A New Method for Classification of Chinese Herbal Medicines Based on Local Tangent Space Alignment and LDA

Huiqin Chen*, Dehan Luo
School of Information Engineering
Guangdong University of Technology
Guangzhou 510006, China
Email: 1033566530 {at} qq.com

Hamid GholamHosseini
School of Engineering
Auckland University of Technology
Auckland 1142, New Zealand

Abstract—**Controlling the quality of Chinese herbal medicines (CHMs) is a challenging issue due to the complex and diverge specification of components in herbs. The main purpose of this study is to develop an algorithm for species identification of CHMs. An electronic nose (E-nose) was employed to collect the smell print of different groups of CHMs with different kinds and production batches. A combination of local tangent space alignment (LTSA) and linear discriminant analysis (LDA) methods was adopted for the classification of CHMs. First, the nonlinear manifold learning algorithm LTSA was employed to reduce the dimension of the feature data. The goal of this dimensionality reduction is to discover the hidden structure from the raw data automatically. Then in the reduced space, the LDA algorithm based on Fisher criterion was employed to implement a linear classifier. The results show that, the combination of LTSA+LDA algorithm can well distinguish six different kinds of CHMs and three different production batches of the same kind with 100% recognition rate of all tested samples.**

Keywords-Electronic nose (E-nose); Chinese herbal medicines; Manifold learning; LTSA+LDA; Classification and identification

I. INTRODUCTION

Chinese herbal medicines (CHMs) with a profound cultural background have been used for the prevention and treatment of disease for thousands of years in the traditional Chinese medicine. However, due to wide varieties and complex sources, some poor quality of precious and rare medicines are often appear in adulterants of CHMs. This leads to a decline in the quality of CHMs with serious impact on the reputation of CHMs in the growing market. Therefore, the type identification of CHMs is an important issue for the quality control and boosting the treatment feature of CHMs.

However, the identification methods based on human senses will be inevitably influenced by factors such as physiology, experience, emotion and environment. These factors are subject to poor reproducibility, low accuracy and strong subjectivity. Therefore, it is difficult to form a standard identification procedure.

While physicochemical methods such as gas chromatography (GC), mass spectrometry (MS) and flame ionization detection (FID) take much longer pretreatment time, it is difficult to directly connect the obtained data with the odor of samples. Therefore, modern analysis techniques which can fully characterize the color, gas and flavor of CHMs have been considered as preferred methods for the quality control of CHMs.

In recent years, with the rapid development of sensors, computers and signal processing technologies, the machine olfaction system - the electronic nose (E-nose), has been developed inspired by the human sensory conduction mechanism. The E-nose consists of gas sensors, signal processing and pattern recognition components, which try to simulate the human/animal's olfactory organs of perception, analysis and judgment of odor. Firstly, the gas sensitive sensors respond to the chemical reaction of different odorant molecules and convert it into an electrical signal that can be measured. Then the signal processing component processes the generated odor signal. Finally by multivariate statistical classification or neural network methods it identifies the measured odor and changes to sensory evaluation index [1]. Compared with the traditional odor analysis technology, E-nose has the advantage of being able to reflect the "odor characteristics" of CHMs with the advantage of determining fast, sensitive, accurate and nondestructive of the processed odor signals [2]. E-nose has been also widely used in food [3], medicine [4], agriculture [5], environment [6] and public safety monitoring [7].

In this paper, CHMs with different odor characteristics have been selected and PEN3 E-nose has been employed to collect odor information of different kinds of CHMs. Moreover, common linear analysis methods such as principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA) have been studied for pattern recognition of the smell prints. It was found that although these methods with dimensionality reduction are easy to implement but they fail to discover the underlying nonlinear structure of the high-dimensional E-nose data which is often nonlinear.

Recently, manifold learning algorithm, which can learn the low-dimensional manifold in high-dimensional data, has been

used as an effective method for nonlinear dimensionality reduction [8]. Local tangent space alignment (LTSA) algorithm is a good nonlinear manifold learning algorithm that can effectively learn the global embedding coordinates reflecting the low-dimensional manifold structure of a data set [9]. The LDA also ensures that after the projection, the pattern sample has the minimum within-class distance and maximum between-class distance in the new space. This property can be considered as the best separability feature of the algorithm for the patterns in the new space.

