Journal of Chromatography, 384 (1987) 145–152
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROMSYMP. 1031

USE OF PATTERN-RECOGNITION TECHNIQUES TO ANALYZE CHROMATOGRAPHIC DATA

MOSES E. COHEN*
Department of Mathematics, California State University, Fresno, CA 93740 (U.S.A.)
DONNA L. HUDSON
Section on Medical Information Science, University of California, San Francisco, CA (U.S.A.)
LEWIS T. MANN
Veterans Administration Medical Center and University of California, San Francisco, CA (U.S.A.)
JACK VAN DEN BOGAERDE
Veterans Administration Medical Center, San Francisco, CA (U.S.A.)
and
NORMAN GITLIN
Veterans Administration Medical Center and University of California, San Francisco, CA (U.S.A.)

SUMMARY

A pattern-recognition technique has been established using a new class or orthogonal polynomials, developed by Cohen. The method is based on a supervised learning approach, and allows classification of data into two or more categories. In this paper, the usefulness of the method in the analysis of chromatographic data is illustrated by its application to the diagnosis of bacterial infection of patients with liver disorders by use of chromatograms obtained from ascitic fluid withdrawn from the patients.

INTRODUCTION

Decision-making algorithms, often denoted expert systems, have been designed which utilize different approaches to automated reasoning[1,2]. These include statistical discrimination techniques[3–6], and decision analysis utilizing methods of pattern recognition[7–11].

Work in pattern recognition of chromatographic data has been done by Albano *et al.*[12], and Wold and Johansson [13], using the SIMCA method. The SIMCA method involves fitting a hyperplane to each class. The method has been applied in various areas, including obtaining chromatographic profiles of human brain tissues[13], classification of fungi[14], and classification of human cancer cells[15]. Unlike the SIMCA method, the techniques described in this paper do not rely on the restrictive properties of hyperplanes in n-dimensional space, but rather permit the fitting of non-linear surfaces to each category using non-statistical methods. A pattern-recognition technique has been developed based on a new class of Cohen orthogonal polynomials which can be used to classify diverse types of data. The method utilizes a supervised learning approach, consisting of a number of phases: feature extraction, iterative development of a separating hypersurface, and testing of the resulting hypersurface for accuracy. The method is applicable to the analysis of chromatographic data where combinations of either the occurrence of a peak at certain retention times or the magnitude of the peaks is of importance in classifying the chromatogram into two or more categories. The method will be illustrated as applied to the analysis of chromatograms obtained from the analysis of ascitic fluid of patients with liver disorders.

In the study described here, chromatographic analysis is used to analyze ascitic fluid, a substance which accumulates in the abdomen of patients with liver disorders. Under some circumstances, patients sometimes develop a life-threatening disorder, known as spontaneous bacterial peritonitis (SBP)[17]. The objective of the procedure described here is to use the chromatographic data to determine at an early stage if a patient has SBP.

In the first phases —feature extraction— parameters are identified which may be useful in separating data into categories. In the case of the ascitic fluid, the objective was to divide the samples into two catagories: those from patients with SBP, and those from patients without SBP, by analyzing the pattern of organic acids which occurred in the patients with liver disorders for which prompt diagnosis is crucial. Peaks were identified by establishing time intervals, denoted buckets, which consisted of the most likely occurrence of a peak ± a fixed interval adjustment. There were 47 possible peaks present, which were treated both qualitatively as the presence or absence of a peak at a designated time, and quantitatively as the area under each peak. Thus, there were 94 possible features.

Available data of known classification were divided into two groups: a training set and a test set. The training set was then used, in conjunction with the orthogonal function, to obtain weighting factors for each feature. The data set contained 46 samples, of which 26 were positive for SBP. A training set of 12 was selected, with 6 positives. A separating hypersurface was then obtained.

METHODOLOGY

A sample of ascites fluid was withdrawn by paracentesis from the abdomen of a patient, usually in the course of his treatment. The ascitic fluid was routinely cultured to ascertain the presence of infection, and if positive, the identity of the pathogen. A portion of the ascites fluid was frozen for future study and served as the source of the samples analyzed in this study.

