ED 442 866 TM 031 286

AUTHOR	Finch, Holmes; Huynh, Huynh
TITLE	Comparison of Similarity Measures in Cluster Analysis with Binary Data.
PUB DATE	2000-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Algorithms; *Cluster Analysis; Monte Carlo Methods; Responses; Sample Size; Simulation
IDENTIFIERS	*Binary Data Analysis; Dichotomous Variables; *Similarity (Concept)

ABSTRACT

One set of approaches to the problem of clustering with dichotomous data in cluster analysis (CA) was studied. The techniques developed for clustering with binary data involve calculating distances between observations based on the variables and then applying one of the standard CA algorithms to these distances. One of the groups of distances that are designed for binary data is known collectively as matching coefficients. There are several incarnations of matching coefficients, but all take as their main goal the measurement of response similarity between any two observations. Thus, distance and similarity come to express the same concept with respect to the observations. This study examined four measures of association that are common to four previous studies. Using Monte Carlo simulation, cluster analysis was conducted using the four distance measures. Under the conditions of this study, the four measures performed very much the same in terms of correctly classifying individuals into two clusters based on dichotomous variables. Another interesting result is that clustering solutions were virtually identical for samples of size 240 and 1,000. (Contains 6 tables, 6 figures, and 12 references.) (SLD)

Comparison of Similarity Measures in Cluster Analysis with Binary Data

Holmes Finch

Huynh Huynh

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in New Orleans
April 2000

Comparison of Similarity Measures in Cluster Analysis with Binary Data

Holmes Finch
Department of Statistics
Huynh Huynh
College of Education
University of South Carolina

Holmes Finch
Department of Statistics
University of South Carolina
Columbia, SC 29208
(803) 777-5074
finch@stat.sc.edu

Introduction

Cluster Analysis (CA) is an analytic technique used to classify observations into a finite (and ideally) small number of groups based on scores taken from two or more measures. Sometimes there are hypotheses regarding the number and make up of such groups, but more often there is little or no prior information, thus making CA an exploratory analysis. There are a number of clustering algorithms available, all having as their primary purpose the measurement of mathematical distance between individual observations, and the subsequent clustering of these individuals based on the distances. Distance between observations in this context is often (though not always) expressed as Euclidean distance, or some similar measure of difference between individuals on a set of measured variables Johnson and Wichern (1992). One of the primary assumptions underlying these standard methods for calculating distance is that the metrics are all continuous in nature, either interval or ratio Anderberg (1973). However, some research situations involve the use of a mixed set of variables, some continuous and others categorical (either nominal or ordinal), or a set containing only categorical variables. In such situations, the standard Euclidean measures of distance are inappropriate, and must be replaced by some other statistic Dillon and Goldstein (1984). It is the purpose of this paper to investigate one set of approaches to the problem of clustering with dichotomous data, recognizing that the data structure discussed here is but one of many and that alternative approaches might be more valid in other circumstances.

Distance measures

The techniques developed for clustering with binary data involve calculating distances between observations based upon the variables and then applying one of the standard CA algorithms to these distances. One group of these distances that are designed for binary data is known collectively as matching coefficients Dillon and Goldstein (1984). There are several incarnations of matching coefficients, but all of them take as their main goal the measurement of response similarity between any two observations. The logic underlying all of these techniques is that two individuals should be viewed as similar to the degree that they share common attributes, Snijders, et al, (1990). Thus, distance and similarity come to express the same concept with respect to the observations. While it is recognized that there may be other approaches, which can be used to cluster nominal or ordinal data and which may not place this constraint on the variables, they are beyond the scope of this paper and thus will not be addressed here.

In order to describe these methods, please refer to the contingency table below. The rows represent presence or absence (1,0) of a set of traits for a single observation, I, and the columns represent presence or absence of the same traits on a second observations, j, where $i \neq j$.

Table 1

	Observation 2	
Observation 1	1	0
1	*a*	*b*
0	*c*	*d*

Thus, cell a includes the count of the number of variables for which the two observations both had the attribute present, while cell b includes the count of the number of variables for which the first observation had the attribute present and the second observation did not, and so on. The primary difference between the measures of association that are described here is in the way that they manipulate these cell counts.

