for high screenout in many types of query,[4] their general utility may be somewhat less than expected. While the actual screenout figures obtained here are specific to the particular file, queries and screen set selection procedure used, analogous results would seem to be applicable to any procedure which assigns screens to each and every bond in a structure. Accordingly, taking the other points above into consideration, significant increases in screenout performance above some point are unlikely to be gained unless the number of screens available for assignment is considerably enlarged, or alternative types of descriptor are used.

CONCLUSIONS

For the set of substructural queries and the screen selection procedure used here, there is a noticeable tradeoff between the number of screens available for assignment to a structure file and the resolving power of the screen sets. Although discrimination increases with increasing screen set size, improvements in retrieval effectiveness above a certain point are likely to be gained only at the expense of a large increase in the number of screens or of alternative bases for screen selection. The exact point at which the marginal increase in discrimination is outweighed by increased storage and search times will depend on, inter alia, the nature of the fragments, the size of the file, and the efficiency of the iterative search algorithm used for exact matching as well as computer hardware limitations.

ACKNOWLEDGMENT

Thanks are due to David Cooper and Michael Lynch for valuable discussions, to Chemical Abstracts Service for the provision of the structure file and to the Department of Education and Science for the award of a British Library Postdoctoral Research Fellowship.

REFERENCES AND NOTES

(1) M. F. Lynch, "Screening Large Chemical Files" in J. E. Ash and E. Hyde, Eds., "Chemical Information Systems", Ellis Horwood, Chichester, 1975.

(2) W. Graf, H. K. Kaindl, H. Kries, B. Schmidt, and R. Warszawski, "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System", *J. Chem. Inf. Comput. Sci.*, **19**, 51–55 (1979).

(3) J. F. B. Rowland and M. A. Veal, "Structure-Text and Nomenclature Text Searching for Chemical Information: an Experiment with the Chemical Abstracts Integrated Subject File and Registry System", *J. Chem. Inf. Comput. Sci.*, **17**, 81–89 (1977).

(4) G. W. Adamson, J. A. Bush, A. H. W. McLure, and M. F. Lynch, "An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments", *J. Chem. Doc.*, **14**, 44–48 (1974).

(5) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).

(6) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).

(7) P. Willett, "A Screen Set Generation Algorithm", *J. Chem. Inf. Comput. Sci.*, **19**, 159–162 (1979).

(8) G. W. Adamson, V. A. Clinch, and M. F. Lynch, "Relationship between Query and Data-Base Microstructure in General Substructure Search Systems", *J. Chem. Doc.*, **13**, 133–136 (1973).

Fisher Discriminant Functions for a Multilevel Mass Spectral Filter Network

G. T. RASMUSSEN, G. L. RITTER, S. R. LOWRY, and T. L. ISENHOUR*

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

Received February 20, 1979

Fisher linear discriminants are described and applied to classification problems using mass spectral data. A two-level network of discriminants is used to improve classification. These discriminants provide a useful basis for two-dimensional projections of multidimensional patterns.

In many experimental problems the results are displayed in two-dimensional graphs, which are frequently plots of two measured physical or chemical properties. In the context of classification problems, some obscure property is to be determined on the basis of the information in the two-dimensional graphs. Often more than two properties are measured and all information cannot be contained in a two-dimensional form. This paper presents a novel and useful method of projecting multidimensional data onto two dimensions in a way that maintains discriminating information. The method relies on Fisher linear discriminant functions used in a two-level filter network. It is particularly amenable to use in interactive pattern recognition systems having graphic displays, such as those systems described by Sammon and by Koskinen and Kowalski.[1-3] The availability of such low-dimensional graphs of high-dimensional data allows the chemist to assume a greater role in the interpretation of the data.

The classification of organic compounds by using mass spectral data is the example selected to illustrate the Fisher ratio method. The general problem is to determine the presence or absence of specific molecular substructures in organic compounds from mathematical analysis of the low resolution mass spectra of the compounds. Each mass position represents a separate dimension and the measured property is the intensity of the peak at each mass position. One hopes to discriminate between compounds which do or do not contain a specific structural feature by applying a pattern recognition method. In the past, a variety of pattern recognition techniques have been applied to classification problems using mass spectral data. Linear learning machines were used in early studies.[4-7] Adaptive digital learning networks, simplex methods, and progressive filter networks have also been used.[8-10] Recently some of these methods and others, including *k*-nearest neighbor methods and Bayesian discriminant analysis, have been reviewed and compared.[11,12] Linear discriminants which maximize the Fisher ratio offer a useful complement to these methods.

