Chemometrics and Intelligent Laboratory Systems 143 (2015) 1–6

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Rapid identification between edible oil and swill-cooked dirty oil by using a semi-supervised support vector machine based on graph and near-infrared spectroscopy

Yang Zhou [a,b,*], Tiebing Liu [a], Jinrong Li [a]

[a] College of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, PR China
[b] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, PR China

ARTICLE INFO

Article history:
Received 3 September 2014
Received in revised form 21 January 2015
Accepted 3 February 2015
Available online 7 February 2015

Keywords:
Swill-cooked dirty oil identification
Near-infrared spectroscopy
Semi-supervised
Graph

ABSTRACT

It is a challenge task to identify the swill-cooked dirty oils from various kinds of edible oils by using near infrared (NIR) spectroscopy. Due to the diversity and deficiency of standard swill-cooked dirty oils samples, the classification model involves complex liner and nonlinear relationships between class label and spectral distribution. Moreover, the small sample problems in the calibration set leads to failure of traditional supervised method such as support vector machine (SVM). A powerful semi-supervised learning method, the semi-supervised support vector machine (GS3VM), is used for classification between swill-cooked dirty oil and edible oil. The GS3VM bases on manifold assumption and approximates the distribution of spectra from both labeled and unlabeled oil samples. Comparing with the PLSDA and SVM, the experimental results show that incorporating unlabeled samples in training process improves the prediction results when insufficient training information is available. Furthermore, excessive numbers of labeled or unlabeled oil samples are helpless for classification performance of GS3VM, which solves the small sample problem and saves the cost of swill-cooked dirty oil samples. Experiment results have established that it is possible to identify the swill-cooked dirty oil from various kinds of edible oils by using the proposed GS3VM approach and NIR data. We hope that the idea of semi-supervise learning obtained in this study will help further investigations in NIR spectra analysis.

1. Introduction

Swill-cooked dirty oil seriously jeopardizes people's health because of contamination by bacteria, heavy metal and harmful chemicals during the picking, decoloration, deodorization and other processing [1,2]. After refine processing, it is difficult to distinguish between edible oil and swill-cooked dirty oil by using the traditional physical–chemical indicators such as acid value, solid fat, cholesterol, heavy metals, conductivity, polyaromatic hydrocarbon, aldehyde/ketone volatile component, and etc. [3]. Therefore, developing an effective and fast authentication method between edible oil and swill-cooked dirty oil is a challenging and significant work.

Near infrared spectroscopy (NIR) has been widely used in oil identification, and it also has many merits such as non-destruction, low-costing and environment protecting. The typical applications involve the authentication of adulterated vegetable oils [4], detecting an adulterant in high quality sandalwood oil [5], discriminating soybean oil adulteration in camellia oils [6], and so on. So far, most studies on NIR have focused on identification between two kinds of oils or adulterated authentication in one specific kind of oil. But in practical application, the NIR technology needs to identify the swill-cooked dirty oils from various kinds of edible oils, which is singularly studied or reported.

The swill-cooked dirty oil is made from discarded kitchen waste, or refined from animal meat, viscera and leather. In most case, the swill-cooked dirty oil is the mixture of normal edible oil and illegal cooking oil [7]. Until now, there have been no standard swill-cooked dirty oil sample, and almost all the calibration samples are ferreted out by law enforcement. It leads to a small sample calibration problem in complex spectrum system [8]. Because of the diversity and deficiency of standard swill-cooked dirty oil samples, the high dimensional distribution of NIR spectral datasets are more complicated, and the relationships between datasets and class information were nonlinear [9]. The traditional classification method, such as principal component analysis (PCA), K nearest neighbor clustering (KNN) and partial least squares discriminate analysis (PLSDA), can't complete the task of identifying the swill-cooked dirty oils from various kinds of edible oils [10]. Our research team had combined PCA with cluster discriminate analysis (CDA) method for distinguishing between refined recycled cooking oil and edible vegetable oil by using UV spectroscopy, but the prediction results showed that the proposed approach didn't perfectly meet the real-world application requirements [11].

* Corresponding author at: College of Information and Electronic Engineering, Zhejiang University of Science and Technology, 318# LiuHe Road, Xihu District, Hangzhou City 310023, China. Tel.: +86 13868077650.
E-mail address: zybuaa@163.com (Y. Zhou).

http://dx.doi.org/10.1016/j.chemolab.2015.02.004
0169-7439/

Support vector machine (SVM) is an effective supervised method for fitting linear and nonlinear relationship between class information and NIR spectroscopy [12], but it needs sufficient labeled samples of each class for classification [13]. Due to the lack of swill-cooked dirty oil samples, the authentication task can't satisfy the needs of SVM. The semi-supervised learning (SSL) uses both of a few number of labeled and a large number of unlabeled samples for classification which has achieved considerable performance in computer vision, speech recognition, and information retrieval [14]. However, this approach has only few applications in NIR spectroscopy analysis. Therefore, in this paper, we propose spectral classification method based on SSL for authentication between edible oil and swill-cooked dirty oil in order to solve the problems in a complex and small sample calibration spectrum system.

Under the manifold assumption [15], the NIR spectral data of edible oils and the swill-cooked dirty oil are mapped to a specific high dimensional space, and these data lies on a low-dimensional manifold embedded in such space. The intrinsic geometric structure of spectral distribution can be approximated by the graph from both labeled and unlabeled spectral samples. The early graph-based methods like TSVM [16] are transductive and it is hard to deal with out-of-sample problem. The manifold regularization framework [17] combines the probability distribution estimating of spectral data and learning process of classification model by adding manifold and ambient regularizer at SVM optimal problem [18]. The manifold regularizer restrains the decision function for classification between edible oil and swill-cooked dirty oil along with the low-dimensional manifold by mining the information from the unlabeled spectral samples. Integrating with the kernel methods, the graph based semi-supervised support vector machine (GS3VM) is proposed to establish the rapid NIR classification model between edible oil and swill-cooked dirty oil.

In our investigation, the NIR spectra of different kinds of edible oils and swill-cooked dirty oil were collected. In order to simulate the real-world application, the calibration model for identification between edible oil and swill-cooked dirty oil was established on a few number of labeled and a large number of unlabeled oil samples by using GS3VM. Comparing with the supervised method like PLSDA and SVM, the GS3VM obtained a better performance whose prediction accuracy of several random tests is approximately 98%. It proved that the GBS3VM method can effectively used in swill-cooked dirty oil identification. The idea of semi-supervised model may provide reference to other NIR applications.

2. Theory and implementation

2.1. Support vector machine classification (SVM)

Support vector machine (SVM) was a widely used chemometrics method and it was successfully applied to NIR spectroscopy classification. SVM considered the l labeled oil spectral samples $\{x_i, y_i\}_{i=1}^{l}$ for calibration model, where x_i was the NIR spectroscopy of *ith* oil sample and the y_i represented the class label of the *ith* sample (edible oil or swill-cooked dirty oil). SVM searched the hyper plane of the largest intervals between classes by solving the following optimization problem listed in Eq. (1).

$$\min_{f\in H_k} \frac{1}{l}\sum_{i=1}^{l} \max(0, 1-yf(x)) + \gamma\|f\|_K^2 \tag{1}$$

H_k was the Hilbert space and penalized the classifier complexity in kernel space. Although there were different forms of H_k in Eq. (1), the solution of such quadratic programming problems were the same as Eq. (2).

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{2}$$

K represented the kernel function and the weight α_i was solved by sequential minimal optimization method (SMO) [19]. The class label of spectroscopy x in prediction set was predicted by substituting the x into Eq. (2). When training the classifier, the SVM needed large number of labeled samples in each class to ensure accuracy of classification. The traditional SVM didn't play a part in classification of edible oil and swill-cooked dirty oil on account of insufficient of swill-cooked dirty oil samples.

2.2. Manifold regularization

For classification between edible oil and swill-cooked dirty oil, both l labeled samples $\{x_i, y_i\}_{i=1}^{l}$ and u unlabeled samples $\{x_i\}_{i=1+1}^{l+u}$ were added to the calibration set. Manifold regularization was a useful tool to exploit the geometry from both labeled and unlabeled samples, and its constraint was carried out by adding a manifold regularizer $\|f\|_I^2$ to the objective function of SVM optimization problem:

$$\min_{f\in H_k} \frac{1}{l}\sum_{i=1}^{l} \max(0, 1-yf(x)) + \gamma_K\|f\|_K^2 + \gamma_I\|f\|_I^2. \tag{3}$$

The regularizer $\|f\|_I^2$ can be approximated by Laplacian Matrix that was considered as an effective graph composition method:

$$\|f\|_I^2 = f^T L f \tag{4}$$

where $f = [f(x_1), \ldots, f(x_{l+u})]^T$, and $L = D - W$ was the graph Laplacian matrix [20]. In a KNN neighboring areas, the affinity matrix W was defined as $W_{ij} = \exp(\|x_i - x_j\|^2/2\sigma^2)$ and the matrix D was a diagonal matrix with elements $D_{ii} = \sum_{j=1}^{l+u} w_{ij}$.

2.3. Graph based semi-supervised support vector machine (GS3VM)

With the constraint of manifold regularizer, the SVM optimization was transformed to graph based semi-supervised support vector machine (GS3VM) as listed in Eq. (5).

$$\min_{f\in H_k} \frac{1}{l}\sum_{i=1}^{l} \max(0, 1-yf(x)) + \gamma_K\|f\|_K^2 + \gamma_I f^T L f \tag{5}$$

Comparing with Eq. (2), the quadratic programming problem of Eq. (5) can be solved by representer theorem [21] and both $l + u$ samples were used in classification model.

$$f(x) = \sum_{i=1}^{l+u} \alpha_i K(x, x_i) + b \tag{6}$$

The weight α_i was figured out by Lagrangian multipliers technique after converting the Eq. (5) to its dual form. Due to the large complexity in computation of the optimizing and the regularizer, the preconditioned conjugate gradient and stopping strategy posed by Stefano Malacci [22] were used in performing the GS3VM solver in its primal problem. The main steps of GS3VM were described as follows.

(1) The spectra of oil samples which consisted of l labeled oil samples $\{x_i, y_i\}_{i=1}^{l}$ and u unlabeled oil samples $\{x_i\}_{i=1+1}^{l+u}$ were collected;
(2) The affinity matrix $W_{ij} = \exp(\|x_i - x_j\|^2/2\sigma^2)$ and the diagonal matrix $D_{ii} = \sum_{j=1}^{l+u} w_{ij}$ were calculated between neighboring oil spectra;
(3) The graph Laplacian matrix $L = D - W$ was computed;
(4) The γ_K, γ_I and K kernel function were chosen;
(5) The weight α_i in Eq. (6) was figured out by GS3VM solver;