LITERATURE CITED

(1) Williams, M. E., and Schipma, P. B., "Design and Operation of a Computer Search Center for Chemical Information," *J. Chem. Doc.*, **10**, 158–162 (1970).

(2) Grunstra, N. S., and Johnson, K. J., "Implementation and Evaluation of Two Computerized Information Retrieval Systems at the University of Pittsburgh," *J. Chem. Doc.*, **10**, 272–7 (1970).

(3) Park, M. K., Carmon, J. L., and Stearns, R. E., "The Development of a General Model for Estimating Computer Search Time for CA Condensates," *J. Chem. Doc.*, **10**, 282–4 (1970).

(4) Roberts, A. B., Hartwell, I. O., Counts, R. W., and Davila, R. A., "Development of a Computerized Current Awareness Service Using Chemical Abstracts Condensates," *J. Chem. Doc.*, **12**, 221–3 (1972).

(5) Wilde, D. U., and Starke, A. C., "A Chemical Search System for a Small Computer," *J. Chem. Doc.*, **14**, 41–4 (1974).

(6) Schipma, P. B., "Computer Search Center Statistics on Users and Data Bases," *J. Chem. Doc.*, **14**, 25–9 (1974).

Interactive Pattern Recognition in the Chemical Laboratory

JAMES R. KOSKINEN and BRUCE R. KOWALSKI*

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

Received March 6, 1975

An interactive pattern recognition system has been developed for utilization in the chemical laboratory. This system has been designed so that the user need not know computer programming. Actual data analysis is done on a time-sharing host computer, and communication with the chemist is via an intelligent computer graphics terminal. The graphics terminal also displays projections of *n*-space data structure as two- or three-dimensional plots on the display screen. These plots can be manipulated by the chemist in real time to provide an approximate view of the *n*-space data structure. By examining the results of the various pattern recognition methods using the display terminal, the chemist can direct the application interactively, thereby increasing operational efficiency and allowing the gain of new insights into *n*-space ($n > 3$) data analysis applications.

Pattern recognition has been demonstrated to provide a powerful method for interpreting chemical data.[1] It has provided a general approach to solving a class of data processing problems commonly encountered in experimental chemistry. A statement of the general problem is: can an obscure property of a collection of objects (elements, compounds, mixtures, etc.) be detected and/or predicted using indirect measurements, made on the objects, that are known to be related to the property via some unknown relationship?

Obviously, the problem is not only to find and predict the property, but also to try to find the mathematical relationship that links the measurements to the property. Therefore, the problem can be considered as a mapping of objects from measurement space into property space. For pattern recognition methods, the objects, usually called patterns, can be considered as points in an n-dimensional hyper-space, where n is the number of measurements made on each object. Likeness among the objects is assumed to be reflected via the measurements as nearness of corresponding points in n-space. Thus, pattern recognition can also be described as a collection of methods that analyze n-space plots.

There are four major branches of pattern recognition which correspond to the four types of operations performed on the n-space data structure. First, the measurements can be preprocessed[2] by forming linear or nonlinear combinations of measurements to generate features. These features can be fewer in number than the measurements. This process is called feature reduction and is often a necessary step. The features are selected to yield a new information representation that is more amenable to data analysis by the other pattern recognition methods. An example will help demonstrate the need for preprocessing. Suppose a chemist submits a sample for NMR analysis and receives an intensity vs. time plot from a free induction decay experiment. The plot will have very little meaning to the chemist until it is preprocessed using a Fourier transform to change the information representation to the frequency domain. There are several preprocessing methods available to the chemist that do scaling operations, weighting, etc. Choice of a particular method is dependent upon the application.

The second type of operation comes from the scientist's excellent ability to recognize patterns in the familiar two- or three-dimensional space. Since the scientist cannot view n-space when n is greater than 3, the data structure (relative position of points) in n-space can be mapped to two-space or three-space by display methods[3] that seek to minimize information loss.

