Determination of Luzhou-flavor liquor ages by three-dimensional fluorescence spectroscopy combined with NMF*

MA Chao-qun (马超群)**, **XU Rui-yu** (徐瑞煜), **CHEN Guo-qing** (陈国庆), **and ZHU Zhuo-wei** (朱焯炜)
Jiangsu Provincial Research Center of Light Industrial Optoelectronic Engineering and Technology, School of Science, Jiangnan University, Wuxi 214122, China

(Received 27 July 2018; Revised 26 September 2018)

The determination of Luzhou-flavor liquor ages is carried out by three-dimensional fluorescence spectroscopy combined with non-negative matrix factorization (NMF). 37 samples of aged liquors with weighted ages of 15, 20 and 25 years were prepared by blending three Luzhou-flavor original base liquors with storage ages of 10, 20 and 30 years in different proportions. The fluorescence spectra of the samples were measured, and then factorized into basis matrix and coefficients matrix by multiplicative iterative NMF. The fluorescence spectra, reconstructed from the basis matrix, are similar to the original spectra. The coefficients matrix is taken as the input of support vector machine (SVM) to establish a prediction model for the determination of liquor ages. Compared with the principal component analysis, the prediction model based on SVM has a predicted accuracy better than 91.7%. This method can provide helps for the market supervision on the aged liquors.
Document code: A **Article ID:** 1673-1905(2018)06-0452-5
DOI https://doi.org/10.1007/s11801-018-8123-9

Chinese liquors, which are made from fermented grains, have a history more than 6 000 years[1,2]. The types and quantities of chemicals in the liquors vary during the aging process, which makes the aged liquors have a mellow taste and be favored by the customers[3]. However, there is still no national uniform standard for the aged liquors[4]. Some distilleries mislead or even cheat the customers, which seriously damages the healthy development of the aged liquor market. Thus, it is of great importance to strengthen the supervision of aged liquor market and develop a rapid, accurate and reliable method to discriminate the authenticity of the aged liquors. Conventionally, the aged liquors fall into two types. One type is blended by a single original base liquor stored for certain years, and the liquor age is defined directly by the storage years. The other type is blended by several original base liquors with different ages in a certain proportion, and the liquor age is defined by calculating the weighted sum of the storage years.

As different aged liquors share a large amount of physical-chemical properties and flavor components in common[5,6], it is difficult to discern one aged liquor from the others. The discrimination of aged liquors is still in the early study stage. Scientists have made efforts to resolve this problem. Cheng et al[4] determined the storage years of the aged liquors by using bomb-pulse ^{14}C. Yang et al[7] identified the aged liquors by measuring contents of some metal ions using inductively coupled plasma spectroscopy. Xu et al[8] used gas chromatography to measure the 21 major aroma components in the aged liquors and discriminated the liquor age by principal component analysis and cluster analysis. Xu et al[9] proposed an approach based on the volatilization coefficient to determine the Jiannanchun liquor age. The techniques referred above are effective for discrimination of liquor age. However, all of them are time-consuming and require complex pretreatments before the determination process.

To achieve rapid determination, the fluorescence spectrometry which has advantages of simple operation, low cost, reliable high sensitivity and high accuracy[10,11] is used in this paper. We measure the three-dimensional (3D) fluorescence spectra of the aged liquors which are blended by three Luzhou-flavor original base liquors with storage ages of 10, 20 and 30 years in different proportions. Besides, the non-negative matrix factorization (NMF) is applied to process the spectral data. The fluorescence spectral data are factorized into two non-negative matrices of basis matrix and coefficient matrix. The coefficient matrix is taken as characteristic

* This work has been supported by the National Natural Science Foundation of China (No.61378037), the Fundamental Research Funds for the Central Universities (Nos.JUSRP51628B and 1142050205180920), and the National First-class Discipline Program of Food Science and Technology (No.JUFSTR20180302).
** E-mail: machaoqun@jiangnan.edu.cn

parameter to be put into support vector machine (SVM) to establish the determination model for Luzhou-flavor aged liquors.

NMF is an effective approach for large-scale data processing which was first proposed by Lee and Seung[12] and has been widely used in image processing, computer vision and pattern recognition[13,14]. Compared with principal component analysis (PCA), NMF is much easier to be used because it dispenses with the need for orthogonal basis matrix, and it can be directly used without meeting the principle of maximum variance and centered process. Moreover, the elements of matrix obtained by NMF are non-negative, which has a real meaning for some practical problems. In this paper, we use NMF to factorize the fluorescence spectral data which are all positive.

