Classification of Mass Spectra via Pattern Recognition

JAMES R. MCGILL and B. R. KOWALSKI*

Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195

Received June 24, 1977

The classification of low-resolution mass spectral data is one of the oldest and most debated problems in the chemical applications of pattern recognition. This paper reviews the problem and applies all the major preprocessors and classifiers suggested by previous workers, along with some that have not been applied to mass spectral data, to a single, well-defined database. The effects of six different preprocessors, six nonprobabilistic classifiers, and two forms of the first chemical probabilistic classifier are studied and their implications considered. The autocorrelation preprocessor and the *k*-nearest-neighbor classifier are found to be superior.

INTRODUCTION

The use of pattern recognition as an aid in the analysis of low-resolution mass spectra has been discussed since the introduction of the first pattern recognition method to chemistry.[1-3] The difficulties experienced with these data caused workers both to look more deeply into the meaning of the results[2,4,5] and to consider methods of modifying the linear learning machine to increase its usefulness.[6-17] Methods of transforming the spectra to increase their usefulness have also been studied. These included the removal of peak intensity information,[2] taking the log of the intensity to reduce the dynamic range,[18] factor analysis,[19,20] the application of various forms of the Fourier transform[21,22] and its square wave equivalent, the Walsh or Hadamard transform,[23] and the calculation of mass moments.[24] Thus, low-resolution mass spectral problems have become the informal chemical standard for testing new pattern recognition ideas, particularly in the area of preprocessors. Unfortunately, because of this informality, very few of the various papers are comparable, as only a minority of them use the same data sets. Thus the claims of the various authors are contradictory, and a great deal of discussion has ensued.[25-28] Also a number of classification methods other than the linear learning maching have been adapted to chemistry, not all of which have been applied to the mass spectral problem. These include the *k*-nearest neighbor classifier,[29] the pairwise least-squares separator,[3,12,30,31] and statistical isolinear multiple component analysis (SIMCA).[32] With the exception of SIMCA, all of the above-mentioned classifiers do not make explicit use of category models or the underlying probability density information. Therefore, they do not, and should not, report a probability that a given compound contains a particular functional group. With a need for probabilistic measures in several applications, we have investigated two forms of Bayesian discriminate analysis.[33]

In this paper most of the preprocessors mentioned above and all of the classification methods are applied to a single, low-resolution mass spectral data set. This allows some conclusions to be drawn about the comparative effectiveness of these preprocessors and classifiers relative to the mass spectral problem.

EXPERIMENTAL

The data set used in this study consisted of 539 spectra from the Mass Spectrometry Data Centre.[34] These spectra were of CHO compounds with six to eight carbons: 180 compounds contained no oxygens atoms; 180 contained one oxygen; and 179 contained two oxygens. The training set consisted of 150 of each of the three types of compounds, for a total of 450, and the test set consisted of the rest of the compounds, for a total of 89. To test the stability of the classifier results, the training and test sets were randomly selected four times and the classifiers reapplied. The effect of spectral errors was studied by inducing a 10% random variation in the data[35] four times and again applying all the classifiers.

The initial number of features varied from 130 for the raw data to 256 for the Fourier transforms. To allow comparison of the methods, each preprocessed data set was variance weighted,[26] and the 20 most highly weighted features were used for classification; 20 features were chosen because 20 was the smallest number which would include all the significantly weighted features for all the preprocessors. All computations were done on the University of Washington CDC-6400/Cyber-73 dual mainframe computer, and all of the programs used are part of the program library ARTHUR[36] with the exception of the Fourier, Hadamard, and autocorrelation transforms.

The preprocessing methods that were tested are (1) the autoscale transform;[26] (2) the Hadamard transform;[23] the (3) real (cosine), (4) imaginary (sine), and (5) power spectrum generated by the Fourier transform;[21,22] and (6) the autocorrelation spectrum. The data transformed by the latter five methods were autoscaled after transformation to prevent biasing of the classifiers.

