Classification Results of Mass Spectra of Toxic Compounds by Class Modeling

Long-Biao Yin

Department of Chemistry, East China University of Science and Technology, Shanghai 200237, PR China

Received February 22, 1993[⊗]

Classification of mass spectra of toxic compounds is based on the numerical features derived from spectral data. Computerized classification on these data provides various parameters, which determine the corresponding compound(s) belonging to a certain class of toxic compounds. Pattern recognition studies on mass spectra data are often used for principal components (PC) modeling of the autocorrelation transformed mass spectra. The significant number of principal components is determined by cross validation. With this technique, for the 119 toxic compounds targeted in ambient air, the predictive accuracy for the four classes obtained by GC/MS from training and calibration data is 76.5%.

INTRODUCTION

For the study of chemometrics, the effect of pattern recognition during computerized information process, has been followed with great interests.[1,2] In particular, statistical pattern recognition has its extensive applications in chemistry.[3] Computer methods of pattern recognition can easily be applied to classification of mass spectra data.[4] Pattern recognition based on principal components modeling processes the sample data according to sample class.[5] It provides a set of parameters that characterizes each class and is the basis for interpreting other descriptive quantities of the data structure.[6,7] It allows the quantitative estimation of so called external variables. The resulting residuals can be used to calculate the distance between the classes, and the standard deviation (SD) of the residuals directly corresponds to the distance between the object and the class.[8] For application of pattern recognition to complex mixtures, principal components modeling is extremely powerful,[9,10] such as the prediction of carcinogenicity of N-nitrous compounds on the basis of their physiochemical properties[11,12] and the analysis of environmental polychlorinated biphenyl and air pollutant residues.[13,14] The approach here applies above methods to the classification of the mass spectra (analysis) data of a target list of 119 compounds sought in ambient air. The classification accuracy of 76.5% has been obtained.

METHODS

Apparatus. Data recording and manipulation were provided by an IBM PC microcomputer. The data stored on the computer were analyzed by the UNIPALS software from W. Dunn in Chicago and the PBM/STIRS database of mass spectra.[15,16]

Procedure. The 119 target compounds are listed in Table 1. Four classes of compounds are represented in the target and training sets. Class 1 is of the chloro compounds, class 2 is of the nonhalobenzene and the polycyclic aromatic hydrocarbons, class 3 is of the bromo (fluoro) compounds, and class 4 is of the nonhalolinear hydrocarbons. In the training set, the highest mass is 350, and the interval is from 35 *m/z* to 350 *m/z*. The flowchart of computer-assisted approach to the entire mass spectra classification scheme is shown in Figure 1. Four classes of compounds are listed in Table 2. The flowchart for the identification scheme (that is, the sequence of pattern recognition) is shown in Figure 2.

RESULTS AND DISCUSSION

Classification and Modeling. According to similar structure condition, all compounds are divided into four classes. Within each class, some compounds are selected to model, and then all of the 119 compounds are fitted to this model. On the basis of the calculated results, the correct classification prediction by the model to corresponding compounds can be obtained as well as the accuracy. All of the results, including the summary of classification results, are listed in Tables 4 and 5. It is shown by the results that the number of compounds of each class during classification is related to modeling. Plots of the three-dimensional eigenvector projections were used for tentative classification of all the compounds, and by observing the class structure, eight clusters of feature points in the space were obtained. Thus this in turn leads to the eight classes.

Summary of Classification Results. Unsatisfactory results of pattern recognition methods in spectroscopy can be mainly due to the following reasons:[17] (1) primitive generation of pattern features from spectra, (2) exclusion of chemical knowledge about spectra interpretation, and (3) the classification problems being investigated are oriented too much to traditional chemical substructures yet not to cluster structures studies in the pattern space.

In this study, 52 chloro compounds of the 119 toxic compounds are classified under one category (class 1), and 50 nonhalobenzene and polycyclic aromatic hydrocarbons are put under another category (class 2). About one-third of the 52 chloro compounds does not fit the classification, and not all these compounds are polycyclic hydrocarbons. It has been reported that pollutants can be identified from mass spectral field data obtained during the process of ambient air monitoring.[18]

[⊗] Abstract published in *Advance ACS Abstracts,* September 1, 1994.