In this paper, we employ the advantages of both LTSA and LDA, as the LTSA+LDA method, to analyze and process nonlinear high-dimensional odor data collected by E-nose, in order to achieve good classification and identification of CHMs with different kinds and different production batches.

II. MATERIALS AND METHODS

A. Materials and instruments

First, The CHMs samples used in our experiments were provided by Guangzhou University of Chinese Medicine with similar morphology. These samples are difficult to be distinguished without a prior knowledge. The selected CHMs types are: Amomum Cardamomum, Atractylodes macrocephala, Atractylodes lancea, Heracleum kansuense, Alpinia oxyphylla and Curcuma aromatica. We chose three kinds of Alpinia oxyphylla of different production batches: Anhui Alpinia oxyphylla, Guangdong Alpinia oxyphylla and Hainan Alpinia oxyphylla.

The instrument used in this experiment is PEN3, a portable E-nose made by German AIRSENSE Company. The PEN3 E-nose is an analytical instrument that consists of a set of complex chemical sensors and recognition software. It consists of 10 metal oxide semiconductor (MOS) sensors with 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 a target gas and G0 represents the resistance while each sensor is exposed to the zero gas filtered by the standard activated carbon.

B. Experimental conditions and methods

The laboratory temperature maintains at 25 ~ 27 ^{o}C and the relative humidity maintains at 50 ~ 60% during the experiments. The static headspace sampling method is used to collect the samples of odor information by PEN3 E-nose. The weight of each sample is 10g, the headspace generation time is 60 min and the size of static headspace space is 250 ml. The sampling time is set to 120 s, the cleaning time of the sensor array is set to 200 s and the sampling interval is set to 1s.

For six different kinds of CHMs, each kind of CHMs was continuously sampled 16 times, a total of 96 sample sets. We chose 60 sets (6x10) for training and 36 sets (6x6) for testing. For three kinds of Alpinia oxyphylla of different production batches, each kind was continuously sampled 12 times, a total of 36 sample sets. We chose 30 sets (3x10) for training and 6 sets (3x2) for testing.

C. The principle of LTSA+LDA algorithm

1) Local Tangent Space Alignment

Manifold learning algorithm as an important method in data mining can discover the low-dimensional hidden structure from the high-dimensional raw data and achieve dimensionality reduction and pattern classification of high-dimensional data. It has the advantages of less algorithm parameters, fast calculation speed, good dimensionality reduction effect and keeping the topology of the original data space [10-12]. The LTSA algorithm [9] is a kind of manifold learning algorithm based on local tangent space, using approximation of the tangent space of each sample point to construct the local geometry of the low-dimensional manifold. It then uses the local tangent space alignment to find out the global low-dimensional embedding coordinates. For a given set of sample points $\{x_1, x_2, \cdots, x_N\}, x_i \in R^m$, the local tangent space alignment algorithm can be described as follows:

- Selection of neighborhood by calculating the neighborhood of each sample point x_i and record $X_i = [x_{i_1}, \cdots, x_{i_k}]$ as the k nearest neighborhood points including the sample point x_i .
- The local linear projection for the neighborhood of each sample point by calculating the right singular vectors corresponding to the d largest singular values of center matrix $X_i - \overline{x_i} 1_k^T$ and let the d right singular vectors form the matrix V_i .
- Alignment of local coordinate system by constructing the permutation matrix $\Phi = \sum_{i=1}^{N} S_i W_i W_i^T S_i^T$, among which, $W_i = I - [1_k / \sqrt{k}, V_i][1_k / \sqrt{k}, V_i]^T$. Calculate the eigenvectors $u_1, \cdots, u_d$ corresponding to the d smallest nonzero eigenvalues of matrix Φ , where $T = [u_1, \cdots, u_d]^T$ is the embedding result of the calculation.

LTSA can well recover the subset of the equidistant low-dimensional space of manifold while there is no claim for the subset being convex. LTSA can also well recover the low-dimensional structure of the "empty" manifold.

