RESEARCH ARTICLE | NOVEMBER 04 2019

Ovarian cancer data classification using bagging and random forest FREE

A. Arfiani; Z. Rustam

AIP Conf. Proc. 2168, 020046 (2019)
https://doi.org/10.1063/1.5132473

Ovarian Cancer Data Classification Using Bagging and Random Forest

A. Arfiani and Z. Rustam[a)]

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

[a)]Corresponding author: rustam@ui.ac.id

Abstract. Ovarian cancer is the fifth most common cause of cancer deaths in women worldwide. Most cases of ovarian cancer occur in women entering menopause or in age of 50 years onwards. One step to reducing mortality from ovarian cancer is timely detection and effective treatment. Accurate and efficient method is needed for the gaining insight on ovarian cancer, particularly in classification of benign or malignant, as the focus of this paper. There have been many ways used to classify ovarian cancer including machine learning methods. In this paper, we proposed the machine learning method, namely Bagging and Random Forest for the classification into the benign or malignant of ovarian cancer. Bagging method is known to maximize classification and prevent overfitting. Whereas, the Random Forest can produce low errors, and is an effective method for estimating missing data. We use microarray data obtained from UCI Machine Learning Repository downloaded on September 2018. Simulations on training data were carried out with various percentage. In each simulation, accuracy and running time were calculated. The final score of experimental result confirmed that bagging reached 100 % accuracy for 90 % training data, while the Random Forest achieved an accuracy of 98.2 % for 90 % training data.

Keywords: Bootstrap aggregating, microarray datasets, random forest

INTRODUCTION

Cancer is one of the main diseases that cause death in the world [1]. Cancer can attack all parts of the body [2]. If it attacks the ovary, it is called ovarian cancer. Ovarian cancer ranks fifth as a cause of death in women worldwide [3]. If a cancer cell is found are found in one part of the ovary, then the other part of the ovary must be examined in detail [4]. Ovarian cancer is often not detected in the early stages until it spreads to the pelvis and abdomen. At an advanced stage several symptoms appear, but may not specifically, such as loss of appetite and weight loss. In the final stage, ovarian cancer will be more difficult to treat and can cause death. Surgery and chemotherapy are usually used for the treatment of ovarian cancer [5].

Rapid detection is needed to determine whether a patient has ovarian cancer or not. When it is known for certain that a patient has ovarian cancer, a proper treatment is required. There are many techniques that have been carried out and used for ovarian cancer detection and seek for appropriate treatment recommendation. One of them includes the use of computational techniques. In this paper, we proposed are Bagging and Random Forest to identify whether a patient has ovarian cancer or not. We used microarray data from hospital laboratory test results.

Microarray data is a type of medical data that is used because in the human body, the expression of genes can be numerically shown through microarray data. In machine learning method, the machine will study the gene patterns of microarray patient data to classify other microarray data [2]. There are two characteristics of microarray data [6],

Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018)
AIP Conf. Proc. 2168, 020046-1–020046-6; https://doi.org/10.1063/1.5132473
Published by AIP Publishing. 978-0-7354-1915-5/$30.00

that is, consists of many features and small sample size. In microarray data, feature is defined by the genes of the patient's body. Data analyzed in this paper consisted of 15154 features from sample of size 266.

The bagging method proposed in this paper is one of the latest intensive and successful international methods to improve unstable classification or classification schemes [7]. Introduced by Breiman [8], the method objective is to improve classification accuracy by randomly generated training sets. Preventing overfitting and reducing variance is an advantage of the bagging method. Although usually applied to decision tree methods, bagging can be applied to other machine learning methods.

Based on a decision tree and combined with aggregation and bootstrap, random forest method was introduced in 2001 by Breiman [9]. This method can handle missing values, scalable to high dimensional data, and maximize the prediction accuracy through variance reduction [10]. Previous studies have shown bagging performance and its application, namely random forest, but it was found that the random forest results were superior to bagging. Beside bagging and random forest method, the classification of cancers that have been reviewed previously uses many other machine learning methods such as the Support Vector Machine (SVM) method [2], Clinical and Gene Expression Integrative Approach [3], Fuzzy C-Means with Feature Selection [11], Normed Kernel Function-Based Fuzzy Possibilistic C-Means (NKFPCM) [12], Markov Blanket-Embedded Genetic Algorithm [13], and Binary Black Hole Algorithm [14]. This paper used ovarian cancer Arcene data obtained from the UCI Machine Learning Repository [15]. We used 10-fold-CV and the level of accuracy and running time are considerations of our comparison.