The basic idea of the NMF is transforming the original data matrix into a product of two non-negative matrices[15]. We can use $\boldsymbol{X}$ to denote an $n\times m$ original data matrix. All the elements in the matrix $\boldsymbol{X}$ are non-negative. There are two matrices $\boldsymbol{W}(n\times r)$ and $\boldsymbol{H}(r\times m)$ that satisfy

$$\boldsymbol{X}=\boldsymbol{W}\times\boldsymbol{H}, \tag{1}$$

where $\boldsymbol{W}$ and $\boldsymbol{H}$ are the basis matrix and coefficient matrix, respectively. r is the number of components which is an empirical value under the condition $(n+m)r<<nm$. According to Eq.(1), the column vectors of matrix $\boldsymbol{X}$ can be treated as the weighted sum of the column vectors of matrix $\boldsymbol{W}$, while the elements of the matrix $\boldsymbol{H}$ are treated as weight coefficients. In order to save storage space and decrease computation time, the original matrix can be replaced by the coefficient matrix to reduce the data dimensionality.

Solving Eq.(1) is an optimization process that $\boldsymbol{W}$ and $\boldsymbol{H}$ can be worked out by iterative method. The initial $\boldsymbol{W}$ and $\boldsymbol{H}$ are randomly assigned, thus the matrix $\boldsymbol{WH}$ is reconstructed. The performance of reconstruction is evaluated by an objective function. Commonly, the Frobenius norm is chosen to be objective function, which is expressed as

$$\min_{W,H\geq 0} f(\boldsymbol{W},\boldsymbol{H})=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\boldsymbol{X}_{ij}-(\boldsymbol{W}\times\boldsymbol{H})_{ij}\right]^2. \tag{2}$$

There are several iterative methods, including multiplicative iterative algorithm[16], alternating least squares algorithm[17], projected gradient algorithm[18], and so on. Among them, the multiplicative iterative algorithm, which has advantages of simplicity and high computing power, is the standard algorithm for NMF. Here, we use the multiplicative iterative algorithm to iteratively update $\boldsymbol{W}$ and $\boldsymbol{H}$ to minimize the objective function. The iteration of Eq.(2) should follow the functions[15] expressed as

$$H_{\mathrm{rm}}\leftarrow H_{\mathrm{rm}}\frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})_{\mathrm{rm}}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H})_{\mathrm{rm}}}, \tag{3}$$

$$W_{\mathrm{nr}}\leftarrow W_{\mathrm{nr}}\frac{(\boldsymbol{X}\boldsymbol{H}^{\mathrm{T}})_{\mathrm{nr}}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})_{\mathrm{nr}}}. \tag{4}$$

SVM is a kind of pattern recognition method which was first reported by Cortes and Vapnik[19]. This method is capable of solving both linear and non-linear problems. The main idea of SVM for non-linear problems is introducing a non-linear function to map the low-dimensional input space to the high-dimensional feature space, so that the low-dimensional non-linear problem is transformed into high-dimensional linear problem that can be easily solved.

In the sample preparation, three Luzhou-flavor original liquors with storage ages of 10, 20 and 30 years were taken as base liquors to be blended in different proportions to obtain the aged liquors. The weighted age of the aged liquors can be calculated according to

$$Y=10a+20b+30c, \tag{5}$$

where Y stands for the weighted age, while a, b and c are the percentages of base liquors with ages of 10, 20 and 30 years in the aged liquors. By adjusting the values of a, b and c, we obtained 37 samples, including 10 samples of 15 years aged liquors, 16 samples of 20 years aged liquors and 11 samples of 30 years aged liquors.

The experiment was carried out in a multifunctional fluorescence spectrometer FLS920P (Edinburgh Instrument, Britain). The range of excitation wavelength λ_{ex} is 215—500 nm with step of 5 nm, while the range of emission wavelength λ_{em} is 210—650 nm with step of 2 nm. The slit width in front of the monochromator is 5 nm. All the programs of PCA, NMF, and SVM were written by Matlab 7.0.

The 3D fluorescence spectra of the original base liquors and 37 samples of blended aged liquors were measured. Fig.1 shows their fluorescence contour maps. Fig.1(a)—(c) are the contour maps of the original base liquors with ages of 10, 20 and 30 years, and Fig.1(d)—(f) are the averaged contour maps of the blended aged liquors with weighted ages of 15, 20 and 25 years. According to the contour maps, the fluorescence spectral characteristic parameters are extracted, as listed in Tab.1.

