DOCUMENT RESUME

ED 076 685 TM 002 693

AUTHOR	Hubert, Lawrence
TITLE	Approximate Evaluation Techniques for the Max Hierarchical Clustering Procedure.
PUB DATE	73
NOTE	22p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Cluster Analysis; *Evaluation Techniques; Goodness of Fit; *Statistical Analysis; Technical Reports

ABSTRACT

A technique for testing the hypothesis that a hierarchical sequence of partitions constructed by the max method could have been obtained solely on the basis of "noise" is discussed. The evaluation procedure imvolves comparing a rank-order goodness-of-fit measure to the tabled percentiles obtained from an approximate cumulative permutation distribution of the measure. One of the rank orderings of the object pairs is derived from the originai similarity values between the objects to be partitioned; the second rank ordering of the object pairs is obtained from the partition hierarchy itself. (Authoɪ/KM)

APPROXIMATE EVALUATION TECHNIQUES FOR THE MAX HIERARCHICAL CLUSTERING PROCEDURE

Lawrence Hubert*

Abstract

This paper presents a simple technique for testing the hypothesis that a hierarchical sequence of partitions constructed by the max method could have been obtained solely on the basis of "noise." The test procedure involves comparing a rank-order goodness-of-fit measure (Goodman-Kruskal γ statistic) to the tabled percentiles obtained from an approximate cumulative permutation distribution of the measure. One of the rank orderings of the object pairs used in defining γ is derived immediately from the given similarity values between the objects to be partitioned; the second rank ordering of the object pairs is obtained from the partition hierarchy itself. The tested hypothesis is simply that the given set of similarity values have been assigned randomly to the object pairs.

APPROXIMATE EVALUATION TECHNIQUES FOR THE MAX HIERARCHICAL CLUSTERING PROCEDURE

1. INTRODUCTION

In recent years a substantial number of applied researchers have attempted to use the max hierarchical clustering scheme as a general data analysis technique. Representative applications may be found in Miller [11], Anglin [1], Sokal and Sneath [13], Johnson [8], and Hubert [5,6]. Variously called the max [8], complete-link [13], furchest-neighbor [9], or hierarchical linkage technique [10], this particular clustering procedure has been used primarily as a descriptive device since there is no standard way of statistically evaluating the adequacy of the obtained sequence of partitions. Although the lack of an elegant methodology is understandable given the combinatorial problems posed by the method, approximate statistical procedures can be developed now in terms of randomization and sampling theory until the more exact assessment methods become available in the future.

As a way of presenting a brief summary of what the max clustering method does, suppose S is a set of n objects labeled $o_1,\ldots,o_n$ and $\{s_{ij}\}$ is an n by n matrix containing similarity measures between all objects o_i and o_j.[1] For a rather weak initial requirement, it is assumed that the elements of $\{s_{ij}\}$ satisfy three constraints:

(i) Symmetry: $s_{ij} = s_{ji}$ for all $o_i,o_j \varepsilon S$;

(ii) Positivity: $s_{ij} \geq 0$ for all $o_i,o_j \varepsilon S$;

(iii) Nullity: $s_{ij} = 0$ for all $o_i = o_j$.

In the discussion below, similarity will be treated as a primitive term; the interested reader can consult Sokal and Sneath [13], or Jardine and Sibson [7] for a more thorough presentation.

The general aim of most hierarchical clustering techniques is to produce an "optimal" sequence of partitions of the basic object set S. More precisely, the commonly used clustering methods produce a sequence of partitions $(\ell_0,\ldots,\ell_{n-1})$ with the following properties:

(i) ℓ_0 is the trivial partition containing an object class for each element in S;

(ii) ℓ_{n-1} is the trivial partition containing a single all-inclusive object class;

(iii) ℓ_k is an immediate refinement of ℓ_{k+1}, $0 \leq k \leq n-2$.

It is possible to characterize inductively one general paradigm for the construction of a partition sequence that will include most of the familiar clustering methods. Suppose the level k partition, ℓ_k, has been obtained and some real-valued function f defined on the Cartesian product of the power set P(S) is evaluated for all pairs of subsets defining ℓ_k. That pair of subsets at level k minimizing (or in some other way optimizing) the function f are then united to form a new object class in the partition ℓ_{k+1}. All remaining subsets in ℓ_k are merely transferred to ℓ_{k+1}.

As an illustration, the max method and an alternative min method [8] are obtained by the following two interpretations of f:

if $L_r, L_{r'} \varepsilon P(S)$, then

$f_{max}(L_r, L_{r'}) = \max \{s_{ij} | o_i \varepsilon L_r, o_j \varepsilon L_{r'}\}$;

$f_{min}(L_r{:}L_{r'}) = \min \{s_{ij} | o_i \varepsilon L_r, o_j \varepsilon L_{r'}\}$.

The max method uses f_{max} and attempts to minimize subset diameters; the min method uses f_{min} and minimizes a standard topological measure of similarity between subsets.

Since both the max and the min techniques depend only upon the rank order of the similarity values, either of these two procedures can be interpreted as a way of reranking the object pairs. Specifically, the partition rank for each object pair $\{o_i, o_j\}$ is defined as the level at which that pair first belongs to the same subset in a partition. Symbolically, the partition rank for the pair $\{o_i, o_j\}$ can be expressed as

$$\min \{k | \{o_i, o_j\} \text{ belongs to the same subset in } \ell_k\}.$$

By comparing the set of all partition ranks to the original similarity ranks, the adequacy of the partition hierarchy in capturing the structure underlying the matrix $\{s_{ij}\}$ can be assessed. The measure used in the following sections for quantifying this agreement is the γ statistic developed by Goodman and Kruskal [4]; although this choice is somewhat arbitrary, the γ statistic has a number of desirable properties with regard to probabilistic interpretation in the case of tied ranks that the more standard measures of rank correlation do not possess.

2. COMPARISON OF THE MIN AND THE MAX METHODS

Jardine and Sibson [7] provide a very strong axiomatic argument for the use of the single-link (min) as opposed to the complete-link (max) clustering procedure. Although their presentation is mathematically elegant, a number of other researchers in the field, notably the "Australian school" (see, for example [14]), have criticized the min method on pragmatic grounds. As a way of introducing a more extensive discussion of the max method per se, this section will present one simple illustration to point out the differences between the min and the max technique in terms of the γ statistic.

In the example given in a later section an object set with cardinality 9 was defined with 30 distinct similarity values assigned to 36 object pairs. Since both the min and the max procedures depend solely upon the rank order of the similarity values, this is equivalent to assigning 30 distinct ranks to the 36 object pairs. Using this fixed set of ranks, 1000 permutations were randomly selected with replacement from the set of all possible permutations of the object pairs. For each permutation, the min and the max hierarchies were obtained along with the two corresponding γ values.

It is obvious from the cumulative distribution of γ given in Table 1 that, on the average, the max procedure provides the more adequate representation of the original similarity values. This result holds true in general and is not an artifact of the cardinality of the object set used in this example.

Table 1 here

The accuracy of the distribution obtained with a sample of 1000 can be evaluated in a number of ways. First, by using tolerance intervals we can say that with probability greater than .999, 99 percent of the complete permutation distribution is less than the maximum observation (see Table 5 in [2]). Thus, if a γ value greater than .82 were obtained for a max hierarchy based upon similarity values of the form used in constructing Table 1, there would be little doubt as to the significance of the result. In particular, if the null hypothesis is one of randomness in the assignment of similarity values to the object pairs, then each permutation of the object pairs should be equally likely to occur a priori. Consequently, if a γ value larger than .82 were calculated for the actual data using the max hierarchy, this particular null hypothesis could be rejected at a significance level close to .01.

A second way of assessing the accuracy of the sampling procedure is in terms of Kolmogorov-Smirnov theory. Using a sample size of 1000 the following statement can be made conservatively since the underlying distribution of γ is discrete: with probability at least .99, the maximum absolute deviation between the sample and the population cumulative distribution function is less than .05 (see [3], p. 81). Thus, if the γ value obtained for the real data lies at the $1 - \alpha$ percentage point of the permutation distribution, the null hypothesis of randomness can be rejected conservatively at a significance level of about $\alpha + .05$.

Obviously, these measures of accuracy for the sampling procedure could be improved upon further; for practical purposes, however, a sample size of 1000 was used in deriving the tables given in the next section for n = 4 through 16. For larger values of n, a normal approximation based upon an estimated mean and variance of the permutation distribution appears to be adequate.