METHODS

The data was obtained from UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Arcene, downloaded on September 2018. Let the data be denoted by $L_i = (\boldsymbol{X}_i Y_i)(i = 1, ..., n)$ with Y_i response of real value and $\boldsymbol{X}_i$ p-dimensional explanatory variable for the i-instance. In this paper, Y_i represents a target variable, and $\boldsymbol{X}_i$ represents the predictor variable, p represents the number of the dimension to p, instance refers to the amount of the observation. The aim of the data analysis is to determine an observation of a patient ovarian cancer entering into benign or malignant. This is conducted using the methods of bagging and random forest, which are presented in the subsequent section.

Bagging

Bagging is one of the newest and most successful computational methods to improve unstable classification or classification processes. This is very useful for large data if it is difficult to determine the model and classify data in one step because large data has a complex level of complexity. Bagging was first introduced by Breiman in 1994 to reduce the difference of the characteristic of the predictor variable. The final prediction is obtained through averaging (for regression case), or voting (for classification case) from each of these basic models. This has attracted a lot of attention and is often applied in classification problems, although lacking in terms of deep theoretical insight. Theoretically, the provisions of bagging are as follows [7]:

I. Empirically from distribution $L_i = (\boldsymbol{X}_i Y_i)(i = 1, ..., n)$, make a sample of bootstrap randomly $L_i^* = (\boldsymbol{X}_i^* Y_i^*)(i = 1, ..., n)$.

II. With the principle of plug-in; which $\hat{\theta}_n^*(x) = (L_1^*, ..., L_n^*)(x)$, where $\hat{\theta}_n(x) = (L_1, ..., L_n)(x)$ is the samples selected, compute the predictor of bootstrap $\hat{\theta}_n^*(x)$.

III. $E^*[\hat{\theta}_n^*(x)]$ is the predictor.

Breiman [8] explained the heuristic bagging performance as follows. The bagged estimator variance $\hat{\theta}_n$ is equal to or less than the original estimator $\hat{\theta}_n(x)$. The bias value is almost the same for the bagged and the original estimator. This implies that, at approximately the same time for a "stable" scheme, bagging improves significantly in the mean squared error for the "unstable" predictors. Breiman [8] defines heuristic instability as the predictor is "unstable" if a small change in data can cause a large change in the predicted value. The bagging algorithm [8] is shown in Fig. 1.

Pre-state: A training set $L := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the ensemble Z = $\emptyset$, and T number of trees in bagging to train.

function Bagging (Z, L)
  for $a \leftarrow 1$ to T do
    $G_a \leftarrow A$ bootstrap sample of L
    Using G_a as the training set, construct a classifier Z_a
    To the current ensemble, $Z \leftarrow Z \cup Z_a$, add the classifier
  end for
  return Z
  Run $Z_1, \ldots, Z_T$ on the input x
  As a label for x, the class with the maximum number of votes is selected
end function

FIGURE 1. Algorithm of the Pseudocode of Bagging

Random Forest

Random forest is an ensemble decision tree and bagging method developed by Breiman in 2001 [9]. This method has a good performance in terms of classification accuracy because each randomized decision tree is built based on samples taken randomly with replacement from training data. In the feature selection process as a node of a randomized decision tree, the selected feature is no longer the best feature of all features, but the best feature of the randomly selected feature set.

Notated with L= $\{(x_1,y_1), ..., (x_n,y_{1n})\}$ a set of learning independent observations of random vector ($\boldsymbol{X}$,Y). $\boldsymbol{X} \in R^p$ is a vector of predictor of $Y \in \mathcal{Y}$, which Y is a label class for classification. A classifier s is a mapping s: $R^p \rightarrow \mathcal{Y}$ [16].

By definition, a random forest classifier contains a structured tree classifier $\{h(x,\theta_k), k = 1, ...,\}$ where $\{\theta_k\}$ are independent and identically distributed random vectors that are distributed separately and each tree gives the unit vote for the most popular class on input **x**. The random forest algorithm can be seen in Fig. 2. [9]:

Pre-state: A training set $L := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, M features, and T number of trees in forest.

function Random Forest (L, M)
  $Z \leftarrow \emptyset$
  for $a \in 1, \ldots, T\ do$
    $L^{(a)} \leftarrow A$ bootstrap sample from L
    $r_a \leftarrow$ RandomizedTreeLearn ($L^{(a)}, M$)
    $Z \leftarrow Z \cup \{r_a\}$
  end for
  return Z
end function
function RandomizedTreeLearn (L, M)
  On each node:
    $m \leftarrow$ the smallest subset of M
    From m, split the best feature
  return the learned tree
end function

FIGURE 2. The algorithm of pseudocode of the random forest.

