Some Measures of Agreement Between Close Partitions

Genane Youness, Gilbert Saporta

Student (2004), Vol. 5 No. 1, 1-12

Some Measures of Agreement Between Close Partitions

Genane Youness

CEDRIC CNAM
BP 11 - 4661
Beirut, Lebanon

Gilbert Saporta

Chaire de Statistique Appliquée
CEDRIC, CNAM, 292 rue Saint Martin
75003 Paris, France

Abstract: In order to measure the similarity between two partitions coming from the same data set, we study extensions of the RV-coefficient, the kappa coefficient proposed by Cohen (in case of partitions with same number of classes), and the D2 coefficient proposed by Popping. We find that the RV-coefficient is identical to the Janson and Vegelius index. We compare the result coming from kappa's coefficient to the ordination given by correspondence analysis. We study the empirical distribution of these indices under the hypotheses of a common partition. For this purpose, we use data coming from a latent profile model to formulate the null hypothesis.

Key words : RV, Janson and Vegelius index, Cohen's Kappa, Popping' D2, Correspondence analysis, latent class

1 Introduction

The problem of measuring the agreement between two partitions of a same data set has attracted interest and reappears continually in the literature of classification.

The most useful agreement measured proposed by Rand (1971) has been rediscovered and modified by [2] . Based on the comparison of object triples, a measure of partition correspondence was introduced by [4] . A generalized Rand-index method for consensus clustering was proposed by [3] for finding an amalgamated clustering

of a set of contributory partitions. By using mathematical and statistical aspects in the standardization of the comparison coefficients, [8] shows how to take into account the relational constraint, which results from the partition structure.

A simple way to compare partitions is to study the empirical distribution of a measure of agreement under some null hypothesis. It is necessary to define measures of agreement between partitions, before testing the agreement. In previous papers [13, 14, 15] , we have examined the following indices: Rand, Mc Nemar, and Jaccard. In this paper, we present new approaches to find the similarity between clustering. The *RV*-coefficient introduced by Robert and Escoufier [12] is proposed and written in term of paired comparisons. We prove that *RV*-coefficient is identical to the JV index [9] .

The kappa coefficient proposed by Cohen [1] is another way to measure the agreement between partitions with equal number of classes. Kappa can be used only in situations where categories are specified in advance, which is not the case of partitions where the labels of the classes are arbitrary: for that, we identify the permutation of classes of one of the partitions by maximizing the kappa values. We compare the optimal permutation with the ranking given by correspondence analysis: the results are the same.

Finally we also study the D_2 index proposed by Popping [10] . This less known index is based on comparison of pairs of subjects. We find the empirical distribution of this agreement measure when the partitions only differ by chance.

2 Notations

Let P_1 and P_2 be two partitions (or two categorical variables) of the same subjects with p and q classes. If K_1 and K_2 are the disjunctives $n \times p$ and $n \times q$ tables and N the corresponding contingency table with elements n_{ij}, we have: $N = K_1'K_2$.

Each partition P_k is also characterized by the $n \times n$ paired comparison table C^k with general term $c^k_{ii'}$:

$$c^k_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are in the same class of } P_k \\ 0 & \text{otherwise} \end{cases}$$

Note that $c^k_{ii} = 1$, for every $i \in \mathbb{N}$. We have $C_1 = K_1K_1'$ and $C_2 = K_2K_2'$. Given n subjects, $n(n-1)/2$ pairs of subjects can be compared. When both partitions assign pairs of subjects to the same classes, we consider this as an agreement.

