NIR Spectroscopy and NMF Algorithm for Identification of Oil Pollutants in Water

Ailing Tan
College of Information Science and Engineering
Yanshan University
Qinhuangdao, P. R. China
The Key Laboratory for Special Fiber and Fiber Sensor of Herbei Province
Yanshan University
Qinhuangdao, P. R. China
e-mail: tanailing@ysu.edu.cn

Xuan Guo
College of Information Science and Engineering
Yanshan University
Qinhuangdao, P. R. China
The Key Laboratory for Special Fiber and Fiber Sensor of Herbei Province
Yanshan University
Qinhuangdao, P.R. China
e-mail: guoxuan@ysu.edu.cn

Yong Zhao
College of Electrical Engineering
Yanshan University
Qinhuangdao, P. R. China
e-mail: zhaoyong@ysu.edu.cn

Abstract—Oil pollutants is one of the major pollution sources in water. Accurate, rapid, and convenient detection method of oil pollutants in water has very important theoretical value and practical significance. The combination of near-infrared spectroscopy (NIR) and chemometrics is ideal for such a situation. NIR spectroscopy is a powerful and effective technique. traditional NIR methods do not take full account of the absorbance data non-negative characteristics, resulting in the analysis lack of reasonable explanation. In this paper, the qualitative discriminate method of single species oil contaminants based on non-negative matrix factorization feature extraction combined with support vector machine classification algorithm is studied. Non-negative matrix factorization algorithm and support vector machine classifier parameters on classification accuracy are discussed in depth to optimize NIR qualitative classification model. The present method has a good identification effect and strong generalization ability，and can work as a new method for rapid identification of oil pollutants in water.

Keywords-NIR Spectroscopy; Non-negative matrix factorization; oil pollutans;; Support Vector Machine; Genetic algorithm

I. INTRODUCTION

With the rapid development of economy, various causes of oil pollutants caused severe damage to the natural environment and ecological resources, which also caused serious damage to human body health. the oil pollution in the water becomes an urgent topic in the environmental protection Different categories of the oil spills have dissimilar contamination degree, therefore it is important to discriminate and predominate the categories , source and pollution extent of the spilled oil fleetly so that the environment departments can take effectual clean-up and protection measures. Owing to the versatility and complexity of the oil pollution, the detection of oil spills is a challenging task for the past years[1-2]. So it is essential to investigate and establish an accuracy and rapid oil spill discrimination method.

.Near infrared spectral analysis technology has extensively been applied for analyses in diverse fields including petroleum chemical industry,agriculture, food, medicine and environmental monitoring and other fields in recent years, with its high measuring efficiency, without damaging, as well as good stability and repeatability[3-6]. However, NIR spectra often contain serious systematic variation that is unrelated to the response data set, and the analysis of interest absorbs only in small parts of the spectral region, Among these applications, we must admit that the use of chemometric methods for the qualitative or the quantitative analysis of the unknown samples is essential in establishing effective models.

There are many multivariate statistical analysis in chemometrics methods, such as the classical Principal Component Analysis(PCA), Independent Component Analysis(ICA), Kernel Principal Component Analysis(KPCA), Singular Value Decomposition(SVD), etc. The input matrix is decomposed into a multiplied two low-rank matrix form, the original input matrix is represented by the encoding decomposed matrix, thus achieve the purpose of dimensionality reduction and feature extraction.The conventional chemometrics method does not guarantee a non-negative element is obtained after the decomposition, although a negative value in the mathematical results is correct, but in many practical problems, a negative value element does not correspond to real physical meaning. Therefore, for specific data types and applications, a dimensionality reduction method which can decompose non-negative data, while the decomposed data is still keeping non-negative is needed. The non-negative matrix factorization is a decompositon method to meet this requirement..

