PATTERN RECOGNITION TECHNIQUES

FOR USE WITH

^{252}Cf-PLASMA DESORPTION MASS SPECTROMETRY

BY

ANDREW M. WARD JR.

SUBMITTED TO THE DEPARTMENT OF

BIOENGINEERING

TEXAS A&M UNIVERSITY

APRIL 1976

TABLE OF CONTENTS:

I. Introduction 1

II. Literature Review 2

III. Theory 10

IV. Data Input Format 13

V. K-Nearest Neighbor 14

VI. Binary Pattern Classifier 17

VII. Literature Cited 28

INTRODUCTION:

Pattern recognition, as used in this paper, reflects a technique of machine data analysis. In this capacity the machine receives the raw data, processes it for the purpose of separating the important features from the background, and finally makes a decision which it bases on these features. For most purposes the decision is categorical (i.e. does compound A contain nitrogen?; is a particular peak indicative of an arginine-guanine complex, etc.). Since the author has an independent interest in the application of pattern recognition to the analysis of mass spectrometry data, this vehicle will be used to describe the techniques of applying pattern recognition in the laboratory. It should be kept in mind that these techniques may be applied to any data source that meets or can be made to meet the machine data format.

LITERATURE REVIEW

Pattern recognition (PR) is a technique of machine data analysis. At this time it is difficult to formulate a more precise definition, since PR seems to mean somthing different to each author. This delima is graphically demonstrated in an article by Verhagen (1). He states, "A survey... ...of definitions and descriptions, taken from the literature, concerning the terms: patern; recognition; pattern recognition; and related terms... appear not to have an identical meaning with different authors." With this state of affairs it is necessary to risk a couple of definitions so that pattern recognition may be discussed in general.

Grenander in his "Foundations of Pattern Analysis" presents an excellent definition of what consititutes a pattern (2).

> "By patterns we shall understand the following. Starting from a set of objects (called images) and a set of rules by which we can transform an image into others, we shall say that two images are similar if one can be transformed into the other by applying some of our rules successively. By a pattern we could mean a class of mutually similar images."

Closely allied with Grenander's definition of a pattern is Sebestyen's concept of PR. In his excellent work "Decision-Making Process in Pattern Recognition" he states:

> "PR is a process of decision making in which a new input is recognized as a member of a given class by comparison of its attributes with the already known pattern of common attributes of members of that class."

In view of the diversity of PR literature this author feels that it is wise to narrow the survey to those applications of PR that deal with information processing in the laboratory. This limit is sufficiently broad as to place no restriction on the techniques used by the researcher.

The choice of PR technique is very sensitive to the format and

source of the data. This is made very clear in Calvert and Young's text "Classification, Estimation and Pattern Recognition"(4). This volume coordinates the techniques developed by industry and researchers to overcome PR problems in their area. The basic theory of operation of each technique is explained in detail, though the reader is advised to consult the primary sources once a method has been selected. The references cited in this work are excellent and thorough, being the most up to date of all texts consulted. The reader seeking only a superficial knowledge of PR is advised to read a short, but concise article by L.R. Carlson, et al (5) which covers much of the material in Calvert and Young's book. The detail is not near as complete, yet the reader undoubtedly profits from his discussion of PR techniques. On a more practical tangent J.B. Justice and T.L. Isenhour rate the effectiveness of six different PR techniques on their ability to extract information from a set of mass spectra (6). Of the methods examined, the K-nearest neighbor and the Binary Pattern classifier were given the highest marks. Also discussed was the application of data transformations to improve the performance and predictive ability of these two methods. Data transformations improve the sensitivity of the classifier by extracting features from the raw data. Features may be considered a subset of the pattern that allows the pattern to be classified. In general they are functions of the measurements and have a lower dimensionality than the raw data. This lower dimensionality aids in classification and reduces the size of the data base. Feature extraction theory is treated very thoroughly in Calvert and Young's book (4) and an example of application is given in P.C. Jurs article (17). It should be noted that the ideal feature set will contain the same amount of informa-

tion as the original data set.

There exist many techniques that could be applied to the analysis of mass spectra, but most of these demand large data bases and extensive use of the computer for their proper function. Such requirements are a great disadvantage when the data originates from a new source such as the Plasma Desorption Mass Spectrometer (PDMS) developed by R.D. Macfarlane and D.F. Torgerson (21). Here the principal inconvenience is the large dimensionality which leads to the two disadvantages cited above. It is therefore thought wise to adopt the conclusion reached by Justice and Isenhour that the K-nearest neighbor and the binary pattern classifier give the best results for data of this type. This decision is justified as both methods are linear, non-parametric decision machines that do not assume knowledge or existance of distribution statistic beyond the requirement of linear separatability. Even this requirement may be relaxed under certain conditions as explained in N.M. Frew, et al's article on the application of piecewise-linear pattern classifiers to mass spectra (22).

These two methods fall into a larger collection of methods (or machines) known as linear pattern classifiers. They should be very attractive to the reasearcher that can not justify a rigerous statistical approach, but could support a decision to train a pattern classifier. The requirements of these machines as well as the theory and development that supports them can be found in two small bookds authored by men who did the most to forward their progress.

The first book is Learning Machines by N.J. Nilsson (7). This is the most often cited work this author has encountered and it is somewhat a classic in the field of PR. In this work Nilsson outlines the various types of linear pattern classifiers and gives detailed discussions on the

methods used to train them. The second text is authored by G.S. Sebestyen. In his book Decision-Making Processes in Pattern Recognition, while the author develops some of the linear pattern classifiers, his greatest contribution to the field deals with data and feature transformations (8). This work was originally written for communications engineers but has since found a place in many a researcher's library if its citation frequency is a good indicator as such. Both books share a common virture of clarity and compactness (less than 180 pages) not found in the other works.

There is also a third text edited by J.T. Tou and R.H. Wilcox, which contains many articles that are often cited by writers in the PR field. This work: Computer and Information Sciences, appeared in 1964 after a symposium sponsored by the U.S. government "to fill a gap in the information sciences created by WW II." The text was also used as a reference at the office of Naval Research.

The literature of the K-nearest neighbor technique that deals with its application to laboratory data is rather limited. This is due to the large number "knowns" that must be retained for classification purposes. While this method is excellent for the identification of postal "ZIP" codes, it fails to be employable in the laboratory for economic reasons.

