Journal of Classification 25:177-183 (2008)
DOI: 10.1007/s00357-008-9023-7

On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index

Matthijs J. Warrens

Leiden University, The Netherlands

Abstract: It is shown that one can calculate the Hubert-Arabie adjusted Rand index by first forming the fourfold contingency table counting the number of pairs of objects that were placed in the same cluster in both partitions, in the same cluster in one partition but in different clusters in the other partition, and in different clusters in both, and then computing Cohen's κ on this fourfold table.

Keywords: Correction for chance agreement; Partitions; Clustering method; Matching table; Simple matching coefficient; Similarity indices; Resemblance measures.

1. Introduction

It is shown in this note that two resemblance measures used in different realms of research are in fact equivalent. The first index is used in cluster analysis for comparing two partitions obtained from different clustering methods. There seems to be some agreement in the cluster community that the preferred measure for comparing two partitions is the Hubert-Arabie adjusted Rand index, proposed by Hubert and Arabie (1985) (see, for example, Steinley 2004). The second measure is often used in psychological research to assess the agreement of judgments made by two observers, and is known as the kappa statistic proposed by Cohen (1960). Cohen's κ is appropriate

The author thanks Willem Heiser, Mark de Rooij, Marian Hickendorff and three anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this article.

Author's Address: Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands, e-mail: warrens@fsw.leidenuniv.nl.

Published online 6 December 2008

for testing whether agreement exceeds chance levels for binary and nominal ratings.

One step in comparing two clusterings, is to obtain the 2×2 contingency matrix summarizing the matching table of two clusterings. A similar fourfold table may be obtained in psychological research as the cross classification of judgments by two observers on the presence or absence of a trait. It turns out that the same coefficient value is obtained when one applies either Cohen's κ or the Hubert-Arabie adjusted Rand index to these types of fourfold tables.

2. Comparing Two Partitions

In cluster analysis, one may be interested in comparing two clustering methods (Rand 1971; Fowlkes and Mallows 1983; Hubert and Arabie 1985; Steinley 2004; Albatineh, Niewiadomska-Bugaj and Mihalko 2006). Suppose we have two partitions of m data points. To compare these two clusterings, a first step is to obtain a so-called matching table $\mathcal{M} = \{m_{ij}\}$, where m_{ij} indicates the number of data points placed in cluster i ($i = 1, 2, ..., I$) according to the first clustering method and in cluster j ($j = 1, 2, ..., J$) according to the second method. The total number of points being clustered is given by $m = \sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij}$. The cluster sizes in the two clusterings considered are the row and column totals of table $\mathcal{M}$, given by

$$m_{i+} = \sum_{j=1}^{J} m_{ij} \quad \text{and} \quad m_{+j} = \sum_{i=1}^{I} m_{ij}.$$

Furthermore, we define the quantity

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{m_{ij}}{2} = \frac{1}{2} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij}^2 - m \right],$$

which equals the number of object pairs that were placed in the same cluster according to both clustering methods, and the three quantities

$$P = \sum_{i=1}^{I} \binom{m_{i+}}{2}, \quad Q = \sum_{j=1}^{J} \binom{m_{+j}}{2} \quad \text{and} \quad N = \binom{m}{2}.$$

The quantity N equals the total number of pairs of objects given m points.

As a second step, one may calculate some sort of resemblance measure that summarizes the information in table $\mathcal{M}$. A well-known measure for the similarity of two partitions is the Rand index (Rand 1971), given by

$$\mathrm{R} = \frac{N + 2T - P - Q}{N}.$$

The Rand index may be adjusted for agreement due to chance (Cohen 1960; Hubert and Arabie 1985; Albatineh et al. 2006). In general, a similarity index S after such correction has a form

$$\text{AS} = \frac{\text{S} - E(\text{S})}{\max(\text{S}) - E(\text{S})}. \tag{1}$$

The expectation $E(\text{S})$ in (1) is conditional upon fixed sets of marginal numbers corresponding to the table of which S is the summary index. The quantity $\max(\text{S})$ in (1) is the maximum value of index S regardless of the marginal numbers (Hubert and Arabie 1985).