$P_1 \backslash P_2$	**Same class**	**Different class**
Same class	**a** Agreement (same)	**c** Disagreement
Different class	**d** Disagreement	**b** Agreement (Different)

Table 1: The four cases

3 RV-coefficient

Robert and Escoufier [12] have derived a measure of similarity of two configurations, called RV-coefficient or the vector correlation coefficient; it allows to measure the similarity between two data tables X_1 and X_2 of the same observations by comparing the scalar product between individuals associated to the two tables. If the scalar matrices product $X_i X_i'$ is noted as W_i with dimension $n \times n$, the RV-coefficient is defined as:

$$RV(X_1, X_2) = \frac{\text{tr } (W_1 W_2)}{\text{tr } (W_1^2)\text{tr } (W_2^2)}.$$

When applied to the disjunctive tables associated to two partitions, we find:

$$RV(P_1, P_2) = \frac{\text{tr } (C^1 C^2)}{\text{tr } (C^1)^2 \text{tr } (C^2)^2} = \frac{\sum_{i \neq i'} \left(c^1_{ii'} - \frac{1}{p}\right)\left(c^2_{ii'} - \frac{1}{q}\right)}{\sqrt{\sum_{i \neq i'} \left(c^1_{ii'} - \frac{1}{p}\right)^2 \sum_{i \neq i'} \left(c^2_{ii'} - \frac{1}{q}\right)^2}}. \quad (1)$$

A high value of RV may lead to the conclusions that partitions are almost identical.

Lazraq, A. and Cleroux, R. [6] have proposed a test for the null hypothesis that the theoretical (population) RV is zero, but only in the case of numerical data, which cannot be applied to indicator variables.

3.1 JV index

The JV index is a measure of association proposed by Janson and Vegelius [9] . Its initial expression is:

$$JV(P_1, P_2) = \frac{pq \sum_u \sum_v n^2_{uv} - p \sum_u n^2_{u\cdot} - q \sum_v n^2_{\cdot v} + n^2}{\sqrt{[p(p-2) \sum_u n^2_{u\cdot} + n^2]\,[q(q-2) \sum_v n^2_{\cdot v} + n^2]}}.$$

It has been proved that, in terms of paired comparisons:

$$JV(P1, P2) = \frac{\text{tr } (C^1 C^2)}{\text{tr } (C^1)^2 \text{tr } (C^2)^2} = \frac{\sum_{i,i'} \left(c^1_{ii'} - \frac{1}{p}\right)\left(c^2_{ii'} - \frac{1}{q}\right)}{\sqrt{\sum_{i,i'} \left(c^1_{ii'} - \frac{1}{p}\right)^2 \sum_{i,i'} \left(c^2_{ii'} - \frac{1}{q}\right)^2}}.$$

Idrissi [5] used this formula to study the probability distribution of JV under the hypothesis of independence. If the k classes are equiprobable, one finds that $\sum_{i \neq i'} (c^1_{ii'} c^2_{ii'})$ follows a binomial distribution $B(n(n-1), 1/k^2)$. Idrissi claims that the JV index between two categorical variables with k equiprobable modalities has an expected value equal to:

$$E(JV) = \frac{k-1}{n}.$$

The RV-coefficient and the JV index are identical, in terms of paired comparisons.

4 Cohen's Kappa

The kappa coefficient for computing nominal scale agreement between two raters was first proposed by Cohen [1] . He defined his index as "the proportion of agreement after chance agreement is removed from consideration".

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where

$$P_o = \frac{1}{n}\sum_{i=1}^{k} n_{ii} \qquad \text{and} \qquad P_e = \frac{1}{n^2}\sum_{i=1}^{k} n_{i.}n_{.i}.$$

In this formula, P_o is the amount of observed agreement. In case of independence, given the marginals, the expected amount of agreement is P_e. Therefore, a correction is made for this amount agreement. In order that the index assumes the value 1 in case of perfect agreement, the correction is also made in the denominator. It measures the deviation of objects on the diagonals in the agreement table.

We use the equivalent formula of the kappa coefficient to compute the agreement between two partitions having the same number of classes:

$$\kappa = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.}n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.}n_{.i}}.$$

An important condition to use the kappa coefficient is for the situation in which the identity of classes per observations is known in advance. In our case, when we compare partitions coming from the same data set and found by the clustering methods, the numbering of classes is totally arbitrary: so we propose to identify the classes of the partitions from the maximum value of κ. We permute the columns or rows in the cross table and we calculate κ at each time; we take the numbering of permutation with the maximum κ.

