Discrimination of Chinese Herbal Medicine by Machine Olfaction

Dehan Luo*, Yawen Shao
School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, P.R China.
*Corresponding author, e-mail: dehanluo@gdut.edu.cn

Abstract

"Small Sample Size" (SSS) problem would occur while using linear discriminant analysis (LDA) algorithm with traditional Fisher criterion if the within-class scatter matrix is singular. The combination of maximum scatter difference (MSD) criterion and LDA algorithm for solve SSS problem is described. It is employed to detect three kinds of Chinese herbal medicines from different growing areas by machine olfaction. Compared with PCA or PCA + LDA algorithm, the classification result was enhanced. It works out that only a few samples of Anhui Atractylodes are classified incorrectly, however, the classification rate reaches 97.8%.

Keywords: *linear discriminant analysis; maximum scatter difference criterion; Chinese herbal medicine; machine olfaction*

1. Introduction

Atractylodes is an asteraceae medicine with special smell, and their quality is affected by place of origin, harvest time, breed and other factors, among the origin factors is one of the most important criteria in judging the quality. With people's increasing quality requirements of Chinese herbal medicine, the identification of medicinal herbs is particularly important.

Research of electronic nose began in the 1990s, it is a part of the specificity with the composition of the gas sensor array and pattern recognition system is composed of the appropriate instruments, mainly used to identify simple and complex odors [1]. There are a lot of researches and social applications in the food industry [2-4], medical diagnostics [5-7], and environmental monitoring [8-10] at home and abroad, but in the field of Chinese herbal medicines are rarely reported in the current.

The Chinese herbal medicine Atractylodes is the object in this paper, and detected by electronic nose. In pattern recognition with the electronic nose, the principal component Analysis (Principal Component Analysis, PCA) and LDA has been widely used [8]. The outstanding feature of LDA is it can ensure that after the projection, model sample has the smallest within-class distance and maximum between-class distance in the new space, that model has the best separability in the space. However, there is also not applicable in the "small sample problem" and other shortcomings. In response to this shortcoming, many scholars have used a method of combination with PCA and LDA [9], the advantages of the PCA and LDA together fully integration, and it can not only solve the problem of PCA algorithm is not sensitive to the different training sample data problem, but also LDA algorithm when the within-class scatter matrix is singular, and obtain a better classification results. In this paper, maximum scatter difference criterion and LDA will be integrated together, it solved the problem of small samples, and there is a better classification result than PCA and PCA + LDA.

2. Research Method

2.1. Electronic nose (E-nose)

Experiments were performed with a commercial E-nose (PEN3). It is provided by WMA AIRSENSE Analysentechnik GmbH (Schwerin, Germany). Table 1. summarizes the sensitivity of different sensors in PEN3.

Received August 20, 2012; Revised December 22, 2012; Accepted January 11, 2013

PEN3 included an array of 10 different MOS sensors, and the sensor response is defined as the ratio of conductance: G/G0. Where, G represents the resistance of each sensor in the chamber after exposing to the target gas and G0 represents the resistance while each sensor is exposed to the zero gas filtered by active carbon. The electronic nose consists mainly of the following sections: computer、sampling channel、sensor channel, as showed in Figure 1.

Table 1. The sensitivity list of 10 sensors in PEN3

Number in array	Sensor name	Sensitive to
S1	W1C	Aromatic components
S2	W5S	Nitrogen oxides, very sensitive
S3	W3C	Ammonia and aromatic components
S4	W6S	Mainly hydrogen, selectively, (breath gases)
S5	W5C	Alkanes and aromatic components
S6	W1S	Propane
S7	W1W	Sulfur organic compounds
S8	W2S	Ethanol
S9	W2W	Aromatic components and organic-sulfides
S10	W3S	Propane (selective sometimes)