Fowlkes and Mallows (1983) and Hubert and Arabie (1985, p. 197) note that, under the generalized hypergeometric assumption with respect to table $\mathcal{M}$, the expectation $E(T)$ is given by

$$E\left(T\right) = \frac{PQ}{N}. \tag{2}$$

Using (2), the expectation $E(\text{R})$ corresponding to the Rand index is given by

$$E(\text{R}) = 1 + \frac{2PQ}{N^2} - \frac{P+Q}{N} \qquad \text{(Hubert and Arabie 1985, p. 198).}$$

Using R and $E(\text{R})$ in (1), we obtain the Hubert-Arabie adjusted Rand index (Hubert and Arabie 1985, p. 198), which is given by

$$\text{AR} = \frac{T - PQ/N}{\frac{1}{2}(P+Q) - PQ/N} = \frac{2(NT - PQ)}{N(P+Q) - 2PQ}.$$

3. Reformulation of the Rand Index

As noted in, for example, Steinley (2004) or Albatineh et al. (2006), the information in a matching table $\mathcal{M}$ of two clustering partitions on the same data points, can be summarized in a fourfold table like Table 1. In Table 1, a is the number of object pairs that were placed in the same cluster according to both clustering methods, b (c) is the number of pairs that were placed in the same cluster according to one method but not according to the other, and d is the number of pairs that were not in the same cluster according to either of the methods. It then holds that

$$a + b + c + d = N \tag{3}$$

where $a = T$, $b = P - T$, $c = Q - T$ and $d = N + T - P - Q$, and $p_1 = a + b = P$ and $q_1 = c + d = N - P$. The four different types of object

Table 1. 2×2 Contingency Table Representation of a Matching Table $\mathcal{M}$.

	Second partition		
First partition	Pair in same cluster	Pair in different cluster	Total
Pair in same cluster	a	b	p_1
Pair in different cluster	c	d	q_1
Total	p_2	q_2	N

pairs are also distinguished in Hubert and Arabie (1985, p. 194). However, these authors expressed their formulas in terms T, P, Q, and N, instead of the quantities a, b, c, and d.

The information in Table 1 can be summarized by some sort of resemblance index or similarity coefficient, and a vast amount of formulas can be found in the classification literature (see, for example, one of the following papers that have appeared in the Journal of Classification: Gower and Legendre 1986; Baulieu 1989; Batagelj and Bren 1995; Albatineh et al. 2006). A well-known resemblance measure is the simple matching coefficient (Sokal and Michener 1958), given by

$$\text{SM} = \frac{a+d}{a+b+c+d} = \frac{a+d}{N}. \tag{4}$$

Expressing the Rand index in terms of the quantities a, b, c, and d we obtain the formula in (4). Thus, if the Rand index is formulated in terms of the quantities a, b, c, and d, it is equivalent to the simple matching coefficient (see, for example, Steinley 2004, or Albatineh et al. 2006).

4. Similarity Between Ratings

In psychological research, one may be interested in the degree of consensus between two observers who rate each of a sample of subjects on a nominal scale. Table 1 can be obtained as a cross classification of judgments by two observers on the presence or absence of a trait: a is the number of times a trait was present according to both observers, b (c) is the number of times a trait was present according to one observer but not according to the other, and d is the number of subjects for which a trait was absent according to both observers.

One may want to calculate a resemblance measure that summarizes the information in Table 1. A possible resemblance measure is the simple matching coefficient in (4). However, a coefficient that is often used for rater data, is the kappa statistic introduced by Cohen (1960). Cohen (1960)

considered correction for agreement due to chance for the simple matching coefficient. Cohen assumed that the data are a product of chance of two different frequency distributions underlying the rows and columns of Table 1. The expectation of the quantity a in Table 1 is then the product of the marginals corresponding to a, p_1 and p_2, divided by N, that is, $E(a) = p_1 p_2 / N$. The complete case of statistical independence is presented in Table 2.

The expectation $E(\text{SM})$ corresponding to the simple matching coefficient given independence, is given by

$$E(\text{SM}) = E\left(\frac{a+d}{N}\right) = \frac{E(a+d)}{N} = \frac{p_1 p_2}{N^2} + \frac{q_1 q_2}{N^2}.$$

Expectation $E(\text{SM})$ can be obtained by considering all permutations of the observations of one of the two raters, while preserving the order of the observations of the other rater. For each permutation the value of SM can be determined. The arithmetic mean of these values is $(p_1 p_2 + q_1 q_2)/N^2$.

