Research on plant leaf classification and retrieval method based on machine learning

Ying Zhang*, Jianxiong Tang, Yi Wang
College of Informatics, Huazhong Agricultural University, Wuhan, China

*Corresponding author e-mail: zy@mail.hzau.edu.cn

Abstract. In this article, a variety of leaf images are taken as the research objects. The images were preprocessed and the color, shape and texture features of the leaf images were extracted. Five machine learning algorithms were used to classify and retrieve the feature values of these leaf images, and the recognition effects of each algorithm was obtained.

Keywords: Plant classification, Plant retrieval, Machine learning, Feature extraction

1. Introduction
The beginning and the most important part of the protection of green plants is to classify plants efficiently and quickly [1]. Due to limited human cognitive ability and a great variety of plants species, manual classification and identification methods of plants are very inefficient and difficult. In recent years, because of the rapid development of computer technology, image recognition and artificial intelligence research becomes research focus, and image retrieval and classification technology have a wide range of applications in various fields [2].

In this article, relevant leaf image extraction methods are applied to complete the extraction of relevant feature values of leaf images. Then the machine learning classifier is trained. And finally, classification test is carried out to achieve leaf classification and retrieval. By comparing the retrieval rate and recognition rate, a relatively good machine learning algorithm can be obtained.

2. Classification of plant leaf
Generally speaking, the leaf recognition process can be divided into two stages, namely the training stage and the testing stage. During the training stage, relatively mature training results can be formed by learning the feature vectors of the leaves in the data set. During the testing stage, the feature vectors of the test data set are input into the classifier for prediction, and at last the corresponding plant species are obtained. The machine learning algorithm completes the classification of plant leaf through the above two stages of training and testing.

2.1. Leaf feature extraction

2.1.1. Leaf image collection and preprocessing. The data set used in this article is a leaf data set collected on the Internet which includes about 1,500 leaf sample images. The preprocessing of the original image includes image binarization processing and leaf contour extraction. The gray
processing method proposed in this article is weighted average method. The formula for calculating the weighted average value is as follows:

\[ f(i, j) = 0.30 \times r(i, j) + 0.59 \times g(i, j) + 0.11 \times b(i, j) \]

Where \( f \) represents the gray value of each pixel, \( r \) represents the value of the red component of each pixel, \( g \) represents the value of the green component of each pixel, \( b \) is the value of the blue component of each pixel. Leaf image contour extraction can be divided into two parts: contour recognition and contour drawing. This article extracts the contour by calling the library function in Opencv. The result of contour extraction is shown in Figure 1.

![Image](image.png)

**Figure 1.** Contrast before and after contour extraction

2.1.2. **Color feature extraction.** In this article, the method of color feature extraction is to calculate the percentage of several main colors on the leaves and to complete feature extraction in the form of color histogram. The cv2.calcHist () function in OpenCV was used to obtain the color histogram features of the image.

| Feature number | Feature name                          | Feature value |
|----------------|---------------------------------------|---------------|
| F[3]           | Percentage of pixels 1                | 0.1563        |
| F[4]           | Percentage of pixels 2                | 7.9958        |
| F[5]           | The mean value of tonality            | 4.7244        |
| F[6]           | The mean value of saturation          | 0.1587        |

2.1.3. **Contour feature extraction.** Table 2. Shows a group of contour features extracted from a certain leaf.

| Feature number | Feature name                                      | Feature value |
|----------------|---------------------------------------------------|---------------|
| F[7]           | Aspect ratio                                      | 13.2056       |
| F[8]           | The ratio of the area of the contour to the area of the bounding rectangle | 15.8277       |
| F[9]           | Convex hull ratio of area                         | 0.9643        |
| F[10]          | The diameter of a circle with an area equal to the contour | 0.9766        |
| F[11]          | Perimeter Convexity                              | 0.1258        |
| F[12]          | Sphericity                                        | 0.8674        |
| F[13]          | Eccentricity                                      | 1.6347        |
| F[14]          | Shape parameter                                   | 2.0474        |
| F[15]          | The ratio of axial diameter                       | 1.3006        |

2.1.4. **Texture feature extraction.** In this article, the gray level co-occurrence matrix method is used to extract the texture features of leaves. On the basis of extracting the gray level co-occurrence matrix, four feature values including contrast, energy, entropy and correlation were extracted to describe the texture features of plant leaves.
Table 3. A set of leaf texture features

| Feature number | Feature name | Feature value |
|----------------|--------------|---------------|
| F[11]          | Contrast     | 2.6492        |
| F[12]          | Energy       | 0.8111        |
| F[13]          | Entropy      | 0.8654        |
| F[14]          | Correlation  | 3.3062        |

2.2. Machine learning algorithm
After the feature extraction of leaf image completed, five commonly used machine learning algorithms in sklearn package are called for training and testing. The trained algorithm is used to predict the type of test data set, and then the accuracy rate and log loss value are calculated by comparing the predicted value type with the actual type of set of testing leaves. These parameters are used to compare the advantages and disadvantages of these machine learning algorithms. (1) k-Nearest Neighbour algorithm, (2) Support vector machine algorithm, (3) Decision tree algorithm, (4) Random forest algorithm.