While there are many of these indices available, Hands and Everitt (1987), this paper will only examine the 4 measures of association that are common to Anderberg (1973), Dillon and Goldstein (1984), Lorr (1983) and Snijder, et al (1990). This limitation is more a function of limited space than any technical determinations regarding the fitness of these indices. Indeed, Anderberg (1973) suggests using several of these indices in order to get a feel for the nature of the clustering in the data. However, perhaps it can be assumed that the indices that appear in four different texts are considered by more than one writer to be reliable for use in clustering problems similar to the one described here.

The first of these measures of association is known as the Russell and Rao index. It can be expressed in terms of the cells of table 1 described above as:

$$\frac{a}{a+b+c+d}$$

This index is simply the proportion of cases in which both observations had the trait of interest, with the denominator including all cells of the 2x2 table. In contrast to this is the Jaccard coefficient, which is similar but which excludes cases where neither observation indicates having the trait of interest. The equation for the Jaccard coefficient is:

$$\frac{a}{a+b+c}$$

A third variation on this theme, called the matching coefficient, includes both matched cells in the numerator as well as in the denominator:

$$\frac{a+d}{a+b+c+d}$$

The final index to be examined here, Dice's coefficient. It can be expressed as:

$$\frac{2a}{2a+b+c}$$

Dice's coefficient is closely related to the Jaccard coefficient, with additional weight being given to the cases of agreement. While this calculation removes the notion of a proportion from interpretation, it has been justified as follows:

> ...for 0,1 mismatches the zero is just as trivial as in the 0,0 case. Mismatches should then like about midway along the scale of significance between the 0,0 and 1,1 cases respectively. The number of mismatches in the coefficient should by this reasoning be multiplied by 1/2. (Anderberg, 1973).

In short, agreement is more important than disagreement and thus should receive a greater weight in the calculation of association.

As an example of how these methods work, assume two observations, each of which have measurements for 7 binary variables, where presence is denoted by 1 and absence is denoted by 0. Given this structure, the following table could be created.

Table 2

Observation	X1	X2	X3	X4	X5	X6	X7
1	1	1	0	0	1	0	0
2	0	1	0	1	1	1	0

These data could then be used to produce the following 2x2 contingency table.

Table 3

	Observation 2	
Observation 1	1	0
1	2	1
0	2	2

It is the information from this contingency table that is used to calculate the measures described above. Taking these data, we can calculate the distance between the two observations using each of the four measures described here.

Table 4

Coefficient	Equation	Result
Russell/Rao	$\frac{2}{2+1+2+2}$	$\frac{2}{7}$
Jaccard	$\frac{2}{2+1+2}$	$\frac{2}{5}$
Simple Matching	$\frac{2+2}{2+1+2+2}$	$\frac{4}{7}$
Dice	$\frac{2(2)}{2(2)+1+2}$	$\frac{4}{7}$

The greatest similarity, and thus the smallest distance, was calculated using Dice's coefficient and the matching statistic. The greatest distance was found using Russell/Rao. Were we to use these results in a clustering algorithm, we would input the value (1 – coefficient) into a distance matrix which in turn would be entered into the clustering algorithm.

Methodology

Monte Carlo data for binary variables were generated using a 2 Parameter Logistic (2PL) model. The value of the latent theta variable was generated using a standard normal distribution, as was the difficulty parameter, while the discrimination parameter was generated using a uniform (0,1) distribution. The probabilities calculated using the 2PL model were then compared against random uniform (0,1) values, with the dichotomous variables being assigned a 1 if the probability was larger than the uniform and a 0 if the probability was smaller. Two sets of unidimensional dichotomous variables were simulated, with each set being based upon one of two thetas. In all cases, two clusters were simulated, being differentiated by the mean of the two thetas. There were two pairs of thetas, -2/2 and -.5/.5, with one of the clusters having the high value of the pair and the other cluster having the low value. Furthermore, the variance of the thetas was varied as well, at either 1.5 or .5, with the same variance for each cluster. The number of dichotomous variables was either 10 (5 for each theta) or 24 (12 for each theta), and the number of subjects was either 240 (120 per cluster) or 1000 (500 per cluster). Note that all levels of each variable was completely crossed with all levels of the other variables, and that each combination of variables was represented by 1000 monte carlo data sets.

