Pattern Recognition Vol. 12, pp. 177–187.
Pergamon Press Ltd. 1980. Printed in Great Britain.

0031-3203/80/0601-0177 $02.00/0

STABILITY OF A HIERARCHICAL CLUSTERING

STEPHEN P. SMITH and RICHARD DUBES*
Computer Science Department, Michigan State University,
East Lansing, Michigan 48824

(*Received* 24 *July* 1979)

Abstract – Clustering algorithms have the annoying habit of finding clusters even when the data are generated randomly. Verifying that potential clusterings are real in some objective sense is receiving more attention as the number of new clustering algorithms and their applications grow. We consider one aspect of this question and study the stability of a hierarchical structure with a variation on a measure of stability proposed in the literature.[1,2]

Our measure of stability is appropriate for proximity matrices whose entries are on an ordinal scale. We randomly split the data set, cluster the two halves, and compare the two hierarchical clusterings with the clustering achieved on the entire data set. Two stability statistics, based on the Goodman–Kruskal rank correlation coefficient, are defined. The distributions of these statistics are estimated with Monte Carlo techniques for two clustering methods (single-link and complete-link) and under two conditions (randomly selected proximity matrices and proximity matrices with good hierarchical structure). The stability measures are applied to some real data sets.

Cluster stability Hierarchical clustering Single-link method Complete-link method
Random graph.

INTRODUCTION

Exploratory data analysis is for situations when little prior information is available about a set of data and one wants to 'look' at the data and study its 'structure'. Clustering,[3–7] also called 'unsupervised pattern recognition' and 'classification', is an important tool in exploratory data analysis. Cluster validity[8] is concerned with the objective interpretation of the results of clustering algorithms and tries to separate 'true' structures from artifacts of clustering algorithms. This paper is restricted to the stability of clustering structures generated by hierarchical clustering algorithms operating on rank-order proximity matrices. Most of the quantitative results in cluster validity have been generated under these restrictions.[9–14]

A proximity matrix is an $N \times N$ symmetrical matrix in which each row and column represents a data item, or pattern to be clustered, and whose entries, called proximities, express the degree of similarity (e.g., correlation) or dissimilarity (e.g., Euclidean distance) between data items. Furthermore, the original proximities are replaced by their rank orders, with no ties.

A clustering method, in the form of a suitable algorithm, is applied to the rank-order proximity matrix and a sequence of N-1 nested partitions of the N data items is formed. The two clustering methods treated in this paper, single-link and complete-link,[15,16] depend only on the rank order of the proximities, so the assumption of ordinal data is not restrictive. A dendrogram, or binary tree, records the successive merging of clusters as the algorithm proceeds from the disjoint clustering (N data items in N individual clusters) to the conjoint clustering (all data items in a single cluster), assuming an agglomerative algorithm is employed. The actual rank orders at which the mergings occur are also recorded.

By way of example, a rank-order proximity matrix is shown in Fig. 1 in which rank 1 means the closest pair. This matrix was derived by selecting six points at random from a unit square and ranking the Euclidean distances between all pairs of points. The points themselves, along with the single-link and complete-link proximity dendrograms are shown in Fig. 1.

Four general problems immediately appear when one tries to evaluate the results of a hierarchical clustering.[8] The order in which these problems are attacked depends on the strategy employed for validating the hierarchy.

(1) Clustering tendency. Do the entries in the proximity matrix indicate a homogeneous structure or does some sort of clustering exist among the data items?

(2) Global fit of the hierarchy. Is the data set well represented by the hierarchy?

(3) Global fit of a partition. Are any of the partitions achieved in the hierarchy good summaries of the data?

(4) Validity of individual clusters. Ling[12,13] describes a 'real' cluster as one that, for its size, is born early in the dendrogram and lasts a long time. Which, if any, of the individual clusters are 'real'?

Solutions to these problems depend on several

* Research supported by NSF grant ENG 76-11936A01.

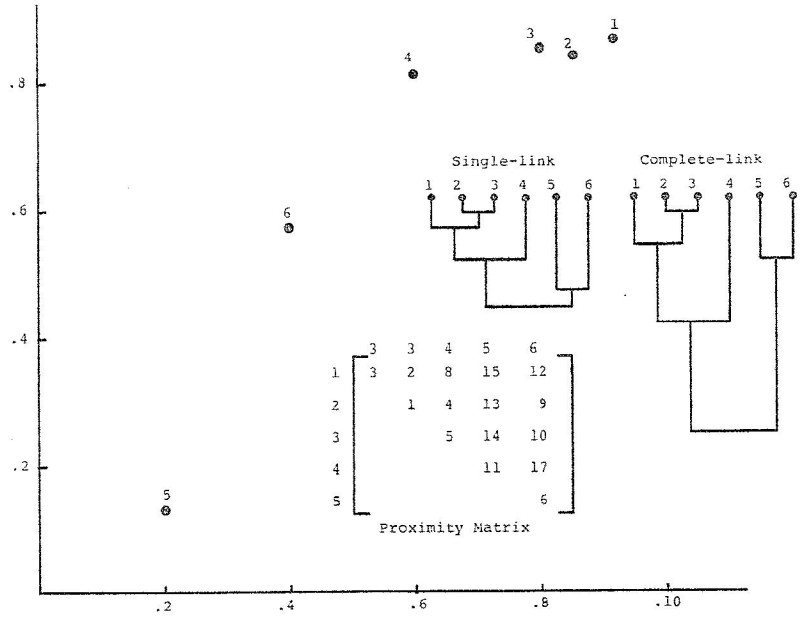

Fig. 1. Example of clustering methods.

