(4) Roth, D. L. "The Role of Subject Expertise in Searching the Chemical Literature...and Pitfalls That Await the Inexperienced Searcher". *Database* **1985**, *8* (1), 43–46.

(5) Lear Siegler, Inc. v. Aeroquip Corp. *U.S. Pat. Q.* **1984**, *221*, 1025–1034.

(6) Fisanick, W. "Requirements for a System for Storage and Search of Markush Structures". In "Computer Handling of Generic Structures"; Barnard, J. M., Ed.; Gower: Hampshire, U.K., 1984; pp 106–127.

(7) *Current Abstracts of Chemistry and Index Chemicus* is published by the Institute for Scientific Information. Philadelphia, and is available as Index Chemicus Online through Telesystems/Questel. *APILIT* is published by the American Petroleum Institute's Central Abstracting and Indexing Service and is available online through ORBIT Search Service. *Ringdoc* is a pharmaceutical literature abstracting service published by Derwent Publications Ltd., London, and is available online through ORBIT Search Service and DIALOG Information Services. *Theilheimer's Synthetic Methods of Organic Chemistry* is published by S. Karger, Basel, Munich, Paris, London, New York, Tokyo, and Sydney, as a yearbook based upon the Chemical Reactions Documentation Service, produced by Derwent Publications Ltd., London, and available online through ORBIT Search Service.

(8) *Central Patents Index* is published in print, microfilm, and magnetic tape forms by Derwent Publications Ltd., London, and is available online as a component of the World Patents Index database through ORBIT Search Service, Telesystems/Questel, and DIALOG Information Services. *CLAIMS Uniterm* is published in print and magnetic tape forms and *CLAIMS Comprehensive Database* is published in magnetic form by IFI/Plenum Data Co., Alexandria, VA, and both are available online through DIALOG Information Services.

A Numerical Index for Characterizing Data Set Separation[†]

DIANA HUNTER LAFEMINA and PETER C. JURS*

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

Received August 14, 1984

A method is reported for assessing the degree of separation between two clusters of points representing chemical data. The method, based on trend vectors, has been tested on both randomly generated and actual data sets. The index is intended to be a quick method for detecting the relative degree of separation between data sets (i.e., suitability of descriptors being used) in structure–activity studies or other pattern-recognition studies.

In pattern-recognition studies, a data set is represented as a set of points in a high- (greater than three) dimensional space whereby clustering, mapping and display, or discriminant development methods are commonly used to investigate the data. One of the basic assumptions underlying pattern recognition is that the distance between points is related to the similarity between points. This assumption applies to structure–activity relation (SAR) studies where the structure of organic compounds of common biological activity are encoded by numerical descriptors and represented as points in high-dimensional space. Compounds of similar activity are expected to be grouped in the same general region of the data space, where the groups may be separated or overlapped. The greater the degree of overlap, the less information can be gained from the system, which reflects the unsuitability of the descriptors being used. There is also the added problem of intuitive judgements being made as to class separability. Visual interpretation of class information is not feasible because most structure–activity studies are done in more than three dimensions. Therefore, a method to quantify the degree of class separability is needed.

This study focuses on the development and use of numerical index for characterizing data sets, analyzing both the effect of size variation and distance separation between data sets. The strategy presented can also be used to assess the degree of similarity between compounds in an assigned class, as well as between a new compound and well-defined classes.

When attempting to understand the structure of a data set that appears to be disordered, clustering routines prove to be very useful.[1] Although there are several ways of defining a cluster,[2,3] distance metrics are often used because the distance between points is a convenient method for establishing similarity between patterns in Euclidean space.[4,5–10] The distances between points can be used to include or exclude a point from a given class. Points grouped in the same class are considered to be similar, while those points that lie outside the class are different.[11]

A variety of cluster-defining techniques have been reported in the literature. Reviews of many of these techniques can be found in Sneath and Sokal,[8,9] Everitt,[12] Hartigan,[2] and Späth.[10] For example, Ling[13] proposed a specific definition of a cluster and also two indices for measuring compactness and relative isolation of these clusters. The nearest-neighbor algorithm of Sneath[14] is another formulation of the same approach. The majority of these clustering techniques calculate similarities and distances between data points so as to summarize information regarding their possible relationships,[12] while we have focused on developing a general numerical index of characterizing data set separation. The trend vector distance ratio index (the *R* index) yields an initial, quick indication of the relative degree of class separation/overlap (i.e., descriptor suitability) occurring in a data set.