The last two branches of pattern recognition are called supervised learning and unsupervised learning[2] and are divided on the basis of what must be learned from the objects via the measurements. Supervised learning assumes that the sought-for property is known for some of the objects. Those objects are "tagged" as knowns (collectively called the training set) and are used to develop a classification rule that can be used to predict the property for unknown objects. Developing classification rules to separate known active anticancer drugs from inactive drugs and then classifying drugs that have not been tested in biological screening systems is an example of such an application.[4] If no training set exists and the goal of the study is to discover a useful property of the objects, unsupervised learning methods are used. The discovery of the periodicity of the elements from properties of the elements is an excellent example of unsupervised learning even though it was done before the advent of computers.

There are many different pattern recognition methods described in the literature. Yet most of the applications of

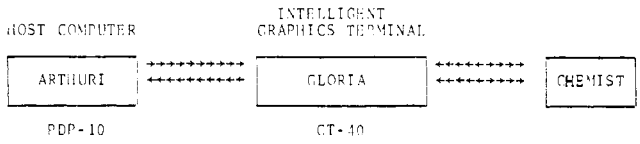

Figure 1. Block diagram of interactive pattern recognition system.

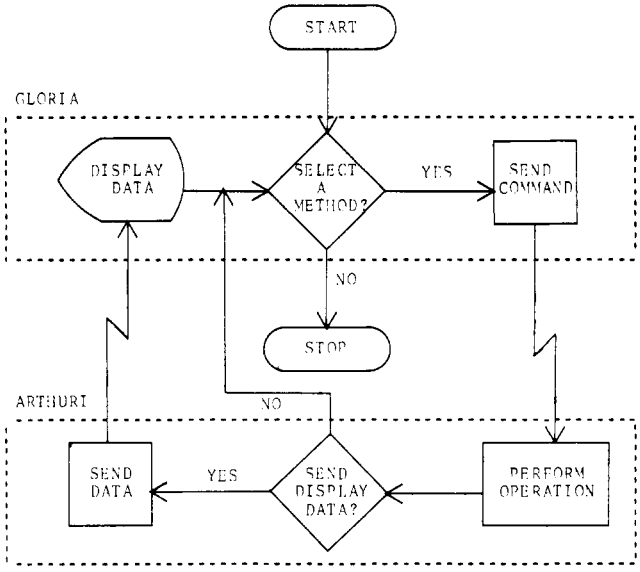

Figure 2. Flow diagram of interactive pattern recognition system.

pattern recognition by chemists have been limited to the use of only one method, the linear learning machine, to analyze various types of spectroscopic data. The true value of pattern recognition for the chemist is realized when several of the methods are used in combination as a system. Such a system, ARTHUR†, has been developed and is in use at the University of Washington.

ARTHUR contains more than 30 different pattern recognition methods which are written in Fortran to run in batch mode on a large computer. Experience with ARTHUR has demonstrated the applicability of pattern recognition to solve a wide range of chemical data analysis problems. Although this system is very useful, three deficiencies were observed during its use.

First of all, the results of one method usually led to the application of other methods. This is rather time-consuming in batch mode because the job had to be set up again and resubmitted to the computer. This process often led to a lack of continuity during the application. The second problem is a result of the "data reduction" fallacy. For example, an analysis of 100 patterns with 20 measurements per pattern amounts to 2000 numbers, which often leads to several pages of numbers generated by the various pattern recognition methods. The chemist ends up wading through several pages of numbers to get the desired results. The third problem comes from the large dependence on the display methods in ARTHUR to allow the chemist to supervise the results of various methods. It has been observed that projections and mappings from n-space to two-space typically retain 50 to 70% of the n-space data structure. Mapping to three-space instead of two space can retain an additional 5 to 25% of the data structure, thereby providing a more accurate display.

The solution to the first problem was first discussed by Sammon[5] in his description of OLPARS. The problem can be alleviated by making the system truly interactive with the user. The solution to the second problem was covered

† Available from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Wash. 98195.

