Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Fast discrimination of the geographical origins of notoginseng by near-infrared spectroscopy and chemometrics

Hui Chen[a,b], Zan Lin[a,c], Chao Tan[a,*]

[a] Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin, Sichuan 644000, China
[b] Hospital, Yibin University, Yibin, Sichuan 644000, China
[c] The First Affiliated Hospital, Chongqing Medical University, Chongqing 400016, China

ARTICLE INFO

Article history:
Received 1 August 2018
Received in revised form 22 August 2018
Accepted 25 August 2018
Available online 27 August 2018

Keywords:
Notoginseng
Near-infrared spectroscopy
Partial least squares-discriminant analysis

ABSTRACT

Notoginseng is a type of highly valued Traditional Chinese medicine (TCM) due to its hemostatic and cardiovascular functions. Notoginseng of Yunnan in China usually commands a premium price and is often the subject of fraudulent practices. The feasibility of combining near-infrared (NIR) spectroscopy with chemometrics was investigated to discriminate notoginseng of different geographical origins. A total of 250 samples of four different provinces in China were collected and divided equally into the training and test sets. Principal component analysis (PCA) was used for observing possible trend of grouping. Two chemometric algorithms including partial least squares-discriminant analysis (PLSDA) and soft independent modeling of class analogy (SIMCA) were used to construct the discriminant models. Standard normal variate (SNV) and first derivative were used for pre-processing spectra. On the independent test set, the PLSDA model outperforms the SIMCA model. When combining both pre-processing methods, the constructed PLSDA model achieved 100% sensitivity and 100% specificity on both the training set and the test set. It indicates that SNV+first derivative pre-processing and PLSDA algorithm can serve as the potential tool of fast discriminating the geographical origins of notoginseng.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Traditional Chinese medicine (TCM) has played an important role in clinical therapy due to its ability to cure disease with considerable side effects [1]. Notoginseng, known as äspirin in TCM”, is one of the most common herbs used TCM and health-maintaining products [2,3]. However, even for the same species, its chemical composition can be influenced by many factors, such as climate and soil [4,5], so that quality and efficacy owing to different geographical origins are somewhat different. Attribute of notoginseng according to geographical origins is often recognized and appreciated by the consumers. Notoginseng of Yunnan is recognized as the best efficacy and its protected geographical indication (PGI) has been approved by Chinese authorities. Notoginseng of Yunnan province usually commands a premium price and is often the subject of fraudulent practices by replacing them with notoginseng of other regions at low cost, which lead to an unfair competition in the pharmaceutical markets and harm the interests of consumers. Therefore, a rapid and reliable analytical method to determine the origins is of great importance.

However, it is not easy and almost impossible to determine the geographical origins by existing analytical tools or through visible inspection [6]. Also, since there are tens of chemical components, which are slightly different according to geographical origins, it is not scientific to select only several specific components as the index. Its pharmacological activity is from the mixture of its constituents and not the presence of a single compound [7]. Even if there exist a method able to analyze many chemical components, it is time-consuming, laborious, expensive or require highy-skilled operators. Near-infrared (NIR) spectroscopy can be a potential tool for quality control of medicines and foods including detecting the geographical origins since it is rapid, and non-destructive, reagent free and requiring minimal or no sample preparation [8–15]. NIR has attracted considerable attention on discrimination between samples of similar biological materials such as food [16,17] and gasoline [18]. The NIR spectroscopy cover the wavelength range of 800–2500 nm and mainly reflects information on hydrogen-containing bonds such as C—H, N—H and O—H of organic molecules.

Although NIR spectra can not provide some significant differences of spectral peaks like mid-infrared (MIR) spectra, they include abundant chemical and structural information of samples. How-

[*] Corresponding author.
E-mail address: chaotan1112@163.com (C. Tan).

https://doi.org/10.1016/j.jpba.2018.08.052
0731-7085/© 2018 Elsevier B.V. All rights reserved.

ever, owing to the overlaps and the systematic noise in NIR spectra, it is necessary to apply appropriate pre-treatment methods and chemometric methods for qualitative or the quantitative analysis [19,20]. Also, these methods are maybe decisive. Chemometrics algorithms are numerous, and the challenging task is to choose the most appropriate algorithm. Among these algorithms, some are more popular. For example, partial least squares-discriminant analysis (PLSDA) [21] is a popular classification method that combines partial least squares (PLS) regression with the discrimination ability of a classification technique. It is based on the PLS regression algorithm and focuses on searching for latent variables (LVs) with a maximum covariance. The soft independent modeling of class analogy (SIMCA) [22] is a classical class-modeling tool by incorporating principal component analysis (PCA) to reduce the dimensions of the data. Recently, NIR and mid-infrared (MIR) spectroscopy have been successfully used to detect adulteration of notoginseng powder [23,24]. To the best of our knowledge, no research on the discrimination of geographical origins of notoginseng by NIR technique have been reported so far.

In the present work, the feasibility of combining NIR spectroscopy with chemometrics was investigated to discriminate notoginseng of different geographical origins. A total of 250 samples of four different provinces in China were collected and divided equally into the training and test sets. PCA was used for observing possible trend of grouping. Two chemometric algorithms including PLSDA and SIMCA were used to construct the discriminant models. Standard normal variate (SNV) and first derivative processing were used for pre-processing spectra. On the independent test set, the PLSDA model outperforms the SIMCA model. When combining both pre-processing methods, the constructed PLSDA model achieved 100% sensitivity and 100% specificity on both the training set and the test set. It indicates that SNV + first derivative pre-processing and PLS-DA algorithm can serve as the potential tool of fast discriminating the geographical origin of notoginseng.

