

Study and Analysis on Data Mining Technologies for Diabetes Mellitus Prediction

Zarana Gajjar¹, Purvi Prajapati²

Assistant Professor^{1,2}

GEC, Gandhinagar, Gandhinagar, India¹

CSPIT, Charusat, Changa, India²

Corresponding Authors' emails: *zarana.gajjar18593@gmail.com¹, purviprajapati.it@charusat.ac.in²*

Abstract

Data mining now-a-days plays a dynamic role in prediction of Disease. Data Mining is frequently defined as the process of determining patterns, correlations, trends or relationships by searching through a huge amount of data stored in repositories, databases, and data warehouses. Diabetes mellitus is an enduring disease and a foremost public health challenge all over. Diabetes affected over 246 million people worldwide with a common of them being women. According to the WHO report, by 2025 this number is projected to increase to over 380 million. This paper provides a survey and analysis of data mining methods that have been commonly applied to diabetes data analysis and prediction of the disease.

Keywords: *Data mining, Diabetes mellitus, Classification, C4.5, Naïve Bayes, CART, Bayes Network, Pima Indian Diabetes Data(PIDD)*

INTRODUCTION

Data mining is a relatively new concept that emerged in the middle of 1990's as a new approach to data analysis and knowledge discovery[19]. Data mining is one of the numerous applications of machine learning.

It signifies a process developed to examine large amounts of data routinely collected. Data Mining, the extraction of hidden predictive information from large Databases, is a powerful new technology with great potential to help Medical

practitioners for extracting hidden medical knowledge[5,9]. Data mining technique is used to predict possible future trends or to discover hidden patterns in the behavior of the data[1].

Diabetes mellitus has grown into a leading global public health problem in recent time. Diabetes affects when blood sugar level is increased in the body to certain level. It is a chronic disease in which body does not produce insulin or use it properly. This increase the risk of developing kidney disease, hypertension, blindness, nerve damage, blood vessel damage and contribution to heart disease or even death [8]. Diabetes disease affects more on women than on men because of their minor survival rate and lower quality of life. As per WHO reports almost 1/3 of the women who suffer from diabetes disease have no knowledge about it. The effect of diabetes is unique in case of mother because the disease is spread to their unborn children[10].

A. Types of Diabetes

There are three types of diabetes, viz. Type 1, Type 2 and Gestational [10]. Type1 diabetes, which used to be called juvenile diabetes, grows most often in young people;

however, type 1 diabetes can also develop in adults. It is usually affected by an autoimmune reaction where the body's defense system attacks the cells that produce insulin.

In this type of diabetes, the pancreatic cells that produce insulin have been damaged [6,10]. Type 2 diabetes, which used to be so-called adult-onset diabetes, can distress people at any age, even youngsters. However, type 2 diabetes progresses most often in middle-aged and older persons. People who are overweight and inactive are also more likely to develop type 2 diabetes. At this point, pancreas doesn't make the required amount of insulin [10,13]. Gestational diabetes can develop when a woman is pregnant having high glucose level. Pregnant women make hormones that can lead to insulin resistance. All women have insulin resistance late in their pregnancy. If the pancreas doesn't make enough insulin during pregnancy, a woman grows to gestational diabetes. All types of diabetes need suitable treatment and if they are identified at an early state, it can avoid the complications related with them [10,13].

B. Dataset

The Data set used in this paper is excerpted from the UCI Machine Learning Repository. The original owner of this dataset is the National Institute of Diabetes and Digestive and Kidney Diseases. Dataset is composed of 768 instances. Each patient is characterized in data set by 8 attributes and all are female at least youngest one is

21 years old to the oldest one is 81 years old of Pima Indian Heritage. All attributes have numerical values. And last attributes have diabetes prediction result either it tested_positive or tested_negative.