2) Linear discriminant analysis

Although, manifold learning can well recover the intrinsic low-dimensional space of the original data, but the space may not be the best recognition space. Therefore, it is necessary to carry out the analysis in the low-dimensional space and map the data to the best discriminant space for classification and recognition. LDA is a common pattern classification algorithm, which constructs the discriminant function by the linear combination of the original data, divides the multidimensional space into some subspaces and distinguishes different sample

sets to the maximum extent. It is easy to realize the performance of classification as effective. Suppose the number of known pattern classes is N as $G_1, G_2, \cdots, G_N$, pattern $x \in R^n$ is n-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, S_w is the total within-class scatter matrix, S_b is the between-class scatter matrix. This specific algorithm can be described as follows:

- Calculate the sample mean vector m_i

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

- Calculate the total within-class scatter matrix S_w

$$S_w = \sum_{i=1}^{N} \sum_{x \in G_i} (x - m_i)(x - m_i)^T, i = 1, 2, \cdots, N \quad (2)$$

- Calculate the between-class scatter matrix S_b

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

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

- Find the optimal projection direction using Fisher criterion function defined as:

$$J_F(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_w \omega} \quad (5)$$

The Fisher discriminant criterion regards the ratio of between-class scatter and within-class scatter as a comprehensive measure of the data separability after projection, the Fisher optimal projection direction is the vector which makes the comprehensive separability measure and achieves the e maximum, i.e.

$$w^* = \max imize \frac{\det(S_b)}{\det(S_w)} \quad (6)$$

3) The LTSA+LDA algorithm

The LTSA is a new manifold learning algorithm, which can effectively learn the global embedding coordinates that reflect the low-dimensional manifold structure of the dataset. LTSA algorithm can be performed on the training sample set, but because there is no explicit mapping relationship, it is difficult to carry out training on test samples. Also, the low-dimensional space obtained by LTSA algorithm may not be the best recognition space. Therefore, it is not perfect for the identification of test samples. When using LDA algorithm directly to deal with the high-dimensional data, there may exist small sample size problem. While the dimension of the sample feature data collected from the E-nose is greater than the total number of samples, thus leads the within-class scatter matrix S_w to be singular, LDA algorithm will not proceed. We can make S_w nonsingular by dimensionality reduction method to solve the small sample size problem.

This paper employs a combined LTSA and LDA algorithm, as the LTSA+LDA to recognition the low-dimensional space of manifold. It can effectively deal with new samples and also avoid the small sample size problem when LDA was used directly. We propose the use of nonlinear manifold LTSA as the first step to reduce the dimension of the feature data and simplify the data and optimize the feature vectors. Then we employ the feature matrix obtained by LTSA algorithm as the input matrix of LDA algorithm and design a linear classifier based on Fisher criterion to complete the classification and identification of CHMs.

III. RESULTS AND DISCUSSIONS

A. Sensors response

The response of PEN3 E-nose to six different kinds of CHMs is shown in Fig.1. The horizontal axis represents the sampling time of 0 to 120 s, and the vertical axis is the sensor response value (G/G0). The PEN3 sensor array shows different response curves to different kinds of CHMs. Each sensor has the characteristic of cross sensitivity to the same odor and the response characteristic of each sensor to the same odor is different (Fig.1). Moreover, the sensitivity of the senor array to Atractylodes macrocephala and Heracleum kansuense is higher than other four kinds of CHMs. The sensitivity of the senor array to Amomum Cardamomum, Alpinia oxyphylla and Curcuma aromatica is low as they belong to the Compositae family of CHMs.

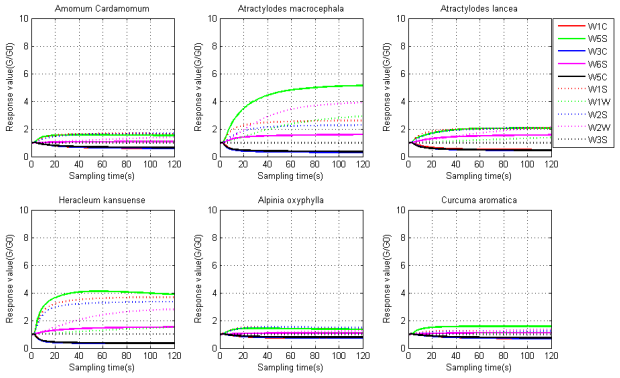

Figure 1. The response curves of six different kinds of CHMs.

