Screening of Tobacco's Effective Image Features Based on a Semi - supervised Clustering

Ge Jin[1, a*], Lin Qi[1, b] and Hang Li[1, c]

[1]Zhengzhou University, No.100 Science Avenue, Zhengzhou City, China

[a]lonezhizi@163.com, [b]ielqi@zzu.edu.cn, [c]lh121144@163.com

*The corresponding author

Keywords: Clustering; Discriminant function; Feature selection; Support vector machine; Tobacco classification

Abstract. In order to reduce the number of extracting tobacco's image features and the computational complexity of hierarchical model, and to increase the speed and accuracy of tobacco classification, this paper presents a feature selection method that based on the semi - supervised clustering. First, define the discriminate function R which can distinguish good features from bad features, and delete bad features according to the R-value. Then set up 42 levels' SVM hierarchical model using all features and screened features. Experimental results show that the feature selection method constructed in this paper can select effective features, which can raise the tobacco's classification speed under the premise of correct discrimination.

Introduction

Tobacco is one of the agricultural products with important economic value [1]. The quality of the tobacco has a direct impact on the economic income of tobacco industry and the health of smokers. We mostly rank the tobacco in an artificial way in the process of tobacco's purchase at this stage. This classification method with subjective factors not only affects the correct rate of classification, but also causes unnecessary disputes. Intelligent tobacco classification method has the characters of high speed and high accuracy, and it can avoid the subjectivity of artificial classification. At present, the intelligent classification is mainly based on tobacco image information [2, 3] or spectral information[4,5]. Image information can better reflect factors, such as length, color, damage rate, and texture which is closely related to tobacco's level.

It needs reduce the dimension because that the collected image features are high-latitude and large-redundancy. At present, there are some common methods including principal component analysis method, wavelet analysis method [6], continuous projection method [7], minimum interval two-multiplication method [8], independent component analysis method [9], Particle swarm optimization [10], Genetic algorithm [11] and so on, and we can use these methods to reduce the dimension of original features and to select features. These methods can effectively reduce the input dimension of the hierarchical model，which can reduce the complexity of the hierarchical model but cannot reduce the time of collecting original features ,which greatly affect the speed of whole recognition. This paper filters valid features directly from original features and only collect effective features while collecting features, which can not only reduce the complexity of classification model but also reduce the number of data collection. This paper screens effective features by Semi - supervised Clustering, and test the collected features in 42 tobacco's levels by SVM classifier.

The Principle of Features' Selection

For the collected image features, it can be known from the clustering idea: The smaller inner-class difference of same features is, the better it is. The bigger inter-class difference of same features is, the better it is. Some collected image features cannot reflect the clustering idea very well. This paper considers of the same feature's inner-class difference and inter-class difference at the same time, constructs a ratio function of features' inner-class and inter-class discrete degree, and gets rid of features that bad for or have small impact to classification. There are some specific methods:

Suppose the training set is P dimension vector of C class, and there is N sample in the class, the i-th sample in k-th class is expressed as $X_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \dots x_{ij}^{(k)} \dots, x_{iP}^{(k)})', 1 \le \mathrm{i} \le C_k, 1 \le \mathrm{j} \le \mathrm{P}, 1 \le \mathrm{k} \le \mathrm{C}$. The mean value of j-th feature in k-th class is expressed as $\overline{X}_j^{(k)} = \frac{1}{C_k}\sum_i x_{ij}^{(k)}$. The inner-class dispersion degree function of j-th feature is:

$$\alpha_j = \frac{1}{C}\sum_k\left(\frac{1}{c_k-1}\sum_i\left(x_{ij}^{(k)} - \overline{X}_j^{(k)}\right)^2\right) \tag{1}$$

The feature of values of inner-class dispersion is feature's aggregation, and the smaller it is, the more favorable for the classification based on the idea of clustering. $Y_j = (\overline{X}_j^{(1)}, \overline{X}_j^{(2)}, \cdots, \overline{X}_j^{(k)}, \cdots, \overline{X}_j^{c})$ is expressed as the mean value of j-th feature in all levels, and the inter-class dispersion degree function of j-th feature is:

$$\beta_j = \frac{1}{C-1}\sum_k\left(\overline{X}_j^{(k)} - \frac{1}{C}\sum_k \overline{X}_j^{(k)}\right)^2 \tag{2}$$

Inter-class features' discrete degree reflects inter-class difference, and the bigger it is, the more favorable for the classification while classifying. The smaller discrete degree that has same features among the same kind is, the better it will be. The bigger discrete degree that has same features among different kind is, the better it will be. In this paper, we consider the discrete degree of feature of inner-class and inter-class at the same time, and construct an inner-class and inter-class features' discrete degree ratio function. The dispersion degree ratio function of j-th feature is:

$$R_j = \frac{\alpha_j}{\beta_j} \tag{3}$$

Calculate the ratio of all features according to function(3),rank from small to large, and delete features on inflection point's right side in an semi-supervised way.