3. Leaf image retrieval method
Common image retrieval methods mainly include text-based image retrieval (TBIR) and content-based image retrieval (CBIR). The TBIR method marks the features by hand and then retrieves the features.

The CBIR method is generally divided into three parts, namely user interface, database image processing and image retrieval module. Feature extraction is also the basic part of image retrieval, including color feature, texture feature, shape feature and space feature, etc. After the feature value of images is extracted, a certain measure method is needed to compare the similarity between the feature vectors of the retrieved image and the features in the image index feature database. The quite common measurement methods are: Minkowsky distance; Histogram; Euclidean distance. Through training and learning, leaf images can be retrieved more efficiently. The foundation of these algorithms is statistical theory. The common machine learning methods include: KNN algorithm, SVM support vector machine algorithm and decision tree algorithm, etc [3-5].

4. Results and analysis
In order to achieve the purpose of the study, over 1500 leaf images from more than 30 plant species were selected for the experiment, half of which were testing data sets and half were training data sets. In this article, Python language was used for programming experiments under the PyCharm environment. The color feature, shape feature and texture feature were extracted by calling all kinds of functions in OpenCV library. Then several machine learning algorithms of the Sklearn library were called to classify, and the corresponding accuracy rate and log loss value were calculated by comparing the sample test value with the predicted value obtained by the machine learning algorithm. The results are as follows:

![Figure 2. Partial feature extraction results](image-url)
Figure 3. Accuracy and loss value results of several classifiers

The above figures are partial operation result figure of this study, figure 5 shows the results of partial feature extraction, figure 6 shows the results of classifier operation, table 4 shows the performance comparison of five machine learning algorithms, figure 7 and figure 8 are the bar charts showing the accuracy rate and log loss results. The results showing, in terms of accuracy, the decision tree algorithm performs best among all, reaching 100%; followed by random forest algorithm, KNN algorithm, nusvc algorithm, all over 90%; the worst one is svc algorithm, only 83%. From perspective of log loss value, svc and nusvc algorithms have distinct log loss value, while other 3 algorithms have no obvious log loss value.

![Classifier performance comparison](image-url)

Figure 4. Accuracy histogram of several classifiers
Table 4. Performance comparison of several machine learning algorithms

| Machine Learning Algo     | Accuracy   | Log loss        |
|---------------------------|------------|-----------------|
| KNeighborsClassifier      | 94.9495%   | 0.132794878583505 |
| SVC                       | 83.5017%   | 4.612996756132111  |
| NuSVC                     | 95.2862%   | 2.029712998498699  |
| DecisionTreeClassifier    | 100.0000%  | 9.889619987133689  |
| RandomForestClassifier    | 99.8316%   | 0.220091022323875  |

The above figures are partial operation result figure of this study, figure 5 shows the results of partial feature extraction, figure 6 shows the results of classifier operation, table 4 shows the performance comparison of five machine learning algorithms, figure 7 and figure 8 are the bar charts showing the accuracy rate and log loss results. The results showing, in terms of accuracy, the decision tree algorithm performs best among all, reaching 100%; followed by random forest algorithm, KNN algorithm, nusvc algorithm, all over 90%; the worst one is svc algorithm, only 83%. From perspective of log loss value, svc and nuvc algorithms have distinct log loss value, while other 3 algorithms have no obvious log loss value.

5. Conclusion
In this article, we find that the accuracy rate can reflect the effect of machine learning well for a given test data set. As can be seen from the results, the decision tree algorithm is significantly better than the other four algorithms, while the SVC algorithm is not as good as others. However, there are many methods and indexes to evaluate machine learning, and the two indexes used in this article cannot fully reflect the advantages and disadvantages of these algorithms. It should also be evaluated comprehensively from professional statistical indicators such as ROC-AUC curve and GINI coefficient when comes to the classification effect of certain machine learning algorithm. In addition, by comparing classification effect of modified algorithm and original one is another effective research angle.

Acknowledgments
This work was financially supported by YingZhang’s 2019 Independent Science and Technology Innovation Fund Projects of Huazhong Agricultural University.

References
[1] Meilong Chen, Shengkui Dai. Image Texture Feature Analysis Based on GLCM Algorithm [J]. Communication Technology, 2012, 45 (2): 108-111.
[2] Jiao Ding, Dong Liang, Qing Yan. Image Recognition of Multi-Feature Plant Leaves Based On D-LLE Algorithm [J]. Computer Engineering and Applications, 2015, 51 (9): 158-163.
[3] De En, Shengqiang Hu. Leaf Identification Method Based on Integrated Neural Network [J]. Journal of Zhejiang Agriculture, 2015, 27 (12): 2225-2233.
[4] Lijun Wang, Yongjian Huai, Yuecheng Peng. Species Identification of Foliage Plants Based on Multi-Feature Fusion of Leaf Images [J]. Journal of Beijing Forestry University, 2015, 37 (1): 55-61.
[5] Ning Zhang, Wenping Liu. A Leaf Recognition Method Based on Clonal Selection Algorithm and K-Nearest Neighbor Algorithm [J]. Computer Application, 2013, 33 (7): 2009-2013.