The classification methods used were (1) the pairwise and (2) the multicategory linear learning machine, (3) the piecewise and (4) the multiple least-squares discriminant function, (5) the Bayes rule and (6) the orthogonalized Bayes rule classifier, (7) SIMCA, and (8) the *k*-nearest neighbor classifier. These methods belong to three families of classifiers: linear separators, category modelers, and distance measures. Linear separators attempt to pass an $(n - 1)$-dimensional hyperplane through the n-dimensional data space produced by the n variables measured on each member of the data set. These hyperplanes can be calculated in several ways, two of which are negative feedback and least-squares minimization. There are also several ways of approaching separations, two of which are pairwise separations and separation of one category from all others simultaneously. The linear learning machine or hyperplane separator is randomly initialized and uses the error correction negative feedback approach to iterate until it either finds a plane that separates the categories or exceeds a preset limit. In this application the limit was 500 iterations per category pair in the pairwise case, and 2000 iterations overall in the multicategory case. The least-squares minimization directly calculates the best hyperplane for pairwise separation of categories using the generalized inverse method.

Classifiers which model categories use the feature values from the training data to form distributions of feature values for each category, and then classify the training and test sets

0095-2338/78/1618-0052$01.00/0 © 1978 American Chemical Society

using these models. The Bayesian classifier produces probability distributions directly from the feature values for each category. Bayes rule then calculates the probability that a given pattern i is a member of category k, using feature j and the probability distribution associated with feature j for category k ($P[X_{j,k}]$)

$$P_j[X_{j,k}|x_{i,j}] = \frac{(\text{PROBk})(\text{RISKk})P[x_{i,j}|X_{j,k}]}{\sum_{n=1}^{\text{NCAT}} (\text{PROBn})(\text{RISKn})P[x_{i,j}|X_{j,n}]}$$

where PROBk is the a priori probability of a given pattern being a member of category k (i.e., the number of patterns in k divided by the total number of patterns) and RISKk is the risk associated with misclassifying a member of k as a member of another category (set to 1.0 in this application). After this probability has been calculated for each category and each feature has been similarly treated, the total probability that the pattern is a member of category k is calculated by multiplying together the probability that each feature indicates that the pattern is a member of that category

$$P_{\text{total}}[X_k|x_i] = \prod_{j=1}^{\text{NVAR}} P[X_{j,k}|x_{i,j}]$$

and assigning the pattern to the category of highest probability.

Bayes rule was originally developed assuming that the features were independent, which is certainly not the case when dealing with mass spectral data. To provide this orthogonality the eigenvector transform was applied to the whole data set before the Bayesian classifier is used.

A different method of modeling is used by SIMCA. In this method the eigenvector transform is applied to each category, and then the training and test sets are classified by transforming them into each category and classifying them as members of the category that they fit the best.

The single distance measure classifier, k-nearest neighbor, calculates the distance between each point and every other, and then classifies the training and test set members as members of whatever category their k-nearest neighbors belong to. It should be noted that as k approaches infinity, this method becomes a slightly modified version of the Bayesian classifier.[29]

RESULTS AND DISCUSSION

While it would be rhetorically tidy to separate the results into two sections, one dealing with preprocessors and one dealing with methods of classification, it is neither easy nor honest to do so. Different classification methods exist because no one method is optimal for every situation. Different algorithms are sensitive to different data structures, so one would expect that different types of preprocessors would optimize separation for each algorithm. One might also expect that the separations produced by each algorithm would probe the structure and particular problems of a data set.

Figure 1 and Figure 2 are two versions of the same plot, showing the results of applying each of the eight classifiers to data transformed by each of the six preprocessors. Figure 1 expresses the results in terms of the preprocessors, and Figure 2 expresses the results in terms of classifiers. The abscissa is the percent of the training set correctly classified by the method, and the ordinate is the percent of the test set classified correctly. In Figure 1 the autocorrelation and the real and the imaginary parts of the Fourier transform are each represented three times in the upper right hand section of the diagram, which seems to indicate that preprocessors that concentrate frequency information are better for mass spectral information. Figure 2 indicates that the k-nearest neighbor classifier is best, followed by a group consisting of the Bayesian

