CHINESE JOURNAL OF ANALYTICAL CHEMISTRY
Volume 42, Issue 9, September 2014
Online English edition of the Chinese language journal

Available online at www.sciencedirect.com

ScienceDirect

Cite this article as: Chin J Anal Chem, 2014, 42(9), 1379–1386.

REVIEW

Algorithms, Strategies and Application Progress of Spectral Searching Methods

CHU Xiao-Li*, LI Jing-Yan, CHEN Pu, XU Yu-Peng
Research Institute of Petroleum Processing, Beijing 100083, China

Abstract: In recent years, many modern spectral databases for complex materials (such as soil, feed, forensic evidence materials, pharmaceuticals, oils, and so on) have been established on the basis of molecular spectroscopy (UV, infrared, near infrared, Raman and fluorescence), which are playing more and more important roles in the agricultural, industrial production and science research. Spectral searching method is one of the key techniques to make full use of the molecular spectral database. This paper reviewed the progress in the basic and modified algorithm, strategy and application of molecular spectral searching methods, and discussed the scientific and technological problems that need attention and further research.

Key Words: Molecular spectroscopy; Spectral searching algorithm; Correlation coefficient; Moving window; Review

1 Introduction

In recent years, with the improvement of equipment manufacturing and the popularization of chemometric methods, molecular spectroscopy analysis technology, especially the infrared, near infrared and Raman spectroscopy, has been widely applied in many fields due to the advantages of convenient test, fast speed, rich information, on line analysis and so on. By the methods of pattern recognition, clustering or recognition of molecular spectra was used for the analysis of complex system, such as oil, grain, fruit, and medicines[1].

In chemometrics, as shown in Fig.1, the modern pattern recognition methods of molecular spectroscopy analysis include three categories[2]: (1) unsupervised methods, such as principal component analysis, clustering method, K-means clustering method, and self-organizing neural network; (2) supervised methods, such as Linear discriminat analysis (LDA), Soft independent modeling of class analogy (SIMCA), Discriminant partial least squares (DPLS), and Support vector machine (SVM); all the methods above are based on the sample types, when a new sample is added to the database, the qualitative models must recalibrated. (3) spectral searching methods, such as correlation coefficient, cosine, Euclidean distance, and spectral information divergence. Based on the spectra of unknown samples, spectral searching methods can qualitatively and quantitatively analyze samples by searching the most similar one or more samples from the built spectral library.

Previously, the spectral searching methods were mainly used for spectral identification of pure compounds, such as Sadtler and Aldrich infrared spectrum database. In recent years, the modern molecular spectra databases of complex mixtures established gradually in many fields (such as soil, feed, evidence materials, medicines, and oil)[3–5], and spectral searching algorithm was more and more popular[6]. New searching algorithm and strategies emerged, and the accuracy and reliability of the spectral searching results were significantly improved. Compared with the unsupervised and supervised pattern recognition method, the spectral searching method has the advantages of simple operation, visual information and convenient maintenance of library, which plays an important role in practical applications. In this review, novel molecule spectral searching algorithms, strategies and their applications are introduced. The scientific and technological challenges are also discussed.

Received 13 May 2013; accepted 2 July 2014

* Corresponding author. Email: cxyuli@sina.com

DOI: 10.1016/S1872-2040(14)60768-4

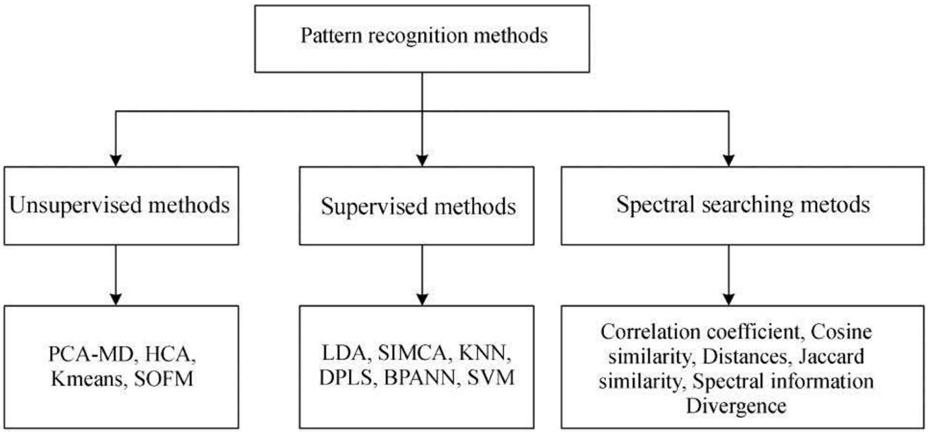
Pattern recognition methods
Unsupervised methods
Supervised methods
Spectral searching metods
PCA-MD, HCA, Kmeans, SOFM
LDA, SIMCA, KNN, DPLS, BPANN, SVM
Correlation coefficient, Cosine similarity, Distances, Jaccard similarity, Spectral information Divergence

Fig.1 Classification diagram of pattern recognition methods

2 Basic spectral searching algorithm

For the ***x*** spectrum of unknown sample, the aim of spectral searching is to find the most similar one (or more) sample spectrum to ***x***, based on certain algorithms and rules. If the properties matrix ***Y*** is known in spectral library, the properties of unknown sample can be predicted according to the spectral searching results.