factors. The choice of clustering method will strongly influence the details of the dendrogram. An exception is when the proximity matrix is ultrametric,[16,17] in which case the single-link and complete-link clustering methods produce the same dendrogram and the dendrogram generated matches the proximity matrix exactly, in that the matrix can be exactly reconstructed from the dendrogram. Real data seldom satisfy the very tight constraints imposed by the ultrametric inequality. Besides the clustering method, other factors affecting the choice of cluster validity measures are: the meaning of 'no clustering', the prior concept of an ideal cluster, the size of the proximity matrix, and the ultimate use of the results.[8]

The concept of stability is not well defined, yet it has application to each of the above four problems.[24,25] One possible interpretation of stability applies to (4) above. When evaluating a cluster generated by some clustering method, we can slightly change the characteristics of the test data set, for instance by choosing more patterns from the underlying population from which the original patterns were chosen. If the new proximity matrix is then clustered, one would intuitively expect the membership of the cluster under consideration to remain stable with, possibly, a few new patterns added to the cluster. Similarly, for evaluating an entire clustering [(3) above] one would expect the cluster membership among all the original clusters to remain the same, with no switches occuring among the original patterns. For stability of a hierarchical structure, one would expect the old structure to be reproduced, with the added patterns only contributing new binary merges in the dendrogram.

Since choosing more patterns is, in many instances, either impractical or impossible, we consider a slightly different notion of stability using only the original patterns in the proximity matrix. Strauss *et al.*[1,2] proposed two methods for testing the stability of a clustering. The first idea is to randomly split the sample of N data items and cluster each half independently. Cluster memberships in the two halves should be similar to memberships in the entire sample if the clustering is stable. The second idea is to add or delete some of the variables used to compute the original proximity measure and compare cluster memberships.

This paper studies a variation of Strauss' first idea. Although strategies could be devised to use this idea in attacking any of the four basic problems, we choose to concentrate on a measure of global fit. We reason that a 'stable' hierarchical structure contains 'real' hierarchical structure while an 'unstable' one shows little global fit. We ask the question: When is a hierarchical structure unusually stable?

The objective validation of clusterings is necessary for the scientific application of clustering methodology. Nevertheless, relatively little attention has been paid to cluster validity in general,[8] and to the concept of stability in particular. Jardine and Sibson [17, p. 110] consider the stability of clusters as a function of the number of features and the sensitivity of clustering methods to data errors and missing data.

[17, p. 86] However, they provide no statistical or empirical studies of stability measures. Mezzich,[19] as part of a study comparing clustering algorithms, identified 'clustering replicability' by randomly halving the data and assessing similarity between clusterings of the halves. Mezzich computed a correlation coefficient between corresponding entries of cross-classification tables obtained on the two halves. Rand[20] evaluated two clustering methods by assessing similarity between clusterings; his new data was conceived from the old by drawing new individuals from the underlying population. Baker,[21] who is again interested in the stability of a clustering method rather than the stability of a particular data set, evaluates the effect of removing a few objects from the data set on the complete-link clustering structure. His results apply only for 16 data items.

DEFINITION OF STATISTIC

Suppose the N data items have been partitioned into two groups. How can the clustering structure of the two groups be compared to the structure of the original data to judge stability? We propose using the Goodman–Kruskal gamma statistic.[18, also see Appendix 1] Hubert[11] has used the gamma statistic previously for a similar purpose. Stability measures based on gamma statistics are defined below.

Let P denote the given $N \times N$ rank-order proximity matrix. The entries are called proximity ranks. Let the N data items be partitioned into two sets of size N_1 and N_2, $N = N_1 + N_2$. Let P_1 be formed from the N_1 rows and columns of P corresponding to data items in the first group. Similarly, let P_2 be formed from P, using the N_2 data items from the second group.

Suppose a hierarchical clustering method is applied to P. Let U be the resulting phenetic matrix whose entries are called partition ranks and satisfy the ultrametric inequality.[15, 16] Define U_1 and U_2 from U in the same way that P_1 and P_2 were formed from P.

Finally, let the same clustering method that was applied to P be applied to P_1 and P_2. Let W_1 and W_2 be the corresponding phenetic matrices. Two gamma statistics can now be defined. Let γ_1 be the gamma statistic for comparing the partition ranks from U_1 and the partition ranks from W_1 and let γ_2 be that for U_2 and W_2. The larger the values of these statistics, the more stable the hierarchy. We propose using

$$\gamma_m = \min(\gamma_1, \gamma_2) \quad \text{and} \quad \gamma_a = (\gamma_1 + \gamma_2)/2$$

as measures of hierarchical stability. An example is given below.

A rank-order proximity matrix P is shown above the diagonal and the phenetic matrix, U^{CL}, derived from the complete-link method is shown below the diagonal. The (i, j) entry, $i > j$ of U^{CL} is k if data items i and j first belong to the same cluster at level k by the clustering method.

U^{CL} \ P	1	2	3	4	5	6	7	8	9	10
1	\	1	5	9	13	21	39	38	23	36
2	1	\	10	7	14	40	22	33	35	37
3	10	10	\	3	15	41	26	34	28	24
4	10	10	3	\	16	42	25	27	31	45
5	16	10	16	16	\	44	43	30	32	29
6	45	45	45	45	45	\	2	6	11	17
7	45	45	45	45	45	2	\	12	8	18
8	45	45	45	45	45	12	12	\	4	19
9	45	45	45	45	45	12	12	12	\	20
10	45	45	45	45	45	20	20	20	20	\

For purposes of this example, let $N_1 = N_2 = 5$ and choose five odd numbered data items to form the first subset and five even numbered are contained in the second subset.