Prior to attempting to determine the organic acid profile of a particular sample using high-performance liquid chromatography (HPLC), it was necessary to isolate the acids from other potentially interfering constituents typically present in ascitic fluid. The following extraction procedure was developed for this project:

To 1.0 ml of sample was added 0.2 ml of saturated K_2HPO_4 to achieve a slightly basic pH and the resulting solution thoroughly mixed with 3 ml of an organic solvent consisting of *tert.*-butyl ether–isopropyl alcohol (95:5). Following centrifugation, the overlying organic layer containing many of the interfering constituents was suctioned off and discarded. The pH of the aqueous phase was then lowered by the addition of 0.2 ml of 9 *M* sulfuric acid and the undissociated organic acids extracted by thorough agitation into 5.0 ml of the *tert.*-butyl ether–isopropyl alcohol solvent. Following centrifugation to enhance complete separation of the two phases, 4.5 ml of the overlying organic layer was pipetted off and retained; the remainder, as well as the aqueous phase, were discarded; 2.0 ml of 0.1 *M* sodium hydroxide was added to the organic phase and the organic acids back-extracted into the aqueous phase via thorough mixing. Following centrifugation, the overlying organic acids in the aqueous phase, the water was removed via freeze-drying under vacuum. Finally, an aqueous solution of the organic acids was reconstituted by dissolving the freeze-dried residue in 0.5 ml of water and 0.1 ml of 1 *M* sulfuric acid. A water blank was simultaneously carried through the extraction procedure so that any artifacts introduced by the extraction could subsequently be identified and eliminated from consideration.

Separation of the organic acids in the extracted and reconstituted sample of ascites fluid was performed with HPLC, primarily via the ion-exclusion mode, on a 220 × 4.6 mm Brownlee Labs. column, packed with Polypore H, a 10-μm, macroporous styrene–divinylbenzene resin. The column was maintained at a uniform temperature of 39°C. The mobile phase or eluent consisted of 0.01 *M* sulfuric acid, prepared from deionized water, deaerated and filtered through a 0.45-μm filter before use. Eluent was delivered to the column by a Consta Metric III metering pump from Laboratory Data Control at a rate of 0.6 ml/min. Samples were introduced with a 50-μl syringe, via a Rheodyne injection valve, into a 20-μl loop in the eluent flow. Effluent from the HPLC column passed into the flow cell of a Laboratory Data Control Spectro Monitor III spectrophotometer, where the organic acids were detected in the UV at the carboxyl absorption band of 210 nm. Detector output was recorded on a Spectra-Physics SP4270 chromatography integrator, which subsequently analyzed the individual organic acid peaks for retention time and area. The water blank was also subjected to HPLC and the areas of the peaks that appeared were subtracted from the areas of the corresponding peaks in the sample chromatogram.

It was presumed that an organic acid could be identified by the retention time of its chromatographic peak. Individual standard solutions of known concentrations of biologically important short-chained mono-, di- and tri-protic organic acids were prepared and the retention time of each determined by HPLC under the aforementioned experimental conditions. Retention times for the following organic acids were identified in this manner: oxalic, oxaloacetic, citric, isocitric, ketoglutaric, tartaric, pyruvic, ascorbic, malic, succinic, lactic, 3-hydroxybutyric, formic, glutaric, acetic, fumaric, propionic, isobutyric, *n*-butyric, 3-methylbutyric, 2-methylbutyric, valeric and caproic acids.

Ascitic fluid was analyzed by the above method. In addition, some of the fluid was cultured for the presence of bacteria for 72 h. Chromatographic analysis and pattern recognition took approximately 3 h.

Although the automatically integrated chromatogram contained continuous analog data, two aspects of these time series data are of significance: (1) the area under each peak; (2) the elution time of each peak. In each chromatogram as many as 47 separate areas and elution times may be present. Thus there are potentially 94 distinct elements which are of significance for analysis of the chromatogram.

Initially, the area under each peak and the elution time of that peak were manually entered for each chromatogram into a data base program on the VAX 11/750 computer. This method is time-consuming and error-prone. Hence, an automated system was established. In this system, an analog to digital converter was connected to the chromatograph, which then sampled the analog data at fixed time intervals, and transmitted all data to an IBM PC microcomputer. Each chromatogram required approximately 1 h to run. The data were then compressed on the IBM PC, and were transmitted via a hardwire connection to the VAX 11/750, where numerical integration was performed prior to the transfer of the data to the data base described above. The correct classification was then entered for each case from results of the bacteriological examination.

Once the data base on the VAX was completed by one of the above methods, pattern classification was begun. The objective in this case was to obtain a two-category classification: Class 1, patients with SBP; Class 2, patients without SBP. Only a subset of the 94 mentioned above will prove useful in obtaining the correct classification.

