http://dx.doi.org/10.21577/0103-5053.20190105

J. Braz. Chem. Soc., Vol. 30, No. 9, 1927-1932, 2019
Printed in Brazil - ©2019 Sociedade Brasileira de Química

Article

Rapid and Automatic Classification of Tobacco Leaves Using a Hand-Held DLP-Based NIR Spectroscopy Device

Zhang Jianqiang,[a] Yang Panpan,[b,c] Liu Weijuan,[a] Yang Yanmei,[,a] Yuan Tianjun,[b] Hou Ying[c] and Li Changyu[a]*

[a]*Yunnan Reascend Tobacco Technology (Group) Co., Ltd, 650000 Kunming, China*

[b]*Yunnan Comtestor Co., Ltd, 650000 Kunming, China*

[c]*Southwest Forestry University, 650000 Kunming, China*

A hand-held near infrared (NIR) spectroscopy device is much more convenient than a traditional desktop NIR instrument. Thus, it is more suitable for the practical application. An automatic and rapid tool for grading tobacco leaves on the spot using a hand-held digital light processing (DLP)-based NIR spectroscopy device is proposed in this paper. Firstly, the spectral data of the samples is scanned with a hand-held NIR device directly from the tobacco leaves without any samples preparation procedures. Then, the training model of different classes is built and the class of each test sample is predicted by using sparse representation classification (SRC) algorithm. Comparing with the traditional linear discriminant analysis (LDA) and support vector machine (SVM) algorithms, the classification accuracy of SRC method is the highest and has the least computation time. The results show that hand-held NIR spectroscopy technology could be a novel classification tool for grading tobacco leaves in the purchasing on the spot.

Keywords: hand-held DLP-based NIR, sparse representation classification, automatic and rapid classification, practical application

Introduction

Flavour and aroma are two important smoking characters in tobacco and they are mainly depended on the classes of tobacco leaves. In China, the main method of grading tobacco leaves is depended on the position and the color. Usually, the positions of a tobacco plant are divided into lower (X), middle (C) and upper (B) portion three categories. The color is divided into red-brown (R), orange (F) and lemon (L) three different hue categories. The group of different classes of tobacco leaves is firstly formed by combination of the position and color categories and then each group will be further divided into 3 or 4 grades according to the quality. However, the most grading work is mainly manually operated by the experience of tobacco experts at present, which is slow, laborious and not objective. The efficiency and the stability are also dissatisfying.[1] Therefore, it is necessary to develop a new method which is fast, high-efficiency and more objective.

*e-mail: yangyanmei@reascend.com

Near infrared (NIR) spectroscopy is a useful analytical chemistry tool[2] and it has the advantages such as accurate, fast, and non-destructive. It has been widely used in various fields such as agriculture, medical, oil, tobacco and so on.[3-5] Bin *et al.*[6] and Wang *et al.*[7] have already used NIR spectroscopy technology in the classification of tobacco leaves and the classification results are satisfying. However, the instrument they used is the traditional desktop NIR spectroscopy device, which is expensive, big and cannot be brought to the purchasing spot. Comparing with the traditional desktop NIR device, a hand-held NIR spectroscopy device is small, cheap, flexible and much more convenient. It is suitable for the tobacco purchasing spot.[8] However, the spectrum range of the hand-held NIR (900-1700 nm) is usually smaller than the traditional desktop NIR instrument (1000-2500 nm), how to build a model that is fit for the device is also of great importance.

As one of the deep learning methods, SRC algorithm can capture the essential of the signal or data and has been widely used in signal de-noising,[9,10] decoding,[11] compressed sensing[12] and machine learning.[13,14] Here it is mainly focused on how to classify the spectral data which is acquired by using

a hand-held NIR device. The SRC (sparse representation classification) algorithm was proposed by Li and Ngom[15,16] and it can reflect the relevance between feature and target under the control of other features. It is robust to noise and suitable for the classification of hand-held NIR spectral data.

