Journal of Image Processing Theory and Applications (2021) Vol. 4: 56-59
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2021.41009
ISSN 2560-6239

Research on identification of medicinal materials by near and mid infrared spectroscopy

Yifan Cheng

Linyi University, Linyi, Shandong, 276005, China

Keywords: characteristics and differences, mid-infrared spectrum, Potential semantic analysis model

Abstract: In this paper, the characteristics and differences of medicinal materials from different origin were analyzed and the origin was identified according to the data of mid-infrared spectrum. Firstly, data integration was carried out, which was divided into 11 producing areas according to data standards,and the mid-infrared spectrograms of medicinal materials[1] from different producing areas were drawn by MATLAB. According to this, it was found that the spectrograms of the same medicinal materials from different producing areas had small differences, and the absorbance of some producing areas was higher than that of others, and there were slight differences in the trend of waveform changes. Potential semantic analysis model was established to obtain similarity function and correlation calculation to obtain the result of origin identification.

1. Background

The spectrum characteristic differences presented different kinds of Chinese herbal medicine, even the same materials, different regions affected by the chemical composition of inorganic elements, organic matter and other factors, in under the irradiation of near infrared, middle infrared spectrum will present a different spectrum characteristics, thus can use these characteristics to identify the types of Chinese herbal medicine as well as the origin[2]. The spectral difference of different kinds of Chinese medicinal materials is obvious.

2. Modeling and solving of problem 1

2.1 Model Establishment

2.1.1 Latent semantic analysis model

Latent semantic analysis is an algebraic model of information retrieval, which is often used for information acquisition and theoretical presentation.By analyzing a large number of text sets with statistical methods, the mapping rules between terms are extracted and quantified.

One middle infrared spectrum reflects the information of a sample of Chinese herbal medicine, and the crest section reflects the material composition information of the sample[3]. For the spectrogram, the absorbance is a function of wavelength, i.e.$a = f(x)$, and the important information of the spectrum is represented by a certain crest segment. Now, a peak search function G is used to find the peak value P:

$$p = G(y) = G(f(x)) \tag{1}$$

For this function, construct a rectangular window function:

$$y = \psi(p, \lambda) \tag{2}$$

$$T = \psi(G(f(x)), \lambda) \tag{3}$$

Where p is the wavelength at a wave peak, λ is the width parameter controlling the interception window, and its value range is [0,1]. When λ is larger, the window is wider.According to the scheme, a full-band is selected when λ =1.When λ is smaller, the window is narrower, and the peak value is selected when λ =0.

2.1.2 The questions correspond to the whole sampleT

For a question form q, calculate its vocabulary:

$$T_q = \psi\left(G\left(f_q(q), \lambda_q\right)\right) \tag{4}$$

λ_q can be set to a common value,or it can be set to:

$$\lambda_q = \sum_{j=1}^{d} \lambda_j / d \tag{5}$$

If the union of T_q and T_s is taken, the meaning is that the key bands of questioning appear together in each sample of each type of Traditional Chinese medicine, denoted as (basis of questioning matrix):

$$T_{sq} = T_s \cap T_{q'} \tag{6}$$

If there are a total of m bands in T_{sq} (m must be less than t), it can be written as:

$$T_{sq} = \left\{T_{v_1}, T_{sq2}, T_{s_q3}, \dots, T_{sqmt}\right\} \tag{7}$$

$T_s = \{T_{s1}, T_{s2}, T_{s3}, \dots, T_{st}\}$ is compared to $T_{sq} = \left\{T_{sq1}, T_{sq2}, T_{sq3}, \dots, T_{sqm}\right\}$,since T_{sq} is a subset of T_s, the element of T_{sq} is compared to the element of T_s one by one. If an element in T_{sq} has an intersection with an element in T_s, the item is set to the intersection of the two elements, otherwise it is set to 0. Rewrite T as:

$$T = \{T_1, T_2, T_3, \dots, T_t\} (1 \le i \le t) \tag{8}$$

For a certain band, the corresponding absorbance of T_i is expressed as:

$$a_i = f(T_i) \tag{9}$$

The corresponding absorbance of n samples of class j can be expressed as:

$$a_{j1} = f_1(T_i), a_{j2} = f_2(T_i), a_{j3} = f_3(T_i), \cdots, a_{jn} = f_n(T_i) \tag{10}$$

The absorbance of the Kth sample can be expressed as:

$$a_{jk} = f_k(T_i) \tag{11}$$

The approximate degree of the absorbance corresponding to a given band T_i and the absorbance of the Kth sample in the band T_i can be measured by different functions.