Support Vector Machine (SVM) Recognition Model

Support vector machine is a method of processing high dimensional data and dividing small samples into more categories. When the classifier is established, we not only consider the minimum experience risk, but also consider the minimum structure risk and excellent promotional ability. The core idea is to map the vector to high dimensional space through a kernel function, to structure optimal hyper plane in high dimensional space and to make the distance of samples between different classified maximum. In this paper, the input vector is mapped into a high-dimensional vector by means of linear kernel function, and then set up a linear classifier in high dimensional space. The decision function of linear classifier is:

$$g(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i\, d_i K(x_i, x) + b\right) \tag{4}$$

$K(x_i, x)$is the kernel function, which maps the input samples from low to high dimension.$x_i$is the support vector of trained samples,x is the sample to be classified, the value of d_i is 1 or -1,which inputs correct type of sample correspondingly.

Support vector machine (SVM) is two-stage classifier. If we need multiple classifications, we need multiple classifiers to compete together. In this paper, the structure of the tree branches is adopted as is shown in Fig. 1.A level is split off at a time, the N-th classification needs to create N-1 SVM two-stage classifiers .Such as the first-level classifier classes A1 class and other classes into two types, the output is 1 for the A1 class and -1 for other classes at the end of classification. When the second-level classifier is working, it classes A2 class and other classes into two types, and so on. If the input sample belongs to N-1 class or N class, all N-1 classifiers are required to obtain the final Classification result.

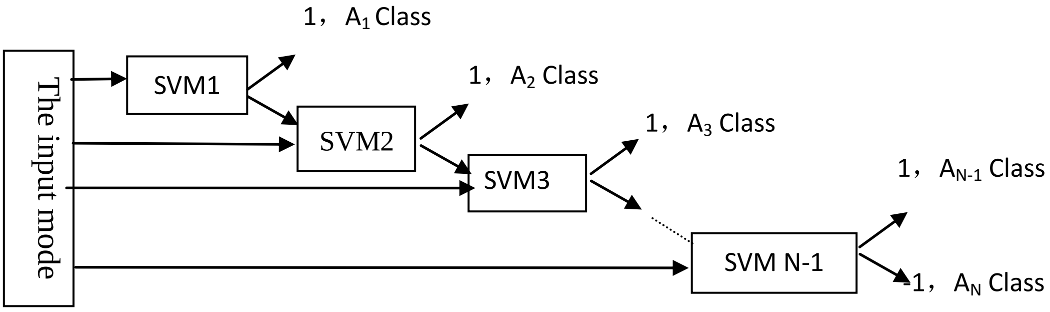

Figure 1. Fork type structure

Achieved Results and Analyses

Tobacco in 42 levels can be divided into green, greenish yellow, slightly green, lemon, orange, reddish brown, variegated and so on by color，and upper, middle, lower three grades by location. The number of leaf of the 42 levels' tobacco is 1500 in this experiment. We select one-third leaves in each level randomly as training samples, and the rest act as testing samples to verify the extension of classification model.

The image features of the extracted tobacco leaves are as follows: morphological features including length, width, aspect ratio, area, perimeter, damage rate, roundness and rectangle; Color features including RGB mean value, RGB variance, HSI mean value and HSI variance; Texture features including inertia, energy, entropy and relevance; Vein features including length, width, area ratio, RGB mean value, RGB variance, HSI mean value and HSI variance. There are 39 features in all.

Feature selection was conducted by formula (3), and the result was shown in Fig.2. The results with * in the figure were deleted features. From the figure we can see that the deleted features were rectangularity, G mean value, S mean value, I variance, inertia, venation-R mean value, venation-G mean value, venation-R variance, venation-G variance, venation-H mean value, venation S-mean value, venation-I mean value and venation-I variance. The accuracy of classification increased from 85.26% under all features to 89.65% under 26 features.

