Survey on Different Techniques to Predict Diabetes

Abhijit Anil Khadiye[1], Arbaz Ayub Khan[2]

[1]P.G STUDENT, MASTERS IN COMPUTER APPLICATIONS, ASM IMCOST, THANE, MAHARASHTRA, INDIA
[2]P.G STUDENT, MASTERS IN COMPUTER APPLICATIONS, ASM IMCOST, THANE, MAHARASHTRA, INDIA

---***---

Abstract - *Diabetes is one of the foremost common diseases worldwide where a cure isn't found for it yet. Thus the foremost important issue is that the prediction to be very accurate and to use a reliable method for that. Prevention and prediction of diabetes are increasingly gaining interest in the healthcare community. Although several clinical decision support systems have been proposed that incorporate several prediction techniques for diabetes prediction. This paper aims at finding solutions to diagnose the disease by analysing /survey on different techniques to predict diabetes. The research hopes to give more insight into different techniques of diagnosing the disease, leading to the timely treatment of the patients. Due to development in technologies, it is now easy to predict the glucose level in the blood, and multiple machine learning techniques are used to predict the diabetes-like Artificial Neural Network (ANN), classification techniques, and data mining techniques. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, resulting in the timely treatment of the patients.*

***Key Words*: Artificial Neural Network, Scale Vector Machine (SVM), Naive Byes, Random Forest, Decision tree, Diabetes, Prediction**

1. INTRODUCTION

Diabetes is a long-lasting disease that happens when the pancreas fails to create insulin, insulin is a hormone that controls sugar level in the blood. Diabetes causes severe damage to many organs, nerves, and blood vessels in a long span of time. Due to the lack of diagnosis of symptoms in patients for a long time may even threaten the life of the patient. Therefore, many studies have been done in the field of predicting for several diseases.

Multiple machine learning techniques are used to predict diabetes-like Artificial Neural networks (ANN), classification techniques, and data mining techniques. Artificial neural network (ANN) is a computational model that consists of several processing elements that receive inputs and deliver outputs based on their predefined activation functions. Artificial Neural Network (ANN) with a vector of features then we use it to predict the future value of blood glucose (after the 30 minutes from the current value). Classification algorithms are an efficient and widely used technique in various applications, such as medical diagnosis of diabetes patients. There are various techniques implemented for the classification of diabetes patients, such as Support Vector Machine (SVM) and Naive Byes, etc. SVM is a supervised machine learning algorithm that can be used for classification problems. It plots each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes to predict the variation in the range value. Naïve Bayes is a well-known type of classification problem, i.e., of programs that assign a class from a predefined set to an object or case under consideration based on the values of descriptive attributes. In a probabilistic approach, i.e., they try to compute conditional class probabilities and then predict the most probable class. Data mining can be called sifting through very large amounts of data for useful information. Some of the most important and popular data mining techniques are Random Forest and Decision tree. RF is a multifunctional machine learning technique it can perform the tasks of prediction and regression. Decision tree process of classification instances. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature and class space.

The main objective of this paper is to analyze the multiple techniques to predict diabetes in a more efficient and accurate way.

2. Techniques Used

2.1 Data Mining Techniques:

Data processing is the process of extracting hidden knowledge from large volumes of data. The knowledge must be new, not obvious, and one must be ready to use it. data processing has been defined as “the nontrivial extraction of previously unknown, implicit, and potentially useful information from data. it's “the science of extracting useful information from large databases”.

2.1.1. Decision Tree

A decision tree may be a basic classification and regression method, Decision tree uses tree structure and therefore the tree begins with one node representing the training samples. Value of the present decision node attribute, the training samples are divided into several subsets, each of which forms a branch, and there are several values that form several branches. the standard algorithms of the choice tree are ID3, C4.5, CART, and so on. The J48 decision tree in WEKA. J48 another name is C4.8, which is an upgrade of C4.5. J48 may be a top-down, recursive divide, and conquer strategy. This technique selects an attribute to be a basic node, generates a branch for every attribute value, divides the instance into multiple subsets, and every subset corresponds to a branch of the basis node then repeats the method on each branch.

The decision tree may be a supervised approach that uses a group of if-then rules to classify samples into categories of interest. The algorithm finds the foremost important experimental variable and sets it because the root node, which is followed by bifurcating to subsequent best variables. The tree flows during a top-down manner from the basis node through the interior nodes (the independent variables) and eventually to the terminal leaf nodes (the class prediction). the arrogance factor was set to 25% because it had been demonstrated to figure reasonably well. Rigorous assessment of the predictive performance was made by separating the info set into three sets: training set, 10-fold cross-validation set, and external validation set. the primary 2 sets, comprising of 90% of the info set, were used for assessing the interior performance, while the last set, consisting of 10% of the info set, was used for assessing the external performance. Statistical parameters to gauge the predictive performance of the choice tree include the sensitivity, specificity, accuracy, positive predictive value and negative predictive value

2.1.2. Random Forest

RF may be a classification by using many decision trees. This algorithm proposed by Breiman. RF may be a multifunctional machine learning method. It can perform the tasks of prediction and regression. Additionally, RF is predicated on Bagging and it plays a crucial role in ensemble machine learning.