In this paper, a novel SRC method of tobacco leaves based on hand-held NIR spectroscopy technology is proposed. The method makes the large tobacco leaf NIR spectral data to be sparse and adopts a transformation matrix to reduce dimensionality. With the advantage of de-noising, avoiding overfitting, and being convenient in practice, it is shown that this method could improve the efficiency of the classification of tobacco leaves in the purchasing on the spot.

Experimental

Samples

In the following research, two experimental sets from different locations will be chosen. The first one was harvested in 2018 from Zhanyi City, Yunnan Province of China. It has 210 pieces of tobacco leaves and contains X2F, C3F and B2F three different categories. Each category has the same number. The second one was harvested in 2018 from Luoping City, Yunnan Province of China. It has 320 pieces of tobacco leaves and contains X2F, C2F, C3F, B1F and B2F five different categories. Each category has the same number. In the experimental process, we chose half of the total samples randomly as the training set and the test set for the two data sets respectively by means of using 2-fold cross-validation method.[17] The details of the two data sets are shown in Table 1.

Equipment

The DLP NIRscan Nano is an ultra-portable spectrometer evaluation module utilizing DLP (digital light processing) technology to meet lower cost, smaller size, and higher performance than traditional architectures. The replacement of a linear array detector with DLP digital micromirror device (DMD) in conjunction with a single point detector adds the functionality of programmable spectral filters and sampling techniques that were not previously available on NIR spectrometers. The detail of the technology different from the current one is shown in Figure 1. The device was provided by Texas Instruments.

Table 1. The detail of data set 1 and data set 2

Data	Year	Location	Class[a]	Feature	Total of samples	Total of training samples	Total of test samples
Data set 1	2018	Zhanyi, Yunnan	X2F	320	70	35	35
	2018	Zhanyi, Yunnan	C3F	320	70	35	35
	2018	Zhanyi, Yunnan	B2F	320	70	35	35
Data set 2	2018	Luoping, Yunnan	X2F	320	64	32	32
	2018	Luoping, Yunnan	C2F	320	64	32	32
	2018	Luoping, Yunnan	C3F	320	64	32	32
	2018	Luoping, Yunnan	B1F	320	64	32	32
	2018	Luoping, Yunnan	B2F	320	64	32	32

[a]The different classes of tobacco leaves, formed by combination of the position and color categories, followed by division into grades according to the quality. X: lower; C: middle; B: upper; F: orange.

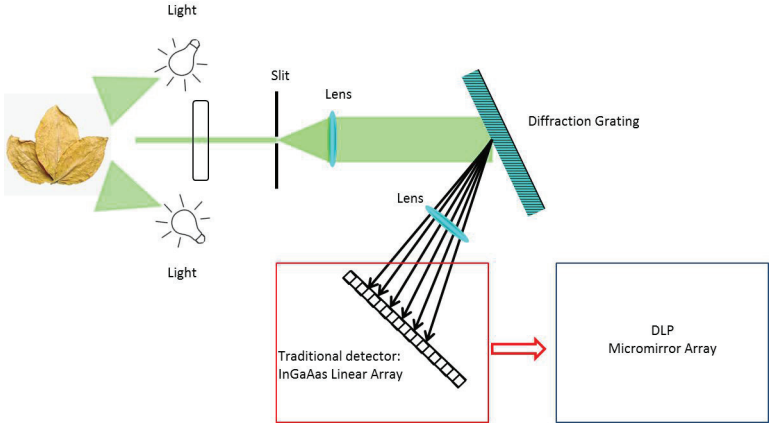

Figure 1. Comparison of DLP-based NIR technology and current NIR technology.

The spectral data was collected with a hand-held DLP-based spectroscopy device with 16 scans co-added. The spectral range of the device is from 900 to 1700 nm and the resolution is 5.85 nm. A total of five absorbance spectral data were collected on each tobacco leaf sample (shown in Figure 2). The final spectral data of each sample was calculated by averaging the five spectra. The NIR spectral data of the two sets collected by hand-held NIR device can be seen in Figures 3a and 3b.

