Weighted Random Forest Algorithm Based on Bayesian Algorithm

Xue Zhang¹ and Mei Wang¹, *

¹School of Computer and Information Technology, Northeast Petroleum University, Daqing 103318, China

*Corresponding author: Mei Wang: wangmei@nepu.edu.cn

Abstract. The random forest(RF) algorithm is a very efficient and excellent ensemble classification algorithm. In this paper, we improve the random forest algorithm and propose an algorithm called ‘Bayesian Weighted Random Forest’(B-RF), focus on the problem that inaccurate decision tree classification caused by the same voting weights in the traditional random forest model. The main idea underlying the proposed model is to replace the supernmajority voting of random forests into weighted voting, fully consider the difference of classification ability of each decision tree, using the Bayesian formula to dynamically update the weight value for each tree, so that the strong classifier has higher voting power and effectively improves the overall performance of classification. Through the verification of UCI database, the results show that the classification accuracy of the proposed weighted random forest model is higher. This illustrate the outperformance of the proposed model in comparison with the general random forest algorithm.

1. Introduction

With the development of the Internet and big data, the richness and coverage of data have gone far beyond the scope of manual observation and summary, and there are various classification problems[1]. Classification methods have been widely studied, but the accuracy of traditional classification methods is limited and the scope of application is narrow. There are many algorithms to solve the classification problem, such as naïve Bayes algorithm, support vector machine algorithm, decision tree algorithm. However, when using a single classifier for classification prediction, it is easy to suffer from overfitting problems. In order to improve the prediction accuracy, multiple classifier models are gathered together, which is called ensemble learning algorithm. The Bagging algorithm is a technique to reduce generalization error by combining several models. The most representative is the random forest algorithm, which was established by Leo Breiman from the University of California in 2001 as an ensemble learning method in the field of machine learning that combines bootstrap aggregating theory and stochastic subspace method with machine learning algorithm [1-3]. Random forest algorithm has good performance, which can deal with both classification and digital features. Because the proposed random forest algorithm needs fewer parameters to choose, the implementation of random forest classifier is relatively simple. It can also process high-dimensional data without feature selection, and the generalization error can be estimated internally without deviation, which has good generalization ability. Among the commonly used machine learning algorithms, random forest algorithm is widely used in cryptography [4], bioinformatics [5], ecology [6] and other fields.

However, with the widely application of the random forest, also gradually exposed some problems
of the algorithm in random forest classification. In recent years, scholars at home and abroad have put forward many different methods to improve the stochastic forest in order to improve the prediction accuracy. Literature[7] two optimization processes are added on the basis of the traditional random forest algorithm. The classification accuracy of the random forest model can be improved by abandoning the decision trees with those bad and selecting those with low correlation. However, it is easy to screen only the decision trees with good classification for a certain category by mistake, thus affecting the final classification result of the category. Literature Literature[8-9] adopted the out-of-bag accuracy rate as the decision tree voting weight. But the diversity of out-of-bag samples affects the fairness of the vote. Based on the above problems, the Bayesian formula is used in this paper to comprehensively consider the correct prediction probability of each tree as the weight of the decision tree, and the difference between strong classifier and weak classifier is fully considered in voting according to the principle of weighted voting, so as to improve the accuracy after integration.

The study is arranged as follows. In Section 2, we give a short description of main approaches to random forest algorithm and the main idea of B-RF and the method to calculate the optimal weight of the classifier. In Section 3 and 4, we assess the algorithm we proposed against RF on 4 data sets from UCI repository and discusses our practical study. Concluding remarks are provided in Section 5.

2. Bayes-based Weighted Random Forest

2.1. Construction of a traditional random forest

The traditional random forest algorithm takes the classification decision tree without pruning constructed by CART algorithm as the base classifier, and combines Bagging and random feature selection to increase the diversity of decision tree models[10]. Its principle is, first of all, from the original sample set using the Bootstrap method to extract the training set, and then training on each training set a decision tree model, the decision tree of the independent identically distributed into a forest, for a new input sample and the all the decision tree is used to determine the ownership of the forest classification, and then to an absolute majority vote on the decision tree selection mechanism to determine the final classification results[11].

It is assumed that the given data set is\( D = \{X_i, Y_i\} , X_i \in R^K , Y_i \in \{1,2,...,C\} \),and the random forest is a combined classifier obtained after integrated learning based on N decision trees on this data set. CART algorithm adopts binary partition method, recursively bisects each feature, thus the feature space is divided into finite units, and the predicted probability distribution is determined on these units. CART algorithm uses Gini coefficient to select features [12]. Assuming that given data set \( D \) has \( K \) categories and the number of \( K \) categories is \( C_k \), then the coefficient of data set \( D \) is

\[
Gini(D) = 1 - \sum_{k=1}^{K} \left( \frac{C_k}{D} \right)^2
\]

(1)

based on \( N \) training data sets, a corresponding number of decision trees were trained, and \( T_1,T_2,...,T_N \) decision trees were integrated to obtain the random forest model. The number of decision trees is set and the integrated decision trees constitute a random forest. Integrated learning method is a method to make joint decision on the same problem after combining different classifiers to improve classification accuracy [13]. The voting method used in the random forest algorithm is simple voting method, also known as majority voting method. Each decision tree is given the same weight, \( N \) decision trees vote on each sample point according to their own training diagnosis results, and the final classification result with the most votes is the classification result of this sample point.

