International Journal of Pure and Applied Mathematics
Volume 119 No. 10 2018, 585-595
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

Illustration of Random Forest and Naïve Bayes Algorithms on Indian Liver Patient Data Set

[1]M. Aiswarya, [2]Swathi Srinivas and [3]A.G. Hari Narayanan

[1]Dept of Computer Science & IT,
Amrita School of Arts & Sciences, Kochi,
Amrita Vishwa Vidyapeetham, India.
aiswaryamurali002@gmail.com

[2]Dept of Computer Science & IT,
Amrita School of Arts & Sciences, Kochi,
Amrita Vishwa Vidyapeetham, India.
swathi.srnvs1221@gmail.com

[3]Dept of Computer Science & IT,
Amrita School of Arts & Sciences, Kochi,
Amrita Vishwa Vidyapeetham, India.
hariag2002@gmail.com

Abstract

Big data is one of the most trending technologies in today's highly technical world, comprising of very huge data sets, which on analysis can be used to reveal different patterns, which can in turn help in efficient decision making. The various techniques used in Big Data such as association, classification, regression, machine learning etc helps to produce nearly accurate results, and hence can be best used in areas such as the medical field where accuracy plays a crucial role in the prediction and diagnosis of diseases. The sole purpose of this paper is to compare the performance of the classification algorithms–Random Forest and Naïve Bayes, in identifying the nature of liver diseases based on the Indian Liver Patient Dataset (ILPD). The factors responsible for the cause of disease are identified from this dataset. These results can be used for the development of Expert Liver Diagnosis Systems using Big Data Analytics tools.

Key Words: Big data, indian liver patient data set, naïve bayes algorithm, random forest algorithm.

1. Introduction

The past few years have seen big advances in the amount of data that is being generated and gathered. Merging latest technologies with data we get Big Data and its applications are huge in this highly industrial world. Big Data applications include various techniques and methodologies, which have been developed for handling data that is too large and complex for any traditional data application to process. These methodologies have been used in significant areas such as Healthcare, Banking, and Media and so on. Rather than just improving profit, Big Data in healthcare can be used to predict diseases and also improve the quality of life. With the increase in population and evolution in technology, methodologies used for treatment are also fast changing. Liver is an organ in the human body responsible for many vital functions. There are several factors that can knowingly or unknowingly result in liver damage such as alcohol consumption, obesity and diabetes. Different symptoms must also be looked into as a cause for liver diseases. Multiple tests can be done to detect a disease. The common liver disorders known are Hepatitis, Cirrhosis and Liver Cancer.

2. Related Work

Prof. M.S. Prasad Babu, K. Swapna, Tilakachuri Balakrishna, and Prof. N.B.Venkateswarulu (2014) in their paper focus on the analysis of the performances of different hierarchical clustering algorithms for the ILPD (Indian Liver Disease Prediction Dataset). The result showed that these algorithms gave more accurate results when applied on this data set than the already existing system. The WEKA tool was used to get the results for the above dataset. (Prof. M.S. Prasad Babu, K. Swapna, Tilakachuri Balakrishna, and Prof. N.B.Venkateswarulu, 2014). Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof. N. B. Venkateswarlu(2012) analyzed the percentage of the population in India and USA with liver diseases. It was observed that the variance in the performance of the classifiers used was directly dependent on the percentage of patients within the population. Three different experiments were conducted on the datasets and each experiment provided varying results of different accuracy (Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof. N. B. Venkateswarlu, 2012).Wangchao Lou, Xiaoqing Wang, Fan Chen, Yixiao Chen, Bo Jiang, Hua Zhang(2014) created a new method for the prediction of the DNA-binding proteins, by using random forest and the wrapper-based feature selection. A method introduced called DBPPred, made use of Naïve Bayes because it gave better outcomes than the other algorithms tested. . (Wangchao Lou, Xiaoqing Wang, Fan Chen, Yixiao Chen, Bo Jiang, Hua Zhang, 2014). Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof.N.B.Venkateswarlu(2011) applied different classification algorithms on the data set, and a result was obtained which projected the better performing algorithm. The performance of these algorithms were measured based on several criteria such as precision. With the selected dataset, it was known that KNN, Back propagation and SVM gave better results than the other algorithms (Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof. N. B. Venkateswarlu, 2011).

