Hindawi Publishing Corporation
Journal of Sensors
Volume 2008, Article ID 262501, 8 pages
doi:10.1155/2008/262501

Research Article

Gas Chromatography Data Classification Based on Complex Coefficients of an Autoregressive Model

Weixiang Zhao, Joshua T. Morgan, and Cristina E. Davis

Department of Mechanical and Aeronautical Engineering, University of California, One Shields Avenue, Davis, CA 95616, USA

Correspondence should be addressed to Cristina E. Davis, cedavis@ucdavis.edu

Received 22 March 2008; Accepted 12 June 2008

Recommended by Pietro Siciliano

This paper introduces autoregressive (AR) modeling as a novel method to classify outputs from gas chromatography (GC). The inverse Fourier transformation was applied to the original sensor data, and then an AR model was applied to transform data to generate AR model complex coefficients. This series of coefficients effectively contains a compressed version of all of the information in the original GC signal output. We applied this method to chromatograms resulting from proliferating bacteria species grown in culture. Three types of neural networks were used to classify the AR coefficients: backward propagating neural network (BPNN), radial basis function-principal component analysis (RBF-PCA) approach, and radial basis function-partial least squares regression (RBF-PLSR) approach. This exploratory study demonstrates the feasibility of using complex root coefficient patterns to distinguish various classes of experimental data, such as those from the different bacteria species. This cognition approach also proved to be robust and potentially useful for freeing us from time alignment of GC signals.

1. INTRODUCTION

Many modern chemical sensors produce extremely complicated signal outputs that require specialized algorithms to interpret. Gas chromatography/mass spectrometry (GC/MS) is currently considered the "gold standard" analysis system for chemical analysis, and is especially useful to analyze complex chemical samples. Due to GC/MS popularity as a chemical analysis tool, various chemometrics algorithms have been utilized to classify chromatograms from the instrument output [1–3]. We have used examples from this common analysis system to illustrate a pattern recognition approach based on autoregressive modeling, which can be broadly applied to many categories of solid-state chemical sensors.

In general, a good classification algorithm has two major parts: feature identification and extraction, and a recognition algorithm. Compared with recognition algorithms, much less work has been reported in the literature to optimize feature extraction in GC/MS data sets. Several widely used feature extracting methods for chromatogram classification include: principal component analysis (PCA) [4], genetic algorithms (GAs), and simulated annealing (SA) techniques [5, 6]. GA and SA are mainly used to select "representative" regions of MS or chromatographic data which provide the basis for classification [7–9]. These algorithms generally provide highly accurate models of chemical sensor output data, depending on the true differences between the samples. However, these feature extracting methods often require alignment in the time domain as a data preprocessing step, and this is especially true for GC/MS chromatogram signals.

There are also several other limitations of these existing feature extraction methods. For PCA, a new learning sample may result in recalculation of entire set of principal components from the system. GA and SA often require significant search times to find chromatogram classification markers due to the complexity of the data signals. Also, all of these machine learning processes are directed by mathematical optimizations and are not guaranteed to find the chemical markers with sufficient physical meaning. Because of all of these reasons, it is important to develop concise and reliable feature extracting methods for real-time signal analysis and chromatogram classification.

Recently, an autoregressive (AR) model (filter) has been successfully introduced to smooth and denoise chromatographic data from complex chemical mixtures [10]. AR

modeling has been a useful method for signal processing fields, such as audio processing [11, 12]. It is also an efficient tool for chemical sensor analysis [13–15], since a typical data format of sensor output is time series. However, to date there has been no report of the application of the AR model to extract the feature of GC data for system characterization and classification. Here we have introduced the AR model as a feature extracting method for GC/MS chromatogram data, and we have tested the feasibility and effect of this method on the data taken from biological systems.

Briefly speaking, the AR model uses n (model order) regression coefficients to represent a whole time series of data, so one significant advantage of this feature extracting method is its independence on the time dimension of the original sample. For GC/MS, variation of the experiment conditions can sometimes result in small time shifts and misalignment of the chromatogram signal. Therefore, we hypothesized that AR modeling of the chromatogram data could serve three purposes: provide noise filtering, effectively compress data while retaining important signal features, and provide feature extraction capability without signal preprocessing to align the chromatograms.

