Article ID: #30678087. Received: August 24, 2021. Revised: September 3, 2021. Accepted: December 24, 2021

Journal of Bioinformatics and Neuroscience (JBINS), 7 (2): 352-354, December 31, 2021
e-ISSN: 2432-5422, p-ISSN: 2188-8116
Applied Science and Computer Science Publications

Effect of the random forest with recursive feature elimination for breast cancer classification using a WDBC dataset

Yoshihiro Mitani[1] and Yuki Ono[1]

National Institute of Technology, Ube College
Department of Intelligent System Engineering,
Japan[1]
Email: mitani@ube-k.ac.jp

Abstract: A breast cancer is the most dangerous disease of the death cause among aged 40-55 women. We need a computer aided diagnosis system for breast cancer classification. In the previous study, the random forest which is known as an ensemble learning method was reported to be one of promising classifiers for classifying breast cancers using a Wisconsin Diagnostic Breast Cancer(WDBC) dataset. This paper presents the effect of the random forest with a recursive feature elimination for breast cancer classification on the WDBC dataset, compared to the state of the art ensemble learning techniques, such as XGBoost and LightGBM.

Keywords: Wisconsin Diagnostic Breast Cancer(WDBC) dataset; Breast cancer classification; Random forests; XGBoost; LightGBM; Feature selection; Recursive feature elimination

I. INTRODUCTION

A breast cancer is the most dangerous disease of the death cause among aged 40-55 women[1]. Therefore, in order to detect breast cancer early, designing a computer aided diagnosis system for breast cancer classification has been desired. In this study, we used a Wisconsin Diagnostic Breast Cancer(WDBC) Data Set[2]. This data consists of 212 malignant and 357 benign, totally 569 samples. This data has 30 features, which have 10 attributes, radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, and three types of statistics, mean, error, and worst. Note that "error" and "worst" mean the standard deviation and the average of top 3 sizes of each attribute in descending order, respectively. In the previous study[1], the random forest[3] which is known as an ensemble learning method was reported to be one of promising classifiers for classifying breast cancers on the WDBC dataset. According to the results[1], the random forest was the best among SVM(linear), k-NN(k=3), Naive Bayes, J48, random forest, and multilayer perceptron, in terms of the accuracy. In this study, we focus on using a random forest classifier for breast cancer classification. Furthermore, we examine the effect of the random forest with a recursive feature elimination(RFE)[4] for breast cancer classification, compared to the state of the art ensemble learning techniques, such as XGBoost[5] and LightGBM(LGBM)[6]. The experimental results showed the effectiveness of the random forest with the RFE for breast cancer classification on the WDBC dataset.

II. METHODOLOGY

A. Random forest classifier

First, we describe random forests[3]. From the previous study[1], we found that the random forest classifier was the best among several classifiers to classify breast cancers on the WDBC dataset. Therefore we focused on the random forest for breast cancer classification using the WDBC dataset. The random

Corresponding Author: Yoshihiro Mitani. Department of Intelligent System Engineering, National Institute of Technology, Ube College, Japan, 2-14-1 Tokiwadai, Yamaguchi Prefecture, Ube City, Japan, E-mail: mitani@ube-k.ac.jp

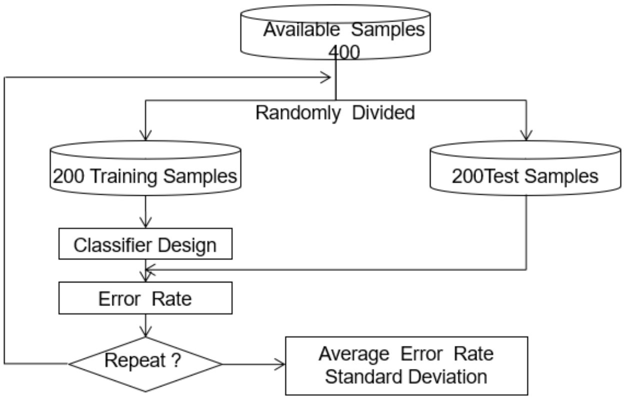

Fig. 1 Error rate estimation

Table 1 Average error rate for the number of the decision trees of the random forest

10	50	100	200
5.92	5.06	4.82	4.87
1.26	1.49	1.32	1.46

Upper: average error rate(%), Lower: standard deviation

Table 2 Average error rate of each classifier when the feature size is 30

RF	XGBoost	LGBM
4.82	5.34	4.92
1.32	1.93	1.68

Upper: average error rate(%), Lower: standard deviation

Table 3 Average error rate of each classifier with the RFE when the feature size is 15

RF	XGBoost	LGBM
4.66	5.16	4.92
1.50	1.54	1.36

Upper: average error rate(%), Lower: standard deviation

forest is one of the ensemble training approaches. The random forest classifies a class based on the multiple outputs performed by each decision tree[3]. A decision tree itself causes overtraining easily. Therefore, the random forest used by multiple decision trees is expected avoid overtraining. The more numbers of the decision trees, the more robust the classification performance of the random forest is. Depending on the pattern classification problem, we could set the number of the decision trees of the random forest. Considering not only improving the classification performance but also reducing the design cost, we would like to develop the random forest classifier. Thus, we should try choosing the smaller optimal number of the decision trees in terms of the average error rate.