Another method to find the optimal ordering is to use the order of the categories of both variables given by the first factor in the simple corresponding analysis. This method maximizes the weight of the diagonal of the cross table by permuting their lines and columns.

We compare the result found by the simple corresponding analysis to our method. Often, the same value of κ can be observed.

5 Popping's D2

The D_2 index was proposed by Popping [10] and is based on the same principles of kappa coefficient. It is used to study the agreement between the nominal classifications by two judges who independently categorize the same subjects in case that the categories are not known in advance. The D_2 index is based on the comparison of pairs of subjects and start from the situation of same agreement (Table 1).

The index contains a correction for agreement that can be expected on the basis of chance given the marginals of the original classification.

Under the null hypothesis of independence, the D_2 is a transformation of the Russel and Rao's coefficient

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

The D_2 index is defined as:

$$D_2 = \frac{D_o - D_e}{D_m - D_e}$$

where

$$D_o = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(n_{ij}-1)}{n(n-1)},$$

$$D_e = \frac{2\sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}}{n(n-1)},$$

$$c_{ij} = g_{ij}(h_{ij} - 0.5g_{ij} - 0.5) \quad \text{where } h_{ij} = \frac{n_{i.}n_{.j}}{n} \quad \text{and } g_{ij} = \text{integer } (h_{ij}),$$

$$D_p = \frac{\sum_{i=1}^{p} n_{i.}(n_{i.}-1)}{n(n-1)},$$

$$D_q = \frac{\sum_{i=1}^{p} n_{.j}(n_{.j}-1)}{n(n-1)},$$

and

$$D_m = \max(D_p, D_q).$$

Note that in D_2 one considers De as a reasonable minimum [10] , so for situations where we use the g_{ij} in D_e, we have no empirical proof to get an expected value smaller than the minimum. This is in favor of g_{ij} instead of the fractional number h_{ij}. However, this results in the fact that D_e is biased: the results are too high. The consequence for D_2 is that the index will take conservative values.

The D_2 index has been compared to the kappa coefficient to the JV index, in the particular following case:

Categories	+	−	Total
+	u	$v-u$	v
−	$v-u$	u	v
Total	v	v	$2v$

Table 2: The particular case found by Popping.

Popping obtained the results:

$$\kappa = \frac{2u-v}{v},$$

$$D_2 = \left[\frac{2u-v}{v}\right]^2 = JV.$$

In the general case we found that there are a strong correlation between these indices in the simulation study.

We find a relation between the Jaccard's index known as measure of similarity between objects described by presence-absence attributes [14] to the D_2 coefficient, we find the following relation:

In case of Integer$(hij) = h_{ij}$

$$D_2 = \frac{(c_n^2 - b)J}{\max(a+c, a+d) - C_n^2}$$

where

$$J = \frac{a}{a+c+d} \qquad \text{and } C_n^2 = \frac{n(n-1)}{2}.$$

6 Empirical distribution

We use the algorithm presented in [13] for finding the empirical distribution of the indices of association when the two partitions only differ by chance. The difficulty consists in conceptualising a null hypothesis of "identical" partitions and a procedure to check it. Note that departure from independence does not mean that there exists a strong enough agreement.

Now we have to define the sentence "two partitions are identical". Our approach consists to say that the units come from a same partition, where the two observed partitions are noised. The latent class model is well adapted to this problem for getting partitions. Note that Green and Krieger [3] have used it in their consensus partition research. More precisely, we use the latent profiles models because we have numerical variables.

For getting "near-identical partitions", we suppose the existence of a common partition for the population according to a latent profile model. The basic hypothesis is the independency of observed variables conditional to the latent classes that gives similar partitions from one or another groups of variables:

$$f(x) = \sum_k \pi_k \prod_j f_k(x_j/k).$$

The π_k are the class proportions and x is the random vector of observed variables, where the component x_j are independent in each class.

