Journal of Data Science **3**(2005), 85-100

Comparison of Distance Measures in Cluster Analysis with Dichotomous Data

Holmes Finch
Ball State University

Abstract: The current study examines the performance of cluster analysis with dichotomous data using distance measures based on response pattern similarity. In many contexts, such as educational and psychological testing, cluster analysis is a useful means for exploring datasets and identifying underlying groups among individuals. However, standard approaches to cluster analysis assume that the variables used to group observations are continuous in nature. This paper focuses on four methods for calculating distance between individuals using dichotomous data, and the subsequent introduction of these distances to a clustering algorithm such as Ward's. The four methods in question, are potentially useful for practitioners because they are relatively easy to carry out using standard statistical software such as SAS and SPSS, and have been shown to have potential for correctly grouping observations based on dichotomous data. Results of both a simulation study and application to a set of binary survey responses show that three of the four measures behave similarly, and can yield correct cluster recovery rates of between 60% and 90%. Furthermore, these methods were found to work better, in nearly all cases, than using the raw data with Ward's clustering algorithm.

Key words: Cluster Analysis, dichotomous Data, distance measures.

1. Introduction

Cluster analysis (CA) is an analytic technique used to classify observations into a finite and, ideally, small number of groups based upon two or more variables. In some cases there are hypotheses regarding the number and make up of such groups, but more often there is little or no prior information concerning which individuals will be grouped together, 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 groups of observations. Distance in this context can be thought of in the Euclidean sense, or some other, comparable conceptualization (Johnson and Wichern, 1992).

One of the primary assumptions underlying these standard methods for calculating distance is that the variables used to classify individuals into groups are continuous in nature (Anderberg, 1973). However, some research situations, such as those involving testing data, may involve other types of variables, including ordinal or nominal. For example, in some situations, researchers are interested in grouping sets of test examinees based on their dichotomously scored responses (correct or incorrect) to individual test items, rather than on the total score for the exam, especially for identifying cases of answer copying (Wollack, 2002). The clustering of observations based on dichotomous variables can be readily extended beyond the realm of psychological testing to any situation in which the presence or absence of several traits are coded and researchers want to group the observations based on these binary variables. These could include economic analyses where individual firms are classified in terms of the presence or absence of various management practices, or situations where binary coding is used to describe industrial processes. In such situations, the standard Euclidean measures of distance are inappropriate for assessing the dissimilarity between two observations because the variables of interest are not continuous, and thus some alternative measure of separation must be used (Dillon and Goldstein, 1984). It is the goal of this paper to investigate four measures of distance designed for clustering using dichotomous data, and to compare their performance in correctly classifying individuals using simulated test data. A fifth approach, using the raw data rather than these distance measures, will also be included. The paper begins with a description of the four distance measures, followed by a discussion of the study design and the Monte Carlo simulation. Next, is the presentation and discussion of the results followed by a description of the implications for practitioners using dichotomous variables for clustering, and finally, weaknesses of the study.

2. Distance Measures for Dichotomous Variables

There are several techniques for conducting CA with binary data, all of which involve calculating distances between observations based upon the observed variables and then applying one of the standard CA algorithms to these distances. A popular group of these measures designed for binary data is known collectively as matching coefficients (Dillon and Goldstein, 1984). There are several types of matching coefficients, all of which take as their main goal the measurement of response set similarity between any two observations. The logic underlying these techniques is that two individuals should be viewed as similar to the degree that they share a common pattern of attributes among the binary variables (Snijders, Dormaar, van Schurr, Dijkman-Caes and Driessen, 1990). In other words, observations with more similar patterns of responses on the variables of interest are seen as closer to one another than are those with more disparate response patterns. An advantage of these measures is that they are easy to effect using available statistical software such as SAS or SPSS.

In order to discuss how these methods work, it is helpful to refer to an example. In this case, Table 1 below will be used to demonstrate how each of the four measures are calculated. The rows represent the presence or absence (1,0) of a set of K traits for a single observation i, and the columns represent the presence or absence of the same set of K traits for a second observation, j, where $i \neq j$.

