International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Predictive Modeling of Clinical Data Using Random Forest Algorithm and Soft Computing

Sanika Shah[1], M. A. Pradhan[2]

[1] Pune University, All India Shri Shivaji Memorial Society's College of Engineering, Kennedy Road, Pune-1

[2] Professor Pune University, All India Shri Shivaji Memorial Society's College of Engineering, Kennedy Road, Pune-1

Abstract: ***Clinical data which includes data of patients and their symptoms is growing largely these days. Detection of a disease in some cases is expensive in terms of money and amount of effort spent. Predictive modeling aids in the early detection of a disease by using health records (HRs). By applying such techniques on an available clinical dataset, a prediction of the current state of a patient's disease can be made. The predictive model, in this paper is a classifier, which uses a combination of the random forest algorithm and the genetic algorithm. Each record from the HRs serves as an input to the classifier. The results of classification show that the random forest algorithm and soft computing techniques give better results.***

Keywords: Predictive modeling, clinical data, health records, random forest algorithm, soft computing.

1. Introduction

In clinical care, predictions under uncertainty, including risk assessment with diagnosis have to be made. If those predictions can be made easily, the clinical care is likely to be better. Computer based clinical prediction methods are giving opportunity to improve diagnosis. Such methods are likely to be the predominant components of decision support systems. The data in HRs can be used to construct prediction models using different methods. Soft computing techniques used in computer field help to devise methods that give a solution which is acceptable and at a low cost. Cost is measured in terms of time, space or memory and money.

2. Literature Review

Simon Bernard, Sébastien Adam, Laurent Heutte, have stated that the random forest algorithm fall in the category of classifiers which rely on a combination of different trees [1].

In their paper, the authors A. Hapfelmeier, K. Ulm state that the random forest algorithm deals with high dimensional data and missing values to achieve better accuracy [2].

The paper, suggests the use of the random forest algorithm for feature selection and then applying the genetic algorithm, to build a model for crash detection [3].

Andrew Kusiak and Anoop Verma, compare the random forest algorithm with the support vector machines, neural network and boosted tree algorithms and the results are in the favor of the random forest algorithm which gives better performance [4].

Akin Ozcift, mentions in the paper, that the random forest classifier is based on the ensemble method and it is found that this classification strategy is more accurate than the other supervised classification algorithms. [5]

The paper by Jasper Diesel, Stefan Winter gives the readers an insight on the random forest algorithm for the dataset consisting of BCG signals for detecting atrial fibrillation in a patient [7].

3. Preprocessing Techniques Used

Preprocessing of data has to be done in order to remove the noisy data. This kind of data includes missing values, values which are out of range, null values, etc. The noisy data has to be smoothed out. Data smoothing is done so as to remove any sort of noise prevailing in the dataset. The following methods are available for data preprocessing.

3.1 Removal of Missing Values

Any given dataset is likely to have missing values in it. Attribute values for a specific record may be blank or missing. These values are referred to as missing values. If such a dataset is provided to the predictive model without any modifications, the results will not be accurate. The training phase for the model will not be precise and hence the testing results for any non tested record may vary significantly. Missing values are replaced by values predicted by certain methods like mean, average, mode, median, etc

3.2 Normalization

The dataset available at hand may be raw data. This dataset has different attributes which define the characteristics of the available records. Thus, each of these attributes may different data types (like numerical, character, etc.). The range of values that each of these attributes take may vary widely. Thus, the predictive model is likely to be biased toward the high or weighted values. In order to minimize this bias, all the values in the dataset are normalized (generally in the range of 0 to 1). This technique of reducing the bias toward a particular range of values is called normalization.

3.3 Discretisation

The dataset may have numerical values that are continuous in nature. These continuous values may not be able to precisely

predict the missing values. Thus the predictive model makes use of the discretization technique to avoid such situations. Discretization is a technique where all the values which are continuous in nature are discretized. This method converts continuous values into non continuous values.

4. Methods Used for Classification

4.1 Classification and Regression Tree (CART)

CART is used for creating predictive models [8]. It is a machine learning method where the models are obtained by partitioning the data space recursively. So, the partitioning can be represented as a decision tree. Classification trees take input of for dependent variables that take limited number of values, with prediction error. Regression trees use non-ordered or sometimes continuous values.

4.2 The J48 Algorithm

The J48 algorithm focuses on the generation of decision trees. Decision trees also are those trees where there is a root node and there are a number of sub nodes. The tree contains leaf nodes which. Leaf nodes have no branching. They lie at the last level such that there is no other different branching emerging from them. It is an algorithm that extends ID3 and can be used for the purpose of classification. It is a statistical classifier.

4.3 Bagging

Bagging is an ensemble method in machine learning also called bootstrap aggregation. This method makes use of majority votes dictated by the classifiers. All the classifiers give their votes and the best or the majority vote is given such that the best class for the testing record is obtained by the method of voting. Votes of these classifiers are combined and the final verdict is the majority of these votes.

4.4 The AdaBoost Algorithm

Similar to bagging, the AdaBoost[6] is also an ensemble method in machine learning. However, the difference between aging and boosting lies in the manner in which the vote is dictated by the classifier. Unlike bagging, the AdaBoost algorithm gives weighted votes. Thus, the result is a boosted classifier.