NMF(Non-negative Matrix Factorization) is an algorithm for data dimensionality reduction and feature extraction by looking for a low dimensional feature space of non-negative factors, which has the advantage of clear principle, simple structure and good interpretive results[7-8] Now NMF has been widely used in various fields of pattern recognition, image procssing, text retrieval, etc., it is also very suitable for multivariate analysis in chemometrics.

II. PRINCIPLE OF NMF ALGORITHM

Non-negative matrix factorization algorithm was originally developed by D.D.Lee and H.S.Seung whose paper name was "Leaming the Parts of Objects by Nonnegative Matrix Factorization" in "Nature", it is based on the "local constitute whole" theory, for a given non-negative matrix, looking for two non-negative matrix, their matrices is approximately equal to the given matrix[9]. NMF algorithm is formally described as follows:

Given a non-negative matrix $V \in R_+^{M\times N}$ and a positive integer r, finding two non-negative matrix $W \in R_+^{M\times r}$ and $H \in R_+^{r\times N}$, which meeting the following conditions:

$$V \approx WH \tag{1}$$

in which each column of W is called basis image, while H is the coefficient matrix, the basis number r is usually chosen to satify $(m+n)r < mn$.

Formula (1) is equivalent to the following optimization problem:

$$\min F(W,H) = \frac{1}{2}\|V-WH\|_F^2 = \frac{1}{2}trace(V-WH)^T(V-WH)$$
$$st. \quad W \geq 0, H \geq 0 \tag{2}$$

From the point of view that linear sparse coding, a better sparse representation will acquired by adding sparse constraint in the matrix decomposition process, therefore NMF algorithm with sparse constraints was proposed by combing sparse constraint coding with NMF algorithm[10]. The objective function is expressed as:

$$\min_{W,H}(V \mid WH) = \sum_{ij}(V_{ij} - (WH)_{ij})^2 \tag{3}$$

for any i, exisiting the following constraints:

$$sparseness(w_i) = sW, \forall i \tag{4}$$

$$sparseness(h_i) = sH, \forall i \tag{5}$$

where sW and sH is sparse constraints factor of marix W and H respectively. The sparse constraint factor is continuously changing values in the range of 0~1, in order to represent different degress sparse constraint.

III. MATERIALS AND METHOD

A. Apparatus and software

For NIR measurement, we used a Brucker MPA spectrometer controlled by OPUS (version6.5) for windows software from Brucker Optics (Bremen, Germany). The times of scanning spectrum was 64, and the resolution was 8cm^{-1}, The spectrum covered the range from 12,000 to 4000 cm^{-1}.Room temperature was monitored using a mercury thermometer with a precision of ±0.5℃ and it did not vary significantly during acquisition of the spectrum of all the samples.

For the spectral acquisition, OPUS 6.5, which is attached with the NIR instrument, was used in this experiment.

B. Preparation of samples and Spectrum collection of sample

All oil samples of gasoline, diesel fuel and kerosene came from two different gas station every week during two months, and the sea water was collected from Qinhuangdao sea area for three times. The following procedure was carried out: 100ul pure oil and 100ml seawater were shaken together respectively. After 15minutes, the oil and the seawater were completely immiscible. About 95ml seawater were discarded, then we used CCl4 to extract the surplus emulsions and the near infrared spectrum were recorded.

Each sample was collected three times. The mean of the three spectrum that were collected from the same oil sample was used in the following analysis step. The NIR spectrum of 35 gasoline samples , 35 diesel fuel samples and 35 kerosene samples are shown in Fig .1. The entire sample set is divided into calibration set and validation set using the 5-fold cross validation method.

As can be seen from the Fig .1, each oil's spectrum in the near infrared spectral has four peaks, due to the oil complex components, each cluster peaks are likely to be doubling frequency and combined frequency absorption of several different base frequency, so the bands attributable are more difficult.