We use here the latent class model to generate the data and not to estimate parameters. Once having choosen the number of classes, we first generate their frequencies from a multinomial distribution with probabilities π_k, and then we generate observations in each class according to the local independence model in other words a normal mixture model with independent components in each class. Then we split arbitrarily the p variables into two sets and perform a partitioning algorithm on each set. The two partitions should thus differ only by random. We calculate the indices for the two partitions: we repeat the procedure N times to find a sampling distribution of kappa, RV (or J index), and D_2. Our algorithm has the following steps:

1. Generate the sizes n_1, n_2,..., n_k of the clusters according to a multinomial distribution $M(n; \pi_1 \ldots \pi_k)$.

2. For each cluster, generate n_i values from a random normal vector with p independent components.

3. Get two partitions: P_1 of the units according to the first p_1 variables and P_2 according to the last $p - p_1$ variables.

4. Compute association measures (for RV, J and D_2) for P_1 and P_2.

4'. Permute the columns of the cross table (P_1, P_2) to find the maximum value of kappa, so the numbering of classes.

5. Repeat the procedure N times.

7 Numerical Applications

We applied the previous procedure with 4 equiprobable latent classes, 1000 units and 4 variables. We obtained the two partitions P_1 (with X_1 and X_2) and P_2 (with X_3 and X_4) with 4 classes by the k-means methods. We present only one of our simulations (performed with S+ software).

Class 1	Class 2	Class 3	Class 4
X1 N(1.2,1.5)	X2 N(4,2.5)	X3 N(7,3.5)	X4 N(10,4.5)
X1 N(-2,1.5)	X2 N(-4,2.5)	X3 N(-6,3.5)	X4 N(-10,4.5)
X1 N(-5,1.5)	X2 N(-10,2.5)	X3 N(-13,3.5)	X4 N(-20,4.5)
X1 N(-8,1.5)	X2 N(-15,2.5)	X3 N(-20,3.5)	X4 N(-30,4.5)

Table 3: The normal mixture model.

The following figure shows the spatial distribution of one of the 1000 iterations for the two partitions P_1 and P_2.

The cross table of the two partitions in one of the simulations is represented as follows:

1	**2**	**3**	**4**
248	0	0	2
1	198	27	9
2	6	43	202
0	58	192	12

Table 4: Cross tabulation of P_1 and P_2 for one of the simulation.

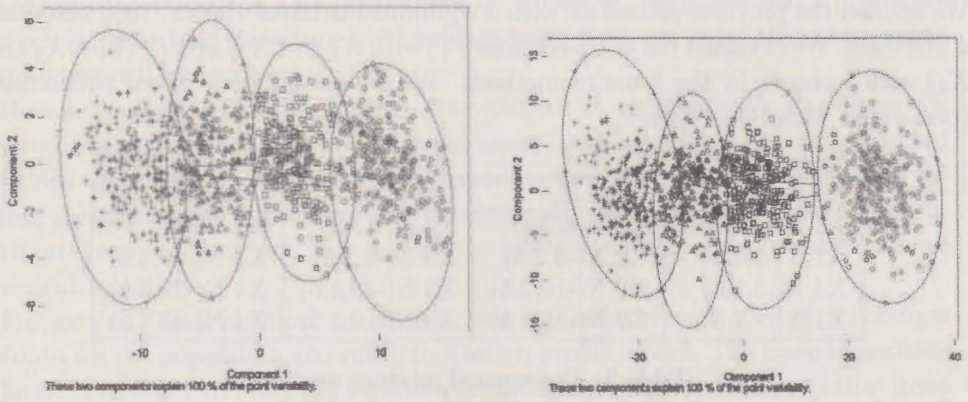

Figure 1. The first two principal components of one of the 1000 samples for the P1 and P2

We find a value of the kappa coefficient equal to 0.335, and a value of JV index (or RV) equal to 0.648. The value of the D_2 index is equal to 0.647.