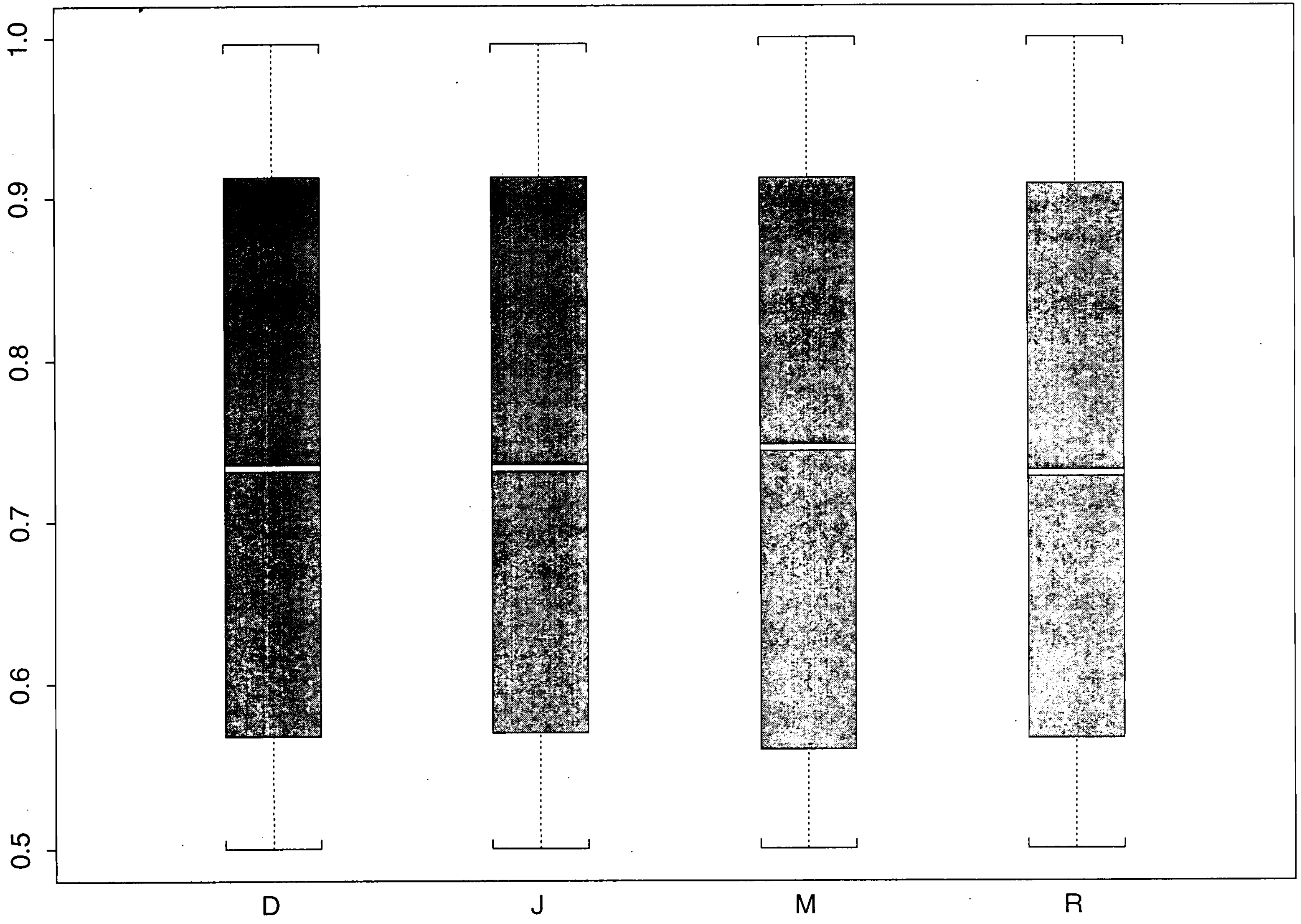
Cluster analysis was conducted using the four distance measures described above, in conjunction with Ward's method. Ward's was selected based upon results indicating that of the major clustering methods (excluding Model Based Clustering), it typically performs the best at population recovery of clusters Kuiper & Fisher (1975); Blashfield (1976); Overall, Gibson & Novy (1993). In addition, Hands and Everitt (1987) found that it performed the best at cluster extraction when used in conjunction with the matching coefficient. The results of the clustering solutions were compared using percent of cases correctly classified together and the kappa coefficient.