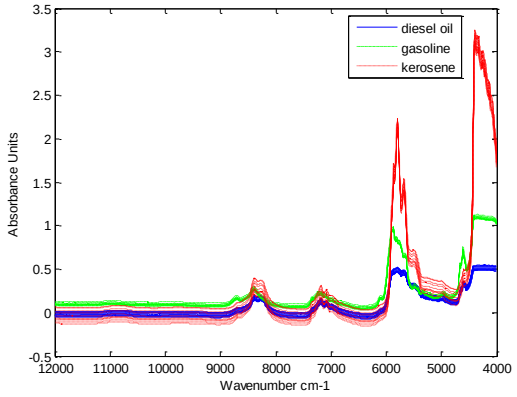


Figure 1. Near infrared spectra of three kinds of simulated spilled oil samples

IV. RESULTS AND DISCUSSION

A. NMF of near infrared spectroscopy

Suppose there are N samples of analog oils pollutants, each sample has M wavelength points, thus we can obtain $M \times N$ matrix of near-infrared spectrum, which is indicated by $V_{M\times N}$, The individual element of the matrix is the absorbance value of a sample in a particular wavelength, the absorbance value is non-negative, i.e. $V_{ij} \geq 0$, we can get the non-negative matrix factorization results of the absorbance matirx for all the samples matrix:

$$V_{M\times N} = W_{M\times r}H_{r\times N} \tag{6}$$

in which, $W_{M\times r}$ represents a non-negative base spectrum matrix, namely $W_{M\times r} \geq 0$; $H_{r\times N}$ represents the non-negative

feature matrix, $H_{r\times N} \geq 0$; r is the number of base spectrum of feature space, which is the size of the feature subspace after dimensionality reduction, the v_j and h_j is a j-th column vector of the original spectral data matirx V and the non-negative feature matrix H respectively, then the formula (6) can be expressede as:

$$v_j = W_{M\times r} h_j \qquad (7)$$

Formula(7) can be interpreted as : the column vector of each sample' s near infrared spectra is the non-negative linear combination of the r column of the non-negative base spectrum matrix, and the combination coefficients are elements of eigenvectors, each column of the base spectrum matrix can be regarded as a local feature of the original spectrum to some extent, so each sample spectrum can be regarded as linear combination of base spectrum, and this if consistent with the basic idea of the NMF "the local constitute whole". The NMF representation model with NMF method for NIR spectra data is shown in Fig .2.

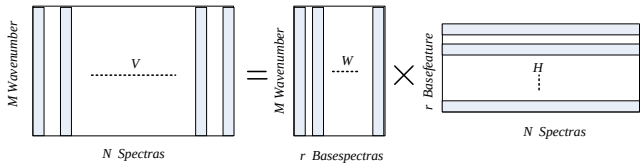
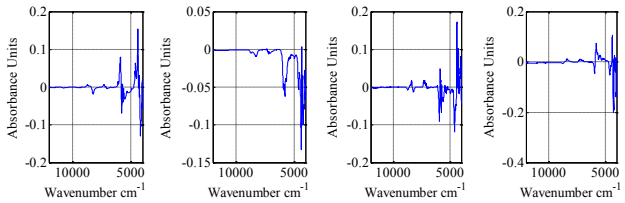


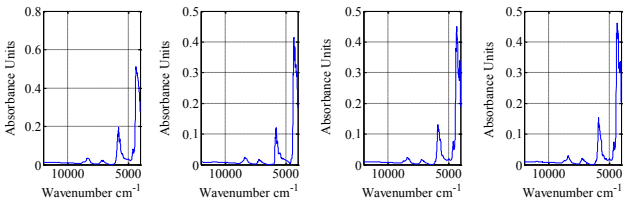
Figure 2. The representation model with NMF method for NIR spectra data

B. *Base spectrum of PCA and NMF feature extraction methods*

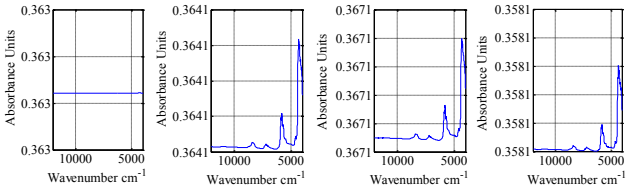
According to the scanning range and resolution of the spectrometer, the oil sample's near infrared spectrum has 2074 absorbance value. The original spectrum data matrix of calibration set were decompose with PCA, NMF and NMFSC algorithm respectively, when the base spectrum number of the feature space r is 4, the base spectrum acquired by three algorithm were shown in Fig .3(a), (b) and (c) respectively.



