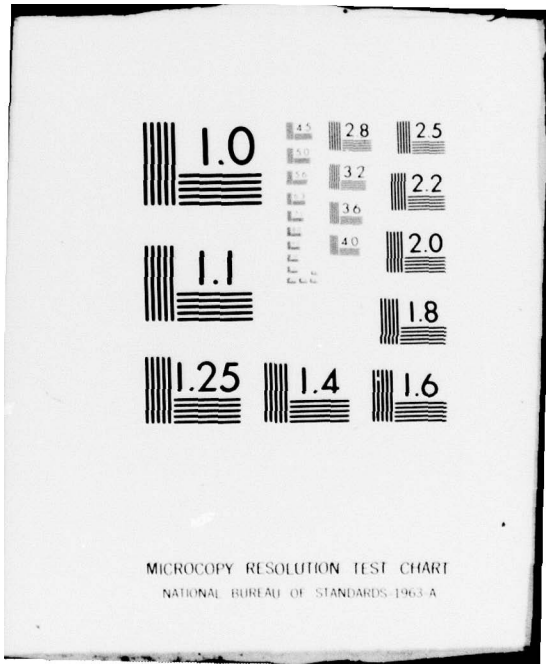
AD-A079 554 MASSACHUSETTS UNIV AMHERST DEPT OF MATHEMATICS AND S--ETC F/G 12/1
SIMILARITY MEASURES ON BINARY ATTRIBUTE DATA - II.(U)
DEC 79 M F JANOWITZ N00014-79-C-0629
UNCLASSIFIED TR-J7902 NL

1 OF 1
ADA 079 554

END
DATE FILMED
2-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Technical Report Number J790

SIMILARITY MEASURES ON BINARY ATTRIBUTE DATA. II

LEVEL II

M. F. Janowitz
Department of Mathematics & Statistics
UNIVERSITY OF MASSACHUSETTS
Amherst, Massachusetts

December 1979

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) TR-J7902	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) Similarity Measures on Binary Attribute Data - II,	5. TYPE OF REPORT & PERIOD COVERED (9) Technical rept., / 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) M. F. Janowitz	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-79-C-0629
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Massachusetts Amherst, MA 01003 (12) 21	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 121405
11. CONTROLLING OFFICE NAME AND ADDRESS Procuring Contracting Officer Office of Naval Research Arlington, VA 22217 (11)	12. REPORT DATE Dec [redacted] 79 / 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Resident Representative, Harvard University Gordon McKay Laboratory, Room 113 Cambridge, MA 02138	15. SECURITY CLASS. (of this report) Unclassified / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Numerical taxonomy, Cluster analysis, Similarity Measure, Special clustering, Optimality measures, Cophenetic correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The exact nature of the coefficient of special similarity is investigated, and its ability to recognize Gilmour natural classifications is compared to that of various other similarity measures by means of a computer simulation of the classification problem.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

411 377

Similarity Measures On Binary Attribute Data - II

M. F. Janowitz*

This report has its genesis in the work by Farris (1977) in which a claim is made that the phylogenetic system of classification is superior to the phenetic. I pointed out certain flaws in Farris' reasoning (Janowitz,1979), and Farris' reply appears in Farris (1979). I enlarged upon my views in Janowitz (1979a), and the current report should be regarded as a continuation of that work. The terminology and notation will follow that of Janowitz (1979a), and much of the data will be taken from there. Despite this, it will prove useful to briefly redefine the similarity measures that will be considered, and restate some of the issues that are under contention.

The input data is a set of p objects to be classified, together with a set of n attributes. For purposes of the present work, I shall assume that each attribute has two states, and that they are coded 0 and 1. For objects J and K, one can now define the following symbols:

symbol	No. attributes for which
a	J and K have state 1
b	J has state 1 and K has state 0
c	J has state 0 and K has state 1
d	J and K have state 0.

*Research supported by ONR Contract N00014-79-C-0629 as well as by grants from the University of Massachusetts Computer Center.