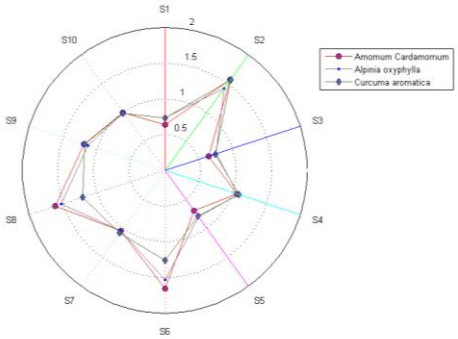

Figure 2. The radar plot of Amomum Cardamomum, Alpinia oxyphylla and Curcuma aromatic.

The radar plot of Fig.2 shows different response characteristics of the sensor array to these three kinds of CHMs. Some sensors (such as S2, S6 and S8) show a better sensitivity than others.

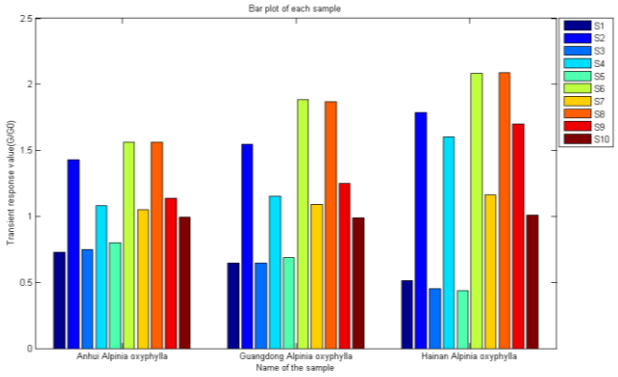

Figure 3. The bar plot of three kinds of Alpinia oxyphylla with different production batches.

Fig.3 shows the bar plot of the response characteristics of E-nose to three kinds of Alpinia oxyphylla with different production batches. In Fig.3, we can clearly see similar pattern in the response characteristic of the sensor array for these three kinds as they belong to the same kind of CHMs. However, due to the difference in the place of collection and harvesting time their response characteristic is different.

B. Classification results of six different kinds of CHMs

The classification results of using LTSA and LTSA+LDA algorithms as applied to the odor samples of six different kinds of CHMs collected by E-nose are shown in Fig.4 and Fig.5.

In Fig.4, the horizontal axis represents the eigenvector corresponding to the minimum nonzero eigenvalue obtained by LTSA algorithm and the vertical axis represents the eigenvector corresponding to the second nonzero eigenvalue. As shown in Fig.4, when LTSA algorithm was used directly to reduce the high-dimensional odor data to the two-dimensional space, the within-class distance is large and the between-class distance is smaller. Atractylodes macrocephala and Heracleum kansuense CHMs can be distinguished from other four kinds of CHMs. There are a few sample points overlap between Atractylodes lancea and Curcuma aromatic. Also, Amomum Cardamomum and Alpinia oxyphylla have some overlapping sample points. Two test samples of Atractylodes macrocephala have been correctly identified as Atractylodes macrocephala category through the LTSA algorithm. To a certain degree, the LTSA algorithm can be used for classification and identification of six different kinds of CHMs by analyzing and processing the high-dimensional of their odor data.

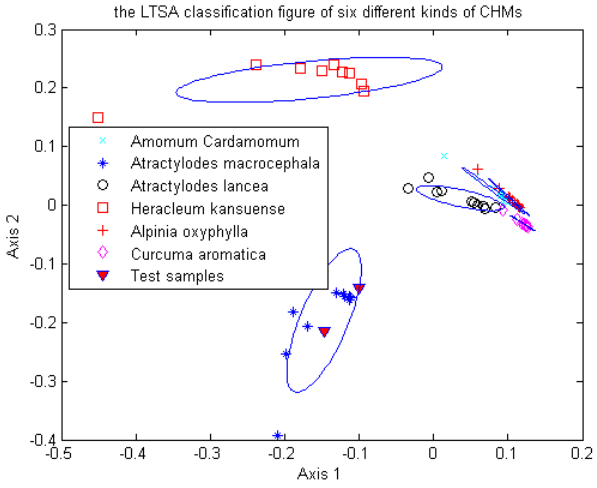

Figure 4. The LTSA classification results of six different kinds of CHMs.

