Data Mining Techniques for CPD of Diabetes

N. Jayanthi
Research Scholar, CSE
K L University, Vijayawada

Dr. B .Vijay Babu,
Professor of CSE
K L University, Vijayawada

Dr. N. Sambasiva Rao
Principal
SRIT, Warangal

Abstract: - Disease diagnosis is very important for saving patient life. There are some diseases which are chronic for example today's major and serious health problem is diabetes. This is known as modern society disease. Even though huge medical data is available absences of disease diagnosis keep an expert to opine about the grade of disease with confidence. Medical professionals need a prediction method to diagnose diabetes. In this situation Data Mining techniques are very useful for classification, prognosis and diagnosis of a disease. Early detection of diabetes in patients helps them for prevention of the disease to some extent. This paper gives a summary of old and recent techniques used for classification, prognosis and diagnosis of diabetes.

Key-Words: - Diabetes, data mining, classification, Prediction, diagnosis

1 Introduction

Diabetes is common disease now days. Diabetes is a condition where insulin plays a key role. Insulin is a hormone produced in the pancreas, which regulates the amount of glucose in the blood. Either lack of insulin production or improper usage of insulin leads to diabetes. Basically identifying such disease in early stage is an important point in research. Data mining is a one of the solutions for this problem. Data mining aims at extracting useful information from a vast data. This information is used to predict future.

This paper is divided in to three sections. Where first section gives about classification of diabetes, second section gives details about prognosis and third section gives details about diagnosis of diabetes followed by conclusion.

2 Survey on Data Mining Techniques used in Diabetes

Anuja kumara and chitra in their work stated that diabetes is divided in to two types TYPE I and TYPE II. Type I diabetes means body fails to produce insulin. Type II diabetes means body fails to use insulin. in their paper they have used SVM with Radial basis function kernel for classification. The performance parameters such as the classification Accuracy, Sensitivity, and Specificity [1] of SVM and RBF are found and used them for classification process. The results made to opt SVM and RBF as the classifier for diabetes are given in Table 1.

Table 1
PERFORMANCE OF SVM CLASSIFIER

Dataset	Accuracy	Sensitivity	Specificity
Diabetes	78%	80%	76.5%

ROC curve for SVM is plotted between false positive rate and true positive rate and it shows a positively predicting measure for the disease which is shown in Fig 1

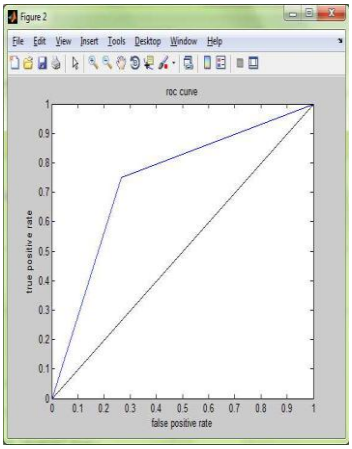

Fig. 1: ROC of SVM using Diabetes data set

Ganandeep Singh and Gurpreet Singh in their paper used simple K- Means and nearest neighbour heirarchial clustering [2] to find the effects of diabetes on the people grouped by age.

Ravi sanakal and Smt T jayakumari in their paper used two techniques for prognosis of diabetes. In this paper they used Fuzzy C Means Clustering and Support vector machine with some additions. Fuzzy C Means Clustering depends on the basic idea of K-Means with small difference. Support vector machine was used along with SMO [3]. Out of the two methods FCM produced the best result of prognosis of diabetes which is shown in Table 2

Table 2
COMPARISON OF FCM AND SMO

Result	**FCM**	**SVM**
Accuracy	94.300518%	59.5052%
Sensitivity	95.384615%	77.4%
Specificity	93.750000%	26.1194%
Positive prediction	88.571429%	-
Negative prediction	97.658537%	-

Sapna , Tamilarasi, Pravin Kumar used genetic algorithm for the for predicting diabetes[4]. They used the following algorithm for predicting diabetes.

Generic Genetic Algorithm
Procedure GA_IPD_Run
Initialize_Population (Pold)
// fills the chromosome of population Pold with 0's and 1's randomly. while termination criteria not satisfied do for each chromosome ci in Pold do Evaluate (ci, Pold) // runs chromosome ci against every member of Pold includes itself to compute fitness end
Generate_New_Population (Pnew, Pold)
// generate new population using
Pold Pold -> Pnew end , end

Algorithm for Generating New Population
Procedure Generate _ New _ Population (Pold, PNew)
PNew ->0
while Size (PNew) < Size (Pold) do
// Selection
c1 <- Select (Pold) c2 <- Select (Pold)
// Crossover
if Pc < r(.) then // return random nos. in the interval (0,1)
// Pc : Crossover Probability Crossover (c1, c2) // implements uniform crossover end // Mutation
for i = 1 to chromosome_length do
if r(.) < Pm then // Pm Mutation Probability
// Chromosome swapping each bit at the corresponding position with fixed probability usually 0.5 percent
c1i <- c1i // ith bit of the Ist chromosome
end
if r(.) < Pm then c2i <- c2i end, end
PNew -> PNew <|c1 <| c2.