In details, ***x*** represents unknown spectrum, organized as 1×*m* vector, *m* represents wavelength points; ***R*** represent all spectra in the library, organized as *n* × *m* matrix, where *n* is the number of sample; $\boldsymbol{r}_j$ represents the *j*th sample in the spectral library, organized as 1×*m* vector, *j* = 1, 2, ..., *n*; ***Y*** represents properties of corresponding spectra in library, organized as *n* × *p* matrix, *p* represents the number of properties; $\boldsymbol{y}_j$ represents property value of the *j*th sample in spectral library, organized as 1 × *p* vector.

In order to obtain the optimal results, spectral pretreatment and wavelength range selection are needed. Pretreatment methods include derivative, vector normalization, standardization, and wavelet transform. Based on chemistry knowledge and mathematics, wavelength selection methods can identify those spectral intervals which are strong characteristic, high in signal to noise ratio, and less vulnerable to external factors. There are several references related to the common used spectral pretreatment and wavelength selection methods[7].

2.1 Distance based algorithm

The basic principle of this algorithm is that the more similar the spectra of two samples are, the shorter the distance between two samples in the spectra. There are various forms of spectral distance, and the absolute distance is the Absolute distance between the sample spectrum ***x*** and the *j*th sample, represented as $\boldsymbol{r}_j$ in the spectra library, can be expressed as following:

$$\mathrm{d}(\boldsymbol{x}, \boldsymbol{r}_j) = \sum|\boldsymbol{x} - \boldsymbol{r}_j| \quad (1)$$

The Euclidean distance, known as the least square distance, is defined by the formula:

$$\mathrm{d}(\boldsymbol{x},\boldsymbol{r}_j)=\sqrt{(\boldsymbol{x}\text{-}\boldsymbol{r}_j)(\boldsymbol{x}\text{-}\boldsymbol{r}_j)^{\mathrm{T}}} \quad (2)$$

2.2 Similarity algorithm

There are two parameters for evaluation of the similarity between two spectra: cosine and correlation coefficient.

The cosine between ***x*** and $\boldsymbol{r}_j$ is expressed as follows:

$$\cos(\boldsymbol{x},\boldsymbol{r}_j) = \frac{\boldsymbol{x}\boldsymbol{r}_j^{\mathrm{T}}}{\sqrt{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}}\sqrt{\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}}} \quad (3)$$

The basic principle of this algorithm is that the smaller the cosine angle is, the greater the similarity of two samples is. If the two spectra are entirely identical, the cos(***x***, $\boldsymbol{r}_j$) = 1, the two samples in the pattern space get close to one point; if the two spectra are completely different, cos(***x***, $\boldsymbol{r}_j$) = 0.

The correlation coefficient between the ***x*** and $\boldsymbol{r}_j$ is expressed as follows:

$$R(\boldsymbol{x},\boldsymbol{r}_j) = \frac{(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{r}_j-\bar{\boldsymbol{r}}_j)^{\mathrm{T}}}{\sqrt{(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}}\sqrt{(\boldsymbol{r}_j-\bar{\boldsymbol{r}}_j)(\boldsymbol{r}_j-\bar{\boldsymbol{r}}_j)^{\mathrm{T}}}} \quad (4)$$

The x and r_{j} are average values of ***x*** and $\boldsymbol{r}_j$, respectively. If the value is closer to 1, it means that the two spectra are more similar; if the value is closer to 0, it means that the two spectra are more different.

2.3 Algorithm based on information theory

Spectral information divergence (SID) [8] can be used to evaluate spectral similarity by relative entropy of spectral information:

$$\mathrm{SID}(\boldsymbol{x}, \boldsymbol{r}_j) = \mathrm{D}(\boldsymbol{x}\|\boldsymbol{r}_j) + \mathrm{D}(\boldsymbol{r}_j\|\boldsymbol{x}) \quad (5)$$

where, D(***x***‖$\boldsymbol{r}_j$) is the relative entropy with $\boldsymbol{r}_j$ to ***x***, and D($\boldsymbol{r}_j$‖***x***) is the relative entropy with ***x*** to $\boldsymbol{r}_j$:

$$\mathrm{D}(\boldsymbol{x}\|\boldsymbol{r}_j) = \sum_{i=1}^{m} q_i \lg\left(\frac{q_i}{p_{j,i}}\right) \quad (6)$$

$$\mathrm{D}(\boldsymbol{r}_j\|\boldsymbol{x}) = \sum_{i=1}^{m} p_{j,i} \lg\left(\frac{p_{j,i}}{q_i}\right) \quad (7)$$