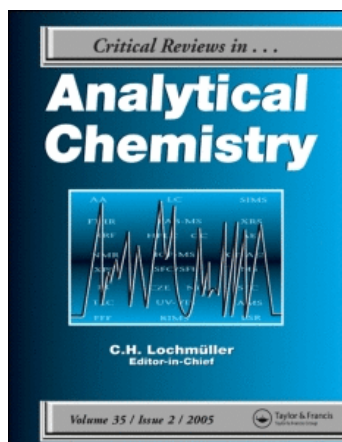


Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



<http://www.informaworld.com/smpp/title~content=t713400837>

T. L. Isenhour; B. R. Kowalski; P. C. Jurs; Louis Meites

**URL:** <http://dx.doi.org/10.1080/10408347408542669>

PLEASE SCROLL DOWN FOR ARTICLE

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## APPLICATIONS OF PATTERN RECOGNITION TO CHEMISTRY

Authors: T. L. Isenhour  
University of North Carolina  
Chapel Hill, N.C.

B. R. Kowalski  
University of Washington  
Seattle, Wash.

P. C. Jurs  
Pennsylvania State University  
University Park, Pa.

Referee: Louis Meites  
Clarkson College of Technology  
Potsdam, N.Y.

"Pattern recognition" is difficult to define because various terms have come to have different meanings to various groups. References to "pattern recognition," "learning machines," "machine decisions," "artificial intelligence," "empirical decision processes," and other names are scattered throughout the literature. Different groups of users, such as electrical engineers, physical scientists, neurobiologists, linguists, and psychologists, have adopted different phrases and have attached their own interpretations to them. Even in chemistry conflicts in usage occur, as evidenced by the early use of "pattern recognition" in mass spectrometry to mean the direct comparison of one mass spectrum to a library of standards as an attempt to identify compounds. Indeed, one might say that all attempts to correlate information in any form amount to the recognition of patterns that represent that information. The problem becomes one of restricting the definition so that it can have some useful connotation.

The current usage of "pattern recognition" in

chemistry might be best described by realizing that all methods of interpreting data stem from either theoretical or empirical procedures or from some combination of the two. The theoretical development of cause-and-effect relations is highly desirable in science, since it indicates that at least some understanding has been achieved of the processes involved. However, this is frequently impossible or impractical, and empirical methods must be used to arrive at certain useful solutions.

We shall, therefore, use the term "pattern recognition" to denote the use in data-interpretation processes, developed by empirical examinations of data from known sources, to elucidate relations from unknown data. Furthermore, we will use the term "learning" to mean the improvement of a decision or classification process based upon its experience. This improvement may be effected by various ways of adding more information to the system or extracting more information from it. However, learning will in general involve some "feedback" mechanism in which the process

that produces a successful result is encouraged (positive feedback) or the process that produces an undesirable result is discouraged (negative feedback).

The general problem of data interpretation is that of placing the data into categories. Defining the categories may in itself be a rather complex process. Still, once they have been defined, the problem becomes one of finding a method that will, to the degree of success desired, place the data into those categories.

In most of the applications of pattern recognition to chemistry to date, data have been represented as patterns in hypergeometric form. That is, data are represented as points in a space of sufficient dimensionality. For example, if only the melting points, boiling points, and molecular weights of a series of compounds were known, each compound could be completely described by a point in a three-dimensional space in which one dimension was devoted to each of the properties. Even though spaces of higher dimensionality are difficult to visualize, the concept can be extended as far as is necessary to represent the data. Even for data sources where the abscissa is continuously variable, such as frequency or wavelength in spectroscopy, the hyperspace representation is completely valid as long as there is some finite limit to the resolution of the measurement. In this case, the dimensionality of the space may be given as the maximum number of meaningful resolution units of the spectrum, and then a single point can represent the complete spectrum. For instance, if an infrared spectrum is recorded from 0.2 to 15 nm and if the resolution is 0.1 nm, then an entire spectrum can be represented by 149 digitized values of absorbance or by a single point in 149-dimensional space. Multidimensional vectors are also frequently referred to in pattern-recognition applications. (The vector refers to lines from the origin to the point described earlier.)

The hypergeometric notation has some very useful features in pattern recognition. In fact, most of the applications that will be discussed are based on one method or another of finding relations among points in hyperspace. Such methods include simple approaches like determining the location of a hyperplane (the so-called threshold logic unit), that will divide points into two classes or, as in nearest-neighbor techniques, determining which of two sets of points is closer to an unknown. In fact, most of the successful

methods of pattern recognition, at least in the area of chemistry, are conceptually very simple. The difficulty of appreciating the methods probably arises chiefly from the fact that most chemists think in terms of theoretically based interpretations rather than in terms of empirical ones.

In short, the process of pattern recognition can be defined as one of transforming patterns from measurement space into classification space. That is, the typical geometric notation described above produces a set of points in the space defined by the measurement space. If any real relation exists among points of a given class, then the problem is to find the transformation that will cause the points to fall in the desired regions of classification space.

In order to develop useful classification processes and to determine their success, it is customary to employ some data for which the correct classifications are known. A common procedure is to use one set of data to develop the classification process and another to test it. Usually the known data are divided into groups by some random selection process to avoid accidental bias. (An alternative procedure is to use all of the data sets but one to develop the classifier, and the remaining one to test it. This is repeated, using every pattern as the test point. This procedure might seem to be prohibitively long, but there are circumstances under which it is practical.) The set of data used to develop a decision process is referred to as a training set, and the success of training is referred to as recognition. (Recognition is usually computed as the percentage of the members of the training set that are correctly classified by the final classifier, although there are cases where it is meaningful to weight recognition based on the classes.) The set of data used to test the developed classifier is referred to as the prediction set, and prediction is defined as the percentage of correct classifications out of the total number of predictions attempted.

As this is written, pattern recognition is far from being a well-developed and fully organized approach to the interpretation of chemical data. Not only has it been introduced too recently to have attained that status, but there are other good reasons for believing that it is in a very early stage of development. Advances being made in various diverse areas, such as speech and handwriting analysis, image interpretation, and others, suggest that new approaches should be forthcoming from

a variety of areas. Therefore, this review will not be presented in the fashion that would be appropriate to a more mature subject that has become well defined. Furthermore, basic points of the techniques will not be covered in detail here, because the authors have participated in four recent publications that have been at least partially dedicated to introducing the subject.<sup>1-6</sup> (These reviews also include references that introduce the pattern-recognition literature outside chemistry.) Rather, we will present a roughly chronological development of the applications of pattern recognition to chemistry.

In 1964, Tal'roze and coworkers published a paper on the minimum sufficient information to identify individual organic substances by coincidence of their mass-spectral lines.<sup>7</sup> It was shown that for 900 spectra the intensity ratios of only two mass values (39 and 41) sufficed to establish that a given compound was one of twelve possible compounds within the library. Three ratios were shown to be sufficient for unequivocal identification, assuming a 5% error in the measurement. While this work does not fall neatly within our definition of pattern recognition, it was a precursor to another publication by these workers in 1966, which may have been the first modern application of pattern recognition to chemistry.<sup>8</sup> Raznikov and Tal'roz developed a mass-spectral classification technique based on a truncated mass spectrum. The lines to be included were selected as being characteristic of the sets to be classified. (The principal case was the discrimination between saturated hydrocarbons on the one hand and monoolefins and cycloparaffins on the other.)

The vectors generated from the truncated spectra were normalized to unit length, and then cones were sought that would contain the two groups. Several algorithms were used in the development of classifying cones, and recognitions as high as 90% were obtained.

\* In 1968, Drozdov-Tikhomirov presented a learning approach for qualitative structural group analysis using infrared spectra.<sup>9</sup> He studied infrared spectra in an attempt to separate carbonyl compounds. Classification was accomplished by using a potential function that produced positive results for one class and negative ones for the other. Huge expenditures of computer time were involved — 8 hours for complete classification of 194 spectra. It was recognized that decreasing the dimensionality of the data could substantially

reduce convergence time, although this might be achieved only at the expense of some deterioration of performance. However, the important feature of this work was that complete training (100% recognition) was accomplished.

Two papers by Crawford and Morrison on the identification of mass spectra appeared in 1968.<sup>10,11</sup> In the first, "unknown" mass spectra were matched against those in a library, and it was shown that mass spectra are so highly specific that successful search routines can be devised even when samples are impure or measurements are noisy. In the second, Crawford and Morrison applied a cluster-analysis method to the identification of molecular classes from a data set of 182 mass spectra. The vector representation of spectra was used, and all of the mass spectra were normalized to generate points on the surface of a hypersphere. This was done by setting the sum of the squares of the mass-spectral peak intensities equal to unity. The similarity between two spectra was defined as the Euclidean distance between them. The centers of gravity of the clusters of points were calculated for various classes of chemically related compounds, such as ethers and amines. The Euclidean distances between centers of gravity of different clusters were calculated, as were the distances between individual points and the centers of gravity of the clusters to which they belonged. These distances were reported. The mass spectra were also reduced to 14-peak spectra by adding the intensities in each successive 14-a.m.u. interval to those in the previous such intervals. Distances between cluster centers and individual "unknowns" were also calculated for these reduced spectra. Correlations were observed between intercluster distances and chemical similarities, and the "unknowns" could often be put in their correct chemical classes by this method.

In 1969 the first of a series of papers was published by the present authors and their coworkers under the main title of "Computerized Learning Machines Applied to Chemical Problems." The first paper<sup>12</sup> used threshold logic units in a branching decision tree to classify mass spectra according to molecular formulas. The data base consisted of data from the American Petroleum Institute Project 44 on 346 compounds containing from one through seven carbon atoms per molecule.

The threshold logic unit amounts to a weight vector ( $W$ ) multiplied by the mass spectrum

represented as a pattern vector ( $Y$ ). An additional constant allows the threshold to be conveniently set at zero. In the hypergeometric representation, which is a very convenient way to think of this method, the weight vector becomes a normal vector representing a hyperplane, which is sought so that it will dichotomize the training set. The dot product between these two vectors provides a simple way of finding whether a pattern vector lies on the same side of the decision plane as the weight vector or on the other side.

$$s = W \cdot Y = |W| |Y| \cos \theta = \sum_i W_i Y_i \quad (1)$$

$$\begin{aligned} -90^\circ > \theta > 90^\circ \quad \cos \theta > 0 \text{ and } s > 0 \\ 90^\circ > \theta > 270^\circ \quad \cos \theta < 0 \text{ and } s < 0 \end{aligned} \quad (2)$$

Hence a simple calculation (the multiplication of two arrays) serves to determine which class a pattern lies in. The problem then was to find a decision surface, if one existed, that would completely separate the two sets. A version of a linear learning machine was applied for this purpose. The term "learning" is used to describe a decision process that improves its performance with experience. Negative feedback was applied in this work to move the decision surface whenever it misclassified a point, so that it would now lie on the correct side of that point and as far from it as it had previously been on the wrong side. This is accomplished by adding to the weight vector the product of the pattern vector and the correct coefficient.

If

$$s' = -s \quad (3)$$

and

$$W' = W + cY \quad (4)$$

since

$$W \cdot Y = s$$

$$W' \cdot Y = s'$$

then

$$c = -2W \cdot Y / Y \cdot Y \quad (5)$$

While this calculation always corrects the weight vector for the misclassified point, it may or may

not improve the situation with respect to any other point. However, if there is a linear decision surface for the given sets of data, then this method will eventually converge to give complete training.

The importance of this publication lies in the fact that it demonstrated that mass spectra could be successfully separated by this method. Actually, 26 different decision vectors were developed to answer the question of molecular formula within this closed set of data. Figure 1 shows the branching tree used. Note that each decision vector represents a separate problem and is trained to 100% recognition.

Furthermore, the reliability of the technique was tested by using distorted spectra from the same set. Normally distributed (Gaussian) errors were imposed on the data. In 1,000 such trials, even with a standard deviation as high as 20%, still 95% of the molecular formulas were correctly determined. Since it requires an average of about ten decisions to arrive at a molecular formula, these results correspond to an extremely low individual error rate. This constitutes further proof of the extremely high redundancy of mass-spectral data on some questions and also demonstrates that pattern recognition is capable of using this redundancy.

The second paper in the series<sup>13</sup> introduced the concept of predictive ability to the classification of chemical data. Mass-spectral data were used again.

Predictive ability is probably the most exciting application of pattern recognition. When an empirically derived classifier can successfully classify unknown data — that is, data that were not used in any way in the development of the classifier — then it is reasonable to assume that the pattern-recognition process has extracted some of the fundamental relation between the data and the categories. Furthermore, this may be used as a demonstration that some relation actually exists between the data and the categories, an interesting and often useful result in itself.

In the predictive process the entire data set is divided, using some random selection device, into two sets: one for developing the decision maker, and the other for testing its predictive ability. In the simple binary case, a predictive ability of 50% would be expected for random guessing, although with small data sets predictive abilities may fluctuate quite widely around 50%. In Reference 13, predictive abilities of about 90% were obtained by

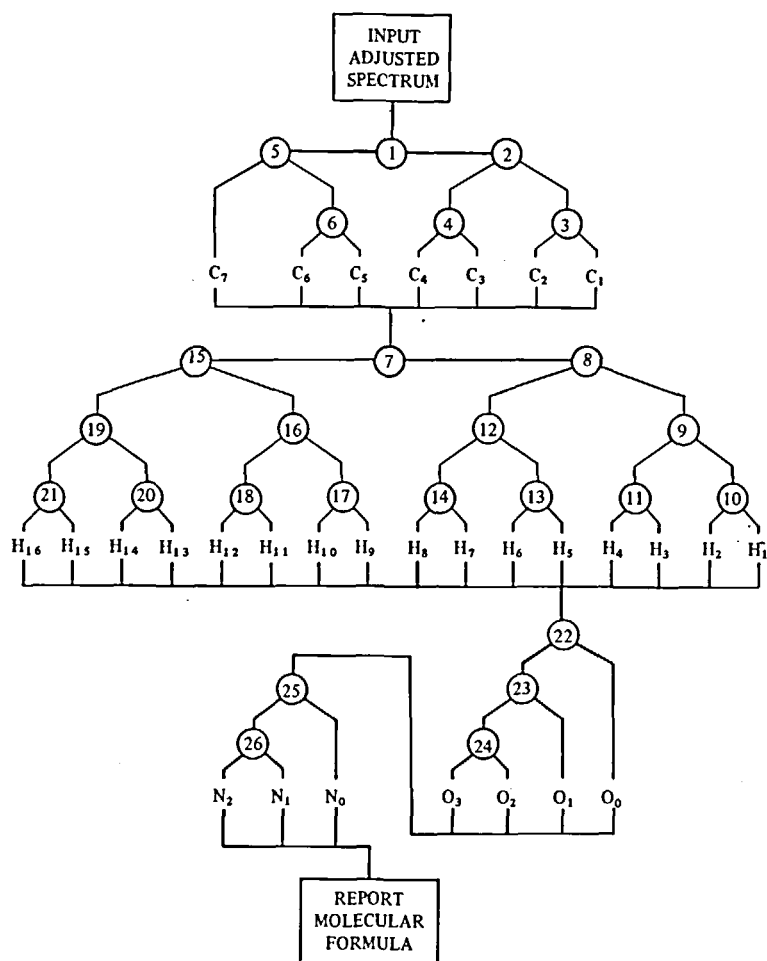


FIGURE 1. Branching tree of binary-decision makers for determining molecular formula. (From Jurs, P. C., Kowalski, B. R., and Isenhour, T. L., *Anal. Chem.*, 41, 21, 1969. With permission.)

several approaches in compounds as containing or not containing oxygen on the basis of their mass spectra. (Later work has considerably improved this prediction.)

Good predictive results can always occur by accident, and this is true for theoretical predictions as well as for empirical ones. However, a predictive ability that is substantially better than random guessing gives rise to confidence in the meaningfulness of the results. A paper<sup>43</sup> to be discussed later treats this problem in some detail by actually generating random sets and testing both recognition and prediction.

Also in Reference 13 are further indications of the high redundancy of mass spectra and one approach to the use of pattern recognition to select the important components of data to be classified.

If predictive ability, convergence rate, and recognition do not suffer unduly, decreasing the number of adjustable parameters used by the learning machine (i.e., the dimensionality of the pattern and weight vectors) is desirable, because the number of computations per classification is linearly related to the number of parameters. Table 1 shows the rate and predictive ability as a function of reduction in dimensionality. Below 35 parameters training was terminated, because full recognition was not obtained within a preset number of iterations. The predictive ability remains within the noise level, but it is fairly constant, despite a drastic reduction in the number of parameters. The performance of the classifier is quite impressive, even with only 35 parameters. The classifier gains complete recognition of the training set and still correctly predicts 90% of

TABLE 1

Convergence Rate and Prediction of Oxygen Presence as a Function of Number of Parameters

Training set = 300		
Parameters	Average recognition	Average percent predicted
155	300	90.6
95	300	91.0
65	300	89.9
50	300	90.2
35	300	90.2
20	284	86.2
10	245	72.2

From Jurs, P. C., Kowalski, B. R., Isenhour, T. L., and Reilly, C. N., *Anal. Chem.*, 41, 690, 1969. With permission.

previously unseen spectra, even though individual spectra may have only ten or twenty peaks occurring within the 35 mass positions. As expected, the predictive ability falls markedly as the number of parameters becomes very small.

The results shown in Table 1 were obtained by using the following criterion for casting out parameters. After each training sequence the resulting weight vector components were multiplied by the sum of the peaks appearing in the corresponding mass position. The resulting numbers indicate the contribution made by each mass position to the overall determination. The weight positions with the smallest indicators, in absolute value, are dropped (in groups of 60, 30, 15, 15, 15, and 10) to reduce the dimensionality of the system. This is felt to be a better criterion for dropping parameters than using the absolute values of the weights, because it takes into account the possibility that a large weight may correspond to only small peaks or, conversely, a small weight may correspond to very large peaks.

Another approach to classification of mass spectra used by these authors was the application of least squares to generate a multicategory classifier.<sup>14</sup>

In general, the least-squares procedure did not have as high a predictive ability as the linear learning machine. However, the least-squares procedure does have the advantage of being a single, albeit long, calculation rather than an iterative scheme whose convergence rate is unpredictable. Furthermore, least squares will produce some sort

of classifier even in cases where the data are inseparable by a linear classifier. (Later work will discuss a way to overcome this problem with linear classifiers.) Hence, least squares usually produces better prediction on inseparable data than a simple linear classifier whose training was terminated arbitrarily on the basis of computer time or number of training steps.

