(22) Isenhour, T. L., "Rapid Memory-Conserving, Compiler-Level Search Algorithm for Mixture Spectra", *Anal. Chem.*, **45**, 2153-4 (1973).

(23) "Codes and Instructions for Wyandotte-ASTM Punched Cards, Indexing Spectral Absorption Data", American Society for Testing and Materials, Philadelphia, Pa., 1964.

(24) Silverstein, R. M., and Bassler, G. C., , "Spectrometric Identification of Organic Compounds", 2nd ed, Wiley, New York, N.Y., 1967, p 237.

(25) Denk, J. R., and Gunn, J., "ISIS-Infrared Spectral Information System—User's Manual", Triangle Universities Computation Center Document No. LSR-98, Research Triangle Park, N.C., 1970.

A Feature Selection Technique for Binary Infrared Spectra

S. R. LOWRY and T. L. ISENHOUR*

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

Received June 18, 1975

Feature selection is frequently employed in the computer classification of spectral data to overcome storage limitations and to decrease computation times during the development of discriminant functions. This paper describes a method of choosing a subset of the important features in binary data based on the a posteriori probabilities as determined by Bayes rule. When this technique was applied to 2600 infrared spectra taken from the ASTM file, the number of 0.1-μm intervals was reduced from 139 to 32 with an overall loss of about 2% in classification ability. The simplicity of this feature selection technique and its success with binary infrared spectra demonstrate that it should be considered in situations involving binary data.

INTRODUCTION

In computer classification of chemical infrared spectral data, a training set of spectra whose classifications are known is frequently employed to empirically develop discriminant functions.[1-4] These discriminant functions can then be applied to an unknown spectrum to give information concerning molecular structure.

Because of the limited memory size of computers and the large amounts of time often required to develop discriminant functions, rarely are all possible spectra incorporated into the training set. In order to overcome storage limitations and to decrease computation times some form of information compression is needed. One method is to optimally pack each spectra as a series of bits, each representing a peak maximum in the infrared spectra. An alternate method is to retain only the information necessary to differentiate among classes. The number of peaks important in a particular classification problem is normally much smaller than the total number of peaks present in an infrared spectrum. If some method is used to select only the important peaks, the "size" of the stored spectra can be greatly reduced. While sophisticated feature selection techniques have been reported in the pattern recognition literature[5,6] and several have been applied to chemical data containing intensity information,[7,8] very little has been reported concerning feature selection of binary data (peak–no peak).[9] This problem is quite significant when one tries to develop discriminant functions using the ASTM file of 91,875 infrared spectra. In this file, each spectrum is stored as a string of ones and zeros. A one indicates that a peak maximum is present in the corresponding wavelength interval of the spectrum. If the number of intervals (features) can be reduced from 100 to 25, the size of the training set could be increased by a factor of 4. Correspondingly, the number of arithmetic operations occurring in the computations could be reduced by as much as a factor of 4, thereby greatly increasing the speed of calculating a discriminant function. However, if considerable interclass information is lost in the feature selection process, the reduced spectra will be less useful as a training set for any type of pattern recognition technique.

To select a set of features that retain most of the discriminating information, we have taken a training set of binary infrared spectra and progressively deleted those wavelength intervals from each spectra which contain the least interclass information. By comparing the classifying ability of discriminant functions developed on the reduced spectra, we obtain a measure of goodness for the particular reduced set of features. This paper describes a method to evaluate the usefulness of each feature in classifying unknown compounds and to assign a relative order of goodness to the features. By selecting the best 32 features out of a total of 139 features, the predictive ability of several types of discriminant functions approached values found using all 139 features.

DATA SET

The data for this study were obtained from the file of 91,875 binary infrared spectra assembled by the American Society for Testing and Materials and made accessible by the Triangle Universities Computation Center (TUCC), the North Carolina Educational Computing Service, and the R. J. Reynolds Tobacco Company. Thirteen mutually exclusive classes were chosen with the main criterion for their selection being that they were similar to ones reported in previous work,[10] thus simplifying the comparison of results. Compounds containing C, H, O, and N atoms exclusively and with a carbon content ranging from C_1 to C_{15} were the only ones selected for this study. From those spectra belonging strictly to each of the 13 classes, 200 were randomly selected from each class, resulting in a data set of 2600 spectra. The range 2.0 to 15.9 μm was divided into 139 intervals of 0.1 μm each. Computations were done on the TUCC IBM 370/165 teleprocessing with the University of North Carolina Computation Center IBM 360/75 using Fortran IV and PL/I computer programs.

* Author to whom correspondence should be addressed; Alfred P. Sloan Fellow, 1971–1975.

THEORY

From the spectra in the training set we approximate the a posteriori probability $P(c_j|x_i = 1)$ of a particular compound belonging to each of the 13 classes (c_j) given that a certain wavelength interval (x_i) contains a peak maximum. Thus for each of the 139 wavelength intervals we have 13 probabilities, one corresponding to each of the 13 functional groups. If the 13 probabilities for a particular interval are similar, then the interval does not contribute to the classification. Those intervals with the most variance among the 13 a posteriori probabilities are the most significant for classification and should be retained.

The a posteriori probability is calculated by using Bayes rule

$$P(c_j|x_i = 1) = \frac{p(x_i = 1|c_j)P(c_j)}{p(x_i = 1)} \quad \begin{matrix} i = 1, 2, \ldots, 139 \\ j = 1, 2, \ldots, 13 \end{matrix} \tag{1}$$

where $p(x_i = 1|c_j)$ is the probability of a peak maximum appearing in interval i given that the compound belongs to class j. $P(c_j)$ is the a priori probability of a compound belonging to class j, and $p(x_i = 1)$ is the probability of a peak maximum appearing in interval i for all spectra.

The conditional probability can be approximated by the expectation value for each interval i and each class j

$$p(x_i = 1|c_j) = \frac{1}{mj} \sum_{n=1}^{mj} X_{n,i} \tag{2}$$

$$i = 1, 2, \ldots, 139$$

$$j = 1, 2, \ldots, 13$$

$$n = 1, 2, \ldots, 200$$

where $x_{n,i}$ is the value (either 1 or 0) of the ith interval of the nth spectrum of class j and mj is the number of spectra in class j (200).

The a priori probability, $P(c_j)$, is equal to the frequency of appearance of each class. Since the classes are the same size, $P(c_j)$ is the same for each class ($^1\!/_{13}$). The denominator term in eq 1 is simply the expectation value of a peak maximum appearing in an interval over all 13 classes

$$p(x_i = 1) = \frac{1}{2600} \sum_{n=1}^{2600} x_{n,i} \quad i = 1, 2, \ldots, 139 \tag{3}$$

Once these three quantities are calculated for each class, the variance among the 13 probabilities for each wavelength interval in the spectral region is easily determined.

$$\sigma_i^2 = \frac{1}{13} \sum_{j=1}^{13} [\bar{P} - P(c_j|x_i = 1)]^2 \tag{4}$$

where $\bar{P}$ is the average of the 13 values. Initially this value was used as the measure of goodness; however, we found that often an interval with a large value of σ_i^2 appears in very few of the 2600 spectra. When a peak maximum appeared in this interval, the classification was accurate but this occurred so seldom that the wavelength interval was really useless. To compensate for this situation, we include a cost factor for each interval. This weighting factor (w_i) is proportional to the number of times a peak appears in interval i for all 2600 spectra. Thus the function used to evaluate the goodness of a particular feature is

$$G_i = \sigma_i^2 w_i \tag{5}$$

It is important to note that when the weighting factor is included the calculation of G_i is greatly simplified. Since the probabilities $P(c_j)$ are the same for all cases, they contribute nothing to the relative variance. The weighting term cancels out the denominator term. This leaves the value of G_i proportional to the variance among the conditional probabilities or average class spectra

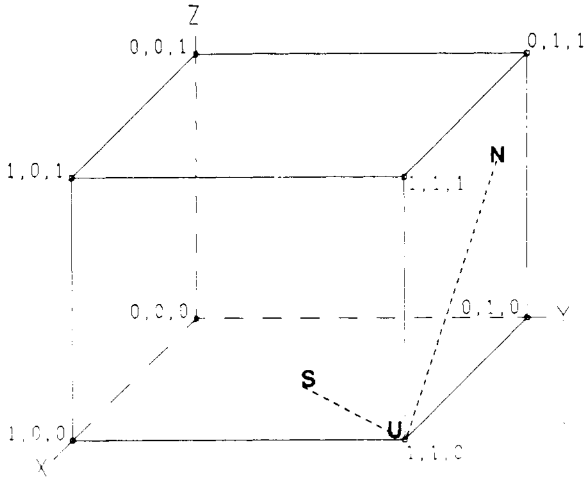

Figure 1. Three-dimensional example of a distance measure for binary data. S is the average class point. N is the average nonclass point. U is the point representation of a binary spectra.