Model order is a key parameter of AR modeling, and yet in most cases, it is difficult to predict the optimum value of this parameter. Moreover, the optimal AR model order may be different between individual chromatogram samples or between classes of chromatogram data. Therefore, one important task of this study was to test if the AR model-based pattern recognition method is robust with respect to AR model order. Three types of neural networks were employed in this classification study: back-propagation neural networks (BPNNs), radial basis function—principal component analysis (RBF-PCA), and radial basis function—partial least squares regression (RBF-PLSR). These learning algorithms were used to classify the resulting complex coefficients from the AR modeling. Covering two most widely used neural networks and their modifications, these three types of neural networks proved the wide suitability of the chromatogram feature composed of AR model coefficients to various classifiers.

We devised a simple experimental system to test the AR modeling and pattern recognition algorithms using the headspace samples of four bacteria species cultured in enclosed vials. It is generally well accepted that the complex chemical gases that are produced from proliferating pure bacteria cultures are unique to a bacteria species [9, 16], and by culturing the bacteria in vials we captured these gasses for chemical analysis. In this exploratory study, we expect to reach the following goals by analyzing chromatogram data for four different species of bacteria: (i) to verify the feasibility of AR model-based pattern recognition strategy for chromatogram data, (ii) to test the robustness of this feature extracting strategy, and (iii) to compare the effects of various learning algorithms based on this feature extracting strategy. This is the first extensive study to apply AR modeling to extract the feature of chromatogram data for classification. The success of this exploratory study provides a novel feature extracting method with the following significant advantages: no requirement of chromatogram signal time alignment, potential real-time classification in automated chromatography instrumentation systems, and the ability to effectively deal with background signals in chromatography outputs.

2. EXPERIMENTAL METHODS

2.1. Bacteria cell culture

Four closely related bacteria species were acquired from American Type Culture Collection (ATCC, Bethesda, MD, USA): *Bacillus subtilis* (ATCC no. 10774), *Bacillus cereus* (ATCC no. 13061), *Bacillus licheniformis* (ATCC no. 12759), and *Bacillus mycoides* (ATCC no. 6462). To prevent background culture conditions from introducing chemical artifacts into our signals, each species was cultured under identical conditions as described previously [10]. Briefly, the bacteria were cultured at 37°C on standard LB agar plates, and colonies were selected to proliferate in liquid LB media for headspace analysis. The bacteria were seeded into 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, USA) to capture the headspace gasses emanating from the proliferating cultures. The headspace gas was analyzed after the culture proliferated for 2 hours at 37°C. The samples were then uniformly chilled to 4°C in order to minimize additional bacterial growth until GC/MS analysis of the headspace was complete.

2.2. GC/MS headspace gas analysis

The headspace gas above the proliferating cultures was analyzed sequentially using standard 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. The chemicals in the headspace were then extracted for 30 minutes using an SPME fiber with an 85 μm polyacrylate coating (Supelco, Inc., Bellefonte, PA, USA). The fiber was desorbed for 15 minutes into a Varian 4000 GC/MS (Varian, Inc., Palo Alto, CA, USA) for analysis. The GC oven was initially held at 40°C for 10 minutes, then ramped at 2.5°C/m, with 5-minute holds at 100°C, 125°C, 150°C, and 175°C. The cycle concluded with a 10-minute 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.

3. DATA ANALYSIS

3.1. Autoregressive modeling of GC/MS chromatograms

The AR model is an all-pole model (filter). A p-order AR model can be expressed by the following transfer function:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}}. \quad (1)$$

Thus, the nth value ($\mathbf{x}(n)$) can be predicted by its previous p values: $\mathbf{x}(n-1), \mathbf{x}(n-2), \ldots, \mathbf{x}(n-p)$. A p-order AR model is equivalent to a p-order linear prediction model below:

$$\hat{\boldsymbol{x}}(n) = -\sum_{i=1}^{p} a_i \mathbf{x}(n-i), \tag{2}$$

where a_i $(i = 1, \ldots, p)$ are AR coefficients. 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.

Previously, we have shown that to effect real-time modeling of a chromatogram, it is reasonable to consider the chromatography data to be in the frequency domain [10], and so the inverse Fourier transformation was applied to each chromatogram before the autoregressive modeling of the signal. A p-order AR model generates p complex roots as the AR coefficients of each inverse Fourier transformed chromatogram profile. For chromatographic signal enhancement and noise filtering, these p complex roots would normally be used to reconstruct the chromatogram data after filtering, but in this study we are applying pattern recognition approaches to these feature vectors to test if we can provide highly accurate classification of the chromatograms using the complex roots.

