Classifications Based Decision Tree and Random Forests for Fanjing Mountains’ Tea

Lina Yuan*, Huajun Chen and Jing Gong
College of Data Science, Tongren University, Tongren, China.

*Corresponding author e-mail: 893422817@qq.com

Abstract. Fanjing Mountains’ tea is made from the mountains of Fanjing, which contains green tea, black tea, white tea, black tea, oolong tea, instant tea and other series, and green tea is mainly to play. Thanks to the natural ecological protection of Fanjing Mountains, and the classic exquisite processing technology and profound cultural deposits, it has formed the unique quality of Fanjing Mountains’ tea. This article takes the idea of classification to analyze the impact factors for Fanjing Mountains’ tea, such as climate, rainfall, geographical position, and etc., and simulates by matlab with the methods of decision tree classification and random forest classification in machine learning. The simulation results show that the latter is better and more accurate.

1. Introduction
At present, Tongren has more than 1.4 million mu of tea area, which is related to 7 counties (i.e., Shiqian, Yinjiang, Songtao, Jiangkou, Sinan, Dejiang, and Yanhe), and 115 towns and villages. The scale of tea garden is the first in the whole province. In the first selection activities about "the tea and tea garden of Chinese ecological civilization", Tongren was awarded the title of the first "the tea of Chinese ecological civilization", and ecological tea garden of Fanjing Mountains won the name of the first "the tea garden of Chinese ecological civilization ". By 2020, Tongren will produce 2 million mu of tea garden, 200 thousand tons of tea production, and more than 10 billion Yuan of annual output value, which will be built into the production base with a national green, organic, safe and healthy tea. Fanjing Mountain tea is produced in gold industry belt of China green tea, i.e., Wuling mountain peak-Fanjing Mountain, which is the best preserved green treasure house of the earth and latitudes, a member of "protection area network of human and biosphere circle" in the United Nations, and a national nature reserve. Fanjing Mountain and its surrounding area have mild climate, abundant rainfall, high mountains with thick fog, fresh air, and protruding microclimate, tea plantations is located in the region between 800-1500 meters above sea level, and its produced Fanjing Mountains’ tea has excellent quality. Series products about Fanjing Mountains’ tea won more than 60 awards in the "green cup", "national drink cup", " tea cup in Guizhou”, and other appraisal activities of famous tea. So how to make the tea really become industry of enriching people and counties, it is necessary to further study the production of the tea is closely related to which environmental factors, and then according to the results using various effective measures to put the tea become bigger and stronger, and promote the construction process of agricultural industrialization in the counties.
The organized structure of this article is as follows. Section II introduces the idea of classification. The basic and process principle of decision tree, and cross validation are expounded in Section III. In Section IV, Bootstrap sampling, Bagging algorithm and random forest algorithm are described. Section V uses MATLAB to do simulations with comparing the two methods. At last, the full text is summarized in Section VI.

2. Classification

Classification is to realize the classification of things, which reflects the positioning of data from "quantity" to "quality". To achieve this goal, it needs to firstly accumulate the ability to classify categories. In terms of tasks about data analysis, from a large amount of data training sets through certain methods of classification modeling, discovery and summary laws, establish a reasonable and effective classification model, and then achieve classification goals according to these models. In order to improve the accuracy, effectiveness and scalability of the classification process, the data set can be pretreated before data sets are to be classified modeling. Specifically, when a typical classification system is actually established, the classification process can be divided into five stages, as shown in Fig. 1.

![Figure 1. A typical system of classification.](image)

In Fig. 1, the first stage of data acquisition uses to collect the necessary data information of establishing an effective classifier, which is the source and foundation of the whole system. The second stage of feature (attribute) selection selects the obviously distinguishing and meaning features, according to the nature of the specific problem domain, which is a crucial step in the establishment process of classification system. The third stage is model selection, which is the same as the second stage that is a key link of classification system. The former of good or bad will directly affect the availability and effectiveness of the classifier, which needs the support of prior knowledge to judge the methods of classifying models. The fourth phase of training uses the selected classifying model to establish the process of classifiers on the collected data sets. The last phase is evaluation, which comprehensively evaluates the performance of classifiers, based on the evaluating index of a certain classifier.

The above five stages constitute a closed feedback system, which has well performance of classification by continuous feedback, adjustment, and established classification systems. However, due to the complexity and variability of the application, not all established classification systems have good classification performance, and not all problems can be found in a fine classification model.

The evaluation and comparison criteria of classification methods are mainly composed of the following five indexes. First, accuracy of prediction: this involves the ability of the models to correctly predict the class label of new or previously unseen data. Second, speed: this refers to the calculation cost of generating and using the model. Third, robustness: this involves the ability of models to correctly predict a given noise data or data with a vacancy value. Fourth, scalability: this refers to the ability to effectively construct models with a large number of data. Last, interpretability: this involves the level of understanding and insight provided by the learning model. In addition, the contribution to the classification and prediction of data mining in the database research community has been always emphasized the index of scalability, especially for the method of decision tree.