The training data set of 4,948 records, contains 18 input predictors, was wont to model the choice Tree, Bagging with Decision Tree-based classification, and typical Random Forest. It is often seen that Random Forest with 18 attributes yielded the simplest accuracy among the three classification models. However, Random Forest utilizes the ensemble method by randomly selecting subsets of attributes to create decision trees. Thus, all 18 input predictors would have an equal chance to be in each predictor. Sometimes, input attributes could also be irrelevant features, defined as those features not having any influence on the response classes. Therefore, we further analyzed the info set using feature selection algorithms to get rid of some irrelevant predictors from these 18 attributes. The Gain Ratio feature selection algorithms were utilized in this paper. The ranking information was wont to model our proposed Random Forest method with Feature Selection because the following steps illustrate.

(a) Rank all variables consistent with a gain ratio ranking

(b) for every time (backward elimination), Remove the last feature from the training data set Rebuild the Random Forest model using only the remaining features.

(c) Select the subset which maximizes prediction.

The classification results show that the Random Forest gave better results for the tiny number of attributes. From the results, the simplest percentage accuracy was (94.743%) by using the primary 14 ranked attributes because the input attributes.

2.2 Classification Techniques:

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier.

2.2.1. Support Vector Machine (SVM)

SVM Model Generation SVM may be a set of related supervised learning methods utilized in diagnosis for classification and regression. SVM simultaneously minimize the classification error and maximize the geometric margin. So SVM is named Maximum Margin Classifiers.

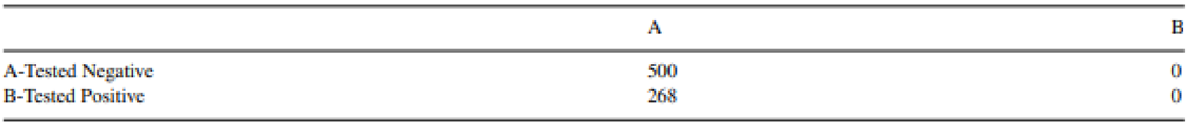
SVM is one of the quality set of supervised machine learning model employed in classification. Given a two-class training sample, the aim of a support vector machine is to seek out the simplest highest-margin separating hyperplane between the 2 classes. For better generalization hyperplane shouldn't lie closer to the info points belong to the opposite class. Hyperplane should be selected which is way from the info points from each category. The nearest point to the margin of the classifier is the support vectors. The Accuracy of the experiment is evaluated using the WEKA interface.

The SVM finds the optimal separating hyperplane by maximizing the space between the 2 decision boundaries. Mathematically, we'll maximize the space between the hyperplane which is defined by wT x + b = −1 and therefore the hyperplane defined by wT x + b = 1

This distance is adequate to 2 w. this suggests we would like to unravel max 2 w. Equivalently we would like min w | 2. The SVM should also correctly classify all x(i), which suggests yi (wT xi +

b) >= 1, ∀i ∈ {1, ¢¢, N}. The evaluated performance of the SVM algorithm for prediction of Diabetes using Confusion Matrix is as follows:

Table 1. Confusion Matrix of SVM



	A	B
A-Tested Negative	500	0
B-Tested Positive	268	0

2.2.2. Naive Byes

Naive Bayes may be a classification algorithm, which uses Bayes theorem of probability for prediction of unknown class. It uses probability to make a decision which class a test point belongs to. Naive Bayes may be a purely statistical model. This algorithm is named Naive thanks to the idea that the features/ attributes within the datasets are mutually independent.

The Bayesian classifier is predicated on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the opposite attributes. This assumption is named class conditional independence. it's made to simplify the computation involved and, during this sense, is taken into account ”naive.

” Let X = {x1, x2, xn} be a sample, whose components represent values made on a group of n attributes. In Bayesian terms, X is taken into account ”evidence.” Let H be some hypothesis, like that the info X belongs to a selected class C. we've to work out P(H|X), the probability that the hypothesis H holds given the ”evidence,” (i.e. the observed data sample X). consistent with Bayes' theorem, the probability that we would like to compute P(H|X) are often expressed regarding probabilities P(H), P(X|H), and P(X) as P(H|X) = P(X|H) P(H) / P(X)

2.3. Artificial Neural Network [ANN]:

An ANN may be a mathematical representation of the human neural architecture, reflecting its “learning” and “generalization” abilities. For this reason, ANNs belong to the sector of AI. ANNs are widely applied in research because they will model highly non-linear systems during which the connection among the variables is unknown or very complex.

