must be some way to select molecules on the basis of their structural features. Hence, search routines will continue to be an important area of computer applications to chemistry as long as structural features of molecules are of interest.

ACKNOWLEDGMENT

The authors wish to thank J. R. Koskinen and C. J. Appellof for the development of the PDP-11/05 program used for graphical input of molecular structures. The authors wish also to thank the National Science Foundation for partial financial support during the summers for Mr. Blom and Mr. Saxberg under the NSF Undergraduate Research Participation program, Grants EPP 75-04525 and SMI 76-03095.

LITERATURE CITED

(1) S. J. Frycki, F. F. Giarrusso, P. A. Roskos, D. E. Dancsecz, S. J. Lucania, and D. M. O'Brien, "Computerized Monitoring of the Inventory and Distribution of Research Chemicals", *J. Chem. Doc.*, **13**, 136–145 (1973).
(2) E. H. Eckermann, J. F. Waters, R. O. Pick, and J. A. Schafer, "Processing Data from a Large Drug Development Program", *J. Chem. Doc.*, **12**, 38–40 (1972).
(3) G. Redl, R. D. Cramer IV, and C. E. Berkoff, "Quantitative Drug Design", *Chem. Soc. Rev.*, 273 (1974).
(4) F. G. Stockton and R. L. Merritt, "The Shell Chemical Structure File System", *J. Chem. Doc.*, **14**, 166–170 (1974).
(5) J. L. Schultz, "Handling Chemical Information in the DuPont Central Report Index", *J. Chem. Doc.*, **14**, 171–179 (1974).
(6) H. Skolnik, "A Notation Symbol Index for Chemical Compounds", *J. Chem. Doc.*, **11**, 120–124 (1971).
(7) H. Skolnik, "A Chemical Fragment Notation Index", *J. Chem. Doc.*, **11**, 142–147 (1971).
(8) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", American Elsevier, New York, N.Y., 1971, pp 68–69.
(9) J. E. Crowe, P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "The Searching of Wiswesser Line Notations by Means of a Character-Matching Serial Search", *J. Chem. Doc.*, **13**, 85–92 (1973).
(10) E. Meyer in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974, pp 108–111.
(11) E. H. Sussenguth, "A Graph-Theoretic Algorithm for Matching Chemical Structures", *J. Chem. Doc.*, **5**, 36–43 (1965).
(12) W. S. Woodward and T. L. Isenhour "Computer Controlled Television Scan System for Direct Encoding of Chemical Structure Models", *Anal. Chem.*, **46**, 422–426 (1974).
(13) A. J. Hopfinger, "Conformational Properties of Macromolecules", Academic Press, New York, N.Y. 1973.
(14) J. Villarreal, Jr., E. F. Meyer, Jr., R. W. Elliot, and C. Morimoto, "CRYSRC: A Generalized Chemical Information System Applied to a Structural Data File", *J. Chem. Inf. Comput. Sci.*, **15**, 220–225 (1975).
(15) K. C. Chu, R. J. Feldmann, M. B. Shapiro, G. F. Hazard Jr., and R. I. Geran, "Pattern Recognition and Structure-Activity Relationship Studies. Computer-Assisted Prediction of Anti-tumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System", *J. Med. Chem.*, **18**, 539–545 (1975).

Reliability of Nonparametric Linear Classifiers

A. J. STUPER and P. C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received June 18, 1976

The consequences of developing nonparametric linear discriminant functions using data sets which have a small ratio of samples to variables per sample are investigated. The lower limit to the ratio of samples to measurements per sample is dependent on the number of variables used but is approximately 3:1. Studies have been done with fully characterized computer generated data sets. Results are reported which indicate that classifiers developed with data sets having a ratio of less than 3:1 are likely to form relationships where none exist, and they are unable to distinguish relationships which do exist. Also, below this limit, feedback feature selection processes are shown to lose their effectiveness.

In the past five years a number of articles concerning applications of pattern recognition have appeared in the chemical literature. Many of these have reported the utility of nonparametric discriminant functions in providing insight into relationships contained within sets of chemical measurements. An assumption inherent in the use of such functions is that the ability to correctly dichotomize the data into classes is meaningful. Alternatively, successful classification is thought to imply that the classification parameters are related to the observed properties through some indirect or complex function. These assumptions can be tested only if certain criteria are met. These criteria deal with the minimum ratio of samples to measurements per sample required to demonstrate a relation within the data. This paper will investigate these criteria and discuss parameters which indicate the reliability of such relations.