The weight vector ( $W$ ) developed in a binary pattern classifier is actually a linear combination of some or all of the patterns of the training set, developed in such a fashion that the dot product of the weight vector and the  $i$ th pattern ( $W \cdot Y_i$ ) gives a scalar ( $s_i$ ) whose sign indicates to which of two categories the pattern belongs. The principle of the multiclassification method is to develop a weight vector that produces values of  $s_i$  having both magnitudes and signs that place the patterns in one of several categories. The correct category ( $s_i^*$ ) for each pattern of the training set is defined by some arbitrary value. For example, in the case of developing a weight set to differentiate between compounds having zero, one, two, three, or four oxygen atoms, the values of 0, 1, 2, 3, and 4 may be assigned to the  $s_i^*$  for the corresponding compound. A least-squares procedure is employed to select the set of weights that yields values of  $s_i$  so as to minimize

$$\sum (s_i - s_i^*)^2.$$

$$Q = \sum_{i=1}^n (s_i - s_i^*)^2 = \sum_{i=1}^n \left( \sum_{j=1}^m w_j y_{ij} - s_i^* \right)^2 \quad (6)$$

where  $m$  is the number of mass positions,  $n$  the number of patterns in the training set, and  $Q$  the sum of squares of deviations.

The normal equations to minimize  $Q$  are developed by taking the partial derivatives of  $Q$  with respect to each of the weights and setting these equal to zero

$$\frac{\partial Q}{\partial w_k} = 2 \sum_{i=1}^n \sum_{j=1}^m (w_j y_{ij} - s_i^*) (y_{ik}) = 0 \quad (7)$$

for  $k = 1, 2, 3, \dots, m$ .

The normal equations are solved in the conventional fashion to determine the optimum values of  $w_j$  and hence  $W$  by the least-squares criterion.

When the number of categories is much smaller than the number of patterns, as in the determination of the number of oxygen atoms from mass spectra, each  $s_i$  is considered in the category of the nearest meaningful value. For example, if oxygen

numbers 0, 1, and 2 are given  $s_i^*$ -values of 0, 1, and 2, respectively, an  $s_i$ -value of 1.68 is classified as meaning two oxygen atoms. However, arbitrarily selecting 1.5 as the discriminating value between one and two oxygen atoms may not give the best results. Hence, after the least-squares step, a "best line" routine is used to compute the discrimination line between the two categories that gives the best answer for the training set. On the other hand, when large numbers of categories exist, as in molecular-weight determination, the category is no longer as important as the nearness of  $s_i$  to the correct answer. Also, the "best line" calculation may be quite lengthy in cases containing many categories, such as hydrogen number, and has little value for these.

The actual time consumed by the least-squares calculation increases as the square of the dimensionality of the problem and as at least the first power of the number of patterns. For a small number of categories it is usually much slower than a series of binary dichotomizers, but because the length of the calculation is independent of the number of categories, the least-squares method becomes much more practical as the number of categories increases.

\* Infrared spectra were also studied with the basic linear learning machine.<sup>15</sup> The data had been classified into three peak "intensities" based on a resolution of 0.1  $\mu\text{m}$ . If a peak occurred in an 0.1- $\mu\text{m}$  band, it was given "intensity" 1; if it was the largest peak in a 1.0- $\mu\text{m}$  band, its "intensity" was assigned as 2; and if it was the largest peak in the spectrum, it was given "intensity" 3. (All other dimension positions then had an intensity of 0 for a given spectrum.) While this resulted in a considerable decrease of intensity information, such a decrease had been shown previously not to be a serious compression in the case of mass spectra and provided a far simpler calculation for the dot-product operation. Each pattern was, therefore, rewritten as a series of integers as follows

$$n_1, p_1, p_2, p_3, \dots, p_{n_1}, n_2, q_1, q_2, q_3, \dots, \\ q_{n_2}, n_3, r_1, r_2, r_3, \dots, r_{n_3}$$

where  $n_1$  is the number of peaks with amplitudes of 1.0 followed by the dimensions (or positions) of those peaks,  $n_2$  is the number with amplitudes of 2.0 followed by the dimensions of those peaks, and  $n_3$  is the number of peaks with amplitudes of 3.0 followed by the dimensions of those peaks.

The resulting patterns required less than one third as much storage as would have been needed for the uncompressed spectra. Furthermore, the dot-product process used to form the scalars necessary for classification could be performed as follows: If  $W$  is the weight vector and  $Y$  is the pattern for which the dot product is to be formed, then

$$W \cdot Y = w_1 \cdot y_1 + w_2 \cdot y_2 + w_3 \cdot y_3 + \dots + w_{d+1} \cdot y_{d+1} \quad (8)$$

However, if  $Y$  contains only values of 0.0, 1.0, 2.0, and 3.0, the dot product may be computed by

$$W \cdot Y = 1.0 \times \sum_{j=1}^{n_1} w_{qj} + 3.0 \times \sum_{j=1}^{n_2} w_{rj} + w_{d+1} \quad (9)$$

For the patterns used, this method of computing dot products decreased computation time by roughly a factor of twenty.

There were nine chemical classes for which there were enough compounds to form balanced training and prediction sets of three hundred compounds each, and these were used to test the learning-machine method on infrared data. The results are given in Table 2. For each chemical class two weight vectors were developed, one with all initial components set equal to +1.0 and the other with all initial components at -1.0. Com-

TABLE 2  
Training Using Even Categories

Training Set 300(150/150)  
Prediction Set 300(150/150)

Chemical class	Percent prediction for weight vector initially	
	positive	negative
Carboxylic acids	74	74
Esters of carboxylic acids	76	76
Linear amides	78	80
Ketones	73	73
Primary alcohols	69	69
Phenols	76	82
Primary amines	71	69
Ethers and acetals	63	62
Nitro and nitroso compounds	81	82

From Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilly, C. N., *Anal. Chem.*, 41, 1945, 1969. With permission.



plete training was accomplished in every case, with the highest prediction around 80%.

Another available piece of information that bears on the confidence that can be placed in the classification of any particular unknowns is the amplitude of the scalar produced by the dot product. Only the sign of the dot product is used in the binary decision maker, but its size should nevertheless be some indication of how close the pattern resembles the typical pattern for a given category. In other words, small scalars correspond to patterns that lie close to the no-decision region and are difficult to classify correctly, while those with large scalars lie far away from the decision surface, on one side or the other and are likely to be correctly classified. For carboxylic acids, the numbers of right and wrong classifications were tabulated against the scalar amplitude, and the result is given in Table 3. While the function is not smooth, probably because of the relatively small number of patterns in each range of values of the scalar, it is certainly evident that the confidence increases rapidly as the scalar moves away from the center zero value and approaches 100% for very large values.

Another interesting feature of the pattern-recognition approach is that it does not require that all of the data patterns be constructed from a single type of measurement. It is completely reasonable, in the hyperspace approach, to assign a dimension to each resolvable measurement by any of a variety of techniques. One study<sup>16</sup> combined infrared and mass spectra with melting and boiling points to gain greater predictive success than either infrared or mass spectroscopy could produce alone. No special consideration need be given to the combined data in the development of decision surfaces or in their application. Furthermore, as shown in that work, it is possible to use the dimensionality-reduction approach to produce patterns of even lower dimensionality that retain some of the characteristics of both data sets. This feature is one interesting approach to trying to determine what aspects of each technique are important.

To our knowledge, the application of linear learning machines to the semiquantitative interpretation of gamma-ray spectra from mixed samples was the first attempt to analyze mixed chemical samples by pattern recognition.<sup>17</sup> Badly overlapping sets gave rise to problems that were solved by the use of training sets in which the concentra-

TABLE 3

## Confidence Intervals for Carboxylic Acid Prediction

Calculated scalar	Number correct	Number incorrect	Total	Percent confidence
-16	1	0	1	100
-14	8	0	8	100
-12	11	1	12	92
-10	22	2	24	92
-8	21	7	28	75
-6	39	9	48	81
-4	41	13	54	62
-2	41	17	58	57
0	NEG (37)	POS (29)	66	—
+2	48	24	72	67
+4	47	27	74	63
+6	43	11	54	80
+8	27	7	34	79
+10	26	5	31	84
+12	13	4	17	76
+14	11	2	13	85
+16	5	0	5	100
+18	2	0	2	100
+20	0	0	0	—
+22	1	0	1	100

From Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilly, C. N., *Anal. Chem.*, 41, 1945, 1969. With permission.

tions of the element of interest fell into well-separated groups. Success was considerable with some elements, but quite limited with others. (This is, of course, common in activation analysis, where its nuclear properties and the resolvability of its radiations make each isotope a distinct case.) However, it is interesting to consider adding the feature of at least semiquantitative analysis to the pattern-recognition approach.

It should be realized that the previously cited work of Kowalski et al.<sup>14</sup> on least-squares classification is quantitative in a certain sense, in that the attempt was made to classify mass spectra directly according to the number of oxygen atoms in the compound. However, this is not the same as quantitative analysis in the usual sense of determining the amount of each component of a mixture.

An early application of the linear learning machine was an attempt to determine as many structural parameters as possible from mass spectra.<sup>18</sup> In this work 43 weight vectors were developed for structural properties of hydrocarbons and 65 for compounds containing oxygen and/or nitrogen. The overall predictive success was

about 90%, including the five hydrocarbon classes that did not reach complete training within a reasonable number of calculations.

This predictive information was actually assembled in an attempt to determine the total structure of the compound. However, the number of incorrect classifications was large enough in the prediction set (about 10% on the average) that there was little chance of predicting the total structure correctly even though much correct structural information could be obtained. (It is felt, however, that there would be a very good chance of predicting complete structures correctly if predictive abilities of perhaps 98 to 99% could be attained.)

Further work on mass-spectral pattern recognition has included a specific treatment of feature selection.<sup>19</sup> Feature selection is the process of selecting specially important components of the data in order to provide for the actual classification method an effectively more meaningful data set with which to operate. In this approach two different weight vectors (caused by different initializations of the starting-weight vector) were generated, and the individual components of the final vectors were compared. The signs of the individual components of the weight vectors are equivalent to votes for or against inclusion in each of the classes, and components that had differing signs were accordingly discarded. By these means the dimensionality was reduced from 119 to 31 with no loss of predictive ability. In other words, the removal of "ambiguous" *m/e* positions did not degrade the performance of the classifier.

Other attempts to improve performance in applications to mass spectra were included in a study on prediction and reliability.<sup>20</sup> Various linear transformations were applied to the input data; a linear transformation is an operation that treats each dimension independently. A "dead zone" classification method was also used, in which only those points that fell at least a minimum distance from the decision surface were classified. (This subject will be further discussed below, in connection with Reference 24.) A committee approach to a layered decision maker was also used, in which outputs of the first layer to threshold logic units to the majority-rule decision maker, were used as inputs to the second layer, etc.

Sybrandt and Perone applied the learning-

machine approach to qualitative analyses of mixtures by stationary-electrode polarography.<sup>21</sup> The limiting effects of concentration ratio, degree of peak overlap, and peak potential variation on both the overall statistical accuracy and the specific categorization procedure were investigated. The conclusion was that about as good accuracy could be obtained from severely overlapped peaks as from second-derivative measurements.

One non-linear transformation method that has been applied to mass spectra is the Fourier transform,<sup>22</sup> from which the basic information that is needed for classification can still be easily obtained. Five classifications of hydrocarbons that had not been found to converge in earlier work with learning machines converged readily with certain forms of the Fourier transform. However, prediction showed no improvement, and indeed it was worse in many cases. Thus the Fourier transform was not shown to be of particular use in mass spectrometry. However, such transformations may be more applicable to other forms of data, where phase shifts along the abscissa cause a classification problem (see Reference 23).

Pattern recognition has been applied to high-resolution nmr data for the purpose of detecting the presence of various features of molecular structure.<sup>23</sup> The training set used to calculate weight vectors for classification comprises pattern vectors derived from calculated nmr spectral frequencies and intensities. The spectra are pre-processed with the autocorrelation function to remove the translational frequency variance produced by variations of chemical shift from spectrum to spectrum. Truncation of the autocorrelation function, which is necessary to keep the pattern dimension relatively small, is possible because of a redundancy in the information. Weight vectors for ethyl, *n*-propyl, and isopropyl groups were trained by a regression procedure and tested on spectra that were "unknown" (not in the training set).

As was indicated above, prediction can be improved by adding a no-decision region. Providing a general no-decision region can also lead to considerably improved recognition and prediction in inseparable cases.<sup>24</sup> This is one example of a way in which the linear decision maker can be extended to render it applicable to linearly inseparable cases. The requirement for classification becomes

$w \cdot y > \Delta$  positive class

$w \cdot y < -\Delta$  negative class

where  $\Delta$  is positive to improve prediction in separable cases, but is negative to exclude the region of overlap in inseparable cases. One begins with a large negative value of  $\Delta$  and progressively makes it more positive until it approaches the highest value that still allows convergence.

In some cases prediction can be increased as much as 20% over the standard approach, in which  $\Delta = 0$ . Of course, the time consumed by the training calculation is increased noticeably by the necessity of seeking the best value of  $\Delta$ .

In almost all of the early applications of pattern-recognition techniques to mass spectrometry, linear threshold logic units were used. These systems were linear in that they used the mass-spectral peaks independently of one another. However, both the theory of mass spectrometry and pattern-classification considerations suggest that second-order interactions (cross terms that consider relationships between peaks) could be used to advantage in performing such classifications. Reference 25 reported on the use of a similarity measure to develop two types of cross terms (intraset and interset) from low-resolution mass spectra. It was shown to be highly probable that the interset cross terms derived in this way were correlated with the molecular features that defined the categories of classification. The method was implemented with threshold logic unit pattern classifiers derived from several sets of mass spectrometric data. The cross terms were shown to increase the power of the pattern classification systems by speeding up convergence and/or raising their predictive ability.

Figure 2 shows a mass spectrum in the form of a general labeled graph. In this symbolism the nodes  $v_1, v_2, \dots, v_i, v_j, \dots, v_p$  represent the peaks that appear in the mass spectrum, and the arrows that connect the nodes (and that are also called edges) represent the reaction pathways giving rise to the fragment ions. Thus,  $v_1$  in Figure 2 represents the parent, or molecular, ion from which all other ions are derived. For a graph with labeled nodes an adjacency matrix  $A = [a_{ij}]$  can be formed. If there are  $p$  nodes, such a graph is a square symmetric  $p$ -by- $p$  matrix. Its components are derived as follows:  $a_{ij} = 1$  if the node  $v_i$  is adjacent to the node  $v_j$  in the graph, and  $a_{ij} = 0$

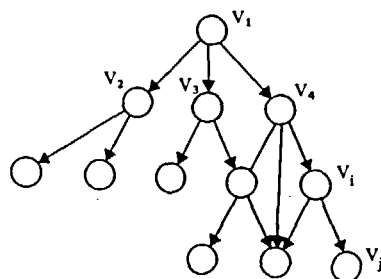


FIGURE 2. A generalized acyclic, directed graph. (From Jurs, P. C., *Appl. Spectrosc.*, 25, 483, 1971. With permission).

otherwise. The  $p$ -by- $p$  adjacency matrix is a complete, unambiguous representation of a graph with  $p$  nodes.

Many types of data (such as the mass-spectral data of interest) can be expressed in vector form. That is, each mass spectrum in a data set can be expressed as  $X = (x_1, x_2, \dots, x_p)$ , where each component of the vector corresponds to one peak in the spectrum, so that  $x_{31}$  represents the intensity of the peak at  $m/e = 31$ . From a set of such vectors representing a data set, two quantities can be calculated:  $b_i$ , the number of vectors containing a non-zero term  $x_i$ , and  $b_{ij}$ , the number of vectors having non-zero values for both  $x_i$  and  $x_j$ . For example, with a mass-spectral data set containing 100 vectors,  $b_{1,5}$  would equal 50 if  $x_{1,5}$  were non-zero for half of the vectors. A value of 40 for  $b_{15,30}$  would mean that 40 vectors have peaks at both  $m/e = 15$  and  $m/e = 30$ .

These quantities are used to calculate the components  $c_{ij}$  of a term-term similarity matrix. If the vectors in the data set are all of length  $p$ , then the term-term similarity matrix is a square  $p$ -by- $p$  matrix. Each element is computed as follows:

$$c_{ij} = b_{ij} / (b_i + b_j - b_{ij}). \quad (10)$$

Each element  $c_{ij}$  represents the degree to which the  $i$ th and  $j$ th components of the collection of vectors are related. This measure of similarity is especially appealing for use with low-resolution mass spectra because it deals explicitly with the locations of components in the vectors rather than with their amplitudes.

The term-term similarity matrix constructed with Equation (10) consists of elements  $c_{ij}$  between zero and one, where larger numbers indicate increased relatedness. This matrix can be converted into an adjacency matrix by comparing

each  $c_{ij}$  with a threshold value  $T$  and setting  $c_{ij} = 1$  if  $c_{ij} > T$  and  $c_{ij} = 0$  otherwise. The number of non-zero elements of the resulting adjacency matrix can be investigated as a function of the threshold value. Each 1 appearing in the adjacency matrix developed by the thresholding process corresponds to a single cross term that appears in the data set often enough to exceed the threshold value. Such cross terms might be useful features of the data for the threshold logic units to use in classifying the data. Features developed in this manner are clearly intraset features, because the entire set of vectors has been used together.

An approach for developing interset features is also based on the term-term similarity matrix. In this method the set of available data is split into the two subsets that the pattern classifier will be trained to detect. Using Equation (10), similarity matrices are then formed for each of the two subsets, yielding matrices with terms  $c_{ij}^1$  and  $c_{ij}^2$ . The absolute values of the differences between corresponding terms in these two matrices

$$\Delta c_{ij} = |c_{ij}^1 - c_{ij}^2| \quad (11)$$

are then taken to obtain a third matrix having terms  $\Delta c_{ij}$ , so that the magnitude of each element expresses the difference between the similarity measures of the cross terms formed from the  $i$ th

and  $j$ th components in the two subsets of the data. This method of choosing interset features has been applied to low-resolution mass spectra in choosing cross terms to be incorporated into the set of features given to the pattern classification system.

For a set of 630 low-resolution mass spectra a term-term similarity matrix was developed using Equation (10). When different threshold values  $T$  were applied to the similarity matrix, the number of cross terms (or edges) present in the data set depended on the value of  $T$  as shown in curves a and b of Figure 3. The uppermost labeling of the x axis was used in plotting curves a and b. The top curve refers to cross terms using all combinations of  $m/e$  positions, while the bottom curve, which has the same general shape, refers to the cross terms formed from two  $m/e$  positions that differ by more than 10 units. With 119  $m/e$  positions there are  $(119)(118)/2 = 7021$  possible second-order cross terms, so the 1246 cross terms with thresholds of  $T > 0.2$  are only a fraction of all the possibilities.

The curves on the right-hand side of Figure 3 show how the number of nodes in the largest cluster depends on the threshold value. A cluster is defined as a group of nodes that are completely interconnected by edges; any node in a cluster can be reached from any other node in the same cluster by traversing a series of edges. It is seen

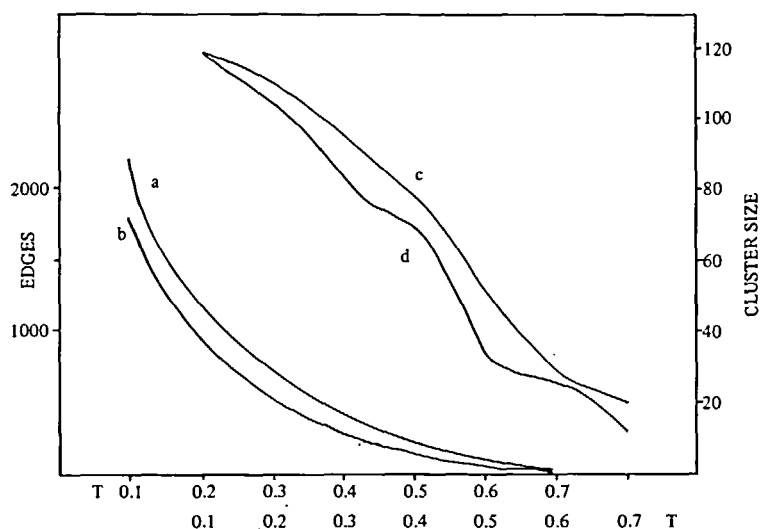


FIGURE 3. Dependences of number of total edges and sizes of largest clusters on threshold for mass spectra with 119  $m/e$  positions. (a) All  $m/e$  positions; (b)  $m/e$  positions differing by at least 10 units; (c) all  $m/e$  positions; (d)  $m/e$  positions differing by at least 10 units. (From Jurs, P. C., *Appl. Spectrosc.*, 25, 483, 1971. With permission.)

that the number of nodes in the largest cluster decreases as the threshold value  $T$  increases.

In order to find interset cross terms that are correlated with two chemical classes, the method described above for finding interset classifier tests was applied to a data set of 450 spectra, each having 132  $m/e$  positions. Three sets of tests were run with the three classifications used for Table 4. To start each test, a linear pattern classifier was trained, using all 132  $m/e$  positions. A randomly selected training set of 250 and a prediction set of 200 were used. The weight-sign feature-selection routine was used to reduce the pattern-space dimensionality by discarding  $m/e$  positions not helpful for the classification being performed. This is done without degrading either the convergence rate or the predictive ability. Then the cross terms, chosen as described above, were inserted in the vacated positions, the classifier was trained again, and the predictive ability was checked. The results of this investigation are shown in Table 5. All tests were run in triplicate, with three different random training sets labeled A, B, and C. A dead zone of 50 was used during all training and for all prediction. The intensities of all peaks underwent a logarithmic transformation. Cross-term intensi-

ties were computed by multiplying the intensities at the two  $m/e$  positions involved and taking the square root of the result to maintain normalization.

Table 5 presents the results of this procedure for oxygen-presence determination. The pattern classifier is being trained to detect whether the compound whose mass spectrum is being classified contained oxygen in any functional form. The first column labels the triplicate runs by training set. The second and third columns give the number of linear  $m/e$  positions and the number of cross terms used by the pattern classifier. The fourth column gives the number of feedbacks required to attain 100% recognition; each of the two values refers to one of the duplicate runs with different weight vector initialization (first all +1's, then all -1's). The fifth column gives a measure of the improvement in convergence rate obtained when the cross terms are included; it is equal to the ratio of the total number of feedbacks used during training with the cross terms to the total number of feedbacks used during training with only linear terms; thus, a value of 1.0 means the convergence rate was unchanged; a value of 0.5 means that convergence was twice as fast with cross terms as

TABLE 4  
Intersect Cross Terms

Threshold	Oxygen presence (120)	Oxygen absence (165)	H>14 (164)	H≤14 (286)	2C -<1 H (141)	2C -≥1 H (309)
1.0	2	0	10	0	21	0
0.95	28	65	283	15	81	21
0.90	69	137	559	38	131	68
0.85	136	273		79	194	124
0.80	220	432		125	292	242
Number of cross terms investigated	220	220	200	200	292	242
Number of cross terms selected	13	14	11	10	10	11
Largest $\Delta a_{ij}$	0.769	0.688	0.498	0.652	0.551	0.367
Average $\Delta a_{ij}$	0.738	0.658	0.456	0.560	0.514	0.313
Smallest $\Delta a_{ij}$	0.701	0.613	0.413	0.393	0.501	0.301
Largest $a_{ij}$	0.886	0.912	0.987	0.903	0.932	0.852
Average $a_{ij}$	0.818	0.898	0.972	0.791	0.914	0.826
Smallest $a_{ij}$	0.805	0.881	0.968	0.753	0.901	0.814

From Jurs, P. C., *Appl. Spectrosc.*, 25, 483, 1971. With permission.

TABLE 5  
Oxygen-presence Training

Training set	Number of m/e positions	Number of cross terms	Number of feedbacks	Convergence improvement	Predictive ability	Average predictive ability	Number not predicted	Number of discarded features
A	95	—	157/151	0.54	96.9/98.5	97.7	6/8	11
	95	27	79/89		95.4/95.4	95.4	3/3	19/1
B	95	—	186/164	0.77	98.4/97.9	98.1	10/9	12
	95	27	147/122		98.0/96.9	97.5	5/4	14/1
C	95	—	284/320	0.58	98.9/98.4	98.7	11/8	13
	95	27	173/167		99.5/99.0	99.3	5/10	49/0

Training set	Training set			Prediction set		
	+	—	+	+	—	—
A	74	176	46	154		
B	68	182	52	148		
C	73	177	47	153		

From Jurs, P. C., *Appl. Spectrosc.*, 25, 483, 1971. With permission.

without them. The sixth column gives the predictive abilities exhibited by the two pattern classifiers, and the seventh gives their average. The eighth column gives the number of patterns (out of 200 in the prediction set) that were not classified because their dot products fell within the dead zone. The ninth column gives the number of features that were in the patterns but were discarded by the feature-selection routine after training. (Features are discarded for which the corresponding weight vector components of the two weight vectors disagree in sign.) For the training using cross terms, the two figures in the ninth column refer to the numbers of linear and cross terms discarded, respectively. The training and prediction set populations are given at the bottom of the table.

For the oxygen-presence training the linear feature-extraction routine reduced the number of m/e positions from 132 to 95. With 95 m/e positions, the pattern classifier quickly converged to perfect recognition and displayed predictive abilities of 97.7%, 98.1%, and 98.7%. With the addition of 27 cross terms (those discussed above in connection with Table 4), the performance of the pattern classifiers changed. The convergence-improvement figure ranges from 0.54 to 0.77 for the oxygen-presence determination tests. This major increase in convergence rate demonstrates that there is a strong correlation between the cross terms that were selected and the presence or absence of oxygen. The predictive ability was only slightly affected for training sets B and C, and it dropped slightly for A. The eighth column shows that in each case the pattern classifiers attempted more predictions when using the cross terms than when using only linear terms. The ninth column shows that in each case more linear features of the patterns were discarded by the feature-selection routine for the training using cross terms than for the linear training. Nearly all the cross terms were found to be useful by the feature-selection routine.

Tests similar to the one shown in Table 5 with cross terms selected by the experimenter or on the basis of  $c_{ij}$  values (intrasets cross terms) showed that such cross terms were not helpful to the pattern classifier. Thus, it has been shown that selection of intersets cross terms on the basis of  $\Delta c_{ij}$  values yield second-order features that are useful to the pattern classifier.

Frew et al. approached the linear inseparability

problem by developing a piecewise-linear multi-category pattern classifier that adds additional subclasses as necessary to resolve the training set.<sup>26</sup> They chose a number of categories  $R$  relating to the information sought in the data and assumed that the members of each category are arranged together in the pattern space. The actual distribution for each of the  $R$  categories may be quite complex and may exhibit a number of local maxima corresponding to the subcategories  $R, L_i, (i = 1, \dots)$ . A simple two-dimensional example, which excludes the  $(d + 1)$ th dimension, is depicted in Figure 4, where  $R = 2, L_1 = 3$ , and  $L_2 = 1$ .

The classification problem is then to develop a set of discriminant functions,  $S_i$ , which in this case are piecewise-linear functions that describe  $d$ -dimensional hyperplanes segregating the members of a given category from those of all other categories. The most extensively studied piecewise discriminant functions have the form

$$S_i = \max_{j=1, \dots, L_i} \{(S_i^{(j)})\} \text{ for } i = 1, \dots, R \quad (12)$$

where the  $S_i^{(j)}$  are known as subsidiary discriminant functions, defined by

$$S_i^{(j)} = f_i^{(j)}(X) = w_{i1}^{(j)}X_1 + w_{i2}^{(j)}X_2 + \dots + w_{id}^{(j)}X_d + w_{i,d+k}^{(j)}X_{d+1} \quad (13)$$

Using such an arrangement, a pattern is assigned to category  $k$  if  $S_k$  has the largest value of all the discriminant functions,  $S_i, i = 1, \dots, R$ .

An iterative training process using an error-correction method adjusts the individual weights that are included in the weight  $W_i^{(j)}$  until all patterns are correctly classified. Suppose that a pattern belonging to the  $k$ th category is presented to the classifier and that  $S_i$  has a larger value than that of any other of the discriminant functions  $S_i$ . The generalized error-correction procedure is then

$$\begin{aligned} W'_k &= W_k + c \cdot Y \\ W'_i &= W_i - c \cdot Y \end{aligned} \quad (14)$$

where  $c$  is a positive correction increment. Several rules for choosing  $c$  have been used.

While the number of categories  $R$  is usually known when formulating the classification problem, the number of subcategories,  $L_i$ , is in general unknown. The conventional method for

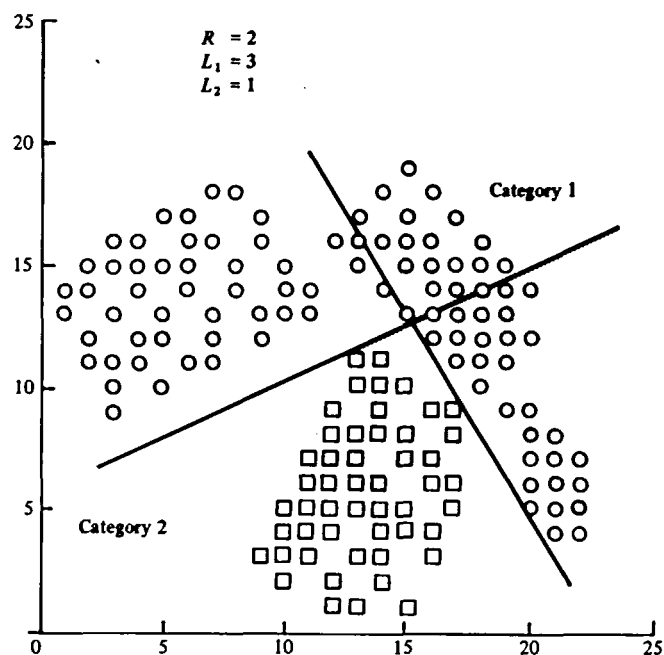


FIGURE 4. Two-dimensional classification problem with  $R = 2$ ,  $L_1 = 3$  and  $L_2 = 1$ . (From Frew, N. M., Wangen, L. E., and Isenhour, T. L., *Pattern Recognition*, 3, 281, 1971. With permission.)

dealing with this uncertainty is to assign an adequate but fixed number of weight vectors to each category in advance of the actual training process. Although such a procedure leads to a solution, it allows for the possibility that the number of vectors will exceed the number actually required; if it does, unnecessary calculation will be introduced both in the training process and in classification tasks after training. In addition, the use of an excessive number of weight vectors can produce "overfitting" of the data distribution of the training set, which leads subsequently to poor performance on "unknown" patterns.

In designing the present classifier, the intention was to provide a simple means for internal generation of new weight vectors as dictated by the training process itself; thus, hopefully, the pattern classifier would evolve to the level of complexity that just sufficed to provide the described solution. Any pattern classifier, having a given complexity and using an error-correction training procedure, will exhibit an oscillatory behavior when presented with an insoluble problem; for example, a classifier using a single linear discriminant function would never converge to a solution for the problem illustrated in Figure

4, but the orientation of the single decision surface would instead, undergo major oscillations for an indefinite period. It is logical, then, to look for means of detecting such oscillations, which would indicate the need for a more complex decision surface.

The method proposed here involves periodic evaluation of a function based on the following quantity:

$$\sum_{m=1}^M \frac{(S_k - S_l)}{M} \quad (15)$$

Here  $M$  is the total number of patterns in the training set; it is understood that  $k$  represents the category of the pattern  $m$  and that  $S_l$  is the largest of all the discriminant functions  $S_i, i \neq k$ . Thus, if the pattern  $m$  is correctly classified, it will contribute a positive term to the sum, whereas that term will be negative if the pattern is misclassified. Although no rigorous argument can be offered to support the use of the above quantity or to indicate how its absolute magnitude should vary during training, relative changes in its value have been found empirically to reflect oscillations in the decision surface and were deemed worthy of investigation.



The actual function evaluated had the following form:

$$P_t = \frac{1}{T}[P_{t-1}] + \sum_{m=1}^M \frac{(S_k - S_t)}{M} \quad (16)$$

In a multicategory application, the piecewise classifier was trained on carbon/hydrogen ratio ( $2m + 2$ ,  $2n - 2$ ,  $2n - 4$ ,  $2n - 6$ ), and its results were compared with those obtained with a binary classifier (Table 6). The overall predictive ability of the multicategory classifier averaged 97.2% for five trials; when broken down into individual categories, the prediction ranged from 98.3% for ( $2n + 2$ ) compounds to 92.4% for ( $2n - 4$ ) compounds. By comparison, the predictive performance obtained by independently training five binary classifiers for the specific C/H ratios was poorer, particularly for those categories containing only a small number of patterns. For example, while a binary vector trained to recognize ( $2n - 4$ ) compounds as distinguished from all others yields an overall prediction of 96.1%, the percentage of patterns actually having the ( $2n - 4$ ) ratio that were correctly classified was only 65.3%. Thus, given the spectrum of an "unknown" ( $2n - 4$ ) compound, the multicategory classifier would provide a decision that was nearly 30% more reliable than the decision provided by the binary classifier. It is likely that a bias toward the larger category arises in all binary cases where the number of patterns for the two categories is extremely unequal.

A further study of the use of the Fourier transform for preprocessing mass spectra has appeared.<sup>27</sup> The data set employed consisted of 450 low resolution mass spectra of 132 m/e positions each. The sequence of computations applied to the original mass spectra was as follows. The original spectra had peaks up to m/e 200, and the pattern vectors to be transformed were therefore 200-dimensional. The intensities of these peaks were logarithmically transformed. The Fourier transform of each spectrum was taken with the fast Fourier transform algorithm (FFT), resulting in a 256-dimensional complex vector. Only the real part of the Fourier transform vector was saved; because the 256 components of the real part of the Fourier transform vector exhibit symmetry about the midpoint of the vector, only the first half of the vector was saved. Thus the Fourier transform pattern had 128 components,

which was approximately equal to the number (132) of m/e positions that were used. Of course, there are many more components with zero intensity in the mass-spectral patterns than in the Fourier transform patterns.

The first study undertaken in this investigation was to perform feature selection on the Fourier transformed spectra and to investigate how the capabilities of the pattern classifiers depended on the number of descriptors used. As has been pointed out, discarding some of the descriptors from the Fourier transform patterns is equivalent to losing some information about all of the original mass spectra, rather than all information about some of the patterns.

Table 7 shows the results obtained when feature selection was performed on the Fourier transformed spectra. The chemical question being trained for was whether the number of hydrogen atoms in the molecule did or did not exceed twice the number of carbon atoms in the molecule. That is, one category comprises the compounds with more than twice as many hydrogen as carbon atoms, while the other category comprises those for which the number of hydrogen atoms is equal to or less than twice the number of carbon atoms. The first category includes alkanes, alkyl amines, and others; the second includes alkenes, ketones, and aromatics, among others. This is evidently a characteristic of a low resolution mass spectrum that is not immediately apparent.

The first column of Table 7 labels the three parallel runs according to the training set employed. Each training set consisted of 250 spectra randomly chosen from the overall data set of 450 spectra. The remaining 200 spectra in each case were used as the prediction set. The second through fifth columns give several measures of the performance of the pattern classifiers achieved before feature selection. Each Fourier transform pattern has 124 descriptors initially, for the first four descriptors have enormous values and do not contribute to the overall ease of solution of the problem. The training was performed with the threshold  $Z = 10$ ; between 600 and 1,000 feedbacks were required to obtain pattern classifiers that could correctly classify all the numbers of the training sets. Two binary pattern classifiers were trained for each training set; the exact numbers of feedbacks required for training in each case are given in the third column as two numbers pertaining to the two different weight vectors and

separated by a slash. The fourth column gives the number of spectra in the prediction set that were not classified because the scalar fell in the dead zone. The fifth column gives the average percent prediction for the two binary pattern classifiers. It varies somewhat among training sets. The sixth through ninth columns give the capabilities of the pattern classifiers obtained after extensive feature selection. For the three training sets there remained only 76, 70, and 78 of the original 124 descriptors, respectively. The numbers of feedbacks needed to converge to perfect recognition of the training sets were approximately the same as, or somewhat less than, in the previous case. Predictive abilities for the pattern classifiers employing the small descriptor lists are substantially the same as for the complete Fourier transformed spectra. Hence, feature selection has not degraded the performance of the pattern classifiers, but has only speeded up the computations.

Table 8 presents a comparison of the capabilities of pattern classifiers that use spectra or Fourier transformed spectra as their pattern vector inputs, and refers to training for the same chemical classes as Table 7. The left-hand half of Table 8 shows that pattern classifiers trained with mass-spectral patterns containing 111 feature-selected *m/e* positions converge to 100% recognition with 1,200 to 2,000 feedbacks and display predictive abilities of 93 to 95% on complete unknowns. The right-hand half of the table gives the performances of the pattern classifiers trained with Fourier transformed spectra. The Fourier transformed spectra consist of 76, 70, and 78 descriptors per pattern, respectively, for the three training sets. Convergence occurs with 600 to 1,000 feedbacks, only about half as many as were necessary in the previous case. The predictive abilities achieved are better in each of the three cases; the increases are 0.4, 0.4, and 1.0%. Thus, the capabilities of the classifiers dealing with the selected Fourier spectra compare favorably with the classifiers dealing with the original data.

The remainder of the investigation involved testing the reliability of threshold logic units trained with the original mass spectra and the Fourier transformed spectra. It was shown that binary pattern classifiers trained with the Fourier transformed spectra were more reliable.

Sybrandt and Perone further applied pattern recognition to the classification of strongly over-

lapping peaks in stationary-electrode polarography.<sup>21</sup> Feature selection, which is discussed in other work in this review, was used to aid the classification procedure. Parameters derived from zeroth, first, and second derivative polarograms were used for the successful classification of uncomplicated reversible systems. In some cases the peaks overlapped so closely that the derivative data could not be interpreted visually and subjectively. A predictive accuracy of about 90% or better was realized for polarograms for which the peak separations were between 6 and 14 mV, the *n*-values between 1 and 3, and the ratios of peak heights between 1:1 and 20:1.

In the first application of pattern recognition to NMR data, classifications were made by a learning machine trained by a regression procedure.<sup>23</sup> The *K*-nearest-neighbor (KNN) rule has been applied<sup>28</sup> to the direct determination of molecular-structure information from NMR data.

The KNN rule is conceptually and computationally quite simple. An unknown pattern is classified according to its nearness to *K* training-set patterns. The *K* patterns are the nearest ones to the unknown pattern as measured by a distance function, such as the *n*-space Euclidean distance  $d_{ij}$  between patterns  $X_i$  and  $X_j$ , where

$$d_{ij} = \left[ \sum_{k=1}^n (X_{ik} - X_{jk})^2 \right]^{1/2} \quad (17)$$

The procedure can be lengthy, because the distance must be found from the unknown pattern to every pattern in the training set, but this is its only disadvantage. Once the *K* neighbors are found, the unknown pattern is classified according to their majority vote. In other words, the predominating category amongst the *K*-nearest neighbors is assigned to the unknown pattern.

The procedure has several advantages, not the least of which is its firm theoretical foundation. The risk of making an incorrect classification is bounded and can be no worse than twice the optimal Bayes risk. The KNN is also a multiclass method and is not limited to data that are linearly separable.

The origins of obsidian artifacts have been identified for archaeological purposes<sup>29</sup> by using pattern recognition to interpret trace elemental concentrations measured on each artifact. Samples of obsidian, a volcanic glass, were collected from four known quarries in Northern California and

TABLE 6

Multicategory Application: Carbon/Hydrogen Ratio; Comparison by Category with Binary Classifier

Overall	Multicategory		Composition of prediction set	Binary question	Overall	Binary		Composition of prediction set	
	By category					By category			
								Positive	Negative
97.2 (average of 5 trials)	2n + 2	98.3	45/187	2n + 2 vs. other	97.4	95.4	98.0	44	143
	2n	97.1	73/187	2n vs. other	94.6	96.4	94.1	74	113
	2n - 2	98.2	34/187	2n - 2 vs. other	96.2	92.8	97.1	37	150
	2n - 4	92.4	10/187	2n - 4 vs. other	96.1	65.3	96.9	7	180
	≤2n - 6	95.9	25/187	≤2n - 6 vs. other	98.0	88.6	98.8	16	171

From Frew, N. M., Wangen, L. E., and Isenhour, T. L., *Pattern Recognition*, 3, 281, 1971. With permission.

TABLE 7

Feature Selection of Fourier Transform Patterns

Training set	Number of descriptors	Number of feedbacks	Average number not predicted	Average % prediction	Number of selected features	Number of feedbacks	Average number not predicted	Average % prediction
A	124	575/752	21	95.3	76	627/629	23	95.5
B	124	727/670	25	93.4	70	678/817	21	93.6
C	124	952/1063	18	94.6	78	903/1003	17	94.5

From Jurs, P. C., *Anal. Chem.*, 43, 1812, 1971. With permission.

TABLE 8

Comparison of Properties of Pattern Classifiers Using Mass Spectra and Fourier Transform Spectra

Training set	<i>m/e</i> positions	Mass spectra			Average % prediction	Number of descriptors	Fourier transform spectra			Average % prediction
		Number of feedbacks	Average number not predicted				Number of feedbacks	Average number not predicted		
A	111	1155/1666	17		95.1	76	627/629	23	95.5	
B	111	1386/1451	19		93.2	70	678/817	21	93.6	
C	111	1752/2020	19		93.5	78	903/1003	17	94.5	

From Jurs, P. C., *Anal. Chem.*, 43, 1812, 1971. With permission.

analyzed by X-ray fluorescence for ten trace elements. Several artifacts were also collected and analyzed for the same trace elements. Pattern recognition methods were then used to develop classification rules, using the quarry data in order to classify the artifacts as having come from one of the quarry sites. Autoscaling<sup>3</sup> was used to weight each of the measurements equally, and four independent pattern-recognition methods were then used to accomplish the main objective. When cluster analysis was performed, an unexpected but very welcome result was noted. One of the artifacts was quite unlike any of the four sources, suggesting that the material came from outside the region covered by the study. These results were of considerable interest to the archaeologists involved in the study.

The problem of class construction for binary-pattern-classifier implementation was discussed in theoretical terms by Lytle.<sup>30</sup> He discussed the arrangements that could be used for multicategory classifications and showed how arrays of threshold logic units might be arranged to introduce a degree of redundancy into the decision-making process. Several versions of Hamming codes, with and without self-correcting capabilities, were introduced and applied to the problem of carbon-number classification. It was shown that the use of a self-correcting Hamming code made it possible to obtain a correct result from an array even though one of that array's threshold logic units might be in error. The method would also allow organizing the training set into groups of equal size, although it was not made clear whether such training sets would be linearly separable.

Another approach to the problem of linearly inseparable sets is the complex-valued nonlinear discriminant function (CNDF) applied by Justice et al.<sup>31</sup> This employs a generalized Walsh transform in constructing the discriminant function. For a first-order generalized Walsh transform, in which the spectra are allowed 50 integral values of intensity ranging from 0 to 49, each dimension of the pattern or spectrum is transformed by the relation

$$T(I) = (e^{2\pi i \sqrt{-1}/50}, I) \quad (18)$$

where  $I$  is the intensity at each mass position in the spectrum. The vector  $\Phi(x)$  is defined in such a way that it represents the transforms of all intensities (dimensions) in the mass spectrum:

$$\Phi(x) = (T(I_1), T(I_2), \dots, T(I_n)) \quad (19)$$

A discriminant function using  $\Phi(x)$  can be constructed to have the form

$$F(x) = \theta + W^* \Phi(x) \quad (20)$$

The vector sum  $W$  of all  $\Phi(x)$  of the training spectra is given by

$$W = \frac{W_A}{a} - \frac{W_B}{b} \quad (21)$$

where  $a$  and  $b$  are the numbers of spectra in categories A and B, respectively, and  $W_A$  and  $W_B$  are the weight vector components resulting from vector summations of the transformed spectra of the compounds in categories A and B, respectively. On this basis the quantity  $W^*$  in Equation 20 is the conjugate transpose of  $W$ , obtained by changing the sign of the imaginary part of the complex number. Since  $W$  is a vector, transposing it involves no actual operation, but maintains mathematical uniformity with matrix notation.  $\theta$  is a constant related to the relative sizes and variances of the training-set categories A and B.  $F(x)$  is then a complex number and is nonlinear with respect to the components of the mass spectrum. For prediction, the compounds for which the real part of  $F(x)$  is positive are classified in one category and those for which it is negative are put in the other.

From Equation 20 it is seen that evaluation of the discriminant function requires the calculation of  $\Phi(x)$  and  $W^*$ . Since the transformed spectral intensities can assume only the 50 values given by Equation 1, these intensities may be calculated and stored for use in calculating the decision surface, rather than being recalculated for each new spectrum, thereby greatly reducing computation time.  $W$  is calculated directly by Equation 21.

For the interpretation of mass spectra, the CNDF is constructed by transforming the mass spectra of the compounds in each of two classes according to Equation 1. Considering one of the training-set classes to be positive and the other negative, the transformed spectra are summed algebraically to form a weight vector  $W$ , which may then be used in conjunction with Equation 20 to predict the category that contains a compound not included in the training set. For example, if the negative class consists of compounds whose

molecular formulas are given by  $C_nH_{2n}$  and the positive class consists of all other compounds, a compound for which the calculated value of  $F(x)$  is less than zero is predicted to have the molecular formula  $C_nH_{2n}$ . In this manner predictions were made on 630 compounds, using the categories listed in Table 9.

In some cases the predictive abilities of the CNDF were comparable to those of the linear learning machine; in other cases they were considerably higher. Furthermore, the training set is limited only by the size of the data set and can be easily revised as new data are added.

Most training methods for finding useful decision surfaces are of two types: error-correction iterative methods, or methods using fitting through functions. A different training procedure of the latter type was described in Reference 32.

The training method developed in this work was an iterative least-squares method. The patterns were fit to a non-singular function that was linear with respect to its parameters. There are many ways by which this can be done, and the one used here was the linearization or Taylor series method. It uses the results of linear least squares in a succession of stages.

The nonlinear function used in this work is the hyperbolic tangent, chosen because it is well suited for pattern dichotomizers. The value of  $Y_i$  is set equal to +1 if the  $i$ th pattern belongs to category 1, and -1 if the  $i$ th pattern belongs to category 2. Then a weight vector  $W$  must be found so that the discriminant function  $g(X_i)$  will be positive if  $Y_i = +1$ , and negative if  $Y_i = -1$ . The components of the weight vector are developed by means of a linear least-squares technique. The function negative through which the patterns are fitted is  $F(S_i) = \tanh(S_i)$ , where  $S_i$  is the scalar product for the  $i$ th pattern,  $S_i = W \cdot X_i$ . According to the least-squares principle, then, the function to be minimized becomes

$$Q = \sum_{i=1}^N [Y_i - F(S_i)]^2 \quad (22)$$

where  $N$  is the number of patterns used as the training set. The problem is to find the weight vector  $W$  that minimizes  $Q$ .

In order to set up an iterative procedure with  $S_i^0 = W^0 X_i$  as the starting condition,  $\tanh(S_i)$  is expanded in terms of the Taylor series up to and including the first derivative. This expansion contains most of the value of the function. Hence,

$$\tanh(S_i) = \tanh(S_i^0) + \sum_{j=1}^{d+1} \frac{\partial \tanh(S_i)}{\partial w_j} \bigg|_0 dw_j \quad (23)$$

which in terms of the chain rule gives

$$\tanh(S_i) = \tanh(S_i^0) + \sum_{j=1}^{d+1} \text{sech}^2(S_i^0) x_{ij} dw_j \quad (24)$$

in which  $i$  stands for the  $i$ th pattern and  $j$  stands for the  $j$ th component. Hence, the function to be minimized now assumes the form

$$Q = \sum_{i=1}^N [Y_i - \tanh(S_i^0) - \sum_{j=1}^{d+1} \text{sech}^2(S_i^0) x_{ij} dw_j]^2 \quad (25)$$

The minimum for one iterative step is achieved for all such vectors  $dW = (dw_1, \dots, dw_{d+1})$  that satisfy the minimization principle

$$\frac{\partial Q}{\partial w_k} = 0 \text{ for } k = 1, \dots, d+1. \quad (26)$$

Hence

$$0 = -2 \sum_{i=1}^N [Y_i - \tanh(S_i^0) - \sum_{j=1}^{d+1} \text{sech}^2(S_i^0) x_{ij} dw_j] \text{sech}^2(S_i^0) x_{ik} \quad (27)$$

which can be expressed in terms of linear equations:

$$\sum_{i=1}^N (Y_i - \tanh(S_i^0)) \text{sech}^2(S_i^0) x_{ik} = \sum_{j=1}^{d+1} \sum_{i=1}^N \text{sech}^4(S_i^0) x_{ij} x_{ik} dw_j \quad (28)$$

In terms of matrices these equations become

$$A(sW) = b \quad (29)$$

in which

$$a_{jk} = \sum_{i=1}^N \text{sech}^4(S_i^0) x_{ij} x_{ik} \quad (30)$$

and

$$b_k = \sum_{i=1}^N [Y_i - \tanh(S_i^0)] \text{sech}^2(S_i^0) x_{ik} \quad (31)$$

so that the system

$$A(dW) = b \quad (32)$$

must be solved for the solution vector  $dW$ .

The matrix  $A$  is real, symmetric, and positive definite. Hence it is non-singular and will have a unique non-trivial solution. This solution can be

TABLE 9

## Predictive Ability of CNDF

	Cutoff	Positive category <sup>a</sup>	Negative category	% in larger category	$\theta$	% Prediction, no normalization	% Prediction, sum normalization
Oxygen	1	456	174	72.4	-0.622	87.6	88.3
	2	544	86	86.4	-0.326	90.0	90.0
Carbonyl	1	554	76	87.9	-1.614	87.9	88.7
Nitrogen	1	549	81	87.1	-1.111	88.3	90.3
Amine	1	572	58	90.8	-1.422	91.4	92.9
-C=C-	1	389	241	61.8	-0.429	80.0	82.5
	2	522	108	82.9	-1.466	94.9	95.2
	3	555	75	88.1	-2.014	98.3	98.3
$C_nH_{2n}$	—	476	154	75.6	-1.955	91.6	96.8
$C_nH_{2n+2}$	—	541	89	86.0	-1.777	95.9	95.6
Methyl	1	87	543	82.8	1.022	87.5	88.7
	2	523	107	83.1	-0.340	86.5	86.0
Ethyl	1	341	287	54.5	-0.148	71.4	77.1
Phenyl	1	568	62	90.2	-2.311	96.5	96.8
Carbon	5	105	525	83.4	0.458	91.6	92.1
	6	183	447	71.0	0.177	84.0	86.4
	7	279	351	55.7	0.088	77.3	85.2
	8	357	273	56.7	0.014	84.3	88.1
	9	446	184	70.8	0.192	86.7	85.4
	10	544	86	86.4	0.177	90.5	90.5
Hydrogen	9	135	495	78.6	0.311	85.6	87.8
	11	225	405	64.3	0.148	80.2	81.6
	13	317	313	50.3	-0.888	78.9	77.8
	15	425	203	67.5	-0.800	85.6	84.0
	17	501	129	79.5	-0.666	87.8	87.0
	19	554	76	88.0	-0.844	92.1	92.1

<sup>a</sup>Positive category contains compounds whose number of functional groups is less than the cutoff.

From Justice, J. B., Jr., Anderson, D. N., Isenhour, T. L., and Marshall, J. C., *Anal. Chem.*, 44, 2087, 1972. With permission.

obtained by any method, either direct or iterative, that is suitable for solving a set of linear equations.

Since  $F(S_i)$  is dependent upon the value of the desired solution  $W$ , the solution vector  $W^{(1)}$  starting from  $W^{(0)}$  cannot be accurate. The linearization method requires setting  $W^{(1)} = W^{(0)} + dW^{(1)}$ , and then beginning a second iteration. The process is repeated until  $W^{(1+1)} = W^{(1)} + dW^{(1+1)}$  satisfies the condition

$$|dW^{(1+1)}|/|W^{(1+1)}| < \epsilon. \quad (33)$$

That is, the ratio of the norm of  $dW^{(1+1)}$  to the norm of  $W^{(1+1)}$  is less than some arbitrarily chosen small quantity  $\epsilon$ . When this condition is met, the system has become self-consistent, there can be no benefit gained by further computation, and the method terminates.

In this study the domain of  $F$  was limited to the closed interval  $(-2.5, +2.5)$ . The range of the function  $F(S_i)$  for this interval is  $(-0.987, +0.987)$ . The initial values of the weight vector were chosen in a way to insure that  $|S_i| < 2.5$ . When this condition is met, the rate of convergence is increased, because the changes occurring in the components will not be large, and therefore  $W$  will not be subject to large errors.

Many direct methods for the solution of sets of linear equations are available. The method used in this study consisted of Gauss-Jordan elimination with complete pivoting. The subroutine evaluated the inverse of the matrix, the determinant, and the solution vector.

The iterative least-squares method was applied to the test problem of identifying the presence of oxygen in small organic molecules, which had been studied previously. By combining weight-sign feature selection with the error-correction-feedback training method, the number of features was reduced from 132 to 31 with complete recognition and a prediction percentage of 93.9%.

For this problem  $Y_i$  was set to +1 if the  $i$ th spectrum contained oxygen, and to -1 if the  $i$ th spectrum did not contain oxygen. The weight vector components were initialized so that if  $w_i$  had the value  $p$ , the  $w_{i+1}$  was assigned the value  $-p$ . In this problem the value of  $w_i$  was set to +0.01 or to -0.01. These initial values were found to be adequate in maintaining the values of the scalar products within the desired interval  $[-2.5, +2.5]$ . Minimizing the distance between the

clusters has been found to facilitate convergence.

Table 10 gives a list of the average values of several weight vector components after training under different conditions. If the average value of the weight vector component is positive for a particular  $m/e$  position, the appearance of a peak at that position must be positively correlated with the presence of oxygen, while a negative value of the weight vector component signifies that the appearance of a peak at that  $m/e$  position is positively correlated with the absence of oxygen. The correlations obtained through the least-squares method were compared with those obtained through the error-correction-feedback method. The correlations agree for twenty-two peaks and disagree for nine. The overall agreement of the correlations obtained by using two different methods supplies another clear indication as to which peaks correlate with oxygen presence and which peaks correlate with oxygen absence. The above agreement also suggests that there is some validity for the two methods used.

The rapid convergence of this previously feature-selected problem suggested that further feature selection could be performed. This was accomplished successfully, and the number of features was thereby reduced to 22. The mass positions discarded in going from 31 to 22

TABLE 10

Oxygen Presence Weight Vector Components

$m/e$	Average weight vector component	$m/e$	Average weight vector component
14	0.0208	46*	0.0081
15	-0.0250	52	-0.0178
17*	0.0275	59	0.0116
18*	-0.0003	63	-0.0215
24**	0.0007	67**	0.0102
25	-0.0003	69*	-0.0009
27*	0.0049	70	-0.0026
30	-0.0183	73	0.0048
31	0.0211	83	0.0038
37	0.0042	84*	-0.0051
38**	-0.0035	86*	0.0002
39**	0.0195	91*	-0.0039
40	-0.0358	100	0.0084
43	0.0140	128	-0.0063
44	-0.0056	135*	0.0046
45	0.0080		

From Pietrantonio, L. and Jurs, P. C., *Pattern Recognition*, 4, 391, 1972. With permission.

positions are marked with single asterisks in Table 10.

Table 11 presents the results obtained for the oxygen-presence determination problem both with the original 31 peaks and after further feature selection. Recognition is nearly complete, and prediction percentage is of the order of 98%, regardless of the number of features employed.

The lack of agreement of correlations for some peaks after comparing the two different training methods suggested that the peaks whose correlations disagreed could actually be eliminated as in the feature-selection routine. This was done, and the number of features was thereby reduced to 18. The features discarded in going from 22 to 18 positions are marked with double asterisks in Table 10. The results are given in Table 11. Recognition and prediction percentage do not suffer from such a removal. The correlations of the remaining 18 peaks are in perfect agreement with the corresponding correlations obtained from the error-correction-feedback method. This result strengthens the argument that some peaks can be used to identify the absence of oxygen from the molecules, whereas some other peaks identify the presence of oxygen.

The mass spectrometer is one of the most important analytical instruments in petroleum research. The qualities of petroleum products, especially gasoline, are monitored by mass spectrometry because the spectrum of a complicated mixture of hydrocarbons can be "typed" according to various categories. Tunnicliff and Wadsworth<sup>33</sup> used pattern-recognition techniques to determine the average properties of gasoline

samples directly from low-resolution mass spectra. Weight vectors were developed for each hydrocarbon type (paraffins, aromatics, cycloparaffins, etc.) and for each of several structural features ( $-\text{CH}_3$ ,  $-\text{CH}_2-$ , etc.) from a carefully selected training set of known spectra. With these weight vectors, gasoline samples not in the training set could be characterized more accurately than with conventional methods. The research reported in the paper is an excellent example of a practical application of pattern recognition to an important problem.

A completely different application of pattern recognition is the prediction of molecular properties from the molecular structure. The first work in this area dealt with the generation of simulated mass spectra of small organic molecules.<sup>34</sup> Molecular structures were represented in computer-compatible form through the use of a fragmentation code that assigns code designations to specific groups of atoms and/or bonds within the molecules.

The technique of fragmentation coding consists of representing a compound as a composite of its predominant structural fragments and their relationships. These features are then assigned numerical descriptors. The advantages of this method are that it is simple to learn and easy to understand, that it yields a linear descriptor list that is immediately computer-compatible, and that it requires only a moderate amount of computer storage. However, the simplicity of this method is to some degree offset by the loss of a complete description of the molecular structure. Fragmentation techniques do not normally indicate which

TABLE 11

Oxygen Presence Problem

m/e Positions	Training set populations		Percent recognition	Prediction set populations		Percent prediction
	+	-		+	-	
31	42	108	99.3	26	74	98
	39	111	100.0	81	219	96
	43	107	98.7	77	223	98
22	43	107	99.3	77	223	98
18	43	107	99.3	77	223	98

From Pietrantonio, L. and Jurs, P. C., *Pattern Recognition*, 4, 391, 1972. With permission.



fragments are bonded to one another, or what atom of a fragment is bonded to another. Some information is lost in terms of the stereochemistry of the molecule.

Several steps are followed in implementing the binary pattern classifier. Initially, a molecule is chosen from the data set of 600 compounds. Secondly, from its three-dimensional structure, the chemist draws by hand a two-dimensional structural picture. As a third step, the two-dimensional

diagram is checked against a previously chosen descriptor list. After all the compounds have been encoded as pattern vectors, the learning-machine method is applied. If the patterns in the training set are linearly separable, a feature-selection routine that reduces the number of descriptors necessary for linear separability is used.

The 61 descriptors used in this study are listed in Table 12. The first column contains the names of the descriptors. Most of these are self-

TABLE 12  
Molecular-structure Descriptors

Descriptors	Type	Normalization constant	Restrictions
1. Molecular weight	N	0.05	—
2. Largest clump	N	1.00	—
3. Largest cycle	N	1.00	Not monocyclics
4. Carbon number	N	1.00	—
5. Hydrogen number	N	0.50	—
6. Oxygen number	N	3.00	—
7. Nitrogen number	N	5.00	—
8. Number of rings plus double bonds	N	1.00	—
9. Ether	B	5.00	—
10. Ester group	B	5.00	—
11. Ketone	B	5.00	—
12. Alcohol	B	5.00	—
13. Carbonyl group	B	5.00	—
14. Oxygen link	B	5.00	—
15. Hydroxyl group	B	5.00	—
16. Vinyl end group	B	5.00	—
17. Aromatic	B	5.00	—
18. Benzene ring presence	B	5.00	—
19. 1 benzene ring only	B	5.00	Monocyclic compounds only
20. Heteroatom in ring	B	5.00	—
21. Number of C=C	N	2.00	Benzene rings have three
22. Number of C≡C	N	5.00	—
23. Acyclic (no ring present)	B	5.00	—
24. Branch point carbon number	N	2.00	—
25. Number of clumps	N	3.00	—
26. Odd hydrogen number	B	5.00	—
27. Number of <i>n</i> -butyl groups	N	5.00	—
28. Number of methyl groups	N	2.00	—
29. Number of ethyl groups	N	3.00	—
30. Number of <i>n</i> -propyl groups	N	5.00	—
31. Number of carbon without hydrogens	N	3.00	—
32. Carbon:hydrogen ratio $2n + 2$	B	5.00	—
33. Carbon:hydrogen ratio $2n$	B	5.00	—
34. Carbon:hydrogen ratio $2p - 2$	B	5.00	—
35. Carbon:hydrogen ratio $2n - 6$	B	5.00	—
36. Carbon hydrogen ratio $2n - 4$	B	5.00	—
37. Number of $-\text{CH}_2-$ groups in a row	N	1.00	—
38. $\text{C}=\text{C}-\text{C}-\text{CH}_3$ ; methyl beta to a C=C	B	5.00	—
39. $-\text{C}\equiv\text{N}$ group	B	5.00	—
40. $-\text{NO}_2$ group	B	5.00	—
41. $-\text{NH}_2$ group	B	5.00	—

TABLE 12 (Continued)

Descriptors	Type	Normalization constant	Restrictions
42. N bonded to two or more carbons	B	5.00	—
43. Isopropyl presence	B	5.00	—
44. Number of rings	N	4.00	—
45. Size of monocyclic	N	1.00	—
46. Smallest cycle	N	1.00	Not monocyclics
47. Fused rings	B	5.00	—
48. Alpha substitution	B	5.00	Nitrogen atom in a ring
49. Gamma hydrogen	B	5.00	Acyclic compounds only
50. Carboxylic acid group	B	5.00	—
51. Aldehyde group	B	5.00	—
52. 2 electron-donating groups, <i>ortho</i>	B	5.00	6-membered aromatic rings only
53. 2 electron-donating groups, <i>meta</i>	B	5.00	6-membered aromatic rings only
54. 2 electron-donating groups, <i>para</i>	B	5.00	6-membered aromatic rings only
55. Nonfused rings	B	5.00	—
56. 2 electron-withdrawing groups, <i>ortho</i>	B	5.00	6-membered aromatic rings only
57. 2 electron-withdrawing groups, <i>meta</i>	B	5.00	6-membered aromatic rings only
58. 2 electron-withdrawing groups, <i>para</i>	B	5.00	6-membered aromatic rings only
59. 1 $e^-$ donating — 1 $e^-$ withdrawing, <i>ortho</i>	B	5.00	6-membered aromatic rings only
60. 1 $e^-$ donating — 1 $e^-$ withdrawing, <i>meta</i>	B	5.00	6-membered aromatic rings only
61. 1 $e^-$ donating — 1 $e^-$ withdrawing, <i>para</i>	B	5.00	6-membered aromatic rings only

From Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 30, 1973. With permission.

explanatory, but several need some explanation and are defined as follows. The largest clump is the number of carbon atoms in the largest aggregate having each carbon atom bonded to at least one other. The largest cycle refers to the largest number of atoms (C, O, or N) that can be traversed to complete a cycle (ring) while going through each atom only once (e.g., naphthalene has a largest cycle of 10 atoms). The smallest cycle is the smallest number of atoms that can be traversed in a cycle, again going through each atom only once (e.g., naphthalene has a smallest cycle of 6 atoms). The number of rings plus double bonds is calculated from the following formula: the number of carbon atoms plus one half of the number of hydrogen atoms minus one half of the number of nitrogen atoms plus one, or  $(C + H/2 - N/2 + 1)$ . Ether, ketone, and alcohol are defined in the conventional manner, but combinations in one molecule are not allowed (e.g., to be considered an alcohol, a molecule can have only —OH functional groups). Ester, carboxylic, and aldehyde refer to these groups being present in a compound, with combinations being allowed (e.g., methyl terephthalate contains both a carboxylic group and an ester group). Carbonyl presence means that there is a carbon-oxygen double bond in the compound.

An oxygen linkage means that two carbon atoms are bonded together by an oxygen bridge. The "one benzene ring only" descriptor is used only for monocyclic molecules. For purposes of classification, a benzene ring is considered to have three double bonds. The branch point carbon number is the number of carbon atoms, in each molecule of the compound, that are bonded directly to at least three other carbon atoms. Methyl, ethyl, *n*-propyl, and *n*-butyl numbers are the numbers of each of these groups that can be produced from a single molecule, rupturing only one bond for each group produced. The "carbon without hydrogen" category refers to quaternary carbon atoms, which are not bonded to any hydrogen atoms. "Two electron-donating groups, *ortho*" refers to six-membered aromatic rings that contain two or more substituents, of which two are in a position *ortho* to each other. The other substituent descriptors are described in a similar manner. The alpha-substitution category is defined as a methyl group *ortho* to a nitrogen atom in a ring. Gamma hydrogen refers to a hydrogen atom in the gamma position in relation to a carbonyl group located in an acyclic compound. The other descriptors are as defined classically.

The descriptors are of two types, binary and

numeric, as indicated in the second column of Table 12. The binary descriptors (B) can have only two values, corresponding to yes or no. Numeric descriptors (N) can have values up to 202 (the molecular weight of  $C_{10}H_{18}O_4$ ). Therefore, it is necessary to normalize the values of the descriptors.

The normalization constant by which each descriptor is multiplied is listed in the third column. After normalization, the values of the descriptors are in a more convenient range. A normalization constant of 5 is applied to each binary descriptor.

The fourth column lists any restrictions on the descriptors, or any special cases that may be encountered. For example, gamma hydrogen refers only to acyclic compounds.

Binary pattern classifiers have been trained for the 60 m/e positions listed in the first column of Table 13. Each of the binary pattern classifiers was trained to predict the presence or absence of a peak in its respective m/e position in the spectra of the compounds in its training set. To be considered present in an m/e position of a spectrum, a peak must have an intensity greater than a designated threshold value referred to as the intensity cutoff.

For all 60 m/e positions, weight vectors are developed for an intensity cutoff equal to 0.5% of the total ion current. A weight vector so trained can answer the following binary question: does the compound under investigation contain a peak in this particular m/e position that has an intensity greater than 0.5%, or does it contain a peak with an intensity less than or equal to 0.5% of the total ion current?

In addition, for 11 m/e positions, two weight vectors each are developed for intensity cutoffs of 0.1 and 1.0% of the total ion current. The former predicts whether a peak has an intensity greater than 0.1% of the total ion current or less than or equal to 0.1% of the total ion current. The latter is able to answer a similar question for peaks with 1.0% of the total ion current. The choice of these 11 m/e positions was made to insure that there was an adequate number of compounds having a peak with an intensity greater than 1.0% of the total ion current in the training and prediction sets. In all, 82 binary pattern classifiers were trained.

The third column of Table 14 lists the number of descriptors that survived feature selection by

the learning machine. These descriptors were considered important to the presence of a peak in an m/e position, consistent with the intensity cutoff imposed on the data set. Despite the fact that in many cases the number of descriptors remaining is only a small fraction of the original 61, the weight vectors are able to retain essentially the same predictive ability as with all 61 descriptors.

The fourth column of Table 14 gives the number of feedbacks necessary for convergence, with >2,500 indicated for those that have not been completely trained after 2,500 iterations.

A simple test of the ability of the learning machine to classify unknown compounds may be made by comparing its predictive ability with the percentage of compounds in the more populous category of an m/e position. If the predictive ability of the learning machine exceeds the success rate of always guessing the more populous category, then the method is considered to have learned something about the relationship between the patterns and the category to which they are assigned. For example, the percentage of compounds having a peak with an intensity greater than a cutoff of 1.0% for m/e position 29 is 75.5. By always guessing that a compound contained a peak in this m/e position, one would make correct classifications 75.5% of the time. The fifth column lists the percentage of compounds in the data set that occur in the more populous category, consistent with the intensity cutoff of the m/e position as listed in the second column.

The sixth column gives the predictive abilities of the weight vectors developed with the numbers of descriptors shown in the third column. For each of the m/e positions, three threshold logic units were trained with three different training sets. The resulting weight vector with the highest predictive ability was chosen for presentation in the sixth column for each m/e and each intensity cutoff. The average predictive ability for all 82 threshold logic units was 88.8%.

The seventh column gives the difference between the predictive ability of the learning machine, as listed in the sixth column, and the percentage of compounds in the more populous category, as listed in the fifth column. In 73 cases, the predictive ability of learning machine prediction exceeded the fraction of the compounds having peaks in that m/e position. The average



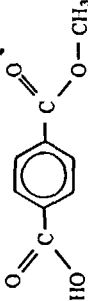
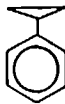
TABLE 13

## Descriptor Lists for Five Selected Compounds

Descriptor number	Compound number					Descriptor number	Compound number				
	1	2	3	4	5		1	2	3	4	5
1	112	88	136	180	118	31	0	1	2	4	1
2	8	2	10	8	9	32	0	0	0	0	0
3	0	0	6	0	6	33	1	1	0	0	0
4	8	4	10	9	9	34	0	0	0	0	0
5	16	8	16	8	10	35	0	0	0	0	0
6	0	2	0	4	0	36	0	0	1	0	0
7	0	0	0	0	0	37	4	1	1	0	2
8	1	1	3	6	5	38	0	0	0	0	0
9	0	0	0	0	0	39	0	0	0	0	0
10	0	1	0	1	0	40	0	0	0	0	0
11	0	0	0	0	0	41	0	0	0	0	0
12	0	0	0	0	0	42	0	0	0	0	0
13	0	1	0	1	0	43	0	0	0	0	0
14	0	1	0	1	0	44	1	0	3	1	2
15	0	0	0	1	0	45	6	0	0	6	0
16	0	0	0	0	0	46	0	0	3	0	3
17	0	0	0	1	1	47	0	0	1	0	0
18	0	0	0	1	1	48	0	0	0	0	0
19	0	0	0	1	0	49	0	1	0	0	0
20	0	0	0	0	0	50	0	0	0	1	0
21	0	0	0	3	3	51	0	0	0	0	0
22	0	0	0	0	0	52	0	0	0	0	0
23	0	1	0	0	0	53	0	0	0	0	0
24	2	0	5	2	2	54	0	0	0	0	0
25	1	2	1	2	1	55	0	0	0	0	1
26	0	0	0	0	0	56	0	0	0	0	0
27	0	0	0	0	0	57	0	0	0	0	0
28	2	2	3	1	0	58	0	0	0	1	0
29	0	1	0	0	0	59	0	0	0	0	0
30	0	0	0	0	0	60	0	0	0	0	0
						61	0	0	0	0	0

TABLE 13 (continued)

## Descriptor Lists for Five Selected Compounds

Compound number	API number	Structure	Chemical name
1	220		1,2-Dimethylcyclohexane
2	326	$\text{CH}_3 - \overset{\text{O}}{\parallel} \text{C} - \text{O} - \text{CH}_2 - \text{CH}_3$	Ethyl acetate
3	466		1,7,7-Trimethylbicyclo-(2,2,1.0(2,6))-heptane(tricyclene)
4	1755		Methylterephthalate
5	1963		Cyclopropylbenzene

From Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 30, 1973. With permission.

TABLE 14

## Feature Selection and Prediction

(1)	(2)	(3)	(4)	(5)	(6)	(7)
m/e	Intensity cutoff (%)	Descrip- tors re- maining	Number of feedbacks	More populous category (%)	Prediction (%)	Columns (6) - (5)
29	0.1	15	270	89.7	91.3	1.6
	0.5	18	1246	80.7	92.0	11.3
	1.0	19	275	75.7	89.6	13.9
30	0.5	19	47	88.3	93.3	5.0
31	0.5	13	52	79.5	89.6	10.1
37	0.5	17	27	84.5	91.8	7.3
38	0.5	39	741	63.2	81.1	17.9
39	0.1	14	42	96.3	97.3	1.0
	0.5	14	20	93.7	96.2	2.5
	1.0	13	36	91.2	93.3	2.1
40	0.5	61	>2500	53.7	79.6	25.9
41	0.1	11	7	96.8	96.4	-0.4
	0.5	15	93	88.5	92.7	4.2
	1.0	25	83	82.7	93.8	11.1
42	0.1	18	>2500	87.5	94.0	6.5
	0.5	17	47	79.9	91.8	11.9
	1.0	61	>2500	56.2	72.1	15.9
43	0.1	19	294	85.1	89.8	4.7
	0.5	61	>2500	71.9	90.0	18.1
	1.0	61	>2500	64.7	84.0	19.3
44	0.5	61	>2500	70.5	76.5	6.0
45	0.5	15	42	81.5	90.8	9.3
46	0.5	13	33	96.1	95.3	-0.8
50	0.5	17	139	72.1	90.6	18.5
51	0.1	61	>2500	80.7	85.9	5.2
	0.5	61	>2500	56.1	93.3	37.2
	1.0	14	166	74.1	89.9	15.8
52	0.5	20	149	74.9	90.1	15.2
53	0.1	18	103	87.0	91.7	4.7
	0.5	26	173	56.4	86.3	29.9
	1.0	61	2349	63.6	86.7	23.1
54	0.5	61	>2500	70.8	82.2	11.4
55	0.1	35	574	79.2	81.7	2.5
	0.5	61	>2500	66.7	82.7	16.0
	1.0	30	1084	56.6	85.5	28.9
56	0.5	61	>2500	55.6	79.3	23.7
57	0.5	61	>2500	50.9	78.1	27.2
58	0.5	61	>2500	75.7	77.7	2.0
59	0.5	28	1058	87.2	83.3	-3.9
63	0.5	21	70	80.8	92.5	11.7
65	0.5	26	>2500	74.4	92.2	17.8
66	0.5	22	336	83.1	90.6	7.5
67	0.1	29	629	59.1	88.4	29.3
	0.5	29	1328	64.2	86.1	21.9
	1.0	34	>2500	74.8	85.4	10.6
68	0.5	61	>2500	76.1	86.3	10.2
69	0.5	61	>2500	56.1	87.0	30.9
70	0.5	61	>2500	61.2	81.4	20.2
71	0.5	61	>2500	76.0	81.3	5.3
73	0.5	18	251	90.1	89.0	-1.1
74	0.5	34	2265	90.1	86.6	-3.5
75	0.5	24	496	92.3	91.3	-1.0

TABLE 14 (Continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
m/e	Intensity cutoff (%)	Descriptors remaining	Number of feedbacks	More populous category (%)	Prediction (%)	Columns (6) - (5)
76	0.5	12	12	94.1	96.3	2.2
77	0.1	15	56	52.0	89.4	37.4
	0.5	21	42	76.7	93.4	16.7
	1.0	18	31	82.5	90.1	7.6
78	0.5	10	161	84.2	94.0	9.8
79	0.1	28	873	54.1	86.7	32.6
	0.5	20	131	79.1	91.3	12.2
	1.0	16	104	87.6	89.9	2.3
80	0.5	18	250	92.9	93.9	1.0
81	0.5	61	>2500	80.6	91.3	10.7
82	0.5	25	>2500	84.1	87.8	3.7
83	0.5	61	>2500	74.8	88.8	14.0
84	0.5	61	>2500	72.8	75.8	3.0
85	0.5	61	>2500	86.4	88.5	2.1
91	0.5	11	491	83.2	92.0	8.8
95	0.5	13	141	88.2	93.2	5.0
97	0.5	61	>2500	84.2	87.3	3.1
98	0.5	61	>2500	78.0	76.2	-1.8
100	0.5	18	230	95.8	96.6	0.8
103	0.5	11	69	82.5	89.5	7.0
104	0.5	27	1365	85.1	90.1	5.0
105	0.5	8	58	85.0	92.1	7.1
106	0.5	15	39	91.1	95.5	4.4
115	0.5	18	68	81.8	93.5	11.7
119	0.5	14	26	83.5	95.4	11.9
120	0.5	18	31	83.9	94.7	10.8
121	0.5	13	27	89.6	92.1	2.5
127	0.5	6	5	93.5	93.2	-0.3
128	0.5	13	8	87.9	92.3	4.4
136	0.5	3	5	83.8	82.1	-1.7

From Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 225, 1973. With permission.

amount by which the predictive ability exceeded the population of the more populous category was 10.4%.