An ANN is constituted by an input layer, an output layer, and a few hidden layers. during this work, the activation function of neurons within the input layer functions while the activation function of output neuron may be a linear function. The network inputs are the present measurement G(t) and therefore the previous blood sugar levels G(t-N*PH). The weights and bias of the network are initialized randomly and updated consistently with the training algorithm Levenberg-Marquardt. The error of actual and predicted blood sugar was back propagated to each layer of the NN, and therefore the optimal weights for minimum error are determined. a neural network contains an input layer with N neurons, a hidden layer with NH neurons, and an output layer with one neuron. ANN is to use the N previous measures to predict subsequent measures. Then, the anticipated measure is going to be used as an input with the previous N- 1 measures to forecast subsequent value, and so on. As a plus, the estimation of subsequent values is incremental, adaptive, and nonlinear because of the activation function nonlinearity of everyneuron

To evaluate the prediction performances, we used the following indices:

1) The Root Mean Square Error (RMSE) that allows the error to be of the same magnitude as the quantity being predicted [15]. The RMSE is expressed as:

$$RMSE = \sqrt{\frac{}{n} \sum_{1}^{n} X_i - X_i \ )^2}$$

2) The Sum of Squares of the Glucose Prediction Error (SSGPE) that measures the discrepancy among the actual data and the predicted data .The SSGPE is expressedas:

$$SSGPE = \sqrt{\frac{\sum_{1}^{n} (X_i - \hat{X}_i)^2}{\sum_{1}^{n} X_i^2}}$$

3) The Relative Error Analysis (e) which gives an indication of how good a measurement is rel- ative to the size of data measured [17]. The expression of “e” is defined asfollows:

$$e = \frac{1 \overset{n}{©} \frac{X - X_i}{X_i}}{n} *100$$

Blood glucose prediction of 12 patients. For the whole database, the average of RMSE (mg/dL) is 6.43, the average of SSGPE is 5.09% and the average of e is 3.71. The evaluated performance RMSE for T1D patient was ranged from 1.14mg/dL to 8.83mg/dL, and SSGPE was ranged from 1.24 to 7.89 and e ranged from0.7 to 5.33.

Likewise, ANN modeling performances were high in patient 1 and patient 3 and low in patients 7 and 10. This is often thanks to the important non-stationarity of those two cases, which is explained by the doctor as an uncontrolled consumption of sugary foods. In fact, patients 7 and 10 don't follow doctor instructions. Thus, it's very difficult to model their blood sugar evolution: it's stochastic and follows no mathematical law.

3. CONCLUSIONS AND FUTURE SCOPE

In this survey paper summarizing the different techniques are used in the field of medical prediction of diabetes are discussed. An artificial neural network was used to predict diabetes. Using artificial neural networks model we can design and implement complex medical processes using the software. The software systems are more effective and efficient in various medical fields including predicting, diagnosing, treating, and helping the surgeons, physicians, and the general population. These systems can be implemented in a parallel way and are distributed in different measures. Decision tree and Random Forest is one of the most powerful and widely applied techniques for classification and prediction. The models with higher performance to classify diabetic patients. The Naive Bayes classification algorithm results determine the adequacy of the designed system with an achieved accuracy of 76.30 %. SVM is useful for the classification model for diabetes, which achieves good classification results it indicates the feasibility of using the information science method.

In the future, the designed system with the used ANN, machine learning classification, and Data mining algorithms can be used to predict or diagnose diseases. The work can be extended and improved for the automation of diabetes prediction including some other machine learning techniques.

REFERENCES

1) Shivakumar, B. L., and S. Alby. "A survey on data-mining technologies for prediction and diagnosis of diabetes." 2014 International Conference on Intelligent Computing Applications. IEEE, 2014.

2) Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint arXiv:1502.03774 (2015).

3) El_Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." (2018).

4) Hamdi, Takoua, et al. "Artificial neural network for blood glucose level prediction." 2017 International Conference on Smart, Monitored and Controlled Cities (SM2C). IEEE, 2017.

5) Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." International Journal of Engineering Research and Applications 3.2 (2013): 1797-1801.

6) Sittidech, Punnee, and Nongyao Nai-arun. "Random Forest Analysis on Diabetes Complication Data." Proceeding of the IASTED International Conference. 2014.

7) Woldemichael, Fikirte Girma, and Sumitra Menaria. "Prediction of Diabetes Using Data Mining Techniques." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018.

8) Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

9) Perveen, Sajida, et al. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

10) Purusothaman, G., and P. Krishnakumari. "A survey of data mining techniques on risk prediction: Heart disease." Indian Journal of Science and Technology 8.12 (2015): 1.