] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu.

[12] Huan Liu, Hiroshi Motoda. Feature selection for knowledge discovery and data mining.

[13] Knowledge Discovery in Databases, http://www2.cs.uregina.ca.

[14] C4.5 Algorithm Description, http://en.wikipedia.org /wiki/C4.5_algorithm.

Authors Profile

Mr.Rajesh. K has completed his B.Tech in Information Technology at Rajiv Gandhi College of Engineering affiliated to Anna University, Chennai, India and completed his M.E. in Computer Science and Engineering at Rajalakshmi College of Engineering, affiliated to Anna University of Technology, Chennai, India. He is a CISCO certified network associate. His areas of interest include Computer Networks, Network Security and Data Mining.

Ms. V.Sangeetha received the B.E.,degree in Computer Science and Engineering from Rajalakshmi Engineering College, Anna University Chennai, in 2006 and received the M.E., degree in Computer Science and Engineering from St.Peter's University Chennai, in 2011.Currently She is an Assistant Professor in the Department of Information Technology at Rajalakshmi institute of Technology, Chennai. Her research areas include Database Systems, Data Mining and Software Engineering.