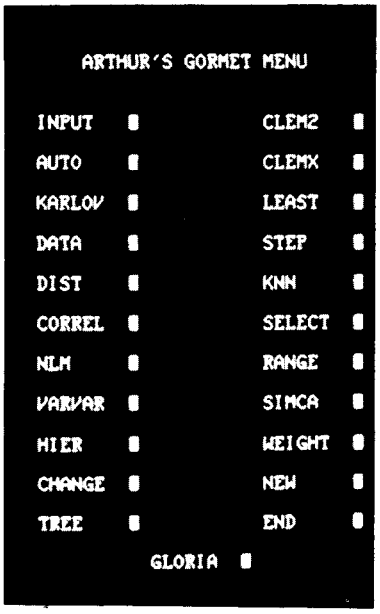

Figure 3. Menu displayed by GLORIA.

in part by Sammon[6] when he described adding computer graphics for display methods in OLPARS. This would mean that the pages of printed output could be replaced with plots and displays of the data. The solution to the third problem is to use three-dimensional plots instead of two-dimensional plots to display the data, thus retaining the maximum amount of information. Combining these solutions into one system should lead to the ideal pattern recognition system. This system would run in an interactive computer environment, using computer graphics to display the data, all under the complete control of the chemist. This paper describes such a system. It is built around an intelligent computer graphics terminal that communicates with a large interactive time sharing computer. The power of this interactive system will be demonstrated by two applications.

DESCRIPTION OF THE SYSTEM

The interactive graphics pattern recognition system is designed so that a chemist with no computer programming knowledge can use it. A block diagram of the system is shown in Figure 1. The system consists of a large host-computer that runs the interactive version of ARTHUR, and an intelligent graphics terminal. The program GLORIA runs in the intelligent graphics terminal and acts as the interface between the chemist and ARTHUR. For the current system configuration the host-computer is a Digital Equipment Corporation PDP-10 computer, and the intelligent graphics terminal is a Digital Equipment Corporation GT-40. The GT-40 is linked to the PDP-10 over a telephone line with a data transmission rate of either 300 or 2400 baud.

Figure 2 is a flow diagram showing how ARTHUR and GLORIA work together to form the interactive system. Program ARTHUR is initiated and sends a command to the terminal which starts GLORIA. GLORIA then displays a menu which contains a list of all the methods available in the current system. Figure 3 is a picture of the menu that GLORIA displays. The chemist selects a method by touching the appropriate light square with the light pen. GLORIA then sends the command to ARTHUR to execute the desired method. ARTHUR obtains additional parameters via dialog with the chemist and then performs the operation. When the particular operation is completed, ARTHUR checks to see if display data are to be sent to GLORIA. If no display

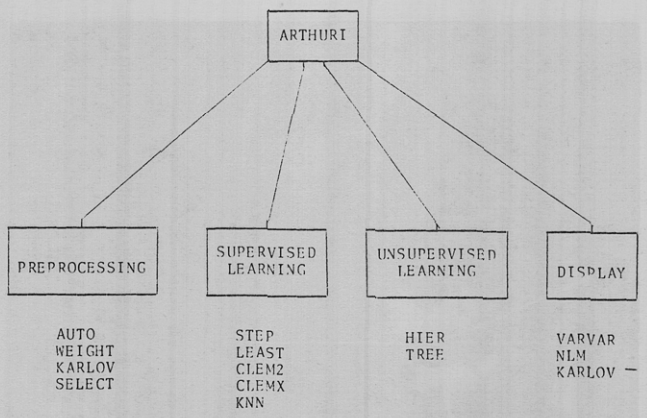

Figure 4. Pattern recognition methods included in ARTHUR (interactive version).

data are to be sent, ARTHUR tells GLORIA to display the menu.

There are two kinds of data sent between the two programs. Some of the methods transmit plot information which is used to display the patterns on the screen in two- or three-dimensional plots. Other methods send cluster information so that the clusters detected by the particular method can be displayed. Once the information has been received, GLORIA displays it on the screen. The chemist signals GLORIA when he is finished looking at the data display and the menu is displayed. At the end of an application, the chemist selects the END routine, and GLORIA sends the command to ARTHUR and ends execution. ARTHUR then queues up the appropriate printed output files and terminates execution.