P_1 =

	3	5	7	9
1	5	13	39	23
3		15	26	28
5			43	32
7				8

P_2 =

	4	6	8	10
2	7	40	33	37
4		42	27	45
6			6	17
8				19

Applying the complete-link method produces the phenetic matrices W_1 and W_2 shown above the diagonal. These matrices are to be compared to the corresponding submatrices of U^{CL}, shown below the diagonal.

U_1^{CL} \ W_1^{CL}	1	3	5	7	9
1	\	5	15	43	43
3	10	\	15	43	43
5	16	16	\	43	43
7	45	45	45	\	8
9	45	45	45	12	\

U_2^{CL} \ W_2^{CL}	2	4	6	8	10
2	\	7	45	45	45
4	10	\	45	45	45
6	45	45	\	6	19
8	45	45	12	\	45
10	45	45	20	20	\

We obtain:

$$\gamma_1^{CL} = 1, \qquad \gamma_2^{CL} = 21/23.$$

Thus, the stability statistics are:

$$\gamma_m^{CL} = 21/23, \quad \gamma_a^{CL} = 22/23.$$

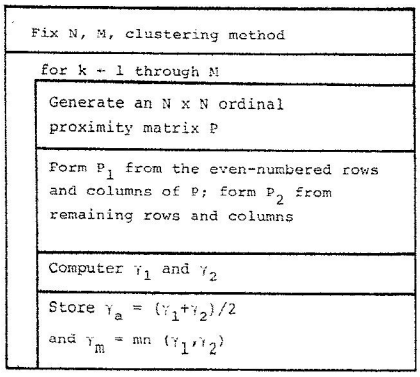

Fig. 2. Monte Carlo procedure for property one.

Similarly, these statistics can be computed using the single-link method which results in:

$$\gamma_1^{SL} = \gamma_2^{SL} = \gamma_m^{SL} = \gamma_a^{SL} = 1.$$

In this example, we obtain high values of our statistics for both the single-link and complete-link methods; γ has a theoretical upper bound of one. Yet we do not know if these high values are unusually high. The distributions of γ_a and γ_m are not known under the conditions of our study. Hubert[11] has empirically studied the distribution of γ between a proximity matrix and its corresponding phenetic matrix for both single-link and complete-link, but these results are not applicable here.

METHOD

If γ_a and γ_m are to measure hierarchical stability, several factors need to be understood. Two properties studied in this paper are listed below.

Property one

The distributions of γ_a and γ_m under a situation involving no clusters must be known to test a 'no-clustering' null hypothesis. We adopt Ling's[12,13] null hypothesis, called the permutation hypothesis;

H_0: All (ordinal) proximity matrices are equally likely.

There are $(N(N-1)/2)!$ ordinal proximity matrices.

The distributions of γ_a and γ_m depend on several items, including the sampled population, the number of data items, the clustering method, and the sizes of the two subsets of data items. We have not been able to derive these distributions so we resorted to Monte Carlo simulations. To generate an $N \times N$ (upper triangular) ordinal proximity matrix, we chose $N(N-1)/2$ uniformly distributed random numbers and replaced the numbers with their rank orders. Several values of N between 16 and 48 and $M = 200$ Monte Carlo trials were used. Hubert[11] indicates that the distribution of γ becomes unreliable for $N < 16$.

The procedure described earlier and given in Fig. 2 was applied to obtain $F_a(x)$ and $F_m(x)$, estimates of the cumulative distribution functions for γ_a and γ_m, respectively, under H_0. Since the entries of P are chosen randomly, the even and odd numbered data items were

Table 1. Mean, variance and Weibull fit to the distribution of the gamma statistics using complete-link clustering: property one

Data items		Mean	Variance	Alpha	Beta	Gamma	Corr.
16							
	Avg:	0.3472	0.0491	0.2776	2.193	−0.1445	0.993
	Min:	0.1738	0.0566	0.3136	2.741	−0.4092	0.985
20							
	Avg:	0.2887	0.0384	0.1947	2.934	−0.2239	0.985
	Min:	0.1603	0.0408	0.1766	2.125	−0.2325	0.992
24							
	Avg:	0.2624	0.0294	0.1243	2.515	−0.1263	0.986
	Min:	0.1569	0.0290	0.1092	2.174	−0.1667	0.973
28							
	Avg:	0.2292	0.0230	0.1514	1.923	−0.5778	0.996
	Min:	0.1286	0.0182	0.0652	2.516	−0.1715	0.971
32							
	Avg:	0.1849	0.0154	0.0619	2.384	−0.0910	0.993
	Min:	0.0978	0.0113	0.0449	2.218	−0.1208	0.993
36							
	Avg:	0.1828	0.0137	0.0627	2.053	−0.0466	0.992
	Min:	0.0981	0.0096	0.0435	2.060	−0.0950	0.996
40							
	Avg:	0.1493	0.0109	0.0503	2.035	−0.0541	0.993
	Min:	0.0770	0.0077	0.0241	2.458	−0.1180	0.985
44							
	Avg:	0.1513	0.0109	0.0445	2.093	−0.490	0.993
	Min:	0.0741	0.0077	0.0218	2.436	−0.1108	0.984
48							
	Avg:	0.1312	0.0079	0.0250	2.406	−0.0604	0.986
	Min:	0.0740	0.0068	0.0233	2.166	−0.0826	0.981

used to form P_1 and P_2. Besides the mean and variance of the empirical distributions for γ_a and γ_m, three-parameter Weibull distributions were fitted to these distributions (Appendix 2) so that the shape of the probability mass could be studied. Also computed is a measure of the goodness-of-fit of the Weibull distribution to the empirical distribution. Values above 0.95 for this fit show that the Weibull distribution is a good summary of the data.[23]

Tables 1 and 2 show the results of this set of Monte-Carlo runs. Table 1 contains the results for the complete-link method, while Table 2 contains the analogous results for the single-link method. We note a few significant facts that occur in these two tables. First, the means and variance of each of the gammas tend to decrease as the number of patterns increases, as expected from the increase in the sample size. Also, for a given number of data items, the statistics computed for the single-link method tend to have larger values than those for the complete-link method. This could possibly be expressed by saying that the single-link method finds more stability in the hierarchy, even when it is not present in the data-set.

