IJURCA: International Journal of
Undergraduate Research & Creative Activities

Volume 12 Article 12

October 13, 2020

A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer

KENDALL LEMONS
Prairie View A&M University, klemons2@student.pvamu.edu

INDIKA WICKRAMASINGHE RATHNATHUNGALAGE
Prairie View A&M University, iprathnathungalage@pvamu.edu

Abstract

Accurate diagnosis of breast cancer is very beneficial as breast cancer is the second-leading cause of cancer death in women after lung cancer in the US. This study compares two machine learning approaches to diagnose breast cancer using a publicly available dataset, which comprises of features computed from a digitized image of a fine needle aspirate (FNA). We employ two different machine learning techniques, namely Naïve Bayes and Random Forest to measure the accuracy of the diagnosis. Using 569 patient's information and 31 features, the above three machine learning classifiers are implemented. According to the findings, the Random Forest classifier performed better than the Naïve Bayes method by reaching a 97.82% of accuracy. Furthermore, classification accuracy can be improved with the appropriate use of the feature selection technique. Furthermore, this section explains the feature selection technique used in the study. The analysis procedure is discussed, and the dataset and the performance indicators are described.

Keywords

Data Classification, Machine Learning, Supervised Learning, Statistics, Biomedical Informatics

Peer Review

This work has undergone a double-blind review by a minimum of two faculty members from institutions of higher learning from around the world. The faculty reviewers have expertise in disciplines closely related to those represented by this work. If possible, the work was also reviewed by undergraduates in collaboration with the faculty reviewers.

Acknowledgments

This study was supported by the Office of Research and Innovations at Prairie View A&M University.

1. Introduction

Breast cancer is considered the most common type of cancer among women throughout the world (World cancer report, 2008). Furthermore, it is estimated that 23 out of 124 women will die due to breast cancer annually (Cancer Statistics Review, 2012). Though Mammography, Fine Needle Aspiration (FNA), and surgical biopsy are the main techniques to diagnose breast cancer, FNA is considered as the most important diagnostic technique to detect breast cancer in early stages (Fiuzy et al., 2012). Further studies about FNA can be seen in (Fiuzy et al., 2012) and (Saxena and Burse, 2012). In this study we aim to utilize machine learning techniques to predict the accuracy of diagnosing breast cancer, using the Breast Cancer Wisconsin Data (Dua and Graff, 2019), which was collected using the FNA technique.

Machine learning is considered as a branch of artificial intelligence, which is considered as a method of data analysis that automates the model building process. Easy identification of patterns among the data and the ability to improve machine learning over time are two of the main advantages of machine learning over traditional data classification techniques.

This study aims to classify subjects, based on the characteristics of their breast biopsies into one of the two groups indicating whether the subject has cancer or not. According to the literature, Naïve Bayes and Random Forest techniques are two of the popular machine learning techniques. Therefore, in this study we employ both Naïve Bayes and Random Forest techniques to classify the above data. The rest of this manuscript is organized as follows. Section 2 discusses the two data classification algorithms, followed by the data analysis in section 3. In section 4, the results are discussed and the section 5 concludes the manuscript.

2. Methodology

In this manuscript, we implemented two machine learning techniques to classify data. They are Naïve Bayes (NB), and Random Forest (RF).

Consider the n-dimensional feature set $X=\left(x_1,x_2,...,x_n\right)$ and let $Y=(y_1,y_2);y_1=True,y_2=False$, be a 2-dimensional vector (classes). In this study, $n=31$ is the number of features.

2.1 Naïve Bayes (NB)

NB is considered as a simple and accurate classification algorithm. Due to the flexibility of the algorithm, a wide range of applications can be seen (Arar & Ayan, 2017). The NB stems from the well-known Bayes theorem, discussed in probability theory. According to the literature, NB is one of the top-performing classifiers used in data mining (Wu et al., 2008).

We aim to estimate the value of Y by maximizing $P\left(Y=y\mid X=x\right)$. NB assumes conditional independence among the features. Therefore, Y={True, False}

$$P(X_1=x_1,X_2=x_2,...,X_n=x_n\mid Y=y)$$
$$=P(X_1=x_1\mid Y=y).P(X_2=x_2\mid Y=y)....P(X_n=x_n\mid Y=y)$$
$$=\prod_{i=1}^{n}P(X_i=x_i\mid Y=y).$$

According to the Bayes theorem, we have $P\left(y\mid X\right)=\dfrac{P(X\mid y)P(y)}{P(X)}$.