ARTHUR (interactive version) currently contains 14 methods that represent the four major branches of pattern recognition (Figure 4).* There are four preprocessing methods in ARTHUR. AUTO is an autoscaling routine that scales the features so that they each have a mean of zero and unit variance. WEIGHT is a routine that weights the features with either the variance or the Fisher weights. KARLOV is the Karhunen–Loeve transformation that can be used for feature selection. The Karhunen–Loeve method creates new features as linear combinations of the original features. A unique ordering is the result of this transformation with the first new feature containing the greatest amount of variance and each successive new feature containing the next greatest amount of the residual variance. SELECT[7] is a new and highly effective method of feature selection, which yields new features which are highly correlated to the sought-for property, linearly independent, yet easily relatable to the original features.

Supervised learning methods are represented by four types of the linear learning machine (LEAST, STEP, CLEM2, CLEMX) and the k-nearest nieghbor method (KNN). LEAST calculates the weight vector for the learning machine using the least-squares method. STEP does a stepwise regression analysis of the features and then calculates the weight vector using the least-squares method. CLEM2 and CLEMX calculate the weight vector using the error-feed-back method. CLEM2 does pairwise separation of the properties, while CLEMX calculates a weight vector for each property.

HIER and TREE are the unsupervised learning methods that are included. HIER is a hierarchical Q-mode clustering method, and uses a similarity matrix which is scanned for the largest value. The two points producing this value are combined. The two combined points are considered as one new point and a smaller similarity matrix is calculated. This process is continued until all the points are in one cluster. TREE is the minimal spanning tree method of cluster analysis described by Zahn.[8] This method forms clusters by calculating the minimal spanning tree for the patterns and then detects inherent separations in the data by deleting edges from the minimal spanning tree which are significantly longer than nearby edges.

* The methods mentioned in this section are those described by Kowalski[2] unless there is a specific reference.

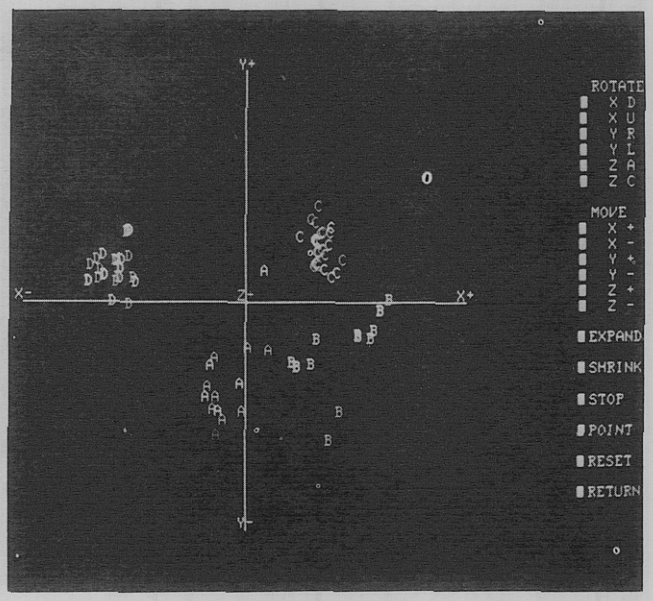

Figure 5. Display of the obsidian data set.

The display methods in ARTHUR are designed to work with GLORIA. They all generate plots in two or three dimensions and then send the plot information to GLORIA. VARVAR is a routine that creates feature by feature plots. Two or three features can be plotted at one time. NLM performs a nonlinear mapping of the features from n-space down to three-space. This method preserves the interpoint distances. KARLOV is also used to provide eigenvector projections to one-, two- or three-space.

The actual display of the plots is done by the program GLORIA. For three-dimensional plots of X axis is horizontal, the Y axis is vertical, and the Z axis is perpendicular to the display screen. Distrance along the Z axis is indicated by the relative brightness of the display point. Data points that are members of property one group are represented on the plots by A's; property two by B's; up to property ten by J's. Any point that is to be classified (test set member) is represented by an X.

GLORIA has three methods of manipulating a plot. These methods are: rotating the plot about an axis, moving the zero point of the plot, and changing the scale factor. All manipulations are done in real time, so that the chemist can monitor and control the changing display. The real time rotation of the display greatly enhances the visualization of the three-dimensional aspects of the data display so that, with absolutely no training or experience, three-dimensions are easily perceived.