Property two

If a hierarchical structure does not match the structure of a proximity matrix to a reasonable degree, conclusions drawn from the hierarchy concerning stability could be biased. To study the effect of global fit, we repeated the simulations for the study of property one, but rejected all proximity matrices P which did not conform to a hierarchical structure. The measure of global fit was γ computed between P and U. The threshold for γ was chosen as the 50th percentile of the null distribution of γ, as determined by Hubert.[11]

Tables 3 and 4 show the results of this set of simulations. Again, the same trends occur as in Tables 1 and 2. Comparing these results with the previous results, it can be seen that values of the statistics tend to be slightly higher than the corresponding values for property one, but as the number of data items increases, it can be seen that global fit is less of a determining factor.

APPLICATION

The stability test is now demonstrated on several 'real' data sets. We illustrate the procedure with data sets from two applications and with a 'random' data set. The application data shows how the test may be used in practice when the number of patterns is the same as one of the Monte Carlo runs in the paper and also when this is not the case. The 'random' data tests the robustness of the test to a different 'no stability' null hypothesis.

The stability test was applied to a data set from a speaker recognition study known to have a good clustering structure.[26] Five samples of speech from each of four speakers were recorded directly into a microphone and, simultaneously, over a telephone line. Thus, a total of 40 speech samples were generated, 20 directly and 20 by telephone. Each set of 20 samples

Table 2. Mean, variance and Weibull fit to the distribution of the gamma statistics using single-link clustering: property one

Data items		Mean	Variance	Alpha	Beta	Gamma	Corr.
16							
	Avg:	0.3991	0.0622	0.5989	3.246	−0.3653	0.994
	Min:	0.2128	0.0764	0.4187	2.312	−0.3947	0.990
20							
	Avg:	0.3866	0.0504	0.8570	4.421	−0.4941	0.994
	Min:	0.2273	0.0641	0.3922	2.533	−0.3851	0.996
24							
	Avg:	0.3797	0.0316	0.4340	4.840	−0.3919	0.996
	Min:	0.2291	0.0436	0.2467	1.901	−0.1914	0.987
28							
	Avg:	0.3500	0.0307	0.5904	5.486	−0.4888	0.987
	Min:	0.2088	0.0403	0.2862	3.198	−0.2960	0.990
32							
	Avg:	0.3651	0.0293	0.1990	3.786	−0.2250	0.996
	Min:	0.2323	0.0402	0.3641	3.775	−0.4590	0.997
36							
	Avg:	0.3324	0.0219	0.1052	3.830	−0.1699	0.984
	Min:	0.2272	0.0279	0.1380	3.021	−0.2342	0.996
40							
	Avg:	0.3016	0.0214	0.0992	3.066	0.1186	0.996
	Min:	0.1924	0.0289	0.1541	2.906	−0.2756	0.994
44							
	Avg:	0.3237	0.0167	0.0639	3.056	−0.0394	0.994
	Min:	0.2212	0.0239	0.1335	3.947	−0.3326	0.993
48							
	Avg:	0.3060	0.0149	0.0556	3.037	−0.0388	0.994
	Min:	0.2005	0.0207	0.1056	3.851	−0.3039	0.992

Table 3. Mean, variance and Weibull fit to the distribution of the gamma statistics using complete-link clustering: property two

Data items		Mean	Variance	Alpha	Beta	Gamma	Corr.
16							
	Avg:	0.3911	0.0504	0.3767	3.183	−0.2675	0.997
	Min:	0.2213	0.0606	0.3305	2.683	−0.3677	0.982
20							
	Avg:	0.3264	0.0367	0.1870	2.184	−0.0833	0.998
	Min:	0.1796	0.0396	0.1886	2.463	−0.2709	0.992
24							
	Avg:	0.2963	0.0320	0.1561	2.160	−0.0774	0.996
	Min:	0.1663	0.0301	0.1268	2.178	−0.1772	0.991
28							
	Avg:	0.2521	0.0235	0.1060	3.155	−0.1873	0.990
	Min:	0.1385	0.0204	0.0773	2.714	−0.2079	0.987
32							
	Avg:	0.2285	0.0199	0.0727	2.531	−0.0869	0.984
	Min:	0.1292	0.0188	0.0661	2.252	−0.1366	0.979
36							
	Avg:	0.1838	0.0139	0.0479	2.484	−0.0776	0.986
	Min:	0.1035	0.0124	0.0393	2.479	−0.1374	0.974
40							
	Avg:	0.1838	0.0151	0.0632	2.090	−0.0536	0.990
	Min:	0.0929	0.0110	0.0417	2.064	−0.0980	0.982
44*							
	Avg:	0.1634	0.0107	0.0342	2.463	−0.0615	0.960
	Min:	0.0870	0.0079	0.0374	1.959	−0.0731	0.988
48*							
	Avg:	0.1351	0.0065	0.0245	2.291	−0.396	0.972
	Min:	0.0613	0.0055	0.0374	0.1959	−0.0781	0.988

* Number of Monte Carlo runs = 50.

