Leaf image recognition and classification based on GBDT-probabilistic neural network

Zhixuan Tang$^{1,2}$
$^1$School of Science, Wuhan University of Technology, Wuhan, 430070, China
$^2$Email: 422970498@qq.com

Abstract. In this paper, the binary images of 100 kinds of leaves are used for leaf recognition. Firstly, we screen 35 important features and use grey clustering analysis to establish the quantitative feature system of leaves. Then we use the gradient descent tree algorithm (GBDT) to select core features and use probabilistic neural network (PNN) to recognize and classify leaves, constructing a hybrid GBDT-PNN model. In the end, we obtain the classification results of leaves to evaluate model performance and the influence of core features on the model. The results show that the accuracy rate of GBDT-PNN model using 12 core features is 92.75%. And the accuracy rate with all 35 features is 93.5%. It illustrates that the model has great performance and core features have high influence on the model. By comparing with other commonly used deep learning algorithms and models, it is verified that the GBDT-PNN image recognition and classification model is effective and has high accuracy.

1. Introduction
Plants are an essential part of maintaining the natural ecological balance and people’s life. However, in recent years, with the expansion of human activities, the growth of plants has been severely challenged. The premise of plant protection is the accurate identification and classification of plants, which is of great significance [1]. The leaves of plants can be considered as two-dimensional shapes, whose retention time is longer than other organs, and they are not easily affected by environment. Nowadays, leaves have become an important feature in plant recognition and classification [2]. Therefore, we can use leaf image to recognize and classify plants.

Nowadays, many scholars have used deep learning algorithms to recognize and classify blade images and have achieved good research results. Wang [3] have developed a system using the support vector machine (SVM) algorithm to realize the leaf recognition, with the recognition rate reaching 91.41%. Zhang [4] proposed a hierarchical convolution deep learning system, with a recognition rate of 91.11% for leaf image. Zhai [5] proposed a method of leaf recognition based on fractal dimension of leaf margin and vein, and added fractal theory to the field of leaf recognition. Backes [6] uses complex network method to analyze and identify blade features, which has good effect on blade image classification.

Although many encouraging advances have been made in the study of leaf classification, there are still various problems. The recognition rate of the traditional method depends to a large extent on whether the artificially selected features are reasonable. However, people often rely on experience in selecting features, which is very blind. These features that people choose are designed for specific data. If we use the same features to process different data sets, the results may be quite different. Therefore, this kind of feature cannot be widely used for recognition and classification of various blades.
In order to solve the above problems, this paper extracts comprehensive leaf features based on the binary images of leaves, and uses grey clustering analysis to construct a quantitative feature system of leaves. We use Gradient Descent Tree (GBDT) algorithm to screen out core features and construct a leaf recognition and classification model based on Hybrid GBDT-Probabilistic Neural Network (PNN) to classify leaves. PNN has a simple structure and fast learning ability, and has been widely used in pattern recognition. Through the blade classification result, we evaluate the model performance and the influence of core features on the model. GBDT-PNN model selects comprehensive blade features and reduces manual intervention to the greatest extent, which is widely applicable to the recognition of various blades. Besides, the model reduces image noise interference and achieves excellent image recognition effect.

2. Leaf recognition and classification model based on GBDT-probabilistic neural network

2.1. Quantitative analysis of leaf image features
According to the relevant theories, leaf features are one of the most important features in plant classification. The binary image contains a number of two-dimensional leaf features. But if all features are extracted, the amount of calculation and storage for image processing is too large, time-consuming and laborious. Therefore, the following features with good recognition effect are selected to establish a leaf feature database.

2.1.1. Geometric parameters and features. The geometric parameters and features can best represent the boundary information. So we quantify 15 geometric parameters and features: blade perimeter, blade area, blade convex hull perimeter, blade convex hull area, blade major axis, blade minor axis, blade minimum circumscribed rectangle, aspect ratio, rectangle ratio, area convexity, perimeter convexity, eccentricity, diameter ratio and shape parameter.

2.1.2. Blade centroid parameters. Centroid is an imaginary point in material system which is considered to be concentrated and has several important features. In this paper, four centroid parameters are selected: centroid coordinate $(\bar{x}, \bar{y})$, phylloid, globular and rotundity [7].

