

Agricultural Environmental Monitoring Data Classification of Fanjing Mountains: A Decision Tree Approach

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 is rich in wildlife resources (such as Guizhou golden monkey, dove tree, and other rare species) and tea resources (for example, Fanjing Mountains cui feng tea, Shiqian moss tea, Vatican black tea, Jin Qiangong black tea, Eight hour black tea, vine tea and other dozens of products), where is completely unable to find traces of artificial scenery, and its main peak "mushroom stone" is a landscape. So it has great important scientific and ecological value in researching the classification of agriculture environment data in Fanjing Mountains, so as to develop various sensors about new crops of growth environment factors (e.g. soil temperature, soil moisture content, air temperature, air humidity, light intensity and carbon dioxide (CO₂) concentration), as well as the biometric information of individual animals (e.g. body temperature, pulse, location information), realize the intellectualization of agriculture, establish administrative policy with the pilot of scientific research, the basis of management and protection, and the guarantee of coordination, better promote the green development of agriculture, and assure the security of resource and ecology. This article analyzes the agricultural environment monitoring data adopting the classification method of decision tree in machine learning.

1. Introduction

The work of constructing agricultural resource environment and rural energy ecology seriously implements the spirit of the 19th party, lifts up the flag of green agriculture development, highlights a green ecological orientation, constantly enhances the development level of green agriculture, and forms the development structure of driving the industry by business and promoting business by industry. And Fanjing Mountains is the "first famous mountain of Guizhou province" and "No. 1 Peak of Wuling", and in the original ecological preservation. Fanjing Mountains was designated as natural reserve by forestry department in 1956, as a world biosphere reserve by the United Nations in 1982, as national nature reserve by the state council in 1986, and the National 4A-class tourist attractions by the national tourism administration in 2012. Meanwhile, Fanjing Mountains national nature reserve has plenty of animal species, wide region, and the complex floristic for the national rare, so it is of great significance for researching on classifying agricultural environment monitoring data of Fanjing Mountains.

This paper aims to classify agricultural environment monitoring data of Fanjing Mountains by means of decision tree. The structure of this article is organized as follows. The next section presents the model of decision-making tree from three perspectives. Section III explains ID3 algorithm and



pruning. Then simulation results by matlab are given in section IV. Finally, Section V concludes the full text.

2. The model of decision-making tree

2.1. The fundamental principle of decision-making tree

The decision-making tree (DT) classifies a sample instance by arranging from the root node to a leaf node. Each non-leaf node on the tree represents the test of an attribute value, and its branches represent each result of the test; while each leaf node on the tree represents a class of classification, and the top node of the tree is the root node. Simply speaking, DT is a tree structure, similar to flow chart, as shown in Figure 1. From Figure 1, it adopts the top-down way of recursion, which, starting from the root node of the tree, compares the test of the attribute values on its internal nodes, and then determines the corresponding branch according to the attribute values of given instance, finally comes to a conclusion in leaf nodes of the DT. This process is repeated in a sub-tree with a new node as the root.

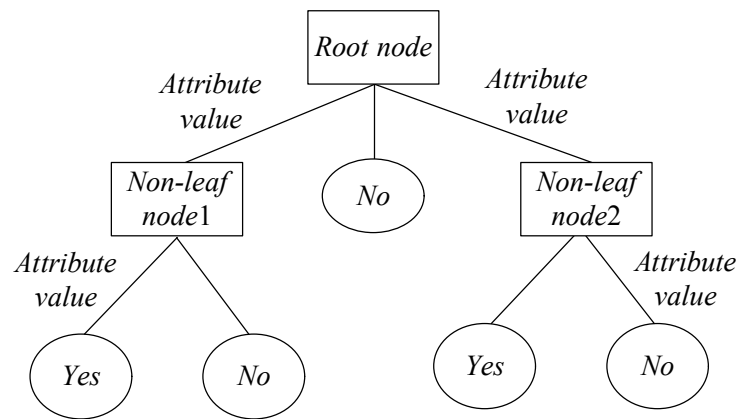


Figure 1. The tree-structure of DT.