3. TABLES FOR ASSESSING THE RESULTS FROM THE MAX CLUSTERING PROCEDURE

Assuming that all similarity values are untied, Table 2 presents selected percentage points of the sample cumulative distribution functions for γ for n = 4 through 16. As mentioned in the previous section, these distributions are based upon a sample size of 1000 and should provide fairly reasonable approximations to the population distribution functions.

Table 2 here

For small values of n the sample permutation distributions are extremely peaked although they are almost perfectly symmetric. The mean values and the variances decrease fairly regularly as n increases; in fact, by merely extrapolating from the means and variances presented in Table 3, reasonable approximations to object sets larger than 16 could be obtained. Instead of extrapolating, however, the estimated means and variances for n = 17 through 25 were obtained with samples of 200 permutations and may be used as parameters of an approximating normal distribution.

Table 3 here

For moderate n, the normal distribution provides a fairly adequate approximation to the underlying sample permutation distribution. For example, Table 4 illustrates the close correspondence with the normal when n = 14. If n is small, however, the sample permutation distribution is considerably more peaked than the corresponding normal distribution.

Table 4 here

4. EXAMPLE

There is a basic problem with the use of Table 2 when tied similarity values are present, since in a strict sense the tabled percentage points are then no longer appropriate. Although a complete discussion of the effect of ties would be valuable, ~~analytically the~~ task seems impossible. We can, however, present an example of what happens to the permutation distribution when ties do occur.

As a way of discussing the problem of ties and presenting an illustration of the use of Table 2, the data collected by Shepard [12] on the confusability of nine colors is ideal. The basic nine by nine similarity matrix in Shepard's paper consisted of the conditional probabilities of confusing one colored circle with eight other possibilities, each of which had the same constant red hue but different values for brightness and saturation. To make the similarity measures symmetric, the values given in Table 5 were obtained by adding the symmetric elements from Shepard's table and subtracting the result from 1.00.

Tables 5 and 6 here

In addition to the similarity values between colors, Table 5 also lists the partition ranks for the object pairs obtained from the max partition hierarchy given in Table 6. The γ value between the partition ranks and the ordered similarity values turned out to be .687 and apparently represents a substantial value. To test whether this γ index is large enough to reject the null hypothesis of a random ordering, 1000 permutations of the object pairs were obtained using the similarity values illustrated in Table 5. The percentage points for this cumulative distribution were given previously in Table 1 and imply that the null hypothesis can be rejected at a significance level of about .01, subject to the variat [illegible] duced by the sampling process itself.

Although obtaining a separate permutation distribution for each similarity matrix is the most ideal alternative, this recommendation defeats the overall usefulness of the percentage points given in Table 2. Most of the time, however, the tabled values will be sufficient to convince the researcher that he is not obtaining a hierarchy based upon noise alone. This can be done merely by breaking the ties in the original similarity values to obtain the largest γ value and then a second time to obtain the minimum value. These bounds are on the γ values that can be obtained from the partition hierarchy assuming the similarity values are untied; but in addition, because of the way in which γ is defined it is also true that the original γ calculated for tied similarity values will lie between these two bounds.

The upper and lower bounds on γ are rather easy to obtain without a complete evaluation of all the possible ways in which ties may be resolved. Each individual set of tied similarity values can be reordered to give a minimum γ value with respect to its own subset of partition ranks and then a second time to give a maximum γ value. When used together, these local operations construct overall orderings of the similarity values that lead to the global upper and lower bounds for the γ index.

Table 5 presents the orderings of the similarity values based upon these two local operations. The minimum γ is .663 and the maximum γ is .688. Since the minimum bound is at a high percentage point in Table 2, the null hypothesis still appears untenable.

In general, if the maximum bound is not sufficient to reject randomness at a reasonable significance level, then a permutation distribution based upon the unique form of the similarity values will lead to the same conclusion (subject, of course, to the sampling variability in the permutation distributions). Similarly, if the minimum bound is sufficient to reject randomness, then the more exact permutation distribution will imply rejection also. However, if either of these two conditions does not occur, using the more exact permutation distribution based upon the exact similarity values is probably the only reasonable procedure to follow.

