International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

VARIOUS DATA MINING TECHNIQUES ANALYSIS TO PREDICT DIABETES MELLITUS

Mr. R. Sengamuthu[1], Mrs. R. Abirami [2], Mr. D. Karthik [3]

[1,2,3]*Assistant Professor, Department of Computer Science, Govt Arts College , Ariyalur.*

Abstract - *Data mining approach helps to diagnose patient's diseases. Diabetes Mellitus is a chronic disease to affect various organs of the human body. Early prediction can save human life and can take control over the diseases. This paper explores the early prediction of diabetes using various data mining techniques. The dataset has taken 768 instances from PIMA Indian Dataset to determine the accuracy of the data mining techniques in prediction. The analysis proves that Modified J48 Classifier provide the highest accuracy than other techniques.*

Key Words: Data mining, Diabetes, Prediction, accuracy, classification

1. INTRODUCTION

Today the buzz word is "Health Care" all over the world. Early Prediction of diseases can reduce the fatal rate of human. There are very large and enormous data available in hospitals and medical related institutions. Information technology plays a vital role in Health Care. Diabetes is a chronic disease with the potential to cause a worldwide Health Care crisis. According to International Diabetes Federation 382 million people are living with diabetes world wide. By 2035, this will be doubled as 592 million. Early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors. Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data mining is a process to extract useful information from large database. It is a multidisciplinary field of computer science which involves computational process, machine learning, statistical techniques, classification, clustering and discovering patterns.

Data mining techniques has proved for early prediction of disease with higher accuracy inorder to save human life and reduce the treatment cost. This paper explores various Data mining techniques such as Navie Bayes, MLP, Bayesian Network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, Homegenity-Based, ANN, Modified J48 etc. are analyzed to predict the diabetes disease. Veena 2014 combined AmalgamKNN and ANFIS to improve the accuracy in prediction. In this K-means and KNN are combined to overcome the computational complexity of large number of dataset. And the training set is verified with fuzzy systems and neural networks to produce better result. Sapna 2012 implemented genetic algorithm with data mining techniques to test the patients affected by diabetes based upon the fitness value and the accuracy chromosome value. Gaganjot Kaur 2014 proposed a new approach for predicting the diabetes using WEKA and MATLAB for generating J48 classifiers with improved existing J48 algorithm. Murat Koklu 2013 formed a decision support system using data mining and artificial intelligence classification algorithms namely Multilayer Perceptron, Navie Bayes classification and J48 to diagnose illness. To achieve good performances in predicting the onset of diabetes, Manaswini Pradhan 2011 suggested and experimented ANN based classification model and Genetic algorithm for feature selection. Hence, this paper mainly focused on Data mining techniques and analyzed its accuracy with various tools.

Diabetes

Diabetes Mellitus (DM) is commonly referred as Diabetes; it is the condition in which the body does not properly process food for use as energy. Most of the food we eat is turned into glucose or sugar for energy. The pancreas, an organ makes a hormone called insulin to help glucose get into the cells of our bodies. When a body is affected with diabetes, it couldn't make enough insulin or couldn't use its own insulin. This causes sugar to build up into blood. Several pathogenic processes are involved in the development of diabetes. These range from autoimmune destruction of the β-cells of the pancreas with consequent insulin deficiency to abnormalities that result in resistance to insulin action. Diabetes is a life threatening disease in rural and urban, then developed and under developed countries. The common symptoms for the diabetic patients are frequent urination, increased thirst, weight loss, slow-healing in wound, giddiness, increased hunger etc. Diabetes can cause serious health complications including heart disease, blindness, kidney failure and low-extremity amputations.

A. Types of Diabetes

Type 1 Diabetes is called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes. Autoimmune, genetic, and environmental factors are involved in the development of this type of diabetes. Type1 mostly occurs in young people who are below 30 years. This type can affect children or adults, but majority of these diabetes cases were in children. In persons with type 1 diabetes, the beta cells of the pancreas, which are responsible for insulin production, are destroyed due to autoimmune system.

Type 2 Diabetes is called non-insulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes. In the type 2 diabetes, the pancreas usually produces some insulin the amount produced is not enough for the body's needs, or the

body's cells are resistant to it. Risk factors for Type 2 diabetes includes older age, obesity, family history of diabetes, prior history of gestational diabetes, impaired glucose tolerance, physical inactivity, and race/ethnicity.

Gestational Diabetes is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood glucose level. The majority of gestational diabetes patients can control their diabetes with exercise and diet. Between 10% to 20% of them will need to take some kind of blood-glucose-controlling medications. In few cases this gestational diabetes may lead to type 2 diabetes in future. It affects on 4% of all pregnant women.

Congenital Diabetes occurs in human due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, and high doses of glucocorticoids leads to steroid diabetes.