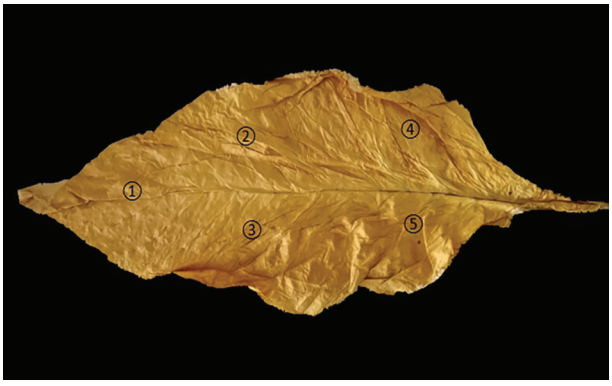

Figure 2. Positions of NIR scan performed on each tobacco leaf sample.

Theory of SRC algorithm

Sparse representation (SR)[9] is a parsimonious principle that a sample can be approximated by a sparse linear combination of basis vectors.[15] Non-orthogonal basis vectors can be learned by SR, and the basis vectors may be allowed to be redundant. There are two techniques in SR. First, given a basis matrix, learning the sparse coefficients of a new sample is called sparse coding. Second, given training data, learning the basis vector is called dictionary learning. It can be statistically formulated as

$$(\boldsymbol{b}|\mathbf{A},\boldsymbol{x},k) = x_1 a_1 + \ldots + x_k a_k + \varepsilon = \mathbf{A}\boldsymbol{x} + \varepsilon \quad (1)$$

In equation 1, $\boldsymbol{b}$ is the training data with samples, $\mathbf{A} = [a_1, \ldots, a_k]$ is defined as a dictionary. a_i is one of the dictionary atoms, $\boldsymbol{x}$ and ε represent the sparse coefficient vector and the error term respectively, k is the model parameter. SR model has the following constraints: (*i*) the error term ε is Gaussian distributed with mean zero and isotropic covariance. (*ii*) The dictionary atoms is usually Gaussian distributed, the coefficient vector should follows a sparsity-inducing distribution. (*iii*) $\boldsymbol{x}$ is independent of ε.

As mentioned above, a SR model usually involves two steps, one is sparse coding and the other is dictionary learning. If all the coefficient vectors of the training dictionary are all non-negative, it can be formulated as the following mode:

$$\min_{x} \frac{1}{2}\|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2^2 \; st. \boldsymbol{x} \geq 0 \quad (2)$$

The sparse coding model of equation 2 is called non-negative least squares (NNLS).[16] NNLS has the advantages of easy interpretable and convenient in practice. It is suitable for dealing with the NIR spectral data. Therefore, the classification of hand-held spectral data based on SRC algorithm has the following steps: (*i*) all the spectral data of the samples are collected by using a hand-held NIR device and they are divided into the training samples and test samples randomly; (*ii*) all the training samples are used to build a training dictionary by using NNLS algorithm; (*iii*) the test samples will be expressed via the above training sparse representation dictionary; (*iv*) the class of a test sample is assigned with the minimum residual by computing the regression residual.

Results and Discussion

Firstly, pre-processing operation is implemented on the original spectral data. Here multiplicative scatter

