International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

SURVEY ON PREDICTION OF DIABETES USING CLASSIFICATION ALGORITHMS

Pawan Toralkar[1], Nagaraj Vernekar[2]

[1]*Student, Computer Science and Engineering, Goa College of Engineering, Goa, India*
[2]*Professor, Computer Science and Engineering, Goa College of Engineering, Goa, India*

Abstract - *The potential of Data mining methods can be used to benefit predictions on medical data. The focus of this research paper is to evaluate various data mining methods used in prediction of diabetes. Diabetes* mellitus, *commonly known as diabetes is a group of diseases which results in high sugar level in the blood which may have a drastic effect such macro vascular and micro vascular complications. Diabetes diagnosed by traditional method such as physical and chemical test may result in inaccurate outputs. To overcome this limitation we make the prediction of disease using a different Data Mining algorithm.*

Key Words: Diabetes mellitus, Normalization, Clustering, classification, K-mean, Decision Tree, SVM.

1. INTRODUCTION

Diabetes mellitus also called as diabetes, which is derived from the Greek word diabetes, which means siphon i.e. to pass through and the Latin word mellitus which means sweet. Excess sugar is found in the blood as well as the urine, hence the name diabetes. In humans, diabetes occurs when blood glucose, also called as blood sugar, is too high. The main source of energy, humans get from the glucose rich food. The pancreas produces insulin that helps glucose from food to get into cells to be used as energy. The most common types of diabetes are type 1, type 2, and gestational diabetes. Type 1 diabetes is incapable of insulin production in the body. The immune system attacks and destroys the cells in the pancreas that make insulin. Children and young adults are commonly affected by Type 1 diabetes, although it can appear at any age. People having type 1 diabetes need to take insulin every day in order to have a longer life span. In type 2 diabetes, the body does not make or use insulin well. Type 2 diabetes can be encountered at any age, even during childhood. Most commonly occurs in middle-aged and older people. Type 2 is the most common type of diabetes. During pregnancy, Gestational diabetes develops in some women. In Most of the cases, once the baby is born this type of diabetes gets cured. There is a higher possibility that if gestational diabetes is present, then there is a greater chance of developing type2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is actually type 2 diabetes. Over the age of 65 diabetes affects 1 in 4 people. About 90-95 percent of cases in adults are type 2 diabetes.

Data mining is characterized as a process for extractingusable data from any larger set of raw data. Essentially, the information gathered from Data Mining helps predict hidden patterns, future trends and behaviors and allow individuals to take decisions. Technically, data mining is the analytical mechanism through which data are analyzed from different perspectives, dimensions, and angles and categorized / summarized into meaningful information. Due to the increased amount of daily data entries, the use of clustering and classification in Data Mining is becoming increasingly common and therefore it requires an efficient framework to manage and coordinate it. A variety of clustering algorithm combinations with classification algorithms plays a vital role in this field. Classification and clustering is used to categorize objects into one or more classes, based on the characteristic. The method of classifying the input instances based on their corresponding class labels is known as classification, while clustering the instances based on their similarity is known as clustering without the aid of class labels. Widely used clustering algorithm is k-means algorithm and widely used classification algorithms are Naïve Bayes and Support Vector Machine.

2. LITERATURE REVIEW

Data mining is applied to medical fields where it effectively and reliably identifies secret patterns from datasets. Such patterns can then be used for the diagnosis and treatment of diseases. Following research papers, the emphasis will be on using various clustering and classification algorithms to determine whether or not a person is diabetic

2.1 Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis

T. Santhanam et al [1] proposed a method using K-means, Genetic Algorithms and SVM to achieve higher accuracy for diabetes diagnosis. Data cleaning was achieved by substituting the missing values for the mean. K-Means is used to eliminate noisy data and genetic algorithms to find the optimal set of features for classification with Support Vector Machine (SVM). Working principle of proposed system shown in Fig. 2.1. During each run, GA selected different attributes from the original set of attributes and the classification accuracy was recorded. To achieve consistent result, the experiment was repeated 50 times and the outcomes were listed in table. The experimental result

shows that, for Pima Indians diabetes in the UCI archive, the proposed model achieves an average accuracy of 98.79 percent. The proposed method also showed that better results were obtained in comparison to the updated clustered K-means data preparation system based on SVM (96.71%).

Outcome of this research are

- Out of 768 instances, K-Means selected 511 samples as correctly classified and 257 samples were detected as outliers. The outlier detection percentage is 33.46.
- The minimum no. of attributes selected using GA is 3 and maximum is 6.
- The minimum and maximum classification accuracy using SVM is 98.43% and 99.21% and the average accuracy is 98.79 %.