3.2. Back-propagation neural network (BPNN)

The BPNN is an effective nonlinear mapping tool and has been widely used for pattern recognition. A typical BPNN is composed of three layers: input, hidden, and output layers. BPNN training can be summarized as an iteration process during which the error between the predicted outputs and the stipulated outputs for each training sample will be back-propagated by a gradient descendent algorithm to adjust the weights of each layer until a convergence criterion is reached. This is usually based on reducing the total error of the training data set.

3.3. Radial basis function-principal component analysis (RBF-PCA)

A significant advantage of RBFN over BPNN is the ability to avoid long-duration training times, but the modeling effect of the RBFN largely depends on the proper determination of radial basis vectors. In order to overcome this drawback, we have integrated the RBFN with multivariate statistical analysis. RBF-PCA is a successful integration approach and has been applied previously in various cheminformatics and bioinformatics fields [17, 18].

The basic concept of RBFN is a radial function-based interpolation problem. Given the sample set of $\mathbf{x}_i$ and the corresponding class index y_i, RBFN aims to seek an interpolation function $f(\mathbf{x}_i)$ to build a map from $\mathbf{x}_i$ to y_i. The RBFN is also composed of three layers. The hidden layer performs a nonlinear transformation to transform the input space into a high dimensional transitional space by radial basis functions. The standard radial basis transformation function is a Gaussian kernel function:

$$\varphi(\mathbf{x}_i, \mathbf{c}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{c}_j||}{\sigma_j^2}\right), \tag{3}$$

where $\mathbf{c}_j$ is the radial basis vector of the jth hidden node, and σ_j^2 is the Gaussian width of the jth hidden node.

The output layer produces linear weighting summation of all the hidden node outputs:

$$f(\mathbf{x}_i) = \sum_{j=1}^{P} \lambda_j \varphi(\mathbf{x}_i, \mathbf{c}_j), \tag{4}$$

where λ_j is the weight connecting the output node and the jth hidden node, and P is the number of hidden nodes. Ordinary least square regression can be used to calculate the weight λ_j.

It can potentially be difficult to determine the proper number of radial basis centers and to estimate these center vectors. An alternative idea is to use all the training samples as radial basis vectors. Supposing there are K training samples, the hidden layer can generate a K-by-K transition matrix $\mathbf{M}_{K\times K}$ each element of which is the output of (3). Thus, all the information of the training set is fully utilized. However, this may result in over-fitting in the regression step, so PCA can be used to extract latent variables from this transitional matrix for regression. The ordinary least square regression of (4) turns out to be the principal component regression (PCR). The detailed processes of PCA and PCR have been documented as potentially valuable learning algorithms in the literature [19]. Clearly, RBF-PCA statistically resolves the problem of determining proper radial basis vectors, avoiding possible local optima that often occur in conventional radial basis vector determination strategies like the K-means algorithm.

3.4. Radial basis function-partial least squares regression (RBF-PLSR)

The PLSR is another generalization of ordinary least squares regression. Compared with PCR, the PLSR usually yields a relatively high modeling accuracy as extracting PLS components not only relies on the independent variable information but also employs the dependent variable information. Therefore, PLSR is another competitive tool for the regression of the transitional RBF response matrix $\mathbf{M}_{K\times K}$ against the stipulated outputs.

One standard PLS algorithm is a nonlinear iterative partial least squares [20]. In this algorithm, we let $\mathbf{X}$ and $\mathbf{Y}$ be independent and dependent data matrices. The entire algorithm consists of two loops. The inner loop is for extracting PLS component $\mathbf{t}$ of $\mathbf{X}$ and the corresponding information vector $\mathbf{s}$ of $\mathbf{Y}$. It has been shown [20] that $\mathbf{t}$ and $\mathbf{s}$ are the components in the X and Y space that have maximal covariance. In each inner loop, the PLS algorithm only yields one pair of components from $\mathbf{X}$ and $\mathbf{Y}$. The outer loop then calculates the residual matrices of $\mathbf{X}$ and $\mathbf{Y}$ by subtracting the extracted components from $\mathbf{X}$ and $\mathbf{Y}$. The whole iteration