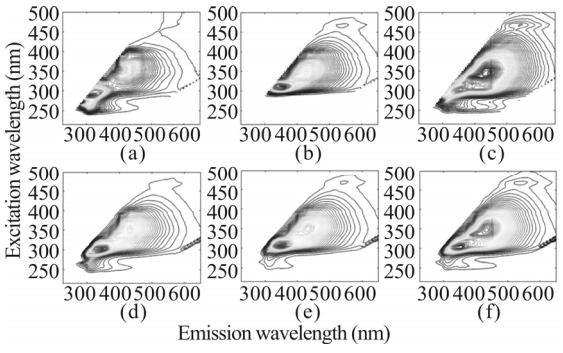

Fig.1 The 3D fluorescence spectra of the original base liquors with ages of (a) 10 years, (b) 20 years and (c) 30 years, as well as the blended aged liquors with weighted ages of (d) 15 years, (e) 20 years and (f) 25 years

Tab.1 Fluorescence spectral characteristics of three base liquors and three blended aged liquors

Liquor ages	Original ages			Weighted ages		
	10	20	30	15	20	25
Peak numbers	2	3	2	3	3	3
Optimal excitation wavelength (nm)	355	350	345	355	350	345
Peak wavelength (nm)	310 432	344 430 536	434 536	336 434 544	340 434 538	340 434 538
Peak intensity ($\times 10^5$)	5.58	6.05	7.64	3.75	5.29	6.60

The results show that the fluorescence spectra of the three original base liquors are similar in shape but have some difference in detail. Each of the base liquors has a fluorescence peak located at 432 nm because these three base liquors belong to the same flavor type of the same brand. It indicates that the chemical substances in the original base liquors are approximately the same. Nevertheless, some trance substances, such as esters and acids, vary with the increase of storage years thanks to the esterification reaction, redox reaction and decomposition reaction, which brings in the differences in detail of the fluorescence spectra. Consequently, the main fluorescence peak of the 10 years base liquor locates at 310 nm while that of the 20 years base liquor locates at 344 nm. Moreover, the fluorescence intensity of the three liquors increases with the storage years.

As the aged liquors were blended by these three base liquors, their fluorescence spectra should inherit the characteristics from those of the base liquors. Their fluorescence peaks are all around 340 nm, 434 nm and 538 nm. It is necessary to introduce metrology method for discrimination.

For the application of PCA, the fluorescence spectral data were pretreated to remove the effects of Rayleigh scattering and Raman scattering. And then, the three-dimensional fluorescence spectral matrix (58×213×37) was transformed into two-dimensional matrix ((58×213)×37). Considering the large volume of spectral data, the wavelet transformation was used for data compression. In this paper, we chose the db7 wavelet function and decomposition level of 4 to decompose the two-dimensional data. The compressed wavelet coefficients of the fourth level were extracted to replace the spectral data for PCA. The scores of the first two principal components are shown in Fig.2.

It can be seen that the scores of these three different aged liquors mingle together. According to the results of PCA, the accumulative contribution rate of the first eight principal components to the total data accounts reaches 100.0%. It indicates that the eight principal components carry the overall data information. Thus, the eight principal components are selected as characteristic parameters for determination of aged liquors.

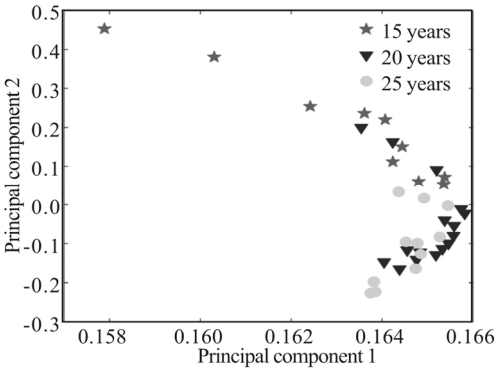

Fig.2 The scores of the first two principal components for these three blended aged liquors with different ages

In comparison, for the application of NMF, the three-dimensional fluorescence spectral matrix is transformed into two-dimensional matrix first. And then, the multiplicative iterative algorithm is used to factorize the two-dimensional matrix into basis matrix ((58×213)×4) and coefficient matrix (4×37) by setting the number of components as 4. The basis matrix indicating the fluorescence spectral characteristics of the aged liquors is reconstructed into four three-dimensional matrices. Each three-dimensional matrix is plotted as a contour map as shown in Fig.3.

