International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 3 (2016) pp 1810-1814

A Survey on Diabetes Mellitus Prediction Using Machine Learning Techniques

Thiyagarajan C,
Assistant Professor, Department of Computer Science,
PSG College of Arts & Science,
Coimbatore, Tamilnadu, India.

Dr. K. Anandha Kumar,
Assistant Professor (SG), Department of Computer Applications,
Bannari Amman Institute of Technology,
Sathyamangalam, Tamilnadu, India.

Dr. A. Bharathi
Professor, Department of Information Technology,
Bannari Amman Institute of Technology,
Sathyamangalam, Tamilnadu, India.

Abstract
Diabetes Mellitus (DM) is a metabolic diseases group where the person will have high blood sugar due to the pancreas unable to produce sufficient insulin or the cell's which are not responding to the insulin produced. Diabetes is a chronic disease and a major public health challenge worldwide.The main drawback is that there is lack of awareness of the people on eating habits. In our country, diabetes patient counts have increased steadily due to this reason. So, there is an increase in interest by various researchers to set up a medical system which can screen a large number of people for life threatening disease such as cardio vascular disease, retinal disorder in diabetic patients.Several data mining and machine learning methods have been used for the diagnosis, prognosis, and management of diabetes. In this work, an effective machine learning algorithm is proposed for the classification of type DM patients. This machine learning algorithm used for classification will find the optimal hyper-plane which divides the various classes. By using this machine learning algorithm, the classification accuracy is achieved for classifying the diabetes patients.

Keywords*:* Diabetes Mellitus, Classification, Machine learning algorithm, Treatments

Introduction

Diabetes Mellitus (DM) is a collection of metabolic infections in which a human being has elevated blood sugar, either for the reason that the pancreas does not generate sufficient insulin, or because cells don't react to the insulin that is generated. This elevated blood sugar makes the conventional signs of polyuria (regular urination), polydipsia (increased need for liquids) and polyphagia (increased starvation). DM includes three major categories, "Type I DM", as a consequence of the human bodies malfunction to generate insulin, and necessitates the individual to insert insulin or carry an insulin pump. This category was previously indicated as "Insulin-Dependent Diabetes Mellitus" (IDDM). The second category of DM is recognized as "Type II DM" as a consequence of insulin confrontation, a situation in which cells are ineffective to exploit insulin appropriately, occasionally merged with an absolute insulin insufficiency. This category also called as "Non Insulin Dependent Diabetes Mellitus" (NIDDM) or "adult-onset diabetes". At last, "gestational diabetes" takes place when conceived women without an earlier

Diagnosis of diabetes increase high blood glucose intensity; it possibly will lead to development of type I DM. Gestational diabetes mellitus (GDM) is characterized by carbohydrate intolerance of varying severity with onset or first recognition during pregnancy. Women with a history of GDM are at increased risk of future diabetes, predominantly type-2 diabetes, as are their children. The extent of this risk depends on diagnostic criteria used to identify GDM.

All DM types will be common in something. Generally, our bodies will break down into carbohydrates and sugar which eats into specific sugar is known as glucose. This glucose will stimulate the cells in the body. But these cells require insulin for taking in the glucose and for energy. Every DM type will treatable because insulin was available in 1921. Types I and II are chronic conditions which are unable to cure. In type I DM, pancreas transplants are tried with the limited success, gastric bypass has been successful in several ways. The DM which is untreated will cause several problems. The complications which include diabetic ketoacidosis and non ketotichyperosmolarcoma are acute. Series long term complications include cardiocascular disease, chronic renal failure, and diabetic retinopathy. So, frequent treatment of the disease is very important in the present situation such as stopping the smoking and maintaining the healthy body weight. Because the cells will not take in the glucose, it starts to create in our blood. High blood glucose will damage the small vessels of blood in our heart, kidneys, eyes or nervous system. Due to these causes, DM will cause the heart disease, kidney disease, stroke, nerve damage and blindness. To avoid this type of causes, early detection of the DM is required with effective method. So, the main aim of this work is to develop

a classification algorithm for DM diagnosis and treatment using a machine learning approach.