The trend vector distance ratio index is based on the concept of trend vectors,[15] where a trend vector is defined as the Euclidean distance between class centers. The centers are representative of the average coordinates for each class point in *n*-dimensional space. The trend is then compared to the sum of the average radii for the two classes being studied. The *R* index provides a measure of class separability and class overlap.

DATA SETS AND METHODOLOGY

In the first step of the study, three data sets (each consisting of five class pairs) were generated to model idealized spherical classes. They provide well-characterized sets in which the effects of separation between class centers and class-size variations could be studied without the added concern of class shape. The data sets were generated to model the two-class situation often found in structure–activity studies. Data set 1 was used to investigate the effect of varying the distance of separation between class centers. Separation distance is defined as the Euclidean distance between two cluster centers. To observe this effect, the size of each cluster was held constant at one standard deviation unit. The ratios of separation to cluster radius were 10:1, 5:1, 3:1, 2:1, and 1:1. These ratios represent separations of one standard deviation (the radius of a cluster) to *n* standard deviations, where *n* varies with the extent of separation desired.

[†] This research was partially supported by Grant CHE-8202620 from the National Science Foundation. The PRIME 750 minicomputer was acquired with the partial financial support of the National Science Foundation.

0095-2338/85/1625-0386$01.50/0

Table I. Trend Vector and Trend Vector Distance Ratio Results

		Data Set 1		
separation-variation ratio[a]	active class distance[a]	inactive class distance[a]	$\|T\|$	R
10:1	432	441	5332	0.16
5:1	636	649	3925	0.32
3:1	733	748	2722	0.54
2:1	775	791	1908	0.82
1:1	803	820	1004	1.61
		Data Set 2		
size-variation ratio[a]	active class distance[a]	inactive class distance*[a]	$\|T\|$	R
10:1	114	1164	561	2.27
5:1	222	1134	1093	1.24
3:1	350	1071	1722	0.82
1:1	686	700	3378	0.41
0.5:1	801	409	3946	0.30
		Data Set 3		
separation-variation ratio[a]	active class distance[a]	inactive class distance[a]	$\|T\|$	R
10:1/10:1	111	1140	1378	0.90
5:1/5:1	220	1124	1359	0.98
3:1/3:1	355	1089	1320	1.09
2:1/1:1	775	791	1908	0.82
1:1/0.5:1	1008	514	1260	1.20

[a] Distance: average point-to-center distance for each class.

Data set 2 focuses on the effect of cluster-size variations. In data set 2, the separation between centroids for each cluster pair was held constant at four standard deviations, one cluster was fixed with a constant size of one standard deviation, and the second cluster's size was varied. The varying cluster's size ranged from 0.5 to 10 times the size of the cluster of constant size.

Data set 3 focused on the effect of simultaneously varying both the size and the distance separation of the clusters. Again, one cluster was held at a radius of one standard deviation in each of the cluster pairs. The second cluster was manipulated to represent changes in size from 0.5 to 10 times the constant-sized cluster as in data set 2. The distance separation was varied from one standard deviation to ten standard deviations as in data set 1. In data set 3, each cluster pair is identified with two ratios. The first ratio is the relative size of the trend vector length to the radius of cluster 1, and the second ratio is the relative size of cluster 2 to cluster 1. The three data sets are described in Table I.

The generation of the data sets used in this study was done with the ADAPT computer software system.[1] ADAPT contains routines for descriptor maintenance during SAR studies, which can also be used for a model study such as this. A class-forming routine that uses points developed by a random number generator was used to generate the point sets from two input parameters: the mean and standard deviation for each of ten descriptors. Ten descriptors were used as this was a large enough number to be a realistic model for actual structure–activity studies but was small enough that the computational burden was not excessive. A total of 100 points was used for each cluster pair, with 50 in each of the two classes. The distance separation for a set of classes can be obtained from a knowledge of the mean for each of the ten descriptors; the standard deviation of a class (i.e., its radius) can be obtained from the points after they have been generated. In these data sets, the classes are spherical; that is, the standard deviation is the same along each of the ten dimensions, with some variability occurring among the points due to the fact that they were generated by a random number generator.