Table 1: 2 × 2 response table

	Subject 2	
Subject 1	1	0
1	a	b
0	c	d

Cell a includes the count of the number of the K variables for which the two subjects both have the attribute present. In a testing context, having the attribute present would mean correctly answering the item. In turn, cell b represents the number of variables for which the first subject has the attribute present and the second subject does not, and cell c includes the number of variables for which the second subject has the attribute present and the first subject does not. Finally, cell d includes the count of the number of the K variables for which neither subject has the attribute present. The indices described below differ in the ways that they manipulate these cell counts. While there are a number of distance metrics available for dichotomous variables (Hands and Everitt, 1987), this paper will examine the 4 most widely discussed in the literature (Anderberg, 1973; Lorr, 1983; Dillon and Goldstein, 1984; Snijders, Dormaar, van Schurr, Dijkman-Caes and Driessen, 1990). It is recognized that other approaches are available, however in the interest of focusing this research on using methods that have been cited previously in the literature as being useful, and that are available to practitioners, only these four will be included in the current study. The first of these measures of distance is the Russell/Rao Index (Rao, 1948). It can be expressed in terms of the cells of Table 1 as:

$$\frac{1}{a+b+c+d} \tag{2.1}$$

This index is simply the proportion of cases in which both observations had the trait of interest. In contrast is the Jaccard coefficient, introduced by Sneath (1957), which has a similar structure but excludes cases from the denominator where neither subject has the trait of interest (cell d).

$$\frac{a}{a+b+c} \tag{2.2}$$

A third variation on this theme, called the matching coefficient (Sokal and Michener, 1958), includes both matched cells a and d: the number of cases where both subjects have both attributes present, and the number of cases where neither subject has the attributes present.

$$\frac{a+d}{a+b+c+d} \tag{2.3}$$

The final index to be examined here is Dice's coefficient (Dice, 1945). It is closely related to the Jaccard coefficient, with additional weight being given to cases of mutual agreement.

$$\frac{2a}{2a+b+c} \tag{2.4}$$

As an example of how these coefficients work, assume two observations, each of which have measurements for 7 binary variables where the presence of some trait is denoted by a 1 and its absence is denoted by a 0. (In the case where these variables represent responses to test items, a 1 would indicate a correct response, while a 0 would indicate an incorrect response). Example data appear in Table 2.

Table 2: Example data for two subjects on 7 variables

Subject	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 can be summed and placed in a format similar to that found in Table 1.

Table 3: Counts of response combinations for subjects 1 and 2

	Subject 2	
Subject 1	1	0
1	2	1
0	2	2

The data contained in Table 3 can be used to calculate values for the distance measures described above.

$$\begin{aligned} \text{Russell/Rao} &= \frac{2}{2+1+2+2} = \frac{2}{7} \\ \text{Jaccard} &= \frac{2}{2+1+2} = \frac{2}{5} \end{aligned}$$

$$\begin{aligned}\text{Matching} &= \frac{2+2}{2+1+2+2} = \frac{4}{7} \\ \text{Dice} &= \frac{2\cdot 2}{2\cdot 2+1+2} = \frac{4}{5}\end{aligned}$$

Distance is determined by taking the results of each calculation and subtracting them from 1. Thus, the largest distance value for these two subjects is associated with the Russell/Rao index, $1-2/7=3/7$, while the smallest distance is associated with the Matching and Dice coefficients, $1-4/7=3/7$. After distances are calculated for an entire set of data, they are combined into a matrix that is entered into a standard clustering algorithm such as Ward's (Ward, 1963).

3. Methodology

In order to compare the performance of these indices in terms of correctly grouping individuals, a set of Monte Carlo simulations were conducted under a variety of conditions, and the 4 distance measures, along with the raw data method, were applied to assess their performance. The data for this Monte Carlo study were generated using a 2-parameter logistic (2PL) model, which takes the following form:

$$P_i(\theta) = \frac{w^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

where $P_i(\theta)$ = probability of having the attributeon variable i, given the level of the latent trait, θ = level of the latent trait, D = scaling constant to make the logistic model match the normal ogive, a_i = ability of the particular attribute to discriminant among subjects with different levels of θ, b_i = the point on the latent trait scale where probability of having the attribute is 0.5, item difficulty.