Classification is one of the most important decision making techniques in many real world problem. Bassamet al (2013) build classification models and tools for diabetes, hypertension andcomorbidity using machine-learning algorithms. In this work, the main objective is to classify the data as diabetic or non diabetic and improve the classification accuracy. For many classification problem, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non diabetic data.

Literature survey

The objective ofKalaiselvi and Nasira (2015) is to examine the association of heart disease and diabetes. The main aim ofJensen*et al* (2014) was evaluating the approach of pattern classification by performance comparison with a newly introduced PCGM algorithm. A new method for Coronary Artery Disease (CAD) detection is presented by Yadav*et al* (2014) using an improved association rule mining. Nuryani (2013) aims to introduce novel computational intelligent techniques for hypoglycaemia detection. The detection is based on electrocardiographic (ECG) parameters. Diagnosing the diabetes criteria is reconsidered bySaudek*et al* (2013) and they recommended criteria of screening for finding the clinicians and patients quickly. The prediction of Gestational Diabetes Mellitus's (GDM) model is developed by Nanda *et al* (2011) from biochemical and maternal characteristics at 11 to 13 weeks gestation. Based on the maternal factors and biochemical and biophysical markers, a model is introduced by Akolekar*et al* (2011) for Pre-Eclampsia (PE) prediction. The risk questionnaire is updated by Alssema*et al* (2011) using clinically diagnosed and screens detected type 2 diabetes and considering the additional predictors. A new system was developed by Sacks *et al* (2011) for grading the evidence and strength overall quality of the recommendation. The objective of Sawaya*et al* (2011) was evaluating the more sensitive echocardiographic measurements and biomarkers could predict future cardiac dysfunction in chemotherapy-treated patients.

Karthikeyan and Vembandasamy (2015) used the association rules using MPSO-LSSVM algorithm. This is the first time classification based association rule is utilized with outlier detection method. Karthikeyan and Vembandasamy (2014) already introduced the association rule mining and enhanced FP-growth algorithm which has similar functions as of ant colony optimization and improve the mining accuracy. Faust *et al* (2012) reviewed the feature extraction algorithm from the digital fundus images. The investigation is done by Nahar*et al* (2013) where the healthy and sick factors are investigate which contribute the heart disease for females and males. For identifying these Factors, association rule mining is used and the UCI Cleveland dataset is considered. A framework is proposed by Kuo*et* al (2007) of data mining which will clusters the data and then follow mining rule. Huang et al (2012) shows that the LS-SVM and PSVM which can be simplified and a unified framework learning and other regularization algorithms can be referred to Extreme Learning Machine (ELM) which can be built.

Comparative study

Author	Description	Pros and cons
Kalaiselvi and Nasira (2015)	The objective is to examine the association of heart disease and diabetes	The Hybrid Particle Swarm Optimization and Library Support Vector Machine Algorithm was developed using categorical variables which is the main advantage.
Jensen et al (2014)	The main aim was evaluating the approach of pattern classification by performance comparison with a newly introduced PCGM algorithm	The pattern classification provides an approach for optimizing the hypoglycemic events identification in PCGM data which is the main benefit. However, the algorithm still needs to be validated on a larger number of subjects, with data including spontaneous hypoglycemic events
Yadav et al (2014)	A new method for Coronary Artery Disease (CAD) detection is presented using an improved association rule mining.	The main advantage is that the generated association rule patterns from this dataset were presented to medical experts in the field.