It is easy to see that Fig.3(b) and (d) are similar to the fluorescence spectra of the 30 and 10 years aged liquors, respectively. In addition, both Fig.3(a) and (c) are similar to the fluorescence spectrum of the 20 years aged liquor. Based on the coefficient matrix, the relationships between different factors are demonstrated in Fig.4. It is obvious that the aged liquors with the same weighted age gathers. The values of Factor 2 of the 25-year blended aged liquors are large, which indicates that the 25-year aged liquors have high level of 30-year base liquor. As for the 15-year blended aged liquors, the situation is

quite the reverse. Therefore, we use the coefficient matrix as the characteristic parameter for the determination of aged liquors.

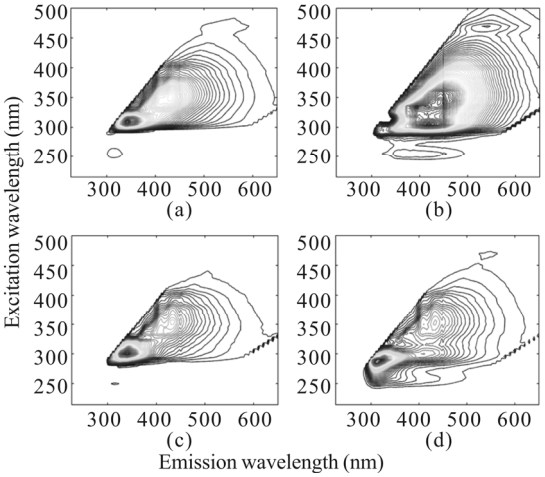

Fig.3 The contour maps of the four matrices derived from the factorization of the fluorescence by NMF

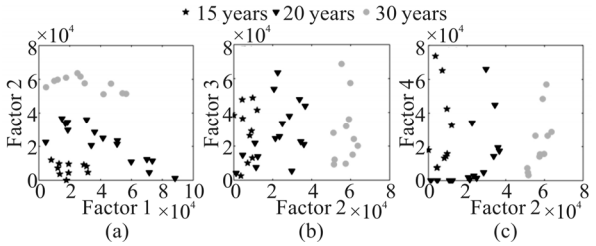

Fig.4 The relationship between the four factors of the coefficients matrix: (a) Factor 1 / Factor 2; (b) Factor 2 / Factor 3; (c) Factor 2 / Factor 4

We randomly choose 25 samples as training set. Among them, 7 samples are from the 15 years aged liquor, 11 samples are from the 20 years aged liquors, and 7 samples are from the 25 years aged liquors. The rest 12 samples are taken as prediction set. Both the data in the training set and prediction set are normalized. The radial basis function is adopted as the kernel function of the SVM. It can be expressed as

$$K\left(X_i, X_j\right) = \exp\left(-\gamma \left\| X_i - X_j \right\|\right)^2, \tag{6}$$

where $\gamma = 1/2\sigma^2$, and σ is a free parameter.

The performance of SVM is quite sensitive to the determination model parameters. The penalty parameter c and kernel function parameter g should be adjusted for better accuracy. In this paper, we use k-fold cross-validation[20] to optimize the parameters. The 25 samples in the training set are divided into 5 subsets. In turn, one of them is taken as validation subset, and the others are taken as training subsets. Consequently, 5 different determination sub-models are established, and the mean value of the determination accuracy of the 5 validation subsets is obtained as the accuracy of the main determination model. Thus, the relationship between log2c, log2g and accuracy is found, as shown in terms of contour maps in Fig.5, to determine the optimal parameters in application of NMF and PCA. As for PCA, the penalty parameter and kernel function parameter are 4 and 2 with mean accuracy of 84.0%. While for NMF, the penalty parameter and kernel function parameter are 8 and 1 with mean accuracy of 100.0%. It is obvious that the accuracy of the determination model based on NMF is more reliable than that based on PCA.