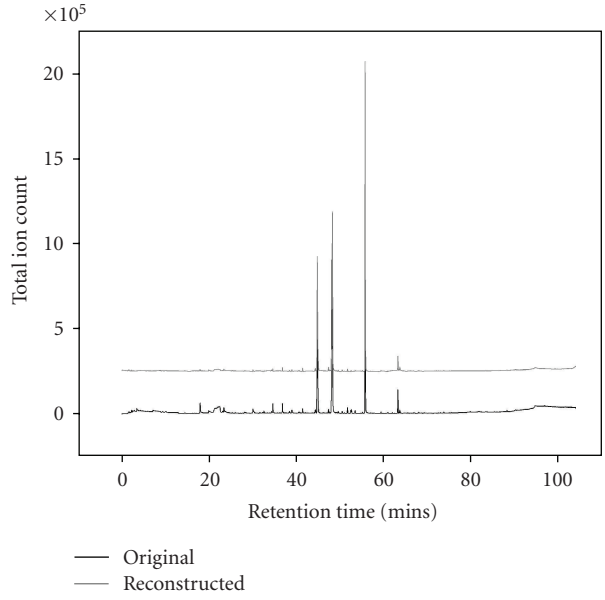

Figure 1: Original versus reconstructed chromatograms based on the 60-order AR filter for a *B. cereus* sample.

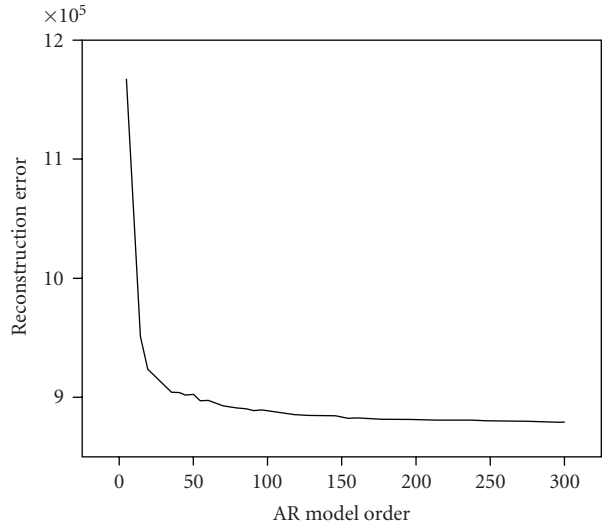

Figure 2: Reconstructed error versus AR model order for the chromatogram data of a *B. cereus* sample.

process continues until a stop criterion is reached or the residual matrix of **X** becomes zero.

In summary, the PLSR generates a transforming matrix **W** to transform **X** to **T** (composed of PLS components **t**). Assuming **Q** is the regression coefficients of T against **Y**, $\mathbf{B} = \mathbf{WQ}$ is the direct regression coefficients of **X** against **Y** and can be used for future prediction.

The RBF-PLSR approach applies PLSR to relate the transitional RBF response matrix $\mathbf{M}_{K\times K}$ to the stipulated outputs. This method was first proposed in 1996 [21], and has been applied to various chemometrics and chemical process problems [22, 23]. All the data analyses and modeling in this study were performed using MATLAB (The Mathworks, Inc., Natick, MA) version 7.3.0.267. Fourier and inverse Fourier transforms were accomplished using the FFT and IFFT algorithms included with the software.

4. RESULTS

4.1. Determination of AR model order

A proper order for an AR model should be able to yield a good data fitting effect while retaining a high data compression ratio. Generally, a plot of the fitting error of data series versus the model order may show the turning point from which the curve will even out. Below is an example to show this decreasing trend by using the chromatogram data from a *Bacillus cereus* chromatogram signal.

Figure 1 is a simple illustration of the original chromatogram of a *Bacillus cereus* headspace sample versus the reconstructed chromatogram of a 60-order AR filter for this sample. As discussed before, the chromatogram data can be considered to be in the frequency domain for the purposes of this experiment [10], so an inverse Fourier transform is needed before AR analysis. In reverse, the reconstructed data series based on AR coefficients need Fourier transform to go back to the original domain. Figure 2 shows the decreasing trend of the reconstruction error against the AR model order. The reconstruction error of an AR model in this study was defined as the distance between the original chromatogram vector $\mathbf{cm}_{\text{ori}}$ and the reconstructed one $\mathbf{cm}_{\text{rec}}$, that is, $\|\mathbf{cm}_{\text{ori}} - \mathbf{cm}_{\text{rec}}\|$.