To facilitate expression, an evaluation function is given:

$$k(x, \beta) = \begin{cases} 1x < \beta \\ 0x \ge \beta \end{cases} \tag{12}$$

2.2 Model solution

2.2.1 The spectrograms of 11 kinds of medicinal materials from producing areas

The spectrograms of 11 kinds of medicinal materials from producing areas are shown in the figure below:

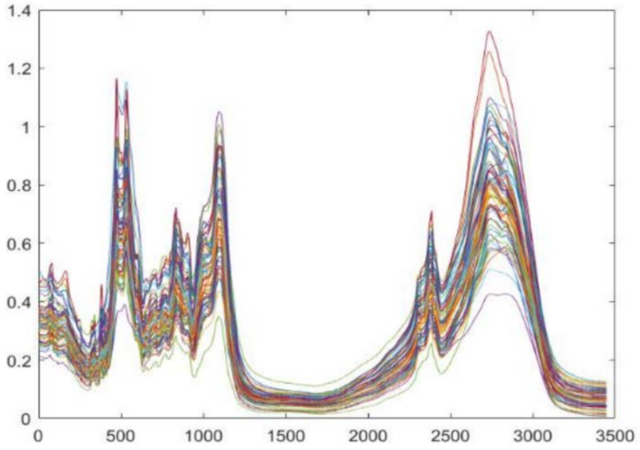

Figure 1: The spectrogram of medicinal materials from origin 1

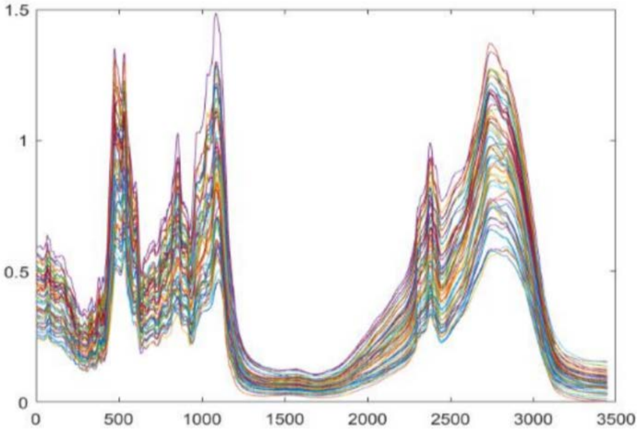

Figure 2: The spectrogram of medicinal materials from origin 2

According to the figure analysis, the spectra of the same medicinal material from different producing areas have certain similarity, but the difference is small.

The absorbance of no.2 in producing area 700 and 2400cm^{-1} was close to 1, which was higher than the other 10, and the peak difference was low.

The absorbance of no.5 and no.10 at 1100cm^{-1} is 1.4, and the waveform changes greatly.

The waveform of no.3 and no.6 region changed slowly from 2000 to 2300cm^{-1}.

2.2.2 Similarity analysis

For a certain question form T_i and a certain text, D_i can calculate the degree of similarity between the current question form and the known text by applying Angle cosine.The calculation formula is as follows:

$$C_{td} = \frac{\sum_{i=1}^{k} T_i D_i}{\sqrt{\sum_{i=1}^{k} (T_i)^2} \cdot \sqrt{\sum_{i=1}^{k} (D_i)^2}} \tag{13}$$

For unsupervised pattern recognition, according to the C_q value clustering, the retrieval results are produced, and the retrieval is completed.For supervised pattern recognition, the one with the highest C_q value is the one to be retrieved.

In the formula, k is the dimension of potential semantic space, and this paper takes 3.The similarity between a certain middle infrared spectrum to be analyzed and 11 known Chinese medicinal materials was calculated.

The identification list of origin is as follows:

Table 1: The identification list of origin

NO	3	14	38	48	58	71	79	86	89	110	134	152	227	331	618
OP	3	1	4	1	8	5	4	6	3	10	9	7	5	2	11

3. Evaluation of Model

3.1 Advantages

(1) The latent semantic analysis model provides the mapping of vocabulary, text, questioning and lexicon - text matrix in nIR spectral analysis, which can realize relevant calculation.

(2) The latent semantic analysis model can make full use of redundant data to improve accuracy.

3.2 Disadvantages

(1) There is no definite index for the number of classification types, and large errors may occur after generating random numbers.

(2) The spectra of the same medicinal herbs from different origins are relatively close, and the error of spectral identification is large.

References

[1] An Shujing, WANG Ting, Niu Dou, et al. Identification and analysis of Cornus officinalis from different producing areas based on mid-infrared spectroscopy combined with stoichiometry [J]. Chinese Journal of Traditional Chinese Medicine, 49(8): 6.

[2] LI Shuifang, Shan Yang, ZHU Xiangrong, et al. Near infrared spectroscopy combined with stoichiometry was used to determine the origin of honey [J]. Transactions of the Chinese Society of Agricultural Engineering, 2011(08): 350-354.

[3] TENG Y.Application of partial least square method in spectral analysis [J]. Application of Integrated Circuits, 37(1): 2. (In Chinese)