(a)Base spectrum in feature space with PCA method



(b)Base spectrum in feature space with NMF method



(c)Base spectrum in feature space with NMFSC method

Figure 3. Base spectrum in feature space with three feature extraction methods(r=4)

As can be seen from Fig .3(a) (b) and (c), in the same subspace dimension conditions, the base spectrum data acquired by NMF and NMFSC extraction method are all non-negative values, which can be interpreted as part of the absorbance values at a particular wave point, but the base spectrum data obtained by PCA method contain negative value, which have not a clear physical meaning. The results illustrate that the NMF algorithm is more suitable for non-negative data.

C. *.Determination of base spectrum number r*

The base spectrum number of feature space is a very improtant parameter in NMF algorithm, for a particular data, the feature space nubmer which is hidden in the inside of the data set is often fixed. For the near infrared spectral matirx, when the selected r is consistent with the essential feature space, the non-negative subspace can best reflect the original features of the near infrared spectral data, the data redundancy is minimized.

There is no no rigorous mathematical theory about the determination of the issue of r so far, generally it is acquired by choosing the best through a large number of experiments. For near infrared spectroscopy to identify different type of oil pollutants ,we also need to determine the base spectrum number of feature space experimentally, in order to obtain the base spectrum which can best indicated the characteristics of the spectrum internal structure, to ensure the best possible classification results. Figure 4. shows the relationship between the base spectrum number and classification accuracy rate when the sparse constraint factor was 0.4, 0.2 and 0.8 respectively..

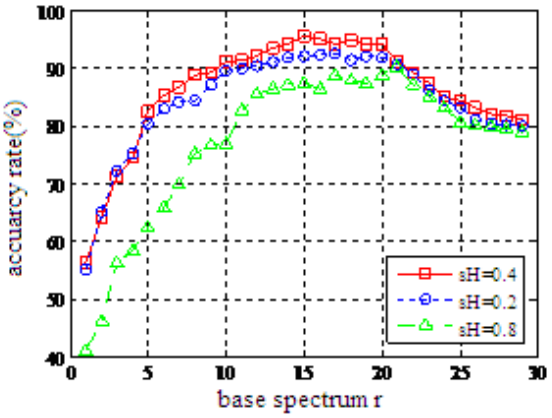


Figure 4. Relationship between number of base spectrum and classification accuracy rate

As can be seen from Fig.4, classification accuracy rate is not high when the base spectrum number r is small, this is due to the small amount of information characteristics,

which can not adequately reflect the rich information of different types of oil pollutats contained in the near infrared spectroscopy; with the increase of r, more feature data can express a sample spectrum, so the correct classification rate gradually increased. However, the classification accuracy rate declines when the base spectrum number is bigger than 15, for the reason of noise will be introduced with excessive base spectrum, resulting in the classification accuracy rate decreases, and the relationship between r and accuracy rate were basically similar for different sparse constraint factor. Therefore, the base spectrum number is determined from 12~18 in this paper.

D. Identification of oils based on SVM and GA algorithm

Among many possible techniques for data classification, SVM are a kind of learning machine based on statistical learning theory. The basic idea of applying SVM to pattern classication can be stated as follows: first, map the input vectors into one features space, possible in higher space, either linearly or nonlinearly, which is relevant with the kernel function. Then, within the feature space from the first step, seek an optimized linear division, that is, construct a hyperplane which separates two classes. It can be extended to multi-class. SVM training always seek a global optimized solution and avoid overfitting. For SVM, the optimal solution is based on the limited sample information. This prominent characteristic of the SVM could perfectly solve the critical problem that the actual analysis results were not as accurate as expected using the traditional approaches when the number of samples for establishing model was smaller. A more detailed description of SVM can be found in [11-12].