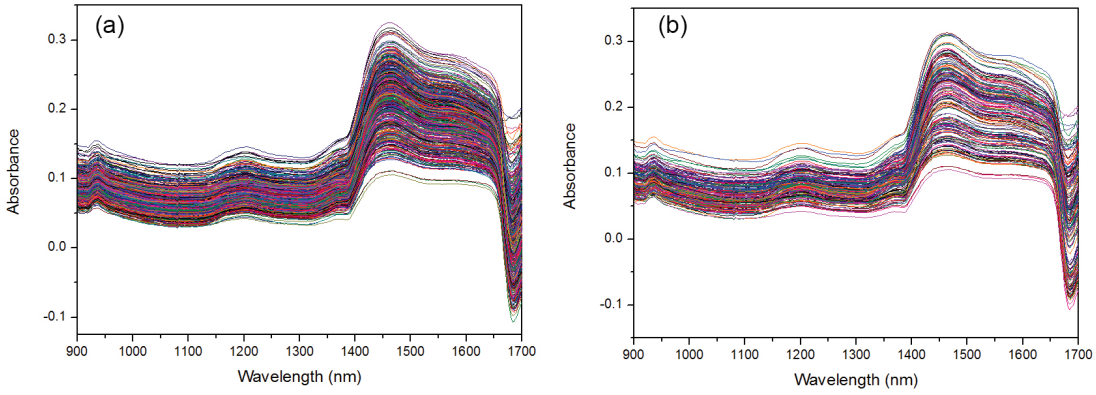

Figure 3. The original spectral data of different classes of tobacco leaves acquired by using hand-held DLP-based NIR device; (a) and (b) are the original spectral of the data set 1 and 2, respectively.

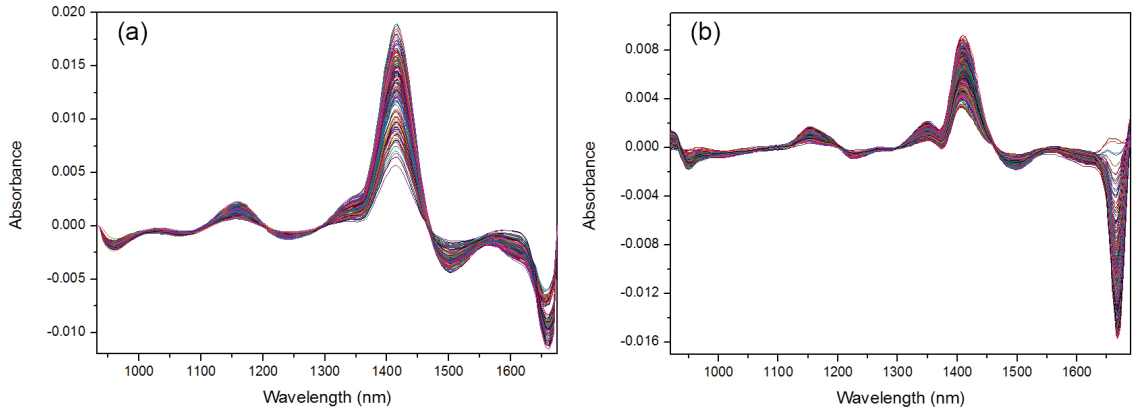

Figure 4. Pre-processing results. (a) The result of data set 1, and (b) the result of data set 2.

correction and gap-segment (first derivative, 5-point window, two polynomial order) methods[18] are chosen. The pre-processing results are shown in Figure 4. The pre-processing results show that the resolution of the spectral data has been improved.

After pre-processing, the dimension of pre-processing data is still huge. Therefore, principal component analysis (PCA)[19] is chosen to reduce the dimension of the data. Table 2 shows the cumulative percentage of variance for the first six principal components of the two data sets and the results indicate that these six PCs can describe the two data sets well. After PCA operation, the dimensionality of the data set 1 has been reduced from 210 × 320 to 210 × 6 and data set 2, from 320 × 320 to 320 × 6.

As mentioned in Experimental section ("Samples" sub-section), two-fold cross-validation method was used to partition all the samples into training and test samples equally. LDA[20] SVM[21,22] and SRC[15,16] algorithms were used to solve the classification problem. We use radial basis function (RBF) as the kernel for SVM algorithm. Besides, particle swarm optimization (PSO) algorithm[23] is used to optimize the accuracy of SVM classifier by randomly generating the parameters and estimate the best value for regularization of kernel parameters for SVM model. All the classifiers ran on the same training and test splits for fair comparison. In order to guarantee every sample would be chosen, all the algorithms were calculated 10 times for each data set. The comparing results are shown in Tables 3 and 4.