This model, which is a standard way to express the probability of an examinee correctly answering a test item (see, for example, Lord, 1952; Birnbaum, 1968), links an underlying ability, θ, with the difficulty of the particular item, b, as well as its ability to differentiate among examinees of different abilities, a. In this context, θ represents the unobservable ability of an individual in a particular subject area. For example, if the test of interest measures mathematics knowledge, θ would be mathematics aptitude. The difficulty parameter, b, will be larger for items that are more difficult, while the discrimination parameter, a, will be larger for items that do a better job of discriminating between those individuals with more aptitude and those with relatively less aptitude. In the analysis of actual testing data, these parameters are estimated using Maximum Likelihood methods. However, for the purposes of this study, they were simply generated randomly, as described below. For the current set of simulations, the value of the latent variable, θ, was generated using a standard normal distribution, as was the b

parameter, while the a values were generated from the uniform (0,1) distribution. For each simulated dataset, a unique set of θ, a and b parameters were generated for each simulated individual and item; i.e., for each simulation there is a unique set of item parameters for all k items and j individuals: a $(a_1, a_2, \ldots, a_k))$ and $b, (b_1, b_2, \ldots, b_k)$ and $\theta, (\theta_1, \theta_2, \ldots, \theta_k)$. The resulting probability of having the attribute was then compared to a randomly generated value from the uniform (0,1) distribution, with the resulting dichotomous variable being assigned a 1 if the probability from the 2PL model was larger than the uniform variate, and a 0 otherwise. Two groups, or clusters, of subjects were simulated with the difference between them being the mean value of the latent trait; i.e., one group was associated with a high value while the other was associated with a low value. There were two sets of these latent variables corresponding to cases of high group separation and low group separation, with high group separation having means of either -2 or 2 and low group separation having means of either -0.5 or 0.5. Other conditions which were varied were the variance of the latent trait (1.5 and 0.5), the number of variables (10 and 24) and the number of subjects (240 and 1000). It should be noted that the two clusters always had an equal number of subjects, and all combinations of simulation condition were crossed with one another.

Cluster analysis was conducted using the four distance measures described above, in addition to the raw dichotomous data with no distance measure applied. Ward's method of cluster extraction was selected for use in this study based upon results indicating that of the major clustering methods, it often performs the best at population recovery of clusters (Kuiper and Fisher, 1975; Blashfield, 1976; Overall, Gibson and Novy, 1993). In addition, Hands and Everitt (1987) found that it performed the best at cluster extraction when used in conjunction with distance measures of this type. For each simulation, the distance measures were calculated and the resulting matrices were submitted to the Ward's method using the SAS software system. In the case of the raw data, Ward's method of cluster extraction was applied directly to the dichotomous variables themselves, with no distance measures calculated. The results of the cluster analyses were compared using the percent of cases correctly grouped together, and the Kappa coefficient of agreement, which takes the form:

$$K = \frac{P_0 - P_E}{1 - P_E},$$

where P_0 = proportion of variables with attribute for both subjects in a pair, $P_E = \sum_i P_{i\cdot} P_{\cdot i}$, expected proportion of variables with attribute for both subjects in a pair, P_{ii} = probability that a given subject will be classified win cluster i and they should be in cluster i, $P_{i\cdot}$ = probability of being in cluster i in the population, $P_{\cdot i}$ = probability of being put into cluster i by the cluster analysis.