DT is a method of nonlinear discriminant analysis, which divides into a series of smaller subgroups through the competition of cause variable. This process is repeated iteration in each branch of the tree, and in some standard, it selects the cause variable that has the strongest association with the result variable to be division. It first processes the data, uses the inductive method to generate readable rules, and then applies this rule to analyze new data. Essentially, DT is a process of classifying data through a series of rules. DT technology finds that the core of data patterns and rules is adopting the greedy algorithm of recursive partitioning. DT mainly has binary split tree and multi-way split tree. Because binary division is more flexible in the method of exhaustive search, it is generally used.

2.2. The idea of constructing DT

DT learning is based on the top-down method of recursion, whose basic idea is to construct a tree with the highest *Entropy* value in terms of *information entropy*, and the *Entropy* value at the leaf node is zero, where the *Entropy* is the measure of average uncertainty, and the lower the average uncertainty, the higher the certainty, thus the more accurate the result. The definition of *Entropy* is:

$$H(X) = - \sum_{x \in X} P(x) \log P(x) \quad (1)$$

The definition of average *mutual information* is that the known information of feature *Y* makes the reduced degree to the uncertainty of the information of the label *X*, and its formulation is

$$I(X;Y) = - \sum_{x \in X, y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad (2)$$

2.3. The division criteria of recursive partitioning

The establishment of a DT begins segmentation from the root node (for continuous variables with be segmented), exhaustively searching various possible ways of partitioning, through the division standard (usually use how much less of mutation in child nodes caused by the result variable as standard) to determine which causal variable as candidate segmentation variables. After the root node is segmented, the child nodes will repeat the segmentation process as the root node, and under the observation of sub-node the partitioning will be stopped until according with a certain condition. The division standard mainly has the following indicators, according to the classification of the result variables as follows.

Firstly, when the result variable is a classification variable, the generated tree is called a classification tree. There are two main criteria for its division, i.e., the index with the reduction of purity (as shown in Fig. 2) and χ^2 testing. In Fig. 2, $i(\cdot)$ refers to the calculation value of the purity of the nodes. The change of *Gini* coefficients and *Entropy* are based on the principle of purity reduction.

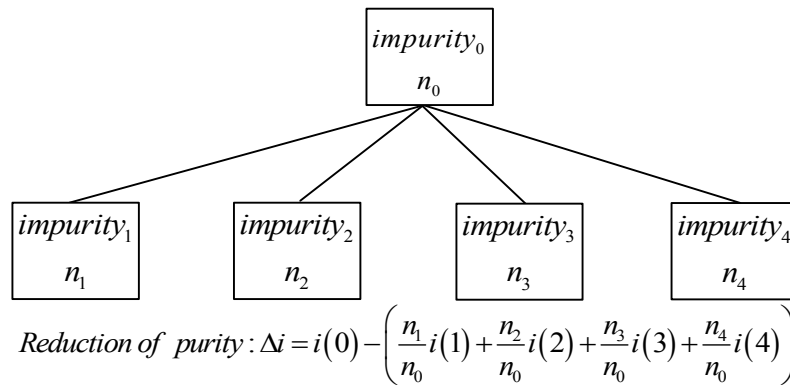


Figure 2. The index with the reduction of purity.

Secondly, when the result variable is a continuous variable, the generated tree is called a regression tree. There are also two main criteria for its division, i.e., the index with the reduction of purity [40] and F examining.

3. ID3 algorithm and pruning

3.1. ID3 algorithm

So far, there have been a lot of kinds of algorithms for DT generation, but Iterative Dichotomic version 3 (ID3) algorithm proposed by J.R. Quinlan is regard as the foremost in the most influential international sample learning algorithm. Quinlan's pioneering work mainly introduced the concept of mutual information in information theory into the DT learning algorithm, to which he referred as the information gain, and as the standard of attributes. To precisely define the information gain, a metric is first defined that is widely used in information theory, which is called *Entropy* that depicts the purity of any sample set.