THEORY

In applying a Fisher linear discriminant, the data set is first divided into two discrete categories or classes. One then attempts to find the direction in the multidimensional space defined by the measured properties such that data points projected onto a line in this direction will be maximally discriminated according to the selected classes. The criterion for discrimination is the Fisher ratio. For a single measured property, the Fisher ratio is a number equal to the square of

0095-2338/79/1619-0255$01.00/0

the difference of the class mean values divided by the sum of the variances for the two classes. In general, a single measured property will not be a good discriminant, so that, for a set of measured properties, one wishes to find a direction in the data space which maximizes the Fisher ratio and thus the discrimination. Methods for determining orthogonal linear transformations which maximize the Fisher ratio have been described by Foley and Sammon.[13,14] For a set of mass spectra including m mass positions, each mass spectrum is treated as an m-dimensional vector, $\vec{I}$, of peak intensities. All spectra belong to one of two classes, and the mean vector, $\vec{M}_i$ for each class is calculated according to eq 1,

$$\vec{M}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\vec{I}_{ij} \qquad (1)$$

where the subscript i denotes the class, N_i is the number of spectra in the ith class, and j is an index variable. Similarly, the within-class variance, $\hat{V}_i$, for each class is computed according to eq 2,

$$\hat{V}_i = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(\vec{I}_{ij} - \vec{M}_i)(\vec{I}_{ij} - \vec{M}_i)^{\mathrm{T}} \qquad (2)$$

where superscript T indicates the transpose. The Fisher ratio, F, for a given m-dimensional discriminant vector, $\vec{d}$, is the ratio of the square of the projected class difference to the sum of the projected within-class variances. This is summarized in eq 3–5,

$$\vec{A} = \vec{M}_1 - \vec{M}_2 \qquad (3)$$

$$\hat{V} = \hat{V}_1 + \hat{V}_2 \qquad (4)$$

$$F(\vec{d}) = (\vec{d}^{\mathrm{T}}\vec{A})^2/\vec{d}^{\mathrm{T}}\hat{V}\vec{d} \qquad (5)$$

where $\vec{A}$ is the difference in the class means and $\hat{V}$, an m by m matrix, is the sum of the within-class variance matrices. The vector, $\vec{d}$, which maximizes the Fisher ratio can be computed from eq 6,

$$\vec{d} = c\hat{V}^{-1}\vec{A} \qquad (6)$$

where c is simply a normalizing factor selected to make $\vec{d}$ a unit vector. For classification of a mass spectrum, the spectrum vector is projected onto the discriminant vector and the resulting scalar value, s, is compared with an arbitrary threshold to assign the spectrum to one class or the other.

$$s = \vec{d}^{\mathrm{T}}\vec{I} \qquad (7)$$

A second discriminant vector which maximizes the Fisher ratio yet is orthogonal to the first vector can be computed according to eq 8,

$$\vec{d}' = c'\left[\hat{V}^{-1} - \left(\frac{\vec{A}^{\mathrm{T}}(\hat{V}^{-1})^2\vec{A}}{\vec{A}^{\mathrm{T}}(\hat{V}^{-1})^3\vec{A}}\right)(\hat{V}^{-1})^2\right]\vec{A} \qquad (8)$$

where $\vec{d}'$ is the second vector and c' is a normalizing factor for this vector.

EXPERIMENTAL PROCEDURE

The two sets of mass spectra used in this study were both large subsets of the Registry of Mass Spectral Data.[15] One set consisted of 1000 randomly selected mass spectra. Discriminant vectors for this set were calculated using data for mass positions from 12 to 200 amu inclusive. The second set consisted of 13 140 mass spectra of organic compounds having peaks at mass positions no higher than 280 amu. For this set, 269-dimensional discriminant vectors were computed using data at mass positions 12 to 280 amu inclusive. Discriminant vectors to identify several molecular features were computed and tested using each set of mass spectra. Two-dimensional projections were plotted on a color television display interfaced to a 64-kbyte Raytheon 704. The use of a color display allowed projected points corresponding to members of different classes to be plotted in different colors rather than with different characters, thus preserving the full resolution of the display with a minimum of overlapped points. (With this display system, classification performance of the discriminants was evaluated interactively.) Computations of discriminant vectors were performed on an IBM 370/165 computer at the Triangle Universities Computation Center using programs written in Fortran IV.