0095-2338/94/1634-1232$04.50/0

Table 1. 119 Target and Training Compounds

no.	compound name	no.	compound name
1	chloromethane	**61**	4-chloro-3-methyl phenol
2	bromomethane	**62**	2-methylnaphthalene
3	chloroethene (vinyl chloride)	**63**	hexachlorocyclopentadiene
4	chloroethane	**64**	2,4,6-trichlorophenol
5	methylene chloride	**65**	2,4,5-trichlorophenol
6	acetone	**66**	2-fluorobiphenyl (supr)
7	carbon disulfide	**67**	2-chloronaphthalene
8	1,1-dichloroethene	**68**	2-nitroaniline
9	1,1-dichloroethane	**69**	dimethylphthalate
10	*trans*-1,2-dichloroethene	**70**	acenaphthylene E1 2,6−dinitrotoluene
11	chloroform	**71**	2,6-dinitrotoluene + acenaphthylene
12	1,2-dichloroethane	**72**	3-nitroaniline
13	2-butanone	**73**	acenaphthalene
14	1,1,1-trichloroethane	**74**	2,4-dinitrophenol
15	carbon tetrachloride	**75**	4-Nitrophenol
16	vinyl acetate	**76**	dibenzofuran
17	bromodichloromethane	**77**	2,4-dinitrotoluene
18	1,2-dichloropropane	**78**	diethylphthalate
19	*cis*-1,3-dichloropropene	**79**	fluorene
20	trichloroethene	**80**	4-(chlorophenyl)phenyl ether
21	dibromochloromethane	**81**	4-nitroaniline
22	1,1,2-trichloroethane	**82**	2,4,6-tribromophenol
23	*trans*-1,3-dichloropropene	**83**	4,6-dinitro-*o*-cresol
24	benzene	**84**	*N*-nitrosodiphenylamine
25	bromoform	**85**	4-(bromophenyl)phenyl ether
26	4-methyl-2-pentanone	**86**	α-BHC
27	2-hexanone	**87**	hexachlorobenzene
28	1,1,2,2-tetrachloroethane	**88**	*β*-BHC
29	tetrachloroethene	**89**	pentachlorophenol
30	toluene	**90**	*γ*-BHC
31	chlorobenzene	**91**	phenanthrene
32	ethylbenzene	**92**	anthracene
33	4-bromofluorobenzene (supr)	**93**	*δ*-BHC
34	styrene	**94**	heptachlor
35	xylenes	**95**	di-*N*-butylphthalate
36	2-fluorophenol	**96**	aldrin
37	pentafluorophenol (supr) and D5-phenol (supr)	**97**	fluoranthene
38	phenol	**98**	pyrene
39	bis(2-chloroethyl)ether	**99**	α-endosulfan
40	2-chlorophenol	**100**	P-P′-DDE
41	1,3-dichlorobenzene	**101**	dieldrin
42	1,4-dichlorobenzene	**102**	endrin
43	benzyl alcohol	**103**	*β*-endosulfan
44	1,2-dichlorobenzene	**104**	P-P′-DDD
45	*o*-cresol	**105**	butylbenzylphthalate
46	bis(2-chloroisopropyl)ethyl	**106**	P-P′-DDT
47	*p*-cresol	**107**	endosulfan sulfate
48	*N*-nitrosodi-*N*-propylamine	**108**	methoxychloro and 3,3′-dichlorobenzidine
49	hexachloroethane	**109**	benzo[A]anthracene
50	nitrobenzene	**110**	chrysene
51	isophorone	**111**	bis(2-ethylhexyl)phthalate
52	2-nitrophenol	**112**	di-*N*-octyl-phthalate
53	2,4-dimethylphenol	**113**	benzo[b]fluoranthene
54	2-chloroethoxy(bis)methane	**114**	benzo[k]fluoranthene
55	benzoic acid	**115**	benzo[a]pyrene
56	2,4-dichlorophenol	**116**	indeno[1,2,3col]pyrene
57	1,2,4-trichlorobenzene	**117**	dibenz[a,h]anthracene
58	naphthalene	**118**	benzo[b,h]perylene
59	4-chloroaniline	**119**	heptachlo epoxide
60	hexachlorobutadiene		

Table 2. Four Classes of Compounds

class	compounds
1	1, 3, 4, 5, 8, 9, 10, 11, 12, 14, 15, 18, 19, 20, 22, 23, 28, 29, 31, 39, 40, 41, 42, 44, 46, 49, 54, 56, 57, 59, 60, 61, 63, 64, 65, 67, 80, 86, 88, 89, 90, 93, 94, 96, 100, 101, 102, 104, 106, 108, 119
2	24, 30, 32, 34, 35, 38, 43, 45, 47, 50, 51, 52, 53, 55, 58, 62, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 81, 83, 84, 91, 92, 95, 97, 98, 99, 103, 105, 107, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118
3	2, 17, 21, 25, 33, 36, 37, 66, 82, 85
4	6, 7, 13, 16, 26, 27, 48

[a] Shown by cmpound number, refer to Table 1 for compound name.