Table 4. Mean, variance and Weibull fit to the distribution of the gamma statistics using single-link clustering: property two

Data items		Mean	Variance	Alpha	Beta	Gamma	Corr.
16							
	Avg:	0.4739	0.0578	1.544	4.747	−0.5295	0.990
	Min:	0.2852	0.0904	0.8350	3.030	−0.5561	0.996
20							
	Avg:	0.4428	0.0450	0.8109	4.627	−0.4310	0.995
	Min:	0.2795	0.0703	0.6171	3.117	−0.4862	0.993
24							
	Avg:	0.4118	0.0380	0.2455	3.070	−0.1533	0.996
	Min:	0.2595	0.0548	0.3722	2.694	−0.3598	0.990
28							
	Avg:	0.3898	0.0263	0.9152	6.518	−0.5298	0.981
	Min:	0.2491	0.0398	0.2624	2.935	−0.3203	0.995
32							
	Avg:	0.3622	0.0282	0.1803	3.764	−0.2107	0.991
	Min:	0.2313	0.0405	0.3456	3.623	−0.4407	0.994
36							
	Avg:	0.3578	0.0262	0.2891	4.998	−0.3586	0.990
	Min:	0.2461	0.0322	0.2973	4.214	−0.4401	0.994
40							
	Avg:	0.3489	0.0188	0.1019	4.545	−0.2037	0.993
	Min:	0.2289	0.0260	0.1358	2.953	−0.2241	0.993
44*							
	Avg:	0.3093	0.0176	0.1017	4.295	−0.2251	0.976
	Min:	0.2222	0.0229	0.2610	5.018	−0.4807	0.968
48*							
	Avg:	0.3348	0.0154	0.0872	4.555	−0.1999	0.990
	Min:	0.2056	0.0207	0.4653	5.840	−0.3201	0.970

* Number of Monte Carlo runs = 50.

can be subdivided into 4 groups by speaker. Using a procedure described elsewhere,[22] a 40 × 40 dissimilarity matrix was obtained. A multidimensional scaling solution indicated that the original 40 patterns could be represented in a 4-dimensional space. An eigenvalue projection in 2 dimensions of this 4-dimensional data is shown in Fig. 3, and clearly indicates 2 clusters, shown with a dotted line. It also indicates various subclusters of these so this data set should provide a good test for a measure of hierarchical stability.

Tables 5 and 6 show the γ_a and γ_m values for complete-link and single-link hierarchies, respectively, along with the global γ statistic proposed by Hubert.[11] Ten separate values of the two gammas are shown, each resulting from a different random division of the 40 × 40 dissimilarity matrix into two 20-node portions. Using the approximation from[11] for the distribution of γ, we have: $P[\gamma^{CL} < 0.230] = 0.995$ and $P[\gamma < 0.166] = 0.995$. Thus the hierarchies produced by both single-link and complete-link methods are unusually well structured when compared to hierarchies from randomly chosen proximity matrices.

Using the Weibull parameters fit to the distributions for γ_a and γ_m for 40 data items, we obtain $P[\gamma_a^{CL} < 0.433] = 0.995$ and $P[\gamma_m^{CL} < 0.291] = 0.995$. In the list of ten numbers for both of these test statistics, none fall below the 99.5 percentile (0.433 and 0.291) of their empirical distributions. This is evidence that the hierarchy is stable and the complete-link dendrogram summarizes the data well.

The Weibull distributions for the single-link method show that $P[\gamma_a^{SL} < 0.656] = 0.995$ and $P[\gamma_m^{SL} < 0.613] = 0.995$. Again, no values fall below these 99.5 percentiles of their empirical distributions. However, we can make the intuitive judgement that the single-link hierarchy appears more stable than that for complete-link due to the consistently high values for γ_a^{SL}. Note however that the percentiles for the single-link method are larger than the corresponding values for the complete-link method. This gives weak evidence that the single-link hierarchy is a more stable representation of the data than the one produced by the complete-link method.

Although the global (γ) gamma value for single-link is below that for complete-link, this conclusion appears to be justified for this data set when pattern class information is considered. The single-link hierarchy separates the patterns into two clusters based on direct recording or telephone recording, while the complete-link hierarchy's clustering with two clusters mixes patterns from the two modes of recording. Also, the internal structure of the two main clusters in the single-link hierarchy consists of sub-clusters representing patterns from the same speaker, with one exception. Several individual speakers occur in different clusters in the complete-link hierarchy.