In general, SVM classification model inculde the regularization parameter C, the kernel parameter σ and the kernel function these three parameters. Regarding the SVM model, the first step is choosing the kernel function, which determines the sample distribution in the mapping space. The radial basis function (RBF) kernel commonly used in massive studies because of its good general performance and few parameters to be adjusted and non-linearly maps the samples into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear. Thus, the RBF function was select as kernel function of SVM in this study. Next, we research the other two parameters optimization based on genetic algorithm. Classification results evaluation use cross validation method.

Genetic algorithm is a stochastic search and optimization algorithm which developed by Darwin's evolution theory and Mendel genetics. This method provides a common framework for solving optimization problems of complex systems, thus it has a wide range of applications in industrial control, data mining, economic management, robotics, image processing [13-14].

The SVM model parameter optimization steps of near infrared spectrum is also including coding, selection, crossover and mutation. Taking into account both speed and efficiency, the parameter optimization interval were set as follows: $2^{-10} \le C \le 2^{10}$, $2^{-10} \le \sigma \le 2^{10}$; First randomly generating initial population, the population size was 20, using binary encoding method to encode the regularization parameter C and the kernel parameter σ, constructing a single chromosome in the order of first C then σ; determing the Fitness function as the classification accuarcy rate, selecting operator was proportional selection method; formating new individual by single point crossover method, the crossover ratio was 0.80, the variation rate was 0.05, the stopping iterations was automatically terminated after a given number of iterations, the evolution iterations number was 50; Finally decoding each chromosome in the population to obtain the best values of C and σ.

According to the above settings, for 5-fold cross validation, each classification results were shown in Table 1, and the optimization process of the optimal parameters was shown in Fig .5.

Table4-1 Parameters optimization and recogniton result with GA method

	first time	second time	third time	fourth time	fifth time
Accuarcy rate of calibration set(%)	97.62	98.81	100	96.43	94.05
optimal value of C	80.058	248.126	66.713	379.113	612.369
optimal value of σ	0.322	0.254	0.160	0.633	0.288
Support vector number	48	47	49	50	47
accuarcy rate of validation set （%）	96.24	95.24	100	90.47	90.47
time(s)	30.37	31.62	31.07	30.99	31.07

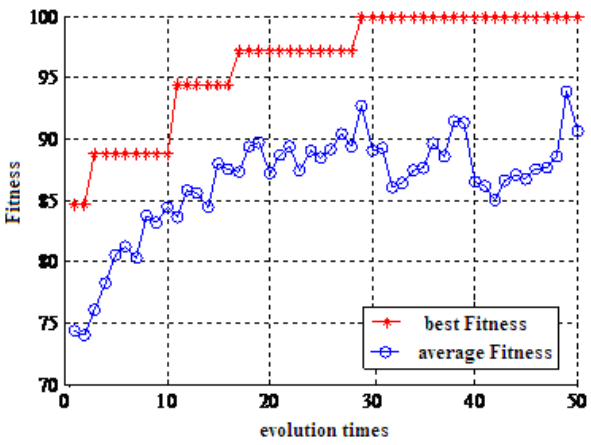


Figure 5. Parameters optimization fitness curve with GA method

As can be seen from Table1 and Fig .5, after the 29 times of the evolutionary process, the target value of population

approached the optima solution, the algorithm tends to converge. After decoding operation, the optimal parameters of the SVM classification model were obtained with C=66.713 andσ=0.160, and the classification accuarcy rate of calibration set and validation set were both achieved 100% with the optimal parameters.