2.1.3. Blade skeleton parameters. Leaf skeleton can well reflect its shape features. The ideal skeleton can be obtained by removing noise from blade binary image. Figure 1 is a schematic diagram of leaf skeleton extraction.

![Skeleton extraction from leaf binary image](image)

Figure 1. Skeleton extraction from leaf binary image.

In this paper, four skeleton parameters are set: skeleton size $l_2$, skeleton branch number, skeleton node number and degree of fullness of skeleton.

2.1.4. Leaf edge features. Leaf edge describes the subtle features, which is of great significance to recognize leaves from multiple angles. Therefore, we set the blade edge features [8]: the ratio of convex residuals to blade area, the average value of convex residuals area and the number of convex residuals. The schematic diagram of convex residuals is shown in Figure 2.
2.1.5. **Invariant moment.** Moment features mainly characterize the geometric features of image regions, which are also called invariant moments due to their natural rotation, translation and scale invariance. In this paper, we use 7 kinds of invariant moments introduced by Feiyu Yao [9] for blade classification.

2.1.6. **Leaf fractal dimension.** Fractal dimension describes the complexity of the image [10] and reflects the relationship between the whole and a part of the blade. When it is applied to leaf binary image, it indicates the complexity of pixel composition and is directly related to people’s visual perception of leaves.

In summary, we select 35 leaf features for leaf recognition and classification.

2.2. **Clustering analysis of quantitative features based on grey correlation analysis**

Clustering analysis is an effective method in classification and identification. The statistics commonly used in traditional clustering analysis methods [11] include distance coefficient, similarity coefficient and correlation coefficient. Correlation in grey system theory is used between two different sequences. Therefore, grey clustering analysis can sort and cluster the indexes according to certain properties, making the result more accurate and flexible. So we use grey clustering analysis to sort and cluster the leaf features. The process is as follows:

**Step 1:** Calculation of Grey Correlation Degree between Features. We have known the standardized features $X = (X_1, X_2, \ldots, X_i)$. The absolute difference of the $p$-th component of feature $X_m$ and feature $X_n$ is $\Delta_{mn}(p) = |X_{mp} - X_{np}|$ for $m, n = 1, 2, \ldots, 35$, $p = 1, 2, \ldots, 1600$. The correlation coefficient between $X_m$ and $X_n$ can be obtained as follows:

$$\sigma_{mn}(p) = \frac{\Delta_{mn}(p) + \xi \Delta_{max}}{\Delta_{mn}(p) + \xi \Delta_{max}},$$

where $\xi \in [0, 1]$. In the grey system theory, the average value of $\sigma_{mn}(p)$ is taken as the correlation degree between features $X_m$ and $X_n$, as shown in equation 2:

$$r_{mn} = \frac{1}{1600} \sum_{p=1}^{1600} \xi_{mn}(p).$$

**Step 2:** Cluster analysis based on grey correlation degree among features. For the clustering of symmetric matrix $E$, we use the method of forming a clustering structure at one time. And the following five principles are followed:

- If the selected pair of features do not appear in the original group, they will form an independent new group.
- If one of the selected pair of features appears in the original group, the other feature is added to the group.
- If the selected pair of features respectively appear in the two groups that have been divided, the two groups are linked together.

![Figure 2. The schematic diagram of convex residuals.](image-url)
If the selected pair of features all appear in the same group, the pair of features need not be grouped separately.

If there are too many features in a group after clustering, cluster it separately again.

According to the above principles, we can get the classification of all features.

2.3. GBDT-probabilistic neural network model

2.3.1. Feature selection based on GBDT. The gradient descent tree (GBDT) is a forward step-by-step algorithm. The basis function is CART. By continuously fitting the gradient of the loss function, the gradient descent method is used to obtain the optimal solution [12]. In order to ensure the accuracy of the results, we first screen the grouped leaf features and select the features that are effective for leaf identification and classification. In this paper, GBDT model is used to calculate the feature contribution degree, which is used to judge whether the feature should be eliminated. We only use the features which make great contributions to leaf identification and classification.

