Automatic discrimination planting areas of flue-cured tobacco based on near-infrared spectroscopy technology and support vector machine improved by whale optimization algorithm

QIU Changgui[1,2], LIU Ze*[3], QI Lin[4], YANG Jingjin[4], WANG Xianguo[4], LIU Jihui[4], WENG Ruijie[4], WEI Qing[1], LIU Jing[1,2], YANG Panpan[1,2], LI Siyuan[4]

1. Yunnan Reascend Tobacco Technology (Group) Co., Ltd., Kunming 650106, China 2. Yunnan Comtestor Co., Ltd., Kunming 650106, China
3. China Tobacco Yunnan Industrial Co., Ltd., Kunming 650231, China 4. HongyunHonghe Tobacco (Group) Co., Ltd., Kunming 650231, China

Abstract

In order to accurately and rapidly identificate planting areas of flue-cured tobacco. A total of 201 flue-cured tobacco samples from three different areas in Kunming, Honghe and Qujing, Yunnan Province were selected for the study, After collecting the near-infrared spectra of different areas and reducing the interference factors through the spectral preprocessing method, followed by principal component analysis (PCA) for dimensionality reduction, and then a whale algorithm (WOA) was established to optimize support vector machine (SVM) parameters to establish an automatic identification method. Results: In the wavenumber range of 8 000 to 4 000 cm^{-1}, the standard normal variable transformation (SNV) combined with the second derivative method (2D) is used for near-infrared spectroscopy preprocessing, and the data after the principal component dimensionality reduction was used as the input variable, and the SOA-optimized support vector parameters could achieve a better recognition effect.

As a result, the classification accuracy rate of the training set is 97.18%, and the classification accuracy rate of the test set is 98.31%. Conclusion: It shows that using near-infrared spectroscopy technology combined with WOA algorithm to optimize SVM can achieve accurate identification of area of flue-cured tobacco.

Objective

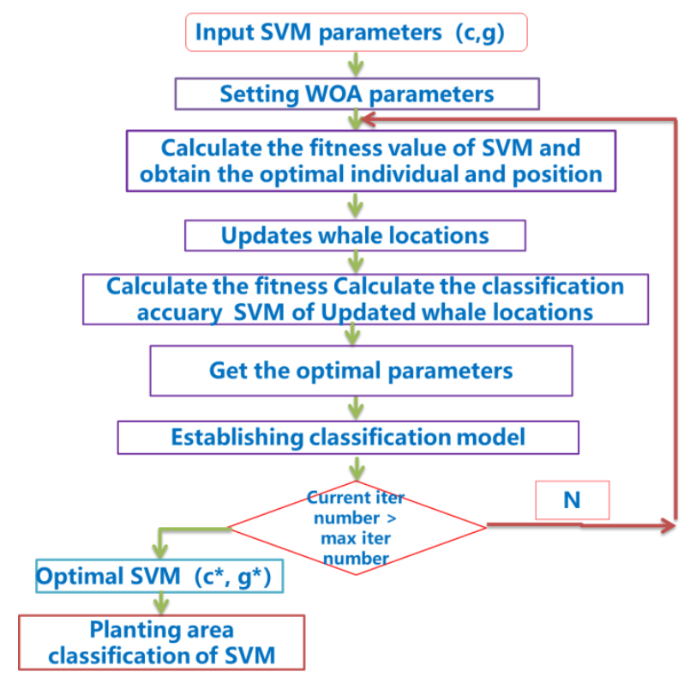
The pattern classification method of SVM parameters was optimized by WOA to identify the difference of flue-cured tobacco characteristics in three major flue-cured tobacco planting areas in Yun-nan Province, aiming to establish a fast and effective method to identify flue-cured tobacco origin, and provide a theoretical basis for the accurate identification of flue-cured tobacco quality characteristics, geographical origin trace-ability, Orientation of Style and Characteristics of Tobacco Leaves

Material and Methods

- A total of 201 samples of C3F primary flue-cured tobacco from Kunming, Honghe and Qujing were used in the experiment. Among them, 71 tobacco samples were from Kunming, 85 were from Honghe and 45 were from Qujing.
- A mean spectrum was then calculated for each sample by averaging the triplicate spectra
- PCA helps to reduce the computational complexity of the model
- The core idea of SVM is to find a hyperplane in space that minimizes the classification error rate.
- WOA has the advantages of fewer adjustment parameters, convergence accuracy, and superiority seeking search ability。