We see from Figure 2 that the optimal AR model order (i.e., turning point) should lie in the neighboring range of 50. However, it is challenging to give an exact value to the optimal order, and may not be feasible for all chromatogram samples to have the same optimal order value. Therefore, it is very important for the AR model-based recognition strategy to be robust in terms of AR model order. In this study, three model orders: 20, 40, and 60 (all in the neighbor range of 50) were used for AR analysis, and their AR complex coefficient vectors were used for bacteria classification.

4.2. Preliminary investigation of AR coefficients using PCA

A p-order AR model generates p complex coefficients for each inverse Fourier transformed chromatogram. Thus, the feature vector of this system is composed of $2p$ elements: p real part coefficients plus p imaginary part coefficients. For illustration, Figure 3 shows the distribution of some representative principal components of the AR complex coefficients obtained by a 20-order AR model for each bacteria headspace sample. The 1st and 2nd principal components clearly separate the *B. subtilis* samples from the others. When the 2nd and 4th principal components area also included, it provides an even better separation of the data. The mild overlaps among classes may indicate possible nonlinearity among them. These results preliminarily showed that it is feasible to use the AR model complex coefficients for the classification of chromatography outputs from complex mixtures of chemicals.

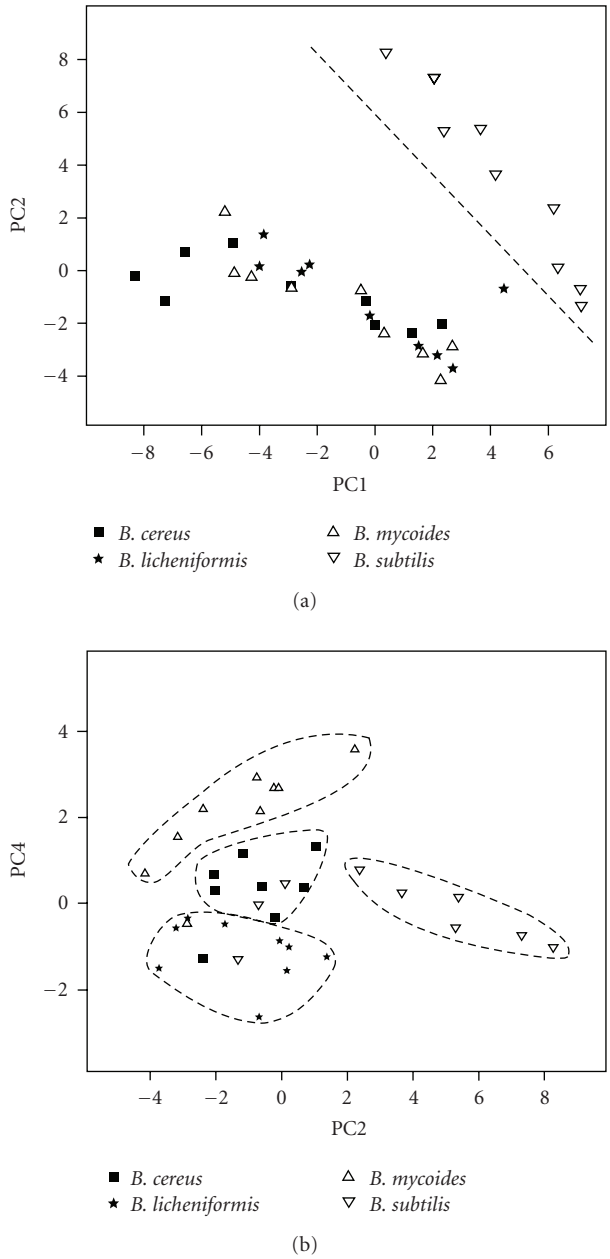

Figure 3: Distribution of some representative principal components of the AR complex coefficients obtained by a 20-order AR model for each bacteria headspace sample.

4.3. Classification based on AR coefficients

In this section, three types of neural networks, BPNN, RBF-PCA, and RBF-PLSR were applied to test their classification effects on the AR coefficients of the bacteria headspace chromatogram data. For this four-class problem, the stipulated output of each sample was designed as a four-dimensional vector in which the element corresponding to the class of this sample was set to be 1 and the other three elements to be 0. For an unknown sample, the element which is the closest to 1 denotes the class of this sample. In this exploratory study, we did not expect to obtain the best network parameters for each type of neural networks but aimed to test the feasibility of the AR coefficients to chromatogram classification by using a number of classifiers.