B. Recursive feature elimination

Second, we show the feature selection of the random forests. One possibility of the feature selection is to reduce a classifier design cost, keeping the average error rate low. The random forest estimates the importance of each individual feature. In order to reduce the number of the features, the RFE technique with the random forest[4] is explored in this study. The RFE algorithm recursively eliminates the variables using permutation importance measure as a ranking criterion[4].

C. XGBoost and LGBM

Furthermore, there are the state of the art classification techniques such as XGBoost[5] and LGBM[6]. These are also known as ensemble learning method alike the random forest. The ensemble learning is based on decision trees. Unlike a single decision tree, the random forest uses multiple decision trees as bagging. This relates to a parallel learning processing. On the other hand, the XGBoost and LGBM use multiple decision trees as boosting. This learning serially processes. In general, the cost of training the XGBoost is relatively high. On the other hand, the learning cost of the LGBM aims to be cheaper than that of the XGBoost. In the experiments, we examine the classification performance of the random forest, XGBoost, and LGBM. We also compare the average error rate of the random forest with the RFE, as well as the XGBoost and LGBM.

III. RESULTS

In the experiments, we used 400 samples: 200 malignant samples and 200 benign samples from the WDBC dataset. In order to evaluate the breast cancer classification performance of the random forest, we estimate the average error rate by a holdout method[7]. The holdout method is known for maintaining the statistical independency between the training and test samples. This means the results are statistically reliable. The error rate is computed as a ratio of the number of test samples misclassified to the number of all test samples. Fig. 1 shows the error rate estimation. First, we randomly divided 400 available samples into 200 training samples and 200 test samples, respectively. Second, we trained the random forest classifier with 200 training samples. Next, using the random forest classifier, we computed the error rate with 200 test samples. By repeating 100 times, we calculated the average error rate and its standard deviation.

Firstly, we examined the influence of the number of the decision trees of the random forest for breast cancer classification. Table 1 shows the average error rate for the number of the decision trees of the random forest. In the table, the upper and the lower show the average error rate(%) and the standard deviation, respectively. From the table 1, the minimum average error rate showed when the number of the decision trees is 100. The average error rate was 4.82%. When it comes to 200 or more numbers of decision trees, not so much the error rate improves than we expected. In the following experiments, we used this random forest with 100 decision trees.

Secondly, we examined the classification performance by the random forest, XGBoost, and LGBM, in terms of the error rate. Table 2 shows the average error rate for each classifier when the feature size is 30. This means we used all the features provided by the WDBC dataset. We set to some parameter values from the preliminary experiments. The result of the random forest showed when the number of the decision trees is 100, as I mentioned above. In the XGBoost, the number of the boost rounds was 100. From the result of the table 2, the random forest is slightly superior to the XGBoost and LGBM classifiers.

Finally, we examined the effect of the feature selection by the random forest, XGBoost, and LGBM with the RFE. We focused on the average error rate when the number of the features is 15. This figure is exactly a half of 30 which is a full of features of the WDBC dataset. Table 3 shows the average error rate for each classifier with the RFE when the feature size is 15. From the result of the table 3, the random forest slightly outperforms the XGBoost and LGBM classifiers.

Considering both experimental results of tables 2 and 3, we see the random forest with the RFE is the best, in terms of the average error rate. The minimum average error rate showed 4.66%.

IV. DISCUSSION

This paper presented a positive effect of the random forest with the RFE for breast cancer classification on a WDBC dataset, compared to the state of the art ensemble learning approaches, such as XGBoost and LGBM. From the results of our limited experiments for breast cancer classification, the random forest classifier with the RFE was slightly superior to or comparable with the XGBoost and LGBM with the RFE. Perhaps the use of the random forests could be better if a computer was capable of parallel computing.

In the pattern recognition field, it is well known that the richer features are, the better the classification performance of a classifier is. Therefore, we should try using richer features to design a pattern recognition system. In the future study, apart from the feature selection of the RFE, we will further compare to another type of feature selection or feature extraction methods. Alternatively, we will try creating richer features of using existing features for breast cancer classification. We would like to try tackling another type of breast cancer data. Furthermore, we should also try addressing other types of cancer.

REFERENCES

[1] Saygili, A. Classification and diagnostic prediction of breast cancers via different classifiers. International Scientific and Vocational Journal, vol. 2, issue 2, 45-56. 2018.

[2] UCI Machine Learning Repository. Breast Cancer Wisconsin(Diagnostic) Data Set. 1992.

[3] Breiman, L. Random forests. Machine Learning, 45, 5-32. 2001.

[4] Gregorutti, B., Michel, B., and Saint-Pierre, P. Correlation and variable importance in random forests. Statistics and Computing, 27, 659-678. 2017.

[5] Chen, T. and Guestin, C. XGBoost: A scalable tree boosting system. Machine Learning, 785–794. 2016. https://arxiv.org/abs/1603.02754.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. Proc. 31st Int. Conf. Neural Information Processing System, 3149-3157. 2017.

[7] Fukunaga, K. Introduction to statistical pattern recognition, Second Edition, Academic Press. 1990.