The K-nearest neighbor method was formulated by Fix and Hodges (10, 11); reviewed by Nilsson (7), Sebestyen (8) and T.M. Cover with P.E. Hart (12). While laboratory applications are few (25,36,37), the K-nearest neighbor method does have an extensive literature in the communications field. Nilsson has shown how this method is actually a special case of the more general Binary Pattern Classifier.

The Binary Pattern Classifier was concieved in 1943 by McCulloch and

Pitts (13) as a mathematical model of the neuron and was shelved for almost eighteen years until Rosenblatt (14) placed it on a more rigorous mathematical basis. It was also Rosenblatt who first suggested that it could be used to classify patterns. Early literature on the Binary Pattern Classifier bears the name Rosenblatt gave it, the Alpha-Perceptron; this should be kept in mind when surveying the literature. Of the two techniques, this one seems to be the most applicable. It has the virtue of simplicity and versatility. Further, it has a well developed literature in the field of laboratory applications. In the arena of mass spectra applications a great bulk of material is available due to the combined efforts of P.C. Jurs, T.L. Isenhour, and B.R. Kowalski. These three men have a virtual monopoly on the application of Binary pattern classifiers (BPC's) to the analysis of I.R., N.M.R. and Mass spectra. Their accomplishments with this machine have been rather impressive. It has been possible to predict the molecular formula of an unknown with an accuracy of 88 to 90% (15,16). Determination of molecular structure of compounds containing only C,N,O, and H, has been accomplished using low resolution mass spectra without recourse to theory (17,18,19). During the course of the experiment it was possible to obtain predictive capability of 98% (18) by the use of a Commitie machine (see Nilsson ref.7) and 95% with the faster Branching tree technique (19).

Other articles that deal with the application of the binary pattern classifier to spectral data are; Jurs use of it to classify and interpret I.R. (23); the use of the BPC to simulate the spectra of small organic molecules (26); the combination of the BPC with the method of least Squares in the development of discriminant functions (24). An assumption

which is inagral to the use of PR to mass spectra is: similar compounds have similar spectra. This has been demonstrated in part by all the articles cited above and also by K.L.H. Ting, et al, who used mass spectra to classify drugs into groups of sedatives or tranquilizers (25). His group used the K-nearest neighbor method first with a dimensionality of 539, which proved difficult to employ and then with data sets of reduced dimensionality. Mapping on to two dimensions was also explored though the results are rather dubious since they rely on a supervisor to oversee the mapping. It has also been possible to combine layers of BPC's into a machine that can grasp its situation. "To be concrete, the machine can correctly recognize patterns having multible meanings as O("Oh" or zero) as the situation may demand." It would seem that the machine accomplishes this feat by increasing the dimensionality of the data (or pattern) and weighing these added dimensions rather heavily. Takashi Nagano also gives an explanation for the machines behavior and insight into the training methods used on layered machines (27). A non-academic application of the BPC of some intrest is discussed by J. Felsen. Felsen spent some time applying the BPC to stock market forcasting with some impressive results. The success of the machine is due to its ability to reduce information that is to complex for the human mind to handle into a form that may be comprehended. This reduced information may then serve as the basis for investment decisions (28).

Having a pattern classifier is often not enough. The K-nearest neighbor method, for example, is very effecitve on data of any dimension; but its storage requirements force one to decide against it. A machine that could select the important features from the data base and in so doing reduce the memory demand, would make this a more practical

alternative. In any case a reduction in dimensionality will have an excelerating effect on the machines performance, regardless of type. For this reason, it is worthwhile to review some of the literature on feature selection. The referance of first choice is Calvert and Young's text (4). Other articles which deal with this problem are those by Jurs which deal with the application of the BPC and the Fourier transform to efficient feature selection (29,30). C.H. Chen explores a recursive computational procedure for feature selection (31). He states in his conclusions that under certain conditions a significant savings in computer time may be obtained by using this algorithm. Finally, if the pattern is thought to represent a signal that is received through a noisy channel then some of the techniques developed by the communications engineers may be employed. Under these conditions texts such as G. Raisbeck's Information Theory (36) or articles like S.A. Kassam's Asymptotically Robust Detection of a Known Signal in Contaminated Non-Gaussian Noise (32).

There are several other PR methods that have been used in the study of mass spectra. One of the most impressive of which is Factor Analysis. Factor Analysis was developed originally for the study of data that came from psycological studies where the relationships between the results and the stimuli are not lucid. As Jurs describes Factor Analysis;

> "The lineraly independent dimensions are determined by Factor Analysis, a multivariate statistical method for studying the nature of high dimensionality data from a large experimental data set. The dimensions are initially the eigenvector solution that diagonalizes a correlation matrix of the experimental variables (the mass spectra). This eigen vector solution may then be rotated to clarify the interpretation of the dimensions with respect to the masses (33)."

In his conclusions Jurs stated that it was possible to establish a clear relationship between functional groups and the mass spectra. Another

article which came out only recently deals with the application of factor analysis to predict physical properties. In his article D.G. Howery explains how this type of statistical analysis is used and gives several examples where it has been used by researchers in the laboratory (34).

The fitting of high order polynomials is a PR method that has found use in modeling environmental systems (3).

Carl Djerassi is another researcher who has applied learning machines to the analysis of mass spectra. His field of interest centers on the study of spectra by relating the spectra to various chemical compounds on the basis of predicted fragmentation patterns (35).

The last few references are cited out of a desire for completness. The deal with techniques for coding mass spectra for the purpose of future comparison (i.e. matching). Stanley Grotch appears to be the expert when it comes to developing search routines and his two articles on the subject are often cited by aurhors in the PR field (39,40). P.C. Jurs has also written an article on the application of hash coding to obtain optimum searching of information files (38).

The conclusion of this review is: Pattern Recognition has shown to be effective in the study of mass spectra. It has sufficient power to become an important analytical tool in the hands of the research scientist and will become more common as computer technology becomes more refined and sophisticated.