Figure 5 is a typical data display from GLORIA. The commands for the various display methods are along the right side of the screen. Any of these methods can be selected by activating the appropriate light square with the light pen. There are six "rotate" commands to allow the display to be rotated about the X, Y, or Z axis in either direction. The "move" commands move the zeropoint of the plot along the X, Y, or Z axis in either the positive or negative direction. The scale factor is changed with the "expand" and "shrink" commands. Any data point on the screen can be identified by using the point command and the light pen (Figure 6). The "stop" command stops whatever manipulation is taking place; the "reset" command resets the screen

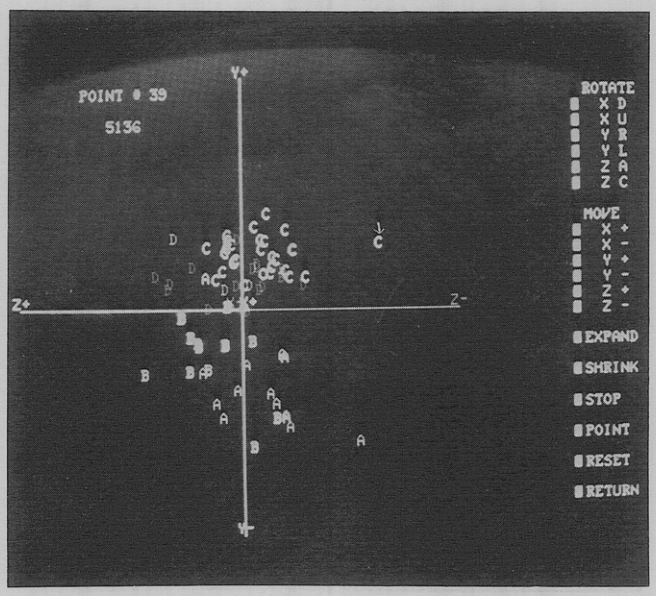

Figure 6. Display in Figure 5, rotated 90° about vertical axis to show hidden point (no. 39).

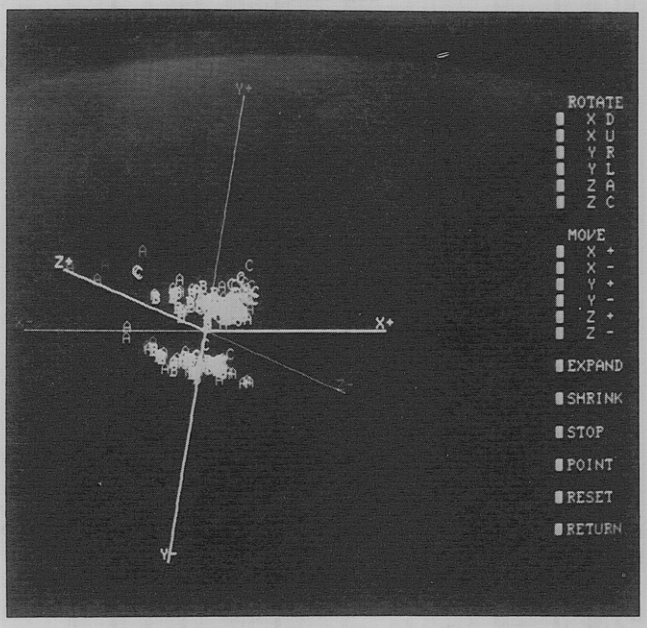

Figure 7. Display of clinical data set with "missing" data.

to the original plot display; and the "return" command returns control to ARTHUR.

GLORIA is also able to show graphically the clusters detected by the clustering methods in ARTHUR. The characters representing patterns in any of the detected clusters (unsupervised learning) or categories (supervised learning) "blink" on and off during execution of the above commands. This effect solves part of the data explosion problem by allowing the chemist the graphical analysis of the results of pattern recognition methods.