Assume that the model GBDT establishes $M$ decision trees. The contribution of feature $T$ to the model is the mean change of $Gini$ impurity when $T$ is used as a splitting node $s$ in the $M$ decision trees. The formula of $Gini$ impurity is as follows:

$$Gini(s) = 1 - \sum_k [p(k|s)]^2,$$

(3)

Where $[p(k|s)]$ is the relative frequency of node $s$ in category $k$. The value of $Gini$ changes after branching are as follows:

$$I_j = Gini(s) - Gini(l) - Gini(r).$$

(4)

In formula (4), $Gini(l)$ and $Gini(r)$ are respectively $Gini$ indices of left and right new nodes obtained by splitting node $s$. If the tree $T$ has a total of $L$ splits, searching all split nodes from the root node to the L-1 layer, we can define the contribution degree of the index in a single tree, which is shown as follows:

$$In_j^T = \sum_{i=1}^{L-1} I_j^i l(s_i = j).$$

(5)

The contribution of variable $j$ is the average contribution on $M$ trees, which is as follow:

$$In_j = \frac{1}{M} \sum_{i=1}^{L-1} In_j(T_i).$$

(6)

From formula (6), we can calculate the contribution of all features and normalize them. The features with great contribution in each group are selected for classification by PNN.

2.3.2. Probabilistic neural network. GBDT is used to screen features to obtain core features, then probabilistic neural networks (PNN) use core features to identify and classify the leaves. PNN is proposed by Dr. Specht in 1989 [13], which has simple structure, good generalization ability and fast learning ability. PNN has been widely used in pattern classification. The deep learning algorithm based on GBDT-PNN can learn blade features autonomously and reduce manual intervention. It can eliminate noise interference for leaf images and improve image recognition efficiency.

The structure of PNN consists of input layer, hidden layer, summation layer and output layer. The output layer of PNN adopts competitive output instead of general linear output. Each neuron only competes for the response opportunity of input mode according to Parzen method, and finally only one neuron wins the competition. The winning neuron indicates the classification of input mode. The step of probabilistic neural network is as follow:

Step1: Determine the center of the radial basis function of hidden layer neurons. We assume that $x_{mn}$ represents the m-th input variable of the n-th training sample, and $t_{mn}$ is the m-th output variable.
of the n-th training sample. Each neuron in the hidden layer corresponds to a training sample. It means the radial basis function center corresponding to n hidden layer neurons is \( C = x' \).

Step2: Determine hidden layer neuron threshold. Threshold of hidden layer neurons is \( B = [b_1, b_2, \ldots, b_N] \), \( b_1 = b_2 = \ldots = b_N = \frac{0.8326}{spread} \). \( spread \) is the expansion speed of the radial basis function.

Step3: Determine weight between hidden layer and output layer. After the radial basis function center and threshold value of hidden layer neurons are determined, the output of hidden layer neurons is shown in formula (7):

\[
a'_i = e^{-K^{-\lambda}f_B^i}, \quad i = 1, 2, 3, \ldots, N,
\]

where \( x_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \) is the connection weight \( W \) between the hidden layer and the output layer of the i-th training sample. \( W \) between the two layers of the PNN is the output matrix \( T \) of the training set.

Step4: Calculate the output of neurons in output layer. After determining the connection weight \( W \), the highest scoring item in the summation layer is the output, the formula is:

\[
\lambda' = LW_{2,1}a', \quad (8)
\]

\[
y' = \text{compet}(\lambda'). \quad (9)
\]

The output result is the classification of leaves.

### 3. Empirical analysis of leaf recognition and classification

In order to verify the feasibility and effectiveness of GBDT-PNN model, this paper carries out leaf recognition and classification. We select 1600 leaf images, including 100 kinds of leaves. Each kind has 16 leaves. The leaf images are from ImageCLEF 2011 database [14] established by Pl@ntNet. This database has three types of leaf images: scan, scan-like, and photograph. We selected 1600 images of scan type and binarized each image. The similarities and differences between leaves are different, which increases the difficulty of leaf recognition and is more conducive to consider the feasibility and effectiveness of the GBDT-PNN model.