2.2 An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles

Saba et al [2] proposed use of multiple ensemble assessment techniques for diabetes datasets. Two diabetes datasets are adapted from the UCI and BioStat databases. Three types of decision trees ID3, C4.5 and CART are used as basic classifiers. As the base classifiers, three decision trees with specific splitting criteria, i.e. information gain, gain ratio and gini index, are used. Entropy is a measure that is used to divide the instances into subsets. It calculates the homogeneity for a given dataset. If it is completely homogeneous, entropy will be zero. Otherwise it is equally divided to have entropy value one. The information gain is based on decrease in entropy. An attribute which has maximum entropy will return highest information gain value. Therefore, a highest information gain attribute will be selected as the splitting attribute. A decision tree will then be constructed based on finding the most homogeneous branch. Decision tree based on Gain ratio(C4.5),It is also commonly known as C4.5. It is a fraction between information gain and its splitting information. It is generally used to reduce the effect of biasness that may occur due to large number of values for a given attribute. A decision tree based on gain ratio outperforms information gain in terms of both accuracy measure and handling complex problems. Decision tree based on gini index (CART), It measures the level of impurity for given data and constructs a binary tree where each internal node outputs exactly two classes for a given attribute. Gini index is calculated for each attribute and then the attribute with lowest gini index is selected as the splitting attribute. The tree is constructed by recursively selecting the attribute with lowest gini index. The basic idea behind ensemble classifiers is to weigh several individual classifiers and then combine them to obtain the result which outperforms every individual classifier. The classification performance and prediction accuracy of ensemble classifier is higher than single classifiers. The ensemble techniques used are Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging. Majority voting is used to classify the unlabeled instances based on the highest number of votes (high frequency vote). This technique is also termed as plurality voting (PV). AdaBoost is a popular ensemble algorithm introduced by Yoav Freund and Robert Schapire and it performs boosting by iterative processing. It focuses on the instances that are difficult to classify using other classification techniques. The level of focus depends on the weight that is associated with instances during each iteration. At start, all instances are assigned equal weight. In each iteration, the weight of misclassified instances is increased whereas weights are reduced for the instances which are correctly classified. Moreover, each classified has an associated weight which measures the overall accuracy of that individual classifier. The classifiers are then combined considering their weight and prediction class. Stacking is an ensemble technique which has achieved greater generalization accuracy. Stacking technique is based on single classifiers and it determines which classifier is reliable and which classifier is unreliable. The idea behind stacking approach is to construct a meta-dataset by taking some instances from original dataset. The predictions made by each individual classifier are used as input attributes instead of original dataset. Bayesian boosting ensemble technique is used to make an ensemble classifier that eventually performs Boolean classification for target attributes. It is an iterative process where weights are assigned and updated in every iteration and then sampling is carried out. The base classifier such as decision tree is applied multiple times sequentially, and then these classifiers are combined into a single coherent classifier. Bagging Also termed as Bootstrap Aggregation. It is most common and well known ensemble method for data classification and prediction. Experimental results and assessment indicate that the technique of the bagging ensemble shows better performance than both individual and other ensemble techniques. Other disease datasets, such as breast cancer, heart disease and liver disease can be exploited with ensemble techniques.

2.3 Predictive Analysis of Diabetes using J48 Algorithm of classification Techniques

Pradeep et al [3] suggested J48 for the diagnosis and prevision of diabetes, developed by Ross Quinlan. J48 can handle non-linear tasks and capitalize on the output of knowledge. It takes less time for training and can handle missing values in diabetes data. UCI dataset has twelve training file. Each file consists of ten attributes and five hundred instances. The dataset is given as input to feature selection. During feature selection process five attributes were selected out of ten attribute. The five attributes were Date, Time, Insulin dose, Glucose level and class variable. The model is trained and evaluated using the feature selected so that the class of invisibly unseen input can be predicted. System architecture consisted of four stages User input, Diagnosis, Classification, Analysis and prediction. The user input is the stage in which user input such as blood glucose is obtained. Diagnosis consists of two stages. They are

feature extraction and supervised learning. In feature extraction process five attributes were selected and in supervised learning process, the model was trained using the dataset and then in classification process was used to classify the unseen records as already explained in block diagram. In Analysis and prediction process, the outcome obtained from classification algorithm is the predicted value for the unseen record. Analysis is trying to find out the number of correctly classified input. The samples collected at 4am- 5am has the highest precision with 76%. The objective of the paper was predicting Blood Glucose level in advance that would facilitate a better diagnosis of diabetes was successfully achieved.