APPLICATION OF CLASSIFICATION ON OVARIAN CANCER DATA

Data Overview

The domain dataset is a microarray obtained from the UCI Machine Learning Repository [15]. This dataset is from the National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS) [17] which contains 266 samples with 15154 features.

RESULTS AND DISCUSSION

In this section, accuracy and running time of classification with bagging and random forest are calculated. In the accuracy formula there are four possible outcomes: true positive (TP) representing number of correctly classified positive class data, false negative (FN) representing number of positive class data that was incorrectly classified into negative class, false positive (FP) representing negative class data that was incorrectly classified into positive class, and true negative (TN) representing number of correctly classified negative class . The accuracy is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \times 100\ \%$$

Program simulation for each method is varied by the percentage of training data, ranging from 10 % to 90 % of the sample data. The results are presented in Table 1. It can be seen in Table 1 that the bagging method produced an excellent level of accuracy for the classification of ovarian cancer data which is 90 % training data and with accuracy 100 % with the running time 0.164187 second because of the data could be separated linearly. Whereas, the lowest accuracy value is in 60 % training data with the final accuracy value of 99.1 % with the running time 1.844037 s.

The result on ovarian cancer data classification using the random forest method can be seen in Table 2. It can be seen in Table 2 that the random forest method produced the best level of accuracy for the classification of ovarian cancer data which is 90 % training data and with an accuracy of 98.2 % with the running time 0.127821 s. Whereas, the lowest accuracy result is at 60 % training data with accuracy results of 92.9 % with the running time 0.123972 s.

Figure 3 and Fig. 4 showing the accuracy and running time of ovarian cancer classification with Bagging and Random Forest. According to Fig. 3, the accuracy results indicate that the bagging method produced the accuracy that tends to be stable but still higher than the random forest, then it is better than the random forest method. Figure 4 shows the running time results from the two methods proposed. Random forest required a shorter running time than bagging.

TABLE 1. Accuracy and running time of ovarian cancer data classification using bagging.

Percentage of Training Data (%)	Accuracy (%)	Running Time (s)
10	99.6	12.812923
20	99.6	9.04004
30	99.3	6.352451
40	99.4	4.7184
50	99.3	3.047705
60	99.1	1.844037
70	99.4	1.0574
80	100	0.412931
90	100	0.164187

TABLE 2. Accuracy and running time of ovarian cancer data classification using random forest.

Percentage of Training Data (%)	Accuracy (%)	Running Time (s)
10	94.1	0.126535
20	93.4	0.111563
30	93.7	0.134389
40	93.0	0.121941
50	93.0	0.114
60	92.9	0.123972
70	96.5	0.128472
80	95.6	0.132737
90	98.2	0.127821

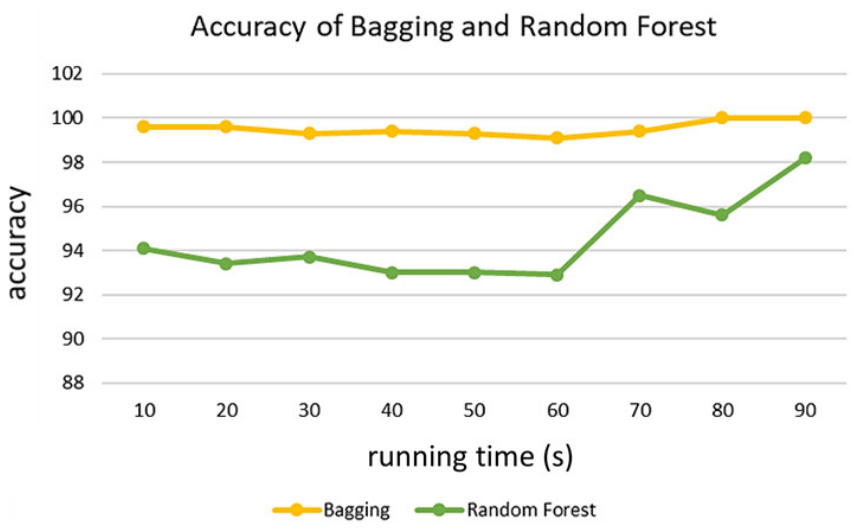

FIGURE. 3 Graph of accuracy ovarian cancer classification using Bagging and Random Forest.