Before the limitations of nonparametric discriminant function development can be established, the framework in which such techniques are employed must be understood. The basic procedure is to take n measurements of each object which is to undergo analysis, treating each measurement as an independent axis. This results in a data set for which each object is represented by a point in the n-dimensional space formed from the measurements made upon it. It is assumed that members of a set most similar to each other will tend to cluster in limited regions in this n-dimensional space. The ability to develop a function capable of separating the groups is thought to imply that the measurements must somehow correlate to these clustering properties.

The ability of a function to separate clusters within a set of data is dependent upon the dichotomization ability of the discriminant function. Dichotomization ability is the total number of two class groupings which can be made by the discriminant function. Different functional forms will have different dichotomization abilities. The dichotomization ability for a linear function is given by the following equation[1]

$$D(N,n) = 2\sum_{k=0}^{n} C_k^{N-1}$$

where $C_k^{N-1} = (N-1)!/(N-1-k)!k!$, N is the number of samples, n is the number of measurements or variables per sample, and k is an index describing how the groupings are taken.

The total number of dichotomies possible for a set of samples, regardless of its n space distribution is 2^N. [The only assumption made is that the data are well distributed. A data set is well distributed if no subset of $n+1$ points lies on a $n-1$ dimensional hyperplane.] Any classifier which could effect all of the 2^N possible dichotomies would always indicate that the desired clusters were present. This behavior is independent

Table I. Results of Developing Classifiers Using Random Data[a]

No. in training set	No. of cases	λ	Theoret. P	Gaussian data: No. trained	Gaussian data: P'	Gaussian data: Predictive ability	Uniform data: No. trained	Uniform data: P'	Uniform data: Predictive ability
35	20	1.67	0.885	16	0.80	51.2	15	0.75	50.1
40	20	1.90	0.625	11	0.55	49.2	13	0.65	51.5
45	30	2.14	0.325	12	0.40	48.2	6	0.20	48.7
50	40	2.38	0.126	2	0.05	49.8	6	0.15	47.8

[a] Data sets contained 250 members of 20 dimensions. Number of cases refers to the number of training sets used to measure P. Number trained refers to the number of sets which were separable. P' is the percentage of the total number of sets which were separable.

of whether such clusters truly exist. Such a classifier invalidates the assumptions concerning the correlation of properties to measurements made in developing the framework for the pattern recognition system. It would, therefore, be useless as a classifier.

In cases for which there are more variables than samples, linear classifiers are able to effect all possible dichotomies. However, the equation above indicates that for cases in which there are more samples than measurements, there are fewer linear dichotomies than total dichotomies. Since the number of possible linear dichotomies is dependent upon both N and n, the existence of a lower limit to the ratio of samples to measurements seems evident. Below this limit the results of discriminant development are of little use. The probability of randomly assigning class memberships which will yield a relation separable by a linear surface is useful in defining this limit. The form of this probability equation is[2]

$$P = D(N,n)/2^N$$

This equation describes the probability of obtaining a linear dichotomy as a function of the total number of samples, N, and the total number of measurements, or variables, n. When there are fewer samples than variables, a dichotomy will be found regardless of whether it truly exists, i.e., $P = 1$. As the number of samples increases with respect to the number of variables, the probability of finding such a dichotomy due to chance will decrease, and the value of P will decrease.

Linear classifiers are generally implemented by adding one extra dimension to the sample space before training. It is therefore convenient to define the variable λ as $\lambda = N/(n+1)$. A plot of P vs. λ is shown in Figure 1. Note that the probability of finding a linear relation is still significant at values of λ greater than 1. For example, P equals $^1/_2$ when λ equals 2. Furthermore, for similar values of λ, the probability of observing random classifications decreases much more rapidly for large values of n. This probability function provides a gauge for determining the likelihood of observing a classifier which is exhibiting random classification. Any linear discriminant developed within this region would not be expected to indicate a relationship within the data.

To demonstrate the significance of random classifications, two data sets consisting entirely of random numbers were generated. Set one was generated using Gaussian distributed random numbers while set two was generated using uniformly distributed random numbers. A total of 250 points were generated for each set. The sets were partitioned into training and prediction sets by randomly choosing members for the training set and placing the remaining members in the prediction set. Results from classifiers developed at various values of λ are shown in Table I. It is evident from the table that the probability of obtaining a linear classification follows that predicted by the equation. It is also interesting to note that the predictive ability for the members of the prediction set is no better than random.