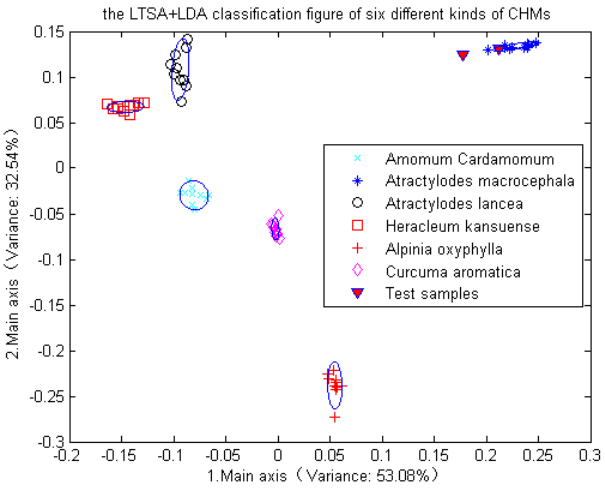

Figure 5. The LTSA+LDA classification results of six different kinds of CHMs.

In Fig.5, the LTSA+LDA algorithm was used for classification and identification of six different kinds of CHMs by analyzing and processing the high-dimensional of their odor data in the two-dimensional space. The horizontal axis represents the first main axis with the variance contribution rate of 53.08%, the vertical axis represents the second main axis with the variance contribution rate of 32.54%. We can see that, for each kind of sample, the within-class distance becomes smaller and the between-class distance becomes larger. Two test samples of Atractylodes macrocephala have also been correctly identified as Atractylodes macrocephala category

through the LTSA+LDA algorithm. Therefore, the LTSA+LDA algorithm can better distinguish these six different kinds of CHMs in compared with the direct use of LTSA algorithm.

For classification of six different kinds of CHMs, each kind of medicine has 10 training samples, with a total of 60 training samples. The E-nose with 10 sensors and the sampling interval 120 s generated a 120 by 10 matrix for each odor sample which can be regarded as a sample point with the dimension of 1200. For processing the high-dimensional odor data of samples, we need to set two important parameters, the number of neighborhood points, k and the intrinsic dimension, d (namely the dimension of the embedding space). The selection of k and d is the key factor in the algorithm as it has a significant effect on the embedding results. If the value of k is too large, LTSA can not reflect the local characteristics, on the other hand if it is too small, LTSA will not keep the topological structure of sample points in the low-dimensional space. If the value of d is too large, the mapping results will contain too much noise; and if too small, the sample points may overlap with each other in the low-dimensional space. In this paper, we used MATLAB V.7.10 (R2010a) to analyze and process the high-dimensional odor data of CHMs. The optimal parameters of k and d were set experimentally to k=15, d=2 for the LTSA algorithm, and to k=15, d=10 for the LTSA+LDA algorithm.

C. Identification results of six different kinds of CHMs

The identification results of the test samples of six different kinds of CHMs based on LTSA algorithm and LTSA+LDA algorithm are shown in Table Ⅰ and Table Ⅱ. The overall recognition rate is defined as the ratio of the number of test samples correctly identified and the number of total test samples.

TABLE I. THE IDENTIFICATION RESULTS OF THE TEST SAMPLES BASED ON LTSA ALGORITHM

Samples	Test samples	Correctly identified	Wrongly identified	recognition rate
Amomum Cardamomum	6	3	3	50%
Atractylodes macrocephala	6	6	0	100%
Atractylodes lancea	6	5	1	83.3%
Heracleum kansuense	6	6	0	100%
Alpinia oxyphylla	6	4	2	66.7%
Curcuma aromatica	6	5	1	83.3%
Total	36	29	7	80.6%

From Table Ⅰ, we can see that, for the 36 test samples, 7 were wrongly identified when directly using the LTSA algorithm. The recognition rate of Amomum Cardamomum and Alpinia oxyphylla were relatively low (50% and 66.7% respectively) compared with other CHMs. The recognition rate of Atractylodes lancea and Curcuma aromatica were both 83.3%. The recognition rate of Atractylodes macrocephala and Heracleum kansuense were the highest with 100%. The overall recognition rate for all test samples is 80.6%.

TABLE II. THE IDENTIFICATION RESULTS OF THE TEST SAMPLES BASED ON LTSA+LDA ALGORITHM

Samples	Test samples	Correctly identified	Wrongly identified	recognition rate
Amomum Cardamomum	6	6	0	100%
Atractylodes macrocephala	6	6	0	100%
Atractylodes lancea	6	6	0	100%
Heracleum kansuense	6	6	0	100%
Alpinia oxyphylla	6	6	0	100%
Curcuma aromatica	6	6	0	100%
Total	36	6	0	100%

As shown in Table Ⅱ, for the 36 test samples, all test samples were correctly identified when using the LTSA+LDA algorithm. The recognition rate of the test samples of six different kinds of CHMs were 100%, with the overall recognition rate of 100%.