Various authors have discussed the interpretation of Kappa (Landis and Koch, 1977; Fleiss, 1981; Gardner, 1995) with some consensus that values less than 0.4 indicate low agreement, and values above 0.7 indicating relatively high agreement. In the context of this study, high agreement corresponds to the clustering solution matching the actual grouping of subjects. In addition to using simulations, a set of real dichotomous data, taken from the National Center for Education Statistics was also cluster analyzed using these five methods. The data are part of the Early Childhood Longitudinal Study, and pertain to teacher ratings of student aptitude in a variety of areas, along with actual achievement scores in reading, math and general knowledge. In this case, 22 questions asked of teachers were used to cluster a sample of 500 first graders. Each of these survey items asked the teacher to rate a specific child on some type of academic performance, such as their ability to clearly compose a story, read words with irregular vowels, group living and non-living things and understand whole numbers. These items were dichotomized into categories representing performance at an intermediate or proficient level, or performance that is still in progress of developing or lower. In order to assess the success of the various approaches, the resulting clusters were examined using student performance on reading, mathematics and general knowledge achievement tests. It is hypothesized that the teacher ratings of student aptitude should result in clusters that clearly demarcate students by performance on these achievement tests.

4. Results

4.1 Results of simulation study

Across all treatment conditions, the percent correctly classified and the Kappa coefficient values are very similar for all five methods, as can be seen in Table 4. Based upon the standards for the Kappa value described above, it would appear that overall, all five approaches to clustering have moderate agreement across all of the study conditions. Indeed, they correctly group just under three quarters of the observations.

Table 4: Kappa and percent correctly classified by distance measure

Distance	Kappa	Percent Correctly Grouped
Dice	0.471	0.735
Jaccard	0.473	0.736
Matching	0.466	0.733
Russell/Rao	0.467	0.734
Raw Data	0.465	0.732

Table 5: Kappa / percent correctly grouped by measure and levels of manipulated variables

	Level of Group Separation	
Distance	Low (0.5/-0.5)	High (2.0/-2.0)
Dice	0.137 / 0.628	0.793 / 0.906
Jaccard	0.139 / 0.637	0.796 / 0.909
Matching	0.124 / 0.568	0.797 / 0.911
Russell/Rao	0.134 / 0.614	0.792 / 0.905
Raw Data	0.103 / 0.472	0.716 / 0.818
	Variance of Latent Variable	
	0.5	1.5
Dice	0.486 / 0.744	0.456 / 0.741
Jaccard	0.489 / 0.749	0.457 / 0.743
Matching	0.485 / 0.743	0.447 / 0.727
Russell/Rao	0.486 / 0.744	0.449 / 0.730
Raw Data	0.543 / 0.832	0.421 / 0.685
	Number of Variables	
	10	24
Dice	0.431 / 0.710	0.513 / 0.776
Jaccard	0.434 / 0.715	0.514 / 0.777
Matching	0.431 / 0.710	0.503 / 0.761
Russell/Rao	0.432 / 0.712	0.504 / 0.762
Raw Data	0.445 / 0.733	0.486 / 0.735
	Number of Subjects	
	240	1000
Dice	0.463 / 0.735	0.478 / 0.748
Jaccard	0.467 / 0.742	0.479 / 0.750
Matching	0.460 / 0.731	0.471 / 0.737
Russell/Rao	0.458 / 0.727	0.476 / 0.745
Raw Data	0.442 / 0.687	0.490 / 0.767

Because the pattern for the Kappa coefficient and the percent correctly grouped is virtually identical for the five methods used in this paper, the results of Kappa are emphasized in subsequent discussion, though the percent correctly grouped is also presented. This choice is made based on the fact that researchers have identified relevant cut off values for Kappa (i.e., what values represent high agreement and what values represent low agreement), making it easier to interpret and generalize than the percent correctly classified. The effect of differences in classification accuracy due to the manipulated study variables, including group separation, variance of the latent variable, number of variables and number of subjects are shown in Table 5.

It appears that the four measures of distance all have similar Kappa values at both levels of group separation, with better performance associated with greater separation, as would be expected. The clustering algorithm is somewhat less accurate when using raw data rather than the distance measures, at both levels of group separation. Similar patterns of performance can be seen among the four distance measures across all treatment conditions. In general, they all have higher Kappa values when the variance of the latent trait is lower, when there are more variables being measured and when there are more subjects included in the sample, though in this latter case the difference from smaller to larger sample sizes is very minor. It is interesting to note that the raw data approach works better than all four distance measures when the variance in the latent trait is low and when the sample size is large. In addition, the impact of increased variance appears to be much greater for the raw data method, as is evidenced by the much sharper decline in Kappa and percent correctly grouped from low to high variance, when compared with the four distance measures. It should be noted that potential interactions among the manipulated variables were examined, and none were found.