The pattern recognition algorithm utilized a non-statistical supervised learning approach. The data obtained from the above procedure were divided into two groups, the training set and the test set; the training set was used to obtain a separating hypersurface. An iterative classification method was used which incorporated the potential function approach to the generation of decision surfaces. The potential function is defined by:

$$\mathrm{P}(\bar{x},\bar{y}) = \sum_{i=1}^{\infty} \lambda_i^2 \, \mathrm{f}_i(\bar{x}) \, \mathrm{f}_i(\bar{y}) \tag{1}$$

where $\mathrm{f}_i(\bar{x})$, $\mathrm{f}_i(\bar{y})$, $i = 1,2,\ldots$ are orthonormal functions, $\bar{x},\bar{y}$ are n-dimensional vectors, and λ_i, $i = 1,2,\ldots$ are real numbers. The decision surface is adjusted iteratively according to:

$$\mathrm{D}_{k+1}(\bar{x}) = \mathrm{D}_k(\bar{x}) + r_{k+1}\, \mathrm{P}(\bar{x},\bar{x}_{k+1}) \tag{2}$$

where $\mathrm{P}(\bar{x},\bar{x}_{k+1})$ is the potential function from eqn. 1, and

$$r_{k+1} = \begin{array}{l} 1 \text{ for } \bar{x}_{k+1} \in \omega_1 \text{ and } \mathrm{P}_k(\bar{x}_{k+1}) < 0 \\ -1 \text{ for } \bar{x}_{k+1} \in \omega_2 \text{ and } \mathrm{P}_k(\bar{x}_{k+1}) > 0 \\ 0 \text{ otherwise} \end{array} \tag{3}$$

where ω_1 represents class i, $i = 1,2$. $\mathrm{P}_0(\bar{x})$ is assumed to be zero, and

$$\mathrm{P}_1(\bar{x}) = \begin{array}{l} \mathrm{P}_0(\bar{x}) + \mathrm{P}(\bar{x},\bar{x}_1) = \ \ \mathrm{P}(\bar{x},\bar{x}_1) \text{ if } \bar{x}_1 \in \omega_1 \\ \\ \mathrm{P}_0(\bar{x}) - \mathrm{P}(\bar{x},\bar{x}_1) = -\mathrm{P}(\bar{x},\bar{x}_1) \text{ if } \bar{x}_1 \in \omega_2 \end{array} \tag{4}$$

$\mathrm{D}(\bar{x}) = 0$ at the completion of the iterative procedure provides the separating hypersurface in the two category case.A new class of orthogonal Cohen functions was used as the potential functions. Two special cases of this general class are considered: $f_n(x)$ and $g_n^{\lambda,\xi}(x)$, where λ and ξ are arbitrary real numbers. These functions are defined by their recurrence relations:

$$\begin{aligned}&f_0(x) = 1 \qquad f_1(x) = 2 - 3x\\&(1 - 2^n)f_n(x) - x(1 + 2^n)f_{n-1}(x^2)\\&+ (1 + 2^{n-1})f_{n-1}(x) + x(2^{n-1} - 1)f_{n-2}(x^2) = 0 \quad n \geqslant 2\end{aligned} \tag{5}$$

$$\begin{aligned}&g_0^{\lambda,\xi}(x) = 1 \qquad g_1^{\lambda,\xi}(x) = \lambda - (\lambda + 1)x\\&\left[\frac{\xi^n - 1}{\xi - 1}\right] g_n^{\lambda,\xi}(x) + x\left[\lambda + \frac{\xi^{n-1}}{\xi - 1}\right] g_{n-1}^{\frac{\lambda+2}{\xi},\xi}(x^\xi) -\\&\left[\lambda + \frac{\xi^{n-1} - 1}{\xi - 1}\right] g_{n-1}^{\lambda,\xi}(x) - x\left[\frac{\xi^{n-1} - 1}{\xi - 1}\right] g_{n-2}^{\frac{\lambda+2}{\xi},\,\xi}(x^\xi) = 0 \quad n \geqslant 2\end{aligned} \tag{6}$$

These new polynomials were compared with the standard Jacobi classical polynomial, which is defined by the recurrence relation:

$$\begin{aligned}&P_0^{a,b}(x) = 1 \qquad P_1^{a,b}(x) = \tfrac{1}{2}[a - b + (a + b + 2)x]\\&2n(n + a + b)(2n + a + b - 2)P_n^{a,b}(x)\\&- (2n + a + b - 1)[(2n + a + b - 2)(2n + a + b)x + a^2 - b^2]P_{n-1}^{a,b}(x)\\&- 2(n + a - 1)(n + b - 1)(2n + a + b)P_{n-2}^{a,b}(x) = 0 \quad n \geqslant 2\end{aligned} \tag{7}$$

