CHEMICAL ENGINEERING TRANSACTIONS
VOL. 55, 2016
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian

ISBN 978-88-95608-46-4; **ISSN** 2283-9216
DOI: 10.3303/CET1655059

Classification and Detection of Adulteration in Olive Oil Using Improved Gaussian Mixture Model and Regression by Artificial Bee Colony Algorithm

Xin Xie[a], Ying Gao[b], Weimin Shi[c], Qi Shen[c]*

[a]Henan Polytechnic College, Zhengzhou, Henan, China
[b]Henan Vocational College of Applied Technology, Kaifeng, Henan, China
[c]The College of Chemistry and Molecular Engineering, Zhengzhou University, Zhengzhou, Henan, China
656271158@qq.com

Gaussian mixture model (GMM) and Gaussian mixture regression (GMR) can be used to detect adulteration in extra virgin olive oil. The estimate of the GMM parameters is commonly obtained from the expectation-maximization (EM) algorithm. EM algorithm has some limitations such as local optimum problems and sensitivity to the initial values. In this paper, artificial bee colony (ABC) algorithm is used to determine the optimal parameters in GMM and GMR. To improve the optimized performance and reduce computational effort of ABC algorithm, the information sharing mechanism among the global best food sources is introduced in ABC. The improved GMM and GMR by artificial bee colony algorithm (GMMRABC) were used to discriminate and quantify the adulteration of extra virgin olive oil with rapeseed oil using FT-IR spectroscopy. It has been demonstrated that the proposed method is an accurate, rapid, stable strategy for identifying and quantifying the extra virgin olive oil.

1. Introduction

Extra virgin olive oil is far more valuable and expensive than most other vegetable oils, so the adulterated olive oil has become the biggest source of agricultural fraud problems in the European Union (Rohman et al., 2010). Edible oil adulteration may cause threat to food security. To ensure food safety for consumers, a set of effective identification and prosecution methods are needed.

The classical methods for detection of extra virgin olive oil adulteration are often complicated with no single test that can accomplish the task. A battery of tests are employed to identify of the adulterant, such as the determination of free acidity, peroxide value, UV extinction, fatty acid composition, sterol composition, triglyceride composition, wax content and steroidal hydrocarbons. Commonly used methods for adulterated oil examination also include different chromatographic techniques (Miloudi et al., 2007), nuclear magnetic resonance method (Alonso-Salces et al., 2010), mid-and near infrared spectroscopic techniques (Wang et al., 2006) and differential scanning calorimetry (Chiavaro et al., 2008). However, these methods are often time-consuming and need high cost and cumbersome operations. Fourier transform infrared spectroscopy (FT-IR) (Rohman and Man, 2011) analysis has been widely used in oil adulteration detecting for its simple sample processing, fast analysis speed and allows for direct and fast determination of several compounds without sample pretreatment. As the spectral of different chemical components are overlaps severely, the FT-IR technique is often coupled with chemometrics methods such as principal component analysis (PCA) (Kamruzzaman et al., 2013), partial least squares (PLS) (Oussama et al., 2012), back-propagation artificial neural network (BP-ANN) (Ni et al., 2012), linear discriminant analysis (LDA) (Sinelli et al., 2007), and support vector machine (SVM) (Caetano et al., 2007).

Gaussian mixture model (GMM) (Jacques et al., 2010) and Gaussian mixture regression (GMR) (Yuan et al., 2014) can be used to model any data distribution using an adequate number of Gaussian distributions. GMM and GMR are widely used statistical tools in pattern classification and nonlinear regression. Determination of the optimal Gaussian distributions parameters is the most important consideration when training GMM and

Please cite this article as: Xie X., Gao Y., Shi W.M., Shen Q., 2016, Classification and detection of adulteration in olive oil using improved gaussian mixture model and regression by artificial bee colony algorithm, Chemical Engineering Transactions, 55, 349-354
DOI:10.3303/CET1655059

GMR. The estimate of these GMM parameters is commonly obtained from the expectation-maximization (EM) algorithm. As EM algorithm is essentially similar to hill-climbing in multiple parameter space, it has some limitations such as local optimum problem and sensitivity to the initial values. One method of determining the optimal parameters is to run the EM algorithm many times with different initial values and to select the best result which maximizes the expected log-likelihood of the data. Even with many times operations, EM algorithm may also lead to overfitting for high-dimensional data. In order to overcome the above defects, some intelligent optimization algorithms (Lu et al., 2014) are introduced and used for the parameters estimation in GMM.

