Clustering Tendency in Chemical Classifications

PETER WILLETT

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Received August 20, 1984

Increasing use is being made of cluster analysis methods for the analysis of chemical data. A clustering method will always result in some degree of grouping when it is applied to a data set, and it is thus of importance to know whether the data set exhibits any real tendency to cluster. This paper tests the applicability of a cluster validity method to several chemical data sets.

INTRODUCTION

Cluster analysis, or automatic classification, methods are increasingly being used for the analysis of chemical data. A characteristic of such methods, and one that is sometimes not appreciated, is that they will identify groupings within a data set irrespective of whether there is actually any cluster structure present. It is therefore of interest to have available some quantitative test to determine whether the clusters identified are, in some sense, real or whether they have arisen solely from the operation of the clustering method. Several such *cluster validity* tests have been described in the literature,[1] and this paper considers the utility of one such test, that due to Ling and Killough,[2] in the context of chemical data sets.

CLUSTER VALIDITY STUDIES

In their excellent review,[1] Dubes and Jain identify three main types of cluster validity study. The first of these seeks to determine whether the clustering tendency is such as to suggest that the data are significantly different from random, since it is clearly inappropriate to consider the application of a clustering procedure if this is not the case; these studies of clustering tendency have generally involved the description of randomness in either graph-theoretic or geometric terms. Once it has been shown that a data set does indeed exhibit a nonrandom clustering tendency, tests may be carried out to determine the extent to which the output from the classification procedure recovers the structure of the data set; this is usually done by using some measure of global fit such as the cophenetic correlation coefficient. Finally, once the overall structure has been confirmed, the validity of individual clusters within a hierarchy or a partition may be of interest.

The *random graph hypothesis* may be considered as a null hypothesis that is invoked to determine whether any significant predisposition to cluster is present within a data set. Given a data set containing n objects, it is assumed that a symmetric $n \times n$ proximity matrix, P, is available, the elements of which, p_{ij}, contain the degree of (dis)similarity between each pair of objects i and j; in the remainder of this paper, it will be assumed that a dissimilarity measure, specifically the Euclidean distance, has been used for the generation of P. The rank matrix R is another $n \times n$ symmetric matrix, the elements of which, r_{ij}, contain the rank orders of the corresponding elements in P after they have been sorted into order of increasing distance; the upper portion of R will accordingly contain all integers in the range 1 to $n(n-1)/2$. The random graph hypothesis is that all $[n(n-1)/2]!$ such matrices are equally likely, and this null hypothesis may be used to study various characteristics of clusterings. One such characteristic is the minimum number of edges, v_{min}, at which a random graph on n nodes becomes connected, i.e., when all of the nodes are contained within a single cluster. Following earlier, and approximate, work by Fillenbaum and Rapoport[3] and Schultz and Hubert,[4] Ling[5] found an accurate method for the enumeration of all connected graphs on n nodes and v edges, and Ling and Killough[2] used this to calculate tables for the probability of observing specific values of v_{min} under the random graph hypothesis given a graph containing n nodes.

The connected subgraphs of a graph may be obtained by setting some threshold distance d and then linking together all pairs of objects for which $p_{ij} \leq d$. The resulting subgraphs correspond to the clusters formed at that threshold by the nearest-neighbor, or single linkage, clustering method, and Ling and Killough's tables may hence be used by studying nearest-neighbor cluster hierarchies. Specifically, a note is made of the rank in R at which all of the objects in a data set become contained within a single nearest-neighbor cluster: this rank may then be compared with the tables that give values for the probability of observing such a value of v_{min} on the assumption that the graph is not a random one. If this probability is greater than some arbitrary threshold value, for which Dubes and Jain[1] suggest 0.99, then evidence exists for the conclusion that R is significantly different from a random graph and that some nonrandom clustering tendency is present in the data set. Thus, the degree of clustering tendency may be tested for simply by determining whether the observed v_{min} is large when compared to the distribution of v_{min} under the assumption of a random graph.

Several points should be made about the use of the test in a practical context. First, the clusters in a nearest-neighbor classification depend only upon the rank order of the distances and not upon the actual value; this characteristic may be used to advantage since most cluster analysis packages, such as the CLUSTAN routines used in the experimental work reported below, are based upon the use of distance matrixes rather than rank matrixes. Specifically, a classification is generated from a distance matrix, and then the distance at which a data set becomes fully connected may be converted to a rank order by identifying its location in a sorted list of the elements of P. Second, the test has at least two limitations. The use of the point at which the graph becomes fully connected reflects the clustering tendency at only a single rank in the evolution of the hierarchy, and the test may thus be strongly affected by outliers that are far removed from the other members of the data set: this point is discussed further below. A further limitation is that the tables can be used for data sets containing only 100 objects or less, this constraint arising from the extremely heavy computation that is required for the enumeration of the connected components as the size of the data set increases.[2] The third and final point that needs to be made is that Ling and Killough's tables provide a means for determining whether a data set exhibits clustering behavior that is noticeably different from that of a random graph. The tables do not provide a formal test of statistical significance for the existence of clusters since this would require a meaningful alternative hypothesis that embodied some explicit definition of cluster structure; in the absence of such a hypothesis, the primary use of the tables is in helping to decide whether clustering techniques are an appropriate means of analyzing a particular data set.

0095-2338/85/1625-0078$01.50/0

Table I. Parameter Values for a Set of 15 Substituents

substituent	π	E_s	M_r	F	R
H	0.00	1.24	1.03	0.00	0.00
n-hexyl	3.05	−0.40	28.87	−0.06	−0.09
SMe	0.61	0.17	13.82	0.20	−0.18
O-*n*-hexyl	2.55	0.02	30.90	0.25	−0.55
CN	−0.57	0.73	6.33	0.51	0.19
Br	0.86	0.08	8.88	0.44	−0.17
OH	−0.67	0.69	2.85	0.29	−0.64
t-butyl	1.83	−1.54	19.62	−0.07	−0.13
NHAc	−0.97	−1.74	14.93	0.28	−0.26
NMe_2	0.18	−1.60	15.55	0.10	−0.92
SO_2Me	−1.63	−1.39	13.49	0.54	0.22
CF_3	0.88	−1.16	5.02	0.38	0.19
NPr_2	2.18	−1.60	34.03	0.10	−0.92
NH_2	−1.23	0.63	5.42	0.02	−0.68
SO_2NH_2	−1.82	−1.38	12.28	0.41	0.19

EXPERIMENTAL RESULTS AND DISCUSSION

The experiments involved taking several chemical data sets from the literature that have been analyzed by clustering techniques, generating the nearest-neighbor classifications, and then determining the level at which all of the members of a data set were contained within a single cluster. As noted above, the measure used to determine the degree of dissimilarity between each pair of objects was the Euclidean distance.

There have been several reports in the literature of methods for the identification of analogues in lead optimization programes that are as widely separated as possible in physicochemical parameter space.[6-9] Ling and Killough's test may be used to determine whether any clustering tendency is evident in a set of substituents that has been selected by one of the methods since, if such a tendency is identified, it may be assumed that the set is not as well separated as desirable. Such a substituent set has been described by Wooldridge[8] who lists 15 substituents characterized by the physicochemical parameters π, E_s, M_r, R, and F; this data set, which is listed in Table I, will be used to illustrate the workings of Ling and Killough's procedure. After standardization of the data, an intersubstituent distance matrix is obtained as shown in Table II. The corresponding nearest-neighbor classification is represented by the dendrogram of Figure 1, from which it will be seen that the 15 substituents become connected into a single cluster at a distance of 0.985. This is the 24th smallest distance in the matrix, and thus v_{min} = 24; the critical value for a graph containing 15 nodes, i.e., n = 15, is 40, and it may hence be concluded that a well-separated set of substituents has been identified in which there is no evident predisposition to cluster.

A practical example of the use of substituent selection techniques is given by Dunn et al.,[7] who synthesized 14 derivatives suggested by the methodology of Hansch et al.[6] in a study of antitumour triazenes. With use of the parameter data given in their paper[7] for π, M_r, M_w, and σ, the nearest-neighbor clustering method was found to give a classification in which all of the substituents were contained within a single cluster at v_{min} = 30. Reference to the tables for n = 14 shows that the critical value is at about 36, and thus, a well-separated set of substituents has indeed been obtained, as with the Wooldridge data set. Conversely, replacement of the σ values by the Swain–Lupton F and R parameters, which are also listed in the paper, results in a hierarchy with v_{min} = 63, this suggesting that a definite clustering tendency is present in the data with these parameter types. However, an inspection of the dendrogram for this classification, which is shown in Figure 2, reveals clearly that an outlier substituent is present that joins the hierarchy long after the other 13 substituents, which are entirely connected at v_{min} =21. This is considerably below the critical value for n = 13, and thus, the nonrandom tendency is due to this solitary outlier.

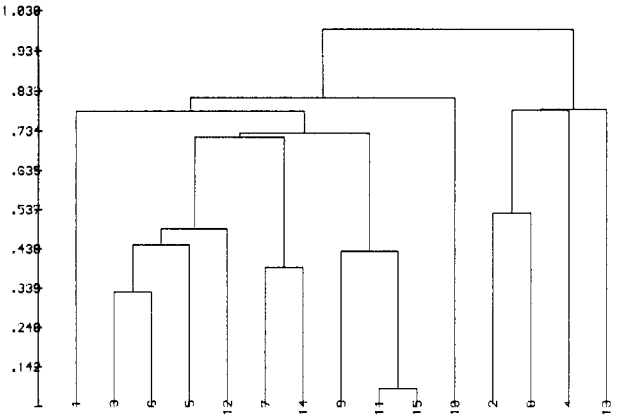

Figure 1. Dendrogram for the nearest-neighbor classification corresponding to the distance matrix of Table II.

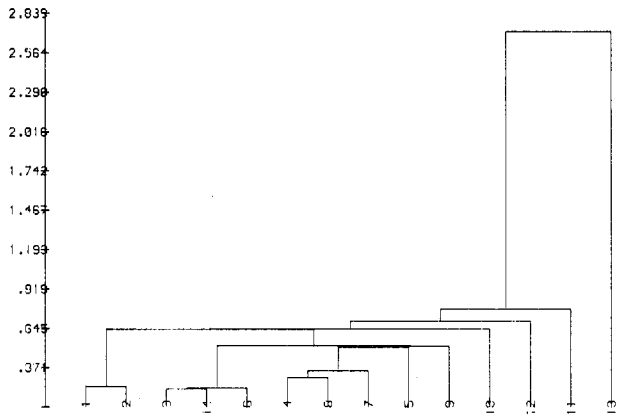

Figure 2. Dendrogram showing the presence of an outlier substituent.

Similar behavior is exhibited by a set of 38 benzodiazepine derivatives, characterized by eight physicochemical parameters, which has been used by Miyashita et al.[10] in a study of clustering methods based on minimal spanning trees. The

Table II. Standardized Distance Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	2.772													
3	0.782	1.324												
4	3.188	0.785	1.058											
5	1.417	3.953	0.912	3.065										
6	1.387	2.408	0.329	1.514	0.447									
7	1.013	3.632	0.719	2.466	1.116	0.718								
8	2.393	0.528	1.075	1.439	3.498	2.023	2.913							
9	2.526	2.649	0.918	2.204	1.774	1.089	1.543	1.317						
10	2.981	2.278	1.325	1.678	3.543	1.894	1.595	1.180	0.815					
11	3.236	4.317	1.621	3.555	1.018	1.196	2.305	3.018	0.674	2.842				
12	1.883	2.605	0.802	2.524	0.914	0.488	1.736	1.603	0.837	2.202	0.801			
13	5.043	1.361	2.286	0.788	5.427	3.117	3.756	1.323	2.232	0.985	4.582	3.688		
14	0.812	3.247	0.925	2.778	2.134	1.615	0.390	2.449	1.746	1.373	3.187	2.527	3.543	
15	2.622	3.852	1.323	3.433	1.060	1.183	1.992	2.461	0.432	2.366	0.088	0.728	4.255	2.529

observed v_{min} value is 299, which is well above the tabulated critical value for a data set of this size. However, the dendrogram identifies two outlier compounds that join the hierarchy long after the other 36 molecules, all of which are connected at $v_{min} = 85$: this value corresponds to the absence of any significant clustering tendency and raises questions as to the suitability of the data set for the purpose for which it was used.

Rather than grouping substituents, several workers have considered the clustering of sets of molecules on the basis of substructural similarities.[11-13] The first such data set examined contained 36 multisubstituted benzenes with variables representing the frequencies of occurrence of each of seven substituents, regardless of position; this data set had been used previously[12] for comparing hierarchic agglomerative clustering procedures. The observed v_{min} is 35 while the tabulated values for $n = 35$ and 40 are 123 and 145 edges, respectively, and it is thus clear that no significant clustering tendency is present. An inspection of the nearest-neighbor dendrogram reveals a complete lack of structure, and the use of the validity test might hence seem to be redundant. However, Adamson and Bawden have shown[12] that the use of other clustering methods, such as the complete linkage or Ward methods, on this set of compounds results in highly structured dendrograms, and an investigator who used these methods, and who had not applied the test, might well conclude that many well-marked clusters were present. It should be emphasized that, although the nearest-neighbor clustering method is used as a basis for the test, it validates not this particular clustering method but the clustering tendency that is present in the data per se; a random result in the test implies that *any* clustering approach is inappropriate and not just the nearest-neighbor method.