The orthogonal relationships for these three polynomials are:

$$\begin{aligned}\int_0^1 x\, f_n(x)\, f_m(x)dx &= 0 \qquad && m \neq n\\&= 1/2^{n+1} && m = n\end{aligned} \tag{8}$$

$$\begin{aligned}\int_0^1 x^{\lambda-1}\, g_n^{\lambda,\xi}(x)\, g_m^{\lambda,\xi}(x)\, dx &= 0 \qquad && m \neq n\\&= 1/\lambda + 2[\xi^n - 1/\xi - 1] && m = n\end{aligned} \tag{9}$$

$$\begin{aligned}\int_{-1}^1 (1 - x)^a(1 + x)^b\, P_n^{a,b}(x)\, P_m^{a,b}(x)\, dx &= 0 \qquad && m \neq n\\&= k_n && m = n\end{aligned} \tag{10}$$

$$\text{where } k_n = \frac{2^{1+a+b}\,\Gamma(1+a+n)\,\Gamma(1+b+n)}{n!\,(1+a+b+2n)\,\Gamma(1+a+b+n)} \tag{11}$$

The above are functions of the one-dimensional variable, x. The orthogonality of these functions permits generation of functions of higher dimensions by the following scheme:

$$\begin{aligned}
\theta_1(x_1,x_2,\ldots,x_n) &= \varphi_1(x_1)\varphi_1(x_2)\ldots\varphi_1(x_n)\\
\theta_2(x_1,x_2,\ldots,x_n) &= \varphi_1(x_1)\varphi_1(x_2)\ldots\varphi_2(x_n)\\
\theta_3(x_1,x_2,\ldots,x_n) &= \varphi_1(x_1)\varphi_1(x_2)\ldots\varphi_2(x_{n-1})\varphi_1(x_n)\\
&\;\;\vdots\\
\theta_{n+1}(x_1,x_2,\ldots,x_n) &= \varphi_2(x_1)\varphi_1(x_2)\ldots\varphi_1(x_n)\\
&\;\;\vdots
\end{aligned} \tag{12}$$

where φ_i, $i = 1,\ldots,n$ are the orthogonal functions defined above. It should be noted that in the strict sense, orthonormality is required, rather than orthogonality. An orthogonal function has the property:

$$\begin{aligned}
\int_a^b \mathrm{w}(x)\mathrm{F}_n(x)\mathrm{F}_m(x)\,\mathrm{d}x &= 0 \qquad m \neq n\\
&= C \qquad m = n
\end{aligned} \tag{13}$$

where C is a constant. Orthonormality requires that $C = 1$. Since the resulting decision surfaces are compared with zero, the equations can be multiplied by the normalizing constant, C, without affecting the comparison.

Five variables were used as features: the standardized lactic acid peak area and the presence or absence of peaks of four selected elution times. The identity of the peaks was unknown. A training set of 12 was used, including 6 positive samples and 6 negative samples. The algorithm was run using three polynomials: $\mathrm{f}_n(x)$, $g_n^{0.5,2}(x)$, and $\mathrm{P}_n^{0,\frac{1}{2}}(x)$. An additional standard discriminant analysis was performed on the same data using the Biomedical Data Processing (BMDP) statistical package[6].

RESULTS

A separating hypersurface was obtained by using the above method. A sample equation, obtained for $\mathrm{g}_n^{0.5,2}(x)$ is given below:

$$\begin{aligned}
\mathrm{D}_g = {} & -3.0 - 3.8x_1 - 1.2x_2 + 1.4x_3 + 3.9x_4 + 2.9x_5 + 1.1x_1x_2 +\\
& +0.9x_1x_3 + 4.6x_1x_4 + 0.5x_1x_5 + 0.2x_2x_3 + 5.6x_2x_4 -\\
& 1.2x_2x_5 - 3.2x_3x_4 - 1.1x_3x_5 + 1.2x_4x_5
\end{aligned} \tag{14}$$

These separating hypersurfaces were then used to classify the samples in the test set.

TABLE I

NUMBER OF MISCLASSIFIED SAMPLES

	$f_n(x)^{\star}$	$g_n^{0.5,2}(x)^{\star\star}$	$P_n^{\frac{1}{2},\frac{1}{2}}(x)^{\star\star\star}$	*Discriminant analysis*§
Class 1 (+)	4	3	2	7
Class 2 (−)	1	1	3	0
Overall	5	4	5	7

* Special case 1 of Cohen polynomial.