Artificial bee colony (ABC) algorithm (Li et al., 2015) is one of the relatively new optimization algorithms which motivated by the intelligent behavior of honey bees. ABC can also be applied in parameters estimation in GMM. With few parameters, ABC is simple to implement and robust compared to other intelligent algorithms (Devos and Duponchel, 2013). It does not need to know about special information for solution of the problem. ABC highlights the global optimum in the whole group via each bee's behavior of local optimization, so it has faster convergence speed. In this paper, ABC was used to determine the optimal parameters in GMM and GMR. To improve the optimized performance and reduce computational effort of ABC algorithm, the information sharing mechanism among the global best food sources was introduced in ABC. The improved GMM and GMR by artificial bee colony algorithm (GMMRABC) were used to discriminate the adulteration of extra virgin olive oil with rapeseed oil using FT-IR spectroscopy. The GMMRABC was also applied to quantify the level of rapeseed oil adulterant present. As comparison, the classic GMM and GMR with EM for parameter estimation, LDA, back-propagation artificial neural network (BP-ANN) and PLS analysis were also used to classify and quantify these oil samples. It has been demonstrated that the proposed method is an accurate, rapid, stable strategy for identifying and quantifying the extra virgin olive oil.

2. Methods

2.1 Gaussian mixture model and Gaussian mixture regression

For the adulterated oil examination problems, GMM assumes that pure oil and adulterated oil can be represented by different Gaussian density and the dataset consists of two underlying Gaussian probability distributions. GMM expresses probability density function of the data (FT-IR spectra) by a weighted sum of K unimodal Gaussian component densities (K=2 in this paper). Each unimodal Gaussian component density is parameterized by a mean vector and covariance matrix. The parameters of GMM are usually obtained from training data set using the EM algorithm. In the prediction stage, the likelihood function of the predicted samples belonging to each model is calculated. The predicted samples are assigned to the cluster with the largest likelihood function.

To calculate the expectation of the level of adulterant, GMR builds several GMM models to obtain the joint probability density of X and Y. In this paper, X is the FT-IR data and Y is the level of adulterant. Each GMM is similarly determined by the mean and covariance matrix. The conditional expectation of the response output can be calculated by the mean and covariance matrix. The contribution of each GMM is represented using a mixing weight. The predicted level of adulterant in extra virgin olive oil is the summation of the weighted conditional expectations of all GMMs.

2.2 Artificial bee colony algorithm

ABC algorithm is a novel optimization algorithm inspired by the foraging behavior of honeybees. In ABC algorithm, the position of a food source represents a solution to the optimization problem and the nectar amount of a food source represents the quality (fitness) of the solution represented by that food source. There are three kinds of bees: employed bees, onlooker bees and scout bees. The number of employed bees or the onlooker bees is equal to the number of food sources.

At first, initial solutions as food source positions are generated randomly. After initialization, the three categories of bees repeat the seeking good food sources process. The process of bees seeking good food sources is that which is used to find the optimal solution. Each employed bee generates a new food source v_i from the old solution x_i and the neighborhood of its previously position x_k as follows:

$$v_{i,j} = x_{i,j} + \phi_{i,j}(x_{i,j} - x_{k,j}) \tag{1}$$

Where k and j are randomly chosen indexes; k has to be different from i; $\varphi_{i,j}$ is a random number in the range [-1,1]. Once the new solution v_i is obtained, a greedy selection operation is applied to compare v_i and the old solution x_i. If the fitness of v_i is equal to or better than that of x_i, the new solution v_i will replace the old one. Otherwise, the old solution x_i is retained. When all the employed bees have finished their exploitation process, they share their obtained food source information with onlooker bees.

Each onlooker bee chooses a food source according to the roulette wheel selection method. Then in a similar way to the employed bees, the onlooker bee generates a new candidate food source. The new candidate solution is evaluated and selected by the greedy selection mechanism.