Considering the limited number of samples ($n = 9$) for each class, the leave-one-out strategy was used to verify the classification effect. Each time 8 samples of each class were used for training and the remained sample of each class was used for testing. The whole training set for each time was composed of 32 ($8*4$) samples while the testing set composed of 4 ($1*4$) samples. This training-testing process was repeated nine times to cover the whole sample set. Therefore, the most ideal classification result for the testing samples of each class is 9/9.

4.3.1. BPNN

The BPNN was first used for this classification problem. The direct inputs for the BPNN were the principal components of the AR complex coefficients (composed of $2p$ elements) of the training samples. Therefore, the testing samples needed to be transformed by the PCA loading matrix of the training samples before being input to the BPNN. One criterion to determine the proper number of principal components is the ratio of the sum of the accumulated eigen values to that of the all eigen values, called PC ratio. In this experiment, we set this to be 99%.

In this experiment, the number of the nodes in the hidden layer of BPNN was set to be 3, 8, and 16 respectively. For each case, the training-testing process was repeated three times with different initial weights. The average accuracy of the three trials was used as the final nominal accuracy of this case. The classification results of BPNN are listed in Table 1. The decimal values mean the average classification rate of three trials.

The AR coefficients generated by the 20-, 40-, and 60-order AR models with different orders all provided good classification results, illustrating the robustness of the AR model-based recognition strategy in terms of the model order. Therefore, it is not necessary to search for a so-called "best" model order because a relatively wide range around the turning point is shown to be equivalently feasible for classification.

4.3.2. RBF-PCA

In each time for the RBF-PCA approach, the AR coefficients of the 32 training samples were used as the radial basis vectors, so the hidden layer generated a 32-by-32 transitional matrix. Then, PCA was applied to create the relationship between this transitional matrix and the stipulated outputs. The PC ratio for this regression model was fixed on 99.5%. As shown in (3), another key parameter of this approach is the Gaussian width for each radial basis. One of the ways to determine this parameter is to set the Gaussian width of each radial basis to be the same value [21]. The results of various width values are listed in Table 2.

The results of the three different order AR models all show the positive effect of this cognition strategy. A proper Gaussian width seems to be an important factor for this

Table 1: BPNN classification effects on the AR coefficients of different order AR models.

AR order	Hidden node number	Accuracy (*B. cereus*)	Accuracy (*B. licheniformis*)	Accuracy (*B. mycoides*)	Accuracy (*B. subtilis*)
20	3	8/9	7.7/9	7.7/9	8.3/9
	8	8.3/9	8/9	7.7/9	9/9
	16	8/9	8/9	8.3/9	9/9
40	3	7.7/9	8/9	7.7/9	9/9
	8	8/9	7.7/9	8/9	9/9
	16	8/9	8/9	7.7/9	9/9
60	3	8/9	8/9	6.3/9	8.3/9
	8	8/9	8/9	7.7/9	9/9
	16	8.3/9	8/9	7.7/9	9/9

Table 2: RBF-PCA classification effects on the AR coefficients of different order AR models.

AR order	Gaussian width	Accuracy (*B. cereus*)	Accuracy (*B. licheniformis*)	Accuracy (*B. mycoides*)	Accuracy (*B. subtilis*)
20	0.005	8/9	8/9	7/9	9/9
	0.01	8/9	8/9	8/9	9/9
	0.02	8/9	8/9	7/9	9/9
40	0.02	8/9	8/9	8/9	9/9
	0.025	9/9	8/9	8/9	9/9
	0.03	8/9	8/9	7/9	9/9
60	0.025	9/9	8/9	7/9	9/9
	0.03	9/9	8/9	8/9	9/9
	0.04	8/9	8/9	8/9	9/9

strategy, as each case shows an increase and then a decrease of the classification accuracy along with the increase of Gaussian width.

4.3.3. RBF-PLSR

In this approach, PLSR was used to build a map from the 32-by-32 transitional RBF response matrix to the stipulated outputs. The correlation between the resolved PLS component **t** from **X** and information vector **s** from **Y** will decrease while the valuable information being extracted from **X** and **Y**. In this study, the criterion to determine the proper number of PLS components was the ratio of the correlation coefficient of the new resolved **t** (from **X**) and **s** (from **Y**) to the sum of the correlation coefficients of all resolved **t**-**s** pairs. Here, this ratio was set to be 0.0001. Similar to the RBF-PCA approach, the Gaussian width for each initial radial basis was set to be the same value. The results of various Gaussian widths are listed in Table 3. The results in this experiment showed the good classification effect of the RBF-PLSR strategy while also indicating the importance of a proper Gaussian width.