A more conservative test can now be applied to both

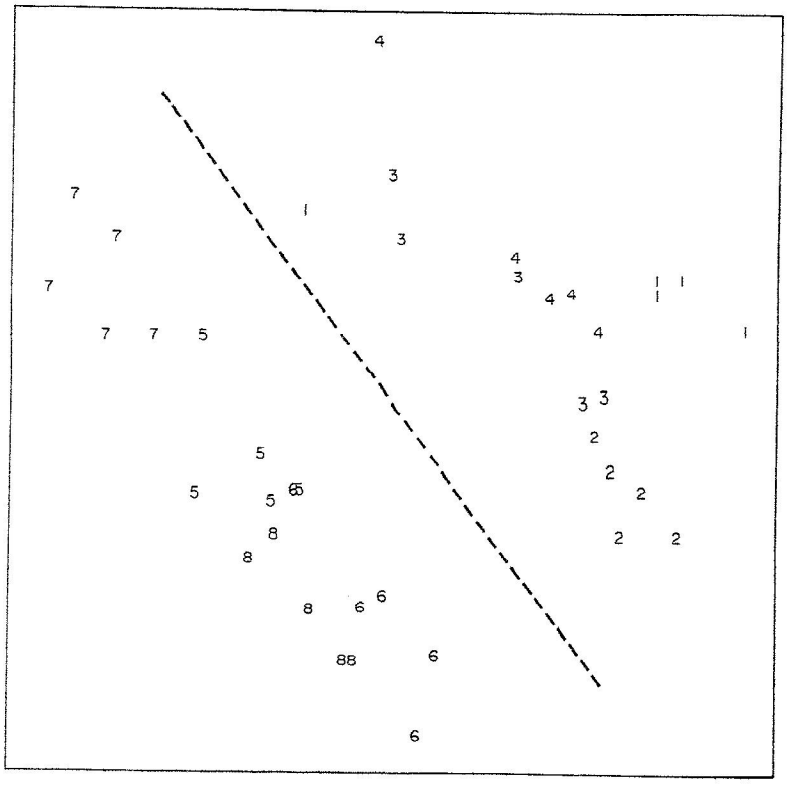

Fig. 3. Eigenvalue projection of the speech data.

Table 5. Complete-link gammas for speech data

Global gamma: 0.75553		
Partition	Gamma average	Gamma minimum
1	0.6801	0.6404
2	0.7879	0.6406
3	0.8636	0.7701
4	0.7990	0.6178
5	0.9567	0.9348
6	0.8445	0.7066
7	0.9732	0.9499
8	0.8420	0.6924
9	0.8842	0.7921
10	0.9075	0.8545

Table 6. Single-link gammas for speech data

Global gamma: 0.73435		
Partition	Gamma average	Gamma minimum
1	0.8158	0.8027
2	0.9311	0.9217
3	0.9248	0.8677
4	0.9421	0.9129
5	0.9522	0.9194
6	0.9129	0.9110
7	0.8385	0.7348
8	0.8319	0.7013
9	0.8360	0.7162
10	0.9431	0.9381

Table 7. Complete-link gammas for 80X data

Global gamma: 0.60263		
Partition	Gamma average	Gamma minimum
1	0.5540	0.2888
2	0.6245	0.4199
3	0.4817	0.3745
4	0.7857	0.7810
5	0.6199	0.5786
6	0.6848	0.6434
7	0.4386	0.4181
8	0.6260	0.5644
9	0.6205	0.5531
10	0.6335	0.6211

Table 8. Single-link gammas for 80X data

Global gamma: 0.63842		
Partition	Gamma average	Gamma minimum
1	0.8686	0.8310
2	0.8533	0.7793
3	0.6894	0.4821
4	0.8735	0.8567
5	0.8688	0.8014
6	0.8875	0.8001
7	0.6944	0.5203
8	0.8122	0.7307
9	0.8325	0.7612
10	0.6918	0.5676

hierarchies by using the distributions found for property two. This again yields no values below the 99.5 percentiles and supports the conclusion that the hierarchies are stable.

The next application involves the 80X data set, in which each pattern represents a character from the Munson handprinted FORTRAN character set. A pattern consists of eight features measured on one of the alpha-numeric characters '8' '0' and 'X'. A total of 15 patterns is compiled from each of these character classes. Thus we have a 45 pattern data set which produces a 45 × 45 dissimilarity matrix based on Euclidian distance. We apply the stability test by randomly splitting this into two proximity matrices, one with 22 patterns, the other with 23 patterns and use the empirical distributions found for 44 patterns to define the percentiles.

Tables 7 and 8 show the results of 10 random division of the 80X data set under complete-link and single-link, respectively. As before, we compute the 99.5 percentiles for the distributions of the global γs using Hubert's approximations. Both global gamma values are greater than this percentile. Next, we compute the 99.5 percentiles of the empirical distributions of the stability statistics using the Weibull fits. Under property 1, no values of γ_a^{CL} or γ_m^{CL} fall below these percentiles for complete-link while two values are below the percentile for γ_m^{SL}. Under the more conservative property two distributions, one value of γ_m^{CL} and one more value of γ_m^{SL} are below the percentiles. These results indicate that neither hierarchy is a stable description of the data, although if one were forced to choose between the two hierarchies, the complete-link hierarchy would be the choice. These results tally with pattern class information. The single-link hierarchy suffers from chaining, and clusters containing patterns from only one class are rare. The complete-link method finds reasonable clusters for most '0's and '8's, but mixes 'X's with some '8's.

The last data set investigated in this paper is an artificial data set generated under hypothesis of 'no stability' other than that used previously. We generated 50 points at random in a unit sphere. This is the 'Random Position' null hypothesis mentioned by Dubes and Jain[8] and it has been shown that this null hypothesis produces different distributions for statistics for cluster validity from the previous 'Random Graph' null hypothesis.[26] Given that we want 'real' structure in the hierarchy representing any data, there should be no stable hierarchy representing random points.