Table II shows the way P varies as a function of n and λ. Each column of Table II shows how P changes for a constant value of λ as n increases. For the probability P to fall below

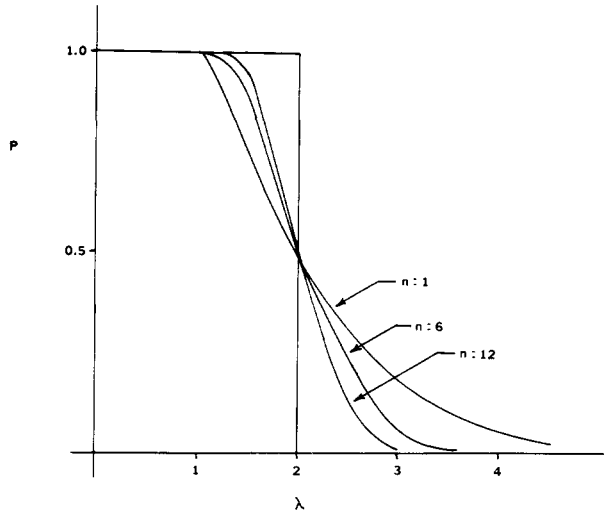

Figure 1. Plot of P vs. λ for several values of n.

Table II. Value of P as a Function of n for Constant λ[a]

n	P (2.25)	P (2.50)	P (2.75)	P (3.00)	P (3.25)	P (3.50)
3	0.3633	0.2539	0.1719	0.1130	0.0730	0.0461
5		0.2120				0.0207
6				0.0577		
7	0.3145	0.1796	0.0946		0.0216	0.0096
9		0.1537		0.0307		0.0045
11	0.2706	0.1325	0.0551		0.0069	0.0022
12				0.0168		
13		0.1147				0.0010
15	0.2498	0.0998	0.0330	0.0093	0.0023	0.0005
17		0.0871				0.0002
18				0.0052		
19	0.2257	0.0762	0.0201		0.0008	0.0001
21		0.0668		0.0030		
23	0.2051	0.0587	0.0124			
25		0.0517				
27	0.1871	0.0456				

[a] Value in parentheses is λ.

1% for $n = 15$ requires $\lambda \geq 3.0$. That is, N, the number of samples, must be at least $(15+1)(3.0) = 48$ for the probability of obtaining a separation between classes to fall below 1%. For data sets where λ is small, P remains large even for large values of n.

A further test was run using a data set consisting of 300 samples, evenly divided between two classes and each sample represented by 40 variables. The data were generated with different variables which define the relationship among them; these variables are called intrinsic variables. Removal of any one intrinsic variable destroys the relationship by essentially converting the data set to random numbers. If only the intrinsic variables are used to represent the samples, then the two classes are always separable with a linear discriminant; if any one of the variables is removed, then linear separability is lost. The method used to generate such a data set has been described previously.[3]

Table III. Results of Exchanging Random Variables and Intrinsic Variables Keeping the Total Number of Variables Constant [a]

Number of variables			Predictive ability						
			40	60	80	100	150	200	250
Total	Intrinsic	Random	(0.976)	(1.46)	(1.95)	(2.44)	(3.66)	(4.88)	(6.10)
40	40	0	47.9	49.4	53.5	58.2	72.8	80.4	83.2
40	30	10	49.0	50.3	53.5	56.8	73.6	81.2	82.8
40	20	20	48.5	48.2	52.8	59.2	76.0	85.0	85.2
40	10	30	50.6	51.0	53.9	59.3	69.6	80.2	80.8
40	5	35	51.9	52.6	58.0	61.7	73.3	81.0	85.6
Average predictive ability			49.6	50.3	54.3	59.0	73.9	81.6	83.5
Standard deviation			1.62	1.66	2.08	1.80	2.29	1.97	1.95

[a] The predictive ability reported is for the average of five independent trainings. Each column of predictive abilities is headed by the number of samples included in the training set and the resulting λ value.

Table IV. Effects of Adding Uniformly Distributed Random Variables to Identical Pattern Vectors[a]

Number of variables			Training set of 100 members		Training set of 200 members	
Total	Intrinsic	Random	λ	Predictive ability	λ	Predictive ability
20	20	0	4.76	76.6	9.52	94.2
25	20	5	3.85	77.0	7.24	91.6
30	20	10	3.23	68.2	6.46	88.6
35	20	15	2.78	63.4	5.56	86.2
40	20	20	2.44	59.2	4.88	85.0

[a] Each line reports the results for the average of five independent runs with five randomly selected training sets. The total data set contains 300 samples.