Results

Figure 1 shows the distribution of percent correctly classified for each of the 4 measures. It appears that the median percent correctly classified is nearly the same across all four measures, roughly 73%. The matching coefficient has a slightly higher median value

Figure 1

Boxplots of Percent Correct by Measure



than the other three measures, but the difference is not very large. The set of all values was the same for the four measures as well, ranging from .5 (basically chance) to 1 (perfect prediction). Table 1, below includes the mean and standard deviation (in parentheses) for percent correct and kappa for each measure.

Table 5
Kappa and Percent Correct by Measure

Measure	Kappa	Percent Correct
Dice	.471 (.344)	.735 (.172)
Jaccard	.473 (.344)	.736 (.172)
Matching	.466 (.351)	.733 (.175)
Russell/Rao	.467 (.344)	.734 (.172)

The differences between the values of kappa and percent correctly classified among the groups are indeed miniscule. Given that across all combinations of the manipulated variables there appears to be very little difference in the performance of the four measures, the second question of interest is whether this pattern is constant across specific levels of these variables. Figures 2, 3, 4 and 5 are interaction plots of the 4 measures with each of the manipulated variables.

Figure 2

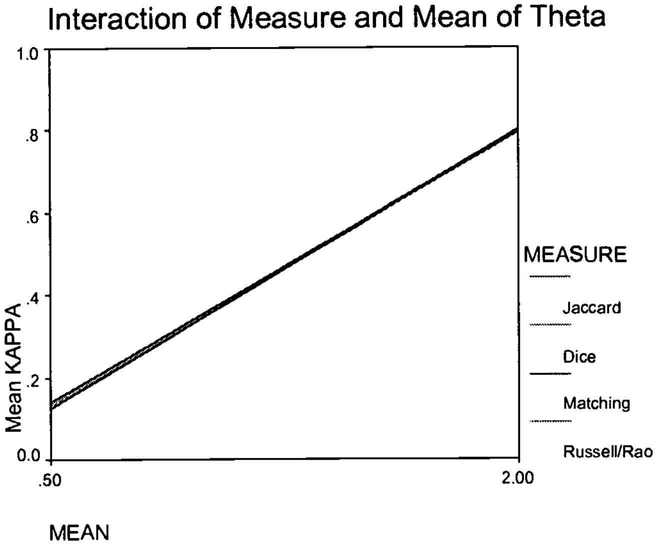


Figure 3

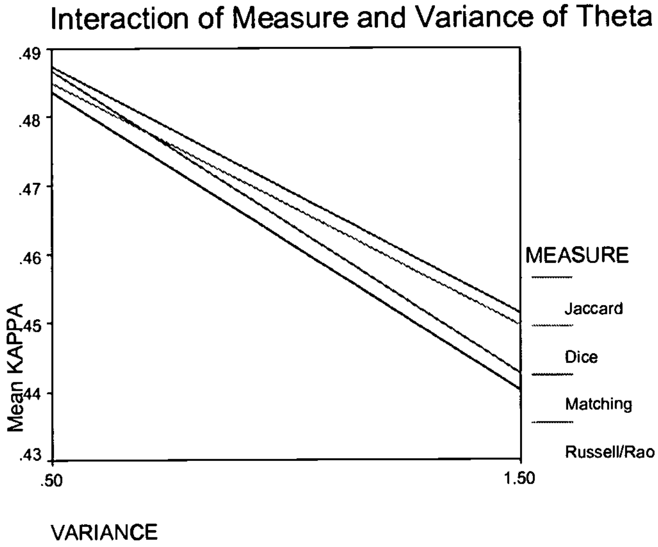


Figure 4

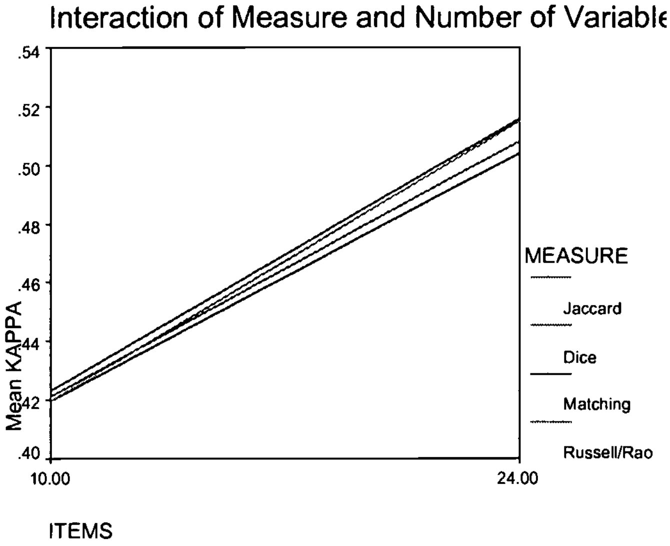
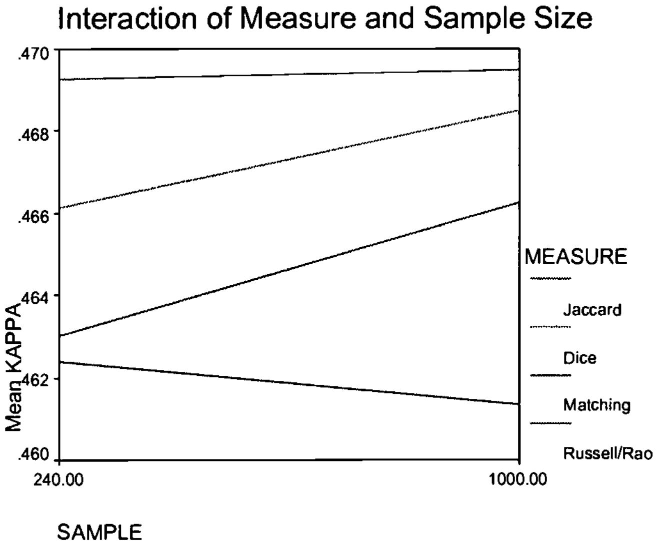


Figure 5



It appears that there is no interaction between the type of measure and any of the manipulated variables. While there is some divergence between the Dice measure and the other three on sample size, the actual values of kappa are within .01 of one another, which would seem to support the lack of any real interaction between the two variables.