The computation complexity of classification task continuously reduces from the root node to the terminal node, while the guarantee of classification accuracy needs powerful means of feature
selection and detection. Because of a series of sub-decision in the decision tree, as long as there are emerging problems of a sub-decision, it will cause the errors of the whole decision-making.

3. Decision Tree and Cross Validation

3.1. Decision Tree.

The decision tree (DT) is an instance-based learning method, which deduces the classification rules of decision tree’s representation with a view to a set of unordered and un-ruled examples. It adopts the top-down recursion, compares the values of properties in the interior nodes of the decision tree, judges with the different attribute values, grows branches from the node down, and comes to a conclusion in the leaf nodes of decision tree as shown in Fig. 2. So a path from the root to the leaf node corresponds to a rule of conjunction, and the whole decision tree is homologous to a set of disjunctive expression rules. Based on learning algorithms of decision tree, one of the biggest advantages is that users don’t need to know a lot of background knowledge in the process of learning, which is also the biggest drawback. As long as the training instances can be expresses by the way of attributes-conclusion, this algorithm can be used to learn. The internal nodes, called the testing attribute, of a decision tree are attributes or a set of attributes. The leaf nodes are the divided class needed. When a decision tree is generated through training a set of training examples, the decision tree can classify an unknown instance set according to the value of the attribute. Using decision tree to classify instances, beginning from the root to gradually test its value for the object’s properties, walk down the branches, until to reach certain leaf node, whose represented class is the class in which the object is located.

Decision tree is a classification technology in very common use, whose working process can be simply described as shown in Fig. 3. In numerous classification methods of decision tree, their fundamental difference lies mainly in different decision-making strategies when generating decision trees, that is, the different of the decision tree’s classification algorithms.
3.2. Cross Validation.
The computed loss from validating sets is sensitive to the choice of validating data. If validating sets are small, it is more difficult. Cross-validation is one method of effectively using existing data sets. It can be seen from Fig. 4 that $K$-fold cross-validation divides data sets into the same size $K$ (or as equal as possible). Each takes turns to act as validating set, and the other $K-1$ as training sets. Thus, the result is that the $K$ loss values are used as the final loss values. The extreme case of $K$-fold cross-validation is when $K = N$, i.e., $K$ exactly equals to the number of observed data in data sets, and each observed data is taken as testing the models obtained by the other $N-1$ objects of training in turn. This special form of cross-validation is called as Leave-One-Out Cross Validation (LOOCV). The mean square validation of LOOCV is

$$L^CV = \frac{1}{N} \sum_{n=1}^{N} \left( t_n - \hat{\theta} x_n \right)^2$$

Thereinto, $\hat{\theta}$ is the parameter estimation of removing the $n$-th training instance.

![Figure 4. Cross-validation. Data sets are as shown the left side of pie chart, and in each K fold, one set of the data points is shift out of the training sets to validate or test models.](image)

4. Random Forest

4.1. Bootstrap Sampling.
Suppose the set $S$ containing $n$ different samples $\{x_1, x_2, \cdots, x_n\}$. If subtract a sample from $S$ each time with replacement, do $n$ time in total to form a new set $S^*$, then the probability of not including certain sample $x_i (i = 1, 2, \cdots, n)$ in $S^*$ is $p = \left(1 - \frac{1}{n}\right)^n$. When $n \to \infty$,

$$\lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368$$

Therefore, although the total number of samples both in the new set $S^*$ and in the original set $S$ equals to $n$, $S^*$ may contain the repeated samples (subtract with replacement). If except the repeated samples, $S^*$ only includes about $1 \times 0.368 = 63.2\%$ samples in the original set $S$.

4.2. Bagging Algorithm.
The abbreviation of boot strap aggregating algorithm is bagging algorithm, which is the first integration algorithm with its flow chart as shown in Fig. 5, and whose specific steps can be
described as follows.

First, use the method of Bootstrap to re-sample, and randomly produce $T$ training sets $S_1, S_2, \ldots, S_T$. Second, create the corresponding decision trees $C_1, C_2, \ldots, C_T$ by each training set. Third, for the sample of testing set $X$, using each decision tree to test, obtain the corresponding category $C_1(X), C_2(X), \ldots, C_T(X)$. Last, through the method of voting, the most class output in $T$ decision tree is regarded as the subordinate category of the testing set sample $X$.

![Figure 5. The flow chart of bagging algorithm.](image)

### 4.3 Random Forest Algorithm

Random forest (RF) algorithm is similar to Bagging, which re-samples based on the method of Bootstrap to produce multi-training set. The difference is that the former adopts to select the way of splitting the property set at random, when building the decision trees. The detailed flow about random forest algorithm is in the following.