4.2 Results of real data analysis

The results of the cluster analysis for each measure are interpreted using both the pseudo F (Calinski and Harabasz, 1974) and pseudo T^2 (Duda and Hart, 1973) measures of cluster fit provided by SAS. These statistics have been identified in simulations (Milligan and Cooper, 1985) to be accurate for identifying the number of clusters present in a dataset, and to be robust in cases where a high degree of error appears in the data. There has been more recent research examining appropriate indices for ascertaining the number of clusters to retain for a set of binary data (Dimitriadou, Dolincar and Weingessel, 2002). However, the decision was made with respect to this study, to use a method that has been proven useful in the past, and that is easily accessed by practitioners using standard software, such as SAS. In general, the number of clusters is determined by examining the pattern of change for both statistics. One should look for a local maximum of the pseudo F statistic, to be accompanied by a local minimum of the pseudo T^2. Using this heuristic, it appears that each clustering approach found three distinct groups of first graders in the data, based upon teachers' ratings of their academic performance. In order to better understand the nature of these clusters, the mean of reading, math and general knowledge test scores taken by the children in the spring of first grade are calculated and appear in

Table 6. Based on these means, it appears that for each of the five approaches, the clustering algorithm finds groups of high, medium and low achieving children. Interestingly, however, the degree of group separation is not uniform across methods. For example, it appears that the Jaccard, Russell/Rao and Dice methods all result in three clearly defined groups based on the test scores. On the other hand, the Matching coefficient cannot seem to distinguish between the highest achieving group, and one that achieves in the middle. It does, however, clearly group the lowest achieving students. Finally, the raw data approach does appear to successfully differentiate the groups based on the reading and math scores, but not general knowledge.

Table 6: Mean scores on achievement tests by distance measure and cluster type

Cluster Type	Reading	Math	General Knowledge
Jaccard			
High	54.57	55.15	54.57
Medium	48.76	46.44	47.06
Low	39.88	34.84	37.67
Russell/Rao			
High	54.80	55.59	54.60
Medium	50.23	48.70	49.85
Low	40.94	35.74	38.12
Dice			
High	54.59	55.37	54.60
Medium	49.65	47.14	48.16
Low	39.49	34.73	37.16
Matching			
High	55.10	55.96	54.97
Medium	53.93	52.02	52.32
Low	45.19	42.01	43.68
Raw data			
High	57.47	57.07	55.56
Medium	52.37	53.84	53.71
Low	45.19	42.01	43.69

Table 7 displays the percent of observations that are grouped in the same clusters by pairs of the clustering methods. It appears that there is high agreement in terms of case clustering for the Jaccard, Russell/Rao and Dice indices.

On the other hand, the groupings created by these three distance measures have somewhat lower levels of agreement with the Matching coefficient, and much lower agreement rates with the raw data. However, the Matching coefficient and raw data approaches appear to have created clusters that are fairly similar, with agreement between the two at 82.6%.

Table 7: Percent agreement in clustering observations

	Jaccard	Russell/Rao	Dice	Matching	Raw data
Jaccard		91.4	96.0	68.6	53.2
Russell/Rao			94.2	74.6	58.6
Dice				71.2	54.6
Matching					82.6
Raw data					