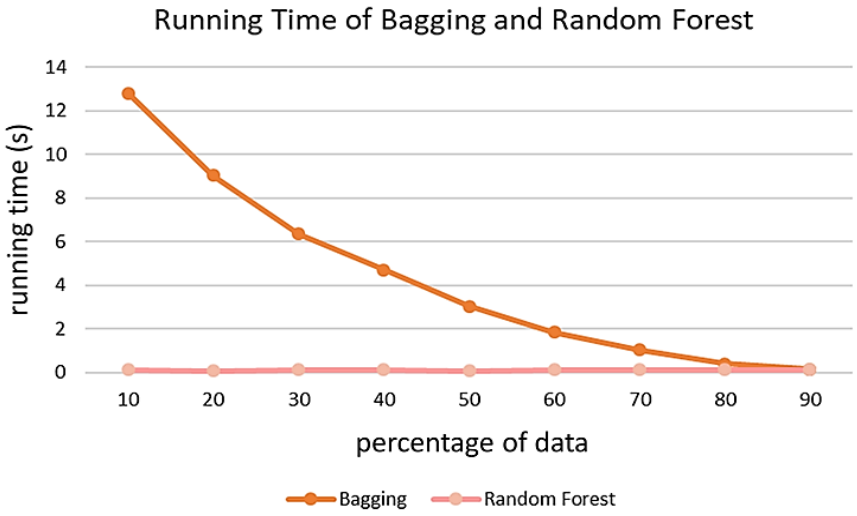

FIGURE. 4 Graph of running time ovarian cancer classification using Bagging and Random Forest.

CONCLUSION

In this paper, we proposed a bagging and random forest based on the work of Breiman method to classify a subject benign or malignant ovarian cancer. Both methods produced high accuracy for classification problems in large data, as the ovarian cancer data used in this paper, with bagging method reached 100 % for 90 % training data and random forest reached 98.2 % of accuracy for 90 % training data. The result showed that the bagging method is able to improve the classification process, producing high accuracy. However, in the context of running time, the random forest method has a relatively shorter computation time than bagging with a computational time difference of 0.036366 s. Overall it can be concluded that the bagging method is better than random forest to classify ovarian cancer data.

ACKNOWLEDGMENTS

This work supported by the Ministry of Research and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PDUPT 2019 research grant scheme, ID number 1621/UN2.R3.1/HKP05.00/2019. We thank to all reviewers for the improvement of this article.

REFERENCES

1. World Health Organization, *10 Facts About Cancer* (2018), available at https://www.who.int/features/factfiles/cancer/en/.
2. V. Panca and Z. Rustam, AIP Conf. Proc. **1862**, 030133 (2017).
3. A. E. Nabawy, N. E. Bendary and N. A. Belal, Procedia Comput. Sci. **131**, 23-30 (2018).
4. National Cancer Institute, *What is Cancer?* (2015), available at http://www.cancer.gov/about-cancer/understanding/what-is-cancer.
5. Alodokter, *Kanker Ovarium* (2018), available at https://www.alodokter.com/kanker-ovarium.html
6. Y. Wang et al., Comp. Biol. Chem. **29**, 37-46 (2015).
7. P. Bühlmann and B. Yu, Ann. Statist. **30**, 927-961 (2002).
8. L. Breiman, Technical Report No. 421 1994-09 (1994) available at https://www.stat.berkeley.edu/~breiman/bagging.pdf
9. L. Breiman, Machine Learning **45**, 5-32 (2001).
10. Y. Qi, *Random Forest for Bioinformatics* (Springer, Berlin, 2012).
11. A. A. Rachman and Z. Rustam, "Cancer Classification using Fuzzy C-Means with Feature Selection", *12th International Conference on Mathematics, Statistics, and Their Application,* Banda Aceh, Indonesia, 2016.
12. A. W. Lestari and Z. Rustam, AIP Conf. Proc. **1862**, 030143 (2017).
13. Z. Zhu, Y. S. Ong and M. Dash, Pattern Recognition, **40**, 3236-3248 (2007).
14. E. Pashaei and N. Aydin, Appl. Soft. Comput. **56**, 94-106 (2017).
15. UCI Machine Learning Repository, *Arcene Data Set* (2018) available at https://archive.ics.uci.edu/ml/datasets/Arcene.
16. R. Genuer, J. M. Poggi, C. T. Malot and N. V. Vialaneix, Big Data Research **9**, 28-46 (2017).
17. E. F. Petrocoin III et al., Lancet **359**, 572-577 (2002).