Screening the Effective Spectrum Features of Tobacco Leaf Based on GA and SVM

Hang Li[1, a], Jinyuan Shen[1, b*], Yinliang Kong[2, a] and Zhongji Cheng[2, b]

[1]Department of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China

[2]Pingdingshan Branch of Henan Provincial Tobacco Company, Pingdingshan 467002, China

[a]lh121144@163.com, [b]18317828142@163.com

*The corresponding author

Keywords: Genetic algorithm; Support vector machine; Tobacco grade; Spectrum

Abstract. To improve the tobacco classification speed, it is necessary to shorten the data acquisition time and reduce the computational complexity of the hierarchy model. In this paper, we take the genetic algorithm to screen the tobacco spectrum characteristics, and set up the support vector machine (SVM) classification mode, then compared the feature selection recognition rate of 13 tobacco leaves grade before and after. The experiment results show that the recognition rate improves greatly after using genetic algorithm for feature selection, and reduce the data acquisition quantity. By using the genetic algorithm method, we can improve the classification speed of tobacco leaves grading on the premise of the correct classification rate.

Introduction

Tobacco intelligent classification has the characteristics of celerity and high accuracy, so it can avoid the subjectivity by the artificial classification. The current intelligent classification method is based on the tobacco leaf image information [1] or the spectral information primarily. The spectral information can reflect some factors related to the level of tobacco leaf like the left thickness, oil contents and the leaf structure. Now the spectrum analysis technique is widely used in the tobacco industry [2, 3].

The spectral characteristics of acquisition have the features of high dimensions and big redundancy. We should deal with the dimension reductions of spectral characteristics when it comes to the classification. The first class of methods to extract features is utilized by the principal component analysis (PCA) [4], wavelet decomposition[5], independent component analysis (ICA) [6], continuous projection[7] and the interval partial least square (IPLS)[8]to reduce the dimensions. All the methods mentioned above can reduce the input dimensions of lassifier, thereby to reduce computational complexity of classification model. But the methods above cannot shorten the original spectral data acquisition time, thus influenced the whole tobacco classification speed greatly. The second class of methods screen effective characteristic spectrum from the original spectrum directly. The methods to filter spectral characteristics including the cluster [9], binary particle swarm optimization algorithm (BPSO) [10]. We just need to collect the characteristic spectrum which had been screened when we collect the data. Using this kind of method, we can not only reduce the computational complexity of hierarchy model, but also the data acquisition quantity. Base on the second kind of method, we will take the genetic algorithm (GA) to screen the effective spectral characteristics, and then verify the characteristics by the SVM classifier. Finally, we will classify the 13 grades of the tobacco leaves.

Genetic Algorithms

Genetic Algorithms (GA) [11] is referenced to the biological natural selection and genetic mechanism, and make selection, exchange and mutation operator operation. With the constantgenetic iterations, the variables with a better objective function are retained, while those with poor variables will be eliminated, ultimately, we get the optimal results. Parameter coding: In

this paper, we use the 0/1 binary string form. As the problem with m parameters, we can representation it with a string of m*p vector. When the P value is 1, the parameters are selected, while the parameter is not selected when the value of P is 0. Initialization of the group: According to a certain limit condition (or random) produce a given initialization group, and the group size changes with the quantity of the parameters. In this paper, we define the initialization as 40. The design of adaptability function: Evaluate the merits of the individual by adaptive function, as the basis of genetic operation later. In this research, we choose the SVM recognition rate as an evaluation function. The design of genetic manipulation: In order to avoid the genetic defect, and improve the global convergence and calculation speed, we choose the high fitness individuals genetic to the next generation directly, or though crossover, mutation to produce new individual to genetic the next generations. We choose the individual by tracking wheel test in this paper. Crossover operation means two pairs of chromosomes swap part of gene in some manner, and form two new individuals. The crossover probability value is 0.6 in this paper. Variation refers to the complementary operation to certain genes in the individual coding string. That means 1turn to 0, or 0 turn to 1.

Convergence criterion: In order to avoid going into the infinite loop, we terminate the operation when the genetic iterations reaches 200.