To identify the labels of classes of P_2 to those of P_1, we permute the columns (4! permutations) of the cross table: The maximum value of the kappa coefficient is 0.787, and the numbering of column's permutation is 1,2,4,3. So the table becomes:

1	**2**	**3**	**4**
1	2	4	3
248	0	2	0
1	198	9	27
2	6	202	43
0	58	12	192

Table 5: Cross tabulation with the new permutation of columns.

Another way of getting the "optimal" ordering is by means of correspondence analysis: categories of both variables are ordered according to their coordinates along the first axis.

198	27	9	1
58	192	12	20
6 43	202	2	
0	0	2	248

Table 6: Order according to the first factor of CA.

If we calculate the kappa coefficient of this last table, we find the value 0.787. Here correspondence analysis ordering gives the same result as our method. The distribution of maximum kappa coefficient values is presented in Figure 3:

With the same choice of normal independent mixtures variables, we find that the kappa coefficient varies between 0.4 and 0.875. Its means is 0.82.

By simulation, the value of the JV index varies between 0.4 and 0.7. The most frequent value is 0.63 and the mean is equal to 0.617. (Fig. 4)

Under the hypothesis of independence $E(JV) = 0.003$, and with 1000 observations, independence should have been rejected for $JV > 0.617$ at 5% level. The 5% critical value is much higher than the corresponding one in the independence case. It shows that departure from independence does not mean that the two partitions are close enough.

The D_2 index takes its values between 0.3 and 0.7 with a mode equal to 0.625. The distribution has a mean equal to 0.610. Bimodality is due to the use of k-means, which gives a local optima [15] .

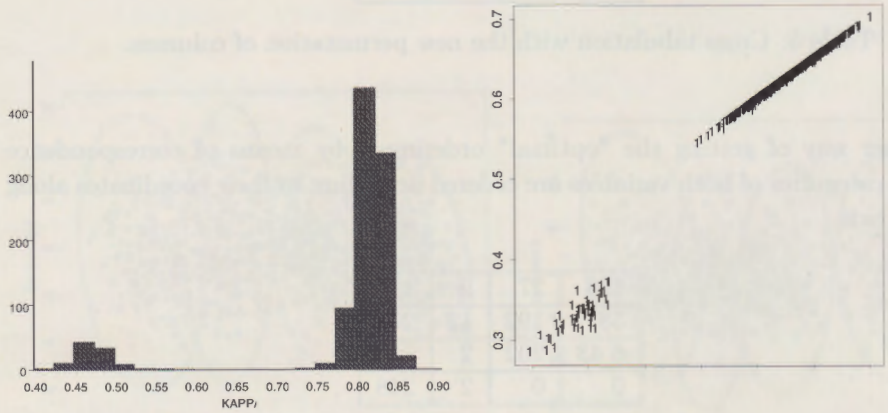

Figure 2. Distribution of kappa for 1000 individuals and 1000 iterations, partitions with 4 classes. The scatter plot of JV against D_2 in the 1000 iterations.

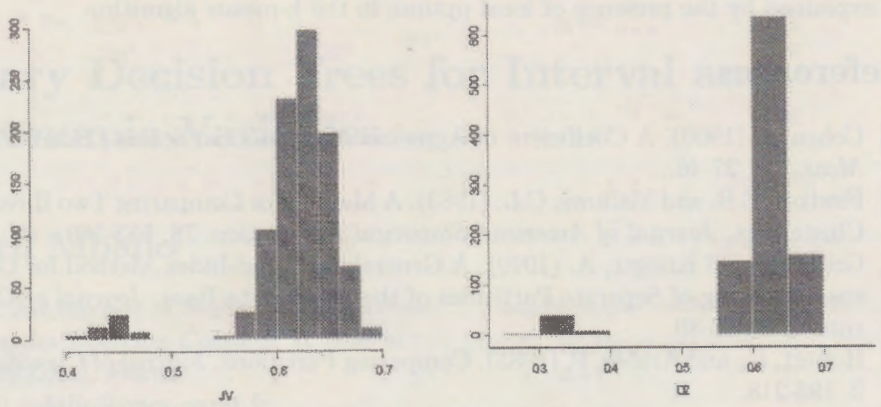

Figure 3. The JV- index distribution and D_2 for the partitions of 4 classes in 1000 iterations.