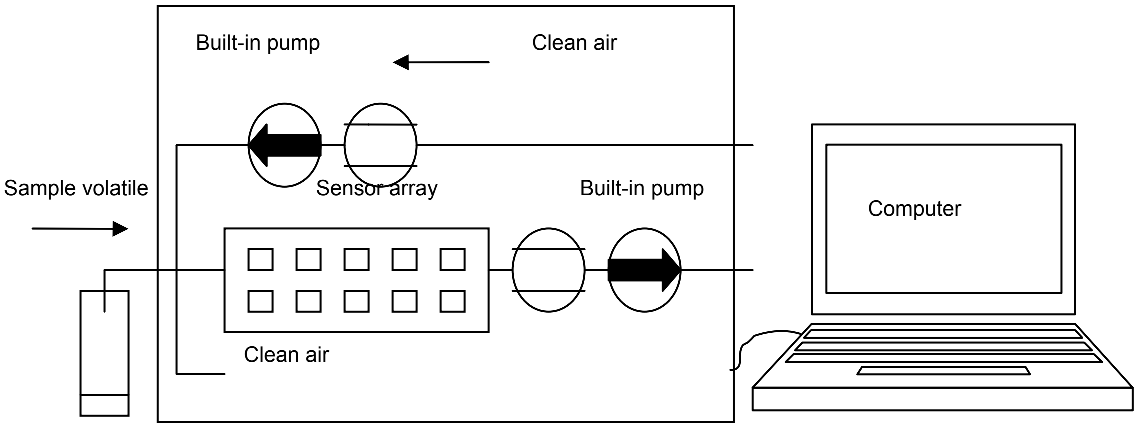


Figure 1. Diagrammatic Layout of Electronic Nose

2.2. Experimental sample

This medicine sample is supported by Guangzhou University of Chinese Medicine. They are irregular clumps of hypertrophy, gas fragrance, sweet and slightly acrid. Atractylodes samples were provided from three kinds of Origin: Baoding of Hebei province, Haozhou of Anhui province, Shaoxing of Zhejiang province.

2.3. Experiment procedure

The experiments were carried out in an air-conditioned laboratory where the temperature was kept at 25±1□ and the humidity at 54±2%. Static headspace sampling method was used because of its accessibility and stability [10].

The samples with different origin were put into four beakers (500ml) labeled Hebei, Anhui and Zhejiang, respectively. The amount of each sample in the beaker was 100g. Then three beakers were hermetically capped with plastic wrap for 70 minutes in order to generate a steady headspace respectively. The sampling time for each sample is 60 seconds, which is enough for each sensor to reach a stable value. The rinsing time is set as 110 seconds, during which the sensors are rinsed with charcoal filtered to force the signals of sensors to baseline. The interval for data collection was one second. One measurement cycle would last for about three minutes. When the measurement was completed, the obtained data was stored in a computer for later analysis. The headspace gas of each beaker of Atractylodes sample was measured 30 times respectively. Thus 90 data sets were collected for all three groups of Atractylodes samples. The 90 samples were divided into two groups: 45 samples (15 samples of each group) for the training set and the rest 45 samples (15 samples of each group) for the testing set.

2.4. Pattern recognition

LDA is one of the widely used classification techniques. However, when the total number of samples is small or the number of selected features is large, SSS problem would occur while using LDA algorithm with traditional Fisher criterion if the within-class scatter matrix is singular. Therefore, an optimized discriminant criterion called maximum scatter difference (MSD) criterion was adopted [11].

Suppose the number of known pattern classes is N as $G_1, G_2, \cdots, G_N$, pattern $x \in R^d$ is d-dimensional real vector, N_i is the number of training samples in i th class, m_i is the mean feature vector of training samples in i th class ,between-class scatter matrix is S_b , within-class scatter matrix is S_W ,and they defined as following respectively:

Mean of samples m_i :

$$m_i = \frac{1}{N_i} \sum_{x \in G_i} x \quad , \; i = 1, 2, \cdots, N \tag{1}$$

within-class scatter matrix S_ω :

$$S_\omega = \sum_{i=1}^{N} \sum_{x \in G_i} \left(x - \mu_i\right)\left(x - \mu_i\right)^T \tag{2}$$

$$i = 1, 2, \cdots, N$$

between-class scatter matrix S_b :

$$S_b = \sum_{i=1}^{N} (m_i - m)(m_i - m)^T \tag{3}$$

among, $m = \frac{1}{N} \sum_{i=1}^{N} m_i$ (4)

Fisher criterion is that the choice makes the maximum of the generalized Rayleigh quotient as the projection direction vector

$$J_F\left(\omega\right) = \frac{\omega^T S_b \omega}{\omega^T S_w \omega} \tag{5}$$

The basic idea of MSD criterion is try to find an optimal projection vectors ω .It is different from Fisher criterion because in MSD, the difference of between-class scatter and within-class scatter is employed as discriminant criterion rather than their ratio. Thus we can define maximum scatter difference criterion function as below:

$$J_M(\omega) = \frac{\omega^T \left(S_b - C * S_\omega\right) \omega}{\omega^T \omega} \tag{6}$$

Where, C is a constant, for convenience, this article is set to 1, to balance maximizing the between-class scatter and minimize the divergence between classes.