Classification Model

Support vector machine (SVM) is a good way to deal with high-dimensional data and make more small-sample classification. When establish a classifier, the conditions we considered not only the empirical risk and the structural risk should reach the minimum, but also it should have good generalization ability. The main idea is to map the vector to a high-dimensional space, then construct the optimal hyperplane in the high-dimensional space, and make the sample's interval largest among the different categories. Through the linear kernel function, the input mode can achieve the high dimensional vector map, and then build a linear classifier in the high-dimensional space. The decision function of linear classifier is as follows as Eq.1.

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

In the Eq.1, $K(x_i, x)$is the kernel function, x_i is the training sample support vector, x is a test sample, $b$is the threshold which determined byx_i , d_i is the label of the training samples, and $\alpha_i$is the Lagrange multiplier. We need to constructed $N(N-1)/2$ classifiers (N is the quantity of categories), and make the multi-category classification through parallel voting. The voting type of classification is shown in Fig. 1.

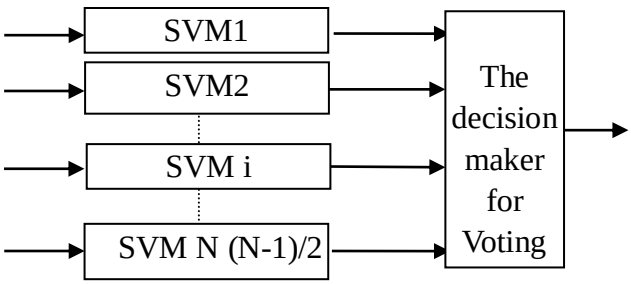

Figure 1. The classifier of parallel voting

The Experimental Data

The 13 levels of tobacco leaf is B2F, B3F, B4F, C2F, C2L, C3F, C3L, X2F, X2L, X3F, X3L, X4F, and X4L respectively. We collect every piece of tobacco leaf's reflectance spectrum by UV3600 models of spectrometer produced by SIMADZU. There are 642 reflection spectrum, the spectral range is 1500~2400nm with the interval of 2nm. We select a third of the samples randomly for

training set, and verify the generalization ability to the model with the rest of the samples. To remove the baseline drift and noise brought by the spectrometer, the collected data was preprocessed by the following Eq.2.

$$y_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \quad (2)$$

In the Eq.2, x_i is the original spectrum without pretreatment.

Results Analysis

All the accuracy of classification, quantity of features, and the time-consuming of grading selected by the genetic algorithm before and after are shown in the Table 1.

Table 1 The result of feature selection and classification

Interval *nm*	Accuracy %		Number of features		Grading time-consuming(*s*)	
	Before	After	Before	After	Before	after
2	90.88	94.39	451	280	0.153	0.088
4	89.49	93.22	226	129	0. 072	0.043
6	89.01	92.76	151	78	0.055	0.03
8	88.55	89.48	113	60	0.046	0.027
10	87.15	89.25	91	46	0.037	0.025

We can draw the following conclusions from Table 1: (1) the recognition rate improves obviously after using genetic algorithm for feature selection. (2) Both the quantity of characteristic and the time of grading reduced greatly. The result shows that fewer characteristics can decrease the acquisition quantity and speed up the real-time classification.

Summary

From what has been researched, we can safely come to the conclusion that the genetic algorithms can select the effective features from the original spectral, and the SVM can be well used for tobacco classification. Feature selection can reduce the quantity of data acquisition effectively, so as to make the tobacco leaves of real-time classification with practical possible.

References

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

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

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

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

[5] D.Q.Peng, J.Y.Shen and J.J.Liu. Tobacco Leaves Grading with Spectrum Based on RBF Network [J]. Journal of Agricultural Mechanization Research, 2009, 53(10):15-18.

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

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

[8] H.L.Zhan g, X.D.Sun and Y.D. Liu. Measurement of soluble solid content in apples using near infrared spectroscopy [J]. Transactions of the Chinese Society of Agricultural Engineering, 2009, 25(2): 340-344.

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

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

[11] Q. M. Kong, Z.B. Su and W.Z.Shen. Research of Straw Biomass Based on NIR by Wavelength Selection of IPLS-SPA [J]. Spectroscopy and Spectral Analysis, 2015, 35(5):1233-1238.