V. CONCLUSION

In this study, we adapt Fourier transform NIR spectrophotometer to collect the spectral data of simulation gasoline， diesel fuel and kerosene oil pollutants. The Sparse Nonnegative Matrix Factorization algorithm was used to extract features. The experimental results illustrate that the NMF algorithm is more suitable for near infrared spectroscopy which belongs t o non-negative datas. the relationship between the base spectrum number and classification accuracy rate was also reasearched, and the results shows that classification accuracy rate will be decrease when r is too small or too big, the base spectrum number is determined from 12~18 in this paper. After feature extraction, we use SVM classification and Genetic algorithm to classify the three categories of oil pollutants. The experimental results show that NMF feature extraction combined with GA-SVM classification for qulaitative identification of oil pollutants in water is feasible.

ACKNOWLEDGMENT

The authors are grateful to the support of Hebei Provincial Science & Technology Research and development project(No.12273302, No.13273305) and the support of Hebei Educational Committee Nature Science Youth Fund (No.QN2014034, No.Q2012129), The authors also thank the support of Hebei Provincial Natural Science Foundation(No.F2014203245, No.F2013203252)

REFERENCES

[1] Marta G C, Jose E, “An assessment of oil pollution in the coastal zone of Patagonia,” Environment Management, vol.40, May. 2007, pp:814-821

[2] Kavanagh R J, Burnison K B, Frank R A. “Detecing oil sands process affected waters in the Alberta oil sands region using synchronous fluorescence spectroscopy,” Chemosphere, vol.76, Jan, 2009, pp:120-126.

[3] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et.al, Modern Near Inf rared Spectroscopy Analytical Technology (Second Edition), Beijing: Chinese Pet rochemical Indust ry Press, 2007.

[4] M. Khanmohammadi, F. Karami, A. Mir-Marqués,et al., “Classification of persimmon fruit origin by near infrared spectrometry and least squares-support vector machines,” Journal of Food Engineering, Vol. 142, Dec, 2014, pp: 17-22

[5] Barry K. Lavine, Nikhil Mirjankar, Stephen Delwiche, “Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm,” Microchemical Journal, Vol. 117, Nov, 2014, pp:178-182

[6] Omar Marín-González, Boyan Kuang, Mohammed Z. Quraishi, et al., “On-line measurement of soil properties without direct spectral response in near infrared spectral range:,” Soil and Tillage Research, Vol. 132, Aug, 2013, pp: 21-29

[7] Inkyung Jung, Dongsup Kim, “LinkNMF: Identification of histone modification modules in the human genome using nonnegative matrix factorization,” Gene, Vol. 518, Apr, 2013,pp: 215-221

[8] Andri Mirzal, “A convergent algorithm for orthogonal nonnegative matrix factorization,” Journal of Computational and Applied Mathematics, Vol. 260, Apr, 2014, pp: 149-166

[9] Jim Jing-Yan Wang, Jianhua Huang, et al., “Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization,” In Press, Uncorrected Proof, Available online 20 Sep, 2014

[10] Patrik O.H. “Non-negative matrix factorization with sparseness constraints,” Journal of Machine Learning Research, vol. 5, 2004, pp: 1457-1469.

[11] Benlan He, Yong Shi, Qian Wan, “Prediction of Customer Attrition of Commercial Banks based on SVM Model,” Procedia Computer Science, Vol. 31, 2014, PP: 423-430

[12] Jan Chorowski, Jian Wang, Jacek M. Zurada, “Review and performance comparison of SVM- and ELM-based classifiers,” Neurocomputing, Vol. 128, Mar, 2014, PP: 507-516

[13] Xin-She Yang, “Chapter 5 - Genetic Algorithms,” Nature-Inspired Optimization Algorithms, 2014, pp: 77-87

[14] D. D'Ambrosio, W. Spataro, R. Rongo, G.G.R. Iovine, “Genetic Algorithms, Optimization, and Evolutionary Modeling,” Reference Module in Earth Systems and Environmental Sciences, from Treatise on Geomorphology, Vol. 2, 2013, pp: 74-97