3. Proposed System

Classification groups data into different classes based on some particular constraints. The performance of Random Forest algorithm and the Naïve Bayes algorithm on the *Indian Liver Patient Data Set* is evaluated with the focus to identify the algorithm giving more accurate value and also better efficiency for Liver disease prediction.

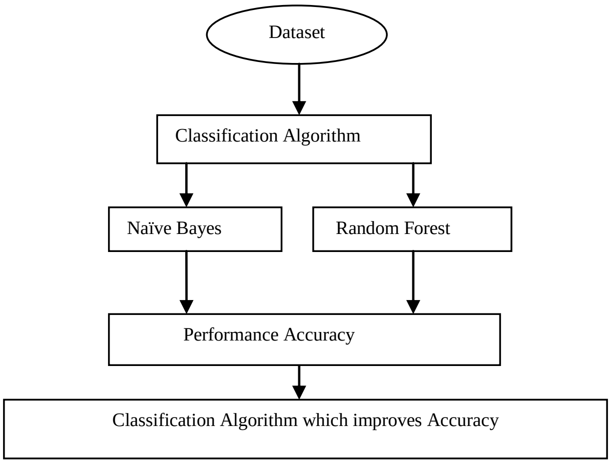


Figure 1 : Overview of Proposed System

Random Forest Algorithm

Several decision trees together constitute a Random Forest. However, using Random Forest it is possible to overcome several problems that exist with normal decision trees such as reduction in over fitting and less variance. Over fitting affects the accuracy when predicting samples that are not part of the training set. As a result, in almost all cases, random forests are more accurate than decision trees. Random forest is a method in which a classifier is constructed by combining several different independent base classifiers and this is one of the main advantages of this algorithm. In short, it can be said that it is best to use the Random Forest algorithm in cases with a high number of variables and having huge sample size.

Consider an example to understand the working of Random Forest algorithm. A bank is sending out a notice for the recruitment of officers. The recruitment process involves 4 rounds – Preliminary test, Main Test, Interview 1, and Interview 2. Only those who clear the preliminary test are allowed to attend the Mains, and those who clear the Mains get selected for the Interview process. Each of the interviews is headed by separate, independent panels. The procedure each panel follows to assess candidates is unique. There is a certain level of Randomness that is met here. Generally it can be said that a group of people makes better decisions than individuals. Each interview panel consists of topics that are diverse in nature, and it is ensured that none of the questions are repeated in corresponding levels of the interview process. After all the rounds are completed, the decision whether to select a candidate depends on maximum votes received from each individual panel. If maximum votes suggest that a candidate should be hired, then the bank goes ahead with selecting the candidate. This is the concept of a random forest.

The following pseudo code can be used to perform prediction using Random Forest algorithm:

1. The feature set is taken and the rules of each individually created decision tree are used to predict and store the outcome or target.
2. Votes are evaluated for each outcome.
3. The target with the highest vote is considered as the final prediction of the Random Forest algorithm.

Naïve Bayes Algorithm

The Bayes Theorem is one theorem with extremely high levels of applicability in varying fields, and the Naïve Bayes is an algorithm directly applying the principles of this theorem. It specifies the outcome of belongingness of records to their corresponding classes. The class with the highest probability will be selected as the most likely class. The formula used is: $P (H | E) = P (E | H) x P (H) / P (E)$.

Why Random Forest?

This paper proposes that the Random Forest algorithm applied on the Indian Liver Patient Dataset will provide better results with greater accuracy than when the Naïve Bayes algorithm was applied on the same dataset. There are several advantages in using Random Forest over Naïve Bayes:

- ILPD is a large dataset, and Random Forest works way more efficiently on larger databases than smaller ones.
- Huge number of input variables can be handled without the deletion of variables.
- An estimation of the important variables in a classification is given.
- A proper estimate of the generalization errors is generated.
- Far more effective in cases handling large number of missing data, and provides better accuracy.
- Once a particular forest is generated, it can be saved and used for other types of data.
- Random Forest can be built easily.
- Involves the concept of parallel execution.

In their paper, Nghia Nguyen, Brandon King and Anand Subramanian(2013), studied several examples and their accuracy were tested on both the chosen algorithms, in order to arrive at the conclusion that Random Forest can be a better option in these cases. (Nghia Nguyen, Brandon King, Anand Subramanian, 2013).