$S_b - C \times S_\omega$ is called matrix of generalized divergence difference as parameters for the *C*.

It can proved that the optimal projection direction ω is to make the maximum scatter difference criterion function $J_M(\omega)$ to take the maximum value of the solution, which the following generalized eigenvalue problem is solved:

$$(S_b - C \times S_\omega)\omega = \lambda \times \omega \quad (7)$$

So, maximum scatter difference criterion can be attributed to the sake of eigenvector problem of the generalized divergence difference matrix $S_b - C \times S_\omega$.

3. Results and Analysis

3.1. Sensors response

Figure 2 shows the typical response curves of 10 sensors to the three selected sample groups. The horizontal axis is the sampling time, and the vertical axis is the sensor response value.

It shows rapid change at the beginning of the sampling time while the response values reach to the steady state soon. After approximately 60 seconds almost all the sensors reached to stable response values. This Figure clearly shows different response signals of sensors array to Atractylodes samples with different growing areas. Each sensor has response to different varieties of Chinese herbal medicines.

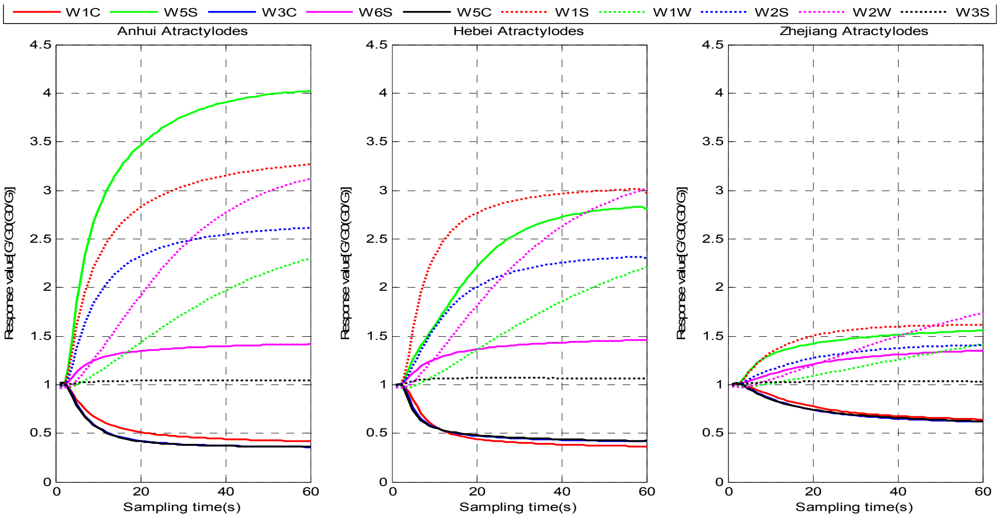


Figure 2. The response curves of Atractylodes samples

3.2. Feature selection

Feature selection is of great importance, which requires the conversion of sample features to patterns that have condense representations, ideally containing only main information.

In this study, initially eight different sub-features were selected as the original feature vector from the sensor response signals:

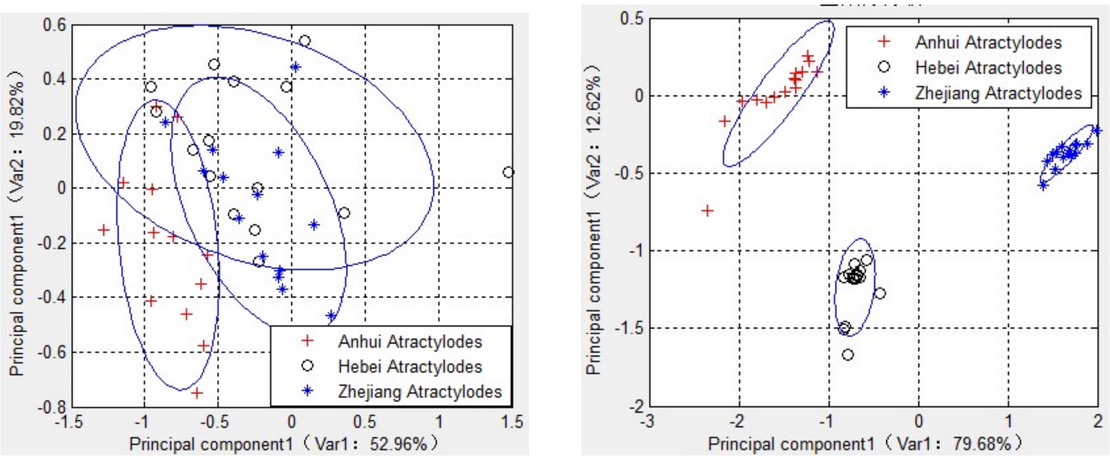
$$T = [\, f_{10}, f_{40}, f_{60}, avg, \max, \mathrm{var}, std, diff \,] \quad (8)$$

where f_i represents the response value at i second of sensor array (i=10, 40, 60); avg represents the average values of each response curve for the duration of 60 seconds; $\max$

represents the maximum values of each response curve for the duration of 60 seconds; var represents the variance of response data for the duration of 60 seconds; std represents the standard deviations of the response signals, presenting the fluctuation around the average values of each response curve; $diff$ represents the differentiation of the response signals.

3.3. Discriminant classification

There is 15 sampling times for Atractylodes training samples of each growing area, so the total of Atractylodes training samples with three different growing areas is 45, and PEN3 has 10 sensors, each sensor measurements are extracted eight characteristic parameters, thus the total characteristics vector dimension are 80-dimensional, then, clearly the total number of training samples are less than the feature vector dimension, a "small sample" problem that arise, at this time LDA algorithm can not proceed at this time. Figure 3 is the PCA and PCA + LDA analysis chart of three Atractylodes training samples.



(a) Analytic result of three groups by PCA

(b) Analytic result of three groups by PCA+LDA

Figure 3. Analytic result of three groups

It can be seen from Figure 3 (a) that the classification results of three sets of training samples with a separate PCA algorithm are not satisfactory, the batch of sample points intertwined, and indistinguishable. The reason is when the difference of sample quality grade is small, there is a big overlap of information or relevance in the differences in the sample that reflect by electronic nose sensor , PCA algorithm to find only the data distribution of spindle orientation [12], retained after dimensionality reduction by the information is not necessarily the most effective for classification . Figure 3 (b) can be seen that the distinguish result of PCA + LDA method is better than using PCA algorithm alone between three groups of training samples, and interspersed with the original training sample points have all been clearly separated. This is because the main idea of LDA algorithm is to minimize the within-class distribution and maximize the spread between classes.

To avoid the small sample problem, we use LDA algorithm based on maximum scatter difference criterion. The results shown in Figure 4:

As can be seen from Figure 4 that the distinguish result of LDA algorithm is better than PCA and PCA + LDA algorithm. While Hebei Atractylodes sample points are mainly concentrated in the lower half of the feature space, Zhejiang Atractylodes sample points mainly in the upper left part and Anhui Atractylodes sample points are concentrated in the upper right part. Various training sample points can be clearly distinguished, and compared to PCA + LDA method, the distribution of sample points within a class is even more concentrated, more obvious the interface between the classes.

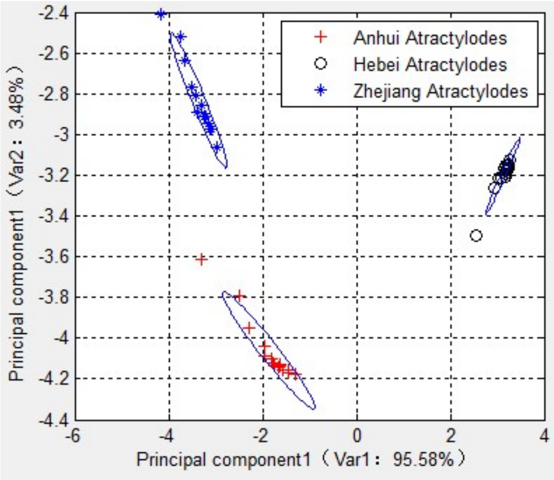


Figure 4. Analytic result of three groups by LDA based on MSD criterion

3.4. Discriminant classification

Table 1 shows the classification result of each test sample in the two-dimensional feature space, and the accuracy rate was calculated by the ratio of the number of correctly predicted samples and the number of total testing samples.