Suppose the number of the properties of samples is $M$, and $m$ is an integer that is greater than zero and less than $M$. First, use the method of Bootstrap to re-sample, and randomly produce $T$ training sets $S_1, S_2, \ldots, S_T$. Second, create the corresponding decision trees $C_1, C_2, \ldots, C_T$ by each training set; before selecting property on each non-leaf node (internal node), randomly subtract $m$ from $M$ properties as the current node of splitting property set, and split this node by the best splitting way in the $m$ properties (In general, the value of $m$ holds constant throughout the growth process of the forest). Third, every tree grows without pruning. Fourth, for the sample of testing set $X$, using each decision tree to test, obtain the corresponding category $C_1(X), C_2(X), \ldots, C_T(X)$. Last, through the method of voting, the most class output in $T$ decision tree is regarded as the subordinate category of the testing set sample $X$.

### 5. Simulation

Assuming that randomly generate data of Fanjing Mountains’ Tea in the following simulations in this paper. The purpose of our simulations is to classify the Fanjing Mountains’ Tea by means of decision tree and random forest, and compare the affection of two methods. The detailed steps of simulations are as follows. First, clear all environmental variables. Second, load the generated data. Third, create training sets/testing sets at random. Fourth, training data. Fifth, testing data. For the classifications of decision tree and random forest, the above five steps are the same, and next describe them respectively.

On one hand, for decision tree, firstly, set up the classification of decision tree, and see its view as shown in Fig. 6(a). Secondly, simulate and test. Thirdly, using cross validating, the minimum number of samples contained in leaf nodes effects on performance of decision tree as shown in Fig. 7. Fourthly, produce the optimal DT by setting the min-leaf is 13 as shown in Fig. 6(b). Lastly, prune as shown in Fig. 6(c).

On the other hand, for random forest, firstly, set up the classification of random fores. Secondly, simulate and test. Thirdly, plot as shown in Fig. 8. Fourthly, the number of DTs
effects on performance in RF as shown in Fig. 9 (first, for each condition, run 100 times and average their value; then, create random forest; last, simulate, test, and plot).

(a) See the view of DT.

(b) Produce the optimal DT by setting the min-leaf is 13.

(c) View the DT after pruning.

**Figure 6.** Various created DTs.

**Figure 7.** The minimum number of samples contained in leaf nodes effects on performance of DT.
6. Summary
Through the simulation results, compared with the classification of decision tree, that of random forest has more advantages, such as less adjusted parameters, faster classification speed, more efficient processing of large sample data, more significance of estimating which characteristic in classification, stronger ability to carry noise, and so on. Therefore, this method is widely used in various fields, for example, feature recognition, biomedicine, remote sensing, the classification of radar signals, and has made excellent results. In our future research, the classification of random forest will be further applied in Fanjing Mountains’ Tea, which reduces the influence on output from the impacts of the outside environment, in order to accelerate the construction process of agricultural industrialization in the counties.

Acknowledgments
This work was financially supported by the project of platform talent in Guizhou in 2016 (Project Number: [2016] 5611), education and cooperation for talent team word in Guizhou in 2015 (Project Number: [2015] 67), the Collaborative Fund Project of Science and Technology Agency in Guizhou
Province Marked by the word LH on 7480 [2014], and partly supported by the National Natural Science Foundation of China (NO. 61562703).

References
[1] Joseph Siryani, Bereket Tanju, Timothy J. Eveleigh. A Machine Learning Decision-Support System Improves the Internet of Things’ Smart Meter Operations. IEEE Internet of Things Journal. Vol. 4, No. 4, p. 1056-1066. (2017)
[2] Sebastian Buschjäger, Katharina Morik. Decision Tree and Random Forest Implementations for Fast Filtering of Sensor Data. IEEE Transactions on Circuits and Systems I: Regular Papers. Vol. 65, No. 1, p. 209-222. (2018)
[3] Z.H.Zhou. Machine Learning. Tsinghua University Press, (2016).
[4] Stephen Marsland. Machine Learning, an Algorithmic Perspective, Second Edition. CRC Press, (2014).
[5] Kyriacos Shiarlis, Joao Messias, and Shimon Whiteson. Rapidly Exploring Learning Trees. 2017 IEEE International Conference on Robotics and Automation (ICRA), (2017), May 29-June 3, Singapore, p. 1541-1548.
[6] Markus Endler, Jean-Pierre Briot, Francisco Silva e Silva, Vitor P. de Almeida, Edward H. Haeusler. Towards stream-based reasoning and machine learning for IoT applications, 2017 Intelligent Systems Conference (IntelliSys), (2017), Sept. 7-8, United Kingdom, p. 202-209.