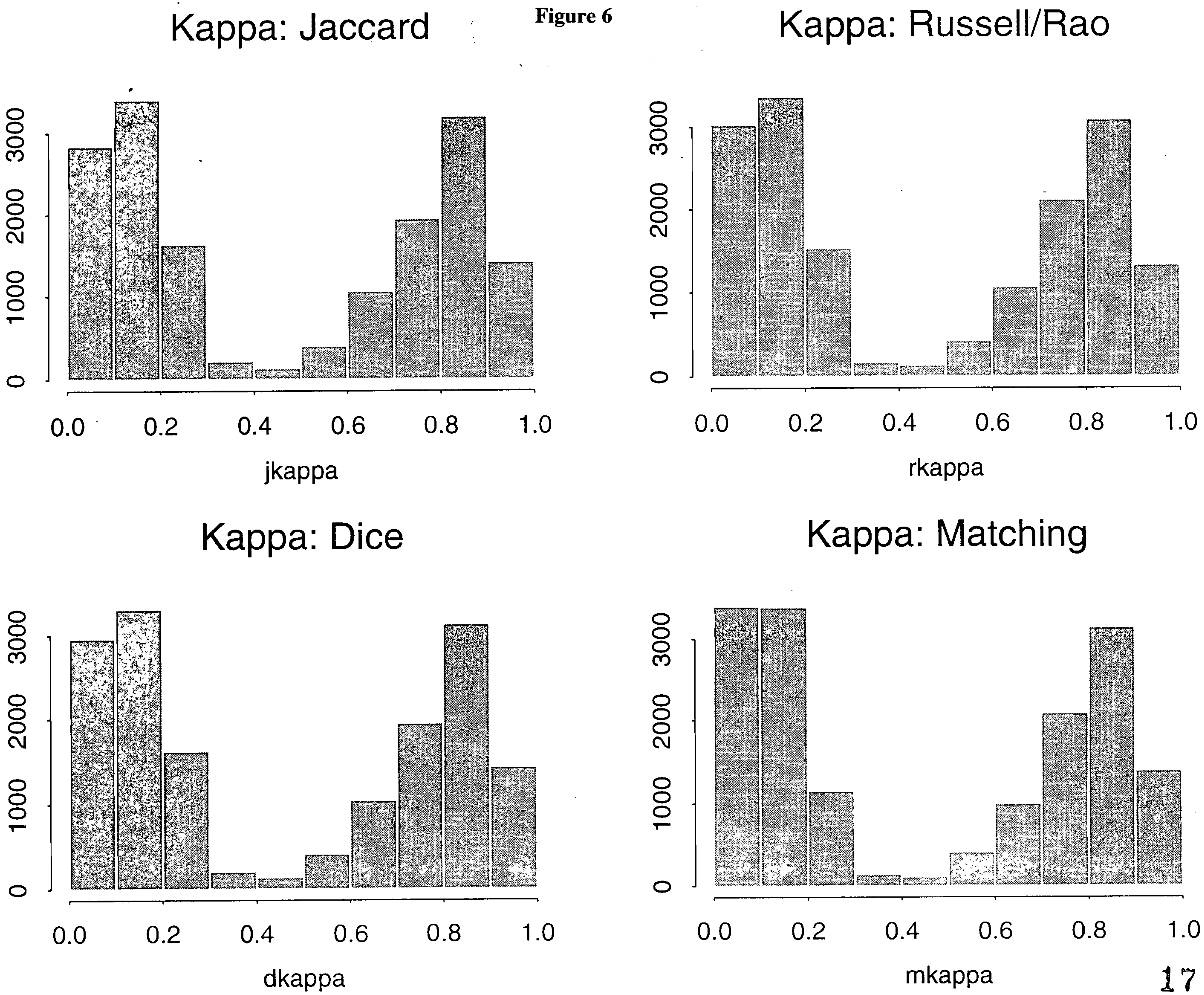
In terms of the manipulated variables, it appears that the major determining factor in terms of successful clustering in this study is the difference in the mean of theta for the two groups. Table 6, includes the mean percent correctly classified and the mean of kappa for each level of theta, variance of theta, number of variables and sample size.

Table 6

Variable		Kappa	Percent Correct
Mean	-.5/.5	.133	.567
	-2/2	.794	.897
Variance	.5	.486	.743
	1.5	.452	.726
Variables	10	.432	.716
	24	.509	.754
Sample size	240	.462	.731
	1000	.476	.738

Based on these results, it appears that in the context of this study, the only factor that greatly influenced the performance of the clustering algorithm was the mean of theta. The larger difference in the means was associated with an improvement of over 30% in correct classification over the smaller mean difference condition. As was noted in the figures above, these patterns were consistent across the 4 measures. Figure 6 demonstrates the effect of the difference in mean; for each measure, kappa takes a

Figure 6


Kappa: Jaccard
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
jkappa
Kappa: Russell/Rao
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
rkappa
Kappa: Dice
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
dkappa
Kappa: Matching
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
mkappa

bimodal distribution, with the mean of the lower part of the distribution being approximately .13, and the mean of the upper part being approximately .79.