It can be seen from Tables 3 and 4 that both, the single and total correct classification numbers or accuracy of SRC algorithm are comparable with LDA and SVM algorithms, especially for the close classes of the same position and color (C2F and C3F, B1F and B2F). The main reason is that the principles of PCA-LDA and PCA-SVM algorithms use the minimum Euclidean distance of the feature space to classify the samples. The classification results are ineffective if the classes and features are close. However, the principle of SRC classification algorithm

Table 2. PCA results of the two data sets

Data set	PC1[a] / %	PC2[a] / %	PC3[a] / %	PC4[a] / %	PC5[a] / %	PC6[a] / %
Data set 1	88.53	98.43	98.89	99.35	99.65	99.78
Data set 2	80.64	97.34	99.47	99.72	99.84	99.88

[a]Cumulative percentage of variance. PCA: principal component analysis.

Table 3. Classification results of data set 1 using LDA, SVM and SRC algorithms

Class[a] (Test samples)	LDA Average classification accuracy / %	SVM Average classification accuracy / %	SRC Average classification accuracy / %
X2F (35 × 10)	331 / 350, 94.57	332 / 350, 94.86	341 / 350, 97.43
C3F (35 × 10)	310 / 350, 88.57	303 / 350, 86.57	307 / 350, 87.71
B2F (35 × 10)	327 / 350, 93.43	336 / 350, 96.00	350 / 350, 100.00
Total average accuracy	968 / 1050, 92.19	971 / 1050, 92.47	998 / 1050, 95.05

[a]The different classes of tobacco leaves, formed by combination of the position and color categories, followed by division into grades according to the quality. X: lower; C: middle; B: upper; F: orange. LDA: linear discriminant analysis; SVM: support vector machine; SRC: sparse representation classification.

Table 4. Classification results of data set 2 using LDA, SVM and SRC algorithms

Class[a] (Test samples)	LDA Average classification accuracy / %	SVM Average classification accuracy / %	SRC Average classification accuracy / %
X2F (32 × 10)	266 / 320, 83.12	290 / 320, 90.62	297 / 320, 92.81
C2F (32 × 10)	210 / 320, 65.62	282 / 320, 88.12	266 / 320, 83.12
C3F (32 × 10)	292 / 320, 91.25	301 / 320, 94.06	283 / 320, 88.43
B1F (32 × 10)	183 / 320, 57.19	229 / 320, 71.56	278 / 320, 86.88
B2F (32 × 10)	112 / 320, 35.00	164 / 320, 51.25	281 / 320, 87.81
Total average accuracy	1063 / 1600, 66.44	1266 / 1600, 79.12	1405 / 1600, 87.81

[a]The different classes of tobacco leaves, formed by combination of the position and color categories, followed by division into grades according to the quality. X: lower; C: middle; B: upper; F: orange. LDA: linear discriminant analysis; SVM: support vector machine; SRC: sparse representation classification.

is using the redundancy characteristic of the dictionary to classify the training samples and test samples. It can capture the essential features of the signal or data and has the strong robustness to noise. As the result, different from LDA and SVM algorithms, it can also achieve the effective classification results even the classes and features are close.

The aim of the paper is how to grade tobacco leaves in the purchasing process on the spot using a hand-held DLP-based NIR device, we want to achieve the classification results as soon as possible. Therefore, computation time is also a very important factor for the field use as it needs to get the classification results immediately. Then, the averaged computation time of each algorithm is recorded and the result is shown as Figure 5. All experiments of the two data sets are performed on an Intel laptop computer (Core TM i7-6700, 3.70 GHz, CPU with 8 GB RAM, with 64-bit Windows 10 Professional operation system). It can be easily seen from Figure 5 that SRC classification algorithm is much more efficient than the other two methods.

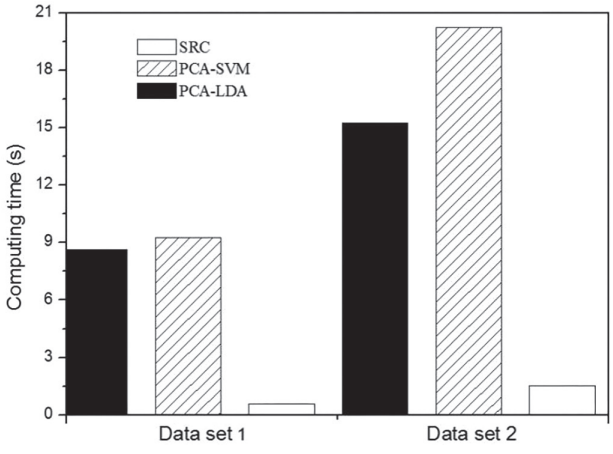

Figure 5. The elapsed execution time of three classifiers.