5. Discussion

As was discussed above, previous work has not been done comparing these indices using Monte Carlo methods, which has left a gap in the literature in terms of assessing their performance under a variety of conditions. Given the results described herein, it appears that under the conditions present in this study, the four measures of distance perform very much the same in terms of correctly classifying simulated observations into two clusters based on a set of dichotomous variables. In turn, the use of the raw data is associated with somewhat lower accuracy unless the sample size is large or the variation of the underlying latent trait is low. The similarity in performance of the four distance measures appears to hold true regardless of the level of group separation, the variation in the underlying latent variable, the number of variables included in the study and the size of the sample. Across all measures, the clustering solutions are more accurate for greater group separation, lower variance in the latent trait, more variables and a larger sample size. With respect to the real data, the five methods did not have perfect agreement in terms of clustering observations. While they all found three clusters, the use of the Jaccard, Russell/Rao and Dice coefficients resulted in very similar solutions, with clusters that were clearly differentiated based on the achievement measures. In contrast, the matching and raw data approaches yielded somewhat different results with less well defined groups.

The relative dearth in previous research of this type leaves little in the way for comparison of these results with comparable ones. However, there has been a small amount of discussion regarding the expected performance of these indices, given their conceptual bases. Hall (1969) made the point that the Dice and

Jaccard coefficients should differ because in the former index, mismatches (the 0,1 or 1,0 cases) lie halfway between matches, either 0,0 or 1,1, in terms of importance, while in the latter they are of equal weight to the 1,1 category. The fact that in the current study these two indices performed very similarly suggests that this distinction is not very important in terms of classification. In other words, giving mismatches equal weight with the 1,1 case does not seem to detract from the ability of cluster analysis to correctly group observations as compared to increasing the weight of the 1,1 agreement. Anderberg (1973) argues that including the 0,0 case should not provide any useful information in terms of classification, and in fact can be misleading. The current research appears to support this assertion, at least in terms of the relatively poor performance of the matching index versus the Jaccard, Dice and Russell/Rao alternatives. It is interesting to note that the latter of these three, which uses the 0,0 category in the denominator but not the numerator does perform as well as either Jaccard and Dice, both of which ignore the matched absence (0,0) category altogether. Therefore, it appears that if inclusion of the 0,0 category in calculating distance does lead to a diminution in the performance of the clustering algorithm, it only does so when this category in included in the numerator and denominator and not the denominator only.

Given the similarities in the structure of the distance measures, it is not totally surprising that they would perform fairly similarly in terms of clustering accuracy. Indeed, the major difference among them is with respect to their handling of an absence of a particular trait for both members of an observation pair. Both the Russell/Rao and matching coefficients include this category in their calculations, though they deal with it somewhat differently. The results presented here would suggest that the number of cases where neither pair has the attribute of interest across the set of variables does not contain useful information in terms of clustering, regardless of how it is handled. Furthermore, the fact that the results for Dice's coefficient (which multiplies the count of joint agreement by 2) are similar to those for the Jaccard index would suggest that the amount of emphasis placed on the number of variables where subjects both exhibit the attribute is not important either. In short, it seems that the simplest of the four measures of distance, the Jaccard index, works as well as its more complicated competitors in terms of correctly grouping individuals based on a set of dichotomous variables.

Another finding of interest in this study is that the use of raw dichotomous data in clustering does not always lead to a worse solution than that obtained by using formal measures of distance, and in some cases may actually work better. In the worst instance, clustering with raw data results in a rate of correct grouping approximately 8

Across all five clustering methods used in this study, the factor that influences the ability to correctly group subjects is the degree of cluster separation. Among the other factors manipulated in this study, none have a particularly dramatic impact on the ability of the procedures to correctly cluster. Indeed, while there is a slight increase in correct grouping for lower variance, more variables and a larger sample size, none of these conditions results in an improvement of more than about 6% in correct classification versus the alternative level for that variable. An interesting implication of these results is that in general, clustering with these measures will work well for samples as small as 240, and for as few as 10 variables. These results may give some insight into the minimum dataset size at which these procedures work properly, particularly if they can be replicated.

The analysis of the real data found that the Jaccard, Russell/Rao and Dice measures are largely in agreement in terms of grouping individuals, and appear to do a better job of finding distinct clusters in the data than either the Matching or raw data approaches. While it is recognized that these results may only be applicable to this dataset, they do reinforce both the apparent lack of importance of cases where neither observation displays the trait, and the slightly better overall performance of distance measure based clustering as opposed to that based on the raw dichotomous data.

