ARTICLE IN PRESS

Chemometrics and Intelligent Laboratory Systems xxx (2017) 1–8

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Authentication and inference of seal stamps on Chinese traditional painting by using multivariate classification and near-infrared spectroscopy

Zewei Chen [a,1], An Gu [b], Xin Zhang [a], Zhuoyong Zhang [a,*]

[a] Department of Chemistry, Capital Normal University, Beijing 100048, China
[b] The Palace Museum, Beijing 100009, China

ARTICLE INFO

Keywords:
Chinese traditional painting
Near-infrared spectroscopy
Multivariate models
Hierarchical cluster analysis
Seal

ABSTRACT

Chinese traditional paintings occupy an important position in Chinese cultural heritage and it is very important for archeologist and artist to identify their authenticity, which is difficult to be realized. Near-infrared spectroscopy (NIRS) coupled with multivariate models was used for authenticating stamps of 12 seals on a Chinese traditional painting in this work. The robustness of linear and nonlinear multivariate models, i.e. partial least squares-discriminant analysis (PLS-DA) and support vector machine (SVM), were evaluated by adding 5 different levels of noise (from 1% to 5%) into 3 original NIR spectra of each the stamps. These spectral data with noise added were fused together with original spectra to establish identification models and then to evaluate the abilities of the two models to tolerate noise disturbance. Accuracies of 92.6% and 100% were yielded by linear PLS-DA and nonlinear SVM methods respectively. The results demonstrate the feasibility of multivariate approaches in authenticating stamps of seals on the Chinese traditional painting. It is also important and necessary to infer the approximate eras of seal stamps on Chinese traditional painting in archeological study. By comparing the Mahalanobis distances between the 12 stamps on the painting, hierarchical cluster analysis (HCA) was adopted to assist the inference of eras for those unknown seal stamps on the Chinese traditional painting. This work demonstrates that NIR spectroscopy combined with multivariate models can be utilized as a non-destructive approach for authentication of stamps on Chinese traditional painting. HCA can also provide useful information to speculate the time period of the stamps of unknown seals on the Chinese traditional painting.

1. Introduction

Chinese traditional painting is considered as one of most important parts of Chinese cultural heritage and is well known for its unique artistic merits and charms all over the world. Such unique artistic merits are not merely characterized by the special artistic style of Chinese traditional paintings, one of the distinctive characteristics of Chinese traditional painting is the use of inscriptions of special seals as part of the painting itself. Chinese traditional paintings with seal are more complete and more distinctive as seals have become the essential constituent part of painting. Therefore, the stamps of seals on the paintings are also regarded as an indispensable part of Chinese traditional paintings, and have a significant role of enlivening the painting in aesthetic value. As seal stamps on a Chinese traditional painting can be printed by painter, or appreciator(s) or collector(s) in different time periods, the stamps of seals on Chinese traditional paintings may provide important historical information and would be meaningful for archeological research. Currently, the authentication and inference of seals on Chinese traditional paintings mainly depend on professional expertise and searching related information of stamps from Chinese ancient literature or Chinese historical materials, while this approach might be limited for some seals due to the lack of information in ancient literature recording. Although techniques related to X-ray have been utilized in the field of archeology, these techniques may result in damages to cultural heritage materials by some extent, especially for some precious antique paintings, and these damages could not be tolerable and reparable. X-ray may threaten the health of the operators. Other than X ray techniques, Raman spectroscopy [1] and FTIR spectroscopy [2] are two techniques most widely utilized in to archeological studies. However, many researches based on these techniques for art and painting studies were mainly focusing on the chemical composition analysis of pigments for paintings and arts [3–6]. In particular, Zhou et al. applied Raman spectroscopy to identify the

* Corresponding author.
E-mail address: gusto2008@vip.sina.com (Z. Zhang).
[1] On leaving for the Center for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Ohio University, Athens, OH 45701, USA.

https://doi.org/10.1016/j.chemolab.2017.10.017
Received 6 May 2017; Received in revised form 18 September 2017; Accepted 18 October 2017
Available online xxxx
0169-7439/© 2017 Published by Elsevier B.V.

molecular structure of the pigments and then the pigments in six imperial China engrave coiling dragon stamps. Sheng et al. proposed a windowed and entropy balance fusion for recognition of Chinese artists [4]. However, few researches on authentication and inference of stamps on Chinese traditional paintings by using spectroscopic techniques were reported. Objective evidence provided by non-destructive spectroscopic methods can be very helpful for identifying and inferring the time period of the Chinese traditional paintings as well as for conserving precious art paintings, and thus important for historical studies.

Recently, near infrared spectroscopy (NIRS) has been increasingly adopted as qualitative and quantitative tools in many fields including energy [5,6], environment [7], medicine [8] and agriculture [9], because NIRS is a fast, simple, cheap technique with nearly no sample preparation. Near infrared light has low radiation energy and no strong light source is used in the spectrometer. NIRS is a truly non-destructive method, and safe for almost all kinds of samples and materials. These advantages of NIRS were found to be perfectly suitable for valuable and in vivo or in vitro biological samples [10]. For the reason on safety, NIRS can also be a perfect approach for non-destructive authenticating the seals on precious art paintings. NIRS analysis involves the rapid acquisition of a large number of absorption values for a selected spectral range within a very short term. Therefore, NIRS is utilized in combination with multivariate approaches, i.e. chemometrics, to perform quantitative and qualitative analysis in many practical applications [11].