As an over-all test of the applicability of this method, 30 compounds were selected randomly out of the data set of 600. All 60 m/e positions were predicted for each compound, using the 82 weight vectors developed above. For 49 of the m/e positions, one can calculate a binary mass spectrum representing the presence or absence of peaks in these m/e positions and having an intensity greater than 0.5% of the total ion current. The use of three intensity cutoffs aids in obtaining a quantitative measure of the intensities of the peaks in the 11 other m/e positions. Each m/e position up to the m/e position one unit greater than the

molecular weight of the compound is predicted for each of the 30 compounds. The average predictive ability was 93%. This figure is slightly higher than the average predictive ability of all 82 threshold logic units because some of the 30 randomly selected molecules were contained in the training sets of some of the threshold logic units. This predictive percentage of 93% demonstrates that the methods employed here are capable of a high predictive ability while using only a fraction of the descriptors available.

Plots were made of a number of predicted and of 11 real peak mass spectra for six widely varying types of molecules. The predicted and real spectra were strikingly similar.

Pattern-recognition methods are used primarily

for the analysis of patterns in  $n$ -dimensional space. Since man is an excellent pattern recognizer, one of the possible approaches is to map the  $n$ -space patterns down to two-space while attempting to preserve the data structure by some selected criterion. Information loss is inevitable, but the methods quite often provide the scientist with an acceptable approximate "view" of his  $n$ -space data structures.<sup>3,5</sup> For example, if two large clusters of patterns exist in ten-space, an isomorphic mapping to two-space might well show the same two clusters. The scientist could then use other pattern-recognition methods to measure the extent of clustering in ten-space and also to investigate the significance of the clusters.

The several methods used to reduce  $n$ -space data to two-space data fall into two categories: projections and mappings. Projections are really linear mappings in that the resultant two-space coordinates are linear combinations of the original  $n$  coordinates. Projections range from trivial plots of two selected dimensions to eigenvector projections based on an orthogonalization process that tries to preserve variance. Mappings include nonlinear operations, because the new two-space coordinates can be nonlinear combinations of the original  $n$  coordinates. Among the many possible mapping algorithms, nonlinear mapping (NLM) is probably the one most often used. The procedure involves using the conjugate gradient method or some other function optimization method to minimize an error function  $E(d_{ij}^*, d_{ij})$

$$E(d_{ij}^*, d_{ij}) = \sum_{i>j} \frac{(d_{ij}^* - d_{ij})^2}{(d_{ij}^*)^2} \quad (34)$$

where  $d_{ij}^*$  is the  $n$ -space Euclidean distance between pattern  $X_i$  and  $X_j$ , and  $d_{ij}$  is the two-space distance function. The latter is a function of the desired coordinates for each point:

$$d_{ij} = [(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2]^{1/2} \quad (35)$$

Therefore, mapping 100 patterns from  $n$ -space to two-space amounts to finding the two coordinates for each pattern (200 unknowns) that minimize  $E(d_{ij}^*, d_{ij})$ .

It should be pointed out that, if interactive computer graphics is available, mappings can be done to three-space, thereby losing far less information. Display methods can be the most useful of the pattern-recognition methods because they allow the human pattern-recognition advantage.

References 36 and 37 describe two preprocessing methods useful for spectral analysis by pattern recognition. The first of these references discusses a method for generating new features from spectral data as approximate class separators. For each pair of classes in the application, a weight vector is trained by a least-squares procedure, resulting in the  $n(n-1)/2$  vectors for the  $n$ -class problem. Each weight vector is multiplied by the spectrum to generate a discriminant that is used as a new feature. The new features are then input to a  $K$ -nearest-neighbor routine, where the actual classification is done. The method should really be applied when several classes are involved in an application.

The second paper<sup>37</sup> used spectral moments and intensity histograms as new features. Contractions from 128 masses per spectrum to 10 new features did not degrade classification performance, but it significantly lessened computation time.

The applicability of these preprocessing methods depends, of course, on the nature of the data. Both have provided drastic reductions in the number of features used for classification while maintaining a high information content. Several other useful analogies to the Fourier theory also exist, and they promise to make the HT a very useful tool for the analysis of chemical data. The HT was used as a preprocessing method for mass-spectral-data analysis, and a significant improvement in classification results was noted.

Another interesting application of the nearest-neighbor technique was to the correlation of gas-chromatographic liquid phases by Leary et al.<sup>38</sup> Using an extensive tabulation for ten solutes measured on each of 226 liquid phases, the authors selected twelve phases that spanned the same range of polarity as the entire set. Also, the nearest-neighbor calculations were used to provide a substitution table for the twelve preferred phases for each of the remaining 214 phases.

An important question to ask before applying a classification method to spectral data is the following: "Is the spectral information in the optional transform domain?" Suppose that a chemist submits a sample for NMR analysis and that the analyst uses a Fourier transform NMR spectrometer but never applies the Fourier transform, then returns the time-domain data to the chemist. The chemist would find the waveform to be almost useless and would ask that the Fourier transform be applied to obtain the frequency



domain. The information is not lost; only the representation is changed.

When a pattern-recognition method is applied to spectral data, the same question must be asked. Several studies have involved changing the representation of mass-spectral data<sup>7,14,17,24,37,38</sup> by using various transforms and other preprocessing methods. Due to the discontinuous nature of raw mass-spectral data, the Hadamard transform (HT) has been used as a preprocessing method prior to analysis by three classification methods.<sup>39</sup> The HT theory is quite analogous to the Fourier transform theory, except that it is based on Walsh functions,<sup>31</sup> which resemble infinitely clipped sine and cosine functions. Figure 5 shows the first functions in each of the two series. Several advantages, most of them beyond the scope of this review, are realized by using the HT. Probably the most important is that the slowest HT is faster than the fast Fourier transform.

While a number of chemically significant ques-

tions, such as the presence or absence of a certain functional group or element, can be answered by making a binary decision, multicategory decisions, as for example about the number of certain groups or atoms per molecule, must also be made. A comparison of multicategory classification methods is through the use of an array of binary classifiers. Perhaps the most obvious manner of combining binary classifiers is the branching-tree method, which has been applied to the determination of empirical formulas.<sup>12</sup> More often, a parallel arrangement of classifiers, each with a different cutoff, has been used.<sup>18</sup> A third method, utilizing Hamming-type binary codes and offering the interesting possibility of self-detection and self-correction of errors, was investigated as well.

While the use of binary codes has several inherent advantages, the feasibility of the method depends upon whether the weight vectors for the pertinent binary pattern classifiers can be successfully trained. The present investigation answers

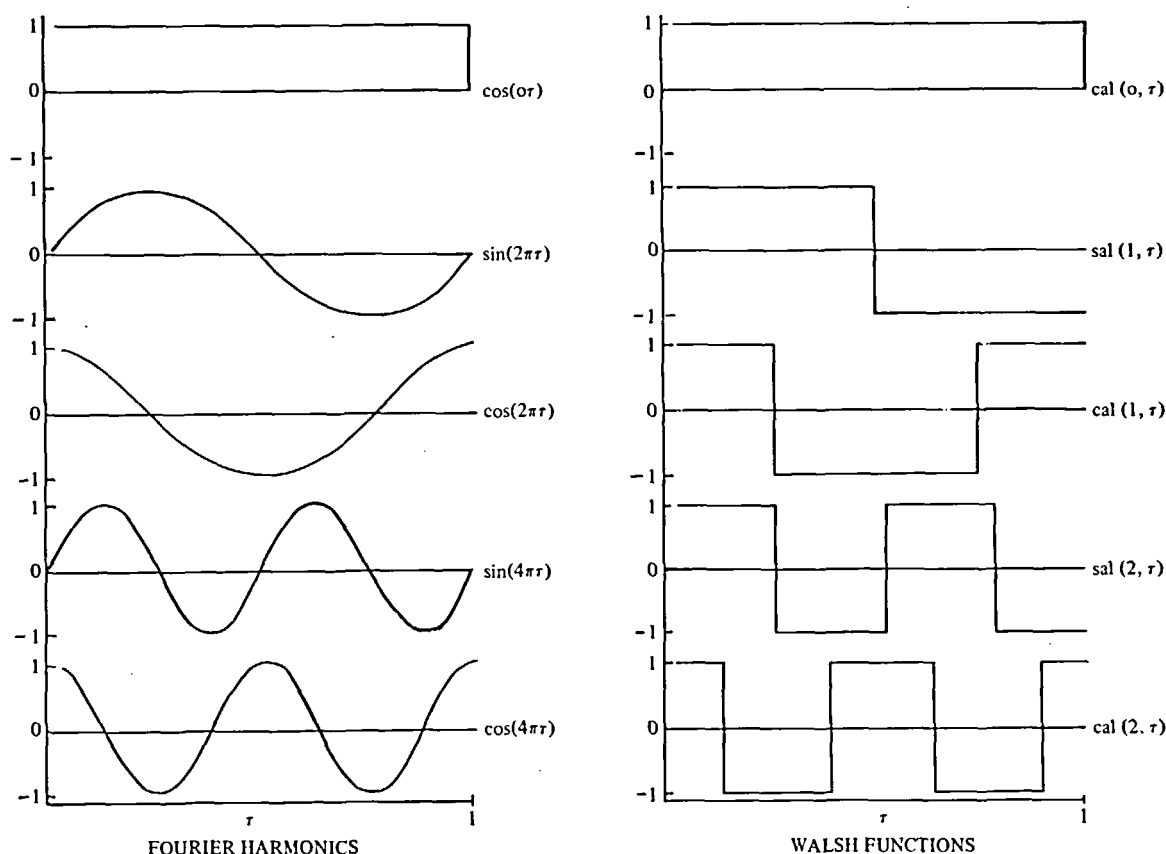


FIGURE 5. Fourier harmonics and Walsh functions. (From Kowalski, B. R. and Bender, C. F., *Anal. Chem.*, 45, 2334, 1973. With permission.

this crucial question, using the test problem of carbon-number prediction. The three methods of pattern classification — branching tree, parallel, and binary code — were then compared with respect to carbon-number prediction as well as to general features.

A set of 372 hydrocarbon spectra and a set of 600 CHON spectra were used. Carbon numbers from three to ten were represented.

In the parallel classification method, each binary pattern classifier is trained to classify pattern vectors into one of two categories separated by a cutoff point. The positive category comprises those pattern vectors having carbon numbers greater than the cutoff, and the negative category contains those with carbon numbers less than or equal to the cutoff. The classifier with a cutoff of 7, for example, was trained to give a positive dot product for carbon numbers 8, 9, and 10, and a negative dot product for carbon numbers 7 and less.

The binary classifiers, each with a different cutoff, were trained and their individual predictive abilities were determined. Here, and throughout this investigation, weight vectors for each question were trained in two ways: starting with initial weights  $w_j$  that were all equal either to +1 or to -1. The trained weight vector giving the better predictive ability was saved.

The set of trained weight vectors was used in a master program to predict carbon number as illustrated in Figure 6. The sequence of  $n$  binary decisions may be written as an  $n$ -digit binary number containing 0's for negative classifications

and 1's for positive classifications. The carbon number is then indicated by the cutoff value giving the first negative classification. For example, in the  $C_4 - C_{10}$  case the number 000000 corresponds to a carbon number of 4, and the 000111 corresponds to a carbon number of 7. The appearance of more than one discontinuity, as in 000101, indicates that one or more erroneous decisions were made and that the pattern cannot be validly classified. For the hydrocarbon data set, 93.0% of the unknown spectra in the prediction set were correctly classified; for the entire data set, 76.0% were correctly classified.

In the branching-tree classification system, the binary pattern classifiers are arranged in a branching network. The method was applied to the hydrocarbon data set, each binary classifier being trained to dichotomize a set of pattern vectors according to the scheme presented in Figure 7. In developing the six binary pattern

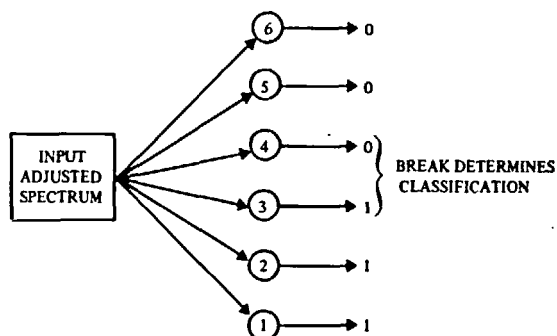


FIGURE 6. Parallel arrangement of binary pattern classifiers. (From Felty, W. L. and Jurs, P. C., *Anal. Chem.*, 45, 885, 1973. With permission.)

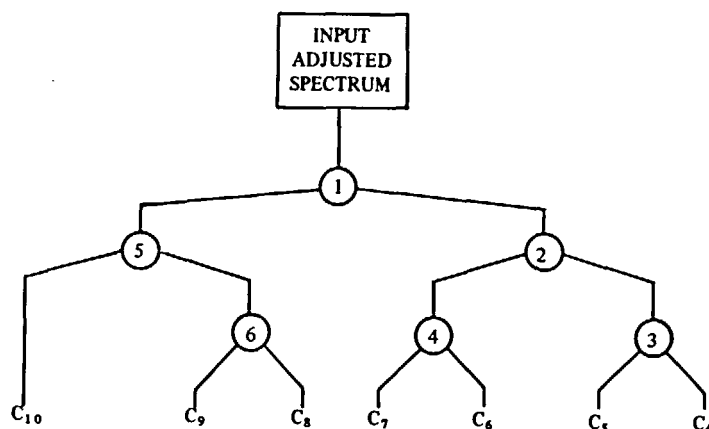


FIGURE 7. Branching tree of binary pattern classifiers. (From Felty, W. L. and Jurs, P. C., *Anal. Chem.*, 45, 885, 1973. With permission.)

classifiers, only those spectra pertinent to each branch point were utilized. For example, the weight vector for branch point number 3 was trained using only those spectra in the overall training set of 200 that correspond to carbon numbers of 4 or 5. The predictive ability of each weight vector was tested on a similarly chosen set of spectra from the overall prediction set of 172. The weight vectors were then used in a master branching-tree program; a predictive ability of 95.4% resulted.

The branching-tree classification was also carried out using weight vectors trained by considering the entire training set for each branch point, i.e., the weight vectors for the parallel scheme above. Only in the case of the first branch point (cutoff of 7) are the two types of weight vectors identical. The resulting overall prediction was 94.2% for this second method.

The branching-tree arrangement of parallel weight vectors was also applied to the entire data set of 600 spectra to give 76.3% prediction. Here a symmetrical tree of seven binary pattern classifiers was necessary in order to classify into eight categories,  $C_3 - C_{10}$ .

A third scheme of classification involves classification by binary code. In order to classify pattern vectors into any number  $m$  of categories, the parallel and branching-tree schemes require  $m - 1$  binary classifiers. An alternate method is to use  $n$  classifiers, where  $2^{n-1} < m \leq 2^n$ , so that when each decision (0 or 1) is used as a digit in a binary number, the decimal equivalent is the proper category number. Up to eight categories can be accommodated by a 3-digit binary number,  $d_3d_2d_1$ . For example  $001 \equiv$  category 1 ( $C_3$ ),  $010 \equiv$  category 2 ( $C_4$ ), ...,  $111 \equiv$  category 7 ( $C_{10}$ ). For the hydrocarbon set, where there are only seven categories, two such codes are possible — the above one, designated variation 1, and another one, called variation 2, where  $000 \equiv C_4$ , ...,  $110 \equiv C_{10}$ . Three binary classifiers were trained to classify appropriate sets of carbon numbers for each variation. Overall prediction of carbon number among the hydrocarbon set was 83.7% and 80.2% for variations 1 and 2, respectively, and 57.8% for the entire data set.

A comparison of the classification schemes is shown in Table 15.

The accuracy of the binary number can be improved by introducing additional digits to form an error-correcting Hamming code. If the

TABLE 15

Summary of Carbon Number Predictions by Different Classification Schemes

Classification scheme	Percent prediction	
	Hydrocarbons <sup>a</sup>	Entire data set <sup>b</sup>
Parallel	93.0	76.0
Branching tree	95.4	—
Branching tree (parallel w. <sup>j</sup> s)	94.2	76.3
Binary codes		
Variation 1: 3-bit	83.7	57.8
6-bit	93.0	68.5
7-bit	95.4	67.0
Variation 2: 3-bit	80.2	
6-bit	93.6	
7-bit	92.4	

<sup>a</sup> $C_4 - C_{10}$ ; random guessing would give a success rate of 1/7 or 14.3%.

<sup>b</sup> $C_3 - C_{10}$ ; random guessing would give a success rate of 1/8 or 12.5%.