Here then are the similarity coefficients that will be considered:

Measure	Name	Similarity Formula	Dissimilarity Formula
DC1	Simple Matching	$(a + d)/n$	$(b + c)/n$
DC2	Jaccard	$a/(a + b + c)$	$(b + c)/(a + b + c)$
DC4	Russell and Rao	a/n	$1 - a/n$
DC10	Yule	$(ad - bc)/(ad + bc)$	$bc/(ad + bc)$

For binary attribute data, DC1 represents Farris' coefficient of overall similarity. Letting s denote the similarity version of DC1, let us see now how Farris defines his coefficient of special similarity (1977:826,836). The idea is to decide that for each attribute, one of the two states is uninformative. An object R, called the reference point, is then defined. R might be one of the objects already under consideration, or it might be a new object. In either case, R has the property that for each attribute, it has the uniformative state. To compute the special similarity coefficient a, one then uses the formula (Farris,1979:836)

$$a(J,K) = \tfrac{1}{2}\{1 + s(J,K) - s(J,R) - s(K,R)\}.$$

This of course produces the similarity version of a: to view it as a measure of dissimilarity, one simple considers $1 - a(J,K)$.

Here is a portion of the argument used by Farris in Farris (1977 and 1979), stated in what I hope is an accurate manner. Pheneticists favor using the coefficient

of overall similarity, and have also favored using the cophenetic correlation coefficient as a measure of how well a cluster technique really works. Farris presented a set of fully congruent attributes that represented a Gilmour natural classification. He then showed that this classification was recaptured by special similarity but not by overall similarity. He concluded that overall similarity could give a wrong classification. He then proceeded to establish that it generally gives the wrong classification by applying the two coefficients to a large number of real life data sets. In each case, the cophenetic correlation coefficient said that special similarity was far superior to overall similarity. He concluded that since the optimality measure favored by pheneticists produced the result that a phylogenetic method was superior to the phenetic method favored by these same people, it must follow that the phylogenetic method is indeed superior.

I argued in Janowitz (1979 and 1979a) that special similarity does not do a very good job of recognizing natural classifications for attribute data that is not fully congruent, and that the cophenetic correlation coefficient cannot be used as a measure of the ability of a similarity measure to recapture such classifications. Indeed, it was argued (1979a) that if the reference point consists of all 0 states, then special similarity does

not perform as well as either simple matching or Yule's coefficient. But Farris did not choose his reference point in this manner. Rather, he took as a reference point the first object he came to (Farris (1977:836) and (1979:210)). This was something that I did not do in 1979a, and something that will be explored in the present work. In section 1, I shall take a careful look at the nature of special similarity, while in sections 2 and 3, I shall compare its performance with various other similarity measures, taking the reference point to be the first object of the set of object to be classified. The comparison will be made both for fully congruent attribute data, and for the more general situation where various types of errors are introduced into the attribute data - the idea being to calculate the ability of various methods to recapture Gilmour natural classifications when they can be recognized.

§ 1. The nature of special similarity. Letting R be the reference point, let's have a close look at a(J,K) where J,K are objects to be classified. Suppose that there are n attributes, and that $h_1, h_2, \ldots, h_8$ are non-negative integers whose sum is n. Suppose further that the following table describes the attribute data for R,J,K:

No. attributes	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
R	0	0	0	0	1	1	1	1
J	0	1	0	1	0	1	0	1
K	0	0	1	1	0	0	1	1

Application of the formula for a shows that

$$a(J,K) = (h_4 + h_5)/n.$$

But the same effect may be obtained by simply recoding the attributes so that R has all 0 states, and then applying DC4. This is shown by the recoded data table

No. attributes	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
R	0	0	0	0	0	0	0	0
J	0	1	0	1	1	0	1	0
K	0	0	1	1	1	1	0	0

From this viewpoint one can instantly see a major difficulty in the interpretation of the output of special similarity. For suppose one is given the input data

object	attribute						
1	1	1	1	1	0	0	0
2	1	1	1	0	1	0	0
3	0	1	1	0	0	1	0
4	0	0	1	0	0	0	1

If this is viewed as presence-absence data, these attributes then reflect the natural classification whose clusters at each level are:

Level 1 12, 3, 4

Level 2 123, 4

Level 3 1234 .