Conclusions

Given the results described above, it appears that under the conditions present in this study, the four measures of association perform very much the same in terms of correctly classifying individuals into two clusters based on dichotomous variables. This pattern seems to hold true regardless of the mean difference in the underlying construct which produces the observed data, the variance of this construct, the number of variables used to cluster or the number of subjects in the sample. Given the similarities in the ways each are calculated, it is not completely surprising that this would be the case. The major difference among them is in how they handle the situation in which a trait is absent for both individuals, and with these results, it would appear that there is not much added information contained in this category. Furthermore, the fact that Dice's coefficient had results similar to the others seems to indicate that the amount of emphasis placed on the number of variables in agreement is not material to the success of the clustering algorithm.

Another interesting result of this study is that the clustering solutions were virtually identical for samples of size 240 and 1,000. If this finding can be replicated, it may give some insight into the minimum sample sizes required in order for the distance measures to work reasonably well. Of equal interest was the modest difference in the performance of the measures when there were 10 variables as opposed to 24. It appears that while having more variables does improve the ability of the cluster analysis to recover the solution, it may well be that in many contexts 10 variables is sufficient.

As with any research, there are weaknesses in this methodology which must be taken into account as the results are interpreted. First of all, only one clustering algorithm was used with the distance measures, which limits the findings to just cases in which Ward's method is used. In order to expand upon these results, other algorithms should be used. Furthermore, the distance measures selected for inclusion in this study are similar in terms of their calculation. Therefore, it would be worthwhile to pursue other measures of distance that treat the problem differently, such as Holley and Guilford's G index. Finally, it might be worthwhile to expand the parameters used in the 2PL model that simulated the data. For example, a larger difference between the two levels of the variance, or a sample size of less than 240 might shed more light on the performance limits of these measures

Bibliography

Anderberg, M. (1973). *Cluster Analysis for Applications*. New York: Academic Press,

Dillon, W. R. & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. New York: Wiley.

Donoghue, J.R. (1995). The Effects of Within-Group Covariance Structure on Recovery in Cluster Analysis I. The Bivariate Case. *Multivariate Behavioral Research*, 30(2), 227-254.

Hands, S. & Everitt, B. (1987). A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering Techniques. *Multivariate Behavioral Research*, 22, 235-243.

Lorr, M. (1983). Cluster Analysis for Social Scientists. San Francisco: Jossey – Bass,

Milligan, G.W. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika*, 45(3), 325-342.

Milligan, G.W. (1981). A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika*, 46(2), 187-199.

Milligan, G. W. and Cooper M. C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research*, 21, 441-458.

Overall, J.E., Gibson, J.M. & Novy, D.M. (1993). Population Recovery Capabilities of 35 Cluster Analysis Methods. *Journal of Clinical Psychology*, 49(4), 459-470.

Scheibler, D. & Schneider, W. (1985). Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms: A Comparison of Hierarchical and Nonhierarchical Methods. *Multivariate Behavioral Research*, 20, 283-304.

Snijders, T. A., Dormaar, M., van Schuur, W. H. , Dijkman-Caes, C. & Driessen, G. (1990). Distribution of Some Similarity Coefficients for Dyadic Binary Data in the Case of Associated Attributes. *Journal of Classification*, 7, 5-31.

Waller, N.G., Underhill J.M. & Kaiser H.A. (1999). A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape, Size, and Orientation. *Multivariate Behavioral Research*, 34(2), 123-142.

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparison of similarity measures in cluster Analysis with Binary Data

Author(s): Holmes Finch and Huynh Huynh

Corporate Source: University of south carolina	Publication Date: 4/24/00

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Holmes Finch	Printed Name/Position/Title: Holmes Finch / Stat Lab Managor	
Organization/Address: Dept. of Statistics, USC Columbia, S.C. 29208	Telephone: (803) 777-5074	FAX: (803) 777-4048
	E-Mail Address: Finch@Stat.Sc.edu	Date: 5/25/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

NA

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

NA

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to: