IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44
www.iosrjournals.org

An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques

B. Tamilvanan[1], Dr.V. Murali Bhaskaran[2]

[1]*Research and Development Centre, Bharathiar University, Coimbatore-641046, TN, India.*
[2]*Principal, Dhirajlal Gandhi College of Technology, Salem-636290, TN, India*

Abstract: *Data mining means to the process of collecting, searching through, and analyzing a large amount of data in a database. Classification in one of the well-known data mining techniques for analyzing the performance of Naive Bayes, Random Forest, and Naïve Bayes tree (NB-Tree) classifier during the classification to improve precision, recall, f-measure, and accuracy. These three algorithms, of Naive Bayes, Random Forest, and NB-Tree are useful and efficient, has been tested in the medical dataset for diabetes disease and solving classification problem in data mining. In this paper, we compare the three different algorithms, and results indicate the Naive Bayes algorithms are able to achieve high accuracy rate along with minimum error rate when compared to other algorithms.*

Keywords: *Data mining, Classification, Naive Bayes, Random Forest, NB-Tree, Weka*

I. Introduction

Data mining techniques and application are used in a wide range of fields, including banking, social science, education, business industries, bioinformatics, weather, forecasting healthcare and big data [9,11]. Nowadays health care industry generates a large amount of data about patients, disease diagnosis, etc. Some different types of approaches to building accurate classification have been proposed (e.g., Naive Bayes, Random Forest, and Naïve Bayes tree). In classification, we give a diabetes data set of example record or the input data, called the test data set, with each record consisting of various attributes.

An attribute can be either a numerical attribute or categorical attribute. If values of an attributes belong to an ordered domain, the attribute is called numerical attribute (e.g.Preg, Plas, Pres, Skin, Insu, Mass, Pedi, Age). A categorical attribute (e.g.Class). classification is the process of splitting a dataset into mutually exclusive groups, called a class, based on suitable attributes.

In this world, different types of Diabetes diseases simply named to as diabetes is a disorder caused when the pancreas does not produce Insulin or body cells no longer respond to Insulin. Our body cells are fueled by Insulin which is one type of hormone which acts as a key that allows glucose from the blood to enter into our cells. If in the pancreas, the insulin-producing beta cells are put down, then the glucose in the blood is not adequately regulated, as a consequence, the glucose level in the blood increases abruptly this causes an individual to be diabetic. On that point are primarily four types of diabetes. Prediabetes is categorized by the glucose level higher than a normal but not yet high enough to be characterized as diabetes. Type1 diabetes is an autoimmune disease that causes the insulin-producing beta cells in the pancreas to be destroyed, inhibiting the body from being able to yield enough insulin to effectively regulate the blood glucose levels. Type 2 diabetes is a metabolic disorder that results from insulin resistance; the body cells no longer react to insulin hormone, a situation in which cells fail to utilize insulin properly. Gestational diabetes refers to higher than normal blood glucose level occurring during gestation in adult females who were not diabetic before pregnancy. It is usually developed between twenty-four and a twenty-eight week of gestation.

This paper is organized accordingly: the relates works and description of the technical aspects of the used data mining methods in section 1. The introduction of the dataset for diabetes in section2. The experimental and comparative results in section 3. And finally, conclude the paper and future works.

II. Related Works And Methods

The Random Forest is appropriate for high-dimensional data modeling because it can handle missing values and can handle continuous, categorical and binary data. The bootstrapping and ensemble scheme makes Random Forest strong enough to overcome the problems of over-fitting and hence there is no need to prune the trees. Besides high prediction accuracy, Random Forest is efficient, interpretable and non-parametric for various types of datasets [18**].**

In [19] we proposed an intrusion detection system model based on K-star and Information gain for feature set reduction. The key idea of this paper is to take advantage of the instance-based classifier and dataset features reduction for intrusion detection system, the model has the ability to recognize attacks with high detection rate and low false negative. Shishupal and Parvat in [20] proposed a layered approach and compared

DOI: 10.9790/0661-1901043944

the proposed layered approach with the Decision Tree and the Naive Bayes classification methods.

Iterative dichotomizer 3 (ID3) which is used to build a decision tree was first developed by Quinlan (1979). It is a top-down approach starting with selecting the best attribute to test at the root of the tree[8]. The selection of the best attribute in ID3 is based on an information theory approach or entropy. Entropy is used to measure how informative a node is. This algorithm uses the criterion of information gain to determine the goodness of a split. The attribute with the greatest information gain is taken as the splitting attribute, and the data set is split for all distinct values of the attribute.

Naive Bayes

The Bayesian classifier is based 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 other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered ”naive.”

Let X = {x1, x2, . . . , xn} be a sample, whose components represent values made on a set of n attributes. In Bayesian terms, X is considered ”evidence.” Let H be some hypothesis, such as that the data X belongs to a specific class C. We have to determine P(H|X), the probability that the hypothesis H holds given the ”evidence,” (i.e. the observed data sample X). According to Bayes' theorem, the probability that we want to compute P(H|X) can be expressed regarding probabilities P(H), P(X|H), and P(X) as

$P(H|X) = P(X|H)\ P(H) / P(X)$

Random Forest

Random forest is an ensemble prediction method by aggregating the result of individual decision trees. In the past decade, various methods have been proposed to grow a random forest1–3,18. Among these methods, Breiman's method1 has gained increasing popularity because it has a higher performance against other methods19. Let D be a training dataset in an M-dimensional space X, and let Y be the class feature with a total number of c distinct classes. The method for building a random forest1 follows the process including three steps18:

Step 1: Training data sampling: use the bagging method to generate K subsets of training data {D1, D2,..., DK} by randomly sampling D with replacement;

Step 2: Feature subspace sampling and tree classifier building: for each training dataset Di (1≤ I ≤ K), use a decision tree algorithm to grow a tree. At each node, randomly sample a subspace Xi of F features (F << M), compute all splits in subspace Xi, and select the best split as the splitting feature to generate a child node. Repeat this process until the stopping criteria are met, and a tree hi (Di, Xi) built by training data Di under subspace Xi is thus obtained.

Step 3: Decision aggregation: ensemble the K trees {h1 (D1 , X1), h2 (D2 , X2), ... , HK(DK, XK)} to form a random forest and use the majority vote of these trees to make an ensemble classification decision. The algorithm has two key parameters, i.e., the number of K trees to form a random forest and the number of F randomly sampled features for building a decision tree. According to Breiman1 , parameter K is set to 100 and parameter F is computed by F = [log2 M + 1]. For large and high dimensional data, a large K and F should be used.

NB-Tree

Cross between Naive Bayes classifier and C4.5 Decision Tree classification and it's best described as a decision tree with nodes and branches [15]. The NB-Tree algorithm is written below with input of T sets of labeled instances and a decision-tree with Naive Bayes category at the output (leaves): 1. For each attribute, Xi, evaluate the utility, u(Xi), of a split on attribute Xi. For continuous attributes, a threshold is also evaluated at this stage. 2. Let J = AttMax(UI). The attribute with the highest utility (Maximum utility). 3. If Uj is not significantly better than the utility of the current node, create a Naive Bayes classifier for the current node and return. 4. Partition T according to the test on Xj. If Xj is continuous, a threshold split is used; if Xj is discrete, a multi-way split is made for all possible values. 5. For each child, call the algorithm recursively on the portion of T that matches the test leading to the child.

III. Diabetes Disease Dataset

The performances of these three algorithms namely NB, Random forest, NB-Tree was tested in a medical database for diabetes Disease dataset from UCI machine learning repository (available at http://archive.ics.uci.edu/ml/datasets/Diabetes) [6]. The data set has nine features of the attributes. Table -1 describes the data for Diabetes. The medical dataset contains data from reviews conducted among patients, each of which has 9 features. All features can be considered as on indicators of diabetes disease for a patient. The dataset holds records of the following attributes.

Table 1: UCI Dataset of Diabetes Disease

Attributes Name	**Attribute Type**	**Description**
Preg	Numeric	Number of times pregnant
Plas	Numeric	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Pres	Numeric	Diastolic blood pressure (mm Hg)
Skin	Numeric	Triceps skin fold thickness (mm)
Insu	Numeric	2-Hour serum insulin (mu U/ml)
Mass	Numeric	Body mass index (weight in kg/(height in m)^2)
Pedi	Numeric	Diabetes pedigree function
Age	Numeric	Age (years)
Class	Discrete	Class variable (0 or 1) class value zero is interpreted as "tested positive, class value one is interpreted as "tested negative.”

IV. Experiment Results And Discussion

In this section, we describe the test database and experimental analysis and the current evaluation results for three algorithms namely NB, Random forest, NB-Tree classifier.

In this experimental analysis, Naive Bayes, Random Forest, and NB-Tree Algorithms performance were compared based on their application in medical datasets. Weka tool is used for research, banking, education and weather datasets. It helps in coordinated activities in machine learning, data mining, text mining and web mining. It supports all the mining process to get a valid and clear visualization of accurate results. Ten-fold cross-validation was to the input datasets in the experiments.

Experimental Step Up

A brief description of the classification process by all three algorithms Naive Bayes, Random Forest, and NB-Tree are given below:

Data mining methods also need an evaluation procedure. This procedure is used to verify the models generated by three algorithms namely Naive Bayes, Random Forest, and NB-Tree. Classification techniques can be evaluated using the data labels in supervised learning methods. The different matrices are used to evaluate the classification algorithm, such as confusion matrices for Precision, Recall, F-measure, and Accuracy.

Table 2: Confusion Matrix

	A	**B**
A	tpA	eAB
B	eBA	tpB

The confusion matrix is shown in Table 2. In the confusion matrix, the diagonal elements are correctly classified data, and the rest of elements are incorrectly classified data. Precision is defined as the ratio between the true positive value and both the true positive and false positive values.

Table 3: Confusion Matrix Using NB-Tree

	A	B	Total
A	409	91	500
B	112	156	268

The confusion matrices for the NB-Tree algorithm are shown in Table-3. This classification uses the class values of A tested negative, B-tested positive. The result from the confusion matrix is discussed in each class as given below

There are 500 items found are classified into class value for A-tested negative, 409 of these items are exactly classified into class A, 91 of these items are incorrectly classified into class B, There are 268 items found are classified into class value for B-tested positive, 112 of these items are incorrectly classified into class A, 156 of these items are exactly classified into class B.

Table 4: Confusion Matrix Using Navie Bayes

	A	B	Total
A	422	78	500
B	104	164	268

The confusion matrices for Naive Bayes algorithm is shown in Table-4. This classification uses the class values of A - tested negative, B-tested positive. The result from the confusion matrix is discussed in each class as given below

There are 500 items found are classified into class value for A-tested negative, 422 of these items are exactly classified into class A, 78 of these items are incorrectly classified into class B, There are 268 items found are classified into class value for B-tested positive, 104 of these items are incorrectly classified into class A, 164 of these items are exactly classified into class B.

Table 5: Confusion Matrix Using Random Forest

	A	B	Total
A	417	83	500
B	110	158	268

The confusion matrices for Random Forest algorithm is shown in Table-5. This classification uses the class values of A-tested negative, B-tested positive. The result from the confusion matrix is discussed in each class as given below

There are 500 items found are classified into class value for A-tested negative, 417 of these items are exactly classified into class A, 83 of these items are incorrectly classified into class B, There are 268 items found are classified into class value for B-tested positive, 110 of these items are incorrectly classified into class A, 158 of these items are exactly classified into class B.