Taking object 1 as the reference point, special similarity acts just like DC4 on the data matrix

object	attribute						
1	0	0	0	0	0	0	0
2	0	0	0	1	1	0	0
3	1	0	0	1	0	1	0
4	1	1	0	1	0	0	1

.

This produces the classification

Level 1 1, 2, 34

Level 2 1, 234

Level 3 1234 .

A common reaction to all of this seems to be that if one views the output of special similarity as an unrooted tree, then the desired natural classification may still be recaptured. The problem is that without prior knowledge of the desired classification, it is difficult to see how to reroot such a tree so as to simultaneously remove the unwanted cluster 34, and produce the desired cluster 12. Similar examples will be considered in the next section.

§ 2. Special similarity on congruent input data.

Special similarity was applied to the data sets that appeared as Tables 6,7,8,9 of Janowitz (1979a). The data in the tables attempt to reflect the following natural classifications (only the nontrivial clusters will be listed at each level):

Table 6		Table 7	
Level 1	12,34,56,78,9-10	Level 1	12, 9-10
Level 2	1-4,56,78,9-10	Level 2	1-3, 8-10
Level 3	1-6, 78, 9-10	Level 3	1-4, 7-10
Level 4	1-8, 9-10	Level 4	1-5, 6-10
Level 5	1-10	Level 5	1-10

Table 8		Table 9	
Level 1	12, 45, 9-10	Level 1	12, 34, 56, 78, 9-10
Level 2	1-3, 4-6, 8-10	Level 2	1-4, 56, 7-10
Level 3	1-6, 7-10	Level 3	1-6, 7-10
Level 4	1-10	Level 4	1-10

When special similarity is applied, the result in each case is an ultrametric. Application of single linkage clustering now produces the classifications shown in Fig. 1. The reader should check for himself to see what a poor job has been done in recapturing the desired natural classifications. Furthermore, when the classifications of Fig. 1 are viewed as unrooted trees, it is difficult to see how to reroot the trees so as to recapture the desired classifications unless one has

prior knowledge of the underlying natural classifications. This is especially well illustrated by Fig. 1(b). How would one know that this tree should be rerooted so as to produce the classification 1-6,7-10, for example?

There really seems to be no way out of this dilemma. If one chooses the first object as a reference point, then special similarity will not recapture a natural classification, even for fully congruent attribute data; on the other hand, if one takes as a reference point an object having all 0 states, then special similarity works fine in the fully congruent case, but as was shown in Janowitz (1979a), it does not perform well in the incongruent case. It remains to be seen how special similarity will perform in the incongruent case when the reference point is taken as the first object to be classified. This will be done in the next section.

§3. Special similarity on incongruent input data.

As a test of this, I repeated a simulation that was done in Janowitz (1979a). I took each of the data matrices from Tables 6,7,8,9 and 1 of that paper, replicated the characters as indicated therein, doubled each character, introduced a 5% random error in reading character states, introduced 6 random characters, and finally discarded 10% of the resulting characters in a random fashion. I used special similarity as well as DC1, DC2, DC4 and DC10 followed by both single linkage and u=.5-clustering. Methods 1 and 4 of 1979a were then used to measure the ability of each cluster method to recapture the classification produced by the unperturbed data. In all cases except that of special similarity, this measures the ability of the cluster method to recapture the desired natural classification. In the case of special similarity, it measures how well the faulty classifications of Fig. 1 are recaptured. Here then are the results, based upon 5 trials with each data set, and using single linkage clustering.