6. Implications for Practitioners and Suggestions for Further Research

There are some interesting implications for those using dichotomous data to cluster individual observations. First of all, the methods described here, all of which are intuitively simple and practically easy to carry out using standard statistical software, typically yield successful clustering rates of around 75% for the simulated data, with the best performance coming with greatest group separation, in which case it can be expected that over 90% of the cases will be correctly grouped together. These results would support the notion that cluster analysis of dichotomous data using these approaches is appropriate, and can be expected to work reasonably well.

Furthermore, the results described herein, particularly those based on the real data, indicate that the Jaccard, Dice and Russell/Rao approaches are all comparable. In addition, it should be noted that using the raw data might be acceptable in some cases, especially when the sample size is large. However, given the different solutions obtained by the raw data with the real data analysis, it seems that further work would need to be conducted before a final decision in this regard is possible.

As noted above, the sample size and number of items did not have a great impact on the ability of these approaches to cluster individuals. These results should help practitioners to define a reasonable size for using dichotomous variables to cluster individuals. Clearly, the payoff for greatly increasing the scope of a study in terms of the sample size and the number of variables is fairly small in terms of clustering accuracy.

Finally, it seems clear that when group separation on the latent trait is relatively low, the Jaccard, Dice and Russell/Rao measures work similarly, and better, than the Matching coefficient or raw data. Indeed, even in the case of low group separation, these three measures are able to correctly cluster over 60% of the subjects.

As with any research, there are weaknesses in this study which should be taken into account as the results are interpreted. First of all, only one clustering algorithm, Ward's, was used. In order to expand upon these results, the data could be replicated and other clustering algorithms used. In addition, the distance measures selected for inclusion in this study are of a particular class, albeit one identified by several authors as among the most useful for clustering with dichotomous data. Therefore, it would be worthwhile to compare these approaches to other measures of distance that are calculated fundamentally 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 latent trait variance could be used, or a sample size of less than 240, so that a true lower bound for adequate performance for this variable could be identified.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (Edited by F.M. Lord and M.R. Novick), chapters 17-20. Addison-Wesley.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin* **83**, 377-388.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1-27.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26**, 297-302.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications.* John Wiley and Sons.

Dimitriadou, E., Dolnicar, S. and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika* **67**, 137-160.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* John Wiley and Sons.

Fleiss, J. L. (1981). *Stastical Methods for Rates and Proportions, 2nd edition.* John Wiley and Sons.

Gardner, W. (1995). On the reliability of sequential data: Measurement meaning and correction, In *The Analysis of Change* (Edited by John M. Gottman). Erlbaum.

Hall, A. V. (1969). Avoiding informational distortions in automatic grouping programs. *Systematic Zoology* **18**, 318-329.

Hands, S. and 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.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Kuiper, F. K., and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**, 777-783.

Landis, J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

Lord, F. M. (1953). *A Theory of Test Scores* (Psychometric Monograph, No. 7). Iowa City, IA: Psychometric Society.

Lorr, M. (1983). *Cluster Analysis for Social Scientists.* Jossey-Bass.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179.

Overall, J. E., Gibson, J. M. and Novy, D. M. (1993). Population recovery capabilities of 35 cluster analysis methods. *Journal of Clinical Recovery* **49**, 459-470.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B* **10**, 159-193.

Sneath, P. H. A. (1957). Some thoughts on bacterial classification. *Journal of General Microbiology* **17**, 184-200.

Snijders, T. A., Dormar, M., van Schurr, W. H., Dijkman-Caes, C. and Driessen, G. (1990). Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes. *Journal of Classification* **7**, 5-31.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**, 1409-1438.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236-244.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Analysis* **5**, 329-350.

Wollack, J. A. (2002). Comparison of answer copying indices with real data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Received October 19, 2003; accepted December 15, 2003.

Holmes Finch
Department of Educational Psychology
Ball State University
Muncie, IN 47304
whfinch@bsu.edu