Application of Data mining Techniques in Diabetes

Medical data can be trained using data mining techniques to predict the diabetes. For this, dataset has to be preprocessed to remove noisy and fill the missing values. Pima Indian Diabetes Dataset was taken to evaluate data mining Classification. The dataset comprises 9 attributes and 768 instances. The following table 1 shows the description of the attributes.

Data mining techniques can be applied to the effective factors such as BMI, DPF, age and skin to predict the diabetes. Insulin and GTT measurement are used for testing diabetes. Pregnancy and BP are also considered as testing factors. The above attributes can be classified and cluster using various techniques such as Navie Bayes, J48, PLS-LDA, SVM,BLR, MLP, K-NN, Bayesian Network. With the help of the above attributes type 1, type 2 diabetes and gestational diabetes can be diagnosed. Obesity, age factor and family history are the main cause for type 2 diabetes. The class variable 1 indicates diabetic test is positive and 0 indicates test is negative. Tanagara, WEKA and MATLAB tools help to do data mining task with all machine learning algorithms. Data mining supervised learning algorithms are used to categorization task. DM technique can predict the hidden patterns from the previous history. Classification is the commonly used technique in medical data mining. The predictive accuracy of the classifier is estimated. The application of data mining technique can minimize the number of test required for detecting disease.

Table 1: Attributes of Diabetes Dataset

Attribute No.	Attribute	Description
1	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2	Pressure	Diastolic blood pressure(mmHg)
3	Skin	Triceps skin fold thickness(mm)
4	Insulin	2-Hour serum insulin (mu U/ml)
5	Pregnancy	Number of times pregnant
6	Mass	Body Mass Index(BMI)
7	Pedigree	Diabetes Pedigree function
8	Age	Age(in years)
9	Class	Class variable(0 or 1)

Various Data mining techniques used to predict diabetes

The diabetic patients suffer with various diseases and also it affects various parts of other organs. If the treatments are not taken to control the disease, it leads the patient to death. Hence, effective measures have to be taken to predict the disease at the earliest and control. In this paper various data mining techniques are analyzed to diagnose diabetes mellitus with the best techniques using various tools. As per the data given in the table 2, Gaganjot compared the accuracy and error rate of various data mining algorithms such as Navie Bayes, MLP, Random forest, Random Tree and Modified J48.The result provides 99.87% accuracy in Modified J48 Classifier. Radha experimented C4.5,SVM,K-NN, PNN and BLR classification techniques to classify the patients with and without diabetes. The result obtained shows that C4.5 decision tree algorithm provides 86% of accuracy.

Mohtaram Bayesian network technique predicts the diabetic patients. The Bayesian Network trained with the given data set and provides 90.4% accuracy in prediction. The genetic algorithm creates an optimal solution to predict the diseases with 80.5% of accuracy. The machine learning algorithm Multilayer Perceptron fed with training dataset and trained to classify the feature vectors. The result obtained in MLP is 97.61%. This paper deals with different data mining techniques with respect to performance of the system to predict diabetes.

Table 2 : Analysis of Different Data mining Techniques to predict diabetes

Author & year	Data mining Techniques	Tools	Accuracy	Best DM techniques
Gaganjot Kaur and Amit Chhabra, 2014 [1]	NavieBayes, MLP, Random Tree, REP tree, RAD, RandomForest, J48, Modified J48 Classifier	WEKA, MATLAB	99.87%	Modified J48 Classifier
P.Radha and Dr. B. Srinivasan, 2014	C4.5, SVM, k-	Tanagara	86%	C4.5

[2]	NN,PNN,BLR			
Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, 2014 [3]	Bayesian Network, Decision Tree	MATLAB	90.4%	Bayesian Network
Sudesh Rao and N. Arun Kumar, 2014 [4]	Genetic algorithm	Clementine	80.5%	Genetic algorithm with fuzzy logic
Veena Vijayan. V and Aswathy Ravikumar, 2014 [5]	EM, KNN, K-means, amalgam KNN and ANFIS algorithm	SharperLig ht	80%	Amalgam KNN and ANFIS
Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufadi, 2013 [6]	MLP	WEKA	97.61%	MLP
K.R. Lakshmi and S. Premkumar, 2013 [7]	C4.5, SVM, k-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, k-means & Apriori	Tanagara	76.78%	PLS-LDA
Murat Koklu and Yavuz Unal, 2013 [8]	Multilayer Perceptron, J48 and Navie Bayes Classifier	WEKA	76.3%	Navie Bayes Classifier
Rupa Bagdi, Prof. Pramod Patil, [9]	ID3 , C4.5 Decision Tree	ID3 , C4.5 Decision Tree	74%	C4.5 Decision Tree
Ashwinkumar.U.M and Dr.Anandakumar. K.R, 2012 [10]	Decision Tree and Incremental learning	WEKA	68%	C4.5
S.Sapna , Dr. Tamilarasi and M. Pravin Kumar, 2012 [11]	Genetic Algorithm	MATLAB	80%	Generic Genetic Algorithm
Manaswini Pradhan and Dr. Ranjit Kumar Sahu, 2011 [12]	Artificial Neural Network, Genetic algorithm	Tanagara	73.438 %	Artificial Neural Network
Muhammad Waqar Aslam and Asoke Kumar Nandi, 2010 [13]	Genetic Programming	GP Lab tool Box	78.5%	Genetic Programming
Huy Nguyen Anh Pham and Evangelos Triantaphyllou, 2008 [14]	Homogeneity-Based algorithm	RapidMiner	80.1%	Homogeneity-Based algorithm