There is a strong correlation between the JV and D_2 equal to 0.983. So we can say that the two indices give same result in comparing partitions (Fig. 3).

Under the null hypothesis of close partitions, all indices have values around the mean, which is close to 0.6 so that we could say that the two partitions P_1 and P_2 are close enough.

8 Conclusion

In this paper, we have proved the identity between the RV-coefficient and the JV-index for comparing two partitions. A latent class model has been used to solve the problem of comparing close partitions and three agreement indices: JV (or RV), kappa and D_2 have been studied.

The kappa coefficient allows to test the similarity between two partitions after permutation and when we they have the same number of classes. We compare the optimal permutation with the order given by correspondence analysis: By simulation, it gives often the same results.

The D_2 index seems useful for comparing the classification of two partitions in the case that the identification of classes is not known in advance. D_2 gives stress on pairs in the same class of both partitions, as Jaccard's index. We found a relation between them.

Since we found a strong correlation between the JV index and the D_2 in our simulation, we can deduce that JV (or RV) and D_2 lead us to the same result in comparing partitions. The distributions of these proposed indices have been found

very different from the case of independence and are bimodal. The bimodality might be explained by the presence of local optima in the k-means algorithm.

References

[1] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales., *Educ. Psychol. Meas.*, 20, 27-46.

[2] Fowlkes, E.B. and Mallows, C.L. (1983). A Method for Comparing Two Hiearchical Clusterings, *Journal of American Statistical Association*, 78, 553-569.

[3] Green, P. and Krieger, A. (1999). A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base, *Journal of Classification*, 16, 63-89.

[4] Hubert, L., and Arabie, P. (1985). Comparing Partitions, Journal of Classification, 2, 193-218.

[5] Idrissi, A. (2000). Contribution à l'Unification de Critères d'Association pour Variables Qualitatives, *Thèse de doctorat de l'Université de Paris 6.*

[6] Lazraq, A. and Cleroux R. (2002). Inférence Robuste sur un Indice de Redondance, *Revue de Statistique Appliquée*, 4, 39-54.

[7] Lerman, I.C (1973). Etude Distributionnelle de Statistiques de Proximité entre Structures Finies de Memes Types; Application à la Classification Automatique, Cahier 19, *Bureau Universitaire de Recherche Opérationelle, Institut de Statistique des Universités de Paris.*

[8] Lerman, I.C (1988). Comparing Partitions (Mathematical and Statistical Aspects), *Classification and Related Methods of Data Analysis*, H.H Bock Editor, 121-131.

[9] Janson S., Vegelius J. (1982). The J- Index as a Measure of Association For Nominal Scale Response Agreement, *Applied psychological measurement*, 16, 243-250.

[10] Popping, R. (1983). Traces of Agreement. On the Dot-Product as a Coefficient of Agreement, *Quality and Quantity*, 17, 1-18.

[11] Popping, R. (1992). Taxonomy on Nominal Scale Agreement, *Groningen; iec ProGAMMA*, 1945-1990.

[12] Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The *RV*-Coefficient, *Appl. Statist.*, 25, 257-265.

[13] Saporta G. and Youness, G. (2001). Concordance entre Deux Partitions: Quelques Propositions et Expériences, in *Proceedings SFC 2001, 8èmes rencontres de la Société Francophone de Classification, Pointe à Pitre.*

[14] Saporta G. and Youness, G. (2002) Comparing Two Partitions: some proposals and Experiments, *Proceedings in Computational Statistics edited by Wolfgang Härdle, Physica- Verlag*, Berlin.

[15] Saporta G. and Youness, G. Une Méthodologie Pour la Comparaison de Partitions, *Revue de Statistique Appliquée*, to appear.