Nuryani (2013)	The author aims to introduce novel computational intelligent techniques for hypoglycaemia detection. The detection is based on electrocardiographic (ECG) parameters.	The main advantage is that several ECG parameters were introduced in this work for hypoglycaemia detection. These parameters show contributions to the detection of hypoglycaemia.
Saudek et al (2013)	Diagnosing the diabetes criteria is reconsidered and they recommended criteria of screening for finding the clinicians and patients quickly.	Specific criteria called HbA1care established for screening, as well as glycemic levels now defined as IFG.
Nanda et al (2011)	The prediction of Gestational Diabetes Mellitus's (GDM) model is developed from biochemical and maternal characteristics at 11 to 13 weeks gestation.	The main advantage is that First-trimester screening can be provided by a integrating of maternal characteristics and biomarkersfor GDM.
Akolekar et al (2011)	Based on the maternal factors and biochemical and biophysical markers, a model is introduced for Pre-Eclampsia (PE) prediction	Effective prediction of PEusing pre-eclampsia (PE) based on maternal factors and biophysical and biochemical markers can be achieved at 11–13 weeks' gestation is the major benefit.
Alssema et al (2011)	The risk questionnaire is updated using clinically diagnosed and screens detected type 2 diabetes and considering the additional predictors	But the disadvantage is that although especially the case for smaller datasets, external validation is needed before implementing a prediction model.
Sacks et al (2011)	A new system was developed to grade the overall quality of the evidence and the strength of the recommendations.	This work provides specific recommendations which are based on published data. Monitoring of glycemic control is performed by self-monitoring of plasma or blood glucose with meters and by laboratory analysis of HbA1c which is advantage
Sawaya et al (2011)	The objective was evaluating the more sensitive echocardiographic measurements and biomarkers could predict future cardiac dysfunction in chemotherapy-treated patients	Early echocardiographic measurements of myocardial deformation and biomarkers are predicted in patients which is the main benefit.
Karthikeyan and Vembandasamy (2015)	They used the association rules using MPSO-LSSVM algorithm. This is the first time classification based association rule is utilized with outlier detection method	One of the most notable aspects of this algorithm is the exercise of some logical operators which permit frequent items to be attained in datasets where there are not many frequent patterns. This mining of association rules permits to attain understandable close relations between items, while these rules are more functional and reasonable.
Karthikeyan and Vembandasamy (2014)	They already introduced the association rule mining and enhanced FP-growth algorithm which has similar functions as of ant colony optimization and improve the mining accuracy	Reducing the scanning of database using optimizing techniqueand improving the quality of rules which van be produced for CACO which is the major advantage.

Faust et al (2012)	They reviewed the feature extraction algorithm from the digital fundus images.	Automated DR detection can reduce the grading cost and thereby make the whole screening process less expensive
Nahar et al (2013)	The investigation is done where the healthy and sick factors are investigated which contribute the heart disease for females and males. For identifying these factors, association rule mining is used and the UCI Cleveland dataset is considered.	The major advantages are CI Heart disease Cleveland dataset used in this research. Gender specific analysis is performed. Gender specific significant factors are determined.
Kuo et al (2007)	A framework is proposed of data mining which will clusters the data and then follow mining rule.	This study has developed a method which is able to discover more useful and accurate rules from the medical database fast. This can not only let the researchers pay more attention on some important groups and find out the hidden relation in the groups easier, but also avoid the important relationship ignored in the large database.

The above table shows the pros and cons of several literatures, different authors give various ideas to identify the diabetes. But there is no effective method for identifying the diabetes. This motivates the new approach for identifying the diabetes.

Problems and directions

During the last twenty years the prevalence of diabetes has increased dramatically in many parts of the world and the disease is now a worldwide public health problem. The regular treatment for the disease is a vital important by maintaining the healthy body weight as well as controlling the blood pressure and lifestyle factors such as smoking, etc., Identifying the diabetes at early stage is an important factor because it will lead to damage the tiny blood vessels in your kidneys, heart, eyes or nervous system. So, an effective method is needed for classifying the diabetes and non diabetes patient quickly. A classification model and risk assessment tools will provide a solution for diabetes, hypertension by approach of machine learning. An effective classification algorithm is developed for DM diagnosis and treatment.

Conclusion

Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. This survey has taken various classification methods and ensemble them to give the new model in the search of finding the better result in terms of accuracy, specificity and sensitivity.The diabetes diagnosis problem is investigated in this research by the performance. Further, the works will be extended for diabetes conclusion by gathering the information from several locales across the world and provide the more precise and general prescient model. The performance can be studied by different parameters and for effective diabetes diagnosis. The work can be extended and improved for the automation of diabetes analysis.

