International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
463 – 466

A Review of Ensemble Machine Learning Approach in Prediction of Diabetes Diseases

Bhavana N[1], Meghana S Chadaga[2]
Department of Computer Science and Engineering
K.S. Institute of Technology, Bengaluru, India.
[1]*bhavanacse1996@gmail.com*,[2]*meghanachadaga@gmail.com*

Pradeep K R[3]
Assistant professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, India
[3]Pradeepkr22@gmail.com

***Abstract*—** Data mining techniques improve efficiency and reliability in diabetes classification. Machine learning techniques are applied to predict medical dataset to safe human life. The large set of medical dataset is accessible in data warehousing which used in the real time application. Currently Diabetes Diseases (DD) is among the leading cause of death in the world. Data mining techniques are used to group and predict symptoms in medical dataset by different examiners. Data set from Pima Indian Diabetes Dataset (PIMA) were utilized to compare results with the results from other examiners. In this system, the most well known algorithms; K-Nearest Neighbor (KNN), Naïve Bayes (NBs), Random Forest (RF) and J48 are used to construct an ensemble model. The experiment's result reveals that an ensemble hybrid model increases the accuracy by combining individual techniques in to one. As a result, the model serves to be useful by doctors and Pathologist for the realistic health management of diabetes.

Keywords- *Diabetes, Machine Learning, Data mining, Ensemble, KNN, NBs, RF, J48.*

*****

I. INTRODUCTION

Diabetes is a chronic disease caused due to abnormally high levels of sugar glucose in the blood. Diabetes is usually referred as "Diabetes mellitus". Diabetes is due to one of two mechanisms, Insufficient production of insulin (which is made by pancreas and brings down blood glucose), or Insufficient sensitivity of cells to the activity of insulin.

Globally, an estimated 422 million adults are living with diabetes mellitus, according to the latest 2016 data from the World Health Organization (WHO) [1].

Diabetes is grouped into two types namely, Type I and Type II diabetes. In Type I diabetes, is a chronic condition in which the pancreas produce little or no insulin, which is also known as insulin-dependent diabetes. In type II diabetes, the human body cannot use insulin the right way, which is also termed as non insulin-dependent diabetes.

II. RELATED WORK

Song et al. [2] Describes and explain different classification algorithms using different parameters such as Blood Pressure(BP), glucose, skin thickness, insulin, BMI, Diabetes pedigree and age. The researchers were not included to predict diabetes diseases. In this research, the researchers were using only small sample data for prediction of diabetes. The five different algorithms used are GMM, ANN, SVM, EM and Logistic Regression. This paper concludes high accuracy is provided by Artificial Neural Network (ANN).

Loannis et al. [3] proposed that machine learning algorithms are very important to predict different medical datasets including diabetes dataset (DDD).The paper proposed SVM, Logistic Regression and Naïve Bayes using 10 fold cross validation to predict different diabetes datasets.

III. A STUDY ON CLASSIFICATION ALGORITHMS

A. Logistic Regression

The classification algorithm aims to develop a model that can map data items to a given category, based on the existing data. It was used to extract significant data items from the model or to predict the tendency of data. The dependent variable of the logistic regression algorithm is binary-classification. It means that the logistic regression algorithm is always used to solve two-category problem. The main purpose of our experiment is to predict whether a person is diabetic or not, which is a typical binary-classification problem. Besides, the logistic regression algorithm is always used in data mining, disease diagnosis and economic prediction, especially predicting and classifying of medical and health problem. It predicts the probability of the outcome that can only have two values that is 0 0r 1. When output is 1 it means the value is greater than the threshold, else the output is 0. The range of output of logistic regression is always between 0 and 1. The main idea of Logistic regression is that it reduces the prediction range and limits the prediction value to 0 or 1.