THEORY:

Pattern recognition is a process that has reached a very high level of sophistication in nature. Every animal is skillful in the recognition of food, and, where necessary, members of the opposite sex. This function is performed automatically and without conscious thought by: observing the environment, subdividing it into sets of known images, and then weighing these images in terms of their relative importance to the problem at hand (i.e. food, shelter, protection, etc.).

This process can best be understood by imagining a young man, who is looking for his wife in a large crowd. He stands there at the edge and scans each face. Without thinking he compares each individual to a cognitive image of his love, then at some instant the realizes that the face he was looking at matches the image in his mind. Unknown to this man, he has made a series of measurements on each person. The mode of dress, or the length of hair was, perhaps the first item compared. If these matched, his mind would proceed further, looking at more subtle details, such as: eyes, nose, chin, etc. This process of comparison and elimination was continued until all the important and unique characteristics of the image of his love and the person he was viewing, agreed. At that instant he realized that he was looking at her. This is an example of a very sophisticated pattern recognition process.

Granted, it may be absurd to compare a simple thing like boy meets girl to the more mechanical aspects of pattern recognition, but it does serve to illustrate a point; animals and people are capable of recognizing elements or patterns in their environment. If it is assumed that there is nothing divine about this process, then it should be possible to

devise a machine that would mimic this function, if only in a limited way. Machines capable of such 'intelligent' processes would become indispensable in the analysis of data. The diagnosis of heart conditions from EKG data or the identification of the components in a mixture from a spectroscopic study, are examples where this procedure has been employed with excellent results. *(4)

Consider, for example the situation of a research chemist trying to determine the structure of an unknown compound. There are galaxies of tests that could be run to give the desired information, but they all take time and skill that he may not have. One way to avoid this situation is to employ some of the more powerful analytical methods which are very simple to run (i.e. I.R., N.M.R., or Mass Spectrometry). These methods give accurate results that contain a great deal of information, but they have the disadvantage of telling the researcher too much at one time. The output issued from these machines could be compared to several tables of physical data on the sample, printed on transparency film and then stacked on top of one another. This may sound odd but it is certainly what happens when data is gathered from these devices. As a consequence, these devices are not employed to their fullest extent. Often, only one or two pieces of information can be extracted from the maze of result; but this is not due to a defect in the device; it is a failure of the researcher.

In theory, it is possible to devise a machine that will separate the data into categories that are meaningful to the researcher. Such a machine is called a pattern classifier and the process that is employs is

---

* As an aid to the reader, this author will use the field of mass spectrometry to relate the techniques of pattern recognition in a simple manner

called pattern recognition. When data is fed into this machine, it is scanned in much the same way as the man surveyed the faces in the crowd. Important facts are noted until a sufficient number are acquired to identify the compound. It should be noted that the method of 'acquiring the facts' is all that separates the different techniques because the pattern recognition procedure is versatile and can be used on any data that meets its input format. For the sake of brevity, this paper will cover only two of the many methods that fall under the heading of pattern recognition. These two techniques are: the k-nearest neighbor method, which was developed from the Euclidian distance formula, and the binary pattern classify, which was developed from an exploitation of the vector dot product. It should be noted that these methods require large training sets and are nonparametric; hence, it is very difficult to establish as error estimate. (7) Still it is believed that their speed and ability to handle non-linear multicategorical data more than compensates for these drawbacks.

The overall process of pattern recognition can be broken down into several steps. The first is the acquisition of raw data. The data is then transformed into vector format which is suitable to the pattern classifier. This vecotr may then be processed to enhance specific features (this step is called processing and is optional). The resulting vector is then fed into the machine and processed to give a result. For most laboratory purposes the result will be a classification of the unknown into one or more known groups. Figure 1 outlines the overall layout of the machine.

DATA INPUT FORMAT:

The normal input format for pattern recognition is the pattern vector. To generate the pattern vector for a mass spectrum, it is necessary to select an interval size that is on half the desired resolution and then calculate the absolute frequency for each interval. Next this table of absolute frequency is normalized to give the relative frequency in terms of the mass spectrum; if an interval size of 1 AMU was used, each interval would contain the normalized value of the ion flux at that mass and the resultant sequence could be written:

$$X = (x_1, x_2, x_3, \ldots x_n)$$

Where X = symbol for the sequence (or pattern vector).
x_i = the intensity of the ith interval (relative frequency). The subscript, in this case, is equal to the corresponding mass for the interval.

This resultant sequence is the pattern vector. The pattern vector is uniquely determined from the mass spectrum and can represent the mass spectrum without information loss.

Transforming the mass spectrum into a vector permits the application of vector analysis to the problem of pattern recognition.

K-NEAREST NEIGHBOR:

The decision rule for the K-nearest neighbor method is: Given several categories, each composed of a set of known compounds represented by their vectors, than an unknown will be classified into a particular category that has the lowest mean scalar difference between it and the vectors of that category.

$$d_i < d_{j \neq i} \tag{1}$$

The scalar difference between two vectors is given by the Euclidian distance formula:

$$d_I = \{\Sigma (x_k - I_{jk})^2\}^{\frac{1}{2}} \tag{2}$$

Where: x_k, = the kth component of the unknown pattern vector

I_{kj}= the kth component of the jth known vector.

Therefore, when two mass spectra are very similar they will have a vector difference of low magnitude. In practice, it is often more useful to have a decision rule that is based on an absolute maximum, than one based on a relative minimum. Such a rule may be derived from the above rule by simple manipulation. If the above expression is squared and expanded then the following results:

$$\begin{aligned} d_i^2 &= \Sigma (x_k - I_{kj})^2 \\ &= \Sigma (x_k^2 - 2x_k I_{kj}^2) \\ &= \Sigma x_k^2 - 2\Sigma x_k I_{jk} + \Sigma I_{jk}^2 \end{aligned} \tag{3}$$

Now, it is recalled that the vector dot product is defined to be:

$$X \cdot B = x_1 b_1 + x_2 b_2 + x_3 b_3 + \ldots + x_n b_n \tag{4}$$

By incorporation of the vector dot product into equation three a simplified equation results:

$$d_j^2 = \overline{X} \cdot \overline{X} - 2\overline{X} \cdot \overline{I}_j + \overline{I}_j \cdot \overline{I}_j \tag{5}$$

