DIABETES PREDICTION BY MACHINE LEARNING OVER BIG DATA FROM HEALTHCARE COMMUNITIES

Dr. R. Vijayakumar[1], Kavin Prrasad Arjunan[2], Manivel Sivasakthi[3], Karthikeyan Lakshmanan[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Vattamalaipalayam, Coimbatore, Tamil Nadu, India
[2][3][4]Student, Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Vattamalaipalayam, Coimbatore, Tamil Nadu, India

---

Abstract - *Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. Diabetes mellitus or simply diabetes is a disease caused due to the increase in level of blood glucose. Various ancient ways, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to help in making predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The main objective is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques and predict diabetes via two different supervised machine learning methods including: SVM for classification, K-means for clustering. This application propose an effective technique for earlier detection of the diabetes disease. Compared to many typical prediction algorithms, the prediction accuracy of the projected algorithmic application reaches 94.9% with a convergence speed that is quicker than that of the prevailing risk prediction algorithmic application.*

Key Words: **Classification Algorithms, Clustering, Dataset, Diabetes Mellitus, K means, SVM**

1. INTRODUCTION

Diabetes is one of deadliest diseases in the world. It is not solely a illness however conjointly a creator of various varieties of diseases like heart failure, excretory organ diseases, blindness etc. The normal distinctive method is that patients are compelled to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, on very occasion they need to induce their diagnosing report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to auto immunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects virtually 20% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other varieties of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L..

Huge amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction. With the event of massive knowledge analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics mechanically from an outsized variety of information to the accuracy of risk classification instead of the past chosen characteristics. However, those existing work mostly considered structured data. The semantics of such cases can vary from generally healthy to seriously ill or it cannot predict the disease in particular. In other words, there is no ground truth available for the "healthy" cases. If we simply treat this set of alive cases as the negative class, it would be a highly noisy majority class. On the other hand, if we take this large alive set as genuinely unlabeled, as opposed to cases with known labels removed, it would become a multi-class learning problem with large unlabeled data. Most existing classification methods on healthcare data do not consider the issue of multimodal disease prediction data

2. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) was initially stated by Vapnik, and SVM is a set of related supervised learning technique forever used in diagnosis for classification and regression. SVM at the same time minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM may be a general rule supported warranted risk bounds of applied math learning theory, thus referred to as structural risk step-down principle. SVMs can perform efficient nonlinear classification victimization what is referred to as the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick permits constructing the classifier while not expressly knowing the feature area. Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have rumored that the SVM (support vector machines) usually have capacity of delivering higher performance in terms of classification accurate than the other algorithms of data classification. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 7.

3. PROPOSED WORK

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if Diabetes Mellitus remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the increase in machine learning approaches solves this major drawback. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Classification strategies are broadly used in the medical field for classifying data into different classes according to some constraints comparatively an individual classifier. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms works better in diagnosing different diseases.

Classification is one of the most important decision making techniques in many real world problem. In this work, the main objective is to classify the data as diabetic or nondiabetic and improve the classification accuracy. For many such problems, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In several 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. The parameters like pregnancies, Glucose, Blood Pressure, skin thickness, insulin, BMI, Diabetes pedigree Function, and Age are used as input. There are many machine learning and applied mathematical techniques which can be accustomed to predict diabetes diseases. Preprocessing, noise removal and clustering is done before the classification. SVM classification techniques for classification of diabetic and non-diabetic data is been used. Thus, it is observed that techniques like Support Vector Machine, K-means clustering are most suitable for implementing this system.

A. Preprocessing and Noise Removal

The dataset that holds the glucose level of various patients are uploaded. The data may be structured or unstructured in Dataset. If dataset will be unstructured means the preprocessing takes place. In preprocessing phase each and every transactions are analyzed and determine the parameters are used in the transactions. Thus, the unstructured dataset is converted into structure dataset.

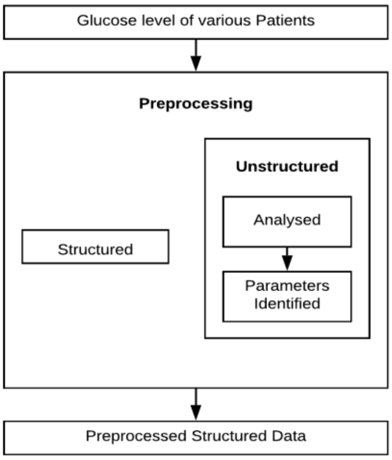


Fig 1. Preprocessing

To obtain the accuracy in predicting the diabetes the noise in the uploaded dataset are removed to rectify the unnecessary fluctuations in results. By removing the empty fields and the unwanted data from dataset, noise removal is done. This helps improving performance and achieving efficiency than the other systems.

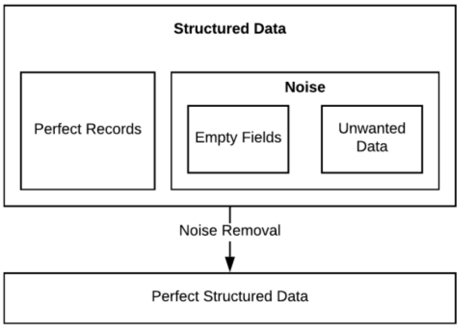
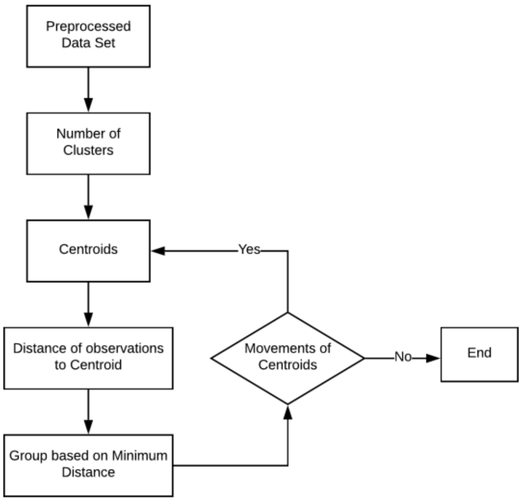


Fig 2. Noise Removal

B. Clustering

We are using K means Clustering for Cluster into three groups data, here *K*-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithmic rule is to identify groups in the given dataset, with the number of groups represented by the variable *K*. The algorithmic rule works iteratively to assign every individual data point to one of *K* groups depending on the attributes that are provided.



.

Fig 3. K means Clustering

C. SVM Classification

Given a collection of training datasets, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one or the other category, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model may be the illustration of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

Our proposed SVM algorithm can be seen as combining the advantages for solving a practical clinical problem of risk prediction from longitudinal health examination data with heterogeneity and large unlabeled data issues. To solve the problem of health risk prediction based on health examination records with heterogeneity and large unlabeled data issues.

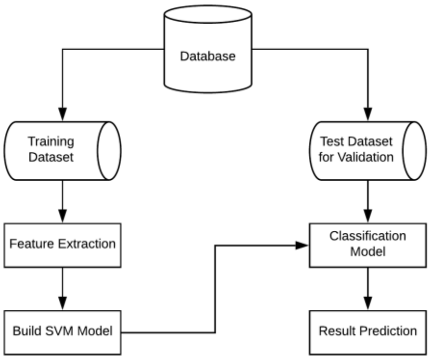


Fig 4. SVM Classification

D. System Architecture

The three processes Preprocessing and Noise Removal, Clustering and SVM Classification are combined together to make this application.

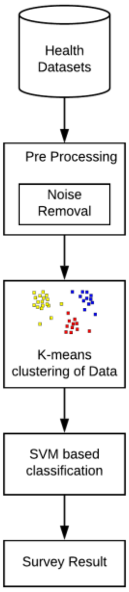


Fig 5. System Architecture

4. DESIGN

A. Input Design

Input design is the method of changing user-originated inputs to a computer-based format. Input design is one of the most valuable phases of the operation of processed system and is commonly the major drawback of a system. Input design is a part of overall style, which requires careful attribute. Inaccurate dataset are the main reason for errors in data processing. The goal of designing input data is to create data entry as simple, logical and free from errors. In the system design phase input data are collected and arranged into groups of comparable data.