In below table attributes for PIDD are defined as:

Table-1: Description of the features in the PIDD

No.	Attribute Name	Description
1	Number of times pregnant	Numerical values
2	Plasma glucose concentration	Glucose concentration in a 2 hour in an oral glucose tolerance test
3	Diastolic blood pressure	In mm Hg
4	Triceps skin fold thickness	Thickness of skin in mm
5	2-Hour serum insulin	Insulin (mm U/ml)
6	Body mass index	$(\text{weight in kg} / (\text{height in m})^2)$
7	Diabetes pedigree function	A function – to evaluate the presence of diabetes
8	Age	Age in years
9	Class	1 – tested_positive 0 – tested_negative

DATA MINING METHODS AND TECHNIQUES

a. Association Rule

Association Rule mining techniques are

used to recognize and discovering relationships between variables in large database. It is more associated when new rules are searched. Association rules are commonly used in discovering the relationship between the symptom and syndrome type. Associations rule mining techniques may generate the probable causes of the particular disease such as Diabetes in the form of association rules which can be used for fast and better clinical decision-making.

Sankara Narayanan .S and Dr Pramananda Perumal. T, in paper [5]used association rule mining for Diabetes prognosis. They used Apriori and FP-Growth for application to Diabetes dataset and generate association rules by using both algorithm. They evaluated that FP-Growth have several advantages over Apriori algorithm and also observed that both algorithms generate the same number of frequent sets as a consequence same number of rules for the same known dataset under the same constraints.

B. M. Patil, in literature [18] introduced a new methodology to generate association rules on numeric data. In their research

authors first handles missing values by preprocessing techniques so it improves the quality of data and applied equal interval binning with approximate values based on medical expert's advice to PIDD. At the end rules are generated by applying apriori association rule algorithm. They also proved that the results obtained by them are very promising.

Sankara Narayanan .S and Dr Pramananda Perumal. T, in research [7] used Clinical Patient records and Pathological test reports inherently to represent data sets and applied Data Mining techniques such as Rule Classification and Decision trees. And applied Rule set classifier with the form as

IF conditions THEN conclusion

And applied this to any healthcare system as (Symptoms)(Previous-history)-> (Cause-of-disease)

And lastly they were identifying the incidence of DM in the form of If_then_rule as example:

i. IF Sex = MALE
AND Age >= 30
AND Mass >= (Height – 100)
AND (Fbgl>= 100 OR Rbgl>= 130

OR Pbg1 >= 155)

AND Bmi >= 27

THEN

Diagnosis = incidence of DMD

ii. IF Sex = FEMALE

AND Pregnant

AND Age >= 20

AND (Fbgl >= 120 OR Rbgl >= 150

OR Pbg1 >= 175)

THEN

Diagnosis = incidence of DMD in
GESTATIONAL period

b. Classification & Clustering

Classification is a data mining (machine learning) technique which is used to predict group association for data instances. It classifies data item into any one of many predefined classes. Clustering is the process of evaluating and grouping the data into different clusters or classes such that objects within the same cluster will be having more similarity to each other, but will be having different properties in other clusters. The standard objective of clustering in this area is to find different groups of Diabetes patients with similar symptoms within a group but different symptoms of other groups [1]. D. Senthil Kumar, G. Sathyadevi and S.Sivaneshat[3] defines

their intelligent medical decision support systems to help physicians using Classification algorithm like C4.5, ID3 and CART.

They experiment on 3 medical dataset that is, heart disease, hepatitis and diabetes datasets. And their aim is to compare with all define attribute selection measure algorithms. Accuracy of these algorithms is observed by producing confusion matrix using 10-fold cross validation method. At the end of research they conclude that CART gives best performance (83.2% Accuracy) among other algorithms in respect of accuracy and time complexity but run time complexity of CART is adequate with compare to C4.5.