Table 6	Method 1		Method 4		Coph. corr.	
	Mean	SD	Mean	SD	Mean	SD
DC1	.8968	.0429	.08162	.0753	.8183	.1112
DC2	.8671	.0703	.1143	.1545	.8290	.1090
DC4	.7107	.2617	.2760	.2293	.7448	.1873
DC10	.9137	.1202	.0875	.1166	.7586	.1909
Spec	.8612	.0738	.1265	.1140	.9367	.0192

Table 7	Method 1		Method 4		Coph. corr.	
	Mean	SD	Mean	SD	Mean	SD
DC1	.8377	.0542	.2279	.0727	.8488	.0358
DC2	.9199	.0387	.2099	.1602	.8750	.0474
DC4	.8583	.0436	.2212	.1198	.7918	.0732
DC10	.8368	.0482	.1469	.1344	.7707	.0305
Spec	.8867	.0520	.0709	.0470	.9678	.0131
Table 8						
DC1	.8482	.1285	.1893	.1948	.8013	.1209
DC2	.8879	.0580	.0923	.0405	.8610	.0575
DC4	.6384	.1147	.4075	.0860	.7281	.1281
DC10	.8938	.0980	.0786	.0242	.7794	.0642
Spec	.8870	.0591	.1137	.0693	.9536	.0238
Table 9						
DC1	.7769	.1598	.2596	.1309	.7858	.0789
DC2	.7891	.1172	.2014	.1779	.8458	.0395
DC4	.6012	.1416	.3990	.1567	.7482	.0649
DC10	.8073	.1819	.1705	.1901	.7330	.0835
Spec	.8872	.0448	.1426	.0948	.9457	.0217
Table 1						
DC1	.8099	.1135	.2256	.2012	.7721	.0790
DC2	.7586	.0736	.2746	.1931	.7967	.0679
DC4	.5676	.0358	.4360	.0453	.7146	.0644
DC10	.7787	.1744	.1799	.1460	.7172	.0716
Spec	.8716	.0456	.0857	.0633	.9585	.0151

The data in the above table pretty well confirm results that were announced in 1979a. With respect to the cophenetic correlation coefficient, special similarity appears to be far superior to any of the others. (Note. This is the result that Farris announced in 1977 and 1979). But it also does well with respect to the criteria of Methods 1 and 4. This shows that it does a good job of recapturing its faulty view of the underlying classifications, and consequently

must do a poor job of actually recapturing those classifications. Ignoring the performance of special similarity, DC10 is best in 3 out of 5 cases with respect to the criterion of Method 1, and is best in 4 out of 5 cases with respect to Method 4. Though it seems pointless to reproduce the results of these 25 trials, it still is informative to reproduce a portion of the actual clusters produced by special similarity. The interested reader may obtain the remainder from the author upon request. Here then are the results from the data of Table 6.

Level	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
1	9-10	9-10	9-10	78	9-10
2	78,9-10	78,9-10	78,9-10	56,78	5 9-10,78
3	56,78 9-10	7-10	7-10	56,78,9-10	5 7-10
4	56,7-10	34,7-10	5-10	5-8,9-10	5-10
5	5-10	34,6-10	4-10	23 5-10	3-10
6	3-10	3-10	2-10	2-10	2-10
7	2-10	2-10	1-10	1-10	1-10
8	1-10	1-10			

The above table indicates the nontrivial clusters only. These results should be compared with the desired natural classifications that appear on p. 7.

The same simulation was then performed using u=.5-clustering in place of single linkage clustering, and here are the results.