From Felty, W. L. and Jurs, P. C., *Anal. Chem.*, 45, 885, 1973. With permission.

evaluated binary number  $d_n \dots d_1$  differs from the correct number by an error in only a single digit — that is, if the Hamming distance is 1 — the error can be detected and corrected by  $k$  check bits, so that  $2^k \geq n + k + 1$  for  $n$  data bits. The Hamming distance is thus increased from 1 in the original  $n$ -digit number to 3 or more in the  $(n + k)$ -digit number. In the present application, three check bits are needed for the three data bits yielding a Hamming (6,3) code. Each check bit,  $c_i$ , is constructed so that when all data bits are correct, the parity of the check bit and two associated data bits, considered collectively, is even. When this is the case, the parity bit,  $p_i$ , is given a value of 0. For odd parity,  $p_i = 1$ . The three sets of associated bits are the following:

$$\begin{aligned} p_1 &: c_1 + d_1 + d_2 \\ p_2 &: c_2 + d_1 + d_3 \\ p_3 &: c_3 + d_2 + d_3 \end{aligned} \quad (36)$$

An error in the data bits gives rise to odd parity in two of the parity groups. The binary number given by the parity digits,  $p_3p_2p_1$ , then assumes a non-zero value and the decimal equivalent is the position of the erroneous bit in the composite number  $d_3d_2p_3d_1p_2p_1$ . For example, in terms of

variation 1, the 6-digit number for carbon number 8 is 100100. If  $d_1$  erroneously equals 0, the two parity checks involving  $d_1$  ( $p_1$  and  $p_2$ ) become odd, giving the number 100011. The binary number  $p_3p_2p_1 = 011$  has a decimal equivalent of 3, indicating that the third digit is in error and consequently should be a 1 rather than a 0.

The additional binary classifiers ( $c_1, c_2$ , and  $c_3$ ) needed for error correction were trained using appropriate subsets of the training set for the positive and negative categories. Carbon-number prediction by the (6,3) codes gave increased prediction of 93.0% and 93.6% for variations 1 and 2, respectively, using the hydrocarbon set, and 68.5% using the entire data set.

All of the weight vectors needed for classification by the different binary codes were found to train completely within a reasonable number of feedbacks. This means that the various sets of carbon numbers were linearly separable, a result that had been far from obvious beforehand. The highest predictive ability was attained with the hydrocarbon data set using the branching tree and one version of the Hamming (7,4) code. With the entire data set, the branching tree method showed the best prediction.

An application of pattern recognition to a very different chemical problem appeared in Reference 41. This paper deals with the use of a parallel array of binary pattern classifiers for the semiquantitative determination of the chlorine dosage necessary for the treatment of water in a municipal water plant. The method of using an array of binary pattern classifiers to perform multicategory classifications was discussed above.

The overall range of chlorine dosages found in the available data set is divided into seven practical intervals with six binary chlorine-dosage cutoffs — 82, 90, 94, 98, 102, and 112 pounds per million gallons. An independent binary pattern classifier is trained by the described method for each of the six dosage cutoffs to split the data set into two subsets: those patterns whose related chlorine dosages exceed the cutoff and those patterns whose dosages do not exceed the cutoff. After training, the six binary pattern classifiers are arranged in sequence from the largest to the smallest cutoff dosages. To make a semiquantitative determination, all the pattern classifiers from the largest to the smallest are used to get a sequence of responses. The recommended dosage for the water sample being classified is determined

by noting where the responses shift from negative to positive.

Seventeen common water-quality variables were utilized in this study. Most of the water-quality data used were taken from the Torresdale water-treatment plant (on the Delaware River), which serves Philadelphia. These seventeen variables are listed in Table 16 with their units and the ranges observed in the raw data. Each pattern point X corresponding to a particular water sample thus consists of seventeen components, which are simply the values of the turbidity, temperature, pH, and the other descriptors listed, arranged in the same order as in Table 16. A total of 104 such pattern points are available for the two-year interval 1964–1965. The entire data set thus consists of 104 vectors, each representing the values of the seventeen water-quality variables for a particular week.

As shown in Table 16, the ranges of the water-quality parameters used in the study are quite different. The following equation was used to normalize the original water-quality data to assure that each variable had an average of 100 and a desirable range:

$$(x_{ij})_{\text{norm}} = 100(x_{ij}/x_{j\text{av}})^{n_j} \quad (37)$$

TABLE 16

## Water-quality Descriptors

Descriptor	Range
Chlorine, lb/mil gal	57–137
1 Turbidity, jtu	13–98
2 Temperature, °F	26–105
3 pH	7.1–7.7
4 Alkalinity, mg/l	17–48
5 Total hardness, mg/l	35–95
6 Color, units	3–35
7 Total residue, mg/l	70–400
8 Iron, mg/l	0.11–2.2
9 Manganese, mg/l	0.00–0.75
10 Ammonia, mg/l	0.02–5.0
11 COD, mg/l	2.8–70
12 BOD, mg/l	0.6–56
13 DO, mg/l	2.8–13
14 DO saturation, %	25–100
15 Total bacteria, count/ml	$5 \times 10^2$ – $2.8 \times 10^5$
16 Coliform bacteria, count/ml	$10^2$ – $6.6 \times 10^4$
17 River flow Q, cfs	1,200–45,000

From Kuo, K. A. and Jurs, P. C., *J. Water Works Assoc.*, 65, 623, 1973. With permission.

where  $x_{ij}$  is the original value of the  $j$ th variable for the  $i$ th week,  $(x_{ij})_{\text{norm}}$  represents the normalized value of that variable,  $n_j$  equals 1/2, 2, 3, or 4, depending on the variable being normalized, and  $x_{j\text{av}}$  is the average value of the  $j$ th variable.

Table 17 shows the results of training using training sets selected by a trial-and-error method. For each dosage cutoff, 30 members were put into the training set and the remainder were left in the checking set. A trial-and-error procedure was employed to find a good training set — good in the sense that a large percentage of the entire set was correctly classified by a binary pattern classifier trained with the training set. The third column in Table 17 gives the percentage of the checking set correctly classified. For example, either of the two weight vectors trained for a chlorine dosage of 82 pounds per million gallons was able to classify correctly 93.2% of the checking set besides all the members of the training set after training.

The ability of the system to perform semiquantitative determinations of chlorine dosages was then tested. The test was performed with 50 randomly selected patterns from the overall data set. Eighty-four percent of the patterns of the entire data set were classified correctly; that is, the chlorine demand was determined correctly by the pattern-classification system for 84 of these randomly selected water samples. In 94% of the cases the chlorine dosage determined by the system was within one subinterval of the correct dosage. In the remaining 6% of the cases, the results were contradictory. Thus, the combined system of six binary pattern classifiers is shown to be capable of performing the overall semiquantitative analysis of chlorine dosages with a high percentage of success. The results are believed to be an encouraging first step toward the possibility of automatic, computer-controlled chlorination of water.

In Reference 42, work on the generation of simulated mass spectra of small organic molecules was extended. A method for training binary pattern classifiers using an iterative least-squares approach was used. Also, a feature-extraction technique known as the attribute-inclusion algorithm was used to investigate the importance of multiple features in the molecular descriptions. The same data set and descriptors were used as in Reference 34.

The iterative least-squares training procedure employed was described above. Table 18 shows

TABLE 17  
Results for Selected Training Sets\*

Chlorine dosages (lb/mil gal)	Number of feedbacks	Checking set <sup>†</sup> (% correct)
82	509	93.2
	447	93.2
90	279	90.5
	639	90.5
94	1,312	90.5
	1,000	90.5
98	1,543	85.1
	1,118	85.1
102	513	87.8
	784	86.5
112	1,730	91.9
	1,570	91.9

\*Different training set members for different dosages.

<sup>†</sup>Training set members: 30; checking set members: 74.

From Kuo, K. A. and Jurs, P. C., *J. Water Works Assoc.*, 65, 623, 1973. With permission.

the results obtained and compares the results of error-correction feedback and iterative least-squares training. The first column lists the ten  $m/e$  positions tested. Each of the BPC's is trained to predict the presence or absence of a peak in its respective  $m/e$  position in the spectra of the compounds in its training set. To be considered present in an  $m/e$  position of a spectrum, a peak must have an intensity greater than a designated threshold value, referred to as the intensity cutoff.

The third column indicates the number of descriptors utilized by the binary pattern classifier whose predictive ability is found in the fourth column. The third and fourth columns give the results of error-correction-feedback training and subsequent feature selection. For each of the ten  $m/e$  positions, three binary pattern classifiers were trained with three different training sets. It is the weight vector with the highest predictive ability that is presented in the third and fourth columns.

The fifth column gives the numbers of descriptors that survived the feature selection of three training sets (including the one for which the results are shown in columns 3 and 4) with

TABLE 18

Comparison of Iterative Least-squares Training and Error-correction-feedback Training

m/e	Error-correction feedback			Iterative least squares		
	Intensity cutoff	Number of descriptors	Prediction (%)	Number of descriptors	Recognition (%)	Prediction (%)
29	0.5	18	92.0	6	94.0	91.3
31	0.5	13	89.6	5	92.0	89.6
45	0.5	15	90.8	4	95.3	93.3
59	0.5	28	83.3	5	89.9	86.3
63	0.5	21	92.5	9	97.2	92.2
73	0.5	18	89.0	8	94.1	90.0
75	0.5	24	91.3	3	94.5	91.6
77	0.1	15	89.4	8	92.8	93.5
77	0.5	21	93.4	8	94.7	91.6
77	1.0	18	90.1	3	96.8	93.0
95	0.5	13	93.2	3	93.5	93.5
115	0.5	18	93.5	10	98.3	93.5

From Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 225, 1973. With permission.

concurrent sign preservation. The pattern vectors used by the least-squares algorithm are constructed from these descriptors. For example, the three training sets for m/e position 45 are feature-selected to 15, 26, and 15 descriptors, among which there are five that are common to all three sets. Only four of these descriptors have the same sign for their respective weight vector components in all three sets, as noted in column 5.

The recognition percentage of the least-squares algorithm is shown in the sixth column of Table 18. Recognition is the ability of a discriminant function to classify correctly the members of a training set.

The predictive ability of the weight vectors developed by least-squares training is chosen from among the three iterations performed on an m/e position. The best result is chosen and appears in the seventh column. Comparisons can be made between the fourth and seventh columns because of the common prediction set for error-correction-feedback and least-squares training of an m/e position. As can be seen, the least-squares algorithm is able to retain a high degree of predictive ability with a limited number of descriptors.

In seven of the twelve cases given in Table 18 least-squares training has a greater predictive ability than error-correction-feedback training; in two

cases it is the same, and in three cases it is worse. For these twelve examples, the predictive ability of the iterative least-squares algorithm is 0.9% higher on the average than that of error-correction-feedback training while it uses only one third the number of descriptors. The correlation of descriptors with m/e positions appears to be real, not only from the outcome of the classification tests, but also from previous physical interpretations.

The second part of this study was an investigation of a method for the development of multiple features in the molecular descriptions.

The feature-extraction method that was used in this study is known as an attribute-inclusion algorithm. Attribute inclusion describes the inter-relationship of attributes in a given set of patterns. In this case, attribute is synonymous with descriptor. The molecular structures are represented by pattern vectors having attribute (descriptor) values as their components. The algorithm used is restricted to binary attributes (0 and 1). Therefore, the pattern vectors that will be presented to the algorithm will be comprised of the 40 binary descriptors selected from the descriptor list of 61 structural fragments.

An attribute is included in another if, whenever the first one is present in any pattern, the second

is also present. Any two attributes satisfying the relation of inclusion belong together in a feature. Correspondingly, all attributes related by successive inclusions can be combined into a single feature. This feature constitutes a multiple feature or multiple descriptor. Therefore, a set of features is described that group together attributes correlated by mutual inclusion. Mathematically, attribute inclusion maps pattern vectors from attribute space into a feature space of lower dimensionality.

The systematic way in which features are developed using the inclusion relation is described and applied to several examples of character reconstruction in an article by Abdali (see references in Reference 42). The method is found to be sufficient for reconstructing patterns from the features it develops.

The feature-extraction algorithm is employed in the following way. After selection of an m/e position and an intensity cutoff, a training set of molecular structures is chosen and divided into two categories. Those compounds that have a peak of sufficient intensity in this m/e position are stored as one class of patterns that is referred to as category 1. Those compounds that lack a peak of sufficient magnitude in this m/e position are stored as a second class of patterns in category 2. Each class of patterns is then presented to the feature-extraction subroutine separately. In this manner, a set of features is derived that is common to "peak" compounds, and another set is developed that is common to "no-peak" compounds.

The feature-extraction subroutine is restricted to the 40 binary descriptors of the descriptor list. The patterns presented to the attribute-inclusion algorithm, whether they belong to category 1 or to category 2, are initially composed of 40 attributes or descriptors. The algorithm begins by removing any descriptors that are not present in any of the patterns of the category being investigated. During the course of development of the features, any descriptor that appears in exactly the same patterns as another descriptor is systematically removed. This condition is known as equality. The features constructed by the algorithm may contain one descriptor, or a number of descriptors. Features that consist of only a single descriptor are normally those that appear in many of the patterns in the category. Multiple features, those that consist of more than one attribute or descriptor, are added to the end of the descriptor list. Those containing only a single attribute are not

added to the end of the list because they already appear in the list. The algorithm is applied twice, once for each category. Any multiple features, from either of the two categories, become additions to the descriptor list, beginning with the descriptor 62.

The entire set of patterns, for both the training set and the prediction set, is examined for the appearance of the multiple features. A multiple feature is present in a compound if each and every one of the descriptors in the feature appears in the molecule. The normalization constant for the multiple features is set at 5.0, the same as that for the binary descriptors.

Table 19 compares the speed of convergence for eight m/e positions using all 61 descriptors with that for all 61 descriptors and any multiple descriptors developed by the attribute-inclusion algorithm.

The first column of Table 19 lists the eight m/e positions studied. For each of the m/e positions the intensity cutoff is 0.5% of the total ion current. The third column gives predictive ability of the binary pattern classifier, using all 61 descriptors, as shown in the second column. The upper number of each pair in the third column corresponds to a starting weight vector equal to +1; the lower number corresponds to a starting weight vector equal to -1. The fourth column contains the number of feedbacks necessary for complete recognition of the training set. The fifth column gives the total number of descriptors used by the binary pattern classifier, including multiple descriptors. This value is equal to 61 plus the number of multiple descriptors constructed by the feature-extraction subroutine for the particular m/e position. The sixth and seventh columns, respectively, show the predictive ability using the multiple descriptors and the number of feedbacks for convergence. The eighth column gives the ratios of the numbers in the fourth and seventh columns. These ratios represent the increases in the rate of training resulting from the addition of multiple descriptors to the pattern vectors.

For m/e position 67, binary pattern classifiers can predict with an accuracy of 85.4 and 84.9% the appearance of a peak with an intensity greater than 0.5% of the total ion current, using 61 descriptors. Training the +1 and -1 weight vectors requires 1,244 and 1,442 feedbacks, respectively. With the addition of 27 multiple features, nearly a 50% increase in the dimensionality of the patterns,

TABLE 19

Effect of Multiple Features on Convergence Rate of Binary Pattern Classifiers<sup>a</sup>

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$m/e$	Without multiple features			Number of descriptors + multiple features	With multiple features		
	Number of descriptors	Prediction (%)	Number of feedbacks		Prediction (%)	Number of feedbacks	Columns 4/7
29	61	88.7	63	87	88.0	57	1.1
	61	88.4	77	87	89.1	51	1.5
31	61	88.9	113	90	90.2	81	1.4
	61	87.1	91	90	88.7	81	1.1
45	61	89.9	130	88	91.7	34	3.8
	61	92.0	74	88	93.3	44	1.7
59	61	82.4	2056	90	83.1	391	5.3
	61	82.8	535	90	82.1	277	1.9
67	61	85.4	1244	88	88.0	844	1.5
	61	84.9	1442	88	81.4	735	2.0
73	61	90.0	664	88	90.5	492	1.4
	61	86.3	379	88	84.5	383	1.0
77	61	91.6	137	89	90.8	119	1.2
	61	92.4	43	89	90.5	33	1.3
104	61	89.2	915	81	88.4	655	1.4
	61	87.1	464	81	87.1	331	1.4

<sup>a</sup>Intensity cutoff for each of the  $m/e$  positions is 0.5% of the total ion current.From Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 225, 1973. With permission.

the binary pattern classifier predicts the appearance of a peak with 88.0 and 81.4% accuracy using 844 and 735 feedbacks, respectively. This is an increase in the convergence rate of 50 to 100% over that required without the aid of the multiple descriptors.

On the average, the rate of convergence for the eight  $m/e$  positions is 80% faster if multiple descriptors are used to augment the original pattern vectors. This increase in dimensionality is normally in the range of 50%. Since the feature-extraction subroutine is quick, the time required for its execution is offset by the savings involved in the increase of the training rate, the slow step in the classification technique.

There is always the question as to whether a

successfully trained pattern classifier has determined real trends between data and categories. Prediction is one method of further establishing the validity of a classifier. Anderson et al. investigated this question for threshold logic units as used in linear learning machines by training on vectors filled with random numbers of Gaussian distribution.<sup>4,3</sup> Trends were observed relating the predictive ability to the ratio  $N/D$  of the training set size  $N$  to dimensionality  $D$ , and also linking the degree of category-set overlap to the frequency with which subsets are successfully trained. The results further suggested that predictive ability may not change significantly despite wide variations of  $D$  if the  $N/D$  ratio is constant and greater than two or three.



A widely different application of pattern recognition to chemistry involved the automatic generation of connectivity tables in computer-compatible format by television scanning of stylized models of molecular structures.<sup>44</sup> A major difficulty that hinders more widespread use of chemical information-retrieval systems is the rapid encoding of chemical structures into computer-compatible format. A great deal of attention has been paid to the development of search systems for retrieval of literature and information from canonical data bases. But a remaining problem is the low-cost, efficient encoding of chemical structures.

In this work, Woodward and Isenhour<sup>44</sup> interfaced a low-cost videcon television camera with a small computer and applied optical character recognition to identify the stylized characters representing a chemical structure and to generate a computer-compatible topological representation of the molecule. This system could be used as a "front end" for further programs, which would go to Wiswesser Line Notation or other linear notations. In this work pattern recognition has been used not to interpret chemical data, but rather to generate the interrogatives for a chemical information-retrieval system.