2. Theory and methods

2.1. PLSDA

PLS regression is a multivariate tool for modeling a relationship between a descriptor matrix **X** and a response matrix **Y** [25]. PLS regression has been widely used in multivariate calibration where the response matrix is quantitative, but it can also been additionally employed for qualitative discrimination/classification in the form of PLS-discriminant analysis (PLSDA) [26–28].

Assume sample i is described by a feature vector $\mathbf{x}_i \in R^p$ where p often takes large values. Feature vectors acquired for n samples constitute a data matrix $\mathbf{X} \in R^{n\times p}$. Each sample has a class label and thus can be described by a binary vector $\mathbf{X} \in R^{n\times p}$ where k is the number of categories (in the case of two classes it is enough to take k = 1). Then for n samples a dummy matrix $\mathbf{Y} \in R^{n\times k}$ might be defined as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2\mathbf{y}, \cdots, \mathbf{y}_n\}$, among which, the class vector $\mathbf{y}_i$ represents the membership of a sample to the k classes) while each entry y_{ik} of $\mathbf{y}_i$ represents the membership of the ith sample to the kth class expressed with a binary code (1or 0). The aim of modeling is to allow the accurate prediction of $\mathbf{y}_{\text{new}}$ from the measurement of $\mathbf{x}_{\text{new}}$.

Broadly, PLSDA can be regarded as a PLS regression between a set of predictors **X** and label responses **Y**, with a binary outcome. PLS defines a new subspace of LVs by an iterative process, which aims at taking a compromise between maximum variance in **X** and maximum correlation to **Y**. The projection of the **X**-matrix into the defined hyperplane is realized by the X-scores (**T**) and is defined in Eq. (1).

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^{\mathrm{T}} + \mathbf{F} \tag{3}$$

T is result of the linear combination of the original variables with the weights **W**, **T** model **X** (Eq. (2)); when multiplied by the loadings **P**, X-scores are good approximations of **X** and the X-residuals, **E**, are usually small. On the other hand, **Y** can be predicted by the X-scores and the matrix **C** (Eq. (3)). The Y-residuals, **F**, are the deviations between the actual and modeled responses. Finally, the relationship between **X** and **Y** that PLS specifies is given by Eq. (4), where **B** is a matrix of the PLS-regression coefficients (Eq. (5)).

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{F} \tag{4}$$

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{C} \tag{5}$$

Similar to the PLS2 regression in the usual way, PLS-DA can provide the estimated values $\left(\hat{y}_{ik}\right)$ for each ith sample and for each kth class. The estimated class values does not have either 1 or 0 values perfectly; but, if $\hat{y}_{ik}$ is closer to zero, then the ith sample is not likely from the kth class, while a value closer to one would imply the opposite.

For making a class assignment, the probability that a sample belongs to the defined classes can be obtained by the estimated class values. Once the probability is calculated for each class and it can be assigned to the class that has the highest probability. Under this approach, samples are always classified to one of the classes. However, a threshold can also be calculated for each class: if he estimated class values is greater than the threshold of the kth class, then the ith sample is assigned to the kth class, otherwise not. Thresholds can be calculated by the Bayes theorem. It assumes that the estimated values follow a normal distribution, which is also comparable to what will be observed in future samples. The threshold corresponds to the point where false positive (FP) and false negative (FN) are minimized. It is possible that the estimated class values for a specific sample are higher or lower than the thresholds of all the defined class. Thus, if higher, the sample would be assigned to all classes and, accordingly, is a confused sample. Conversely, if lower, the sample would not be assigned to any class. In both cases, the sample can be regarded as ‘not assigned’.

2.2. SMICA

Soft Independent Modeling of Class Analogies (SIMCA) is a well-known supervised classification technique. In SIMCA, a training set is modeled by PCA and new samples can be fitted to the model and are classified according to their similarity or dissimilarity to the training set. The theory of SIMCA has already been extensively discussed by several papers [29,30]. Only a brief introduction of SIMCA is therefore presented. In its original form, each class is modeled separately in terms of the similarity of the samples within the class. The model is constructed by PCA with a certain number of PCs. This is described by the following Eq. (6) for one class k,

$$\mathbf{X}_k = \bar{\mathbf{X}}_k + \mathbf{T}_k(n \times r)\mathbf{V}_k^T(r \times p) + \mathbf{E}_k(n \times p) \tag{6}$$

where $\bar{\mathrm{X}}_k$ is the mean centered data matrix, $\mathbf{T}_k(n \times r)$ the score matrix obtained for n objects and r selected PCs; $\mathbf{V}_k^T(r \times p)$ is the loading matrix using the first r PCs on p variables and $\mathbf{E}_k(n \times p)$ is the residual matrix. The selection of a correct number of PCs, r, is a key in SIMCA. Several methods are available for this purpose but cross-validation is the most popular. The class boundaries, or confidence limits, are then constructed around the PC model. They are based on the distribution of the distances measure between the

Download English Version:

https://daneshyari.com/en/article/10134886

Download Persian Version:

https://daneshyari.com/article/10134886

Daneshyari.com