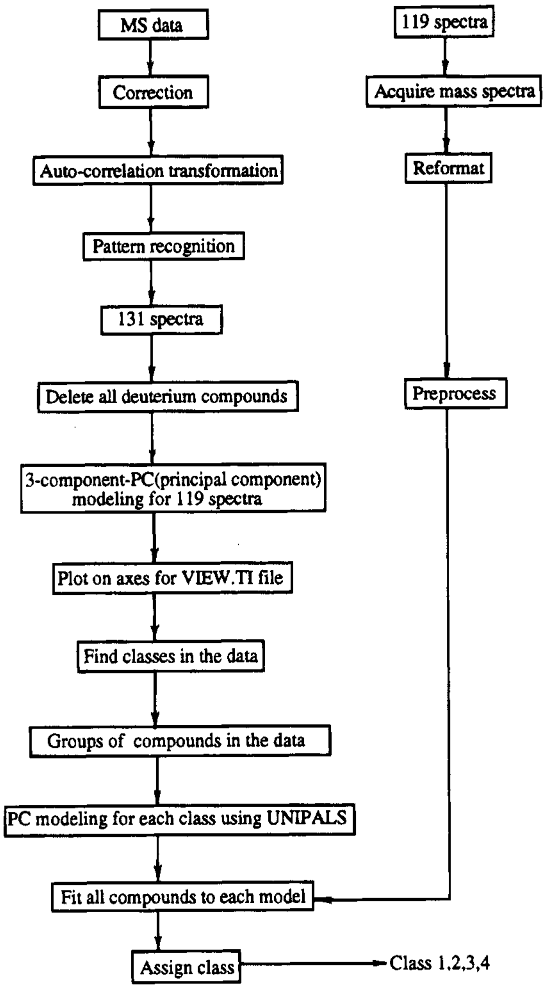
MS data
Correction
Auto-correlation transformation
Pattern recognition
131 spectra
Delete all deuterium compounds
3-component-PC(principal component) modeling for 119 spectra
Plot on axes for VIEW.TI file
Find classes in the data
Groups of compounds in the data
PC modeling for each class using UNIPALS
Fit all compounds to each model
Assign class
Class 1,2,3,4
119 spectra
Acquire mass spectra
Reformat
Preprocess

Figure 1. Mass spectra classification scheme.

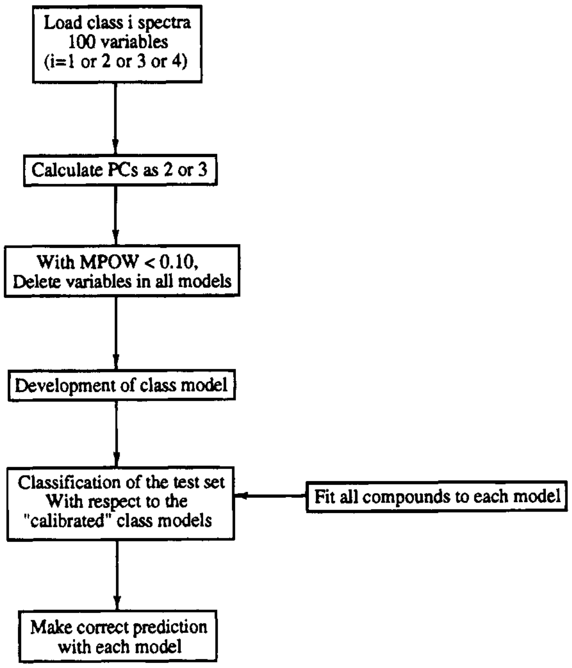

Figure 2. Flowchart of identification scheme.

Actual classification studies indicate that it is very important to select class model and to develop new systems for mass spectra classification. Not only can the compound be identified from this GC/MS data with the principal component modeling but also the method can be used to probe into areas such as its variable selection and classification pattern.[19] In related studies it has been shown to give more accurate results.[20] Pattern recognition method, as a synonym of chemometrics, has been spoken highly of over

Table 3. Eight Classes of Compounds by Three-Axis Image Classification

class	compounds
1	24, 30, 34, 38, 43, 45, 47
2	6, 7, 13, 16, 32, 36, 37, 46, 48, 50
3	1, 3, 4, 12, 26, 27, 33, 35, 40
4	2, 18, 29, 39, 54, 58, 59, 61, 66, 67, 73, 75, 79, 85, 96, 98, 101, 109, 110, 113, 114, 115, 116, 118
5	86, 88, 90, 93, 99, 103
6	20, 60, 63, 89, 107
7	9, 15, 51, 52, 53, 55, 62, 68, 69, 71, 72, 74, 76, 77, 78, 80, 82, 83, 84, 91, 92, 94, 95, 100, 105, 108, 111, 112, 117
8	5, 8, 10, 11, 14, 17, 19, 21, 22, 23, 25, 28, 31, 41, 42, 44, 49, 56, 57, 64, 65, 70, 81, 87, 97, 102, 104, 106, 119

[a] Represented by compound number, refer to Table 1 for compound name.

