Chemometrics and Intelligent Laboratory Systems 96 (2009) 252–257

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Autoregressive model based feature extraction method for time shifted chromatography data

Weixiang Zhao, Cristina E. Davis *

Department of Mechanical and Aeronautical Engineering, University of California, Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 18 September 2008
Received in revised form 23 February 2009
Accepted 25 February 2009
Available online 6 March 2009

Keywords:
Feature extraction
Wavelet analysis
Autoregressive
Chromatogram
PCA
Neural networks

ABSTRACT

This study aims to demonstrate the distinct advantages of a novel feature extraction method based on autoregressive (AR) model for the classification of the chromatography data with time shifts. Two typical classification methods, principal component regression (PCR) and neural networks (NN), were employed to compare the classification effects of three types of feature spaces: the chromatogram AR coefficients obtained through an AR model, the chromatogram wavelet coefficients, and the original chromatography data. The results indicate: (1) for the normal non-time shifted chromatography data, three types of feature spaces show almost equally good classification results; (2) for the time shifted chromatography data, the classification effect based on the AR coefficients is significantly better than those based on the other two feature spaces, especially when an NN model is employed for the nonlinearity in the classification system; and (3) without time alignment, the damage caused by the time shifts to the classification based on the chromatogram wavelet coefficients and the original chromatography data is not easily overcome even by employing an NN model. This study not only demonstrates the distinct ability of this proposed feature extraction method to free us from time alignment for the chromatography data with time shifts, but also illustrates the robustness of this feature extraction method in terms of the AR model order selection. The AR model based feature extraction method provides a novel, fast, and reliable strategy for chromatogram characterization and classification, with a great potential to benefit the development of automated instrumentation systems.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Chromatography is a powerful analytical technique, especially when it is coupled with mass spectrometry to generate instruments such as gas chromatography mass spectrometry (GC/MS). Because of the high dimensional data of chromatography samples, various methods have been utilized to extract features from original data for classification and calibration [1–3]. Principal component analysis (PCA) and partial least squares regression (PLSR) are two widely used methods in this field [4,5]. Wavelet analysis is another typical feature extraction method for chromatography data and other spectral data, simultaneously examining signal in both time and frequency domains [6,7]. Genetic algorithms (GA) [8] and simulated annealing (SA) [9] are also used to select "representative" regions of chromatography data for system characterization and classification [10–12]. However, GA and SA often require long searching time to find discernable chemicals for chromatogram classification and the detected chemicals cannot be guaranteed to have sufficient physical meaning.

More importantly, these current methods usually require time alignment as a data pre-processing step, when chromatography data show time shifts. It is common, and sometimes almost unavoidable, to have time shifts in chromatography data due to a variety of minor variations of experiment conditions or outside disturbances. Therefore, it is very attractive to develop novel feature extraction methods with the ability to free us from time alignment of the original chromatography data.

Recently, an autoregressive (AR) model based filter was introduced to smoothen and de-noise chromatography data [13]. AR modeling [14] is an effective method for signal processing such as audio processing [15,16] and speaker/speech recognition [17,18], but there have been very few reports of the application of the AR model to chromatogram characterization and classification. A recent report has illustrated the feasibility of an AR model to extract features from chromatography data, employing a bacterial species classification problem in which headspace gas above the bacteria cultures was analyzed with GC/MS to determine speciation [19]. However, the most significant advantage of the AR model based feature extraction method has not yet been demonstrated in that paper, namely freedom from signal time alignment.

Intuitively, we understand this concept when considering how AR modeling is used in different disciplines. For example, using an AR model for speaker recognition does not require that speakers must speak at the same exact speed. Likewise, considering chromatogram as a type of time series data, we expect that an AR model based feature extraction method can solve retention time shift related problems in

* Corresponding author. Tel.: +1 530 754 9004; fax: +1 530 752 4158.
E-mail address: cedavis@ucdavis.edu (C.E. Davis).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.02.010

chromatogram data. Therefore, the goal of this paper is to: (1) test if the AR model based feature extraction method can free us from time alignment for the chromatography data with time shifts, by comparing the classification results based on the AR coefficients with those based on the chromatogram wavelet coefficients and the original chromatography data, and (2) illustrate the robustness of this new feature extraction method in terms of both AR model order selection and the subsequent classification process. This is the first extensive study to address how an AR model may solve the time shift problem in chromatogram data classification. The success of this exploratory study will provide a novel and powerful feature extraction method for the development of automated instrumentation systems.

2. Experimental methods

2.1. Bacteria cell culture

We acquired four closely related bacteria cultures from American Type Culture Collection (ATCC; Bethesda, MD): *Bacillus subtilis* (ATCC #10774), *Bacillus cereus* (ATCC #13061), *Bacillus licheniformis* (ATCC #12759), and *Bacillus mycoides* (ATCC #6462). To prevent background culture conditions from introducing chemical artifacts into our signals, each species was cultured under identical conditions as described previously [13]. Briefly, the bacteria were cultured at 37 °C in liquid LB media for headspace analysis. The bacteria were innoculated in 0.5 ml LB media and grown in 10 ml borosilicate glass vials sealed with PTFE/silicone septa and aluminum screw top caps (Agilent; Palo Alto, CA) to capture the headspace gasses emanating from the proliferating bacteria cultures. The headspace gas was analyzed by GC/MS after the culture proliferated for 2 h at 37 °C. The samples were then uniformly chilled to 4 °C in order to minimize additional growth until headspace analysis, which took 1–2 h.