All of the three types of neural networks yielded good classification results in this study. The best accuracy (8/9, 9/9, 9/9, 9/9) was obtained by applying the RBF-PLSR approach with the Gaussian width being 0.01 to the AR coefficients of the 20-order AR models. Meanwhile, in average RBF-PCA showed a slightly better effect than BPNN.

5. DISCUSSION

Three machine learning methods were employed and introduced to verify the feasibility of the AR coefficients for the classification of the chromatogram samples. All of three classifiers yielded good classification results, which demonstrates the feasibility of the proposed classification strategy for chromatograms. AR models were directly applied to the chromatograms, so this proposed AR model-based recognition strategy potentially frees us from the time alignment procedure which is often required in conventional chromatogram classification methods.

It is challenging to decide the "best" order for an AR model, so the proposed AR model-based recognition strategy must be robust in terms of the model order. That is to say, this recognition strategy should be able to yield equivalently good classification effects within a relatively wide range of the AR model order. The results of the AR coefficients obtained by three different order AR models demonstrated the robustness of this recognition strategy, which releases us from searching for a so-called "best" AR model order.

With respect to classifier accuracy, two statistical neural networks yielded slightly higher accuracy than the BPNN approach. This exploratory study was not designed to obtain the best network parameters for each type of neural networks, so the results may not adequately suggest which type of neural networks is superior to the others. Having shown feasibility and robustness of the AR complex

Table 3: RBF-PLSR classification effects on the AR coefficients of different order AR models.

AR order	Gaussian width	Accuracy (*B. cereus*)	Accuracy (*B. licheniformis*)	Accuracy (*B. mycoides*)	Accuracy (*B. subtilis*)
20	0.005	5/9	8/9	8/9	9/9
	0.01	8/9	9/9	9/9	9/9
	0.015	7/9	8/9	7/9	9/9
40	0.005	7/9	8/9	7/9	9/9
	0.01	7/9	8/9	8/9	9/9
	0.015	5/9	8/9	9/9	9/9
60	0.01	7/9	8/9	8/9	9/9
	0.015	7/9	9/9	8/9	9/9
	0.02	6/9	8/9	7/9	9/9

coefficients on the chromatogram classification, the study will aim to increase the classification accuracy by optimizing classifier parameters. Given the ratios to determine principal components and PLS components, the modeling accuracy of RBF-PCA and RBF-PLSR largely depend on the Gaussian width of each training sample.

6. CONCLUSIONS

This exploratory study demonstrates the feasibility of the AR coefficients for the classification of the chromatograms. The experiments on the AR coefficients obtained by three different order AR models illustrate the robustness of this strategy in terms of AR model order. Three types of neural networks: BPNN, RBF-PCA, and RBF-PLSR all showed good classification effects, indicating the wide suitability of this novel chromatogram feature extracting method to various classifiers. This study provides a novel recognition method for the classification of complex chemical samples and other chromatogram represented samples, which is able to free us from possible time alignment for chramatography output signals. This may potentially allow us to classify chromatogram or other chemical sensor outputs in real-time, and reduce the effects of background signals on this category of chemical sensors.

ACKNOWLEDGMENTS

This work was supported in part by several research Grants to C. Davis: the Army Research Office contract W911NF-06-1-0272, the Lawrence Livermore National Laboratory B563119, and by a generous gift from the American Petroleum Institute. This publication was made possible by Grant no. UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Information on NCRR is available at http://www.ncrr.nih.gov. Information on Re-engineering the Clinical Research Enterprise can be obtained from http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

REFERENCES

[1] R. Bucci, A. D. Magrí, A. L. Magrí, D. Marini, and F. Marini, "Chemical authentication of extra virgin olive oil varieties by supervised chemometric procedures," *Journal of Agricultural and Food Chemistry*, vol. 50, no. 3, pp. 413–418, 2002.

[2] G. R. Magelssen and J. W. Elling, "Chromatography pattern recognition of Aroclors using iterative probabilistic neural networks," *Journal of Chromatography A*, vol. 775, no. 1-2, pp. 231–242, 1997.