If a food source is not improved by a predetermined number of iterations, it is abandoned and the employed bee associated with the food source becomes a scout. The scout randomly generates a new food source, and then becomes an employed bee again. The process of bees seeking good solutions will be repeated until a stopping criterion is satisfied.

In standard ABC algorithm, employed bees and onlooker bees update the food source using neighborhood operator. This operator may lead to the weak ability of local exploitation and the poor speed of converging. To improve the optimized performance of ABC algorithm, the information sharing mechanism among the global best food sources is introduced.

The employed bee generates a new food source v_i as:

$$v_{i,j} = x_{i,j} + (x_{i,j} - GBest_j) * (r_{ij} - 0.5) * 2 \quad (2)$$

Where GB_{est} is the global optimal solution found so far. r_{ij} is a random number in the range [0,1].

2.3 Improved Gaussian mixture model and regression by artificial bee colony

The mean vectors and covariance matrixes are the most important considerations when training GMM or GMR. The ABC optimization algorithm is an efficient scheme to obtain the parameters of GMM. To simplify the calculation processes, the covariance matrix was constrained to be diagonal matrix. In ABC, real number strings were adopted to code all the particles. Each real number coded string stands for the parameters of GMM, a set of mean and covariance. The improved Gaussian mixture model and regression by artificial bee colony (GMMRABC) is described as follows.

Step 1. Randomly initialize the initial population of solution vectors with an appropriate size.

Step 2. Calculate the fitness function of each individual of the population. If the best object function of the generation fulfills the end condition, the training is stopped with the results output, otherwise, go to the next step.

Step 3. Update the population using employed bees, onlooker bees and scout bees by applying the ABC.

Step 4. Go back to the second step to calculate the fitness of the renewed population.

In the GMMRABC algorithm, the misclassification rate and root-mean-square error were used as the fitness for classification and quantitation respectively.

2.4 Samples

Different brands of extra virgin olive oil samples and rapeseed oil samples used in the experiment were bought from the local market. The rapeseed oil was used as adulterants in this study. The volume fractions of the rapeseed oils in the extra virgin olive oils were at fifteen levels respectively, 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 35%, 45%, 50%, 55%, 70%, 80%, 90%. FT-IR spectra of extra virgin olive oil samples and adulterated samples were obtained without any chemical pretreatment. A total of 28 extra virgin olive oil samples and 60 adulteration samples were randomly split into a training set consisting of 64 samples, and a test set of 24 samples.

2.5 Instrumentation and Software

All FT-IR spectra were collected on NICOLET 6700 FT-IR spectrometer (Thermo Electron Corporation), equipped with DTGS detector and KBr beam splitter. The spectra were recorded in the range of 4000 to 650 cm^{-1} with resolution of 4 cm^{-1} using ZnSe single-bounce attenuated total reflectance (ATR) accessory. All spectra were recorded at 25℃ using an average of 32 scans. The spectra data was scaled into (0, 1) for chemometrics analysis.

The FT-IR spectra data was then processed using different algorithms written in Matlab 7.8 and run on a personal computer. The algorithms used in this paper include the proposed GMMRABC and GMM with EM for parameters estimation, BP-ANN, LDA and PLS.

3. Results and discussion

3.1 Classification

FT-IR spectra of the extra virgin olive oil samples and adulterated oil samples were shown in Figure 1. The infrared absorption spectra of these oil samples are similar and appear virtually indistinguishable. It is difficult to classify and quantify these olive oil samples, so the proposed GMMRABC was introduced into olive oil sample analysis.

For the 28 extra virgin olive oil samples and 60 adulterated oil samples, two-thirds samples were randomly selected as training set and the remaining samples as the prediction set. The training set consists of 19 extra virgin olive oil samples and 40 adulterated oil samples, the remaining 29 samples were the predicted samples.