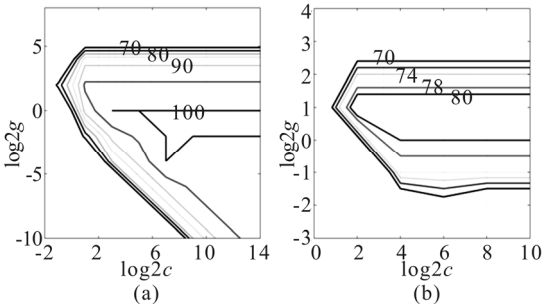

Fig.5 The contour map for pursuit of the optimal parameters of (a) NMF- SVM and (b) PCA- SVM

After the determination of the optimal parameters, the final determination model is established based on the total 25 samples in the training set. And it is used to identify the rest 12 samples in the prediction set. The results, listed in Tab.2, show that the determination model based on NMF-SVM has an accuracy of 100.0%, comparing with that of 50.0% for the method based on PCA-SVM.

Tab.2 The optimal parameters and accuracy of the models for the determination of two brands of liquors

		Model parameters		Training set	Prediction set	
Method	Brand	c	g	Accuracy	Error number	Accuracy
PCA-SVM	I	4	2	84.0%	6	50.0%
	II	64	4	72.0%	6	50.0%
NMF-SVM	I	8	1	100.0%	0	100.0%
	II	2048	0.125	96.0%	1	91.7%

To prove the validity of the model, we chose another brand of Luzhou-flavor original liquors with storage ages of 10, 20 and 30 years to prepare samples of blended aged liquors with weighted ages of 15, 20 and 25 years in the same way. Also, we use the methods based on PCA-SVM and NMF-SVM for determination of aged liquors, respectively. The determination accuracy based on NMF-SVM is 91.7%, higher than that of 50% based on PCA-SVM.

In summary, the determination model based on fluorescence spectroscopy combined with NMF and SVM is demonstrated for the prediction of the weighted ages of blended aged liquors. The fluorescence spectra of the

blended aged liquors were measured and decomposed by NMF to train the SVM. The penalty parameter and kernel function parameter are optimized by *k*-fold cross-validation to establish the determination model. The determination accuracies of two different brands of Luzhou-flavor original liquors reach 100.0% and 91.7%. This method paves a way for the determination of aged liquors and the market supervision.

References

[1] Yao F., Yi B., Shen C., Tao F., Liu Y., Lin Z. and Xu P., Scientific Reports **5**, 9553 (2015).

[2] Han S., Zhang W., Li X., Li P. and Liu J., Food Analytical Methods **9**, 2194 (2016).

[3] Zhu S. M., Xu M. L., Ramaswamy H. S., Yang M. Y. and Yu Y., Scientific Reports **6**, 30273 (2016).

[4] Cheng P., Zhou W., Burr G. S., Fu Y., Fan Y. and Wu S., Scientific Reports **6**, 38381 (2016).

[5] Ding X., Wu C., Huang J. and Zhou R., Journal of Food Science **80**, 2373 (2015).

[6] Ding X., Wu C., Huang J. and Zhou R., LWT-Food Science and Technology **66**, 124 (2016).

[7] Yang T., Li G. and Zhuang M., Liquor Making **35**, 33 (2008).

[8] Xu M. L., Yu Y., Ramaswamy H. S. and Zhu S. M., Scientific Reports **7**, 39671 (2017).

[9] Xu Z., Liquor-Making Science and Technology **2**, 90 (2008).

[10] Ma C., Chen G., Gao S., Chen C., Shi Y. and Gu L., Optoelectronics Letters **7**, 158 (2011).

[11] Zhu Z., Que L., Wu Y., Chen G., Xu R. and Zhu T., Chinese Journal of Lasers **42**, 0615002 (2015). (in Chinese)

[12] Lee D. D., and Seung H. S., Learning the parts of objects by non-negative matrix factorization, Nature **401**, 788 (1999).

[13] Liu Y., Liao Y., Tang L., Tang F. and Liu W., Neurocomputing **173**, 224 (2016).

[14] Guillamet D., Vitria J. and Schiele B., Pattern Recognition Letters **24**, 2447 (2003).

[15] Hoyer P. O., Journal of Machine Learning Research **5**, 1457 (2004).

[16] Eggermont P. P. B., Linear Algebra and its Applications **130**, 25 (1990).

[17] Kroonenberg P. M. and de Leeuw J., Psychometrika **45**, 69 (1980).

[18] Zhang L. and Zhou W., Journal of Computational and Applied Mathematics **196**, 478 (2006).

[19] Cortes C., and Vapnik V., Support-vector networks, Machine Learning **20**, 273 (1995).

[20] Grimm K. J., Mazza G. L. and Davoudzadeh P., Structural Equation Modeling A Multidisciplinary Journal **24**, 246 (2017).