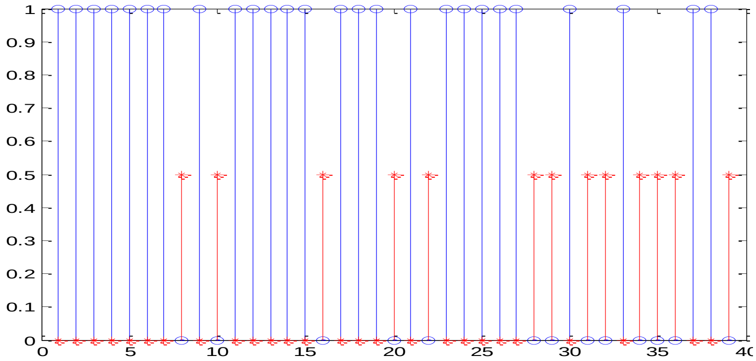

Figure 2. After screening the rest of the characteristics

After the screen by semi - supervised clustering, there remained 26 features, and the recognition rate has obviously improved. In this way it can not only reduce collection quantity of tobacco image information and the complexity of classification model, but also accelerate the speed of tobacco classification, which will provide technical support for the real-time classification of tobacco in acquisition phase.

Conclusion and Outlook

We can obtain following results through above work:

1. The tree-branch SVM classifier is a good model for tobacco classification.
2. Effective tobacco image features can be screened by semi - supervised clustering. It can not

only reduce the computational complexity of classification model, but also reduce the time of collecting data under the premise of ensuring a correct rate.

3. The near-infrared spectrum of tobacco can reflect the inner chemical composition information of tobacco. Combining images and spectral information may get a better classification effect. Hyper-spectra technology can simultaneously reflect images and spectral information, which will lay the foundation of highly-accurate classification of tobacco.

References

[1] X.Wang and L.Y.He. A Synchronous Background Segmentation Method for the Transmission and Reflection Images of Tobacco Leaves**.** Geometrics and Information Science of Wuhan University, 2014, 39(8):998~1002.

[2] K.D.Tian, K.X.Qiu and Z.H.Li,*et al.* Determination of Calcium and Magnesium in Tobacco by Near-Infrared Spectroscopy and Least Squares-Support Vector Machine. Spectroscopy and Spectral Analysis, 2014, 34(12):3262~3266.

[3] X.Ren, C.L.Lao and Z.L.Xu,*et al.* The Study of the Spectral Model for Estimating Pigment Contents of Tobacco Leaves in Field**.** Spectroscopy and Spectral Analysis, 2015, 35(6): 1654~1659.

[4] D.Q.Peng, J.Y.Shen and J.J.Jun,*et al.* Tobacco Leaves Grading with Spectrum Based on RBF Network**.** Journal of Agricultural Mechanization Research, 2009, 53(10):15~18.

[5] H.D.Zhao, J.Y.Shen and R.J.Liu,*et al.* Tobacco Leaf Selection Method of the Near-infrared Spectroscopy Effective Feature Based on the Cluster. Infrared Technology, 2013, 35(10):659~664.

[6] Y.W, X.Ma and Y.D.Wen,*et al.* Near Infrared Spectroscopy and Multivariate Statistical Process Analysis for Real-Time Monitoring of Production Process**.** Spectroscopy and Spectral Analysis, 2013, 33(5): 1226-1229.

[7] K.Yang,J.Y.Cai and C.P.Zhang,*et al.* Analysis of Tobacco Site Features Using Near Infrared Spectroscopy and Projection Model. Spectroscopy and Spectral Analysis,2014, 34(12): 3277~3280.

[8] H.L.Zhang, X.D.Sun and Y.D.Liu,*et al.* Measurement of soluble content in apples using near infrared spectroscopy . Transactions of the Chinese Society of Agricultural Engineering, 2009, 25(2): 340~344.

[9] Z.Y.Hou, W.Wang and W.S.Cai,*et al.* A local regression method based on independent component analysis and its application in near infrared spectral analysis**.** Computers and Applied Chemistry, 2006, 23(3):224-226.

[10] H.Li,H.D.Zhao and J.Y.Shen,*et al.* Screening the effective features in the near-infrared spectroscopy of tobacco leaf based on BPSO and SVM. Physics Experimentation, 2015, 35(6):8~12.

[11] H.R.Wang,W.J.Li and Y.Y.Liu,*et al.* Study on Discrimination of Varieties of Corn Using Near-Infrared Spectroscopy Based on GA and LDA. Spectroscopy and Spectral Analysis, 2011, 31(3):669-672.