If the target attribute has c different values, then the definition of *Entropy* that set S is relative to the classification of c states is re-expressed as

$$\text{Entropy}(S) = \sum_{i=1}^c (-p_i \log_2 p_i) \quad (3)$$

Where, p_i is the proportion of the samples of the i -th attribute value in the sample.

By (3), if all samples in the set S belong to the same class, $Entropy(S)=0$; If the sample size of two categories is not equal, then $Entropy(S)$ is within $(0,1)$. In particular, if the set S is a *Boolean* set, all samples in the set S belong to two different categories, then the following relationship is established: if the sample size of the two categories is equal, $Entropy(S)=1$. Since $Entropy$ has been used as a criterion to measure the purity of training samples, $Gain(S,A)$ is defined as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in V(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (4)$$

Where $V(A)$ is the value domain of attribute A , S_v is the subset that the value is equal to v in the property A of the set S . The information gain $g(D,A)$ of the feature A for the training data D is defined as the difference between the empirical entropy $H(D)$ of the set D and the empirical condition of A under the given conditions of D $H(D|A)$, namely:

$$g(D, A) = H(D) - H(D|A) \quad (5)$$

Obviously, this is the mutual information for training data set D and feature A .

Through all the features, select the most characteristic feature of information gain as the current division characteristics. Next, introduce the basic flow of ID3 algorithm. Assuming that the Examples are set of training samples, and Attribute list is the set of candidate properties. First, create the root node of DT- N . Second, if all samples belong to the same category c , then N is returned as a leaf node and is marked as c . Third, if Attribute list is empty, return N as a leaf node and mark the category with the most categories in the contained sample of that node. Fourth, the information gain of each candidate Attribute in Attribute list is calculated, and the corresponding Attribute* of the maximum information gain is selected, which is marked as root node N . Fifth, according to each value V_i in the value domain of Attribute*, the corresponding branch is generated from the root node N , and S_i is noted as the subset of samples to satisfy the condition of Attribute* = V_i in the Examples set. Sixth, if S_i is empty, the corresponding leaf node is marked as the categories of the largest class in the sample set of Examples. Otherwise, remove the Attribute* from the Attribute list, return (1), and recursively create the sub-tree.

3.2. Pruning

The largest DT can achieve 100% for the accuracy of training sets, but the results often tend to result in over-fitting (for both signal and noise). Therefore, the established tree model is not well extended to other samples in the overall. Similarly, a too small decision tree contains only a few branches, which can lead to under-fitting. A good tree model has a low bias (adaptive signal), and low variance (not adapt to the noise), whose complexity often makes a compromise between bias and variance, and thus the trees are needed to be pruned. The method of pruning is mainly used to deal with the problem of over-fitting, which mainly includes two methods: pre-pruning and post-pruning.

Firstly, pre-pruning is also called top-down stopping the rules. By stopping the structure of the tree in advance, for example, the tree is pruned by the decision to stop dividing at a node. Once the dividing of the node is stopped, it becomes a leaf node, which can take the most categories of classes in its contained subset as the class of the node. top-down stopping the rules mainly includes: limit the depth of the tree, limit the number of fracture (observed samples in each node), examine values of P according to the hypothesis (such as the values of P by testing or F testing, if it does not meet the statistical significance, the tree stops growth).

Secondly, post-pruning is also called down-top stopping the rules. The basic idea is to prune branches of fully grown trees by removing the branches of the nodes and replacing them with leaf nodes. The leaf nodes are usually marked with the most frequent types of subsets. That is, after growing a large tree, then the grown tree is pruned from its down to its top. The post-pruning requires

two conditions: one is to have a classification or predictive performance of evaluating method, and the simplest is to divide the data set into training sets and test sets. The test set is used to compare the models. However, there exists one problem: when the data set is smaller, the data set is split to reduce the degree of fitting. Another condition is to have a criterion of selecting a model, such as accuracy, profit, error, and etc.

Some researchers have conducted many experiments using simulation data, and found that the pre-pruning is usually faster than the post-pruning, but its effect is not as good as that of post-pruning [17, 41].