Table I. Ordered Features in Micrometers

4.5	5.8	3.7	5.9	3.0	3.6	3.8	3.1	6.5	4.4
6.2	6.0	4.6	5.7	6.4	7.1	3.2	2.9	3.5	6.3
6.6	12.5	7.0	7.8	9.6	3.4	8.0	7.4	9.5	9.0
7.3	9.7	9.4	8.5	7.5	9.9	10.7	8.4	6.8	8.9
7.7	6.1	11.3	8.2	6.9	2.8	6.7	3.3	8.8	7.2
9.8	14.6	5.2	13.4	4.1	10.6	13.5	8.3	3.9	11.8
10.8	10.5	10.4	7.9	12.2	8.1	10.0	8.6	4.7	11.4
14.3	2.3	11.9	13.6	2.7	12.0	9.1	5.6	11.1	8.7
13.7	11.7	14.7	12.3	11.0	9.3	12.9	4.0	12.1	13.1
2.4	5.1	11.6	10.2	7.6	10.3	12.4	10.1	4.2	10.9
12.6	5.4	11.5	15.7	13.0	5.0	11.2	15.3	13.3	14.2
4.3	9.2	15.2	5.5	5.3	14.0	12.7	14.1	14.8	4.9
2.5	12.8	15.5	13.2	13.9	14.5	4.8	14.4	2.6	15.0
15.4	13.8	15.8	14.9	2.0	15.1	2.2	15.6	2.1	

$$G_i \propto \sum_{j=1}^{13} [\bar{P} - p(x_i = 1|c_j)]^2 \tag{6}$$

G_i is computed for all 139 features in the binary infrared spectra. The features are then ordered on the basis of G_i values. Table I gives all 139 wavelength intervals in order of goodness as calculated above.

RESULTS

In order to evaluate the feature selection techniques, we developed discriminant functions using progressively fewer of the ordered features. The discriminant function used simply assigned an unknown compound to the class whose average spectrum most closely resembled that of the unknown.[11]

The first type of question investigated is whether the unknown compound contains a particular function group. The average class spectrum (S) is calculated using all the spectra in the training set that belong to compounds containing the functional group. The average nonclass spectrum (N) is calculated from the remaining spectra in the training set. If these two spectra are represented as points in a multidimensional space, an unknown spectrum (U) (also represented as a point) will be assigned to the class which has the nearer point.

$$D = d(\mathrm{S,U}) - d(\mathrm{N,U}) \tag{7}$$

Figure 1 shows a three-dimensional example of this problem. Since all the unknown spectra are binary, they can be represented as the vertices of a cube. The average class point S and the average nonclass point N are represented as points within the cube. One simply calculates the Euclidean distance from the unknown U to S and N. The unknown will be assigned to the closer point (in this case S).

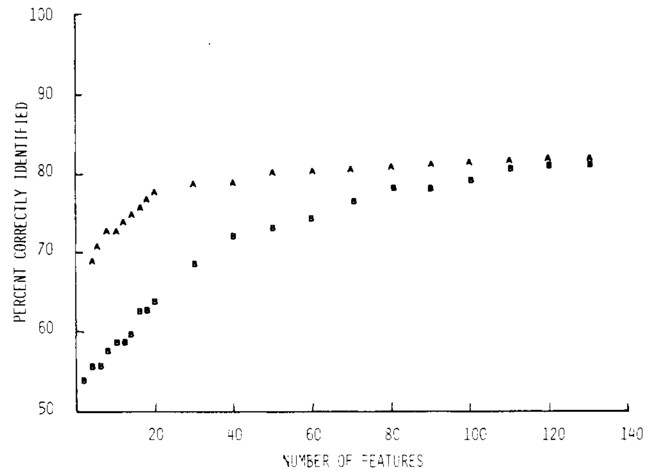

Figure 2. A plot of overall % correct for the 78 two class questions as a function of number of features. Points labeled A were calculated using the selected features. Points labeled B were calculated using random, ordered features.

