Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 130 (2014) 245–249

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Discrimination between authentic and adulterated liquors by near-infrared spectroscopy and ensemble classification

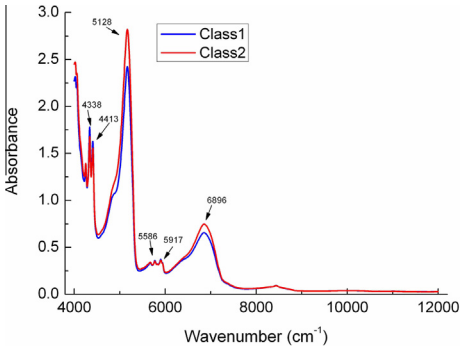
Hui Chen [a], Chao Tan [a,b,c,*], Tong Wu [a], Li Wang [a], Wanping Zhu [a]

[a] Key Lab of Process Analysis and Control, Yibin University, Yibin, Sichuan 644007, PR China
[b] Department of Chemistry and Chemical Engineering, Yibin University, Yibin, Sichuan 644007, PR China
[c] Computational Physics Key Laboratory of Sichuan Province, Yibin University, Yibin, Sichuan 644007, PR China

HIGHLIGHTS

- Near-infrared spectroscopy is used to discriminate liquors.
- Ensemble support vector machines serves as a tool.
- Such a method is fast and objective.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 9 January 2014
Received in revised form 5 March 2014
Accepted 18 March 2014
Available online 2 April 2014

Keywords:
Liquor
Authenticity
Near-infrared
Support vector machines

ABSTRACT

Chinese liquor is one of the famous distilled spirits and counterfeit liquor is becoming a serious problem in the market. Especially, age liquor is facing the crisis of confidence because it is difficult for consumer to identify the marked age, which prompts unscrupulous traders to pose off low-grade liquors as high-grade liquors. An ideal method for authenticity confirmation of liquors should be non-invasive, non-destructive and timely. The combination of near-infrared spectroscopy with chemometrics proves to be a good way to reach these premises. A new strategy is proposed for classification and verification of the adulteration of liquors by using NIR spectroscopy and chemometric classification, i.e., ensemble support vector machines (SVM). Three measures, i.e., accuracy, sensitivity and specificity were used for performance evaluation. The results confirmed that the strategy can serve as a screening tool applied to verify adulteration of the liquor, that is, a prior step used to condition the sample to a deeper analysis only when a positive result for adulteration is obtained by the proposed methodology.

Introduction

Chinese liquor (distilled spirits) is one of the famous distilled spirits and has been consumed for centuries in China [1]. As a complex matrix, liquor is composed of water, ethanol, inorganic elements and many kinds of fragrance ingredients such as esters, among which water and ethanol are the two major constituents and account for about 98% of the total mass. In recent years, counterfeit liquor is becoming a big, expensive issue in the market. Most of counterfeit liquors are prepared by simple dilution of the original liquors, by water or ethanol, or by a mixture of alcohol, water and aroma. Also, age liquor is facing the crisis of confidence

* Corresponding author at: Department of Chemistry and Chemical Engineering, Yibin University, Yibin, Sichuan 644007, PR China. Tel./fax: +86 831 3551080.
E-mail address: chaotan1112@163.com (C. Tan).

http://dx.doi.org/10.1016/j.saa.2014.03.091
1386-1425/

because it is difficult for consumer to identify the marked age, which prompts unscrupulous traders to pose off low-grade liquors as high-grade liquors. The authenticity of liquor is regulated by strict guidelines laid down by the responsible national authorities, which may include official sensory evaluation, chemical analysis, and examination of the records [2]. The identification or classification of liquors, mainly in terms of variety and geographical region of origin, has received increasing attention during the past 10 years. Most methods for screening liquor authenticity has been attempted based several different types of compositional data such as volatile compounds [3,4]. However, all of these methods include the pre-treatment process such as distillation, and the operation are complicated, time-consuming and laborious.

Nowadays, the application of spectroscopic techniques in wine analysis and quality control has developed considerably [5,6]. Especially, Near-infrared (NIR) spectroscopy offers the advantages of simplicity of sample presentation, the speed of collecting the information (spectra) and low cost [7–9]. It is well known that the NIR technique is based on the correlation between chemical properties and absorption of light at different wavelengths in the NIR region, measured by reflectance, transmittance or transflectance. Often, the NIR signal is very weak. So, it is inevitable to use modeling techniques including classification and regression to highlight useful information. Also, the performance of a model is decisive to the availability of NIR-based applications. In recent years, great effort has been made to develop model construction methods for improving the performance.