By rearrangement a new function may be defined;

$$(X) = -(d^2 - \overline{X}\cdot\overline{X})/2$$

$$= \overline{X}\cdot\overline{I}_j - \tfrac{1}{2}\overline{I}_j\cdot\overline{I}_j \qquad (6)$$

A further simplification will result if the augmented vectors for $\overline{X}$ and $\overline{I}_j$ are formed. The augmented vectors are:

$$\overline{X}^* = (x_1, x_2, x_3, .., x_n, 1) \qquad (7)$$

$$\overline{I}^*_j = (I_{1j}, I_{2j}, I_{3j}, .., I_{nj}; \tfrac{1}{2}\overline{I}_j\cdot\overline{I}_j)$$

This new function has its greatest value when $\overline{X} = \overline{I}_j$ (i.e. $K_j(\overline{X}) = \tfrac{1}{2}\overline{I}_j\cdot\overline{I}_j$)
The corresponding decision rule is: $\overline{X}$ is placed in that category which has the greatest mean value for the function $K_j(\overline{X})$. That is:

$$(M_k^{-1}) \Sigma K_{jk}(\overline{X}) < (M_{t\neq k}^{-1}) \Sigma K_{jt}(\overline{X}) \qquad (8)$$

Where: M_k = the number of elements in the kth category

j = a dummy variable of summation.

This rule also has the advantages: greater speed during execution and lower storage requirements.

One of the disadvantages of the K-nearest neighbor method is the assumption that all intervals are of equal value for each category. In the laboratory this may not be the case. Some intervals may have values that deviate greatly. While others may have values that are very compact. The same may be said of the categories. In one group the points may be clustered very tightly together, and in another group these points may be dispersed. Under these conditions a vector X may actually belong to group 1, which is very dispersed, but may be classified into group 2, which is very dense and thus, gives a high $K_j(\overline{X})$.

This condition of varying density may be corrected by weighing the intervals in such a manner to produce a constant density cluster. The

mechanism for this weighing is a coordinate transformation which minimizes intra-cluster distances while maximizing the inter-cluster distance (care is taken to prevent the trivial transformation) (8).

BINARY PATTERN CLASSIFIER:

The binary pattern classifier is a classification machine that gives a positive one (+1) if the unknown compound is to be classified as a member in the group of knowns that possess the characteristics for which the machine was trained and a negative one otherwise. The two principle assumptions used to derive this machine are:

(1) Compounds that have similar physical characteristics will have certain peak characteristics in common.

(2) These similar characteristics will cause the compounds mass spectrum to cluster in the pattern space in a linearly separable fussion.

The first assumption appears to have been validated by K.L.H. Ting and co-workers in a series of drug studies where physiological properties were correlated to mass spectrum shape (25). They were able to classify several unknowns into their drug classes based on mass spectrum alone. The second assumption has been validated by P. C. Jurs, B.R. Kowalski, and T.L. Isenhour. Their work with the binary pattern classifier has enabled them to determine the carbon number, elemental composition, and presence of functional groups (17, 18, 19).

Some of the advantages of the binary pattern classifier are:

(1) its ability to be trained to recognize small portions of mass spectrum that relate to the class automatically.

(2) automatic weighing procedure which selects those features of importance.

(3) fast classification.

(4) may be operated in a supervised or unsupervised manner.