A recent report discussed the use of relocation clustering methods for the study of structure–activity relationships.[13] The experiments involved 11 sets of compounds for which physical, chemical, or biological data were available and evaluated a range of clustering procedures by the extent to which classifications based on substructural similarities also reflected similarities in the chosen property. The nearest-neighbor classifications for nine of these data sets, those containing less than 100 compounds, together with a further five small sets of compounds from the structure–activity literature, have now been analyzed to determine whether they do in fact exhibit any clustering tendency. Each molecule in a data set was characterized by the frequencies of occurrence for the augmented atom types present, and by use of this level of fragment description, random clustering behavior was observed in only two of the sets of compounds, which had been used in earlier studies on molecular connectivity indexes.[14,15] The first of these was a set of 27 aliphatic hydrocarbons, ethers, and ketones for which the property of interest was AD_{100} values for loss of righting reflex in mice; the observed v_{min} value was 60, well below the critical value of 88. The second set of compounds contained antimicrobial activities for 28 substituted phenyl propyl ethers for which the observed v_{min} was 355, as against a critical value for $n = 28$ of 92. However, the observed value arises from an outlier compound that joins the hierarchy long after all of the other members of the data set; accordingly, it is only the presence of this single molecule that causes the compounds to exhibit nonrandom clustering behavior, and the data set should thus be regarded as unclustered. Thus, these two sets of compounds were not entirely suitable for use in a comparative study of clustering procedures.

CONCLUSIONS

This paper has studied the applicability of a cluster validity method to chemical classifications. The method provides a simple and general means of determining whether a data set exhibits a clustering tendency significantly different from that exhibited by a random graph. It is recommended that future chemical applications of automatic classification methods should involve a test of clustering tendency as a precursor to the use of a clustering procedure.

ACKNOWLEDGMENT

Thanks are due to the referees for their comments on an earlier draft of the manuscript.

REFERENCES AND NOTES

(1) Dubes, R.; Jain, A. K. "Validity Studies in Clustering Methodologies". *Pattern Recognition* **1979**, *11*, 235–254.
(2) Ling, R. F.; Killough, G. G. "Probability Tables for Cluster Analysis Based on a Theory of Random Graphs". *J. Am. Stat. Assoc.* **1976**, *71*, 293–300.
(3) Fillenbaum, S.; Rapoport, A. "Structures in the Subjective Lexicon"; Academic Press: New York, 1971.
(4) Schultz, J.; Hubert, L. "Data Analysis and Connectivity of Random Graphs". *J. Math. Psychol.* **1973**, *10*, 421–428.
(5) Ling, R. F. "An Exact Probability Distribution on the Connectivity of Random Graphs". *J. Math. Psychol.* **1975**, *12*, 90–98.
(6) Hansch, C.; Unger, S. H.; Forsythe, A. B. "Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents". *J. Med. Chem.* **1973**, *16*, 1217–1222.
(7) Dunn, W. J.; Greenberg, M.; Callejas, S. S. "Use of Cluster Analysis in the Development of Structure–Activity Relations for Antitumour Triazenes". *J. Med. Chem.* **1976**, *19*, 1299–1301.
(8) Wooldridge, K. R. H. "A Rational Substituent Set for Structure-Activity Studies. *Eur. J. Med. Chem.–Chim. Ther.* **1980**, *15*, 63–66.
(9) Schaper, K. J. "Rational Selection of Test Series for QSAR Analysis". *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1983**, *2*, 111–120.
(10) Miyashita, Y.; Takahashi, Y.; Yotsui, Y.; Abe, H.; Sasaki, S. I. "Application of Pattern Recognition to Structure-Activity Problems. Use of Minimal Spanning Tree". *Anal. Chim. Acta* **1981**, *133*, 615–620.
(11) Adamson, G. W.; Bush, J. A. "A Method for the Automatic Classification of Chemical Structures". *Inf. Storage Retr.* **1973**, *9*, 561–568.
(12) Adamson, G. W.; Bawden, D. "Comparison of Hierarchical Cluster Analysis Techniques for the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 204–209.
(13) Willett, P. "Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29–33.
(14) DiPaolo, T. "Molecular Connectivity in Quantitative Structure-Activity Relationship Studies of Anaesthetic and Toxic Activity of Aliphatic Hydrocarbons, Ethers and Ketones". *J. Pharm. Sci.* **1978**, *67*, 566–568.
(15) Hall, L. H.; Kier, L. B. "Antimicrobial Activity of Substituted Phenyl Propyl Ethers". *J. Pharm. Sci.* **1978**, *67*, 1743–1747.