Tables 9 and 10 show the results of applying the stability procedure 10 times. We now approximate the distribution of the stability statistics for 50 patterns by those found empirically for 48 patterns. At the 0.005 level, both values of the global gamma are significant using Hubert's approximation. However, applying the percentiles for the stability statistics at the 0.005 level using property one, we find one value of γ_a^{CL} and two values of γ_m^{CL} are below their percentiles, while eight

Tape 9. Complete-link gammas for sphere data

Global gamma: 0.52870		
Partition	Gamma average	Gamma minimum
1	0.6754	0.5588
2	0.3438	0.2632
3	0.6013	0.4340
4	0.5011	0.4396
5	0.5261	0.2837
6	0.4971	0.4512
7	0.5375	0.4846
8	0.6722	0.5858
9	0.5880	0.4733
10	0.5587	0.3630

Table 10. Single-link gammas for sphere data

Global gamma: 0.42618		
Partition	Gamma average	Gamma minimum
1	0.4733	0.2229
2	0.6715	0.5964
3	0.5568	0.4204
4	0.5877	0.5632
5	0.3874	0.2848
6	0.5827	0.5400
7	0.4149	0.3438
8	0.6108	0.4500
9	0.7258	0.7153
10	0.5497	0.4656

values of γ_a^{SL} and six values of γ_m^{SL} are below their percentiles. We conclude that the hierarchies produced by the two clustering methods are not stable, as would be expected from random data. This gives evidence that the stability statistics give more information about the data set than can be obtained from the global gamma alone.

CONCLUSIONS

In this paper we have presented a method for testing the stability of the hierarchical clustering structure of a data-set. It is assumed that a stable structure indicates the hierarchy provides a good summary of the data. The procedure involves randomly halving the data-set, reclustering each half, and computing two measures of stability, γ_a and γ_m. Large values of both of these measures show that little change has taken place between the structure of the original hierarchy and that of each half, and this leads us to conclude that the hierarchy is stable.

To apply this procedure to a test data-set, the practitioner must know what constitute large values of γ_a and γ_m. This paper studies this question by using a Monte Carlo technique to assess the values of γ_a and γ_m for random proximity matrices. This study was done for both the single-link and complete-link methods and for various numbers of patterns. Finally, the paper concludes with an example of using these measures in practice. It is shown that the stability statistics provide more information on the global fit *a* hierarchy than does Hubert's γ.

Additional work needs to be done in the general area of stability as a technique for cluster validity studies and on properties of the two measures γ_a and γ_m. The biggest weakness in using the test in practice is for data-sets with more than fifty patterns, since no distributional information is known for this size data set. We are currently working on fitting a possible distribution to these cases, given the information available from this study. Next, if possible, the power of the test defined in this paper should be studied. Also, a study of the distributions of the two gamma values determined from a single matrix divided a number of times should be undertaken. Possibly a different approach to partitioning should be studied, such as a leave-one-out approach. This would bypass the problem of obtaining various gamma values depending upon the specific partition of the test data-set.

SUMMARY

Stability is an intrinsic characteristic of individual clusters, clusterings and hierarchical structures. We have defined a measure of stability for a hierarchical structure, based on Strauss'[1,2] suggestion, and studied its distribution under several circumstances. These distributions permit a quantitative evaluation of the stability of a hierarchical clustering under certain conditions.

The stability statistic measures the agreement between two phenetic matrices, one derived from a clustering imposed on half the data and the other obtained from clustering the entire data set. Randomly halving the data produces two rank correlation coefficients, γ_1 and γ_2. The average, γ_a, and the minimum, γ_m, of these coefficients are the stability statistics. Our estimates of the distributions of γ_a and γ_m depend on the clustering methods, the null hypothesis, the manner in which the data are split, and the sample size. The more stable the hierarchy, the better the hierarchical structure fits the data.

The hierarchical structure imposed by the single-link and complete-link clustering methods depends only on the order of the entries in the $N \times N$ proximity matrix. We limited our study to these two clustering methods and used the Random Graph Hypothesis, which states that all $[N(N-1)/2]!$ proximity matrices are equally likely, as the null hypothesis. We also restricted the proximity matrices to those having a reasonable hierarchical structure to ascertain the effect of global fit. Sample sizes (N) from 16 to 48 in steps of 4 were used. The means and variances of γ_a and γ_m were tabulated along with Weibull parameters describing the distributions of these statistics. The means and variances decreased, as expected, and global fit became less of a factor as N increased.

The stability statistics were computed for ten dichotomies of each of three data sets, which verified the usefulness of studying stability. Several areas for further investigation were also noted.