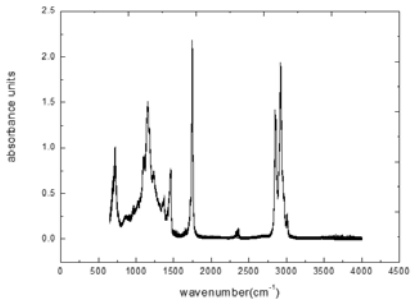

Figure 1: FT-IR spectra (4000-650 cm^{-1}) of pure olive oil samples and adulterated oil samples

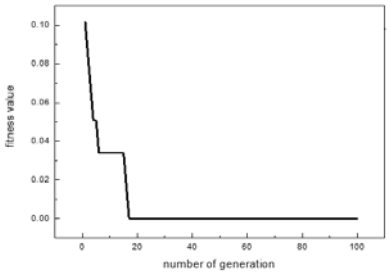

Figure 2: Convergence curves for GMMRABC

When deal with IR spectral classification, the number of samples is small compared with the number of variables (experimental points cm^{-1} per spectrum). This may deteriorate the generalization ability of the learned model and lead to possible overfitting in GMM and GMR analysis. As principal component analysis (PCA) is the most popular linear feature extraction method, we firstly conducted variable selection by PCA. In order to find the optimal number of principal components (PCs), we increased the selected PC one by one, and then the performance of the selected PCs was measured according to an averaged classification error rate over 5-fold cross-validation by the proposed GMMRABC. The misclassification rate was minimum when the number of PCs was two. The first two PCs explained for 90% of the variance. So the first two PCs with the largest eigenvalues were selected for GMMRABC to classify and quantitate the olive oil samples.

In the proposed GMMRABC, ABC was used to determine the mean vectors and covariance matrixes. The population size of ABC was selected as 50 and the ABC was stopped after 100 iterations. The means and covariance matrixes were initialized randomly. The misclassification rates for training set and test set were 0%. That's to say, the GMMRABC model can correctly classify all the training and prediction samples by optimizing mean vectors and covariance matrixes. The convergence curve was showed in Figure 2. One can see that during the updated process, the fitness value decreased until about 20 generations and fitness values dropped quickly in the GMMRABC algorithm.

To evaluate accurately the performance of GMMRABC, the total samples were randomly partitioned into training and test sets 50 times and then the misclassification rates were averaged. The average misclassification rates for training set and test set were 0% and 1.79% respectively. Among the 50 times operations, the misclassification rate for prediction set was 0% for 30 times, the largest misclassification rate was 6.90%. It means that the number of misclassified sample was two at most. The result shows that the GMMRABC model is stable and reliable, even for the adulteration content of 1%.

To compare with the proposed GMMRABC, the classic GMM with EM for parameters estimation, BP-ANN and LDA were also performed on the extra virgin olive oil spectral data. The first two PCs with the largest eigenvalues were selected for these methods. In order to keep consistent with the proposed algorithm, the total samples were also randomly partitioned into training and test sets 50 times and then the misclassification rates were averaged. Using GMM-EM algorithm, the averaged misclassification rates for training samples and prediction samples were 6.85% and 7.38% respectively. Compared with the GMM-EM, it proves that our proposed GMMRABC algorithm has a better performance and less depends on the initial value. It can be seen that the parameters of GMM optimized by the ABC algorithm is an efficient scheme and it can improve the performance of GMM. In the present work, three layer ANN was used with 10 hidden nodes for 2 variables. BP-ANN was stopped after 500 iterations. Results of BP-ANN and LDA were listed in Table 1. The performance of the GMMRABC was better than that of the other classification methods.

Table 1: Results of misclassification rates of different methods.

Method	Misclassification rate	
	Training set	Test set
GMMRABC	0%	1.79%
GMM-EM	6.85%	7.38%
BP-ANN	4.81%	5.79%
LDA	6.71%	7.03%

3.2 Quantification

The GMMRABC model was used to predict the concentration of adulteration of extra virgin olive oil with rapeseed oil. The variable selection and optimized process in the quantification were the same as in the classification. The first two PCs with the largest eigenvalues were chosen for quantification. The correlation coefficient (R) and the root mean square error (RMSE) for training set were 0.9938 and 0.0341, while the R and RMSE for testing set were 0.9936 and 0.0357 respectively. The correlation between the calculated and experimental values of the rapeseed oil contents of the adulterated oil samples was shown in Figure 3.