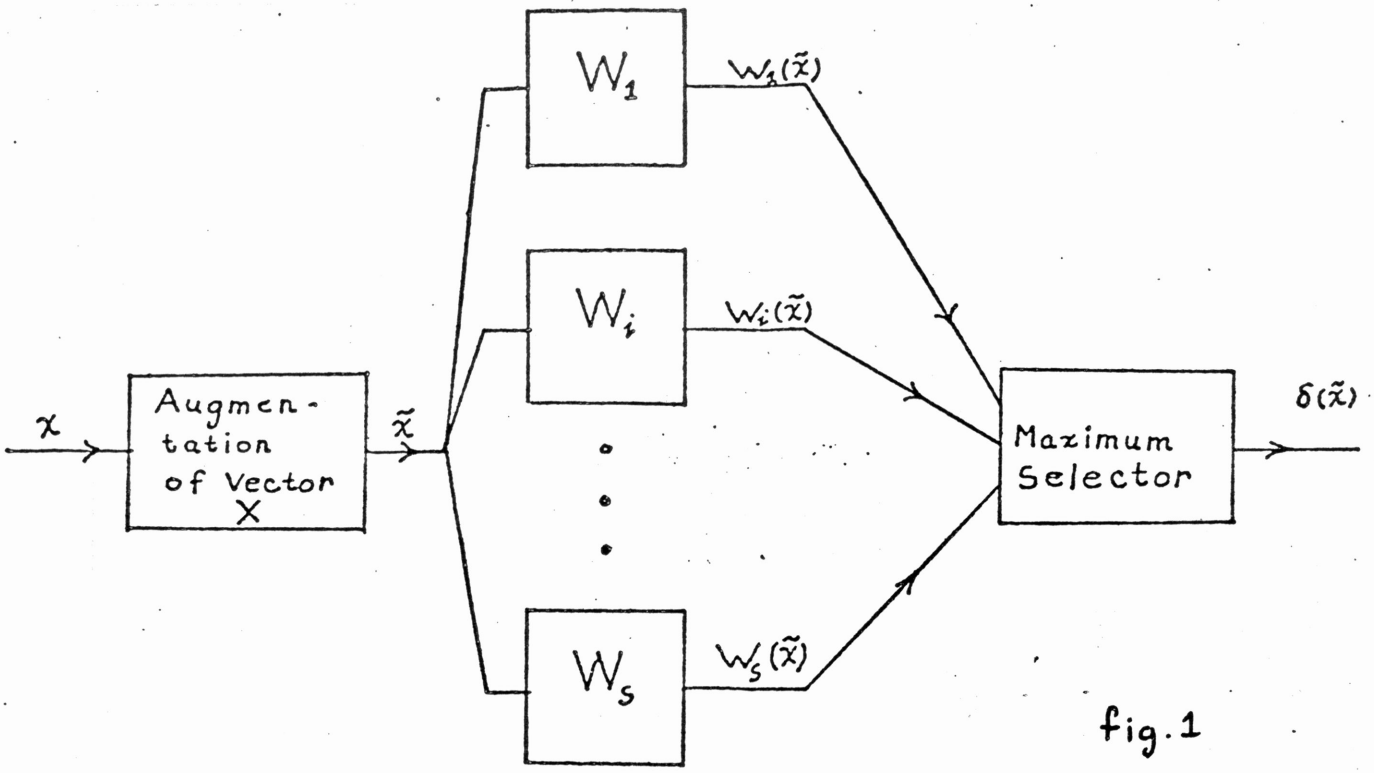


fig. 1

A Pattern Classifier for Multiple Categories

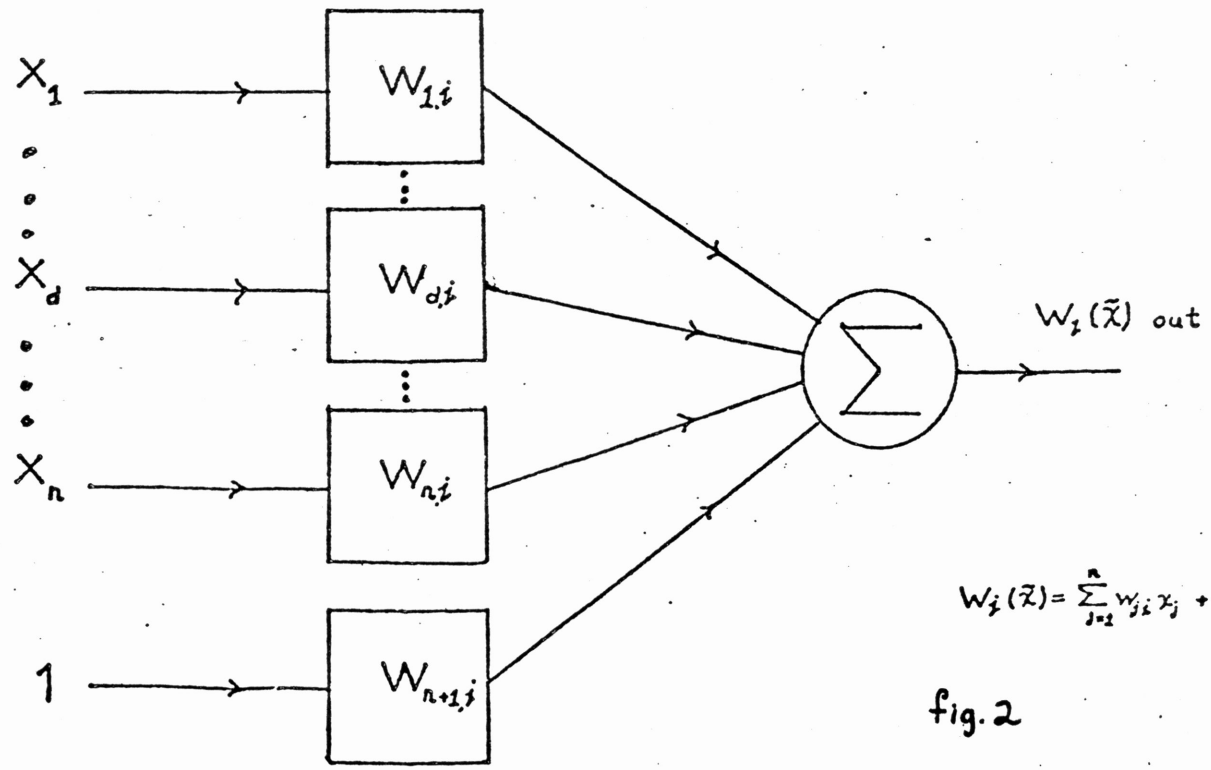


fig. 2

Binary Pattern Classifier

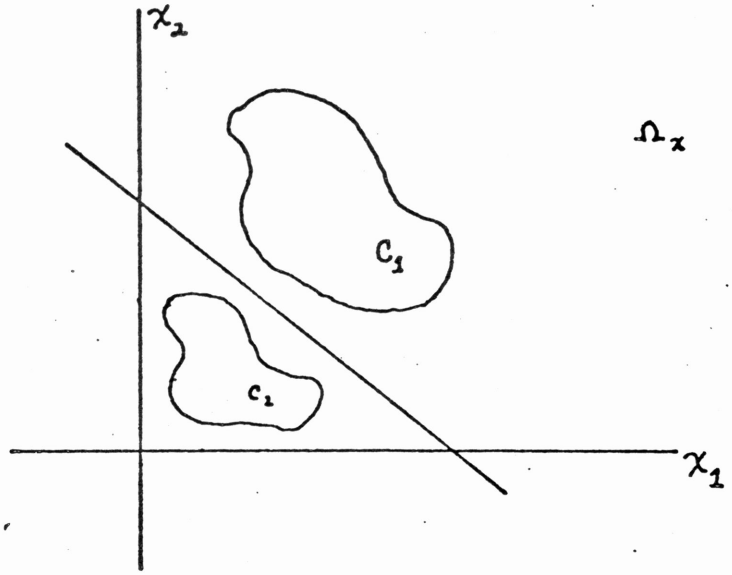
(5) binary pattern recognition may be layered and/or combined to give results for data sets that are not convex.

The main disadvantage is the need for a large training set if the results are to be meaningful.

Calvert and Young demonstrated that the pattern classifier will converge for any training set that has fewer elements than the pattern vector has dimensions (4). The pattern classifier will only be practical if it has been trained on such a large set, that it contains a representative sample of real world data. The effect of this disadvantage can be minimized if the machine is updated after each classification. Such a procedure will permit the user to employ the machine in a secondary capacity until it has obtained the ability to properly classify the data. The machine therefore becomes more precise, the more it is used.