Table III shows the results from linear discriminant functions, developed using a linear learning machine.[4-6] The training and prediction sets were chosen randomly. Each entry in Table III is the average of results obtained for training five sets with the indicated number of intrinsic variables, random variables, and λ values.

It is apparent from Table III that the predictive ability at any one value of λ is unaffected by changing the ratio of intrinsic and nonintrinsic variables as long as the total number of variables remains constant. For sets trained above a λ of 2.4, little if any effect upon the predictive ability would be expected due to chance correlations. In these cases the predictive ability reflects the ability of the training set to represent the entire data set. The results in Table IV indicate that this is the case. Shown are the results for developing discriminants using five randomly selected training sets. If the increase in the predictive ability for the classifiers were due to the decreasing probability of chance correlations, then it would be expected that the set with 200 members would lose its predictive ability more slowly than the 100-member set. A decrease in predictive ability which is similar for each set would indicate that the difference in predictive ability between a set of 100 members and a set of 200 members arises because 200 members are more representative of the data than are 100. As the correlation between the decrease in predictive abilities is 0.98, the latter argument seems the most plausible.

Based upon these observations, one might assume that the lower predictive ability for sets having a λ below 2.4 (Table III) is due to these same effects. However, below 2.4 the predictive ability is reflecting the effects of chance correlations. To demonstrate this, a feedback feature selection method[3] was used to select those variables which the classifier indicated as being intrinsic. None of the training sets with a $\lambda < 2.44$ were able to differentiate between the intrinsic variables and the nonintrinsic variables. This demonstrates that the sets with low values of λ were developing relations in which nonintrinsic variables played a significant part. The low predictive abilities reflect the randomizing influence which the nonintrinsic variables imparted to the classifier. In contrast, feature selection of the data above a λ of 2.44 showed only the intrinsic variables as those being used to classify the data.

SUMMARY

In this paper it was demonstrated how artifactual classifications can arise from improper use of nonparametric linear classifiers. While extensive theoretical treatments of these techniques have appeared in pattern recognition literature, little if any explanation of the proper conditions under which these techniques should be applied has been made in the chemical literature. As a result, several pattern recognition investigations have been conducted under questionable conditions. The reliability of relations derived through use of pattern recognition techniques is dependent upon their proper application. The basic arguments included here hold regardless of the form used for the discriminant function. These arguments concern the number of samples required to ensure that a nontrivial relationship is present.

It was demonstrated that although a relation may indeed be present, classification attempts at low values of λ (<3.0) fail to uncover that relation. At low values of λ the probability of obtaining a separating linear discriminant function using random data is high.

The inability to detect a relationship actually contained in the data was demonstrated through use of a feature selection technique which uses the results of the classification to determine those features which the classifier deemed most important. For the data sets with too small values of λ, the classifier indicated many of the random variables to be intrinsic to the classification process. This suggests that this feature selection technique, like the classification technique, is of little utility when the data are over-determined. One cannot start out with an over-determined data set and use the results of classification to lower the number of features to obtain an acceptable value of λ.

The behavior of the predictive ability of a classifier was also shown to depend upon the value of λ. For extremely low values of λ, the predictive ability reflects the random behavior of the classifier. The predictive ability of a classifier which is not over-determined was seen to be governed by how well the training set represented the data set.

CONCLUSION

Nonparametric linear classifiers developed at λ values below 3.0 are unable to distinguish relationships contained within a set of measurements. Results obtained from classifiers operated in this region are of very little utility in elucidating properties within a set of data.

LITERATURE CITED

(1) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Inc., Reading, Mass., 1974, p 58.

(2) Reference 1, p 61.
(3) G. S. Zander, A. J. Stuper, and P. C. Jurs, *Anal. Chem.*, **47**, 1085 (1975).
(4) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, N.Y., 1955.
(5) P. C. Jurs and T. L. Isenhour, "Chemical Application of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1973.
(6) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, New York, N.Y., 1973.

NEWS AND NOTES

CA Selects: A New Service

Chemical Abstracts Service is initiating a new information service aimed at giving individual scientists and engineers a convenient and affordable way to keep up with current developments in specialized areas of chemical science and technology.