Conclusions

The paper proposes a novel classification tool based on hand-held NIR technology to grade different classes of tobacco leaves. The experimental results show that SRC algorithm works better than LDA and SVM algorithms on both classification accuracy and computation efficiency. The results suggest that a hand-held DLP-based NIR device could be an effective tool for grading tobacco leaves in the purchasing process on the spot.

Acknowledgments

This work was supported by China Postdoctoral Science Foundation (2017M623322XB).

References

1. Bin, F.; Jun, F.; Fan, W.; Zhou, J. H.; Yun, Y. H.; Liang, Y. Z.; *RSC Adv.* **2016**, *36*, 30353.
2. Chaucharda, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V.; *Chemom. Intell. Lab. Syst.* **2004**, *71*, 141.
3. Osborne, B.; Fearn, T.; *Near Infrared Spectroscopy in Food Analysis*; Wiley: New York, USA, 1986.
4. Duarte, F. J.; *Tunable Laser Applications,* 2nd ed.; CRC Press: New York, USA, 2008.
5. Nicolai, B. M.; Beullens, K.; Bobelyn, E. J.; Peirs, A.; Saeys, W.; *Postharvest Biol. Technol.* **2007**, *46*, 99.
6. Bin, J.; Pang, H. R.; Xie, G. Y.; Qin, G. X.; Yin, D. H.; Du, W.; Wang, B.; *Acta Tab. Sin.* **2017**, *23*, 60.
7. Wang, Y.; Ma, X.; Wen, Y. D.; Yu, C. X.; Wang, L. P.; Zhao, L. L.; Li, J. H.; *Spectrosc. Spectral Anal.* **2013**, *33*, 78.
8. Camps, C.; Christen, D.; *J. Food, Agric. Environ.* **2009**, *7*, 394.
9. Hinton, G. E.; Osindero S.; Teh, Y. W.; *Neural Comput.* **2006**, *18*, 1527.
10. Bengio, Y. In *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, vol. 7700; Montavon, G.; Orr, G. B.; Müller, K. R., eds.; Springer: Berlin, Heidelberg, 2012, p. 437-478.
11. Hinton, G. E.; *Trends Cognit. Sci.* **2007**, *11*, 428.

12. Bengio, Y.; Courville, A.; Vincent, P.; *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798.
13. Yoshua, B.; *Found. Trends Mach. Learn.* **2009**, *2*, 127.
14. Ivakhnenko, A. G.; *IEEE Trans. Syst. Man, Cybern.* **1997**, *4*, 364.
15. Li, Y.; Ngom, A.; *BMC Syst. Biol.* **2013**, *4*, 1.
16. Li, Y.; Ngom, A.; *Neurocomputing* **2013**, *118*, 41.
17. Liu, W. Y.; Han, J. G.; *Neural Comput.* **2013**, *108*, 31.
18. Asmund, R.; Berg, F. V.; *TrAC, Trends in Analytical Chemistry* **2009**, *28*, 1201.
19. Lin, Y.; Li, W.; Xu, J.; *Int. J. Pharm.* **2015**, *488*, 120.
20. Guo, Y.; *Biostatistics* **2007**, *8*, 86.
21. Vapnik, V.; *The Nature of Statistical Learning Theory*; Springer-Verlag Press: New York, USA, 1995.
22. Vapnik, V.; *Statistical Learning Theory*; Wiley Press: New York, USA, 1998.
23. Ardjani, F.; Sadouni, K.; *IJMECS* **2010**, *2*, 32.

Submitted: February 1, 2019
Published online: May 21, 2019