In [16] the authors combine K-means clustering and C4.5 and developed hybrid model for classifying PIDD. In model firstly, K-means clustering is used to identify and eliminate incorrectly classified instances and in second stage a fine tuned classification is done using Decision tree C4.5 by taking the correctly clustered instance of first stage. The resultant dataset is used to train and test the diabetic dataset using 2 methods (a) partitioning by 60-40

ratio (training set) method and (b) 10 folds cross validation. 93.33% accuracy is improved by their experiment on hybrid model compared to accuracy of 73.62% using C4.5 alone.

Asma A. Aljarullah, in paper [17], used decision tree methods to predict patients with emerging diabetes. PIDD dataset is first pre-processed to improve the data quality. And then WEKA's J48 decision tree classifier was applied to the improved data set using 10-folds cross validations to construct the decision tree model. And achieve 78.17% accuracy of resulting model. In study [7], Sankarannarayanan. S and Dr Pramananda Perumal. T applying

various data mining techniques in diabetes research as rule set classifiers and decision trees to find out hidden knowledge. Using decision tree it classifies an unknown class data instances. They used some dataset having three predictor attributes as Age, Gender, Glycated Hemoglobin level and one result attributes, that is disease signifying whether the patient can have Diabetes or not. The decision tree tests the intensity of symptom value in the instance. So as per the answer; the instance is pushed down for further process through the corresponding branch and reaches leaf node, where it is classified as whether person having Diabetes or not. In Fig-1 decision tree is shown which is built from small training dataset.

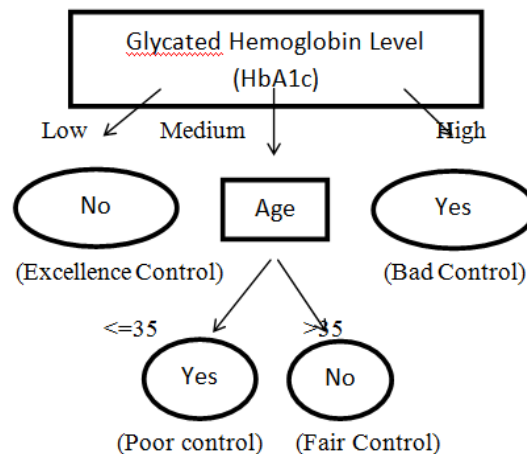


Fig-1: A Decision Tree structure build from analysed data[7]

Yang Guo, Guohua Bai and Yan Hu, in their research [8] proposed Bayes Network to predict patients with developing Type-2 diabetes. They used min-max normalization model to translate the attribute's values to 0 and 1 as a new range. Then they discretized dataset, to present the Conditional probability tables of Bayesian network. Then Naïve bayes basic principle is applied as, from training dataset of patient, marginal probabilities of symptoms $P(s_i)$ and diseases $P(d_j)$, and conditional probabilities of symptoms on all diseases $P(s_i|d_j)$ are calculated by counting frequencies in the data. For Bayesian network they proposed for diabetes they use WEKA to simulate the dataset and fill in the Conditional probability tables. Then after Leave-one-out method was used in their study to evaluate the proposed Naïve Bayes network. The accuracy of the resulting model was 72.3%.

V. Krishna Priya, A. Monika and P. Kavitha, in paper [9] applied C4.5 classification algorithm to discover decision tree model for the prediction of diabetes. They used significant attributes of dataset for prediction purpose. Android based application was developed during their study with aim of identifying whether a person has

diabetes or not, if predicted diabetes system suggest measures for maintaining normal health and if not diabetic it predicts the risk of getting diabetic.

Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, in literature [10] used two algorithms namely, J48 and Naïve Bayes, have been used to create model for diagnosis. They divide data into training set and test set by using percentage split and 10-fold cross-validation technique. As a result conclusion was both methods have a comparatively small difference in error rate through 70:30 percentage split. And also conclude that Naïve Bayes technique gives the least error rate as compared to other J48 implementations.

c. Other Techniques

Palivela Hemant and Thotadara Pushpavathi in paper [14] combines K-means clustering with several different classification methods like Naïve Bayes, SMO, Bagging, J48, Rotation Forest and Random Forest, AdaBoost M1 to predict the positive and negative of disease. Missing values are replaced globally and nominal attributes are converted into binary and then normalized. Initially K-means clustering is used to group

the disease related data into clusters and assigns classes to clusters. For training a support vector classifier, SMO implements the sequential minimal optimization algorithms. Bagging bags classifiers to reduce variance. And the best accuracy for the given dataset is achieved in bagging algorithm compared to other classifiers. Various classification algorithms are used by authors to propose an efficient hybrid model for the prediction of the positive and negative case of the diabetes disease.

C. M. Velu and K. R. Kashwan, in their researched [15] define three techniques of EM (Expectations - Maximisation) algorithm, Genetic Algorithm (GA) and H-means+ clustering. EM algorithm is used for sampling. The simulation results were

noticeable that h-means+ algorithms performed little better compared to EM. Genetic algorithm was used to optimize the chromosome was performed. The simulation results produced samples with diabetes presence in about 35 % of the population of the sample.

ANALYSIS AND DISCUSSION

From above survey we conclude accuracy of various Algorithms. We analyze that [3] CART gives best accuracy over other algorithms based on techniques they used. It also analyzes that decision tree algorithms gives best result over other data mining techniques. This analysis is done after surveying papers.

Table-2: Performance of Analysis Algorithms

Algorithms Used	Accuracy(%)
Naïve Bayes	76.3
C4.5	73.62
J48	78.17
CART	83.2
SMO	77.34

Bagging	74.6
KNN	78
BayesNet	72.3
AdaBoostM1	74.34

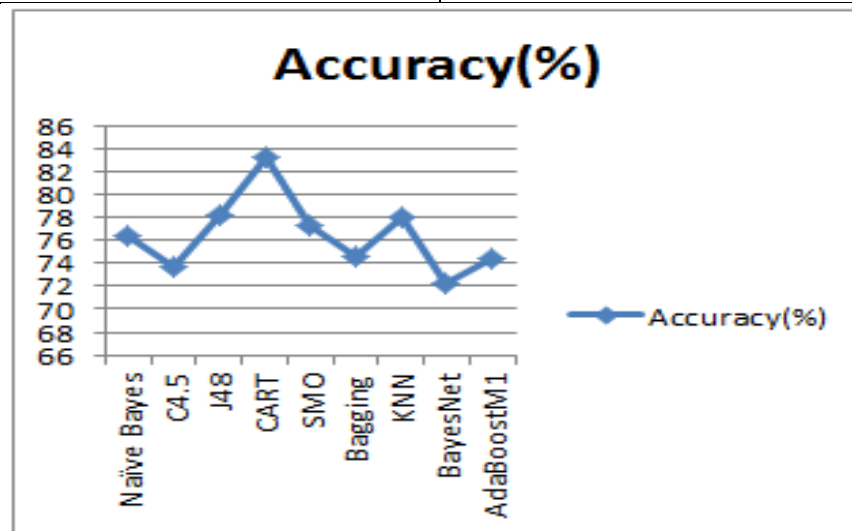


Fig-2: Predicted Accuracy

CONCLUSION

In this paper we presented major research accomplishments and techniques that immersed in the field of Diabetes Mellitus prediction. By using DM techniques the chance of Diabetes can be predicted which is helpful for timely detection of the disease. This paper has delivered the survey of data mining techniques used for Diabetes diseases they classified. It also Focus into the significance of locally frequent patterns

and the mining techniques used for the purpose.

ACKNOWLEDGMENT

I would like to thank my Co-Author Asst. Prof. PurviPrajapati who have guided and helped me throughout the survey work. I owe my deepest gratitude to PurviPrajapati, for her guidance, encouragement. I am grateful thank my institute, Government Engineering College, Gandhinagar, for

giving me this opportunity to work in the great environment.