The service, called *CA Selects*, consists of a series of inexpensive, biweekly publications, each containing the complete *Chemical Abstracts* abstracts and citations for current publications on a specific chemical or chemical engineering topic. The contents of each publication are selected by searching the CAS computer-readable information base with a special search profile developed for the topic.

CA Selects is designed to allow individuals and organizations to tailor their own current-awareness services by subscribing to the service on an appropriate combination of topics. CAS emphasizes that *CA Selects* is strictly a current-awareness service. No indexes of any kind are included.

CAS is offering the *CA Selects* service initially on six topics: organosilicon chemistry, forensic chemistry, photochemistry, high-speed liquid chromotography, mass spectrometry, and psychobiochemistry. Other topics will be added to the service gradually. Service on a particular topic will be continued as long as there is sufficient subscriber interest in the topic to justify the cost of printing and distributing the abstracts.

CAS can now publish specialized collections of abstracts oriented to the interests of relatively small groups of subscribers economically because all of the content of *Chemical Abstracts* is processed in computer-readable form and photocomposed for printing through a highly efficient computer-controlled photocomposition system. Citations pertinent to a particular topic can be identified through a computer search, and the associated abstracts can be selected automatically from the file and photocomposed at a relatively low cost through the same system used to compose *Chemical Abstracts*. Abstracts in the *CA Selects* publication are identical with those published in *Chemical Abstracts*.

The United Kingdom Chemical Information Service pioneered a similar alerting service in the UK several years ago. UKCIS's *Macroprofiles* service, which is derived by searching CAS's *CA Condensates* computer-readable file, offers subscribers biweekly listings of citations, but not abstracts, of new publications on 45 specialized topics. Two of the most popular topics in the *Macoprofiles* series, high-speed liquid chromatography and photochemistry, have been included in the initial *CA Selects* series with the cooperation of UKCIS, which is marketing *CA Selects* in the UK and Ireland.

CAS is seeking suggestions for other topics to be included in the *CA Selects* service. The suitability of a particular topic for the service will be influenced both by subscriber interest in the topic and the number of abstracts on the topic that CAS processes during a two-week period. CAS believes that biweekly publications containing 100 to 200 abstracts will prove to be the optimum size for the service, although some of the initial topics in the series will produce smaller publications. CAS officials feel that biweekly compilations of abstracts on broader topics would tend to be too large to be a convenient current-awareness tool for the individual, while very narrow topics might not be of interest to enough subscribers to recover the cost of composing, printing, handling, and mailing the abstracts. They note, however, that several highly specific topics might be combined in a single biweekly publication that could be marketed economically. They also point out that a *CA Selects* publication could be tailored to replace an internal abstracting bulletin for a company or organization.

Subscription prices for *CA Selects* have been set at $50 per year for the services on forensic chemistry, photochemistry, high-speed liquid chromatography, and mass spectrometry, and at $55 per year for the services on organosilicon chemistry and psychobiochemistry, which will contain larger numbers of abstracts. A quantity discount of $5 per subscription is offered for 25 to 49 subscriptions to any combination of topics delivered to one address and a discount of $10 per subscription for 50 or more subscriptions to the same address. Complimentary issues of *CA Selects* may be obtained by writing the Marketing Department, Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210.

BIOSIS Photocomposition

The abstracts section of *Biological Abstracts* is now being photocomposed through an extension of the computer system used to compose *Chemical Abstracts* in a joint undertaking by BioSciences Information Service (BIOSIS), publisher of *Biological Abstracts*, and Chemical Abstracts Service (CAS). The abstracts section of BIOSIS' monthly specialty publications *Abstracts* of *Entomology* and *Abstracts of Mycology* also are produced through the new procedures, which take advantage of systems compatibility achieved by the two services over the past several years.

BIOSIS converts its magnetic tape record of abstracts to the standard file format for computer-readable information used internally by CAS and forwards the tapes to CAS's Columbus offices twice monthly. The CAS photocomposition system, which has been extended to accomodate the page-layout characteristics of BIOSIS publications, produces photographic page positives, which are returned to BIOSIS where the issues are assembled.

The joint photocomposition effort is the result of several years of cooperative effort between BIOSIS and CAS. It is expected to result in increased processing efficiency and lower production costs for BIOSIS and more efficient utilization of CAS's photocomposition programs and equipment. Cooperative efforts between the two services are continuing, and both expect to derive further benefits from them.