Rashedur et al in their research work presented different classification techniques like Multilayer Perceptron, Bayesnet, Navie Byes, J48graft(C4.5), Fuzzy Lattice Reasoning, Jrip, Fuzzy Inference System, Adaptive Nero-Fuzzy Inference System[5] using three data mining tools such as WEKA, TANAGRA, MATLAB. They show the results of different classifiers on different tools are shown in Fig 2.

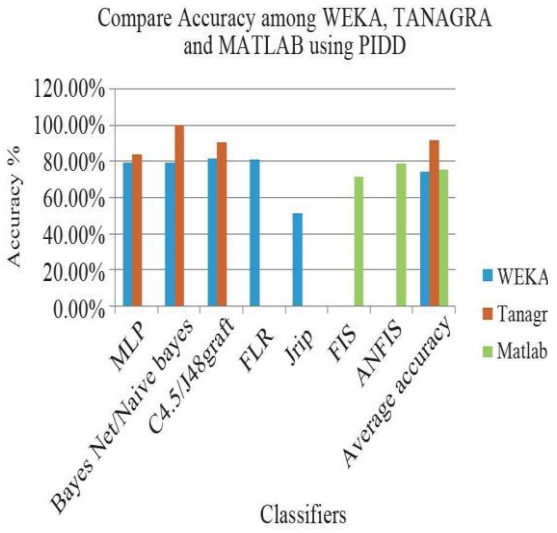

Fig 2: Accuracy of three tools using PIDD

They concluded that the best algorithm in WEKA is J48graft classifier, in TANAGRA is Naive Bayes classifier and ANIFIS in MATLAB.TANAGRA machine learning tool is best compared to other two.
Arwa Al-Rofiyee et al concentrates on predictive analysis of diabetes diagnosis using artificial neural network as a data mining technique. They used

WEKA software as a mining tool for diagnosing diabetes [6].

Table 3
RESULTS

Correct prediction%		
T.S	**TE.S**	**A.S**
87.83	68.18	81.8
93.26	64.93	74.7
97.61	65.58	73.4

Incorrect prediction%			H.L	T.T
T.S	**TE.S**	**A.S**		
12.2	31.8	18.2	50	500
6.7	35.1	25.3	20	1000
2.4	34.41	26.6	20	8000

The above result Table 3 was obtained by using data mining tool WEKA and its technique called Multi Layer Perception. They concluded that highest performance of MPL is 97.61 when the highest layer parameter is set to 20 and training time is set to 8000.

Nahal and Andrew used intelligible support vector machine for diagnosis of diabetes mellitus. In this paper they proposed a new method to overcome the drawback of support vector machine. The drawback of SVM is they are black box models.

Table 4
RULE PERFORMANCE COMPARED TO THE SVM AND DIRECT RULE LEARNERS AT EQUAL MISCLASSIFICATION COSTS

Rule extraction method	# rules/ antecedents	Accuracy %	Rule set fidelity %
SVM	N/A	0.89	N/A
SQRex-SVM	2/1.5	0.94	0.87
Eclectic	5/2.6	0.93	0.85
C5	7/2.14	0.95	N/A
CART	9/3.1	0.94	N/A
Jripper	4/2.5	0.89	N/A

In this paper they used two techniques SQRex-SVM and eclectic methods[7] in Table 4 for rule extraction and to turn SVM black box into more intelligible model.

Rajesh and Sangeetha in their research work aims for mining the relationship in diabetes data for efficient. A proposed an architecture which include feature relevance analysis [8] as shown in Fig 3