Comparing the results of Table Ⅰ and Table Ⅱ we can find that, the recognition rate of the LTSA+LDA algorithm was 19.4% higher than that of the LTSA algorithm in identification of six different kinds of CHMs. Therefore, the LTSA+LDA algorithm made significant improvement in recognition performance.

D. Classification results of three kinds of Alpinia oxyphylla with different production batches

The classification results of using LTSA and LTSA+LDA algorithms as applied to the odor samples of three different kinds of Alpinia oxyphylla with different production batches are shown in Fig.6 and Fig.7.

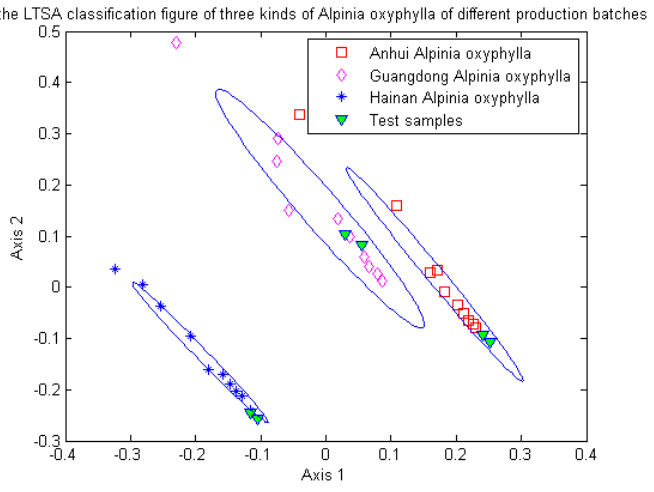

Figure 6. The LTSA classification results of three kinds of Alpinia oxyphylla with different production batches.

As shown in Fig.6, when LTSA algorithm was used directly to reduce the high-dimensional odor data to the two-dimensional space, the within-class distance is large and the between-class distance is small with unsatisfactory classification result. In Fig.6, two test samples of each kind of CHMs were all correctly identified as the corresponding category through the LTSA algorithm.

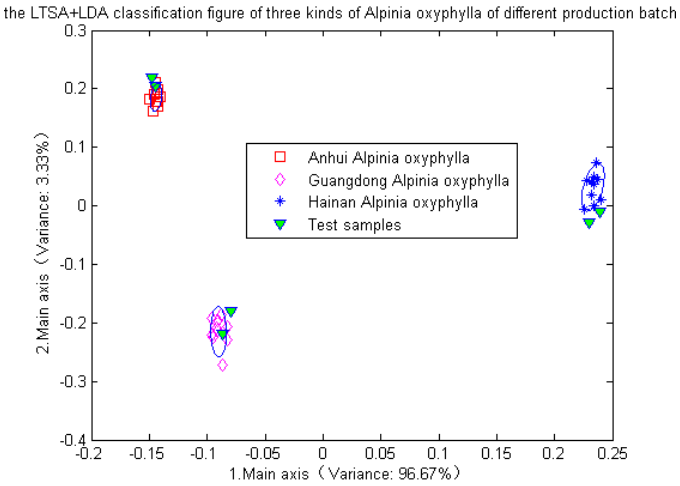

Figure 7. The LTSA +LDA classification results of three kinds of Alpinia oxyphylla with different production batches.

In Fig.7, the LTSA+LDA algorithm was used for classification and identification of three kinds of Alpinia oxyphylla with different production batches by analyzing and processing the high-dimensional of their odor data in the two-dimensional space. The variance contribution rate of the first main axis is 96.67%, the variance contribution rate of the second main axis is 3.33%. We can see that, for each kind of sample, the within-class distance is significantly reduced and the between-class distance is increased. Therefore, the LTSA+LDA algorithm can better distinguish these three kinds of CHMs in compared with the direct use of LTSA algorithm. Also, two test samples of each kind of CHMs were all correctly identified as the corresponding category through the LTSA+LDA algorithm.