Results and Discussion

- In Table 2, SNV+SD pretreatment method had the highest classification accuracy of training set and test set, and the average classification accuracy of training set and test set were 100.00% and 98.33%, respectively. MSC pretreatment method had the worst classification effect.
- In Fig. 2, when the number of principal component factors is 25, the accuracy of the training set is the highest, which is 99.30%; when the number of principal component factors is increased, it remains unchanged; when the number of principal component factors is 29, the accuracy of the test set is the highest, which is 91.53%. That is, when the number of principal component factors is 29, the best WOA-SVM classification model can be obtained.
- In Fig. 3, When the number of iterations is 2, the optimal fitness degree begins to stabilize and becomes stable at 95.77%, while the average fitness degree becomes stable at 54 iterations and becomes stable at 95.77%. It is shown that the combination of two parameters (penalty parameter and kernel function parameter) of SVM achieves the optimal performance, that is, the best penalty parameter c=100 and the best kernel function parameter =100.
- In Fig. 4, when the optimized parameters were used for classification, the classification accuracy of the training set was 97.18%, and the classification accuracy of the test set was 98.31%.


Input SVM parameters (c,g)

Setting WOA parameters

Calculate the fitness value of SVM and obtain the optimal individual and position

Updates whale locations

Calculate the fitness Calculate the classification accuary SVM of Updated whale locations

Get the optimal parameters

Establishing classification model

Current iter number > max iter number

N

Optimal SVM (c*, g*)

Planting area classification of SVM

Fig.1. Using the whale algorithm to optimize support vector machine parameter flow chart