Table 6	Method 1		Method 4		Coph. Corr.	
	Mean	SD	Mean	SD	Mean	SD
DC1	.9250	.0222	.1523	.0717	.8713	.0328
DC2	.9114	.0184	.1648	.1134	.8958	.0285
DC4	.8746	.0433	.1727	.1218	.8317	.0539
DC10	.9058	.0839	.1156	.0899	.8363	.0586
Spec	.8530	.0767	.1903	.0952	.9838	.0216
Table 7						
DC1	.9153	.0793	.0919	.1067	.8474	.0544
DC2	.9220	.0646	.1156	.1604	.8955	.0352
DC4	.8979	.0906	.1588	.1638	.8480	.0375
DC10	.9702	.0160	.0455	.0172	.8300	.0487
Spec	.9424	.0136	.0541	.0356	.9744	.0081
Table 8						
DC1	.7584	.1995	.1887	.1649	.7820	.1241
DC2	.7332	.2699	.2790	.2264	.8031	.1252
DC4	.3979	.1498	.4747	.0949	.6080	.1928
DC10	.7829	.2439	.2008	.2073	.7218	.1181
Spec	.9408	.0227	.0786	.0422	.9434	.1145
Table 9						
DC1	.9119	.0591	.1020	.0814	.8496	.0515
DC2	.8619	.0891	.1283	.1279	.8833	.0323
DC4	.8366	.1196	.1586	.1280	.8131	.0351
DC10	.8987	.0855	.1378	.1231	.8268	.0434
Spec	.9010	.0661	.1471	.1279	.9582	.0124
Table 1						
DC1	.8432	.1570	.1481	.1468	.7886	.0744
DC2	.7985	.1729	.1540	.1506	.8433	.0575
DC4	.6223	.2914	.3233	.2407	.7027	.1292
DC10	.7947	.2287	.2132	.1796	.7496	.1080
Spec	.9170	.0516	.1074	.1122	.9735	.0094

The situation is similar to that of the earlier case. Special similarity has by far the highest cophenetic correlation, while DC4 is significantly the lowest. Since special similarity also performs pretty well with respect to the criteria of Methods 1 and 4, what this shows is that it does a very good job of reflecting a classification that is not the one that we are after. If one ignores special similarity, then DC1 is best in 3 out of

the 5 cases with respect to the criterion of Method 1, and is best in 4 out of 5 cases with respect to Method 4. In all of these instances, DC4 is the worst. Thus whether one regards the first object as the reference point or one takes the reference point to have all 0 states, special similarity is the worst choice of these dissimilarity measures with respect to its ability to recapture a natural classification in the presence of errors.

Having established that special similarity is not a particularly good choice as a similarity measure, at least for this particular simulation, I would like to close this section by presenting 5 more trials on each data set. This time, only DC1, DC2 and DC10 are involved.

Table 6	Method 1		Method 4		Coph. Corr.	
	Mean	SD	Mean	SD	Mean	SD
DC1	.9347	.0414	.0917	.1039	.8853	.0675
DC2	.9190	.0254	.0681	.0461	.9228	.0132
DC10	.9640	.0362	.0329	.0209	.8742	.0298
Table 7						
DC1	.8959	.0674	.1592	.0899	.8566	.0261
DC2	.8608	.0773	.3000	.2283	.8550	.0601
DC10	.8151	.1384	.1784	.1897	.7642	.0534
Table 8						
DC1	.8142	.1151	.2059	.1928	.8035	.0929
DC2	.8062	.0260	.1778	.1536	.8486	.0572
DC10	.8723	.1226	.1327	.1110	.7829	.0584
Table 9						
DC1	.8850	.0699	.0612	.0375	.8459	.0363
DC2	.7907	.0900	.1273	.1164	.8785	.0341
DC10	.9334	.0495	.0651	.0400	.8468	.0530
Table 1						
DC1	.7337	.1409	.2926	.1650	.7854	.1262
DC2	.7242	.1347	.2640	.2113	.8344	.0964
DC10	.7511	.2174	.2714	.1870	.7667	.1192