Case 1: Census Income Prediction

This case study focuses on forecasting the annual income of Los Angeles and Long Beach for 3 different years–1970, 1980, 1990, based on data that is acquired and recorded systematically. The result shows if a person gets a salary which is less than or more than 50000. There are 40 features which were considered here:

age, class, industry code, race, sex, labor union, capital gains, marital status, state of residence, weeks worked per year, etc.

Table 1: The Table Consists of: Training: 199523 Samples, Test: 99762 Samples

	Random Forest	Naïve Bayes
Accuracy:	.952	.827

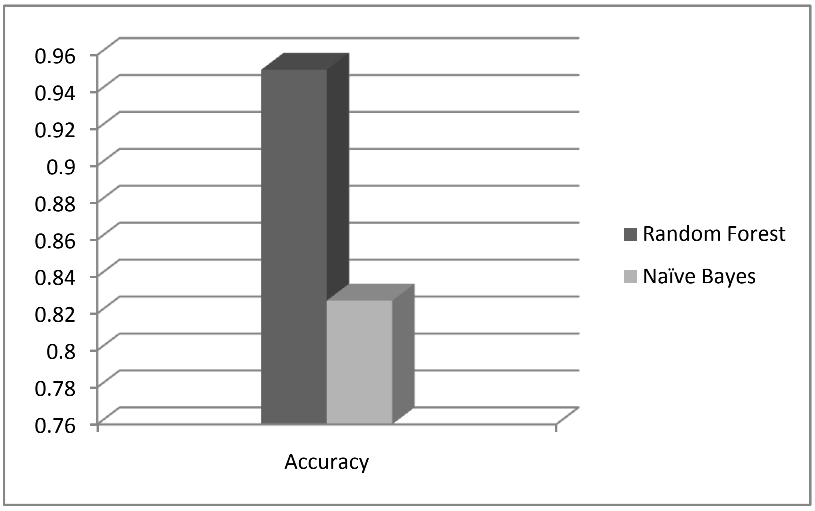


Figure 2: Accuracy Rates for Census Income Prediction

Case 2: ISOLET - Isolated Letter Speech Recognition

150 speakers saying each letter of the alphabet twice. Hence there will be 52 classes. Each of these 150 speakers was divided into groups of 30 and each group was observed separately.

617 Features: Continuous Linguistic attributes (sonorant, contours, and spectral coefficients)

26 Classes: Letters of the alphabet

Table 2: The Table Consists of: Training: 6238 Instances (120 speakers), Test: 1559 Instances (30 speakers)

	Random Forest	Naïve Bayes
Accuracy:	.940	.767

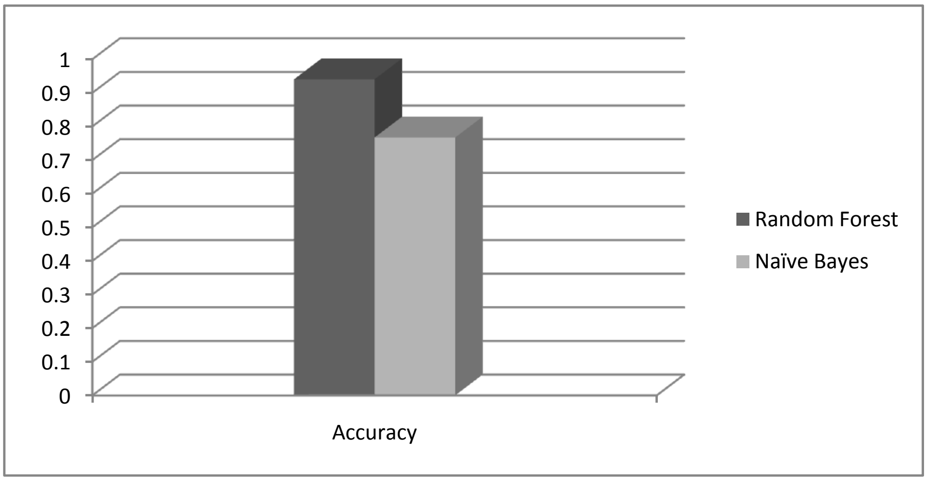


Figure 3: Accuracy Rates for ISOLET

Case 3: Letter Image Recognition

Identifying a letter from images based on their primitive features.