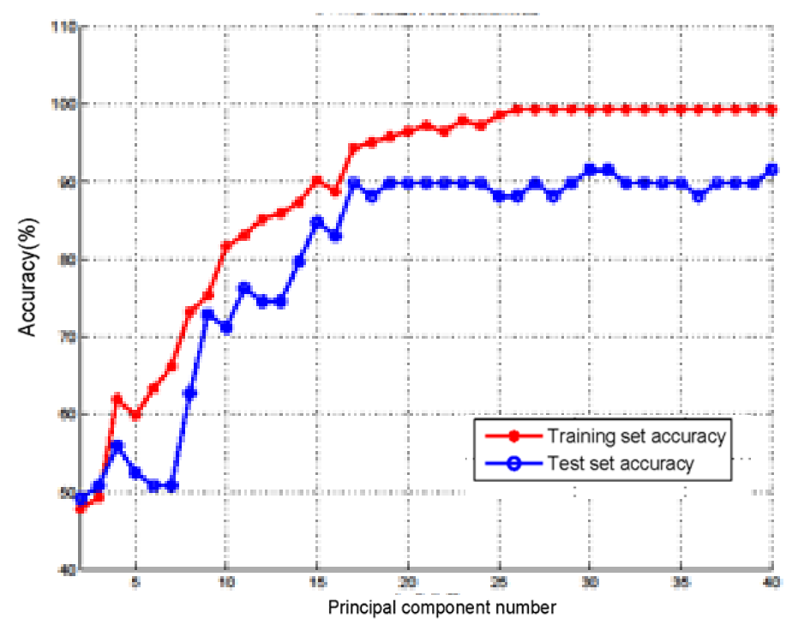


Fig.2. Classification results of WOA-SVM model under different PCs

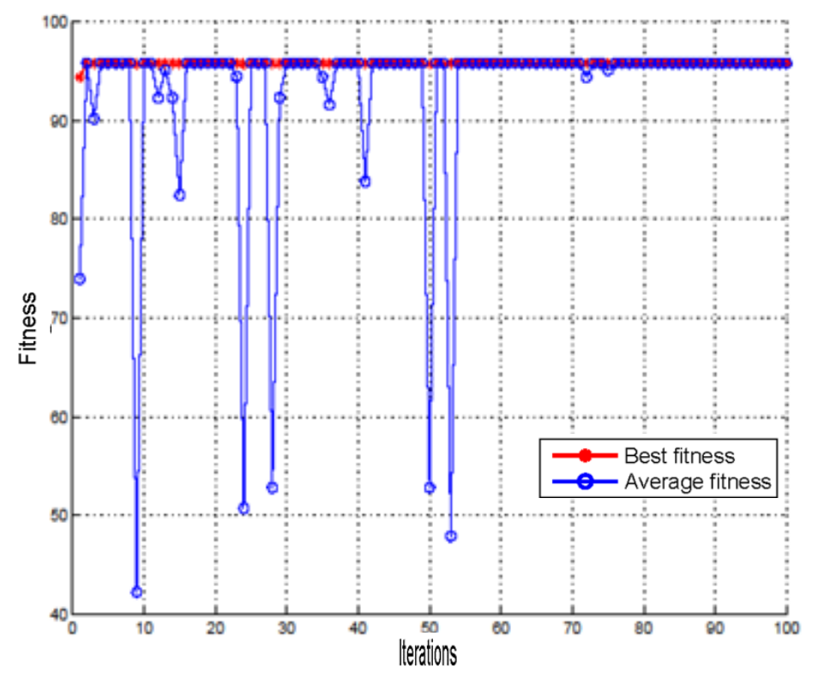


Fig.3 WOA-SVM algorithm fitness optimization process curve

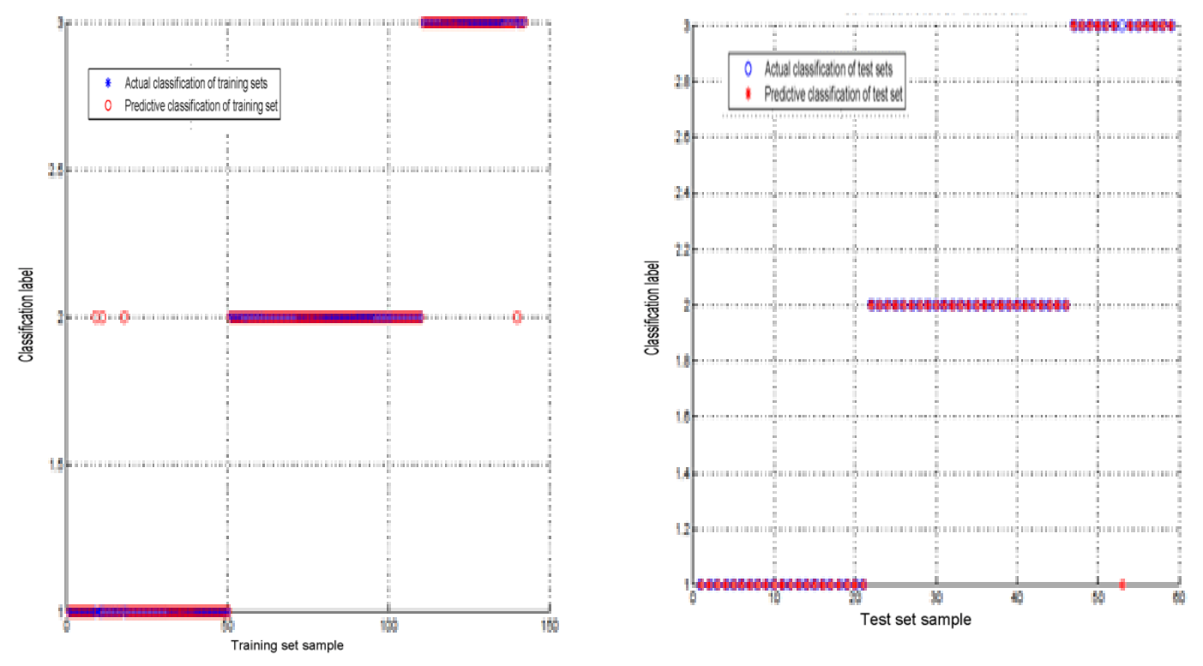


Fig.4. Classification effect diagram of training set and test set

Table 1 Division of the sample set of flue-cured tobacco in 3 planting area

Planting area	Training set	Test set
Honghe	50	21
Kunming	60	25
Qujing	32	13

Table 2 Recognition results of WOA-SVM with different preprocessing method

Preprocessing	nLV	Classification accuracy (number, percent) of training set/%	Classification accuracy (number, percent) of testing set/%
Raw spectra	4	1109/142=76. 76	37/59=62. 71
SNV	8	142/142=100. 00	38/59=64. 41
MSC	2	106/142=74. 65	36/59=61. 02
SNV+FD	29	139/142=97. 89	52/59=88. 14
MSC+FD	29	96/142=67. 61	37/59=62. 71
SNV+SD	29	138/142=97. 18	58/59=98. 31
MSC+SD	29	60/142=42. 25	25/59=42. 37
SNV+FD+SG(15:3)	18	136/142=95. 77	46/59=77. 97
MSC+FD+SG(15:3)	18	93/142=65. 49	37/59=62. 71
SNV+SD+SG(15:3)	29	85/142=59. 86	33/59=55. 93
MSC+SD+SG(15:3)	29	60/142=42. 25	25/59=42. 37

Conclusions

1. The standard normal variable transformation (SNV) combined with the second derivative (SD) was used to pretreat the near infrared spectrum, then principal component analysis (PCA) was used to reduce dimension
2. Whale algorithm optimization support vector machine algorithm combined with near infrared spectroscopy technology can achieve accurate identification of flue-cured tobacco origin.
3. The correct identification rate of flue-cured tobacco training set was 97.18%. The correct recognition rate of the test set was 98.31%.

References

1. Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.
2. Alex J. Smola, Bernhard Schölkopf. A tutorial on support vector regression [J]. Statistics and Computing, 2004, 14(3): 199-222.
3. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. Advances in Engineering Software 2016, 95, 51–67. DOI: 10.1016/j.advengsoft. 2016. 01. 008.