Using SM and $E(\text{SM})$ in (1), we obtain

$$\text{Cohen's } \kappa = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}. \tag{5}$$

5. Equivalence of AR and Cohen's κ

Similar to the Rand index, the adjusted Rand index may be expressed in terms of the quantities a, b, c, and d. Expressing the Hubert-Arabie adjusted Rand index in these quantities, we obtain, following Steinley (2004, p. 388), the formula

$$\text{AR} = \frac{N(a+d) - [(a+b)(a+c) + (b+d)(c+d)]}{N^2 - [(a+b)(a+c) + (b+d)(c+d)]}. \tag{6}$$

Using (3), we have the equalities

$$\begin{aligned} &(a+b)(a+c) = a(a+b+c) + bc = a(N-d) + bc \\ \text{and} \quad &(b+d)(c+d) = d(b+c+d) + bc = d(N-a) + bc. \end{aligned}$$

Using these equalities, the numerator of (6) can be written as

$$\begin{aligned} &N(a+d) - [(a+b)(a+c) + (b+d)(c+d)] \\ &= Na + Nd - a(N-d) - d(N-a) - 2bc \\ &= 2(ad - bc). \end{aligned}$$

Table 2. The Expected Values of Quantities a, b, c, and d in Table 1 under the Assumption That the Data are a Product of Chance of Two Different Distribution Functions.

	Second partition		
First partition	Pair in same cluster	Pair in different cluster	Total
Pair in same cluster	p_1p_2/N	p_1q_2/N	p_1
Pair in different cluster	p_2q_1/N	q_1q_2/N	q_1
Total	p_2	q_2	N

Furthermore, the denominator of (6) equals

$$\begin{aligned} & N^2 - [(a+b)(a+c) + (b+d)(c+d)] \\ &= N^2 - p_1p_2 - q_1q_2 \\ &= (p_1+q_1)(p_2+q_2) - p_1p_2 - q_1q_2 \\ &= p_1q_2 + p_2q_1. \end{aligned}$$

Hence, Equation (6) is equivalent to Equation (5). Thus, if the adjusted Rand index is formulated in terms of the quantities a, b, c, and d, it is equivalent to Cohen's κ. Moreover, expectation $E(T)$ in (2) can be written as

$$E(T) = \frac{PQ}{N} = \frac{(a+b)(a+c)}{N} = \frac{p_1p_2}{N} = E(a).$$

Hence, statistical independence under the generalized hypergeometric distribution function used in Hubert and Arabie (1985) for the matching table of two partitions, is equivalent to the case of statistical independence under the binomial distribution function for the fourfold contingency table.

6. Discussion

A practical conclusion is that we can calculate the Hubert-Arabie adjusted Rand index by first forming the fourfold contingency table counting the number of pairs of objects that were placed in the same cluster in both partitions, in the same cluster in one partition but in different clusters in the other partition, and in different clusters in both, and then computing Cohen's κ on this fourfold table (compare Saltstone and Stange 1996, p. 171).

As pointed out by a referee, it is important to note that, although the formulas of the Rand index and the simple matching coefficient, and the formulas of the Hubert-Arabie adjusted Rand index and Cohen's κ, are the same, the inputs are quite different. The simple matching coefficient and Cohen's κ operate on singletons and require the labels of the clusters for

matching the partitions, or the labels for rating, to be known. However, the Rand index and the adjusted Rand index are based on pairs of observations that appear in the same cluster in each of the partitions. Steinley (2004), for example, considers this one of the biggest advantages of the adjusted Rand index over other measures that require some knowledge about the labeling.

References

ALBATINEH, A.N., NIEWIADOMSKA-BUGAJ, M., and MIHALKO, D. (2006), "On Similarity Indices and Correction for Chance Agreement," *Journal of Classification, 23*, 301-313.

BATAGELJ, V., and BREN, M. (1995), "Comparing Resemblance Measures," *Journal of Classification, 12*, 73-90.

BAULIEU, F.B. (1989), "A Classification of Presence/Absence Based Dissimilarity Coefficients," *Journal of Classification, 6*, 233-246.

COHEN, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement, 20*, 37-46.

FOWLKES, E.B., and MALLOWS, C. L. (1983), "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association, 78*, 553-569.

GOWER, J.C., and LEGENDRE, P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification, 3*, 5-48.

HUBERT, L.J., and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification, 2*, 193-218.

RAND, W.M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association, 66*, 846-850.

SALTSTONE, R., and STANGE, K. (1996), "A Computer Program to Calculate Hubert and Arabie's Adjusted Rand Index," *Journal of Classification, 13*, 169-172.

SOKAL, R.R., and MICHENER, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin, 38*, 1409-1438.

STEINLEY, D. (2004), "Properties of the Hubert-Arabie Adjusted Rand Index," *Psychological Methods, 9*, 386-396.