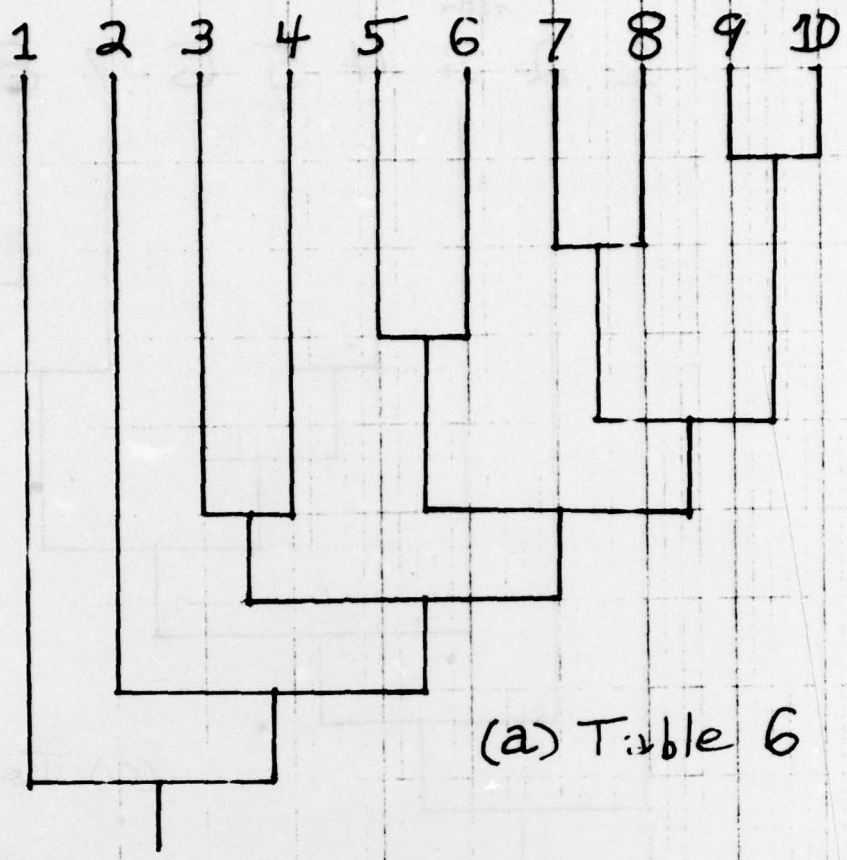
Here it should be noted that Method 1 shows DC10 to be best on 4 out of the 5 data sets, Method 4 shows DC10 to be best on 3 of the 5 sets, while the cophenetic correlation coefficient produces quite different results in that it makes DC10 the worst choice in 4 out of the 5 sets. Again we have proof that the cophenetic correlation coefficient does not provide a meaningful criterion for measuring the performance of a similarity measure.

4. Conclusion. If one decides that it is important for cluster methods to be able to recognize natural classifications in the presence of certain types of error, then the results that were presented in Janowitz (1979a) and the present paper seem to indicate that special similarity is a very poor choice for a similarity measure, while both simple matching and Yule's coefficient seem to be quite good. One must exercise extreme caution, however, in attempting to draw conclusions such as this from specific data sets. The results could very well be data dependent! The only safe conclusion that can be drawn (hence the only conlcusion that I shall draw) is that the results cast considerable doubt on any assertion that special similarity is superior to the simple matching coefficient. The matter should still be regarded as open. One can of course use these results as an indicator that special similarity and Yule's coefficient might be reasonable candidates for use as a measure of similarity, but at the moment that is all I dare say.