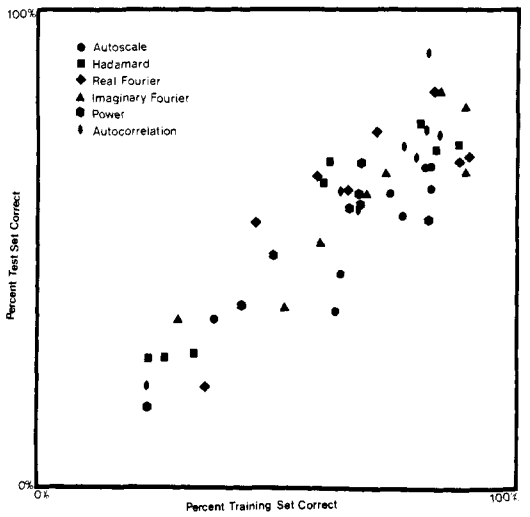

Figure 1. Preprocessor results.

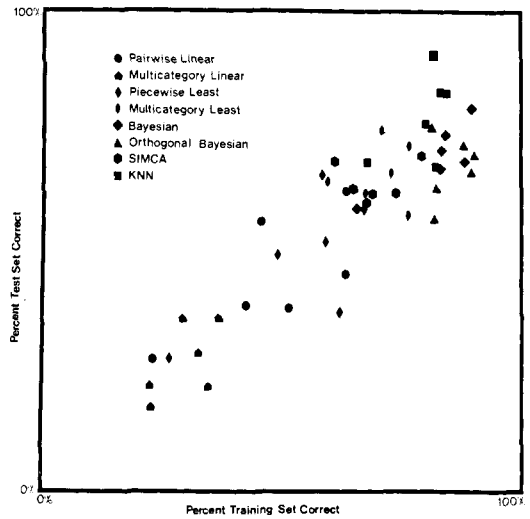

Figure 2. Classifier results.

and the orthogonal Bayes classifier, SIMCA, and the multicategory least-squares discriminant. All of these do fairly well. The hyperplane separator and the pairwise least-squares discriminant do poorly, and the multicategory hyperplane separator does least well of all. Lest one feel that the hyperplane separator was not allowed enough iterations to converge, it should be noted that all the separations reported for both the pairwise and the multicategory versions were essentially stable after less than 100 iterations. Thus the distance classifier is the best, the category modeling classifiers all do well, and the linear separators classify the data at a fair to poor level.

These results are quite interesting, but there is no assurance that the poor showing of the linear separators is not an artifact of the ordering of the data set. To test this possibility, the entire autocorrelation preprocessed data set was shuffled four times, the training and test sets were reselected, and the eight classifiers were applied to the resulting data. The results are shown in Figure 3. The lines connect the results of each classifier in the order that they were produced, with the single symbol marking the first point obtained. As can be seen, the ordering of the classifiers is not significantly altered. The k-nearest neighbor method is always the best, the multicategory hyperplane separator is always the worst, the pairwise hyperplane separator and the pairwise least-squares discriminant perform comparably and not too well, and the Bayesian, orthogonal Bayesian, multicategory least-squares discriminant, and SIMCA do comparably and fairly well. The

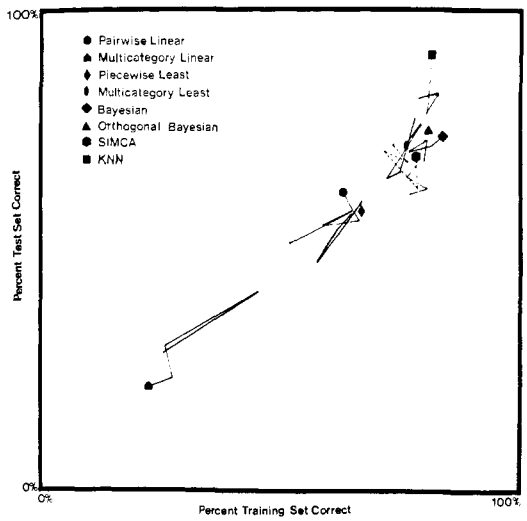

Figure 3. Effect of shuffling data.

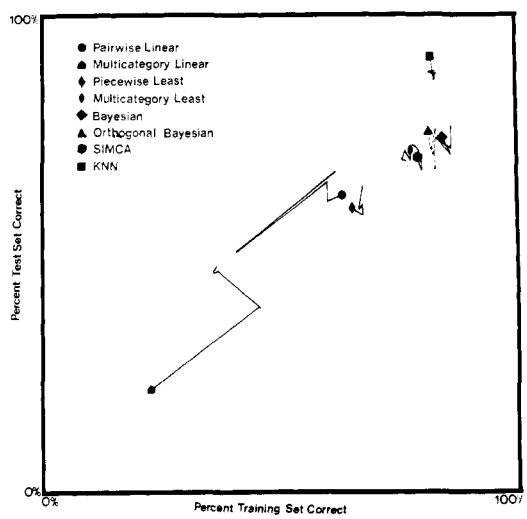

Figure 4. Effect of 10% random error.

mathematical relationship of the various methods is emphasized in Figure 3 by the fact that mathematically similar methods track each other.

Another possible source of concern is whether the results of the classifiers are stable to error in the initial data. To test whether this is the case, 10% random variation was induced in the autocorrelated data four different times and the eight classifiers were applied to the results. Ten percent variation was chosen, as it is slightly in excess of the cumulative errors experienced in the operation of University of Washington's low-resolution mass spectrometer over the course of seven years.[37] Figure 4 is a plot of the results of this operation. Again, the lines connect the points in the order that they were produced, with a symbol marking the first. With the exception of the hyperplane separators, none of the classifiers are as affected by the error as they were by reordering the data.

CONCLUSIONS

The preceding discussion is separated into two sections, with the latter further divided into three subsections. The first section was a comparison of six preprocessors applied to mass spectral data. The results indicate that the autocorrelation preprocessor is the best, followed by the real and imaginary parts of the Fourier transform preprocessor which are both quite good. The power spectrum, the Hadamard transform, and the original data are all considerably less effective preprocessors. Three problems were dealt with in the section on

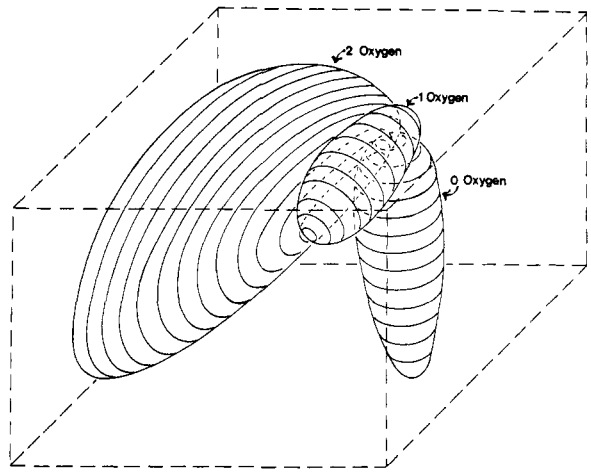

Figure 5. Three-dimensional projection of data space.

classifiers: which classifier of the eight tested was the best for the original data, the affect of reordering the data on the classification results, and how much the classifications were affected by the addition of random variation. The best classifier was the only distance classifier tested, *k*-nearest neighbor. The three modeling classifiers, the Bayesian, orthogonal Bayesian, and SIMCA, also performed quite well. The linear separators did not do well, particularly those based on the negative feedback learning machine. The shuffling of the data did not effect the ordering of the classifiers drastically, and revealed that all the classifiers' results, except those of the multicategory hyperplane separator, do not vary more than about 10% with data set reordering. The addition of 10% random variation to the data, which allows some estimate of the stability of the classification results with respect to spectroscopic errors, causes at most an 8% change in the classification results of six of the classifiers. For the other two classifiers, the multicategory and the pairwise hyperplane separator, the 10% random error produces changes of up to a 30% fluctuation in the classifications.