4. Simulation

Here, the number of environment monitoring data is randomly generated as 569, on behalf of various sensors about new crops of growth environment factors (e.g. soil temperature, soil moisture content, air temperature, air humidity, light intensity and carbon dioxide (CO₂) concentration), as well as the biometric information of individual animals (e.g. body temperature, pulse, location information), which are classified by DT. The total number of training sets and test sets are also randomly generated as 500 and 69, respectively. In this paper, we do three experiments as follows. Firstly, the created DT classifier is as shown in Fig. 3. Then, in Fig. 4, min-leaf value is set as 10 to generate an optimized DT. Lastly, Fig. 5 gives the result of a DT after pruning.

Compared with the above three conditions, it is not hard to find that the generated DT without processing data is more complex than the DT with setting min-leaf value and the pruned DT. And their resampling errors are 4.199999999999995e-02, 5.999999999999992e-03, and 2.199999999999997e-02, respectively, and their cross-validation errors are 5.999999999999984e-02, 6.199999999999983e-02, and 7.599999999999983e-02, separately. Because the best level is four, the resampling error of the pruned DT is the largest and its cross-validation error is the smallest. Therefore, according to the specific situation of the agricultural environment monitoring data and the actual requirements, to choose which decision tree to be used classification, in order to obtain the simpler and more accurate result.

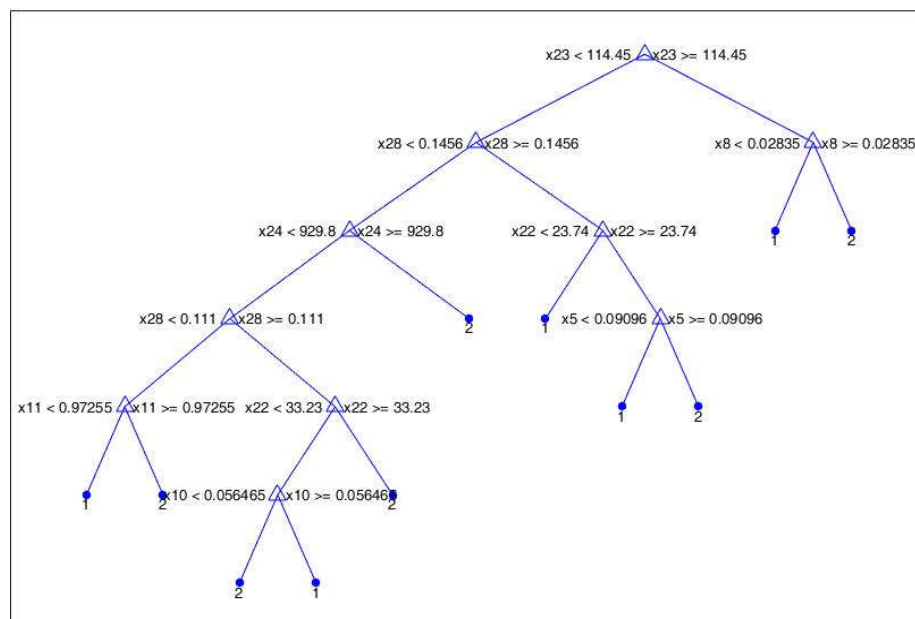
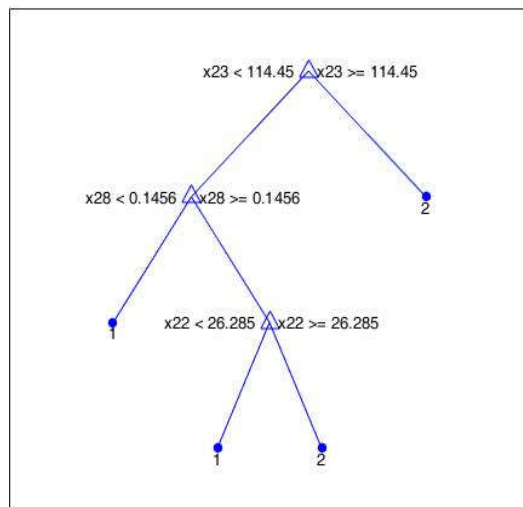
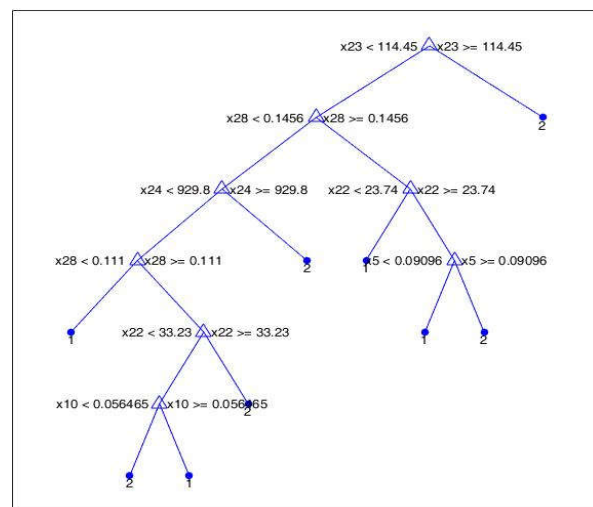