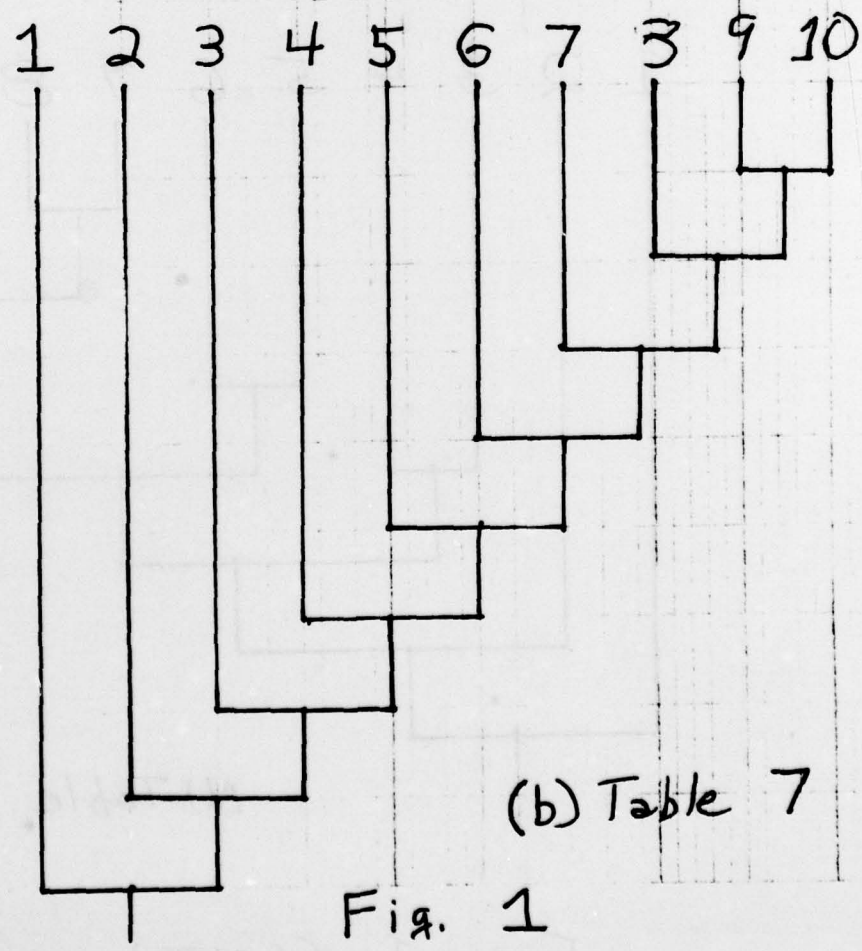
What then is it that Farris has shown? He took a phylogenetic cluster method, applied a measure of optimality that has in the past been favored by pheneticists, and showed that in a large number of trials with different data sets, his phylogenetic method was deemed superior to the method favored by the pheneticists. I have indicated in

great detail why it is that one cannot conclude that this shows the phylogenetic method to be superior to any phenetic method. What it <u>does</u> show is that the optimality measure fails. This is the heart of what Farris has demonstrated. I have little quarrel with his data - only with the conclusion that he has drawn from that data.

In a later paper, I shall examine the question of just why it is that Yule's coefficient performs well on the type of data that I have been examining. I shall also make some concrete suggestions as to which similarity measures ought generally to be used.