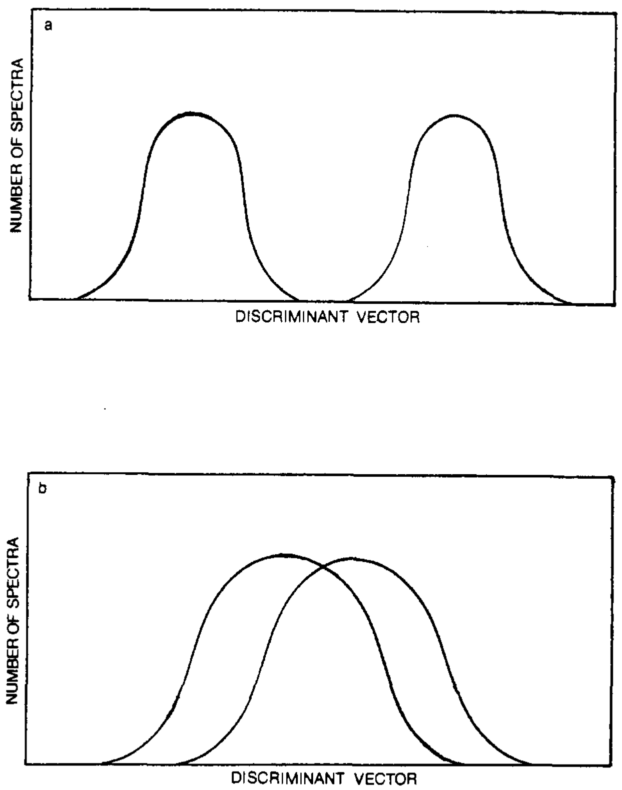

Figure 1. Hypothetical distributions of projected spectra in two classes for (a) an ideal case and (b) a severely overlapped case.

RESULTS AND DISCUSSION

The effectiveness of a discriminant vector can be tested by projecting all mass spectra in a data set onto the vector and observing the distribution of class members along this vector. With an ideal discriminant, the projected class means will be sufficiently different and the projected scatter within the two classes will be sufficiently narrow that the classes will be completely separated. In this case, a threshold can be selected which will correctly classify all spectra. If the projected class means are only slightly different and if the within-class scatter is greater, the distributions of the two classes may overlap so that no threshold can be used to classify all compounds. Figure 1 illustrates these two hypothetical cases with graphs showing the number of spectra projected onto a discriminant vector. Figure 2 is a histogram reflecting the results obtained when all members of the large data set are projected onto a discriminant vector computed to distinguish compounds containing nitrogen from nonnitrogen compounds. The observed performance for this discriminant lies between the extreme cases shown in Figure 1. Although the distributions for the two classes overlap, nitrogen-containing compounds are separated from nonnitrogen compounds.

The occurrence of such overlapping distributions, which were typical of the cases studied, means that a single discriminant vector will have a limited effectiveness for classification. Thus,

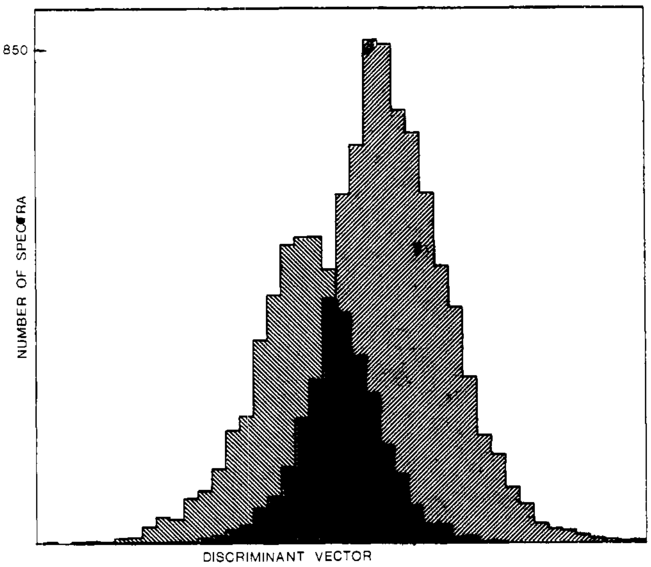

Figure 2. Distributions of 13 140 spectra projected onto a discriminant vector. Hash marks upwards to the left denote nitrogen compounds; those upwards to the right denote nonnitrogen compounds.