REFERENCES

1. J. S. Strauss, Classification by cluster analysis, Chap. 12 of *Report of the International Pilot Study of Schizophrenia*, Vol. 1, pp. 336–359. World Health Organization, Geneva (1973).
2. J. S. Strauss, J. J. Bartko and W. J. Carpenter, Jr., The use of clustering techniques for the classification of psychiatric patients, *Br. J. Psychiat.* **122,** 531–540 (1973).
3. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York (1973).
4. B. Everitt, *Cluster Analysis*, John Wiley, New York (1974).
5. J. A. Hartigan, *Clustering Algorithms*, John Wiley, New York (1975).
6. R. M. Cormack, A review of classification, *Jl R. statist. Soc.* **A134**(3), 321–367 (1971).
7. R. Dubes and A. K. Jain, Clustering techniques: the user's dilemma, *Pattern Recognition* **8,** 247–260 (1976).
8. R. Dubes and A. K. Jain, Validity studies in clustering methodologies, *Pattern Recognition* **11,** 235–254 (1979).
9. F. B. Baker and L. J. Hubert, A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering, *J. Am. statist. Ass.* **71,** 870–878 (1976).
10. S. Fillenbaum and A. Rapoport, *Structures in the Subjective Lexicon*, Academic Press, New York (1971).
11. L. Hubert, Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures, *J. Am. statist. Ass.* **69,** 698–704 (1974).
12. R. F. Ling, On the theory and construction of k-clusters, *Comput. J.* **15**(4), 326–332 (1972).
13. R. F. Ling, Probability theory of cluster analysis, *J. Am. statist. Ass.* **68,** 159–164 (1973).
14. R. F. Ling and G. S. Killough, Probability tables for cluster analysis based on a theory of random graphs, *J. Am. statist. Ass.* **71,** 293–300 (1976).
15. L. J. Hubert, Some applications of graph theory to clustering, *Psychometrika* **39,** 283–309 (1974).
16. S. C. Johnson, Hierarchical clustering schemes, *Psychometrika* **32,** 241–254 (1967).
17. N. Jardine and R. Sibson, *Mathematical Taxonomy*, John Wiley, New York (1971).
18. L. A. Goodman and W. H. Kruskal, Measures of association for cross-classifications, *J. Am. statist. Ass.* **49,** 732–764 (1954).
19. J. G. Mezzich, Evaluating clustering methods for psychiatric diagnosis, *Biol. Psychiat.* **13,** 265–281 (1978).
20. W. M. Rand, Objective criteria for evaluating clustering methods, *J. Am. statist. Ass.* **66,** 846–850 (1971).
21. F. B. Baker, Sensitivity of the complete-link clustering technique to missing individuals, *J. educ. Statist.* **3,** 233–252 (1978).
22. O. I. Tosi, The problem of speaker identification and elimination, In *Measurement Procedures in Speech, Hearing and Language*, Ed. S. Singh, University Park Press, Baltimore (1975).
23. R. C. Dubes, Data reduction with grouping and Weibull models, Interim Report No. 7, Contract No. AFOSR-1023-67B, Division of Engineering Research, Michigan State University (1970).
24. R. Gnanadesikan, J. R. Kettenring and J. M. Landwehr, Interpreting and accessing the results of cluster analyses, *Bull. int. statist. Inst.* **47,** 451–463 (1977).
25. R. K. Blashfield and M. S. Aldendorfer, Cluster analysis literature on validation. Paper presented at the Classification Society Meeting (1978).
26. T. A. Bailey, Cluster validity and intrinsic dimensionality, Ph.D. Dissertation, Michigan State University (1979).

APPENDIX 1

This appendix defines the basic Goodman–Krustal gamma statistic,[18] as used in this paper. In general terms, the gamma statistic, γ, measures the rank correlation between proximity ranks and partition ranks. That is, it measures the consistency of the rankings. To define γ, given a rank-order proximity matrix P and phonetic matrix U corresponding to some clustering method, we first realize that the entries in P are integers from 1 to $N(N-1)/2$ with no ties. The entries of U are a subset of these integers. A list of all n sets of two ranks is then formed. Each set of two ranks contains one rank from P and one from U, taken from corresponding positions. The gamma statistic is defined as:

$$\gamma = [S(+) - S(-)/S(+) + S(-)]$$

where $S(+)$ is the number of 'condordant' pairs, or consistent pairs, while $S(-)$ is the number of 'discordant', or inconsistent, pairs. A 'pair' consists of two items from the list formed above. For example, the pairs (3, 8) and (5, 12) are concordant because $3 < 5$ and $8 < 12$. However, (3, 8) and (6, 7) are discordant because $3 < 6$ but $8 > 7$. Ties are not counted so the pair (3, 5), (3, 7) is neither discordant nor concordant. Writing

$$\gamma = \{S(+)[S(+) + S(-)]\} - \{S(-)[S(+) + S(-)]\}$$

leads to interpreting γ as the probability of a consistent ranking minus the probability of an inconsistent ranking.

The number of pairs that must be examined to compute γ is:

$$n(n-1)/2 = (N+1)N(N-1)(N-2)/8.$$

APPENDIX 2

The Weibull distribution function as used in this paper is defined as:

$$F(a) = 1 - \exp[(-1/\alpha)(a-\rho)^{\beta}] \quad \text{if } a > \rho$$
$$= 0 \quad \text{if } a < \rho.$$

If W is a Weibull random variate, then

$$P[W < a] = 0.99 = 1.0 - 0.01$$

implies that

$$a = [-\alpha \ln(0.01)]^{1/\beta} + \rho.$$

The parameters are named as follows: α is the scale parameter, β is the shape parameter, and ρ is the location parameter.

About the Author – RICHARD C. DUBES was born in Chicago, IL, and received the B.S. degree from the University of Illinois, Urbana, in 1956, and the M.S. and Ph.D. degrees from Michigan State University, East Lansing, in 1959 and 1962, respectively, all in electrical engineering.

In 1956 and 1957 he was a member of the Technical Staff of the Hughes Aircraft Company, Culver City, CA. From 1957 through 1968 he served as Graduate Assistant, Research Assistant, Assistant Professor, and Associate Professor in the Electrical Engineering Department at Michigan State University. In 1969 he joined the Computer Science Department at Michigan State University and became Professor in 1970. His areas of technical interest include pattern recognition, clustering, decision theory, and application of data analysis methods to the medical area.

Dr. Dubes is a member of the Pattern Recognition Society, IEEE, and Sigma Xi.

About the Author – STEPHEN P. SMITH was born in Cincinnati, Ohio on September 4, 1955. He received the B.S. and M.S. degrees in Computer Science from Michigan State University, East Lansing in 1977 and 1979, respectively.

Since 1977 he has been a Graduate Research Assistant in the Pattern Recognition and Image Processing Laboratory in the Computer Science Department at Michigan State University. During the summer of 1979, he was a consultant at Babcock and Wilcox, Inc., working on problems of feature selection and image processing in a non-destructive testing environment. Currently, he is working on a Ph.D. degree at Michigan State University. His current research interests include shape matching and cluster validity. Mr. Smith is a member of the ACM, IEEE and Phi Kappa Phi.