Firstly, we calculate the correlation score of 35 features based on grey correlation analysis, and the average value of this correlation matrix is 0.806. Therefore, we support that the correlation is high. If all the features are used as PNN input data, the calculation amount will be redundant. So after 35 quantified leaf features are numbered (the number is the order in which the feature appears), the grey correlation analysis is used to cluster features. Then we divide the features into 11 categories with average number, and establish a complete quantitative system of features, which lay a solid foundation for the selection of core features and leaf classification in the following text.

We construct GBDT model to calculate the contribution score of each feature and rank all the features. In order to improve the effectiveness and accuracy of the model, we select two features with the greatest contribution in each category as effective features, 22 effective features obtained are shown in Figure 3.
Figure 3. Effective features and their contribution value.

We re-rank the 22 effective features based on the contribution score and obtain the top 12 with the highest score as core features. The core features extracted are shown in Figure 4.

Figure 4. Core features of leaf.

Based on the above selected core features, this paper uses GBDT-PNN leaf recognition method to classify 1600 leaf images. Leaf images in each kind are randomly divided at a ratio of 3:1. So we have 1200 training sets and 400 test sets. Before the specific image processing, we normalize each image, mainly involving the size of the image. We define the size of the leaf image as an array of 100×100 pixels. The results of leaf recognition and classification based on core features and effective features are shown in Figure 5 and Figure 6.
Figure 5. Classification result. (core features)

Figure 6. Classification result. (effective features)

The accuracy of leaf recognition and classification based on core features is 92.75%, which based on effective features is 91.5% and based on all features is 93.5%. The accuracy rate and running speed of the model is shown in Table 1:

| Feature Types     | Leaf Classification Accuracy | Running Speed of the PNN Model |
|-------------------|------------------------------|---------------------------------|
| Core features     | 92.75%                       | 0.11s                           |
| Effective features| 91.5%                        | 0.19s                           |
| All features      | 93.5%                        | 0.25s                           |

From Table 1, the accuracy of leaf classification under the three feature types is high. The accuracy of core features is higher than that of effective features. And the accuracy of 11 core features is 87%.
which shows that reducing one core feature will greatly reduce the accuracy. So the number of core features has a great impact on the model. Besides, with the decrease of features, the running speed reduction rate gradually increases. These improve that the selection of core features in this paper is reasonable. And we can see PNN has a fast running speed. It shows that PNN has high efficiency because PNN is based on one-pass learning algorithm and do not require iterative training.

In order to verify the classification accuracy of GBDT-PNN, we select several widely used leaf classification models to replace PNN based on core features, including BP neural network [15], support vector machine (SVM) [16] and convolution neural network (CNN) [17]. The results are shown in Table 2.

| Model   | Classification Accuracy | F1-measure |
|---------|-------------------------|------------|
| BP      | 82.5%                   | 0.6789     |
| SVM     | 90.75%                  | 0.8196     |
| CNN     | 91.5%                   | 0.8556     |
| PNN     | 92.75%                  | 0.8937     |

From Table 2, GBDT-BP model has the lowest classification accuracy. GBDT-SVM and GBDT-CNN have lower accuracy than GBDT-PNN. And we show F1-measure (parameter $\alpha=1$) for each model. The greater the F1-measure, the higher the accuracy of the method. The value range is [0,1]. The F1-measure of PNN is 0.8937, which is significantly higher than other methods. So Table 2 illustrates PNN has higher accuracy.

| Model                  | Classification Accuracy of leaves |
|------------------------|----------------------------------|
| GBDT-PNN               | 92.75%                           |
| TSLA-STSLA-KLSH[18]    | 91.02%                           |
| CCH-WVF-SVM[19]        | 89.15%                           |
| DFH-GP[20]             | 87.06%                           |

Table 3 shows the classification accuracy of some methods with good performance based on ImageCLEF 2011 database. The accuracy of GBDT-PNN is 92.75%, which is obviously higher than other methods. It well demonstrates the model in this paper has higher accuracy.

In conclusion, the above results analysis, running speed analysis and comparative analysis can prove that GBDT-PNN model has perfect performance such as high rationality, high efficiency and high accuracy.

4. Conclusions

In this paper, a leaf recognition and classification model based on GBDT-PNN is proposed. We use grey clustering analysis to construct a quantitative feature system for leaves. The GBDT algorithm is used to select the core features based on the contribution score. And PNN is used to identify and classify leaves. Based on the classification results of leaves, we evaluate model performance and the effects of core features on the model. Compared with other classification and recognition models, the experimental results show that GBDT-PNN model has the higher accuracy rate and better recognition and classification effect.

We extract leaf information from all aspects of leaf features. The extracted features are comprehensive and easy to classify. Because the GBDT-PNN model established in this paper reduces manual intervention to the greatest extent and has high classification accuracy, it also has recognition and classification function for other types of pictures. The model has wide applicability. The future work is to improve the network structure and enhance the computational efficiency, further improve the recognition rate and apply the model to a wider range of image processing fields.
Acknowledgement
This work is supported by the National innovation and entrepreneurship training program for college students (20191049714013).

References
[1] Gong D X and Cao C R 2014 Classification of plant leaves based on convolutional neural network [J]. Computer and Modernization 04 12-15 + 19
[2] Rahman M M, Islam M S, Sassi R, Akhtaruzzaman M 2019 Convolutional neural networks performance comparison for handwritten Bengali numerals recognition [J]. SN Applied Sciences 1(12)
[3] Wang L J, Huai Y J, Peng Y C 2015 Species identification of foliage plants based on multi-feature fusion of leaf images [J]. Journal of Beijing Forestry University 37(01) 55-61
[4] Zhang S, Huai Y J 2016 Study on plant leaf recognition based on hierarchical convolution deep learning system [J]. Journal of Beijing Forestry University 38(09) 108-115
[5] Zhai C M, Wang Q P, Du J X 2014 Study on the method of plant leaf recognition based on fractal-dimension characteristics of leaf margin and vein [J]. Computer Science 41(2) 170-173
[6] Backes A R, Casanova D, Bruno O M 2009 A complex network-based approach for boundary shape analysis [J]. Pattern Recognition 42(1) 54-67
[7] Li Y, Li Y Y, Luo H C 2017 Online plant left recognition based on shape features Computer Engineering and Applications 53(2) 162-165
[8] Li Y B, Yu X Y, Wang Z J 2013 Research on risk assessment of photovoltaic power generation project using grey correlation degree and TOPSIS method [J]. Power Grid Technology 37(06) 1514-1519
[9] Yao Y F 2011 A method for leaf identification based on fractal dimension [D]. Beijing Forestry University
[10] Yuan J S, Yao Y F 2012 A method for image recognition of blades based on fractal dimension [J]. Computer Engineering and Design 33(02) 670-673
[11] Huang T H, Wang Y L, Wang Z, Cui G S 2008 Distributed traffic flow data prediction system based on Spark [J]. Computer Application Research 35(02) 405-409 + 416
[12] Wang H W, Meng Y 2017 Online comment quality useful feature recognition: based on GBDT feature contribution method [J]. Journal of Chinese Information Technology 31(03) 109-117
[13] Yuan J S, Shang H K 2013 Partial discharge pattern recognition of transformer based on principal component analysis and probabilistic neural network [J]. Power Automation Equipment 33(06) 27-31
[14] Image CLEF-Image Retrieval In CLEF[EB/OL]. http://www.imageclef.org/2012/plant
[15] Zhou Y, Li S J 2020 BP neural network modeling with sensitivity analysis on monotonicity based Spearman coefficient [J]. Chemometrics and Intelligent Laboratory Systems 200
[16] Wei L 2012 Research on classification of plant leaves based on image processing and SVM [D]. Northwest Agriculture and Forestry University
[17] Cai J, Cai J Y, Liao X D, Huang H T, Ding Q J 2015 Preliminary study on gesture recognition based on convolution neural network [J]. Computer System Application 24(04) 113-117
[18] Hong Y 2015 Research on Plant Leaf Recognition Based on Leaf Contour and Corner [D]. Hefei University of Technology
[19] Yanikoglu B A, Aptoula E and Tirkaz C 2012 Sabanci-Okan System at ImageClef 2012: Combining Features and Classifiers for Plant Identification[C]/CLEF (Online Working Notes/Labs/Workshop)
[20] Yahiaoui I, Mzoughi O and Boujemaa N 2012 Leaf shape descriptor for tree species identification[C]. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME 2012 254-259