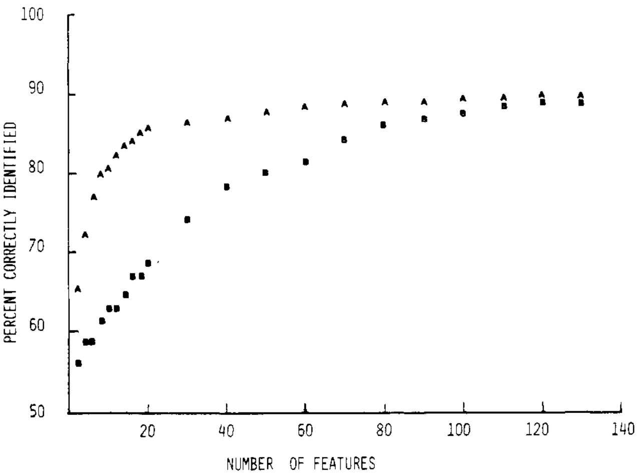

Figure 3. A plot of the overall % correct for the 13 class–nonclass questions as a function of the number of features. Points labeled A were calculated using the feature selection technique. Points labeled B were calculated using randomly ordered features.

Table II. Percentage of Compounds Correctly Identified

Functional group	% positive	% negative	Av
1 Carboxylic acid	81	90	86
2 Ester	85	86	86
3 Ketone	74	78	76
4 Alcohol	79	83	81
5 Aldehyde	80	88	84
6 Amine (prim)	88	84	86
7 Amine (sec)	75	77	76
8 Amine (tert)	72	83	78
9 Amide	74	78	76
10 Urea and derivatives	82	81	82
11 Ether and acetal	78	79	79
12 Nitro and nitroso	81	86	84
13 Nitrile and isonitrile	87	93	90
Overall percent correct	81.8		

Table II gives the results for the 13 functional groups contained in the training set using all 139 features. The % positive column gives the percent of the 200 spectra belonging to each class which were closer to S than N. The % negative column gives the percent of the 2400 nonclass spectra which were closer to N than S. The last column is the average % correctly identified for each class. The overall percent correct is the average of the last column. This is the number used in evaluating reduced sets of features.

The distance calculation described above is now repeated, but fewer of the features are used in finding the distance (the dimensionality of the space is reduced). The points labeled A in Figure 2 are the overall percent correct values for computations done by progressively deleting the less important features. To compare these results to randomly chosen features, the calculations were repeated, but instead of using the selected order, the features were ordered using a random number generator. The points labeled B show the results of this work.

A second question frequently asked in qualitative analysis of infrared spectra data is whether the compound belongs to one class or another (is it an alcohol or is it a ketone?). To look at this type of question, we determine the point corresponding to the average spectrum for each of the 13 classes as before and look at all pairwise class questions (1 vs. 2, 1 vs. 3, . . ., 12 vs. 1). That is, we take an unknown spectrum that we know belongs to either class 1 or class 2 and assign it to the class having the closest point. Table III gives the classification results for all these questions using all 139 features. The final number is the average prediction ability for all 78 questions. Low classification results for a particular question occur when the two points are close together. This indicates that binary infrared spectra belonging to these two classes are very similar and there is great uncertainty in assigning them to one or the other. Figure 3 shows the average prediction ability as a function of the number of features used in developing the discriminant function. The calculations were repeated using a randomly ordered set of features and these results are shown in the lower trace. Once again the selected features significantly improve the prediction ability of the discriminant function over randomly selected features.

CONCLUSION

The results in the previous section show that as little as 2% of the classification ability of the discriminant function is lost when the number of features is reduced from 139 to

Table III. Percent Correctly Identified in All Two Class Problems

	Class A												
Class B	1	2	3	4	5	6	7	8	9	10	11	12	13
1													
2	89.2												
3	83.5	82.2											
4	93.2	97.0	92.2										
5	87.5	90.7	87.0	96.0									
6	94.5	98.0	92.0	90.5	95.7								
7	93.7	95.2	90.0	78.7	90.7	84.5							
8	93.2	92.2	85.7	87.0	89.2	90.0	77.0						
9	86.2	88.7	79.7	87.5	89.2	83.7	85.5	86.0					
10	87.5	94.7	84.2	90.2	88.5	85.2	89.0	90.2	77.5				
11	92.5	90.2	86.7	84.2	90.7	92.2	86.7	78.7	87.5	93.2			
12	90.7	95.2	88.7	91.2	93.0	91.2	87.5	88.7	90.7	90.2	90.7		
13	93.7	93.2	91.2	90.5	92.5	95.0	91.7	90.5	92.2	91.5	92.7	89.2	
Overall percent correct			89.5										

32. By retaining only 32 of the features it is possible to reduce storage requirements and computation times to the point where elaborate pattern recognition problems can be implemented on microcomputers. It should be noted that these results are dependent on the classification problem investigated and the type of discriminant function used.