[3] K. M. Pierce, J. C. Hoggard, J. L. Hope, et al., "Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts," *Analytical Chemistry*, vol. 78, no. 14, pp. 5068–5075, 2006.

[4] M. D. Krebs, R. D. Tingley, J. E. Zeskind, M. E. Holmboe, J.-M. Kang, and C. E. Davis, "Alignment of gas chromatography-mass spectrometry data by landmark selection from complex chemical mixtures," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 74–81, 2006.

[5] D. B. Hibbert, "Genetic algorithms in chemistry," *Chemometrics and Intelligent Laboratory Systems*, vol. 19, no. 3, pp. 277–293, 1993.

[6] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[7] M. D. Krebs, B. Mansfield, P. Yip, et al., "Novel technology for rapid species-specific detection of *Bacillus* spores," *Biomolecular Engineering*, vol. 23, no. 2-3, pp. 119–127, 2006.

[8] E. Llobet, O. Gualdrón, M. Vinaixa, et al., "Efficient feature selection for mass spectrometry based electronic nose applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 2, pp. 253–261, 2007.

[9] M. Shnayderman, B. Mansfield, P. Yip, et al., "Species-specific bacteria identification using differential mobility spectrometry and bioinformatics pattern recognition," *Analytical Chemistry*, vol. 77, no. 18, pp. 5930–5937, 2005.

[10] M. D. Krebs, R. D. Tingley, J. E. Zeskind, J.-M. Kang, M. E. Holmboe, and C. E. Davis, "Autoregressive modeling of analytical sensor data can yield classifiers in the predictor coefficient parameter space," *Bioinformatics*, vol. 21, no. 8, pp. 1325–1331, 2005.

[11] M. Ž Marković, M. M. Milosavljević, and B. D. Kovačević, "Quadratic classifier with sliding training data set in robust

recursive AR speech analysis," *Speech Communication*, vol. 37, no. 3-4, pp. 283–302, 2002.

[12] V. Šmídl and A. Quinn, "Mixture-based extension of the AR model and its recursive Bayesian identification," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3530–3542, 2005.

[13] M. Holmberg, F. A. M. Davide, C. Di Natale, A. D'Amico, F. Winquist, and I. Lundström, "Drift counteraction in odour recognition applications: lifelong calibration method," *Sensors and Actuators B*, vol. 42, no. 3, pp. 185–194, 1997.

[14] E. L. Hines, E. Llobet, and J. W. Gardner, "Electronic noses: a review of signal processing techniques," *IEE Proceedings: Circuits, Devices and Systems*, vol. 146, no. 6, pp. 297–310, 1999.

[15] C. Di Natale, S. Marco, F. Davide, and A. D'Amico, "Sensor-array calibration time reduction by dynamic modelling," *Sensors and Actuators B*, vol. 25, no. 1–3, pp. 578–583, 1995.

[16] V. Rossi, R. Talon, and J.-L. Berdagué, "Rapid discrimination of *Micrococcaceae* species using semiconductor gas sensors," *Journal of Microbiological Methods*, vol. 24, no. 2, pp. 183–190, 1995.

[17] N. Pochet, F. De Smet, J. A. K. Suykens, and B. L. R. De Moor, "Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction," *Bioinformatics*, vol. 20, no. 17, pp. 3185–3195, 2004.

[18] Y. Yamanishi, J.-P. Vert, and M. Kanehisa, "Protein network inference from multiple genomic data: a supervised approach," *Bioinformatics*, vol. 20, supplement 1, pp. i363–i370, 2004.

[19] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[20] A. Höskuldsson, "PLS regression methods," *Jounral of Chemometrics*, vol. 2, no. 3, pp. 211–228, 1988.

[21] B. Walczak and D. L. Massart, "The radial basis functions—partial least squares approach as a flexible non-linear regression technique," *Analytica Chimica Acta*, vol. 331, no. 3, pp. 177–185, 1996.

[22] W. Zhao, D. Chen, and S. Hu, "Detection of outlier and a robust BP algorithm against outlier," *Computers & Chemical Engineering*, vol. 28, no. 8, pp. 1403–1408, 2004.

[23] W. Zhao, P. K. Hopke, X. Qin, and K. A. Prather, "Predicting bulk ambient aerosol compositions from ATOFMS data with ART-2a and multivariate analysis," *Analytica Chimica Acta*, vol. 549, no. 1-2, pp. 179–187, 2005.