What are the conclusions that can be drawn from the foregoing, both about the methods of computation used, and about the structure of the mass spectral data space? Clearly, preprocessors that concentrate and emphasize frequency information are the most effective. Also it is clear that linear separation is not a good approach to this data set, while modeling of categories is better, and classification by local category information is best. This implies that the 0, 1, and 2 oxygen spectra are entwined. Figure 5 is a sketch of a three-space projection, via the eigenvector transform, of the mass spectral data space. This projection contains 63% of the variance of the original 128-dimensional space. As can be seen, the 1 oxygen spectra form a tight, spherical group; the 0 oxygen spectra form a cigar-shaped ellipsoid which penetrates into the sphere; and the 2 oxygen spectra form a delta-shaped wedge, of elliptical cross section, which also penetrates the 1 oxygen sphere, and within which it also penetrates the 0 oxygen ellipsoid.

In order to understand these results, one must consider the origin of mass spectra. A characteristic fragment pattern is generated for each compound analyzed. But many of the peaks of one compound's spectrum are nearly identical with those of a similar compound, and distinguishing between chemically similar compounds is a common analysis problem. Specifically, the spectrum of a C_6H_{14} compound, a $C_6H_{14}O$ compound, and a $C_6H_{14}O_2$ compound can have a large number of peaks in common. To a human, these three spectra are easily distinguished, because the human automatically ignores the similarities. But these three spectra are very similar and can be considered closely related. It is not surprising that pattern

recognition has difficulties with this problem as many methods rely on separable categories. When some of the categories being studied are subsets of others, this assumption breaks down. The more strongly a method depends on independence of categories, the more spectacularly it fails, which explains much of Figure 2's trends.

ACKNOWLEDGMENT

We thank Dr. J. W. Frazer of Lawrence Livermore Laboratory for his aid in initiating this study, A. Harper for helping with ARTHUR, and M. daKoven for criticism.

REFERENCES AND NOTES

(1) P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, **41**, 21 (1969).
(2) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 690 (1969).
(3) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 695 (1969).
(4) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **42**, 1387 (1970).
(5) P. C. Jurs, *Anal. Chem.*, **42**, 1633 (1970).
(6) N. M. Frew, L. E. Wangen, and T. L. Isenhour, *Pattern Recognition*, **3**, 281 (1971).
(7) P. C. Jurs, *Anal. Chem.*, **43**, 20 (1971).
(8) P. C. Jurs, *Appl. Spectrosc.*, **25**, 483 (1971).
(9) L. E. Wangen, N. M. Frew, and T. L. Isenhour, *Anal. Chem.*, **43**, 845 (1971).
(10) J. B. Justice, D. N. Anderson, T. L. Isenhour, and J. C. Marshal, *Anal. Chem.*, **44**, 194 (1972).
(11) F. E. Lytle, *Anal. Chem.*, **44**, 1867 (1972).
(12) W. L. Felty and P. C. Jurs, *Anal. Chem.*, **45**, 885 (1973).
(13) T. J. Stonham and M. A. Shaw, *Pattern Recognition*, **7**, 235 (1975).
(14) G. S. Zander, A. J. Stuper, and P. C. Jurs, *Anal. Chem.*, **47**, 1085 (1975).
(15) T. J. Stonham et al., *Anal. Chem.*, **47**, 1817 (1975).
(16) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, *Anal. Chem.*, **47**, 1951 (1975).
(17) L. J. Soltzberg et al., *J. Am. Chem. Soc.*, **98**, 7144 (1976).
(18) L. Pietrantonio and P. C. Jurs, *Pattern Recognition*, **4**, 391 (1972).
(19) J. B. Justice and T. L. Isenhour, *Anal. Chem.*, **47**, 2286 (1975).
(20) G. L. Ritter, S. R. Lowry, T. L. Isenhour, and C. L. Wilkins, *Anal. Chem.*, **48**, 591 (1976).
(21) P. C. Jurs, *Anal. Chem.*, **43**, 1812 (1971).
(22) L. E. Wangen, F. M. Frew, T. L. Isenhour, and P. C. Jurs, *Appl. Spectrosc.*, **25**, 203 (1971).
(23) B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **45**, 2334 (1973).
(24) C. F. Bender, H. D. Shepherd, and B. R. Kowalski, *Anal. Chem.*, **45**, 617 (1973).
(25) J. B. Justice and T. L. Isenhour, *Anal. Chem.*, **46**, 223 (1974).
(26) B. R. Kowalski, "Pattern Recognition in Chemical Research", *Comput. Chem. Biochem. Res.*, **2**, 1–76 (1974).
(27) L. J. Soltzberg et al., *J. Am. Chem. Soc.*, **98**, 7139 (1976).
(28) N. A. B. Gray, *Anal. Chem.*, **48**, 2265 (1976).
(29) B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **44**, 1405 (1972).
(30) J. Schechter and P. C. Jurs, *Appl. Spectrosc.*, **27**, 225 (1973).
(31) C. F. Bender and B. R. Kowalski, *Anal. Chem.*, **46**, 294 (1974).
(32) S. Wold, *Pattern Recognition*, **8**, 127 (1976).
(33) T. W. Anderson, "An Introduction to Multivariate Statistical Analysis", Wiley, New York, N.Y., 1958.
(34) Mass Spectroscopy Data Centre, Building A8.1A, AWRE, Aldermaston, Reading RG74PR, England.
(35) D. L. Duewer, B. R. Kowalski, and J. Fasching, *Anal. Chem.*, **48**, 2002 (1976).
(36) ARTHUR: available from Dr. B. R. Kowalski, Laboratory for Chemometrics, Chemistry Department, University of Washington, Seattle, Wash. 98195.
(37) Dr. A. L. Crittenden, private communication.

LETTERS TO THE EDITOR

Patent Symposium Papers

Dear Sir:

The August 1977 issue, to me, is one of the most valuable you have ever issued, in terms of interest and relevance to my work. The relevance is of the order of 70–80%. That is, there is useful material for me in nearly every article. Further, each article is well written and sheds light on some rather obscure points of the chemical patent situation. I felt at home with the authors of those articles; they are the sort of people whom I would like to cultivate.

Altogether, my heartiest congratulations for an excellent issue.

(Mrs.) **Elizabeth H. Groot**, Librarian
Schenectady Chemicals, Inc.,
Schenectady, New York 12301

Received September 16, 1977

Bibliometric Problems Associated With The Patent Literature

Dear Sir:

The patent symposium published in the August 1977 issue of the Journal was indeed an informative treat. It was especially important in that it focuses attention on some difficult, but inescapable, bibliographic and bibliometric problems.

Bibliometrics, the statistical study of the literature,[1] has been particularly useful for quantitatively characterizing the flow of information involved in the development of pharmaceuticals.[2-4] In fact, it appears to be useful for making predictions about the ultimate clinical success of a drug long before it becomes a marketed commodity.[5]

On the whole, patents form only a small portion of the developing drug literature, less than 10% of the total publications,[4] with some notable exceptions, as might be expected. Nevertheless, patents are important bibliometrically in that their positions in time relative to other publications often reveal clues to the future of their subject drug. Drugs which have a high probability of clinical success usually have clinical papers published before their patents.[2] That is, a drug which has a patent published before a paper reporting its administration to a human has a low probability of clinical success. This is not a hard and fast rule, but a heuristic probability statement—one that, while maybe not refined enough yet for huge financial ventures, does have a practical use as a routine working hypothesis for predicting, and coping with, future information demands.

Nevertheless, in spite of all their bibliometric value, patents are still an irritating nuisance to contend with. Maynard[6] reassures us that although patents are different from journal articles in format, nature, and purpose, they are still important integral components of the chemical literature. Such a position certainly cannot be rejected.

However, dealing with patents bibliographically is often an exercise in frustration. The only readily available printed source for bibliographic data is *Chemical Abstracts* (CA), and, as Maynard stresses, CA is very selective (i.e., biased) in its inclusion. Furthermore, CA only gives bibliographic data for the basic patent; the concordance only gives patent numbers, not even dates. Perhaps the most stinging blow is that the *Official Gazette of the United States Patent and Trademark Office* does not even give the pages!

Since bibliographic data are required before any bi-