Precision

It is used to represent the fraction of retrieved data from connecting datasets, which are relevant to the search. Precision will be used to represent how many instances have been correctly classified in the confusion matrix table (correct classified data is truly positive and incorrect classified data is error positive).

$$\text{Precision} = \frac{tpA}{tpA + eBA}$$

Where tpA is represented as true positive for the class A and eBA are represented as false positive.

Recall

It is used to represent the fraction of retrieved data from connecting datasets, which are relevant to the query that is successful. It is used to find out the ratio between the true positive and both true positive and false positive values.

$$Recall = \frac{tpA}{tpA + eAB}$$

Where tpA is represented as true positive for the class A and eAB are represented as error positive.

F-measure This is evaluated by the harmonic mean between precision and recall.

$$\text{F} - \text{Measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy This is calculated as the proportion of true positive, true negatives and true results from all the given data.

$$\text{Accuracy} = \frac{tpA + tpB}{tpA + eAB + eBA + tpB}$$

Error Rate= 1 - Accuracy.

These results are shown in table 6,7 and 8. The total records are used in the experiment as shown in tables 3, 4 and 5.

Table 6: Confusion Matrix Using NB-Tree

	Results		
	Precision	**Recall**	**F-Measure**
A	0.785	0.818	0.801
B	0.632	0.582	0.606

Table 7: Confusion Matrix Using Naive Bayesian

	Results		
	Precision	**Recall**	**F-Measure**
A	0.802	0.844	0.823
B	0.678	0.612	0.643

Table 8: Confusion Matrix Using Random Forest

	Precision	**Recall**	**F-Measure**
A	0.791	0.834	0.812
B	0.656	0.59	0.621

Table 9: Average of the Precision, Recall, F-Measure, and Accuracy

Algorithm Names	**Results**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
NB-Tree	0.731	0.736	0.733	0.735
Naive Bayes	0.759	0.763	0.76	0.763
Random Forest	0.744	0.749	0.745	0.748

The results are evaluated by the precision, recall, and f-measure and compared the results with three algorithms namely Naive Bayes, Random Forest, and NB-Tree. It shows the better accuracy for the classification as shown in figure 1. The final result, the Naive Bayes algorithm seems to be a good performance for the supervised learning method.

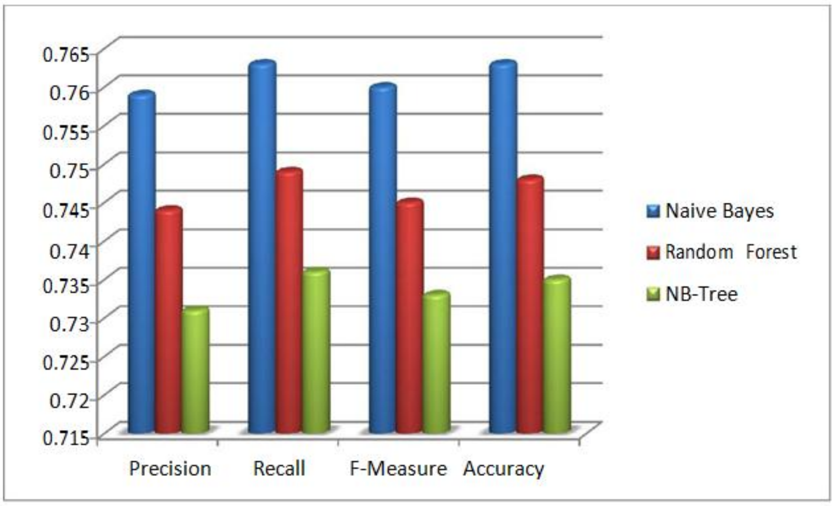

Figure 1: the comparison of classification precision, recall, F-measure and accuracy

V. Conclusion

The overall objective of our work is to predict more accurately the presence of diabetes. The efficient three classification algorithms namely Naive Bayes, Random Forest, and NB-Tree are used to develop the model and all are evaluated with 10 fold cross-validation. These three algorithms are compared, and accuracy is evaluated for true positive and false positive rate. From the experiments, it is observed that Naive Bayes had the best predictive power with high accuracy 76.3% and less error rate 23.7% as compared to Random Forest, and NB-Tree.