B. Output Design

Output design typically refers to the results and knowledge that are generated by the system for several end-users, output is the main reason for developing the system and the basis on which they evaluate the utility of the application. Computer output is the most vital and direct source of knowledge to the user. Output design is very important phase because the output will be in an interactive manner. The report generated in this project is Diabetic Prediction.

C. Database Design

The database design is an essential one for any application developed and more important for the data store projects. Since the chatting method involves storing the message in the table and produced to the sender and receiver, proper handling of the table is a must. In the project, login table is designed to be unique in accepting the username and the length of the username and password should be greater than zero. The different users view the data in various format according to the privileges given.

5. DRAWBACKS OF EXISTING SYSTEM

•Identifying participants at risk based on current and past medical history is important for early warning and preventive intervention. By “risk”, we tend to mean unwanted outcomes like mortality and morbidity.

•It does not contain dead dataset for evaluation.

•Most existing classification methods on healthcare data do not consider the issue of multimodal data. They either have expert-defined low-risk or control classes.

6. ADVANTAGES OF PROPOSED SYSTEM

•It extracts features automatically for structured and unstructured data set.

•It provides accurate possible risk predictions.

•It combines the structured and unstructured data in healthcare field to assess the possibility in occurrence of diabetes.

•It handle’s a challenging multimodal classification problem.

7. EXPERIMENTAL RESULTS

The dataset consists around 800 user’s record which are has been considered for the analysis and that dataset is taken from kaggle. Totally 9 attributes are take place in the dataset. The attribute of the dataset are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age, outcome. These attributes are used to classify and predict the diabetes. And the SVM data mining algorithm is used for both classification and prediction.

The language chosen to design this system is C# the forms and the database connectivity everything is developed using C#. This Application is developed using .Net Framework 4.6.2 so it is purely a windows desktop application and it is not platform independent software.

8. CONCLUSION

Machine learning has the great feature to revolutionize the Diabetes Mellitus risk prediction with the assistance of advanced computational ways and availability of enormous amount of medical specialty and genetic diabetes risk dataset. Detection of Diabetes Mellitus in its early stages is the key for treatment. This work has detailed machine learning approach to predicting diabetes levels.

9. FUTURE ENHANCEMENTS

This technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status. In this study, a systematic effort was made to identify and review machine learning and data mining approaches applied on DM research. DM is rapidly emerging as one of the greatest global health challenges of the 21st century. To date, there is a significant work carried out in almost all aspects of DM research and especially biomarker identification and prediction-diagnosis. The advent of biotechnology, with the vast amount of data produced, along with the increasing amount of EHRs is expected to give rise to further in-depth exploration toward diagnosis, etiopathophysiology and treatment of DM through employment of machine learning and data mining techniques in enriched datasets that include clinical and biological information.

REFERENCES

[1]. Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction

[2]. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering.

[3]. Alan Siper, Roger Farley and Craig Lombardo, "Machine Learning and Data Mining Methods in Diabetes Research", Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.

[4]. Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[5]. Berry, Michael, and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997

[6]. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[7]. Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform." Knowledge-Based Systems 37 (2013): 274-282.

[8]. Agrawal R. and Srikant. R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Databases, 478-499, 1994.

[9]. Kavakiotis I., Tzanis G., Vlahavas I. (2014a) "Mining Frequent Patterns and Association Rules from Biological Data" Biological Knowledge Discovery Handbook: Preprocessing, Mining and Post processing of Biological Data, M. Elloumi, A. Y. Zomaya (Eds.), Wiley Book Series on Bioinformatics: Computational Techniques and Engineering, Wiley-Blackwell, John Wiley & Sons Ltd., New Jersey, USA (Publish.) (2014)