Table 4. Results of Class Modeling

class	no. of compds by modeling	fit all (119) compds to each model	no. of compds of correct prediction by each model to the no. of compds of each class
1,CH1	52	86	35/52
1,CH1	46	107	48/52
1,CH2	35	93	41/52
1,CH3	27	70	35/52
1,CH4	34	87	42/52
1,CH5	42	106	50/52
1,CH6	23	45	29/52
1,CH7	29	59	34/52
1,CH8	14	17	16/52
1,CH9	26	40	24/52
1,CH10	31	57	34/52
2,NH1	50	72	40/50
2,NH2	40	83	42/50
2,NH3	42	77	45/50
2,NH4	38	83	43/50
2,NH5	36	57	39/50
3,BF1	10	108	10/10
3,BF2	5	48	8/10
3,BF3	3	30	6/10
3,BF4	7	100	10/10
4,NH1	7	44	7/7
4,NH2	6	33	6/7
4,NH3	3	12	3/7

the years.[21] Not only has it shown achievement in areas of separation and identification of compounds,[22] but also it has developed an artifical neural network as one of its means.[23]

Table 5. Summary of Classification Results

class	rate of correct classification	accuracy (%)
1	34/52	67.3
2	42/50	84.0
3	9/10	90.0
4	5/7	71.4
total	91/119	76.5

ACKNOWLEDGMENT

A part of this work was done in the laboratory of Professor W. Dunn, University of Illinois at Chicago, using the UNIPALS software. His interest and encouragement is acknowledged.

REFERENCES AND NOTES

(1) Nillson, N. J. *Learning Machine*; New York: McGraw Hill, 1965; p 17.
(2) Batchelor, B. G. *Pattern Recognition-Idea in Practice*; New York: Prentice Hall: 1978; p 23.
(3) Jurs, P. C. et al. *Chemical Application of Pattern Recognition*; New York: John Wiley & Sons: 1975; p 137.
(4) Voorhees, K. J. et al. *Anal. Chem.* **1985**, *57*, 1630.
(5) Wold, S. et al. In *Chemometrics: Theory and Application-ACS Symposium Series 52*; American Chemical Society: Washington, DC, 1977; p 278.
(6) Wold, S. *J. Pattern Recognition* **1976**, *8*, 127.
(7) Albano, C. et al. *Anal. Chim. Acta* **1978**, *103*, 429.
(8) Wold, S. et al. *Chemometrics–Mathematics and Statistics in Chemistry*; Reidel, D. Publishing Company: New York, 1984; pp 17–95.
(9) Dunn, W. J., III et al. *Analysis* **1984**, *12*, 477–485.
(10) Lindberg, W. et al. *Anal. Chem.* **1983**, *55*, 643–648.
(11) Dunn, W. J., III et al. *Bioorganic Chem.* **1981**, *10*, 29.
(12) Dunn, W. J., III et al. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 8.
(13) Stalling, D. L. et al. *Environmental Applications of Chemometrics*; American Chemical Society: Washington, DC, 1985; p 1.
(14) Scott, D. R. et al. *Environmental Applications of Chemometrics*; American Chemical Society: Washington, DC, 1985; p 106.
(15) Glen, W. G., Dunn, W. J., III; Scott, D. R. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Computer Methodology* **1989**, *2*, 349–76.
(16) Glen, W. G.; Dunn, W. J., III; Sarker, M.; Scott, D. R. UNIPALS: Software for Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Computer Methodology*, **1989**, *2*, 377–96.
(17) Varmuza, K. *Computer Application in Chemistry*; Elsevier Science Publishers BV: Amsterdam, 1983; p 78.
(18) Arcos, J. C. *Environ. Sci. Technol.* **1987**, *21*, 743.
(19) Dunn, W. J., III et al. *Environ. Sci. Technol.* **1989**, *23*, 1499.
(20) Scott, D. R. *Anal. Chim. Acta* **1988**, *211*, 11–29.
(21) Brown, S. D. *Anal. Chem.* **1990**, *62*, 84R, 101R.
(22) Yin, L. B. *FENXI HUAXUE* **1992**, *20(2)*, 137.
(23) Bos, A. et al. *Anal. Chim. Acta* **1992**, *256*, 133.