B. K-Means

Cluster analysis aims to partitioning the observations into disparate clusters so that observations within the same cluster are more closely related to each other than those assigned to different clusters [4] . In the first stage, improved K-means algorithm is used to remove the incorrectly clustered data. The optimized dataset is used as input to the next stage. The main idea of K-Means is to divide the given unspecified data into fixed K number of centroids. A centroid is real or imaginary

 ISSN: 2454-4248
463 – 466

center of a cluster. Thus the analysis of these K partitions may provide a better characterization of data and may be of additionally benefited with fast computational than hierarchical structure, when K is small.

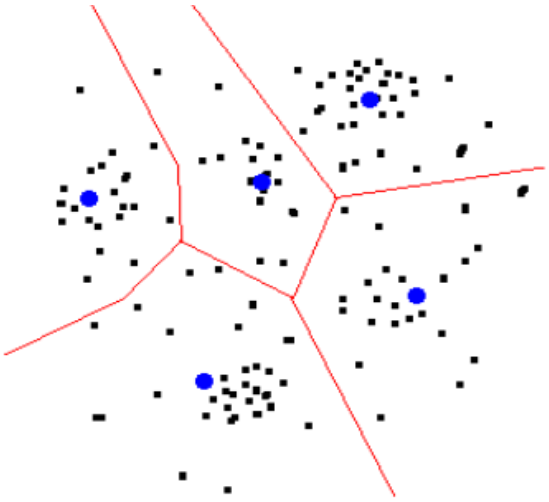

Figure 1:- A cluster representation.

The black dots (smaller dots) denote data points. The red line denotes partition created by K-Means algorithm. The blue dots (bigger dots) denote the centroids.

C. *KNN*

K-Nearest Neighbor is also referred as lazy learning algorithm that classifies datasets based on their similarity with neighbors. 'K' represents number of dataset items that are considered for the classification. KNN is used for both regression problem and classification problem. This technique classifies new belongings based on similarity measures [5]. Some advantages of KNN, it is very simple and intuitive. Good classification if the number of samples is huge enough. KNN also has few disadvantages like it is dependent on K value. Test stage is computationally expensive and need large number of samples for accuracy.

Steps on how to compute K-Nearest Neighbor (KNN) algorithm:

Step 1: Determine K, where K is number of nearest neighbors.

Step 2: Determine the distance between the instances and training samples.

Step 3: Order the distance and determine nearest neighbor based on the K- the minimum distances.

Step 4: Assemble the class of the nearest neighbor.

Step 5: Use majority of the class of closest neighbors as the prediction value of any query instances.

D. *Naive Bayes*

The Naive Bayes (NB) is a quick method for creation of statistical predictive models. NB is based on the Bayesian theorem. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class. In the process training, the probability is computed of each class by taking counting of many times it occurs in the training dataset. This is called the "prior probability" P(C=c). This probability becomes the product of the probabilities of each single attribute. Then the probabilities can be estimated from the frequencies of the instances in the training set. NB computes based on possibility by using Bayes formula [6]. Using Bayes' theorem, the conditional probability can be decomposed as shown in equation (1):

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

Where,
P(c|x) is Posterior Probability
P(x|c) is likelihood
P(c) is Class Prior Probability
P(x) is Predictor Prior Probability

E. *J48*

J48 is an extension of ID3. The extra highlights of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc...It creates a decision tree based on labeled input data. The tress generated can be used for classification and for this reason it is called as statistical classifier. In the WEKA data miming tool, J48 is an open source Java implementation of the C4.5 algorithm. Advantages of J48 algorithm, it is easy and simple to understand [7] and can use both categorical and continuous values.

IV. ENSEMBLE MODELING

It is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications, this helps to improve machine learning results by combining several models. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting).