All the pattern recognition methods described above for ARTHUR are written in Fortran IV as separate subroutines. The main program in ARTHUR (also written in Fortran IV) decodes the commands from GLORIA and calls the requested subroutine. ARTHUR uses overlay methodology to accommodate all the subroutines in 11k of core. The overlay structure also makes it relatively easy to add new methods to ARTHUR. Data sets that are being examined reside on disk files, so that several applications can be done using the same data set without having to reload the data set. The utility subroutines in ARTHUR that provide the communication with GLORIA are written in Macro-10 assembly language. This configuration of ARTHUR has been designed to handle a maximum of 200 patterns. Each pattern can have a maximum of 100 features and the patterns can be divided into a maximum of 10 categories. Larger applications are studied in subsets or by the batch version of ARTHUR. GLORIA is written in Pal-11 assembly language and is designed to display a maximum of 200 data points on the screen in three dimensions. All of the calculations for the various plot transformations are done in real time for the chemist. GLORIA is also designed to allow the addition of new pattern recognition methods or new display manipulation methods.

APPLICATIONS

The system was tested using a data set made up of the concentration of ten trace elements of 75 obsidian samples as determined by X-ray fluorescence.[9] In the original study four pattern recognition methods were used to separate the obsidian source samples into four categories according to their origin using the concentration of the ten elements. At that time the four methods used were implemented as four separate programs. The interactive system described herein obtained the same results and much more in just one short (less than 10 min) session with the chemist.

These data also illustrate the power of three-dimensional display techniques. In earlier studies the data structure in ten-space was mapped down to two-space and plotted. When this mapping from ten-space to two-space was done using the Karhunen–Loeve transformation, only 73% of the data structure information was retained. The same transform, in mapping down to three-space, retains 84% of the information. Figure 5 shows the two-dimensional display of the data. The A's represent one source of obsidian; the B's a second, etc. As can be seen in the plot, the four sources separate fairly well. The results of the Zahn minimal spanning tree cluster method indicates that there is a cluster with only one member. It is very difficult to see which point it is in two dimensions. However, in the three-dimensional plot (Figure 6) it is easy to see. Figure 6 shows the three-dimensional representation of the data. The display has been rotated 90° about the vertical axis to show point no. 39 with the identification number 5136 as the one-member cluster found by the clustering method. The point is overlapped in the two-dimensional plot by all the other C's. The extra 11% of information that is retained when the third dimension is added is very helpful in interpreting the n-space data structure. The visual display quickly shows, and the other pattern recognition methods confirm, that obsidian sample 5136 results from a fifth, but unknown source.

This interactive pattern recognition system has proved to be a very powerful tool. One of the most interesting applications studied to date has been the reevaluation of a clinical chemistry data set.[10] This data set consists of the results from a series of 22 clinical laboratory tests made on 70 individuals. This series of tests was administered to some of the individuals several times. The entire data set is comprised of 272 separate laboratory samples analyzed for 22 tests on a total of 70 individuals. Eight of the individuals were healthy and the rest had a specific disease of the liver.

Early in the study the 22 measurements were mapped to three-space (Figure 7) using the eigenvector projection method (KARLOV). The data appear to be easily divided into two groups. This is confirmed using supervised learning methods on the 22-space. Visual inspection reveals that there are both the healthy individuals and those with a liver disease represented in each group. A closer examination of the data set revealed that some of the zero values

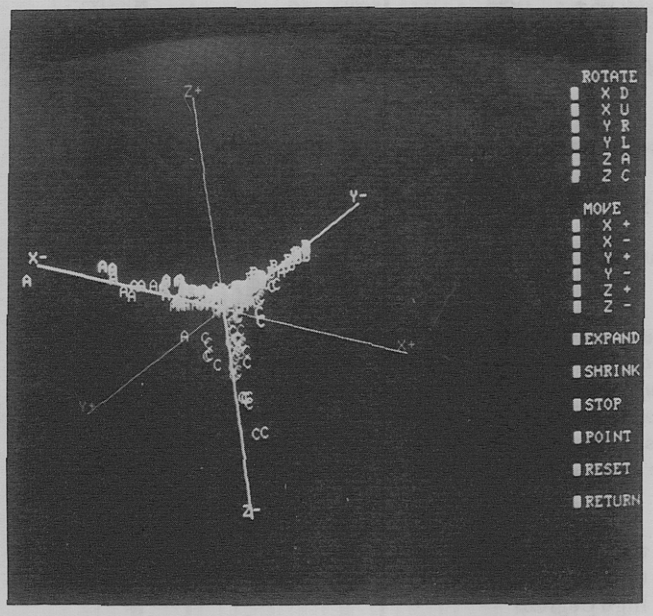

Figure 8. Display of clinical data showing "three-legged starfish" cluster.