In terms of classification, each decision tree in the random forest will predict which classification the latest data belongs to. The predict function \( f(x) \) is the classification has been selected the most which classification the latest data belongs to, i.e.,

\[
f(x) = \arg\max \sum_{a=1}^{n} T_a(x) = y, y = 1,2...c
\]

(2)
The simple voting method does not consider the difference between the decision tree with high classification accuracy and the decision tree with low classification accuracy, and gives the same weight to the base learner, which will reduce the prediction accuracy of the random forest after integration. Therefore, improving the random forest by changing the weight is a better way to improve the prediction accuracy.

2.2. Weighted voting decision based on Bayes

Aiming at the above problems, this paper improves the integration method of random forest algorithm. In order to improve the prediction accuracy of the random forest algorithm, the base learner with high prediction accuracy are given higher weights, and the base learner with low prediction accuracy are given smaller weights to reduce their influence on the prediction results. Bayesian formula is a very important knowledge in statistics. Its application is characterized by the combination of prior probability and practical results to obtain more accurate posterior probability[14]. It is based on the conditional probability to find the cause of the event. For a given training data set, the posterior probability is estimated as accurately as possible based on the limited sample set. The Bayes formula is determined through the prior probability and conditional probability as

$$P(y|l) = \frac{P(l|y)P(y)}{P(l)}$$

where $P(y)$ is the category marker $y$ class prior probability. $P(l|y)$ is the conditional probability of the class of sample $l$ relative to category marker $y$. $P(l)$ is the class-prior probability of the sample $l$.

In 2014, Kuncheva proposed a relationship between classifier weight and prediction accuracy for classification problems[15], assuming a group of classifiers $S_1, S_2, ..., S_n$, the corresponding prediction accuracy rate is corresponding to $P_1, P_2, ..., P_n$, the relationship between the weight of each classifier and the prediction accuracy is

$$\omega_l \propto \frac{P}{1-P}$$

The Equation (4) shows that when the prediction accuracy of a single decision tree is smaller, its weight value will be relatively small, so the influence of base learners on the final prediction result of random forest will also be reduced. On the contrary, the influence of basic learners on the prediction results will also increase.

2.3. Weighted Random Forest

Aiming at the shortcomings of random forest model in decision tree integration and predict accuracy, the optimization was carried out by constructing the weighted voting method. In this paper, the above theory is applied to evaluate the performance of a single decision tree in the random forest. Bayes' value of prediction is used as an evaluation index of the prediction accuracy of the base learner. It is a percentage value, and the corresponding weight can be assigned to the base learner accordingly and the results are normalized as the weight of weighted voting.

Process of weighted random forest classification algorithm is as follows.

1. Draw $N$ bootstrap samples from the original data. Take the bootstrap samples as the training sample of CART decision tree and repeat $N$ times to generate $N$ decision trees to form a random forest.

2. The prediction accuracy is calculated by using Bayesian formula, and all the predicted values of a set of data are averaged as the prediction accuracy of each tree.
\[ acc = \frac{1}{S} \sum_{j=1}^{S} \frac{P(l_j | y) P(y)}{P(l)} \]  

(3) The weight of each CART decision tree in the random forest can be calculated by

\[ \omega_i = \ln \frac{acc}{1 - acc} \]  

(6)

(4) Assuming that \( N \) decision trees in random forest are constructed, the prediction output function is expressed as

\[ H(x) = \arg\max_{\sum_{i=1}^{N} \omega_i h_i(x)} \]  

subject to \( \omega_i \geq 0, \sum_{i=1}^{N} \omega_i = 1 \).

3. Experiments
The motivation of this paper is to improve the prediction performance of random forest. Several experiments are adopted to assess the worth of the algorithm performance our proposed weighted random forest algorithm in different data classification. The experiments, comparing the classification accuracy of the proposed algorithm with that of the improved algorithm under different data sets, illustrates the ascendancy of the algorithm we proposed.