A. *Bagging*

Bagging stands for Bootstrap Aggregating, it is a method to diminish the variance of prediction by generating extra data for training from the original dataset. To improve classification accuracy and unstable classification problems [8]. Over fitting is avoided by bagging method. This is a simple method to understand if the quantity is a descriptive statistic such as a mean or a Standard Deviation (SD).Consider we have a sample

 ISSN: 2454-4248
463 – 466

of 100 values (x) and need to estimate mean of the sample. The mean is calculated as shown in below equation (2):

Mean (x) = 1/100 * sum(x). (2)

B. Boosting

Boosting is a two-step approach. The boosting method uses subsets of the original data to generate a series of averagely performing models. As, the name suggests boosting means it "boosts" the performance by combining them together using a particular cost function. Boosting method will create a strong classifier from many different weak classifiers. In general, this method works by building a model from the training data, and then creating a second model that attempts to correct the errors from the first model.

TABLE 1: COMPARISON BETWEEN BAGGING AND BOOSTING

	Bagging	Boosting
Example	Random Forest	Gradient Boosting Method
Bias/Variance Trade Off	Decreases Variance	Decreases Bias
Training Data	Partition before training	Adaptive data weighting
Base Learner	Complex	Simple
Classifier Strength	Tries to overfit with a flexible model and then average to decrease variance.	Tries to underfit using weak learners and then improve model based on classifier performance.

V. MODELS EVALUATION

Every classification model is fitted utilizing 10-fold cross validation. The assessment metric utilized as a part of this work is precision to assess the adequacy of the investigation models. This metric registered from number of components in confusion network which are generally assessed [9]. The perplexity network, if there should arise an occurrence of two classes forecast contains True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy is characterized by the following equation (3):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

EXPERIMENTS AND RESULTS

The various classification algorithms are applied to the data set in order to find out the best diabetes classifier model. The conceptual framework of 5 steps is depicted in below Figure2:

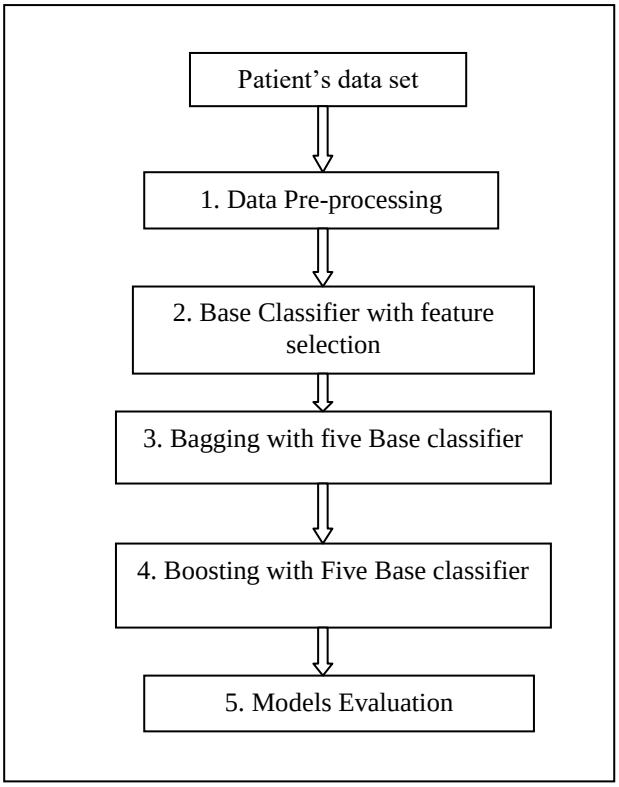

Figure 2. Conceptual Framework.