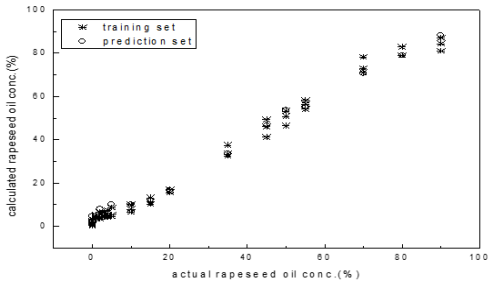

Figure 3: Calculated versus actual rapeseed oil content using GMMRABC

In comparison, GMR with EM for parameters estimation, BP-ANN and PLS were also performed on these oil samples. The optimal number of latent variables for PLS was determined by the predicting residual sum of squares and was set as 7. The first two PCs with the largest eigenvalues were selected for GMR-EM and BP-ANN. Table 2 summarized the results of these regression methods. Using EM for parameters estimation, the GMR resulted model with R=0.9912 for training set. The correlation coefficient and the root mean square error for testing set were 0.9892 and 0.0437 respectively. The BP-ANN model gave the correlation coefficients 0.9900 and 0.9853 for training and test sets respectively. The RMSE for training set was 0.0411 and for testing set was 0.0795. There is a slight symptom of overfitting. We can see that the GMMRABC perform better than the common GMR, indicating that ABC algorithm is better than EM for parameters optimization. The results obtained above suggest that the proposed GMMRABC method is a promising alternative to detect adulteration in adulteration oil samples.

Table 2: Results of quantification of different regression methods

Model	Training sets		Prediction sets	
	correlation coefficient	root mean square error	correlation coefficient	root mean square error
GMMRABC	0.9938	0.0341	0.9936	0.0357
GMR-EM	0.9912	0.0379	0.9892	0.0437
BP-ANN	0.9900	0.0411	0.9853	0.0795
PLS	0.9989	0.0139	0.9941	0.0322

4. Conclusions

The parameter estimation of GMM and GMR by ABC algorithm was proposed in this paper. 88 extra virgin olive oil and adulteration oil samples were classified and predicted by the proposed method. Moreover, several other classification and regression methods were applied for comparison. It proved that the proposed method had a better performance and it was an accurate and stable method for quality and quantification analysis of edible oil.

Acknowledgments

The work was financially supported by the National Natural Science Foundation of China (Grant No. 21575131).

Reference

Ahmed B., Fouad B., Djalil B.A., Mohamed B.B., Abdelouahed T., Bedia E.A., 2016, The Thermal Study of Wave Propagation in Functionally Graded Material Plates (FGM) Based on Neutral Surface Position, Mathematical Modelling of Engineering Problems, 3(4), 202-205, Doi: 10.18280/mmep.030410.

Alam M.S., 2016, Mathematical Modelling for the Effects of Thermophoresis and Heat Generation/Absorption on MHD Convective Flow along an Inclined Stretching Sheet in the Presence of Dufour-Soret Effects, Mathematical Modelling of Engineering Problems, 3(3), 119-128, Doi: 10.18280/mmep.030302.

Alonso-Salces R.M., Heberger K., Holland M.V., Moreno-Roias J.M., Mariani C., Bellan G., Reniero F., Guillou C., 2010, Multivariate analysis of NMR fingerprint of the unsaponifiable fraction of virgin olive oils for authentication purposes, Food Chemistry, 118(4), 956-965, Doi: 10.1016/j.foodchem.2008.09.061.

Almeida M., Vargas-Zerwes F., Ferreira-Bastos L., Costa A., Souza-Schneider R., Machado E., Kohle A., 2015, Cation and anion monitoring in a wastewater treatment pilot project, Revista de la Facultad de Ingeniería, 30(3), 82-89, Doi: 10.17533/udea.redin.n76a10.

Caetano S., Ustun B., Hennessy S., Smeyers-Verbeke J., Meissen W., Downey G., Buydens L., Heyden Y.V., 2007, Geographical classification of olive oils by the application of CART and SVM to their FT-IR, Journal of Chemometrics, 21(7-9), 324-334, Doi: 10.1002/cem.1077.