After the raw descriptors were generated, they were preprocessed for use with the analysis routines. The preprocessing consisted of autoscaling (a common preprocessing method in SAR and other pattern-recognition techniques), where each descriptor is scaled so that it has a mean of zero and a standard deviation of unity, removing any effects of variability in scale between descriptors.

Table II. Real Data Set Information

(Data Set 4) 9-Anilinoacridine Data Set
213 total compounds
153 compounds in active class
60 compounds in inactive class
200 compounds correctly classified (94%)
18 descriptors used

(Data Set 5) *N*-Nitroso Compound Data Set
150 total compounds
112 compounds in active class
38 compounds in inactive class
146 compounds correctly classified (97%)
22 descriptors used

(Data Set 6) Pyrolysis Gas Chromatography Data Set A
144 total chromatograms
72 chromatograms in active class
72 chromatograms in inactive class
138 chromatograms classified correctly (96%)
5 descriptors used

(Data Set 7) Pyrolysis Gas Chromatography Data Set B
144 total chromatograms
72 chromatograms in active class
72 chromatograms in inactive class
130 chromatograms classified correctly (90%)
9 descriptors used

(Data Set 8) Pyrolysis Gas Chromatography Data Set C
144 total chromatograms:
72 chromatograms in active class
72 chromatograms in inactive class
128 chromatograms correctly classified (89%)
13 descriptors used

(Data Set 9) Pyrolysis Gas Chromatography Data Set D
144 total chromatograms
72 chromatograms in active class
72 chromatograms in inactive class
144 chromatograms classified correctly (100%)
12 descriptors used

The analysis of these data sets was done by using a numerical index of class separation. The characterization routine uses vectors and several class-based distance measures in determining the degree of class separation. The trend vector is defined as the vector joining the centers of two classes. The center of each class is calculated by taking the mean of each coordinate for all points within a class. The final trend vector magnitude, $|T|$, for a class pair is the magnitude of the vector between the class centroids.

As the vector does not take into account class size, an additional distance measure was needed. The sum of the average radii for each class was calculated so as to provide an indication of class size. The trend vector distance ratio (R index) was then calculated from

$$R = \frac{r_1 + r_2}{|T|}$$

where r_1 = average radius of class 1, r_2 = average radius of class 2, and $|T|$ = trend vector magnitude. The average radius was calculated as the average of all point-to-center distances for each class. This provides only an approximation of the actual average radius.

From the results obtained with the theoretical data sets, a trend vector distance ratio index (the R index) was proposed. The final stage of the study involved the testing of the proposed index on six real data sets. The data sets were those from two structure–activity relationship studies (data sets 4 and 5) and those from four different sets of descriptors from a pyrolysis

Table III. The Six Real Data Sets: Trend Vector and Trend Vector Distance Ratio Results

data set no.	% correctly classified[a]	active class distance[b]	inactive class distance[b]	$\|T\|$	R^c	index separation prediction
9	100	796	593	2223	0.64	very good separation, minimal or no overlap present
5	97	458	766	1862	0.66	very good separation, minimal or no overlap present
6	96	736	862	1684	0.95	good separation, approaching overlap
4	94	288	749	1931	0.54	very good separation, minimal or no overlap present
7	90	536	849	1243	1.1	poor separation, overlap present
8	89	663	869	1158	1.32	poor separation, overlap present

[a] The percentage of compounds classified correctly decreases as the overlap between the classes increases. [b] Distance: average point-to-center distance for each class. [c] The *R* index predictions correlate well with the classification results reported.

gas chromatography data analysis project (data sets 6–9). These data sets are described in Table II. The percentage of compounds of chromatograms correctly classified by nonparametric linear discriminants is reported for each data set.

The Penn State University of Chemistry PRIME 750 minicomputer was used to perform all the computations done during this study.

RESULTS AND DISCUSSION

Table I shows the values calculated for all the numerical indices described for each of the three model data sets. From the values related to the separation of the classes (see column five), a set of prediction ranges suggest themselves: >1.0, poor separation/much overlap; 0.8–1.0, separation, but approaching significant overlap; <0.8, good separation/minimal or no overlap.