16 Features: Image characteristics such as width, height, # of pixels, etc.

26 Classes: Letters of the alphabet

Table 3: The Table Consists of: Training: 15000 Samples, Test: 5000 Samples

	Random Forest	Naïve Bayes
Accuracy:	.959	.725

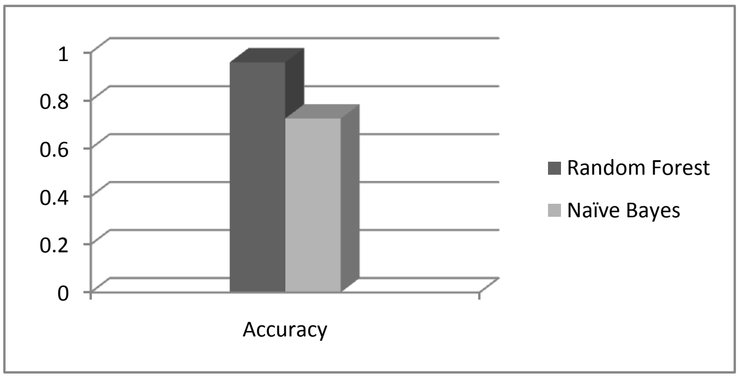
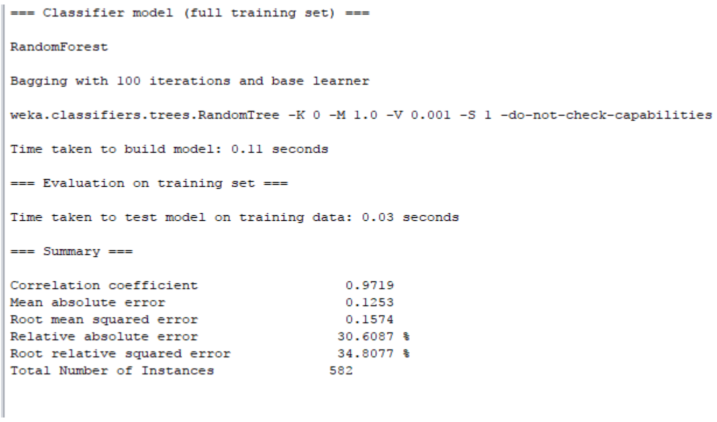


Figure 4: Accuracy Rates for Letter Image Recognition

All these case studies, classifying different types of data arrive at a common conclusion that Random Forest algorithm provides more accuracy than Naive Bayes algorithm. Random Forest best performs with huge data sets. This is one of the major advantages of this algorithm, and hence automatically is the best option while dealing with big data. The above examples showcases that RF can be used with varying types of data sets. But since it deals with huge data sets, the running time will be a little more as compared to Naïve Bayes. But the accuracy the algorithm provides covers this slight drawback.

4. Experimental Result

Figure 5 gives the result obtained when a certain sample of the original dataset was put into WEKA tool and classified using the Random Forest algorithm.



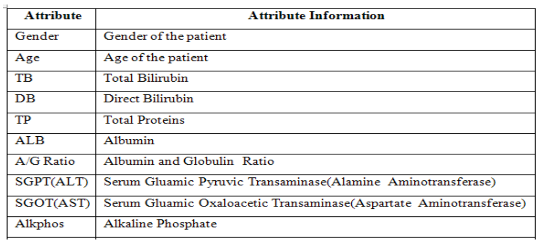
```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.11 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correlation coefficient                  0.9719
Mean absolute error                      0.1253
Root mean squared error                  0.1574
Relative absolute error                 30.6087 %
Root relative squared error             34.8077 %
Total Number of Instances              582
```

Figure 5: ILPD when Put into WEKA Tool for Random Forest

It was found that the error rates while using the RF algorithm was also less, even for this particular dataset with big number of instances, which in turn shows that the number of correctly classified instances are also more. Hence we can predict that the accuracy will also be more when the entire dataset is evaluated.

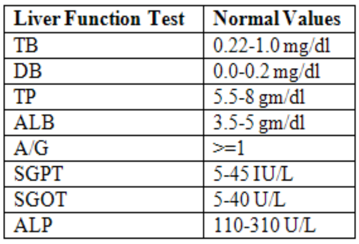
Dataset Description

Figure 6 comprises of the attributes used for prediction and also the normal values for each attribute. Each test result provides an insight on the existence of a particular disease.