Then we can write $P\left(y\mid X\right)$ as follows.

$$P\left(y \mid X\right) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$= \frac{P(Y=y)P(X=x \mid Y=y)}{P(X=x)}$$

$$= \frac{P(Y=y)\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)}{\prod_{i=1}^{n} P(X_i = x_i)}$$

$$\propto P(Y=y)\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)$$

Therefore, our aim is to find y such that the above expression is maximized. This means, we need to find y, which is

$$\arg\max_y P(Y-y)P(Y=y)\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)$$

2.2 Random Forest (RF)

Random Forest is an extension of the popular decision tree algorithm by introducing a higher number of decision trees. This approach aims to reduce the variance of the novel decision tree (Couronné, 2018). The construction of the decision tree is done by selecting a collection of random variables (features). Finally, such a collection of random trees is called a Random Forest, or RF for short. RF is considered as one of the most accurate classification algorithm, due to the higher classification accuracy (Breiman, 2001; Biau and Scornet, 2016). Another characteristic of RF is its significance for unbalanced and missing data (Shah et al., 2014), compared to other alternative techniques. Further experimental and theoretical activities of RF can be seen in Bernard et al. (2007), Breiman (2001).

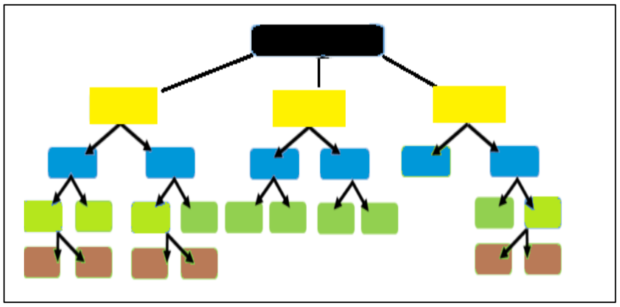

Figure 1: Random Forest Classifier

2.3 Feature Selection Technique

To improve the classification accuracy a feature selection technique was applied. The purpose of this is to reduce the dimension of the dataset. In other words, instead of considering all the variables (31 features) of the data, it attempts to filter the most important features that impact the classification. Here we selected a feature selection technique called, Recursive Feature Elimination Technique (RFE), which attempts to remove the most insignificant features from the data until the pre-specified number of significant features is reached. RFF is easy to configure and to handle due to its effectiveness to select features that have significant relationship to predict the target variable. When using the RFF, the elimination of insignificant features is done in a recursive way using the dependency and the correlation of the variables in the dataset.

3. Analysis

3.1 Dataset

This study was implemented on the Breast Cancer Wisconsin dataset, which was obtained from a publicly available source. The dataset consists of 569 patients data with 31 features. The class variable is the diagnosis of the breast tissues [Benign, Malignant]. The rest of the features have been computed from a digitized image of a process called, fine needle aspirate (FNA) of a breast mass. These features consist of characteristics such as radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension and fractal dimension.

3.2 Quantifying the Performance

To quantify the accuracy of classification, we use sensitivity, specificity, and precision as the performance indicators. All of these performance indicators are based on the confusion matrix, which represents the two states of actual and the predicted.

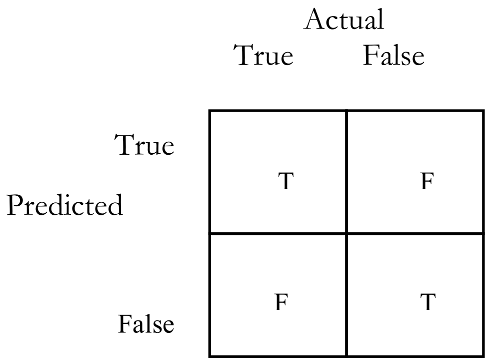

Figure 2: Confusion Matrix

$$Sensitivity = \frac{TP}{TP+FN},$$
$$Specificity = \frac{TN}{TN+FP},$$
$$Precision = \frac{TP}{TP+FP}$$

Here, TP - true positives (true instances predicted correctly), FP - false positives (false instances predicted as true), TN - true negatives (false instances predicted correctly), |N| -the total of true instances and |P| - total of false instances in the testing sample. Furthermore, we define F-measure, which is also called the harmonic mean between the precision and the sensitivity.

$$F\ measure = 2\left(\frac{Precision \times Sensitivity}{Precision + \ Sensitivity}\right)$$

This F-measure represents the weighted average of the two quantities of precision and sensitivity. The maximum value of F-measure is 1, while the minimum value is 0. Another two measures one can use to evaluate classifications method are Classification Accuracy and Error Rate. They are defined as follows.

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN}$$

The use of various performance indicators will bring more insight to interpret the accuracy of the data prediction. For instance, precision is the proportion of related information out of all the retrieved information. This is a valuable indicator in almost all applications. Sensitivity measures the true-positive recognition rate. This becomes very useful where there is a high importance of classifying positives such as in security checking. In contrast, specificity measures the rate of actual negatives and it is useful in areas such as diagnosing health conditions prior to treatments.

4. Results

According to the table 1 and 2, both algorithms perform better with the introduction of the feature selection algorithm. Though specificity and precision show the opposite relationship with NB after introducing the feature selection algorithm, RF shows significant progress with the implementation with the feature selection algorithm. Comparing both of the feature selection algorithms, it is clear that RF outperforms NB.