Table 1. SAMPLE CUMULATIVE PERMUTATION DISTRIBUTIONS OF γ FOR NINE OBJECTS [N = 1000, 30 DISTINCT SIMILARITY VALUES]

γ	Cumulative proportions	
	Min[a]	Max[b]
.02	.002	.000
.06	.006	.000
.10	.016	.000
.14	.032	.000
.18	.065	.003
.22	.122	.006
.26	.202	.021
.30	.314	.058
.34	.429	.120
.38	.570	.223
.42	.706	.350
.46	.801	.485
.50	.862	.654
.54	.919	.784
.58	.956	.882
.62	.985	.942
.66	.996	.980
.70	.998	.990
.74	1.000	.997
.78	1.000	.999
.80	1.000	.999
.82	1.000	1,000

[a] Mean of .37; standard deviation of .12 .

[b] Mean of .47; standard deviation of .10 .

Table 2. PERCENTAGE POINTS FOR THE SAMPLE CUMULATIVE PERMUTATION DISTRIBUTIONS OF γ USING THE MAX METHOD

n	Cumulative proportions[a]							
	Min γ	Max γ	.500	.700	.800	.900	.950	.990
4	.557	1.000	.818(.612)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)
5	.241	1.000	.724(.589)	.862(.757)	.931(.839)	.943(.905)	1.000(1.000)	1.000(1.000)
6	.211	1.000	.612(.516)	.741(.700)	.803(.803)	.837(.902)	.873(.950)	.916(.990)
7	.181	.956	.557(.514)	.652(.701)	.712(.801)	.745(.900)	.784(.952)	.819(.990)
8	.175	.792	.500(.500)	.597(.704)	.648(.803)	.686(.900)	.719(.951)	.744(.990)
9	.193	.782	.451(.500)	.536(.703)	.588(.800)	.624(.900)	.651(.950)	.708(.990)
10	.163	.738	.432(.507)	.505(.702)	.548(.802)	.579(.901)	.615(.951)	.669(.991)
11	.154	.665	.401(.501)	.468(.700)	.506(.801)	.534(.900)	.560(.950)	.605(.990)
12	.157	.677	.378(.505)	.438(.704)	.472(.800)	.496(.904)	.515(.951)	.550(.990)
13	.043	.556	.352(.504)	.414(.702)	.443(.800)	.472(.902)	.501(.950)	.525(.990)
14	.120	.575	.345(.502)	.395(.704)	.422(.802)	.443(.901)	.463(.950)	.488(.990)
15	.140	.553	.322(.505)	.359(.705)	.378(.802)	.405(.900)	.429(.951)	.467(.990)
16	.119	.512	.309(.503)	.354(.705)	.381(.803)	.406(.901)	.428(.950)	.450(.990)

[a] In parentheses next to each γ value are the maximum proportions that would be appropriate.

Table 3. RELATIONSHIPS BETWEEN THE NUMBER OF OBJECTS IN S AND THE SAMPLE MEAN AND STANDARD DEVIATION OF γ

n[a]	Mean γ	Standard Deviation γ
4	.818	.1665
5	.706	.1606
6	.613	.1407
7	.553	.1222
8	.502	.1119
9	.457	.0990
10	.432	.0908
11	.400	.0828
12	.376	.0749
13	.353	.0724
14	.343	.0634
15	.324	.0635
16	.307	.0601
17	.299	.0554
18	.288	.0563
19	.277	.0530
20	.265	.0481
21	.254	.0457
22	.245	.0402
23	.235	.0461
24	.230	.0426
25	.229	.0387

[a] Mean and standard deviation based upon samples of 1000 through n = 16; for larger n, sample sizes are 200.

Table 4. NORMAL APPROXIMATION TO THE (SAMPLE) PERMUTATION DISTRIBUTION OF γ [n = 14]

Sample γ	Sample percentage	Standardized γ [a]
.195	.010	.195
.209	.020	.213
.218	.030	.224
.227	.040	.232
.236	.050	.239
.245	.060	.244
.249	.070	.249
.255	.080	.254
.259	.090	.259
.262	.100	.260
.289	.200	.289
.311	.300	.310
.326	.400	.327
.345	.500	.343
.361	.600	.359
.377	.700	.376
.395	.800	.396
.422	.900	.426
.425	.910	.428
.430	.920	.432
.433	.930	.436
.440	.940	.441
.444	.950	.447
.451	.960	.454
.460	.970	.462
.471	.980	.473
.488	.990	.490

[a] Based upon the sample mean and standard deviation given in Table 3 for n = 14..