Attribute	Attribute Information
Gender	Gender of the patient
Age	Age of the patient
TB	Total Bilirubin
DB	Direct Bilirubin
TP	Total Proteins
ALB	Albumin
A/G Ratio	Albumin and Globulin Ratio
SGPT(ALT)	Serum Gluamic Pyruvic Transaminase(Alamine Aminotransferase)
SGOT(AST)	Serum Gluamic Oxaloacetic Transaminase(Aspartate Aminotransferase)
Alkphos	Alkaline Phosphate

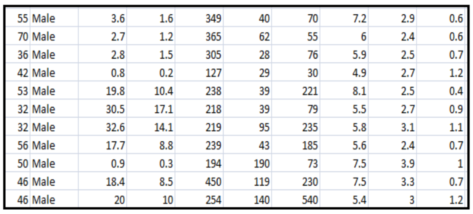
Figure 6: Attribute Description of ILPD



Liver Function Test	Normal Values
TB	0.22-1.0 mg/dl
DB	0.0-0.2 mg/dl
TP	5.5-8 gm/dl
ALB	3.5-5 gm/dl
A/G	>=1
SGPT	5-45 IU/L
SGOT	5-40 U/L
ALP	110-310 U/L

Figure 7: Normal Values of Attributes

Figure 8 provides a sample 11 rows from the ILPD data set, with columns showing Age, Gender, Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, Albumin and Globulin Ratio, SGPT (ALT), SGOT (AST), Alkaline Phosphate. The values for each test is examined separately in order to arrive at a result regarding the presence of a particular disease.



55	Male	3.6	1.6	349	40	70	7.2	2.9	0.6
70	Male	2.7	1.2	365	62	55	6	2.4	0.6
36	Male	2.8	1.5	305	28	76	5.9	2.5	0.7
42	Male	0.8	0.2	127	29	30	4.9	2.7	1.2
53	Male	19.8	10.4	238	39	221	8.1	2.5	0.4
32	Male	30.5	17.1	218	39	79	5.5	2.7	0.9
32	Male	32.6	14.1	219	95	235	5.8	3.1	1.1
56	Male	17.7	8.8	239	43	185	5.6	2.4	0.7
50	Male	0.9	0.3	194	190	73	7.5	3.9	1
46	Male	18.4	8.5	450	119	230	7.5	3.3	0.7
46	Male	20	10	254	140	540	5.4	3	1.2

Figure 8: Sample Dataset of ILPD

5. Conclusion

The merits of Random Forest algorithm when applied on huge data sets is irreplaceable, and results show that when applied for multiple classes, there is no better algorithm than Random Forest. Consistent outputs combined with high accuracy makes this algorithm the best choice in areas demanding “to the point” results, such as the medical field. Since the algorithm deals with large amount of data sets divided into smaller sub trees, the processing time involved can be slightly high when compared to other classification algorithms, but the efficiency and accuracy the Random Forest provides outweigh this small setback. Thus, weighing the pros and cons of this algorithm, it can be concluded that for the prediction of liver diseases, Random Forest is found to be better than Naïve Bayes algorithm.

References

[1] Prasad Babu M.S., Swapna K., Balakrishna T., Venkateswarulu N.B., An Implementation of Hierarchical Clustering on Indian Liver Patient Dataset, IJETCAS (2014).

[2] Lou W., Wang X., Chen F., Chen Y., Jiang B., Zhang H., Sequence based prediction of DNA-binding proteins based on

hybrid feature selection using random forest and Gaussian naive Bayes, PLoS One 9(1) (2014).

[3] Nghia Nguyen, Brandon King, Anand Subramanian, Benchmarking Random Forest against Naive Bayes (2013).

[4] Ramana B.V., Babu M.P., Venkateswarlu N.B., A critical comparative study of liver patients from USA and INDIA: an exploratory analysis, International Journal of Computer Science Issues 9(2) (2012), 506-516.

[5] Ramana B.V., Babu M.S.P., Venkateswarlu, N.B., A critical study of selected classification algorithms for liver disease diagnosis, International Journal of Database Management Systems 3(2) (2011), 101-114.