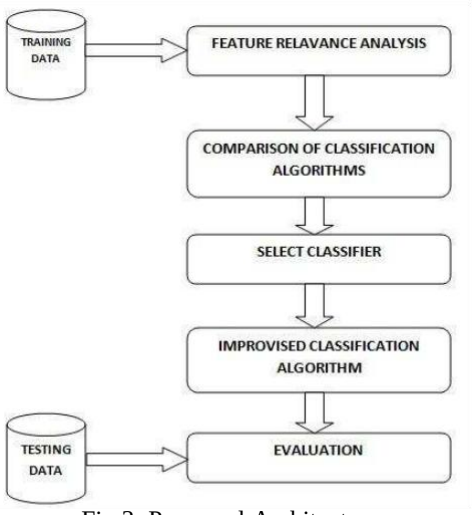

Fig 3: Proposed Architecture

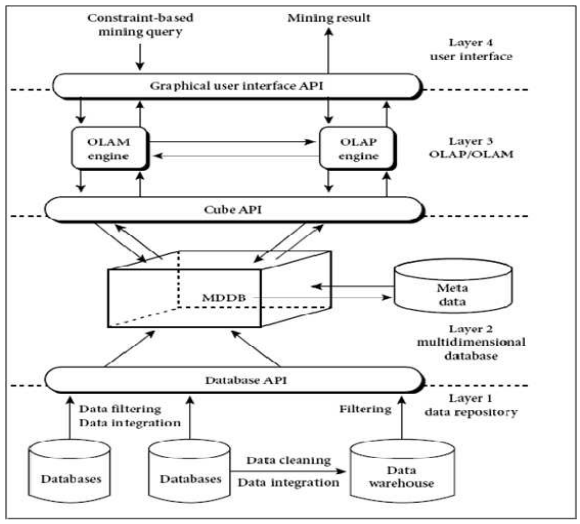

Fig 4: Architecture of Integrated model (OLAP with data mining)

Table 5
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

They used C4.5 a well known decision tree induction learning technique applied in medical data. They have shown that this algorithm can give classification rate of ~91% without feature relevance. They also compares the results of the ID3 and C4.5 decision tree algorithms [9]. They compared different classification techniques to diabetes dataset and obtained results are shown in Table 5. Their system is implemented as shown in the fig 6

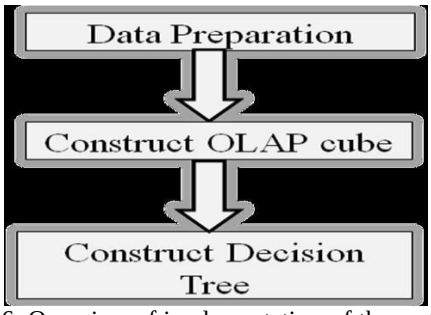

Fig 6: Overview of implementation of the system

B.L.Shivakumar and S.Alby in their research provided there are three types of diabetes TYPE I, TYPE II and GESTATIONAL DIABETES[10]. They also provided methods that have been commonly applied to diabetes data analysis and prediction o the disease. The methods they mentioned by there research are association rules, clustering K- means, cascaded method K-means with decision tree, EM algorithm, H-means clustering , genetic algorithm etc. In their paper they concluded that occurrence of diabetes is having a strong relation with deceases like wheeze, edema, oral disease, female pregnant, increase of age.

3 Conclusion

The purpose of this paper is to give a brief idea about the old and new techniques of data mining used in diabetes. Here various techniques are presented with some improvements to already existing methods. This paper gives a clear idea about types of diabetes, how to predict and how to diagnosis diabetes.

References:

[1] Anuja Kumara and Chitra "classification of diabetes disease using support vector machine", IJERA vol 3, issue 2, march-April 2013.

[2] anandeep Singh and Gurpreet Singh " Diabetes Classification Using K- Means" Apeejay Journal Of Computer Science and Applications ISSN 0974 – 5742.

[3] Ravi Sanakal and T Jayakumari "Prognosis Of Diabetes Using Data Mining Approach Fuzzy C Means Clustering and Support Vector Machine' presented at IJCTT vol 11 no 2 may 2014.

[4] Sapna, Tamilarasi, Pravin Kumar "Implementation Of Genetic Algorithm In Predicting Diabetes" IJCSI vlo 9 issue 1, no 3, January 2012.

[5] Rashedur M. Rahman, Farhana Afroz " Comparison Of Various Classification Techniques Using Different Data Mining Tools For Diabetes Diagnosis" Journal Of Software Engineering and Applications, 2013,6,85-97.

[6] Arwa Al-Rofiyee et al " using prediction methods in data mining for diabetes diagnosis".

[7] Nahal and Andrew "intelligible support vector machine for diagnosis of diabetes mellitus" presented at IEEE TRANCTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE VOL 14, NO4, JULY 2010.

[8] Rajesh and Sangeetha "application of data mining methods and techniques for diabetes diagnosis" IJEIT, vol 2, issue 3, September 2012.

[9] Rupa bagdi and pramod patil "diagnosis of diabetes using OLAP and data mining Integration" IJCSCN vol 2(3), 314-322.

[10] B.L.Shivakumar and S.Alby " A Survey On Data Mining Technologies For Prediction and Diagnosis Of Diabetes" DOI 10.1109/ICICA.2014.44