Table 5. SIMILARITY VALUES AND RERANKINGS OBTAINED FROM THE SHEPARD DATA AND THE MAX METHOD

Object pairs	Similarity value	Partition rank	"Minimum γ rank"	"Maximum γ rank"	Object pairs	Similarity value	Partition rank	"Minimum γ rank"	"Maximum γ rank"
{1,2}	.685	1	1	1	{1,5}	.938	6	19	19
{4,7}	.692	2	2	2	{3,8}	.939	7	20	20
{3,5}	.773	3	3	3	{8,9}	.948	8	21	21
{7,9}	.794	5	4	4	{5,9}	.950	8	22	22
{6,8}	.809	4	5	5	{1,6}	.953	7	23	23
{2,4}	.819	8	6	6	{1,4}	.955	8	24	25
{5,8}	.832	7	7	7	{2,8}	.955	7	25	24
{2,3}	.864	6	8	8	{2,6}	.956	7	26	26
{4,9}	.867	5	9	9	{3,7}	.963	8	27	27
{2,5}	.879	6	10	10	{6,9}	.965	8	28	28
{3,6}	.880	7	11	11	{6,7}	.969	8	29	29
{1,3}	.886	6	13	12	{2,9}	.969	8	30	30
{4,5}	.886	8	12	13	{3,4}	.969	8	31	31
{5,6}	.889	7	14	14	{1,7}	.973	8	32	33
{5,7}	.894	8	15	15	{1,8}	.973	7	33	32
{7,8}	.895	8	16	16	{3,9}	.975	8	34	34
{2,7}	.922	8	17	17	{1,9}	.979	8	35	35
{4,8}	.922	8	18	18	{4,6}	.980	8	36	36

Table 6. PARTITION HIERARCHY OBTAINED FROM SHEPARD'S DATA USING THE MAX METHOD

Level	Partition
1	{{1,2},{3},{4},{5};{7},{8},{9}}
2	{{1,2},{4,7},{3},{5},{6},{8},{9}}
3	{{1,2},{3,5},{4,7},{6},{8},{9}}
4	{{1,2},{3,5},{4,7},{6,8},{9}}
5	{{1,2},{3,5},{6,8},{4,7,9}}
6	{{1,2,3,5},{6,8},{4,7,9}}
7	{{1,2,3,5,6,8},{4,7,9}}

FOOTNOTES

[*] Lawrence Hubert is Assistant Professor, Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin, 53706.

[1] To be consistent with Johnson [8], the term "similarity" is used rather than "dissimilarity."

REFERENCES

[1] Anglin, J.M., The Growth of Word Meaning, Cambridge: The M.I.T. Press, 1970.

[2] Conover, W.J., Practical Nonparametric Statistics, New York: Wiley, 1971.

[3] Gibbons, J.D., Nonparametric Statistical Inference, New York: McGraw-Hill, 1971.

[4] Goodman, L.A. and Kruskal, W.H., "Measures of Association for Cross-classifications," Journal of the American Statistical Association, 49 (December 1954), 732-764.

[5] Hubert, L., "Some Extensions of Johnson's Hierarchical Clustering Algorithms," Psychometrika, 37 (September 1972),

[6] Hubert, L., "Monotone Invariant Clustering Procedures," Psychometrika, 38 (March 1973), in press.

[7] Jardine, N. and Sibson, R., Mathematical Taxonomy, London: Wiley, 1971.

[8] Johnson, S.C., "Hierarchical Clustering Schemes," Psychometrika, 32 (September 1967), 241-254.

[9] Lance, G.N. and Williams, W.T., "A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems," The Computer Journal, 9 (February 1967), 373-80.

[10] McQuitty, L.L., "Hierarchical Linkage Analysis for the Isolation of Types," Educational and Psychological Measurement, 20 (Winter 1960), 55-67.

[11] Miller, G.A., "A Psychological Method to Investigate Verbal Concepts," Journal of Mathematical Psychology, 6 (June 1969), 169-191.

[12] Shepard, R.N., "Stimulus and Response Generalization: Tests of a Model Relating Generalization to Distance in Psychological Space, Journal of Experimental Psychology, 55 (June 1958), 509-523.

[13] Sokal, R.R. and Sneath, P.H.A., Principles of Numerical Taxonomy, San Francisco: Freeman, 1963.

[14] Williams, W.T., Lance, G.N., Dale, M.B., and Clifford, H.T., "Controversy Concerning the Criteria for Taxonometric Strategies," Computer Journal, 14 (May 1971), 162-165.