The abilities of several pattern-recognition

techniques to extract information of importance to chemists from mass spectra were compared by Justice and Isenhour<sup>45</sup> with the results shown in Table 20 and Figure 8. In Table 20 the fraction of correct prediction is recorded for each structural property and for average overall prediction by each method. The results for any one property are similar for all methods. Generally the nearest-neighbor approach gave more accurate results than methods based solely on the means of the classes. This is to be expected, because the decision-vector approach reduces the information from an entire class of vectors to a single vector, whereas the nearest-neighbor method retains all of the original vectors.

In the study of learning machines applied to mass spectra, the data set was divided into two separate data sets consisting of 387 hydrocarbons in one and 243 oxygen and nitrogen compounds in the other. This precludes direct comparison with the results in columns 1 to 5 of Table 20 because one would expect results to be better on the more homogenous data sets. The results do establish an upper bound for learning-machine prediction based on the entire data set, so that some comparisons can be made. As the overall average line in Figure 8 illustrates, the predictive ability

TABLE 20

Predictive Abilities of Six Pattern-recognition Methods

Property	Sum spectra	Binary spectra	Normalized sum spectra	Nonlinear transformation <sup>3 2</sup>	Nearest neighbor	Learning machine <sup>1 2</sup>	
						CH Compounds	CHON Compounds
Oxygen	88.9	91.0	85.6	88.3	91.6	—	93.5
Nitrogen	91.3	88.9	92.5	90.3	94.9	—	91.4
Amine	94.6	91.9	95.4	92.9	95.6	—	92.5
Carbon-4	89.1	91.3	89.7	92.1	89.4	97.3	82.8
Carbon-5	83.2	85.2	86.7	86.4	88.4	97.9	80.0
Carbon-6	80.2	83.8	84.0	85.2	89.5	94.1	83.9
Carbon-7	82.1	84.1	83.5	88.1	89.2	93.6	90.3
Carbon-8	81.9	85.6	—	85.4	88.9	92.5	87.1
Carbon-9	88.6	89.5	88.9	90.5	91.1	89.3	94.6
Double bonds							
1	82.5	82.4	82.1	82.5	86.2	77.5	88.2
2	95.2	95.1	95.2	95.2	95.7	92.5	94.6
3	98.1	98.4	98.3	98.3	97.1	98.4	94.6
Methyl	85.4	87.1	87.9	88.7	88.9	89.3	85.0
Phenyl	97.0	97.0	96.8	96.8	96.8	—	—
C <sub>n</sub> H <sub>2n</sub>	94.1	96.8	97.6	96.8	97.0	96.8	91.4
C <sub>n</sub> H <sub>2n+2</sub>	96.0	97.3	97.8	95.6	95.4	96.8	87.1
Average	89.3	90.3	90.8	91.0	92.2	93.0	89.1

91.0

From Justice, J. B. and Isenhour, T. L., *Anal. Chem.*, 46, 223, 1974. With permission.

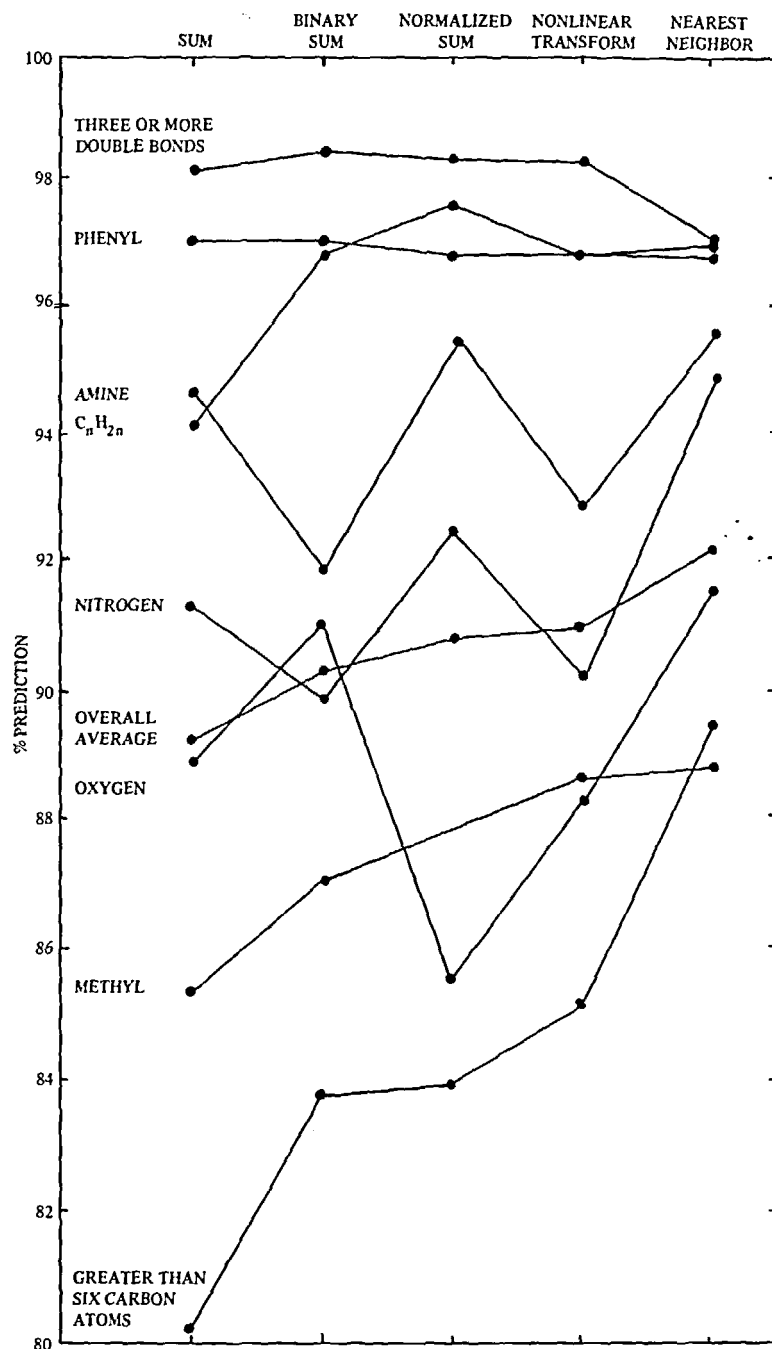


FIGURE 8. Comparison of pattern-recognition methods for classification of mass spectra (From Justice, J. B. and Isenhour, T. L., *Anal. Chem.*, 46, 223, 1974. With permission.)

increases in the following order: sum spectra < peak/no peak < normalized sum spectra < CNDF = learning machines < nearest neighbor.

Calculating the means of the original spectra gave the lowest prediction rate. That the reduction

to binary intensities improved prediction not only indicates that there is a substantial amount of information in a binary spectrum, but also suggests that the original spectra may contain much "noise" that hinders, instead of helping in, the

identification of structural properties. The relative success of each method depends on the particular structural properties studied. Peak/no peak prediction was superior to summation of the original spectra in determining the number of carbon atoms. This indicates that information related to the carbon structural frame of the molecule may be obscured by functional groups whose presence adds nonlinear noise to the relevant intensity data. Normalizing the spectra before summation resulted in improved prediction, as did applying a nonlinear transformation.

Examination of the effectiveness of the methods on individual structural properties reveals that phenyl groups, the presence of three or more double bonds, carbon/hydrogen ratio, and the presence of nitrogen and amine groups were identified with greatest accuracy. The number of carbon atoms and the presence of one double bond were lowest in predictability.

A further study of infrared-spectra interpretation using pattern-recognition methods has appeared.<sup>46</sup> The data for this study were taken from the Sadtler Standard Infrared Spectra and consisted of 212 spectra for small organic molecules in the ranges  $C_{3-10}H_{2-22}O_{0-3}N_{0-2}$ . Each spectrum was divided into  $0.1\text{-}\mu$  intervals across the  $2\text{--}15\text{-}\mu$  region to give 131 intervals and 131 x-y values. The transmittance in each interval was

converted into absorbance and put onto a 0-to-9 scale for convenience.

Table 21 shows the results obtained using the error-correction method with two dead zones for seven representative chemical classes. The "benzene" class contains compounds having benzene rings incorporated in their structures. The weight vector components were initialized with values of 0.1 for training one binary pattern classifier and with values of -0.1 for training a second, and prediction is therefore given as average percent prediction. The binary pattern classifiers whose prediction abilities are shown in the fourth column were trained with spectra containing 131 features. In every case the two sets of spectra were separated. Predictive abilities ranged from 70% to 100%. These high predictive abilities can be observed even though the situation is undetermined (there are more components per pattern than patterns in the training set), because the weight vectors are constrained to be linear summations of the patterns in the training set. The overall prediction is usually improved as the dead zone is increased, but there is a point where an increase in the dead zone will not improve the prediction. Normally one would think that the predictive ability would keep improving, but after a certain point is reached, error-correction-feedback method will only increase the absolute magnitudes of the

TABLE 21

Results of Error-correction-feedback Training and Feature Selection

	Training set size	Prediction set size	Dead zone	Average prediction (131 features) (%)	Number of features retained	Average prediction (%)
Carbonyls	90	122	0	94.6	40	98.4
			50	100.0	70	99.2
Cyclohexanes	23	189	0	76.0	52	73.0
			50	69.9	87	71.2
Alcohols	41	171	0	91.0	27	95.1
			50	95.0	86	95.6
Ketones	26	186	0	81.4	18	85.2
			50	87.6	79	89.5
Esters	41	171	0	91.6	38	94.4
			50	95.6	68	96.2
Benzene	49	163	0	86.9	47	90.5
			50	98.9	93	90.5
Ethers	22	190	0	90.0	41	93.4
			50	99.0	66	98.4

From Liddell, R. W., III and Jurs, P. C., *Appl. Spectrosc.*, 27, 371, 1973. With permission.

weight components of the weight vector. By doubling these components, all the scalar products are increased by a factor of 2; there will be no improvement in the decision surface, even though the numerical separability of the positive and negative elements is increased by this factor of 2.

The fifth and sixth columns show the results of using the feature-selection technique with the same data sets. For each row the following procedure was employed: two weight vectors (with all components initialized at 0.1 and -0.1, respectively) were trained to correctly classify all the members of the training set. Then all the descriptors for which the weight vector components disagreed in sign were discarded and the cycle was repeated a second time. This was repeated until no more descriptors could be discarded, and then the procedure was terminated. The number of feedbacks needed for training increased as a function of the size of the dead zone used. Moreover, as the width of the dead zone increased, it became more difficult to decrease the number of features; it is best to find a good compromise between feature selection and predictive ability by using a dead zone of approximate width. Table 21 shows that the number of features necessary to classify an infrared spectrum into one of these chemical classes can be reduced substantially from the original 131 with little loss of predictive ability.

Comparing the values in the fourth and sixth columns of Table 21 shows that the predictive ability does not decrease after feature selection has been completed; in fact, in the majority of cases the predictive ability is slightly improved. This seems logical enough, for certain peaks that are not useful in the prediction process for a certain chemical class will, under a learning-machine approach, influence the prediction decision. After feature selection, however, these irrelevant peaks will have been eliminated and cannot affect the decision surface.

Comparing characteristic group frequencies, from tables in the literature, with the learning-machine weight vectors after feature selection has been completed, one obtains striking correlations between the tables and the highly positive weight components of the weight vector. These are readily observable, for example, with alcohols and carbonyl compounds. Most alcohols have a strong peak between 2.7 and 3.1  $\mu$ , whereas the carbonyl compounds have strong peaks in the 5.7–6.2- and

7.6–8.4- $\mu$  intervals. For the error-correction-feedback method without a dead zone the alcohols show positive weight components between 2.9 and 3.1  $\mu$ , whereas the carbonyl compounds have positive weight components between 5.8 and 6.0  $\mu$  and between 7.6 and 8.1  $\mu$ .

The remainder of the interpretation of infrared spectra was done with a new algorithmic training routine that also performs feature selection as it trains. A sample of the capabilities of this routine may be seen in Table 22.

Only a few years have passed since pattern recognition was first seriously applied to chemistry. The variety of these applications in this short period can only indicate that much more usage will be made of pattern recognition in the future. It is difficult to guess where the limit lies. Perhaps the researcher will eventually be able to call on a variety of techniques available as software on some automatic system and just wait for the answer to be returned. However, it is also possible that the subject will remain specific to the extent that each problem will require an individual treatment by an expert well versed in the area.

In any event it is certain that more of these applications will be forthcoming. As software becomes more sophisticated and computers more easily available, pattern-recognition systems should become widespread. Perhaps general approaches will emerge for certain areas, such as spectroscopy.

There remains the question, though, whether

TABLE 22  
Classifications Performed by MAX

	Peaks input	Peaks retained	Prediction (%)
Carbonyls	131	42	97.5
Cyclohexanes	131	54	80.4
Alcohols	131	29	98.3
Ketones	131	36	87.6
Esters	131	54	95.9
Benzene	131	35	92.0
Ethers	131	36	93.7
Carbonyls	40	13	99.2
	70	38	98.4
Cyclohexanes	52	26	71.4
	87	35	66.1
Alcohols	27	14	97.7
	86	28	97.1

From Liddell, R. W., III and Jurs, P. C., *Appl. Spectrosc.*, 27, 371, 1973. With permission.

any general approach to problem-solving will come from pattern recognition. It is currently accepted that the only way to find the best pattern-recognition method for a given problem is to try them all.

While it is not possible to project its final place in chemistry, it is safe to say that pattern recognition is here to stay and is likely to solve more important problems in the future than it has in the past.

## REFERENCES

1. Isenhour, T. L. and Jurs, P. C., *Anal. Chem.*, 25, 20A, 1971.
2. Isenhour, T. L. and Jurs, P. C., in *Application of Computer Techniques in Chemical Research*, Hepple, P., Ed., Institute of Petroleum, London, 1972.
3. Kowalski, B. R. and Bender, C. F., *J. Am. Chem. Soc.*, 94, 5632, 1972.
4. Jurs, P. C., *Japan Analyst (Bunseki Kagaku)*, 21, 1276, 1972.
5. Isenhour, T. L. and Jurs, P. C., in *Computers in Chemistry and Instrumentation*, Vol. 1, Mattson, J. S., Mark, H. B., Jr., and McDonald, H. C., Jr., Eds., Marcel Dekker, New York, 1973.
6. Kowalski, B. R., in *Computers in Chemical and Biochemical Research*, Vol. 2, Klopfenstein, C. E. and Wilkins, C. L., Eds., Academic Press, New York, 1974.
7. Tal'roze, V. L., Raznikov, V. V., and Tantsyrev, G. D., *Dokl. Akad. Nauk SSSR*, 159 (1), 182, 1964.
8. Raznikov, V. V. and Tal'roze, V. L., *Dokl. Akad. Nauk SSSR*, 170 (2), 397, 1966.
9. Drozdov-Tikhomirov, L. N., *Opt. Spectrosc.*, 27, 77, 1968.
10. Crawford, L. R. and Morrison, J. D., *Anal. Chem.*, 40, 1464, 1968.
11. Crawford, L. R. and Morrison, J. D., *Anal. Chem.*, 40, 1469, 1968.
12. Jurs, P. C., Kowalski, B. R., and Isenhour, T. L., *Anal. Chem.*, 41, 21, 1969.
13. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., and Reilley, C. N., *Anal. Chem.*, 41, 690, 1969.
14. Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilley, C. N., *Anal. Chem.*, 41, 695, 1969.
15. Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilley, C. N., *Anal. Chem.*, 41, 1945, 1969.
16. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., and Reilley, C. N., *Anal. Chem.*, 41, 1949, 1969.
17. Wangen, L. E. and Isenhour, T. L., *Anal. Chem.*, 42, 737, 1970.
18. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., and Reilley, C. N., *Anal. Chem.*, 42, 1387, 1970.
19. Jurs, P. C., *Anal. Chem.*, 42, 1633, 1970.
20. Jurs, P. C., *Anal. Chem.*, 43, 22, 1971.
21. Sybrandt, L. B. and Perone, S. P., *Anal. Chem.*, 43, 382, 1971.
22. Wangen, L. E., Frew, N. M., Isenhour, T. L., and Jurs, P. C., *Appl. Spectrosc.*, 25, 203, 1971.
23. Kowalski, B. R. and Reilly, C. A., *J. Phys. Chem.*, 75, 1402, 1971.
24. Wangen, L. E., Frew, N. M., and Isenhour, T. L., *Anal. Chem.*, 43, 845, 1971.
25. Jurs, P. C., *Appl. Spectrosc.*, 25, 483, 1971.
26. Frew, N. M., Wangen, L. E., and Isenhour, T. L., *Pattern Recognition*, 3, 281, 1971.
27. Jurs, P. C., *Anal. Chem.*, 43, 1812, 1971.
28. Kowalski, B. R. and Bender, C. F., *Anal. Chem.*, 44, 1405, 1972.
29. Kowalski, B. R., Schatzki, T. F., and Stross, F. H., *Anal. Chem.*, 44, 2176, 1972.
30. Lytle, F. E., *Anal. Chem.*, 44, 1867, 1972.
31. Justice, J. B., Jr., Anderson, D. N., Isenhour, T. L., and Marshall, J. C., *Anal. Chem.*, 44, 2087, 1972.
32. Pietrantonio, L. and Jurs, P. C., *Pattern Recognition*, 4, 391, 1972.
33. Tunnicliff, D. D. and Wadsworth, P. A., *Anal. Chem.*, 45, 12, 1973.
34. Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 30, 1973.
35. Kowalski, B. R. and Bender, C. F., *J. Am. Chem. Soc.*, 95, 686, 1973.
36. Bender, C. F. and Kowalski, B. R., *Anal. Chem.*, 45, 590, 1973.
37. Bender, C. F., Shepard, H. D., and Kowalski, B. R., *Anal. Chem.*, 45, 617, 1973.
38. Leary, J. J., Justice, J. B., Tsuge, S., Lowry, S. R., and Isenhour, T. L., *J. Chromatogr. Sci.*, 11, 201, 1973.
39. Kowalski, B. R. and Bender, C. F., *Anal. Chem.*, 45, 2334, 1973.
40. Felty, W. L. and Jurs, P. C., *Anal. Chem.*, 45, 885, 1973.
41. Kuo, D. A. and Jurs, P. C., *J. Am. Water Works Assoc.*, 65, 623, 1973.
42. Schechter, J. and Jurs, P. C., *Appl. Spectrosc.*, 27, 225, 1973.
43. Anderson, D. N. and Isenhour, T. L., *Pattern Recognition*, 5, 249, 1973.
44. Woodward, W. S. and Isenhour, T. L., *Anal. Chem.*, 46, 422, 1974.
45. Justice, J. B. and Isenhour, T. L., *Anal. Chem.*, 46, 223, 1974.
46. Liddell, R. W., III and Jurs, P. C., *Appl. Spectrosc.*, 27, 371, 1973.