In conventional practice, most methods are based on constructing a single model, which in some cases may result in unsatisfactory performance, especially when the number samples is relatively small, which therefore leads to the emergence of the so-called ensemble technique [10–12]. Ensemble has gained increasing attention in many fields and has made a fundamental shift for model construction, i.e., instead of trying to build a single complex model, one can instead resort to combine a group of models. It enables an increase in generalization performance by combining several individual models trained on the same task. The ensemble advantage has been justified both theoretically and empirically.

An ideal method used to verify the quality and authenticity of liquors should be non-invasive, non-destructive and timely. In addition, it should to accomplish a fast data acquisition and data treatment accurately with relatively low costs. The combination of NIR spectroscopy with chemometric is a good way to reach these premises. In this work, it is proposed a new strategy for classification and verification of adulteration of liquors using NIR spectroscopy and chemometric classification, i.e., ensemble support vector machines (SVM). The results confirmed that the strategy can serve as a screening tool applied to verify adulteration of the liquor, that is, a prior step used to condition the sample to a deeper analysis only when a positive result for adulteration is obtained by the proposed methodology. Three measures, i.e., accuracy, sensitivity and specificity were used for model evaluation.

Materials and methods

Samples and spectra collection

A total of 120 samples of commercial bottled liquors were collected in local stores of China. It contained 42 liquor samples belonging to a brand without marked age from the markets (Class 1) and 78 samples with marked age (Class 2). These liquors belong to a type of distilled spirit made from grains based on a series of process such as fermenting, distilling, storage and blending, etc. The samples of class 2 were actually age liquor, i.e., the high-grade liquor product in China and therefore had a high price and the samples of class 1 were low-grade with low price. It is known that through ageing or maturation, the liquor body becomes more harmonious after undergoing a large number of physical and chemical reactions. Nowadays, age liquor is facing the crisis of confidence because it is difficult for consumer to identify the marked age, which prompts unscrupulous traders to pose off low-grade liquors as high-grade liquors. Each sample was confirmed by expert and composition analysis. The two classes had different distribution and abundance of esters such as ethyl acetate, which reflects characteristic aroma of liquor. All analyses were done in duplicate. For each sample, liquor bottles were opened and subsamples were scanned on in transmission mode using a near-infrared spectrometer coupled with an automated transmission module (Antaris II, Thermo fisher, USA). Spectral data collection was made using Vision software-TQ Analyst. Samples were scanned in a rectangular curette with a 1-mm path length and temperature equilibrated at 25 °C for 2 min in the instrument before scanning. For each sample, the spectrum was recorded in the region of 4000–12,000 cm^{-1} at intervals of 4 cm^{-1}, containing 2074 points and was an average of 32 scans. The absorbance in the region of 8000–12,000 cm^{-1} was very weak.

Ensemble support vector machines

In the past few years, there has been increased attention in the literature on the use of so-called ensemble methods in pattern recognition such as classification [13,14]. These methods have been shown to have interesting properties. Ensemble modeling is a word originated from machine learning. The idea behind ensemble methods is to generate a large number of alternative predictors obtained on independently perturbed data and combine the predictors either by averaging or majority vote strategies. The most well known method in the class of techniques is perhaps bootstrap aggregating (bagging) [15]. In fact, a known problem in machine learning is that classifiers with good training performance often exhibit a poor generalization performance on unseen data. A possibility to overcome these limitations is to create an ensemble of classifiers and average the output of all independent/member classifiers. This idea can be compared to the process of consulting several experts for making a final decision. For a medical practice, it means to seek the opinion of several doctors. By the way, it certainly reduces the risk of making a poor classification based on a single model. Another major advantage of ensemble classifiers is that they are able to successfully address small sample problem.

The mechanism of ensemble modeling can be analyzed in-depth by bias/variance decomposition of the error. More specifically, the generalization error of an ensemble model could be significantly improved if the predictions of different member models disagree and if their residual errors are uncorrelated. The success of the ensemble heavily depends on so-called trade-off between accuracy and diversity of member models in it. Different ways of achieving this give birth to different ensemble methods. Several ways are available. The most common way is based on the re-sampling of the training set to produce diverse subsets.

In this work, we use an effective scheme, i.e., a mixture of bagging and cross-validation. As one of the first successfully applied ensemble-based technique, bagging can improve the classification by combining classifiers trained on randomly generated sample subsets. Wichard [16] extended it by applying a cross-validation (CV) scheme for model selection on each subset and further combine the selected models to an ensemble model. Unlike traditional cross-validation, it adopts random subsets for cross-validation folds. The ensemble scheme consists of the following steps:

1. For the *K*-fold CV, the dataset is divided *K*-times into a training set and a test set, both sets containing randomly drawn subsets of the dataset without replications. The size of each test set was 20% of the entire dataset.