Step 1: Data Pre-processing

Data Pre-processing is a data mining method where it converts all physical data into understandable form. Raw data will be not clear, complete and consistent in features and chances of having many errors. Thus, Data Pre-processing method will resolve all such problems. The data set consists of 8 input attributes and 1 output attribute. The below describes all the attributes under consideration:

- Pregnant: Number of times pregnant
- BP: Diastolic Blood Pressure(mmHg)
- FAST GTT: Glucose Tolerance Test
- INSULIN: Serums Insulin(U/ml)
- SKIN: Triceps Skin Thickness(mm)
- BMI: Body Mass Index(kg/m)
- DPF: Diabetes Pedigree Function
- AGE: Age of the person(years)
- DIABETES: Diabetes diagnosis result(no, yes, pre)

Step 2: Base classifiers with feature selection

The 8 attributes are placed by addition algorithm for choosing attributes that are important for diabetes risk factors. The attributes are placed in the following order: BP, INSULIN, SKIN, BMI and AGE. Then, well suited algorithms are ; Naïve Bayes, K-Nearest Neighbors are made used to construct classification models along with the attributes. All models

 ISSN: 2454-4248
463 – 466

were tested using 10-folds cross validation to dodge model over-fitting.

Step 3 : Bagging with three base classifiers

The bagging method is applied to the base classifiers that are previously mentioned .The method is tested with various percentage of bagging size that ranges between 70-100. Above all, Logistic Regression resulted in high accuracy and it is substantially the most widely used for its good features.

Step 4 : Boosting with three base classifiers

The boosting method is applied to the base classifiers. The method was tried with different values of weight of threshold. Above all, accuracy of boosting with base classifier Logistic Regression resulted high and is the most comprehensively used for its incredible features.

Step 5: Models Evaluation

The bagging method with base classifier Logistic Regression resulted in high accuracy than boosting method with base classifier Logistic Regression. As a result, the ensemble methods provide better performance than base classifiers.

VI. CONCLUSION

In this review, Machine learning methods have different power in different dataset. Machine learning and data mining can be helpful in providing vital statistics, real- time data and advanced analytics in terms of the patient's disease, lab test results and blood pressure and so on to doctors. Using these types of advanced machine learning algorithms we can provide better information to doctors at the point of patient care. The classification algorithm are used to predict what is the chance for pregnant women is having type 2 diabetes are not. Here, an attempt is made to combine individual technique into one on order to increases the performance and accuracy. Thus, Single algorithm provide less accuracy than ensemble one.

ACKNOWLEDGEMENT

We would like to express our great attitude to R&D, Department of CSE, KSIT, Bengaluru for their constant guidance and encouragement.

REFERENCES

[1] World Health Organization, Global Report on diabetes.Genera, 2016 accessed 30 august 2016.

[2] Komi, M., Li, J., Zhai, Y., and Zhang, X, "Application of data mining methods in diabetes Prediction in image , vision and computing (IVIVC)", pp.1008-1010,June 2017.

[3] Kavakiotis, I., Tsave, o., Salifoglou, A, Magalaveras, N., Vlahavas, I., and chauvarda, I, " Machine learning and data mining methods in diabetes research",2017.

[4] Guajun G, Chaoane M, Jianhong w, " In data clustering theory algorithm and application finted", ASA-Siam.M, 2007.

[5] Saran nanthan,K,and Velmurgan, T, " Analyzing Diabetic data using classification algorithm on data miming" ,Indian journal of science and technology, 9(43),2016.

[6] Ramzan, M., "Comparing and evaluation the performance of WEKA classifications on critical disease" ,In information processing (IICIP), 1st Indian International conference on (pp.1-4) IEEE,August 2016.

[7] Patil, B.M., Joshi, R.C., and Toshniwal.D,"Hybrid prediction model for type-2 diabetics Patients". Expect system with applicaton,37(12),8102-8108.//9,2010.

[8] L. Breiman, Bagging predictors, machine leraning,24(2),123-140,1996.

[9] I. Syarif,E. Zaluska, A. Prugel-Bennett and G.Wills, "Application of Bagging, booting and stacking to intrusion dection",MLDM2012,LNAI7376,513-602,(2012)

[10] http://who.int/medicentre/factsheets/fs312/en/index.html.

[11] P. Giudici, S. Figini, " Applied data mining for business and industry", second ed., Italy,2009.

[12] J.Quinlan,Induction of decision tree, Readings in Machine learning,1986.