As the mathematical definition suggests, when the magnitude of *R* decreases, the degree of separation between clusters increases and can be considered as total class separation for magnitudes of 0.55 or less. At a value of 1, the class model consists of two spheres that contain points and whose boundaries are just touching; the sum of the radii of the two classes is equal to the length of the trend vector between class centers. This implies (for any class pair of any size variation) that the classes are theoretically separated. Minimal overlap is predicted as the classes are computer-generated and smooth, continuous outer radii are not expected. Therefore, the determining value for the prediction of overlap was set at a magnitude of 1. A third range, representing the occurrences where classes are approaching significant overlap (values less than 1), eliminates the need for intuitive judgements to be made by the user. Any calculated value greater than 1 represents some degree of overlap so no corresponding range was needed. In this way, the trend vector distance ratio index retains a general, standardized nature and provides a quick physical quantification of the degree of separation occurring between class pairs.

The *R* index was then tested on the six real data sets. The method correctly classified all data sets as to the degree of separation occurring between the inactive and active classes (see Table III). In these data sets, the degree of separation, or overlap, was related to the number of compounds misclassified; the greater the number of misclassified compounds, the greater the degree of class overlap (see Table II for data set information). The class shapes not defined for these cases as two-dimensional plots do not accurately represent the *n*-dimensional classes. The class size may be related to the number of compounds occurring in the class.

In data sets 4, 5, 6, and 9, the class pairs exhibited similar degrees of separation, with the percentage of misclassified compounds being less than 7% in each case. Data sets 7 and 8 exhibited poorer cluster separation, with misclassification being 10% and 11%, respectively. The *R* index results are illustrative of this overlap with the calculated values being greater than 1. The variation of the *R* values was not directly correlated with the percentage of compounds classified correctly (see Table III). The variation may be explained by differences in class shape.

CONCLUSIONS

The expected results for the model data sets were indeed shown (i.e., a decrease in class size will decrease *R*), leading to the *R* index used in characterizing the real data sets. In testing the index on the real data sets, it was concluded that there was no direct correlation between the magnitude of *R* and the number of correctly classified compounds or chromatograms in each set. Cluster separability was related to correctly classified compounds or chromatograms, while the size of the classes was related to the number of compounds or chromatograms in each. All real data sets were correctly classified as to the degree of separation occurring in each.

The trend vector distance ratio index provides a quick, general, and standardized measure of class separability, eliminating the use of intuitive judgements. In quantifying the degree of class separability (or overlap), the *R* index provides an accurate representation of the physical picture of class separability and yields information related to the suitability of the descriptors being used.

This study represents only an initial step in the area of data characterization. Factors such as cluster shape may affect separation quantification results to some degree (and spherical classes only occur with certain types of clustering algorithms), so further research into varying class shapes and comparison of different interclass pairwise point distances coonstitutes the next phase in additional class separability characterizations.

ACKNOWLEDGMENT

D.H.L. thanks John P. LaFemina for his many helpful discussions during this work. Special thanks also to Terry R. Stouch for his constructive comments and time.

Registry No. 9-Anilinoacridine, 3340-22-5.

REFERENCES AND NOTES

(1) Stuper, A. J.; Brügger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley-Interscience: New York, 1979.

(2) Hartigan, J. A. "Clustering Algorithms"; Wiley: New York, 1975.

(3) Barrett, V., Ed. "Interpreting Multivariate Data"; Wiley: New York, 1981.

(4) Tou, J. T.; Gonzalez, R. C. "Pattern Recognition Principles"; Addison-Wesley: Reading, MA, 1974.

(5) Whalen-Pederson, E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 264.

(6) Massart-Leën, A.; Massart, D. D. *Biochem. J.* **1981**, *196*, 611.

(7) Snedecor, G. W.; Cochran, W. G. "Statistical Methods", 6th ed.; Iowa State University Press: Ames, IA, 1974.

(8) Sneath, P. H.; Sokal, R. R. "Numerical Taxonomy"; Freeman: London, 1973.

(9) Sokal, R. R.; Sneath, P. H. "Numerical Taxonomy"; Freeman: London, 1963.

(10) Späth, H. "Cluster Analysis Algorithms"; Holsted Press: New York, 1980.

(11) Dubes, R.; Jain, A. *Pattern Recognition* **1976**, *8*, 247.

(12) Everitt, B. S. "Cluster Analysis", 2nd ed.; Heinemann: London, 1980.

(13) Ling, R. F. *J. Am. Stat. Assoc.* **1973**, *68*, 159.

(14) Sneath, P. H. *J. Gen. Microbiol.* **1957**, *17*, 201.

(15) Carhart, R. E., personal communication.