Chiavaro E., Vittadini E., Rodrigaze-Estrada M.T., Cerretani L., Bendini A., 2008, Differential scanning calorimeter application to the detection of refined hazelnut oil in extra virgin olive oil, Food Chemistry, 110(1), 248-256, Doi: 10.1016/j.foodchem.2008.01.044.

Devos O., Duponchel L., 2011, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemometrics & Intelligent Laboratory Systems, 107(1), 50-58, Doi: 10.1016/j.chemolab.2011.01.008.

Han Z.H., Zhu P.X., Guo Y., Zhou S.G., Fan N.J., 2015, Synthesis and property study of layered ti/tib2 composite electrode materials for wet electrolytic, Mathematical Modelling of Engineering Problems, 2(2), 11-14, Doi: 10.18280/mmep.020203.

Jacques J., Bouveyron C., Girard S., Devos O., Duponchel L., Ruckebusch C., 2010, Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data, Journal of Chemometrics, 24(11-12), 719-727, Doi: 10.1002/cem.1355.

Kamruzzaman M., Sun D.W., ElMasry G., Allen P., 2013, Fast detection and visualization of minced lamb meat adulteration using NIR hyperspectral imaging and multivariate image analysis, Talanta, 103(2), 130-136, Doi: 10.1016/j.talanta.2012.10.020.

Li B., Chiong R., Lin M., 2015, A balance-evolution artificial bee colony algorithm for protein structure optimization based on a three-dimensional AB off-lattice model, Computational Biology & Chemistry, 54, 1-12, Doi: 10.1016/j.compbiolchem.2014.11.004.

Lu S.J., Salleh A.H.M, Mohamad M.S., Deris S., Omatu S., Yoshioka M., 2014, Identification of gene knockout strategies using a hybrid of an ant colony optimization algorithm and flux balance analysis to optimize microbial strains, Computational Biology and Chemistry, 53, 175-183, Doi: 10.1016/j.compbiolchem.2014.09.008.

Miloudi H., Zoubida C., Abdelaziz S., Larbi H., Dominique G., 2007, Detection of argan oil adulteration using quantitative campesterol GC-Analysis, Journal of the American Oil Chemists' Society, 84(8), 761-764, Doi: 10.1007/s11746-007-1084-y.

Ni Y.N., Li B.H., Kokot S., 2012, Discrimination of Radix Paeoniae varieties on the basis of their geographical origin by a novel method combining high-performance liquid chromatography and Fourier transform infrared spectroscopy measurements, Analytical Methods, 4(12), 4326-4333, Doi: 10.1039/C2AY25950H.

Oussama A., Elabadi F., Platikanov S., Kzaiber F., Tauler R., 2012, Detection of Olive Oil Adulteration Using FT-IR Spectroscopy and PLS with Variable Importance of Projection (VIP) Scores, Journal of the American Oil Chemists' Society, 89(10), 1807-1812, Doi: 10.1007/s11746-012-2091-1.

Rohman A., Che M.Y., Ismail A., Hashim P., 2010, Application of FTIR Spectroscopy for the Determination of Virgin Coconut Oil in Binary Mixtures with Olive Oil and Palm Oil, Journal of the American Oil Chemists' Society, 87, 601-606, Doi: 10.1007/s11746-009-1536-7.

Rohman A., Man Y.B.C., 2011, The use of Fourier transform mid infrared (FT-MIR) spectroscopy for detection and quantification of adulteration in virgin coconut oil, Food Chemistry, 129(2), 583-588, Doi: 10.1016/j.foodchem.2011.04.070.

Sinelli N., Cosio M.S., Gigliotti C., Casiraghi E., 2007, Preliminary study on application of mid infrared spectroscopy for the evaluation of the virgin olive oil "freshness", Analytica Chimica Acta, 2007, 598(1), 128-134, Doi: 10.1016/j.aca.2007.07.024.

Wang L., FSC L., Wang X., He Y., 2006, Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR, Food Chemistry, 95(3), 529-536, Doi: 10.1016/j.foodchem.2005.04.015.

Yuan X.F., Ge Z.Q., Song Z.H., 2014, Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression, Chemometrics & Intelligent Laboratory Systems, 138, 97-109, Doi: 10.1016/j.chemolab.2014.07.013.