** Special case 2 of Cohen polynomial.

*** Special case of Jacobi polynomial.

§ Run using BMDP statistical package.

In all, 46 samples were analyzed, 26 of which were positive for SBP. The results are summarized in Table I.

It should be noted that, once the classification algorithm has been run to obtain the separating hypersurfaces, classification of new samples can be made by direct substitution into the separating hypersurface equation. In the above analysis, the resulting classification provides a strictly categoric variable. However, the results obtained from substitution into the hypersurface equations can be interpreted in another way. A numerical value is obtained. Since the separation occurs at $D(\bar{x}) = 0$, the absolute value of the results can be interpreted as a degree of membership in that category: the larger the absolute value, the more certain the classification.

Work is continuing on this application as additional patient data become available. The work described can be extended in two directions. The pattern recognition technique is applicable not only to SBP data but to any analysis in which chromatograms are obtained. In addition, by modifying the knowledge representation and feature extraction phase, the pattern recognition method can be applied to any classification problem and has, in fact, been used in a number of applications[16–18] including diagnosis of SBP by a method which does not involve chromatograms[19,20].

REFERENCES

1 E. A. Patrick, *Syst., Man, Cybern. Rev.*, 6 (1977) 4.

2 E. H. Shortliffe, B. G. Buchanan and E. A. Feigenbaum, *Proc. IEEE*, 67 (1079) 1207.

3 P. Armitage and E. A. Gehan, *Int. J. Cancer*, 13 (1974) 16.

4 M. Ben-Bassat, R. W. Carlson, V. K. Puri, M. K. Davenport, J. A. Shriver, M. Latif, R. Smith, L. D. Portigal, E. H. Lipnick and M. H. Weil, *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-2 (1980) 1481.

5 M. Ben-Bassat, D. B. Campbell, A. R. MacNeil and M. H. Weil, *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-5 (1983) 225.

6 W. J. Dixon (Editor), *BMDP Statistical Software*, University of California Press, Berkeley, 1983, p. 519, p. 537.

7 L. N. Kanal, *IEEE Trans. Inform. Theory*, 6 (1974) 697.

8 E. A. Patrick, *Systm. Man, Cybern. Rev.*, 6 (1977) 4.

9 E. A. Patrick, F. Stelmock, L. Shen, *IEEE Trans. Systems, Man, Cybernetics*, SMC4 (1974) 1.

10 D. Coomans, D. L. Massart, I. Broeckaert and A. Tassin, *Anal. Chim. Acta*, 133 (1981) 215.

11 D. Coomans, I. Broeckaert and D. L. Massart, *Anal. Chim. Acta*, 134 (1982) 139.

12 C. Albano, W. Dunn, U. Edland, E. Johansson, B. Norden, M. Sjöström and S. Wold, *Anal. Chim. Acta*, 103 (1978) 429.

13 S. Wold and E. Johansson, *Anal. Chim. Acta*, 133 (1981) 251.

14 G. Blomquist, E. Johansson, B. Söderström and S. Wold, *J. Chromatogr.*, 173 (1979) 19.

15 E. Jellum, I. Bjørnson, R. Nesbakken, E. Johansson and S. Wold, *J. Chromatogr.*, 217 (1981) 231.

16 M. E. Cohen, D. L. Hudson and P. C. Deedwania, *Use of Pattern Recognition Techniques to Classify Exercise Testing Data, Park City, UT, September 1984*, IEEE, New York, 1984.

17 M. E. Cohen, D. L. Hudson, P. C. Deedwania and N. Gitlin, *Pattern Recognition Using New Orthogonal Functions to Classify Medical Data, Proc., Microcomputers in Medicine, New York, 1984*, ISMM, New York, 1984.

18 M. E. Cohen and D. L. Hudson in M. Gupta (Editor), *Approximate Reasoning in Expert Systems*, North-Holland, 1985, p. 435.

19 M. E. Cohen, D. L. Hudson and N. Gitlin, *Pattern Classification Using a New Orthogonal Function for Recognition of SBP, Proc., American Association of Medical Systems and Informatics, San Francisco, May, 1984*, IEEE, New York, 1984.

20 D. L. Hudson, M. E. Cohen and N. Gitlin, *Pattern Classification of Patients with Spontaneous Bacterial Peritonitis Using New Orthogonal Functions with Extensions to Higher Dimensions, Proc., Symposium on Computer Applications in Medical Care, Washington, DC, November, 1984*, IEEE, New York, 1984.