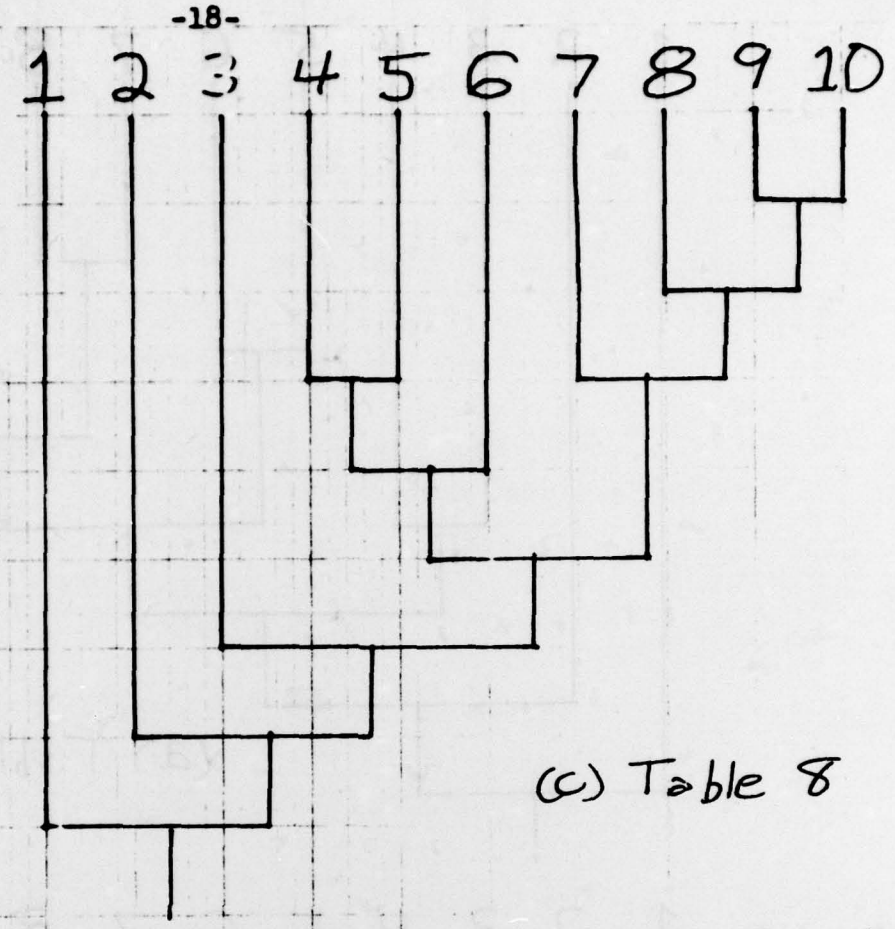


(a) Table 6

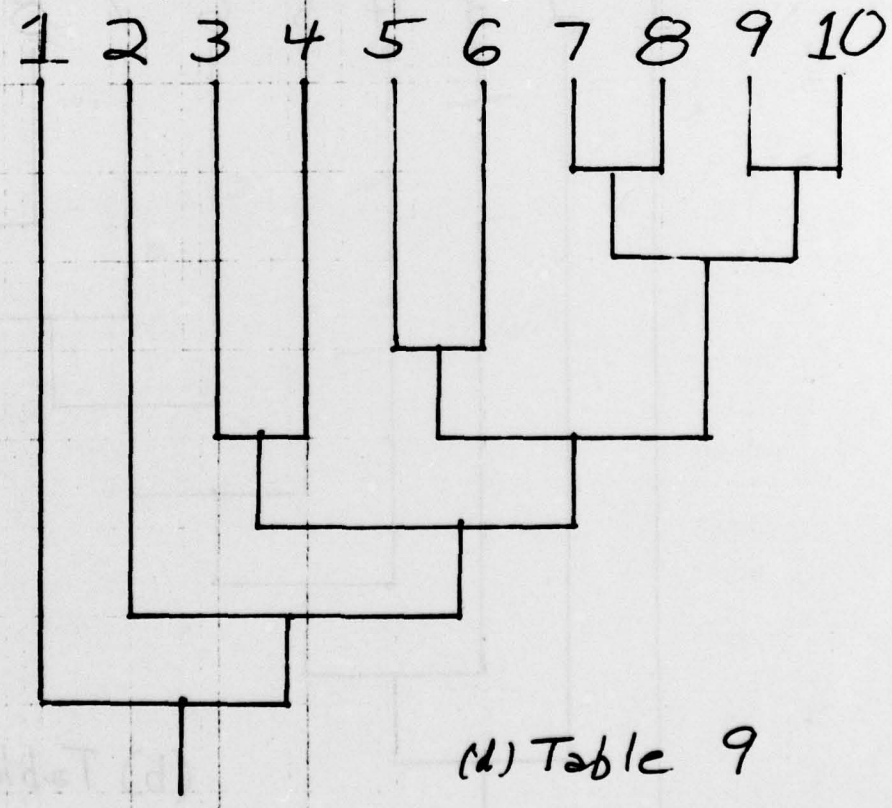


(b) Table 7

Fig. 1



(c) Table 8



(d) Table 9

Fig. 1 (CONT)

REFERNCES

Farris, J. S. 1977. On the phenetic approach to vertebrate classification. In Hecht, M.K., P. C. Goody amd B. M. Hecht (eds.), Major patterns in vertebrate evolution. Plenum, New York, pp. 823-850.

Farris, J. S. 1979. On the naturalness of phylogenetic classifications. Syst. Zool. 28:200-214.

Janowitz, M. F. 1979. A note on phenetic and phylogenetic classifications. Syst. Zool. 28:197-199.

Janowitz, M. F. 1979a. Similarity measures on binary attribute data. University of Massachusetts Technical Report No. J7901.

Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003
USA