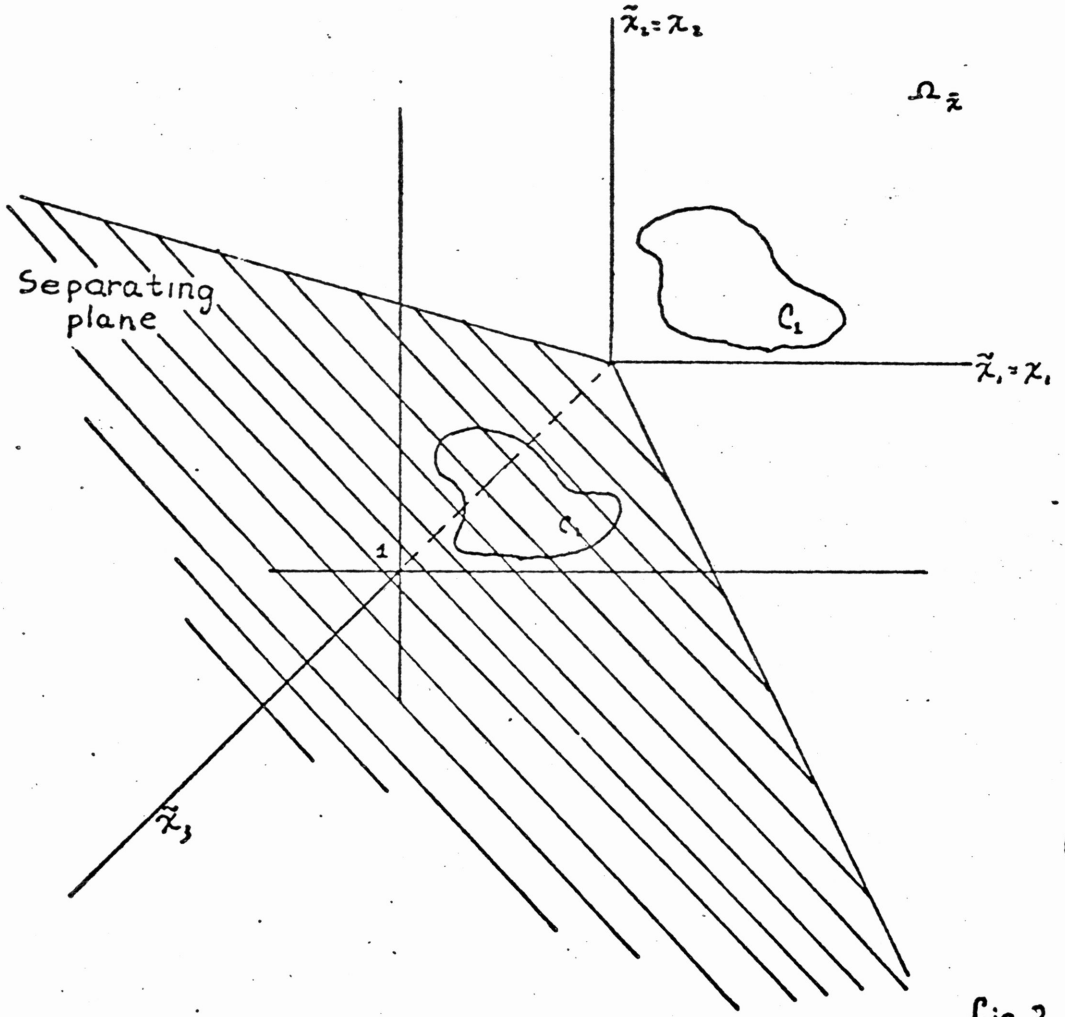
To aid in the understanding of the binary pattern classifier, it will be assumed that for a given mass spectrum it is only necessary to look at two of its many components (X_1,X_2) to determine the presence of arginine. Now, if the plane of all possible values of X_1,X_2 is drawn and the cluster of compounds that do not contain arginine are labeled 'B' and the two assumptions of the binary pattern classifier are fulfilled, then figure 3 results. Assumption two permits a line to be drawn, that separates the plane into two regions. If the decision rule is modified to classify the unknown into one category or the other depending in what region the unknown appears, then the equations for the binary pattern classifier may be derived as follows. If the equation for the separating line is:

$$ax_1+bx_2 + d = 0 \qquad (9)$$



(a)

LINEARLY SEPARABLE CLASSES BEFORE (a) AND AFTER (b) AUGMENTATION



(b)

fig. 3.

then it is apparent that the locus of the line's point set can be given equally well by the vector dot product:

$$N \cdot R + d = 0 = ax_1 + bx_2 + d \qquad (10)$$

Where: $N = (a,b)$
$R = (x_1,x_2)$

This equation could be further simplified if the augmented vectors of $\overline{N}$ and $\overline{R}$ are formed, such that $\overline{X}^* = (x_1,x_2,1)$ and $\overline{W} = (a,b,d)$. Equation 10 becomes:

$$\overline{W} \cdot \overline{X}^* = 0 \qquad (11)$$

Equation 11 is the basis for the binary pattern classifier. If the dot product is negative, then the unknown is placed in category 'B'. If the result is positive, the it is placed in category 'A'. The procedure can be demonstrated if $\overline{X}^*$ is set equal to the augmented mass spectra of an arginine containing compound. If the dot product is formed, the result will be positive and the compound will have been judged to contain arginine.

This method can be applied to any number of features simply by increasing the number of dimensions in the pattern vector and the weight vector such that:

$$\overline{X}^* = (x_1,x_2,\ldots,x_n,1) \qquad (11a)$$

and,

$$\overline{W} = (w_1,w_2,\ldots,w_n,w_{n+1}) \qquad (11b)$$

where: $\overline{X}^*$ = the augmented pattern vector
$\overline{W}$ = the weight vector

The procedure for the calculation of W, so that all the known vectors in the training set are properly classified, has been devised by Nilsson and employed by P.C. Jurs, et al, in the classification of mass spectra. The procedure is relatively simple and is presented without proof below.

The primary step in the learning procedure is the formation of a training set should have an associate set where in the elements contain (represent) the proper classification of all elements in the training set.

The learning phase of the routine requires that several passes be made across the training set to optimize 'W'. Each pass through the set results in a new 'W*' which is more capable of separating the clusters from each other. The optimization procedure is:

$$\overline{W}^* = \overline{W} + c\overline{X}^* \tag{12a}$$

when the dot product $W \cdot X^*$ is incorrectly negative, and

$$\overline{W}^* = \overline{W} - c\overline{X}^* \tag{12b}$$

when the dot product $\overline{W} \cdot \overline{X}^*$ is incorrectly positive. The coefficient "c" is a number to be determined by one of several methods and its function is to insure that the vector $\overline{X}^*$, which was incorrectly classified by W will be properly placed by W*. This procedure will converge only if the Pattern Group, "$\underline{X}$" = $(X_1, X_2, X_3, \ldots, X_n)$ is linearly separable.* To effect the procedure the weight vector W is optimezed on the training set "$\underline{S}$", where the elements of "$\underline{S}$" are members of the pattern group placed in random sequence ad infinitum, i.e., and element X occurs an infinite number of times in $\underline{S}$. The procedure is terminated whe W correctly classified all the elements in the set $\underline{X}$.