Table 1: Algorithm, Feature Selection Technique, Sensitivity, Specificity, and Precision

Algorithm	Feature Selection	Sensitivity	Specificity	Precision
NB	No	0.9634	0.9425	0.9009
	Yes	0.9748	0.9405	0.8962
RF	No	0.9722	0.9722	0.9521
	Yes	0.9763	0.9802	0.9655

Table 2: Algorithm, Feature Selection Technique, F-measure, Classification Accuracy, and Error Rate

Algorithm	Feature Selection	F-Measure	Classification Accuracy	Error Rate
NB	No	0.9574	0.9455	0.0545
	Yes	0.9529	0.9402	0.0598
RF	No	0.9722	0.9648	0.0352
	Yes	0.9782	0.9724	0.0277

Further implementation of the feature selection algorithm with RF indicates the improvement of the classification accuracy. According to the figure 3, the highest classification accuracy (0.9782) of RF is recorded when using 17 features out of the 31 entire features.

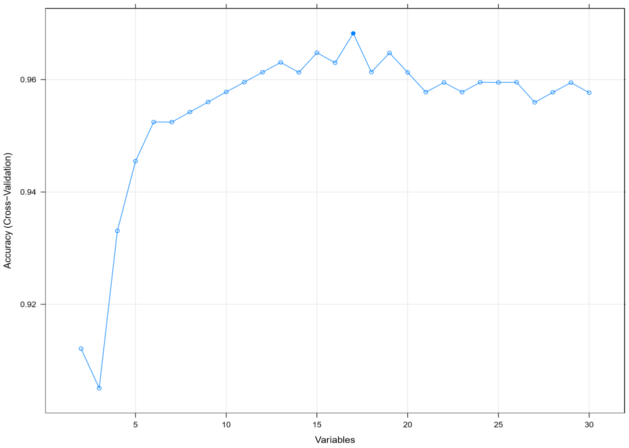

Figure 3: Accuracy vs features using RFE

According to the figure 3, the classification accuracy with RF increases with the increment of the number of features. This goes until number of features is 13. The highest accuracy level is reached with 17 features and the accuracy level becomes stable after considering 20 features.

5. Discussion

At present, the impact of machine learning has reached the majority of areas. When considering the healthcare industry, this influence is immense. The skillset that can be assumed from machine learning by health care professional such as physicians have taken healthcare to a different height. In this manuscript we aimed to use the knowledge of machine learning to diagnose a cancer, based on the characteristics of the biopsy taken from the breast using the Fine Needle Aspiration (FNA) technique. After employing two machine learning algorithms, Naïve Bayes and Random Forest we found that both techniques can be used to classify cancer patients effectively. Out of these two techniques, Random Forest outperformed the Naïve Bayes and the predicting accuracy can be further improved with appropriate selection of feature set. According to the experimental data, the highest accuracy of 97.82% was reached with the Random Forest by selecting only 17 features from a total of 31. In a future study, it is important to consider other possible parameters involved with the classification technique to improve the classification accuracy further. Furthermore, these techniques can be effectively used in other areas of data classification applications.

Furthermore, when choosing a machine learning algorithm for data classification one needs to think about aspects of the execution time and the simplicity of the model. With NB, the model is very simple, fast at the execution, lower risk of overfitting the data, and higher accuracy with categorical data compared to the numerical data. Unfortunately, the issues with probability calculations and assumption of independency with predictors are some of the limitations of the NB model. On the other hand, RF has less variability in the prediction due to the selection of multiple trees, and it can handle a higher volume of data effectively. Two of the notable drawbacks with RF are the complexity of the model and the higher possibility to be over fitted. It is recommended to tune its hyper-parameters involved to minimize the impact of overfitting.

References

Arar, Ö. F., & Ayan, K. (2017). A feature dependent naive Bayes approach and its application to the software defect prediction problem. Applied Soft Computing, 59, 197– 209.

Biau, G. and Scornet, E. (2016). A random forest guided tour. Test 25 197–227.

Breiman, L. (2001). Random forests. Machine Learning 45 5–32.

Couronné, R., Probst, P. & Boulesteix, A. Random forest versus logistic regression: a large- scale benchmark experiment. BMC Bioinformatics 19, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fiuzy, M., Haddadnia, J., Mollania, N., Hashemian, M., & Hassanpour, K. (2012). Introduction of a New Diagnostic Method for Breast Cancer Based on Fine Needle Aspiration (FNA) Test Data and Combining Intelligent Systems. Iranian journal of cancer prevention, 5(4), 169–177.

IARC. World cancer report: International agency for research on cancer. Lyon, 2008

NCI. SEER: Cancer Statistics Review. 2012 Machine Leaning Approach and the advantages of it over conventional methods

Shweta Saxena, Kavita Burse: A Survey on Neural Net-work Techniques for Classification of Breast Cancer Data. International Journal of Engineering and Advanced Technology, 2012

Simon Bernard, Laurent Heutte, Sébastien Adam. Using Random Forests for Handwritten Digit Recognition. 9th IAPR/IEEE International Conference on Document Analysis and Recognition (ICDAR), Sep 2007, Curitiba, Brazil. pp.1043-1047, ff10.1109/ICDAR.2007.4377074ff. ffhal-00436372f

Wu, X., Kumar,V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., et al., "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, vol. 14, no. 1, pp. 1- 37, 2008.