I would also like to thank my parents, my friends and well-wishers who helped me to come out of woods whenever any difficulty was encountered. It was only due to their support, motivation and encouragement that I could steer through the project on an honest course to the splendor of success.

REFERENCES

- 1) J. Han, M. Kamber, 2000, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 1st edition
- 2) Hui-Huang Hsu, 2006, "Advanced Data Mining Technologies in Bioinformatics", Hershey, PA, USA, Idea Group
- 3) D.Senthil Kumar, G.Sathyadevi and S.Sivanesh, 2011, "Decision Support System for Medical Diagnosis Using DataMining",IJCSI International Journal of Computer Science Issues, pp. 147-153
- 4) Binal A. Thakkar, Mosin I. Hasan and Mansi A. Desai), 2010, "Health care decision support System for swine flu prediction Using naïve bayes classifier", IEEE (International Conference on Advances in Recent Technologies in Communication and Computing, pp. 101-105
- 5) Sankaranarayanan.S and DrPramanandaPerumal.T, 2014, "Diabetic prognosis through Data Mining Methods & Techniques", IEEE International Conference on Intelligent Computing Applications, pp. 162-166
- 6) Dr. B.L.Shivakumar and S. Alby, 2014, "A Survey on Data Mining Technologies for Predicion and Diagnosis of Diabetes", IEEE International Conference on Intelligent Computing Applications, 167-173
- 7) Sankaranarayanan.S and DrPramanandaPerumal.T, 2014 , "A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", IEEE World Congress on Computing and

- Communication Technologies, pp. 231-233
- 8) Yang Guo, Guohua Bai and Yan Hu, 2012, "Using Bayes Network for Prediction of Type-2 Diabetes", IEEE the 7th International Conference for Internet Technology and Secured Transactions, pp. 471-475
- 9) V. Krishna Priya, A. Monika and P. Kavitha, 2015, "Android Application to Predict and Suggest Measures for Diabetes using DM Techniques", International Journal of Computer Applications Technology and Research, pp. 291-295
- 10) Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, 2015, "Diagnosis of diabetes using Classification mining techniques", International Journal of Data Mining & Knowledge Management Process, pp. 1-14
- 11) UCI Machine Learning, "Pima Indians Diabetes Dataset", <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>
- 12) https://en.wikipedia.org/wiki/Diabetes_mellitus
- 13) <http://diabetes.niddk.nih.gov/dm/public/type1andtype2/#signs>
- 14) Palivela Hemant and Thotadara Pushpavathi, 2012, "A novel approach to predict by cascading clustering and classification", IEEE 3rd International Conference on Computing Communication & Networking Technologies, pp. 1-7
- 15) C. M. Velu and K. R. Kashwan, 2013, "Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE 3rd International Advance Computing Conference, pp. 1070-1075
- 16) Asha Gowda Karegowda, Punya V., M. A. Jayaram and A. S. Manjunath, 2012, "Rule based Classification for Diabetic patients using Cascaded K-Means and Decision Tree C4.5",

International Journal of Computer
Applications, pp. 45-50

17) Asma A. AlJarullah, 2011,
“Decision Tree Discovery for the
Diagnosis of Type II Diabetes”,
IEEE International Conference on
Innovations in Information
Technology, pp. 303-307

18) B. M. Patil, R. C. Joshi,
DurgaToshniwal, 2010, “Association
rule for classification of type-2
Diabetic Pateints”, 2nd International
Conference on Machine Learning
and Computing, pp 330-334

19) IllhoiYoo, Patricia Alafaireet,
MiroslavMarinov, Keila Pena-
Hernandez, RajithaGopidi, Jai-Fu
chang and Lei Hua, 2011, “Data
Mining in Healthcare and
Biomedicine: A Survey of the
Literature”, Springer,pp 2431-2448