Should the event occur that the raw data set is not linearly separable, it may be possible to effect some form of preprocessing on the data. Preprocessing will rearrange the data to yield clusters that are separable. The type of processing used will depend on the data, but in most cases it is some form of co-ordinate transformation, (8,17). As the gain

---

* any time N N+1 the procedure will converge.

onstant "c" can have a great effect on the speed at which the procedure converges, the different methods for determing 'c' should be expanded. Nilsson gives several methods for the calculation of 'c'.

One method determines that 'c' which is just sufficient to result in proper classification of $\overline{X}^*_i$. To meet this constraint "c" must satisfy one of the equations:

$$0 = (\overline{W}+c\overline{X}^*)\cdot\overline{X}^* \tag{13a}$$

if $\overline{W}\cdot\overline{X}^*$ is erroneously nonpositive or,

$$0 = (\overline{W}-c\overline{X}^*)\cdot\overline{X}^* \tag{13b}$$

is $W\cdot X$ is erroneously non-negative. From the above two equations it is apparent that c is the next integer larger than:

$$\text{abs}(\overline{W}\cdot\overline{X})/\overline{X}\cdot\overline{X} \tag{14}$$

Nilsson also claims (with proof) that this method for calculating c will result in the same weight vector W that would develop if c were set equal to one at the onset. This procedure results in swifter convergence. (For greater detail read chapt. 4 of ref. 7).

P.C. Jurs outlined one method (fractional correction) that gave excellent results on a wide range of categories (18). But first it is necessary to explain a variation of the BPC method employed by Jurs.

The decision surface for "most" BPC's is a hyperplane which has zero thickness, but Jurs has found that a hypersheet of finite thickness gives results which excel those of the hyperplane. The reason for this better performance, is the constraint that all points along the decision surface must be offset from it by some distance d=z. The hyperplane surface required only that the elements of the clusters not reside on the surface. This configuration would give erroneous results if a borderline point of

one category shifted, due to random variation, into the region of another category. The hypersheet avoids these "random" complications somewhat.

The mechanics of the hypersheet decision surface for a sheet of thickness 2z are: X is placed in category one if k z or if k -z X is placed in category two; for the case $-z \neq k \neq z$, the vector X is not classified. When using this method on spectral data he found that the larger thickness gave better results (18).

The advantages of the hypersheet decision surface can be enhanced even further when it is paried with the fractional correction method for evaluating the constant 'c' (equ.13). This constant and the correction routine, i.e. equ. 13, control the rate at which the weight vecotr converges to its final value. Correction of the weight vector after each error results in a vector that is very slow in convergence, because the first few corrections result in a vector that is only a fraction closer to the ideal or correct value. The correction procedure only guarantees that the vector previously misclassified, will be properly classified by the new weight vector. The procedure does not claim that the new weight vector will correctly classify those vectors that came before, nor will it stand a better chance of that vector just missed correctly. It is the "law of averages" that guarantees that, if and only if, the clusters are linearly separable will the process converge after a suitable number of trials, and then its classification ability is only as good as the set it was trained from.

The method of fractional correction is one method that will reduce the number of weight vectors that must be tried before the final value is reached. This method was outlined in Nilsson's book Learning Machines

and has been improved on by Jurs, so that it will yield a series of weight vectors that converge rapidly to the final value. This is desirable because the first training vectors result in vast changes in the form of the weight vector that must be overcome in subsequent training cycles. The number of these cycles can be reduced if the effect of the training vectors can be equalized. With fractional correction all the patterns missed during a pass will contribute equally to the development of the weight vector.

The procedure for fractional correction requires that the weight vector be split up into two sets of components; a positive class $\overline{W}_p$ represented by the vector,

$$\overline{W}_p = (w_{p1}, w_{p2}, w_{p3}, \ldots, w_{pn+1}) \tag{15a}$$

and the negative class W_n, represented by the vector,

$$\overline{W}_n = (w_{n1}, w_{n2}, \ldots, w_{n+1}) \tag{15b}$$

When a pattern is to be classified the sum of W_p and W_n is used,

$$\overline{W}_t = \overline{W}_p + \overline{W}_n \tag{16}$$

thus the calculation of discrimanent value, k, becomes,

$$k = \overline{X}_i \cdot \overline{W}_t \tag{17}$$

The classification of $\overline{X}_i$ is made as before.

The first step of the procedure is to form two summations,

$$K_p = \Sigma k_j \tag{18a}$$

and

$$K_n = \Sigma k_i \tag{18b}$$

where K_p is the sum of those dot products that were actually positive but misclassified, and K_n is the sum of those dot products which were members of the negative class but were classified positive. When these sums have

been obtained the new values of W_p and W_n are calculated by the equations:

$$W_p' = W_p + c_j f_j X_j \tag{19}$$

where X_j is the set of patterns that were positive, but classified negative, and

$$f_j = (1.5)\ (k_j - z)/K_p \tag{20a}$$

and

$$c_j = -2(k_j - z)X_j \cdot X_m \tag{20b}$$

The process is repeated until all the k_j's are exhausted. This same method is used to calculate the new components of W_n', with the only changes being:

$$f_i = (1.5)(k_i - z)/K_n \tag{21a}$$

and

$$c_i = -2(k_i - z)/X_i \cdot X_i \tag{21b}$$

The constant, 1.5, is an emperical parameter that is adjusted to yield the fastest cycle time, i.e. the greatest speed of convergence. The new weight vector $W' = W_p' + W_n'$ is then formed and used to classify all the spectra, are classified without error.