Table2. Predicted results of three testing sets

	Testing sets	Correctly predicted	Wrongly predicted	Accuracy
Hebei Atractylodes	15	15	0	100%
Anhui Atractylodes	15	14	1	93.3%
Zhejiang Atractylodes	15	15	0	100%

The results shows that, for the 45 samples tested, there is only an error to be carried out to determine, the recognition rate of Anhui Atractylodes was 93.3%,while Hebei and Zhejiang Atractylodes recognition rate was 100%. The discriminant results reached 97.8% of correct classification rate for all test samples.

4. Conclusion

More and more studies have shown that the use of electronic nose technology for odor analysis is not only objective and accurate, but also reproducible and convenient. In this paper, PEN3 electronic nose used to test Atractylodes samples of three growing areas, data analysis method using LDA algorithm based on MSD criterion to solve the problem of small samples, also distinguish with three Atractylodes from three different growing areas correctly, and the correct recognition rate of all testing samples reaches 97.8%, furthermore, the classification results clearly superior to the use of PCA or PCA + LDA algorithm. This provides the assurance for the quality of Chinese herbal medicine an effective way.

References

[1] Gardner JW, Bartlett PN. A Brief History of Electronic Nose. *Sensors and Actuators B*. 1994; 15(18): 211-220.

[2] Ghasemi-Varnamkhasti M, Mohtasebi SS. Meat Quality Assessment by Electronic Nose. *Sensors.* 2009; 9(8): 6058-6083.

[3] Brezmes J, Fructuoso MLL. Evaluation of An Electronic Nose to Assess Fruit Ripeness. *IEEE Sensors Journal.* 2005; *12*(8): 97–108.

[4] Wang YW, Wang J. Monitoring Storage Time and Quality Attribute of Egg Based on Electronic Nose. *Analytica Chimica Acta*. 2009; 16(3): 183-188.

[5] Thaler ER, Hanson CW. Medical Applications of Electronic Nose Technology. *Expert Review of Medical Devices*. 2005; 2(5): 559-566.

[6] Mazzone, PJ. Analysis of Volatile Organic Compounds in the Exhaled Breath for the Diagnosis of Lung Cancer. *Journal of Thoracic Oncology*. 2008; 3(7): 774-780

[7] Kateb B, Ryan MA. Sniffing Out Cancer Using the JPL Electronic Nose: A pilot study of a novel approach to detection and differentiation of brain cancer. *Neuroimage*. 2009; 7(4): 5-9.

[8] Szczurek A, Szecowka PM. Application of Sensor Array and Neural Networks for Quantification of Organic Solvent Vapours in Air. *Sensors and Actuators B*. 1999; 8(5): 427-432.

[9] Martin MA, Santos JP, Vasquez H, et al. Study of the Interferences of NO and CO in Solid State Commercial Sensors. *Sensors and Actuators B*. 1999; 8(5): 469-473.

[10] Martinelli E, Zampetti E. Design and Test of An Electronic Nose for Monitoring the Air Quality in the International Space Station. *Microgravity Science and Technol*ogy. 2007; 19(8):60-64.

[11] AZ Berna, J Lammertyn S. Electronic Nose Systems to Study Shelf Life and Cultivar Effect on Tomato Aroma Profile. *Sensors and Actuators B: Chemical*. 2004; 9(7): 324-333.

[12] A Martinez, A Kak. PCA Versus LDA . *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2001; 23(2): 228-233.

[13] MP Marti, R Boque O. Electronic Noses in the Quality Control of Alcoholic Beverages. *Trends in Analytical Chemistry*. 2005; 24(7): 57-66.

[14] FX Song, D Zhang JY. Adaptive Classification Algorithm Based on Maximum Scatter Difference Discriminant Criterion. *Acta Automatica Sinica*. 2006; 32(4) : 541-549.

[15] Scott SM, James D, Ali Z. Data analysis for electronic nose systems. *Microchimica Acta*. 2006; 15(6):183-207.