In order to carry out the comparisons, the experimental data set used in this study is 4 public data sets with different application backgrounds in UCI[16]. Table 1 gives the sample information.

| S. no. | Datasets | Instances | Number of classes(k) | Number of classes(d) | Size of classes |
|-------|----------|-----------|----------------------|----------------------|----------------|
| 1     | Iris     | 150       | 3                    | 4                    | 50,50,50       |
| 2     | Wine     | 178       | 3                    | 13                   | 59,71,48       |
| 3     | Sonar    | 208       | 2                    | 60                   | 97,111         |
| 4     | Ionosphere | 351     | 2                    | 34                   | 126,225        |

Experiments are performed using 70\% of the data sets as the training sets and the classification accuracy and running time of each algorithm under different data sets was used as the experimental evaluation method. In the experiment, the number of decision trees to construct the random forest model is 100 and 200. We repeated the process 10 times, the average of classification accuracy of each algorithm is using as accuracy evaluation method and the average of running times of each algorithm is using as efficiency evaluation method of the experiments.

4. Results and Discussion

4.1. Comparisons at data with different classifiers
Since the B-RF can be viewed as an improvement of the original RF, then our interest in this study is to compare the weighted RF and the original RF. In order to verify the effectiveness of the algorithm, experiments were carried out in Windows 10 and Python 3.7 programming environments. The average of classification accuracy of each algorithm are shown in table 2.
### Table 2. Comparison of classification accuracy at data.

| Datasets   | Number of classifiers | RF       | B-RF*    |
|------------|-----------------------|----------|----------|
| Iris       | 100                   | 0.9311111| 0.9488891|
|            | 200                   | 0.9266667| 0.9422224|
| Wine       | 100                   | 0.9537036| 0.9759257|
|            | 200                   | 0.9629628| 0.9777775|
| Sonar      | 100                   | 0.7428574| 0.8079367|
|            | 200                   | 0.7047622| 0.7904764|
| Ionosphere | 100                   | 0.8924526| 0.9122639|
|            | 200                   | 0.8999997| 0.9160374|

The average of running time of each algorithm are shown in Table 3.

### Table 3. Algorithm running time comparison.

| Datasets   | Number of classifiers | RF       | B-RF*    |
|------------|-----------------------|----------|----------|
| Iris       | 100                   | 5.1175743| 5.3623307|
|            | 200                   | 5.0556058| 5.731324 |
| Wine       | 100                   | 5.2102217| 5.7533174|
|            | 200                   | 5.7125326| 6.1198982|
| Sonar      | 100                   | 5.3929139| 5.5776089|
|            | 200                   | 5.9532001| 6.4164364|
| Ionosphere | 100                   | 5.2841777| 5.6012982|
|            | 200                   | 5.8007777| 6.0363544|

In this experiment, four groups of data sets were used for classification verification. It can be seen from Table 2 that the average prediction accuracy of B-RF for four different data sets is higher than that of random forest algorithm, which fully demonstrates the superiority of this algorithm. As can be seen from Table 3, the average running time of the B-RF algorithm proposed in this paper is slightly increased compared with the original random forest algorithm due to the use of Bayesian algorithm for weight calculation.

### 4.2. Discussion

In the previous section, we compared the prediction accuracy and running times of the proposed weighted random forest algorithm with that of the RF algorithm. In different data sets, the algorithm proposed in this study performs best in the accuracy of prediction results. Therefore, it can be said that the weighted random forest method in this study is suitable for the classification of many types of classification data. However, the study has its drawbacks. First, Bayesian method is used to calculate weights for individual classifiers, which is beneficial to improve the prediction performance of random forest algorithm on data, but also increases the computational complexity of the method. In addition, when the integration scale is large, the weighted voting method needs to estimate multiple weights, which is easy to cause overfitting. Further improvement is needed to solve the above problems.

### 5. Conclusion

In this paper, a weighted voting based random forest (B-RF) method is proposed to improve the random forest algorithm. In the classification model, the difference of the classification ability of each decision tree is fully considered by using Bayesian method, and the voting mechanism of the classifier in bagging ensemble is improved. The weighted voting strategy is used to obtain the final recognition result, which ensures the fairness of the vote and effectively improves the recognition accuracy. The experimental results of different examples show that this method is superior to the traditional random forest method.
Acknowledgments
This work is supported in part by the National Natural Science Foundation of China under Grant 51774090.

References
[1] BREIMAN L 2001 Random forests. *Machine learning*. 45 1 5-32.
[2] BREIMAN L 1996 Bagging predictors. *Machine learning*. 24 2 123–140.
[3] HOT 1998 The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*. 20 8 832-844.
[4] Huang L T, Zhao Z C, Zhao Y Q 2018 A two-stage cryptosystem recognition scheme based on random forest. *Chinese Journal of Computers*. 41 2 382-399.
[5] XU B, GUO X, YE Y 2012 An improved random forest classifier for text categorization. *Chinese Journal of Computers*. 35 5 680-683.
[6] Chen X 2020 Application and extension of Bayes' formula. *Heilongjiang Science*. 11 12 38-39.
[7] Ludmila I. Kuncheva, Juan J. Rodríguez 2014 A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*. 38 2 259-275.
[8] Frank A, Asuncion A 2010 UCI Machine Learning Repository /http://archive.ics.uci.edu/mlS.