In archeology, some multivariate analysis methods have been introduced for various problems [12]. Multivariate analysis methods generally attempt to find the relationships between the objects and variables in a given data set and convert into new latent variable. Multivariate classification and multivariate regression are involved by using multivariate analysis methods, and predictions can be achieved based on the established models [13,14]. Typically, multivariate methods can be classified into supervised and unsupervised ones. The method to be used depends on the prior knowledge and requirement of the problem. Supervised methods try to divide samples into groups based on their characteristics by using a training set, while unsupervised methods can divide up data space into corresponding groups without any predefined training set [15]. Many multivariate models are normally based on supervised learning, including linear and nonlinear supervised methods, for example partial least squares-discriminant analysis (PLS-DA) and support vector machine (SVM). These commonly used supervised approaches have been adopted in different fields, including archeology, to solve various discriminant problems [16–19]. Whereas, the unsupervised approaches are important for some problems of real-world data mining in which the data from practical applications are often unlabeled and no prior knowledge is available [20]. Clustering analysis is one of important parts in unsupervised learning [21], and the hierarchical clustering analysis (HCA) is regarded as a main stream clustering method [20]. As the advantages of supervised and unsupervised methods, both of them were exploited in this paper.

In this paper, we investigated the feasibility of using NIRS coupled with multivariate models as a noninvasive technique to authenticate seal stamps on a Chinese traditional painting. The robustness of linear and nonlinear models, i.e. PLS-DA and SVM respectively, was evaluated by adding different proportional disturbance of noise into the original NIR spectra. In addition, HCA provides useful information to infer the time period of those unknown seal stamps on the Chinese traditional painting based on the similarities between 12 stamps.

2. Methods

2.1. Partial least squares-discriminant analysis (PLS-DA)

PLS-DA, a particular case of PLS algorithm, is a supervised classifier that models the relationship between the feature variables and the target variables like class labels [22]. PLS-DA is a commonly used linear classifier for solving classification problems based on its easy operation and excellent ability for classification [23]. Determining number of the latent variables is critical for PLS-DA. Through extracting the latent variable, the explanation for the variance of the measured variables as well as the correlation with the response matrix that encodes the class membership, data dimension can be reduced and the maximum separation among the classes can be achieved [24,25].

In PLS analysis, selecting an optimal number of latent variables is a critical step. As mentioned above, it is important to determine how many latent variables should be used to yield optimal PLS-DA model. The optimal number of latent variables might vary with the variation of training set data. Although using a single number of latent variables for all predictions is easy to manipulate, it does not yield the optimal model and statistical power of predictions would be low. In this work, a PLS approach with self-determining latent variables was used to determine the optimal number of latent variables from the training data set within each bootstrap Latin partitions (BPLs). BPLs are based on cross validation and random sampling verification for evaluating the accuracy and stability of calibration methods. Five Latin partitions and ten bootstraps repetition according to our expertise were used to evaluate the classification accuracy of the models with latent variables. All samples in training set were divided into five equal parts randomly for each bootstraps repetition and every part was used as test set only once for prediction, and the rest four parts were used as the training set to build the calibration model. The results of the prediction from each partition were pooled. The number of latent variables that yields the least prediction error was chosen and then automatically used for the entire training set to construct the model and for the validation set to predict. These procedures can be implemented with MATLAB code. The number of latent variables with the lowest mean square error was chosen and then used to construct the PLS-DA classifier. By such self-optimization, PLS-DA can not only work as a parameter-free method, but also utilize optimal parameter for modeling.

2.2. Support vector machines (SVM)

SVM is a nonlinear classifier that is commonly used to solve classification problems. More detail description of SVM can be found elsewhere [26,27]. In brief, SVM algorithm attempts to find the optimal hyperplane that can separate all objects of one group from those of the other group with the largest margin between the two groups. The largest margin can achieve the maximal width of the slab parallel to the hyperplane while there is no interior data point in this region. The data points that located far from the boundary are removed from the calculation while those data points on the boundary of the slab will be maintained and determined as the so-call support vectors to obtain satisfactory computational efficiency. The SVM classifier is limited to solving binary problems. For k-class (k is greater than 2) classification problems, a strategy is to construct a SVM model between each two classes. With that strategy, $k\cdot(k-1)/2$ SVM models are established to assign an unknown sample into the closest group based on the weights calculated from the $k\cdot(k-1)/2$ SVM models.

In this work, SVM package was taken from Ref. [28]. The RBF kernel was applied to check the potential of using nonlinear models for analyzing near infrared data. The penalty weight C and g optimized by grid searching method were used as its basic parameters and set as 32 and 0.0034, respectively.

2.3. Hierarchical cluster analysis (HCA)

HCA groups data over a variety of scales by creating a cluster tree or dendrogram. Instead of a single set of clusters, the cluster tree is a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows one to determine which level of clustering is most proper for practical applications and studies. Typically, strategies for HCA can fall into two types: agglomerative and divisive HCA [20]. The former is commonly used in clustering applications because of its