2.2. GC/MS headspace gas analysis

The headspace gas above proliferating bacteria cultures was analyzed using tailored gas chromatography/mass spectrometry (GC/MS) methods. The cultures were heated to 37 °C and agitated at 500 RPM to facilitate chemical release from the liquid culture and equilibrium of the chemicals with the headspace. A 30 min extraction of the chemicals in the headspace was performed using a SPME fiber with an 85 μm polyacrylate coating (Supelco, Inc.; Bellefonte, PA). The chemicals were desorbed from the fiber for 15 min into a Varian 4000 GC/MS (Varian, Inc.; Palo Alto, CA) and GC/MS analysis was performed. The GC oven profile was as follows: initial hold at 40 °C for 10 min, ramped at 2.5 °C/min, with 5 min holds at 100 °C, 125 °C, 150 °C, and 175 °C, then a 10 min hold at 200 °C. The column eluent was fed into a mass spectrometer scanning an m/z range of 35–1000. Ionization was achieved with 70 keV electron ionization. The total ion count was collected against retention time for further analysis. Totally thirty-six (nine for each species) chromatographic profiles are generated for this study.

3. Method description

3.1. Autoregressive model based feature extraction

Briefly speaking, an AR model uses p (i.e., model order) regression coefficients to characterize a time series of data. A p-order AR model can be expressed by the following equation [20–23]:

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + \varepsilon_n \tag{1}$$

where, the nth value x_n can be predicted by its previous p values: x_{n-1}, x_{n-2}, …, x_{n-p}, a_i ($i=1, \ldots, p$) are AR coefficients, ε_n is the fitting error for x_n. The goal of an AR model is to estimate the AR coefficients that can fit the original data as much as possible through an optimization process. The detailed algorithm is well described in the literature [14].

In terms of an entire classification process, the AR model functions as a feature extraction method. Generally, for a complex high dimensional classification problem, feature extraction aims to transfer data from their original high dimensional space to a lower dimensional feature space, expecting to significantly reduce the system (sample) dimensions, intensify the differences between groups, or resist possible disturbance in the original data. For example, PCA, a typical feature extraction method, transforms original data into a new coordinate system in which a number of major principal components usually are enough to reflect the differences between groups and even make the difference more prominent. Using feature extraction methods, we can observe the differences between groups from another perspective (or coordinate system) and expect the differences displayed in the new features (or coordinate system) is more observable and stronger than those in the original data system.

As a feature extraction method for time series data, AR model actually has been widely used for speaker recognition in which each speech frame is represented by a user-defined multi-dimensional AR coefficient vector [17,18]. Similar to speaker recognition, we will observe and investigate the difference between chromatogram profiles from a new perspective (feature system or coordinate system) composed of chromatogram AR coefficients and expect the group differences displayed in the new feature system can be more observable and the chromatogram profiles of different groups can be more distinguishable in the new feature system than in the original chromatogram data system.

It has been proved feasible to consider chromatogram data to be in a frequency domain, and the AR model coupled with inverse Fourier transformation (converting signal from frequency domain to time domain) has been successfully applied to de-noise chromatogram data [13] and extract feature for chromatogram classification [19]. A p-order AR model generates p complex coefficients for each inverse Fourier transformed chromatogram profile, thus the feature vector for each chromatogram sample is a $2p$ dimensional vector (p real parts and p imaginary parts).

3.2. Wavelet analysis based feature extraction

Wavelet is a powerful tool for signal analysis, examining the signal in different scales. Proposed by Mallat for digital signal processing, wavelet analysis has been substantially developed for various fields since Daubechies constructed a set of wavelet orthonormal basis functions [24,25]. Basically, wavelet transformation can represent an arbitrary function by superposing a group of wavelets which are generated from a mother wavelet through dilations and translations. A wavelet function generated at scale a and location b can be described as,

$$\Psi_{a,b}(x) = \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a}\right) \tag{2}$$

Briefly, using wavelet analysis to decompose chromatogram can be viewed as a filtering process. This filtering process is implemented by applying a bank of low-pass and high-pass filters to split the signal into high frequency (detail, D) and low frequency (approximation, A) bands [26,27]. A down-sampling step is employed in the filtering process. High frequency signals may contain more noise related information [6,28], so usually low frequency domain signals will be further decomposed using the similar transformation process and more high frequency signals will be removed. The detailed principle can be referred to in the literature [24,25]. With an n level decomposition process, the original signal (S) can be represented by:

$$S = A_n + D_n + D_{n-1} + \ldots + D_1 \tag{3}$$

Containing the higher frequency domain signals, the details are the major component used to characterize the noise. Determining the type

Download English Version:

https://daneshyari.com/en/article/1181186

Download Persian Version:

https://daneshyari.com/article/1181186

Daneshyari.com