It is of interest to note that expansion of the factor $f_i c_i$ produces the following result:

$$f_i c_i = -3(W \cdot X_i - z)^2 \cdot (X_i \cdot X_i)_{\Sigma(W \cdot X_i)} \tag{22}$$

If this equation is inserted into equation and the appropriate summation terms inserted, an equation for W_f, the final weight vector, results.

$$W_f = W_p + W_n + \Sigma\{3(W \cdot X_i - z)^2 X_i/(X_i \cdot X_i)(\Sigma W \cdot X_i)\} \tag{23a}$$

rearranging terms

$$W_f = W + (-3/\Sigma W \cdot X_i))\ \Sigma((W \cdot X_i - z)^2 X_i/X_i \cdot X_i) \tag{23b}$$

If z is set equal to zero, then W_f becomes;

$$W_f = W - 3 \sum (W \cdot X_i) X_i / X_i \cdot X_i \qquad (24)$$

At this point it is noted that $(W \cdot X_1)/X_i \cdot X_i$ is the term that Nilsson used to calculate the constant c in the training procedure for the binary pattern classifier, with the absolute value function replaced by two separate equations for W_p and W_n. Jurs has, therefore, taken the basic binary training routine and developed a method that converges more rapiedly to the same weight vector W.

Some of the advantages of the fractional correction routine as described here are: Both methods converge to the same weight vector; the weight vector is modified by all misclassified training vectors at one time; and the inclusion of the thickness constant z produces a machine of greater resolution.

The hypersheet may be used in the training method outlined by Nilsson if the constant c is calculated by the equation:

$$c = \text{abs}(z - W \cdot X_i)/X_i \cdot X_i$$

The constant c is then used in Nilsson's training routine as before.

Some characteristics of the BPC are: The weight vector may be initialized to any set of values; but it has been found most convenient to set all the w_i's equal to 1 ( ), though Jurs outlines a method for calculating a "superior" starting value that involves analysis of the features of those elements in the pattern group prior to effect in the optimalization procedure ( ). The value of the constant "c" may be calculated in several different ways. And, finally, several BPC may be joined together to form a greater, somewhat more flexible machine that does not require that the data be convex.

LITERATURE CITED:

1. C.J.D.M. Verhagen, Pattern Recognition, 7,109-116 (1975)
2. U. Grenander, Foundations of pattern analysis, Q. Appl. Math. 27 (1969)
3. J.J. Duffy and M.A. Franklin, IEEE Transactions on Systems, Man, and Cybernetics, SMC-5, 2 (1975)
4. Tzay Y. Young and Thomas W. Calvert, "Classification, Estimation and Pattern Recognition," Elsevier inc., N.Y., New York (1974)
5. L.R. Carlson, C.F. Bender and R.H. Pritchard, R/D, 26, 34-41 (1975)
6. J.B. Justice and T.L. Isenhour, Anal. Chem. 46, 223 (1974)
7. N.J. Nilsson, "Learning Machines," McGraw-Hill Book Co., N.Y., N.Y. (1965)
8. G.S. Sebestyen, "Decision-Making Processes in Pattern Recognition," The Macmillian Co., N.Y., New York (1962)
9. J.T. Tou and R.H. Wilcox, "Computer and Information Sciences," Spartan Books Inc., Washinton, D.C. (1964)
10. E. Fox and U.L. Hodges Jr., Project No.21-49-004, Report No.4, Contract No. AF41(128)-31, USAF School of Aviation Medicine
11. E. Fox and U.L. Hodges Jr., Report No. 11, ibid.
12. T.M. Cover and P.E. Hart, IEEE Transactions on Information Theory, IT-14, No. 1 (1968)
13. McCulloch and Pitts, Bull. Math. Biophys., 5, 115-137 (1943)
14. F. Rosenblatt, Psychol. Rev., 65, 386-407 (1958)
15. P.C. Jurs, B.R. Kowalski and T.L. Isenhour, Anal. Chem., 41, 21-27, (1969)
16. _____, Anal. Chem., 41, 690-700 (1969)
17. _____, Anal. Chem., 42, 1387-1394 (1970)
18. P.C. Jurs, Anal. Chem., 42, 22-26 (1970)
19. W.L. Felty and P.C. Jurs, Anal. Chem., 45, 885-889 (1973)
20. R.W. Liddel III and P.C. Jurs, Applied Spectroscopy, 27, 371-376 (1973)
21. R.D. Macfarlane and D.F. Torgerson, Science, 191, 920-925 (1976)
22. N.M. Frew, L.E. Wangen and T.L. Isenhour, Pattern Recognition, 3, 281-296 (1971)
23. D.R. Preuss and P.C. Jurs, Anal. Chem., 46, 520-525 (1974)
24. L. Pietrantonio and P.C. Jurs, Pattern Recognition, 4, 391-400 (1972)
25. K.L.H. Ting et al, Comput. Biol. Med., 4, 301-332 (1975)
26. J. Schechter and P.C. Jurs, Applied Spectroscopy, 27, 30-40 (1973)
27. T. Nagano, IEEE Transactions on Man,Systems and Cybernetics, SMC-5 (1975)
28. J. Felsen, IEEE Transactions on Man,Systems and Cybernetics, SMC-5 (1975)
29. S. Zander, A.J. Stuper and P.C. Jurs, Anal. Chem., 47, 1085-1093 (1975)
30. P.C. Jurs, Anal. Chem., 43, 1812-1815 (1971)
31. C.H. Chen, Pattern Recognition, 7, 87-94 (1975)
32. S.A. Kassam, IEEE Transactions on Information theory, IT-22, 22-26 (1976)
33. J.B. Justice and T.L. Isenhour, Anal. Chem., 47,2286-2288 (1975)
34. D.G. Howery, American Laboratory, __, 14-25 (1976)
35. D.H. Smith et al, Tetrahedron, 29, 3117-3134 (1973)
36. G. Raisbeck, "Information Theory," the MIT Press, Massachusetts Instsitute of Technology, Cambrige, Massachusetts (1965)
37. C.W. Swonger, "Frontiers of Pattern Recognition,"

38. P.C. Jurs, Anal. Chem., 43, 364-367 (1971)
39. S.L. Grotch, Anal. Chem., 46, 526-534 (1974)
40. _____, Anal. Chem., 43, 1362-1370 (1971)
41. I am very gratful to R.D. Macfarlane and D.F. Torgerson who proof read this paper and gave many helpful comments. And to C. Mahula who typed the original draft.