This system can be further expanded. It can use more number of input attributes table - 1.Other data mining techniques can also be used for prediction (e.g. Clustering, Association Rules).the test mining & web mining can be used to mine a huge amount of unstructured data available in healthcare industry database.

References

[1]. Joachims T., 1998, “Text categorization with support vector machines: Learning with many relevant features,” ECML'98. pp.137-142.

[2]. Quinlan J., 1993, “C4.5:Programs for Machine Learning, ”. The Morgan Kaufmann.

[3]. AI-hegami A. “Pruning Based Interestingness of Mined Classification Patterns.

[4]. International Arab Journal of Information Technology. Vol. 6, No. 4, pp. 336-343, 2009.

[5]. Manish M., Rakesh A., and Jorma R., 1996, “SLIQ: A Fast Scalable Classifier for Data Mining,” Int. Conference on Extending Database Technology(EDBT’96), Avignon, France.

[6]. John S., Rakesh A., and Manish M., 1996, “SPRINT: A Scalable Parallel Classifier for Data mining, ”. proceedings of the 22nd VLDB Conference Mumbai (Bombay), India.

[7]. UCI Machine Learning Repository (2013). Available from: http://archive.ics.uci.edu/ ml/datasets.html

[8]. Breiman L., Friedman J.H., Olshen R.A., and Stone C.J., 1984, “Classification and Regression Trees, ” Wadsworth,Belmont.

[9]. Quinlan J R.,1979, “ Discovering rules by induction from large collections of examples, ” Expert Systems in the Micro Electronic Age, Edinburgh University Press, 168–201.

[10]. Quinlan J R., 1996, “Improved use of continuous attributes in C4.5, ” Journal of Artificial Intelligence Research 4: 77-90.

[11]. Roohallah A., Jafar H., Mohammad J H., Hoda M., Reihane B., Asma G., Behdad B.,and Zahra A S., 2013, “ A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, pp.53-61.

[12]. Sitar-Taut, V.A., et al.,2009, “Using machine learning algorithms in cardiovascular disease Risk evaluation. Journal of Applied Computer Science & Mathematics.

[13]. Wu, X., et al.,2007, “Top 10 algorithms in data mining analysis,” Knowl. Inf. Syst.

[14]. Bennett K.P., and Blue J.A.,1998, “A support vector machine approach to decision tree,” In proceedings IJCNN'98, pp. 2396–2401.

[15]. Han j., and Kamber M., 2006, “Data Mining : Concepts and Techniques,” Morgan Kaufmann Pulishers, San Francisco, CA.

[16]. Ganesan.p., and Sivakumar.s.,2015, “An Experimental Analysis of Classification Mining Algorithm For Coronary Artery Disease,” Research India Publications.

[17]. Deeman Yousif Mahmood*, Dr. Mohammed Abdullah Hussein.,2014, “AnalyzingNB, DT and NBTree IntrusionDetection Algorithms,” In Journal of Zankoy Sulaimani- Part A (JZS-A).

[18]. Chaitrali S Dangare., and Sulabha S.Apte., 2012, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques” International Journal of Computer Applications.

[19]. Yanjun Qi., “Random Forest for Bioinformatics”.www.cs.cmu.edu/~qyj/papersA08/11-rfbook.pdf

[20]. Deeman Y. Mahmood, Mohammed A. Hussein, "Intrusion Detection System Based on

[21]. KStar Classifier and Feature Set Reduction", International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE) Vol.15, Issue 5, PP. 107-112, Dec. 2013

[22]. Rupali S. Shishupal , T.J.Parvat, " Layered Framework for Building Intrusion Detection Systems", International Journal of Advances in Computing and Information Researches ISSN:2277-4068, Volume 1– No.2, April 2012.

[23]. K. Sowjanya, Ayush Singhal, Chaitali Choudhary, MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices", IEEE International Advance Computing Conference (IACC),2015