given as results of the tests were, in fact, an indication that the particular test was not administered. The division of the data set into two groups in the display appears to be an artifact of the missing data. This artificial separation disappeared when the tests with missing data were excluded from the study. Using this complete data set, it is straightforward to separate the individuals with the various liver diseases.

A visual examination of the data representation in three-space shows why the separation was so easy. Figure 8 shows a subset of the data that consists of the healthy individuals and those with one of three liver diseases. The data cluster appears to be a three-legged "starfish". The center of the cluster contains only the healthy individuals, while each of the arms contain only one of the liver diseases. The arms appear to be orthogonal. With this type of clustering it is easy to separate the diseases using supervised learning methods. However, the unsupervised methods have some difficulty. This is the first time that this "starfish" type of clustering has been observed by the authors.

In this clinical application the interactive pattern recognition system has shown its value by quickly pointing out the artificial separation caused by the missing data. It has also revealed a new type of clustering. Both of these observations have led to the current investigation of methods to handle them.

CONCLUSION

The interactive pattern recognition system described above is a powerful and useful tool for the chemist. It allows him to use several pattern recognition methods in one session to examine his problem. The computer graphics enable him to gain new insights into n-space data structures. It is currently being used to examine problems in forensic chemistry and to study chemical structure–biological activity correlations in selected areas of medicinal chemistry. The open-ended design of the system allows easy expansion. New pattern recognition methods are being readied for inclusion as are new display methods. Plans are also being formulated to add on-line chemical instrumentation.

There is little doubt as to the enormous importance of the simple two-space plot in science and engineering. Interactive pattern recognition provides a tool to analyze the n-space plot even though it can only be visualized approximately.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the National Science Foundation, under Grant No. GP-42013 to develop ARTHUR. We also gratefully acknowledge the support of the Office of Naval Research under Contract No. N0014-67-A-0103-0036 to develop interactive ARTHUR and GLORIA. We also acknowledge the additional computer time provided by Dr. J. Sobolewski, Director of the John L. Locke, Jr., Computer Center, to thoroughly test the interactive pattern recognition system. We would also like to acknowledge Dr. P. Strandjord, K. Clayson, and R. Roby for providing us with the clinical data and for their helpful discussion concerning it.

LITERATURE CITED

(1) Kowalski, B. R., and Bender, C. F., *J. Am. Chem. Soc.*, **94**, 5632 (1972).
(2) Kowalski, B. R., "Pattern Recognition in Chemical Research," in "Computers in Chemical and Biochemical Research," Vol. 2, C. F. Klopfenstein and C. L. Wilkens, Ed., Academic Press, New York, N.Y., 1974.
(3) Kowalski, B. R., and Bender, C. F., *J. Am. Chem. Soc.*, **95**, 686 (1973).
(4) Kowalski, B. R., and Bender, C. F., *J. Am. Chem. Soc.*, **96**, 916 (1974).
(5) Sammon, J. W., *IEEE Trans. Comput.*, **C-19**, 594 (1970).
(6) Sammon, J. W., Procter, A. H., and Roberts, D. F., *Pattern Recog.*, **3**, 37 (1971).
(7) Kowalski, B. R., and Bender, C. F., Abstracts, Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, No. 161, 1974.
(8) Zahn, C. T., *IEEE Trans. Comput.*, **C-20**, 6 (1971).
(9) Kowalski, B. R., Schatzki, T. F., and Stross, F. H., *Anal. Chem.*, **44**, 2176 (1972).
(10) Strandjord, P. E., Clayson, K. J., and Roby, R. J., *Hum. Pathol.*, **4**, 67 (1973).