efforts were made to improve the classification results by combining the use of two discriminant vectors. As previously mentioned, it is possible to calculate a second discriminant vector orthogonal to the first discriminant which also maximizes the Fisher ratio. It was thought that by using these two discriminant vectors and projecting the mass spectra in two dimensions, the classification performance might be improved. In practice, however, the projections of mass spectra onto the two discriminants were highly correlated so that projected points fell essentially on a straight line. Figure 3 is a graph of 100 mass spectra projected onto two orthogonal Fisher discriminants computed to classify carboxyl vs. noncarboxyl compounds. Each vector is partially effective as a discriminant as evidenced by the clustering of points representing carboxyl compounds at one end of the line with those points representing noncarboxyl compounds clustering at the other end. However, the use of two discriminant vectors computed in this way does little to improve the classification results. Presumably, this is because of redundant information in the mass spectral data.

As an alternative, a filtering method was used to improve the performance of the classifiers. A first discriminant vector is calculated as before. All mass spectra are projected on this vector, and the distribution of the compounds in the two classes is studied. Two threshold values are selected which divide the discriminant vector into three segments. Those spectra for which the projected scalar values are less than the lower threshold, L, are identified as members of the class having the specific substructural characteristic. Similarly, those spectra for which the projected points lie above the upper threshold, U, are identified as members of the class lacking the characteristic. No classification is made for the projected spectra falling between the two thresholds. Instead, these spectra are used to calculate a second discriminant vector. The same Fisher ratio method is used to compute this vector except that fewer spectra are used. Figure 4 is a histogram for the projection of mass spectra onto a discriminant vector computed to distinguish phenyl compounds from nonphenyl compounds. The vertical lines, L and U, indicate the location of threshold values on the discriminant. All projected spectra lying on the left of L are classified as phenyls while those projected spectra lying to the right of U are classified as nonphenyls. The second discriminant vector is used to classify only those spectra left unclassified by the first vector. Threshold values may be established for the second vector, and thus the discriminants are used in a two-level filtering network. A graph of 100 mass spectra projected on a pair of discriminant vectors computed to distinguish phenyl and nonphenyl compounds is shown in Figure 5. The lines, L and U, mark the threshold values used to define the "no decision" region along the first discriminant. The second discriminant was computed using 331 of the original 1000 spectra. The projected spectra in the region

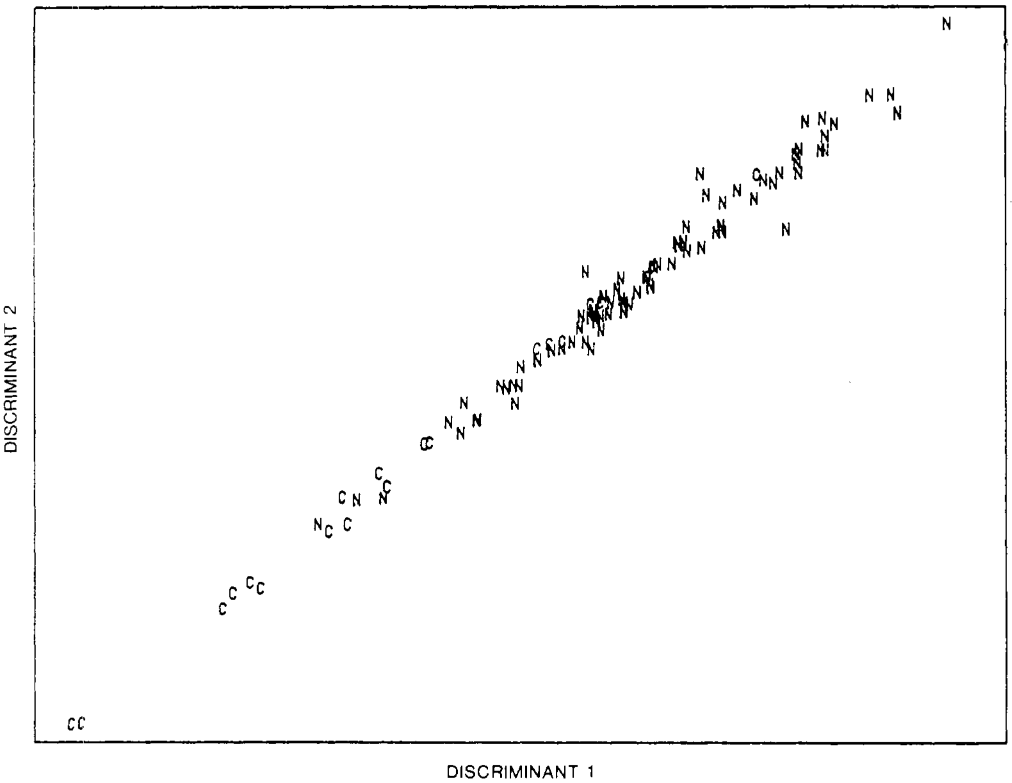

Figure 3. Projection of 100 spectra onto two discriminant vectors. "C" denotes compounds containing a carboxyl group, and "N" denotes those without a carboxyl group.

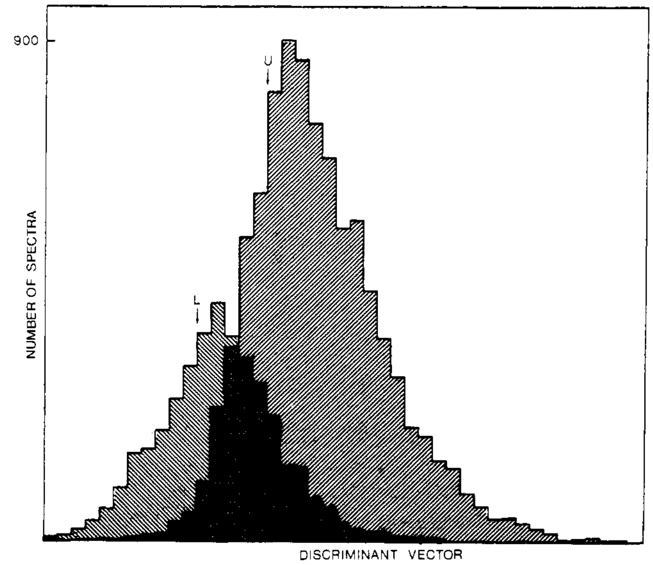

Figure 4. Distributions of 13 140 spectra projected onto a discriminant vector. Hash marks upwards to the left denote phenyl compounds; those upwards to the right denote nonphenyl compounds. The positions of two thresholds are indicated with "L" and "U".

between lines L and U show a noticeable distribution along the vertical discriminant, with phenyls clustered nearer the bottom and nonphenyls clustered nearer the top of the graph. This indicates that the second discriminant is useful in improving the classification. This method has been used to define discriminant vectors for various molecular substructures of organic compounds.

A number of quantitative measures have been used to evaluate the performance of the discriminant functions. The recall is defined as the ratio of the number of compounds correctly identified as belonging to a class to the number of compounds actually in that class. The precision is defined as the ratio of the number of compounds correctly identified as belonging to a class to the total number of compounds assigned to that class by the discriminants. An overall precision can be defined as the fraction of the classifications which were correct. Also, the fraction of the data set classified is reported. Table I summarizes these definitions.

Table I. Summary of Performance Metrics[a]

	Classification by Discriminant		
	structure	structure	
Actual	present	absent	unclassified
structure present	a	b	c
structure absent	d	e	f
	Metrics		
	+	−	overall
recall	$\frac{a}{a+b+c}$	$\frac{e}{d+e+f}$	
precision	$\frac{a}{a+d}$	$\frac{e}{b+e}$	$\frac{a+e}{a+d+b+e}$
fraction classified			$\frac{a+b+d+e}{a+b+c+d+e+f}$

[a] The "+" denotes presence, the "–" absence of the structural feature in question.

These performance statistics are reported in Table II for discriminants computed to classify spectra according to the presence or absence of several molecular features. The discriminants calculated and tested with the smaller data set are listed first. The effectiveness of the discriminants varies according to the problem tried, and this reflects varying degrees of overlap for the projected class distributions. Classification results for the carboxyl problem reflect the independent use of two orthogonal discriminants each computed from the full data set. The discriminants independently perform about equally well, and little improvement is gained from the second

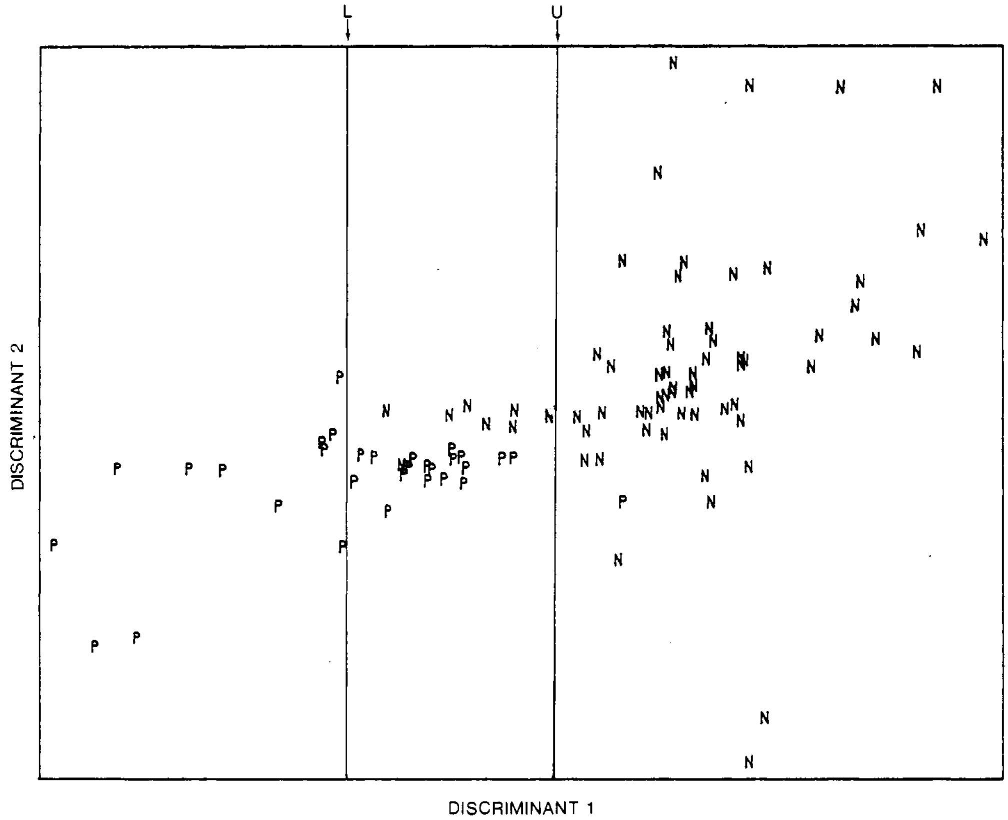

Figure 5. Projection of 100 spectra onto two discriminant vectors. "P" denotes phenyl compounds, and "N" denotes nonphenyl compounds. "L" and "U" are threshold values along the first discriminant which define the central unclassified region.

Table II. Statistics for Classification by Fisher Linear Discriminants

no. of mass spectra	molecular feature	% data set with feature	discriminant used	+precision	+recall	−precision	−recall	% classified	overall precision
1000	carboxyl	19.2	1 level	89.5	22.4	90.3	79.2	75.6	90.2
			1A level	89.5	22.4	90.2	77.8	74.4	90.2
	hydroxyl	19.7	1 level	74.5	17.8	91.9	72.4	63.2	90.7
			2 levels	74.8	58.9	90.4	95.1	100.0	88.0
	nitrogen	35.5	1 level	87.9	49.0	93.1	65.2	64.9	84.9
			2 levels	88.0	55.8	91.0	83.4	81.5	90.2
	phenyl	30.2	1 level	90.0	47.7	94.7	82.2	86.9	89.3
			2 levels	92.7	80.4	92.0	97.3	100.0	92.2
13140	alkene	21.0	1 level	86.7	1.9	90.1	46.6	41.3	90.1
	sulfur	12.0	1 level	81.9	3.7	91.0	96.7	91.1	93.6
	nitrogen	38.4	1 level	88.9	23.8	91.7	57.7	49.1	91.1
			2 levels	87.7	33.3	90.6	69.3	61.7	89.9
	phenyl	29.1	1 level	91.4	34.7	91.9	78.5	72.1	91.2
			2 levels	88.1	47.1	90.0	87.4	82.2	90.0

discriminant. For all other classification problems, the second discriminant is computed for a subset of the data and represents the second level of a filter network. The use of a second-level discriminant vector consistently increases the recall for a given level of precision. The thresholds selected for the hydroxyl and phenyl problems with a two-level filter are such that all spectra in the data set are classified with an overall precision around 90% in both cases. For the nitrogen discriminants, a 90% overall precision can be achieved for slightly over 80% of the data set. If thresholds are selected so that all compounds are classified by the two-level nitrogen discriminants, the overall precision falls to 85%. The somewhat lower precision for the identification of compounds containing hydroxyl groups is probably caused in part by the imbalance in class size. Similarly, the one-level carboxyl discriminant can achieve a precision near 90% on only 22% of the carboxyl-containing compounds. Not unexpectedly, better results are observed for the larger class (which lacks the molecular feature) with each set of discriminants tested.

Also reported in Table II are the results obtained using the discriminant functions computed with the set of 13 140 mass spectra. As before, the discriminants are evaluated by testing their ability to classify all members of the data set. The performance of discriminants for sulfur-containing compounds and for alkenes illustrate the effects of uneven class sizes. In each case the smaller class falls at one "end" of the discriminant, but under the tail of the distribution for the larger class. This indicates that some discrimination has occurred even though a reasonable precision can be obtained for only a small fraction of the compounds in the sulfur-containing and alkene classes. Results obtained with the nitrogen and the phenyl problems can be compared with those observed with the smaller data set. The same levels of precision can be achieved when a somewhat larger unclassified region is allowed. The two-level phenyl discriminants still show a 90% overall precision with 82% of the spectra classified. If thresholds are selected so that all spectra are classified, the overall precision is slightly better than 85% for the two-level phenyl discriminants and 81% for the two-level nitrogen discriminants. These results seem quite respectable considering that few pattern recognition techniques have been tested with data sets containing more than 10 000 mass spectra.

In summary, the use of Fisher linear discriminants offers an effective method for the classification of mass spectra. Improvements in classification effectiveness can be obtained by using discriminants in a multiple-level filter network. Two-dimensional projections of the data on discriminant vectors can be useful in assessing the performance of the discriminants and in aiding the selection of useful threshold values for classification. The method is, of course, not restricted to mass spectra but can be applied to any multidimensional data set which can be divided into two classes.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation.

REFERENCES AND NOTES

(1) J. W. Sammon, Jr., *IEEE Trans. Comput.*, **C-19**, 594 (1970).

(2) J. W. Sammon, Jr., A. H. Procter, and D. F. Roberts, *Pattern Recog.*, **3**, 37 (1971).

(3) J. R. Koskinen and B. R. Kowalski, *J. Chem. Inf. Comput. Sci.*, **15**, 119 (1975).

(4) P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, **41**, 21 (1969).

(5) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 690 (1969).

(6) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 695 (1969).

(7) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 1949 (1969).

(8) T. J. Stonham, I. Aleksander, M. Camp, W. T. Pike, and M. A. Shaw, *Anal. Chem.*, **47**, 1817 (1975).

(9) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, *Anal. Chem.*, **47**, 1951 (1975).

(10) S. R. Lowry, J. C. Marshall, and T. L. Isenhour, *Comp. Chem.*, **1**, 3 (1976).

(11) C. L. Wilkins, *J. Chem. Inf. Comput. Sci.*, **17**, 242 (1977).

(12) J. R. McGill and B. R. Kowalski, *J. Chem. Inf. Comput. Sci.*, **18**, 52 (1978).

(13) J. W. Sammon, Jr., *IEEE Trans. Comput.*, **C-19**, 826 (1970).

(14) D. H. Foley and J. W. Sammon, Jr., *IEEE Trans. Comput.*, **C-24**, 281 (1975).

(15) E. Stenhagen, S. Abrahamsson, and F. W. McLafferty, "Registry of Mass Spectral Data", Wiley-Interscience, New York, 1974.