2.4 Application of Machine Learning in Disease Prediction

Pahulpreet et al [4] is focusing the use of machine learning on three different diseases. The diseases are Breast cancer, which is a very common disease among women, heart diseases, which are the leading cause of deaths in the US, and diabetes, in which blood glucose or blood sugar levels are too high. The datasets can be downloaded from UCI machine learning library. The proposed method consists of four steps. They are Exploration of dataset, Data Munging, Feature Selection and Model fitting and testing. The first step involves exploration of dataset in the python environment. In Data Munging process it tries to find out if there is missing value. If missing values are present then values are replaced by mean value in case of continuous value otherwise mode value in case of categorical value. Feature selection is done to take care of multi-collinearity, remove any redundant features that are highly correlated with each other, therefore, improving the model's performance. Backward selection method is used for feature selection. The attributes with the p-value greater than 0.05 were deleted and the model was refitted with the remaining variables. This process was iterated multiple times until every existing variable for the model was at a significant level. The selected features were given as input to classification algorithm. Classification algorithms are Logistic regression, decision making tree, random forest, vector support and adaptive boosting are classification algorithms. Dataset was divided into training set and testing set i.e. 90 percent for training and 10 percent for testing. Prediction accuracy reaches 87.1 percent by logistic regression of the proposed method in the detection of heart disease, 85.71 percent in the prediction of diabetes through the use of the Help Vectors Model (linear kernels), and 98.57 percent in Breast Cancer Classifier with the use AdaBoost.

2.5 A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree

Wenqian et al [5] proposed a hybrid prediction model to aid diagnosis of Type 2 diabetes. The proposed method consists of four steps i.e. data pre-processing, data reduction, classification using decision tree algorithm and performance evaluation. Pima Indians Diabetes Dataset from the UCI Machine Learning Repository to obtain the experimental test was used. Dataset consist of missing and impossible value. With help of data pre-processing, missing and impossible values were replaced by mean value. Data reduction was done using k-means. Dataset had 768 instances. After using k-means the instances that didn't belong to cluster were removed i.e. 236 instances and dataset was reduced to 532 instances and were given as input to classification algorithm. Finally, with 10-fold cross validation, the J48 decision tree algorithm was used on the data set. Performance evaluation involved calculation of accuracy, sensitiveness and specificity, true positive , true negative, false positive and false negative. The accuracy, sensitiveness and specificity of the proposed model were 90.4%, 87.27% and 91.28% respectively according to the confusion matrix.

2.6 Analysing Feature Importance for Diabetes Prediction using Machine Learning

Debadri et al [6] proposed the study of the diabetes prediction feature of the data set. The features used for making predictions are the most significant part of an algorithm, and some features play an extremely significant role in the prediction. The dataset was divided into train and test set in ratio of 67% train and 33% test. The correlation between each attribute was found to verify the presence of strongly correlated characteristics and thresholds was set to 0.7. Diabetes, BMI, Age, Insulin and Glucose were the essential features extracted from correlation plots. As input for three classification algorithms i.e. logistic regression, SVM and random forest, the extracted features were given. Random Forest is the best algorithm for diabetes prediction, that gives approximately 84% accuracy.

2.7 Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree

Purushottam et al [7] designed a system capable of discovering the rules effectively to predict the patients' risk level based on the given health parameter. Results show that

C4.5 can identify diabetes mellitus in the first fold correctly up to 81.27%. Results of C4.5 and the partial tree evaluation indicate that C4.5 classifier should be used in Pima data set for successful diabetes prediction

3. CONCLUSIONS

Diabetes mellitus (DM), is a series of metabolic disorders in human body due to high level of blood sugar in body. If left untreated, it can cause many complications in the long run. Diabetes is majorly caused due to dis-functioning of pancreas leading to failure in production of required insulin. From the above survey its understood that it is always better to use more than one classifier for predicting the output as its accuracy is higher than accuracy of single classifier. Bagging ensemble classifier should be used in the predicting System so that can obtain better and efficient results.

REFERENCES

[1] T. Santhanam a, M.S Padmavathi b. "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis". Procedia Computer Science, vol.47, pp.76-83, 2015

[2] Saba Bashir, Usman Qamar, Farhan Hassan Khan, M.Younus Javed "An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles". 12th International Conference on Frontiers of Information Technology, DOI 10.1109/FIT.2014.50, 2014.

[3] Pradeep K R, Dr Naveen N C. "Predictive Analysis of Diabetes using J48 Algorithm of classification Techniques" ,978-1-5090-52561/16/$31.00 c_2016 IEEE.

[4] Pahulpreet Singh Kohli, Shriya Arora "Application of Machine Learning in Disease Prediction", 4th International Conference on Computing Communication and Automation (ICCCA) 2018.

[5] Wenqian Chen, Shuyu Chen, Hancui Zhang and Tianshu Wu "A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree", 978-1-5-977/1$31.00©201IEEE.

[6] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh "Analysing Feature Importances for Diabetes Prediction using Machine Learning", 978-1-5386-7266-2/18/$31.00 ©2018 IEEE

[7] Purushottam, Dr. Kanak Saxena, Richa Sharma, "Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree", 978-1-4673-72312/15/$31.00©2015 IEEE.