4.5 The Random Forest Algorithm

The random forest algorithm can also be thought of as an ensemble method in machine learning. The input to a random forest algorithm is a dataset consisting of records, with attributes. Random subsets of the input are created. On each of the random subset created, a decision tree will be constructed. The final class of a test record will be decided by the algorithm which uses the majority vote technique. Random forest algorithm makes use of the out of bag error technique.

5. Results Obtained

A comparison of the results of the experimentation on the heart_disease dataset is shown in the table 1.

Table 1: Results of experimentation on heart_disease dataset

Classifier	*Accuracy (in %)*
CART	76.5306%
J48	79.9320 %
Bagging	79.5918 %
AdaBoost	78.9116 %
Random Forest	80.9524 %

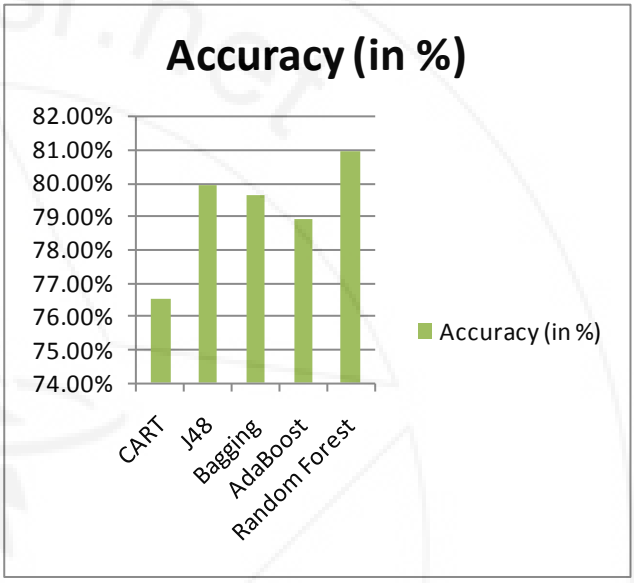

Figure 1: Graphical representation of accuracy

The results of experimentation show that the random forest algorithm gives better performance than the CART, J48, Bagging and AdaBoost classifiers. Similarly, many such algorithms are available, which are the variants of the mentioned algorithms. A comparison is done as shown in figure 1. Form the table we can deduce that the random forest algorithm gives higher accuracy than any other algorithm mentioned in the table 1.

6. Conclusion

As per the analysis of the algorithms namely the CART, J48, Bagging, AdaBoost and Random forest, we find that the result given by the random forest algorithm is better than the other algorithms. The random forest algorithm is suitable for many instances if datasets where there are a large number of records available. This algorithm is scalable even for datasets where the amount of data is large and can be used for many instances of the available dataset. Data preprocessing is needed for obtaining these results.

References

[1] Simon Bernard, Sébastien Adam, Laurent Heutte, “Dynamic Random Forests”, Pattern Recognition Letters 33 (2012) 1580–1586, Elsevier 2012.

[2] A. Hapfelmeier, K. Ulm, "A new variable selection approach using Random Forests", Computational Statistics and Data Analysis 60 (2013) 50–69, Elsevier 2013.
[3] Chengcheng Xu, Wei Wang, and Pan Liu, "A Genetic Programming Model for Real Time Crash Prediction on Freeways", IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 14, NO. 2, JUNE 2013
[4] Andrew Kusiak, Anoop Verma, "A Data-Mining Approach to Monitoring Wind Turbines", IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, VOL. 3, NO. 1, JANUARY 2012
[5] Akin Ozcift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis", Computers in Biology and Medicine 41 (2011) 265–271, Elsevier 2011.
[6] AdaBoost, Available: http://en.wikipedia.org/AdaBoost
[7] Christoph Bruser, Jasper Diesel, Matthias D. H. Zink, Stefan Winter, Patrick Schauerte, and Steffen Leonhardt, "Automatic Detection of Atrial Fibrillation in Cardiac Vibration Signals", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 17, NO. 1, JANUARY 2013
[8] Classification and regression trees, Available: http://en.wikipedia.org/wiki/classificationRegressionTrees
[9] Yali Amit, Donald Geman, "Shape quantization and recognition with randomized trees", Neural Computation Volume 9 Issue 7, Oct. 1, 1997, MIT Press Cambridge, MA, USA.
[10] Sebastien Adam, dblp computer Science bibliography, Available: http://www.informatik.uni-trier.de/~ley/pers/hd/a/Adam:S=eacute=bastien.html
[11] Chesner Désir, Simon Bernard, Caroline Petitjean, Laurent Heutte, " One class random forests", Pattern Recognition 46(12): 3490-3506 (2013)

Author Profile

Ms. Sanika Shah is a student of M.E. (Computer) at AISSMS's CoE from Savitribai Phule University of Pune. She received B.E. degree from Walchand Institute of Technology, Solapur in Computer Science and Engineering.

Mrs. Madhavi Pradhan is an Assistant Professor at AISSMS's CoE, Pune. She received M.Tech (Software Engineering) from S.J.C.E. Mysore and B.E. (Computer Engineering) from Government College of Engineering, Amravati.