References

[1] Akolekar, Ranjit, ArgyroSyngelaki, Rita Sarquis, Mona Zvanca, and Kypros H. Nicolaides. "Prediction of early, intermediate and late pre-eclampsia from maternal factors, biophysical and biochemical markers at 11–13 weeks."*Prenatal diagnosis* 31, no. 1 (2011): 66-74.

[2] Alssema, M., D. Vistisen, M. W. Heymans, G. Nijpels, C. Glümer, P. Z. Zimmet, J. E. Shaw et al. "The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes." *Diabetologia* 54, no. 5 (2011): 1004-1012.

[3] Farran, Bassam, Arshad Mohamed Channanath, KazemBehbehani, and Thangavel Alphonse Thanaraj. "Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study." BMJ open 3, no. 5 (2013): e002457.

[4] Faust, Oliver, Rajendra Acharya, Eddie Yin-Kwee Ng, Kwan-Hoong Ng, and Jasjit S. Suri. "Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review." Journal of medical systems 36, no. 1 (2012): 145-157.

[5] Huang, Guang-Bin, Hongming Zhou, Xiaojian Ding, and Rui Zhang. "Extreme learning machine for regression and multiclass classification." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42, no. 2 (2012): 513-529.

[6] Jensen, Morten Hasselstrøm, ZeinabMahmoudi, Toke Folke Christensen, Lise Tarnow, Edmund Seto, MetteDencker Johansen, and Ole Kristian Hejlesen. "Evaluation of an algorithm for retrospective

hypoglycemia detection using professional continuous glucose monitoring data." *Journal of diabetes science and technology* 8, no. 1 (2014): 117-122.

[7] Kalaiselvi, C., and G. M. Nasira. "Classification and Prediction of Heart Disease from Diabetes Patients using Hybrid Particle Swarm Optimization and Library Support Vector Machine Algorithm."

[8] Karthikeyan, T., and K. Vembandasamy. "A Novel Algorithm to Diagnosis Type II Diabetes Mellitus Based on Association Rule Mining Using MPSO-LSSVM with Outlier Detection Method." Indian Journal of Science and Technology 8, no. S8 (2015): 310-320.

[9] Karthikeyan, T., and K. Vembandasamy. "A Refined Continuous Ant Colony Optimization Based FP-Growth Association Rule Technique on Type 2 Diabetes." International Review on Computers and Software (IRECOS) 9, no. 8 (2014): 1476-1483.

[10] Kuo, R. J., S. Y. Lin, and C. W. Shih. "Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan." Expert Systems with Applications 33, no. 3 (2007): 794-808.

[11] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.

[12] Nanda, Surabhi, Mina Savvidou, ArgyroSyngelaki, RanjitAkolekar, and Kypros H. Nicolaides. "Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks." *Prenatal diagnosis* 31, no. 2 (2011): 135-141.

[13] Nuryani, Nuryani. "Electrocardiogram and hybrid support vector algorithms for detection of hypoglycaemia in patients with type 1 diabetes." (2013).

[14] Sacks, David B., Mark Arnold, George L. Bakris, David E. Bruns, Andrea Rita Horvath, M. Sue Kirkman, AkeLernmark, Boyd E. Metzger, and David M. Nathan. "Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus." *Diabetes care* 34, no. 6 (2011): e61-e99.

[15] Saudek, Christopher D., William H. Herman, David B. Sacks, Richard M. Bergenstal, David Edelman, and Mayer B. Davidson. "A new look at screening and diagnosing diabetes mellitus." *The Journal of Clinical Endocrinology & Metabolism* (2013).

[16] Sawaya, Heloisa, Igal A. Sebag, Juan Carlos Plana, James L. Januzzi, Bonnie Ky, Victor Cohen, SuchetaGosavi et al. "Early detection and prediction of cardiotoxicity in chemotherapy-treated patients." *The American journal of cardiology* 107, no. 9 (2011): 1375-1380.

[17] Yadav, Chetana, Shrikant Lade, and Manish K. Suman. "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining." *Age*30 (2014): 86.