Figure 3. The created DT classifier.

**Figure 4.** The optimized DT.**Figure 5.** The pruned DT.

5. Conclusion

In a word, there are many models to solve classification problems in various fields, and the DT classification model is one of the most widely used methods [107]. In our future research work, all kinds of environmental monitoring data in Fanjing Mountain Areas will be collected, which will be analyzed by various classification methods of DT. Thus, to establish more scientific and effective management policy on the national nature reserve of Fanjing Mountain Areas, in order to achieve more remarkable achievements in different aspects, such as protection management, resource protection, forest fire prevention, tourism development, and so on.

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), and partly supported by the National Natural Science Foundation of China (NO.61562703).

References

- [1] H. Liu, A. Gegov, and M. Cocca, Rule Based Systems for Big Data: A Machine Learning Approach. Switzerland: Springer, 2016.
- [2] Stephen Marsland, Machine Learning, An Algorithmic Perspective, Second Edition: CRC Press, Oct. 2014.
- [3] Z.H.Zhou, Machine Learning, Tsinghua University Press, Beijing, China, 2016.
- [4] Lavinia Aurelian Bădulescu, Data mining classification experiments with decision trees over the forest covertype database, 21st International Conference on System Theory, Control and Computing (ICSTCC), 2017, pp. 236-241.
- [5] Priyanga Chandrasekar; Kai Qian; Hossain Shahriar; Prabir Bhattacharya, Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing, IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2 (2017) 481-484.
- [6] Martin Schuster; Thomas Walther, Valuation of combined wind power plant and hydrogen storage: A decision tree approach, 14th International Conference on the European Energy Market (EEM), 2017, pp. 1-6.
- [7] Xinhuai Zou, Ming Cheng, Cheng Wang, Yan Xia, Jonathan Li, Tree Classification in Complex Forest Point Clouds Based on Deep Learning, IEEE Geoscience and Remote Sensing Letters, 14 (2017) 2360-2364.

- [8] Han Liu, Mihaela Cocea, Weili Ding, Decision tree learning based feature evaluation and selection for image classification, International Conference on Machine Learning and Cybernetics (ICMLC), 2 (2017) 569-574.
- [9] Duncan Macmichael, Dong Si, Addressing Forest Management Challenges by Refining Tree Cover Type Classification with Machine Learning Models, IEEE International Conference on Information Reuse and Integration (IRI), 2017, pp. 177-183.
- [10] Sai Zou, Yuliang Tang, Yanglong Sun, Ke Su, An Identification Decision Tree Learning Model for Self-management in Virtual Radio Access Network: IDTLM, appeared in IEEE Access, 2017.
- [11] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, Privacy-Preserving Inductive Learning with Decision Trees, IEEE International Congress on Big Data (BigData Congress), 2017, pp. 57-64.