In a second application of this feature selection technique, Woodruff[12] calculated the nearest neighbor to each of the compounds on the basis of their binary infrared spectra. He performed these calculations using both 139 features and the 32 best features as chosen by this technique. The results of this work demonstrate that the nearest-neighbor algorithm performs as well or better with the reduced set of features than with all 139. The simplicity of this feature selection technique and its success in these cases demonstrate that it should be considered in any situation where computations involve binary spectral data.

ACKNOWLEDGMENT

Support of this research by the National Science Foundation is gratefully acknowledged.

LITERATURE CITED

(1) Jurs, P. C., and Isenhour, T. L., "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.

(2) Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilley, C. N., "Interpretation of Infrared Spectrometry Data", *Anal. Chem.*, **41**, 1945 (1969).

(3) Preuss, D. R., and Jurs, P. C., "Pattern Recognition Techniques Applied to the Interpretation of Infrared Spectra", *Anal. Chem.*, **46**, 520 (1974).

(4) Liddell, R. W., and Jurs, P. C., "Interpretation of Infrared Spectra Using Pattern Recognition Techniques", *Anal. Chem.*, **46**, 2126 (1974).

(5) Andrews, H. C., "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, N.Y., 1972, p 15.

(6) Tou, J. T., "Feature Extraction in Pattern Recognition", *Pattern Recognition*, **1**, 3-11 (1968).

(7) Kowalski, B. R., and Bender, C. F., "Pattern Recognition II Linear and Nonlinear Methods for Displaying Chemical Data", *J. Am. Chem. Soc.*, **95**, 686 (1973).

(8) Jurs, P. C., "Mass Spectral Feature Selection and Structural Correlations Using Computerized Learning Machines", *Anal. Chem.*, **42**, 1633 (1970).

(9) Wilkins, C. L., Williams, R. C., Brunner, T. R., and McCombie, P. J., "Heuristic Pattern Recognition Analysis of Carbon-13 Nuclear Magnetic Resonance Spectra", *J. Am. Chem. Soc.*, **96**, 4182-5 (1974).

(10) Woodruff, H. B., Ritter, G. L., Lowry, S. R., and Isenhour, T. L., "Density Estimation and the Characterization of Binary Infrared Spectra", *Technometrics*, in press.

(11) Woodruff, H. B., Lowry, S. R., and Isenhour, T. L., "A Comparison of Two Discriminant Functions for Classifying Binary Infrared Data", *Appl. Spectrosc.*, **29**, 226 (1975).

(12) Woodruff, H. B., Lowry, S. R., Ritter, G. L., and Isenhour, T. L., "Similarity Measures for the Classification of Binary Infrared Data", *Anal. Chem.*, **47**, 2027 (1975).

A Method of Structure–Activity Correlation Using Wiswesser Line Notation

GEORGE W. ADAMSON* and DAVID BAWDEN

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received June 19, 1975

Fragment sets generated manually from Wiswesser Line Notation have been used to correlate the chemical structures of a group of 79 penicillins with their serum binding activity, using multiple regression analysis. Statistically significant correlations were found, with results in accordance with the generally accepted nature of the binding. Algorithmic methods for the generation of such fragment sets are proposed and the use of various structural representations for structure–property correlation within chemical information systems is discussed.

INTRODUCTION

The investigation of the relationship between physical, chemical, or biological properties and chemical structures has recently been a field of considerable activity, particularly in the search for a methodology of rational drug design. Four major approaches to this problem may be distinguished:[1]

(i) semiempirical methods, correlating biological activity with physico-chemical properties[2]

(ii) additive mathematical modelling, for series of structurally related compounds[3]

(iii) correlations based on quantum-mechanical studies[4]

(iv) substructure analysis: in which the activity of a chemical species is correlated directly with structural features using methods related to substructure search procedures.

This last approach has the major advantage that, unlike the semiempirical and additive modelling methods, it can be applied to collections of structurally diverse compounds. Also its relative simplicity and compatability with substructure search techniques should enable its use as a routine, large-scale procedure in a manner not presently possible with the more sophisticated quantum-mechanical methods. The structural features used have included connection table fragments[5–7] and substructures from a fragmentation code.[8] Structural features such as standardized heteroatom counts have also been utilized.[9] Both regression analysis and pattern recognition techniques have been applied for structure–activity correlation and property prediction.

Wiswesser Line Notation (WLN)[10] is widely used for chemical structure representation in information storage and retrieval systems, and is also used in systems storing property data.[11] This notation thus has obvious potential for structure–property correlation, as has been noted by