Results & Discussion

The results obtained from the given dataset classified into two classes i.e patients with diabetes and without diabetes using various data mining techniques. The accuracy to predict the diabetes disease using different techniques is shown in graphical representation in the fig1. Based on the results demonstrated, Modified J48 classifier provides highest accuracy 99.87% to predict the diseases. The performance of the algorithm is calculated using the equation for Total Accuracy and Random Accuracy. Here, True positive and True Negative, False positive and False Negative parameters are taken to evaluate the equation. Radha compared classification techniques and found the C4.5 decision tree algorithm gives better accuracy 86% in prediction. Arwa Al-Rofiyee et.al used machine learning algorithm Multilayer Perceptron to predict the disease with 97.61% accuracy.

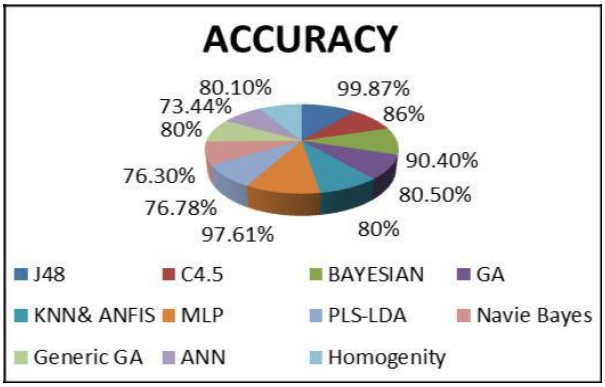

Figure 1: Performance of various data mining techniques

CONCLUSION

In the medical field accuracy in prediction of the diseases is the most important factor rather than the execution time. In the analysis of data mining techniques and tools Modified J48 Classifier gives 99.87% of highest accuracy using WEKA & MATLAB tool. Since the diabetes is a chronic disease it has to be prevented before it affects people. In future the diabetes can be prevented using gene analysis and previous history of the diabetes.

REFERENCES

[1] Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the prediction of Diabetes", International Journal of Computer Applications (0975-8887) vol.98 No.22, July 2014.

[2] P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339

[3] Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes" , MAGNT Research Report, vol.2(5), pp.892-902.

[4] Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.

[5] Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014

[6] Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufad, Dr. Mohammed Abdullah AL-Hagery, "Using Prediction Methods in Data mining for Diabetes Diagnosis", http://www.psu.edu.sa/megdam/sdma/Downloads/Po st ers

[7] K.R Lakshmi, S.Premkumar, " Utilization of Data mining Techniques for prediction of Diabetes Disease survivability", International Journal of Scientific & Engineering Research, vol.4 Issue 6, June 2013.

[8] Murat Koklu and Yauz Unal, " Analysis of a D. population of Diabetic patients Databases with Classifiers", International Journal of medical, Health,Pharmaceutical and Biomedical Engineering", vol.7 No.8, 2013.

[9] Rupa Bagdi, Prof. Pramod Patil," Diagnosis of Diabetes Using OLAP and Data Mining Integration", International Journal of Computer Science & Communication Networks,Vol 2(3), pp. 314-322.

[10] Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

[11] S. Sapna, Dr. A. Tamilarasi and M. Pravin Kumar, "Implementation of Genetic Algorithm in predicting Diabetes", International Journal of computer science, vol.9 Issue 1, No.3, January 2012.

[12] Manaswini pradhan, Dr. Ranjit kumar sahu, " predict the onset of diabetes disease using Artificial Neural Network", " International Journal of Computer Science & Emerging Technologies, vol.2 Issue 2, April 2011.

[13] Muhammad Waqar Aslam and Asoke Kumar Nandi, "Detection of Diabetes using Genetic Programming", European Signal Processing Conference (EUSIPCO-2010), ISSN 2076-1465.

[14] HuyNguyenAnhPhamandEvangelos Triantaphyllou, " Prediction of Diabetes by Employing New Data mining approach which balances Fitting and Generalization, Springer 2008