In the classification of three kinds of Alpinia oxyphylla with different production batches, each kind of medicine has 10 training samples and 2 test samples, which have a total of 30 training samples and 6 test samples. Each sample is a 120 by 10 matrix and can be regarded as a sample point with the dimension of 1200. In this experiment, the optimal parameters of k and d were set to k=10, d=2 for the LTSA algorithm, and to k=16, d=10 for the LTSA+LDA algorithm.

IV. CONCLUSIONS

In this paper, a novel algorithm based on LTSA+LDA which integrates the characteristic of both LTSA and LDA algorithms is proposed. It employs LTSA to learn the global embedding coordinates that reflect the low-dimensional manifold structure of the dataset and uses the advantage of the LDA algorithm that maps the datasets to the best feature space for classification and recognition. Therefore, the proposed algorithm can perfectly solve the problem of dimensionality reduction and discover the hidden structure of the E-nose smell print data for classifying different varieties of CHMs.

When we analyze and process the high-dimensional odor data collected by PEN3 E-nose, the LTSA+LDA algorithm can well distinguish six different kinds of CHMs. Moreover, it can also classify three kinds of Alpinia oxyphylla with different production batches. The proposed algorithm performs correct identification of six different kinds of CHMs and three kinds of Alpinia oxyphylla with different production batches with the correct recognition rate of 100% for all test samples. The classification results and the recognition rate of the proposed algorithm are significantly better than that of applying only LTSA algorithm. The superiority of the LTSA +LDA algorithm is mainly due to the advantage of using LDA algorithm for minimizing the within-class scatter and maximizing the between-class scatter. However, identification other kinds of CHMs with herbs collected at different places or different harvesting time has yet to be further studied.

ACKNOWLEDGMENT

This work was finally supported by the Key Program of Natural Science Foundation of Guangdong Province. (Grant No.2011020002906)

REFERENCES

[1] Du Ruichao, Feng Yi, Xu Desheng, Wang Youjie, Wu Fei, Electronic Nose and Its Application Prospect in Chinese Medicine Industry, Chinese Journal of Experimental Traditional Medical Formulae, 2013, 19(5): 348-351.

[2] Zou Huiqin, Han Yu, Xing Shu, et al. Electronic Nose and Its Application in Chinese Materia Medica. Mode Tradit Chin Med Mater Med, 2012, 14(6): 2120-2125.

[3] Jia Hongfeng, He Jianghong, Yuan Xinyu, Yan Hong, ZHu Limin, Jia Dongying, Quality Analysis of Sichuan Hot Bean Sauce Using Electronic Nose, Journal of Food Science, 2011, 32(12): 178-182.

[4] Zheng Zhezhou, Lin Xuejuan, Study on Application of Medical Diagnosis by Electronic Nose, Mode Tradit Chin Med Mater Med, 2012, 14(6): 2115–2119.

[5] Zhu Jianyun, Zhao De'an, Pan Tianhong, Zhang Xiaochao, Identification of moldy foodstuff based on artificial olfactory system, Transactions of the CSAE, 2005, 21(1): 106-109.

[6] Leilei Pan, SimonX.Yang, A new intelligent electronic nose system for measuring and analyzing livestock and poultry farm odours, Environ Monit Assess, 2007, 135(1-3): 399-408.

[7] Cosimo Distante, Giovanni Indiveri, Giulio Reina, An application of mobile robotics for olfactory monitoring of hazardous industrial sites, Industrial Robot: An International Journal, 2009, 34(1): 51-59.

[8] Yan Zhimin, Liu Xiyu, Manifold Learning and Research of Algorithm,Computer Technology and Development, 2011, 21(5): 99-102.

[9] Zhang ZY, Zha HY, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, SIAM Journal of Scientific Computing, 2004, 26(1): 313-338.

[10] Min Wang, Hong Qiao, Bo Zhang, A New Algorithm for Robust Pedestrian Tracking Based on Manifold Learning and Feature Selection, IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4): 1195-1208.

[11] Hong Qiao, Peng Zhang, Di Wang, Bo Zhang, An Explicit Nonlinear Mapping for Manifold Learning, IEEE Transactions on Cybernetics, 2013, 43(1): 51-63.

[12] Yi Wang, Junan Yang, Hui Liu, Acoustic targets feature extraction method based on manifold learning, Electronics Letters, 2012, 48(3): 139-140.