Comparing two partitions: Some Proposals and Experiments

Gilbert Saporta[1] and Genane Youness[2]

[1] Chaire de Statistique Appliquée-CEDRIC, CNAM, 292 rue Saint, 75003 Paris, France, saporta@cnam.fr
[2] CNAM-ISAE, BP 11 4661, Beirut, Lebanon, genaneb@terra.net.lb

Abstract. We propose a methodology for finding the empirical distribution of the Rand's measure of association when the two partitions only differ by chance. For that purpose we simulate data coming from a latent profile model and we partition them according to 2 groups of variables. We also study two other indices: the first is based on an adaptation of Mac Nemar's test, the second being Jaccard's index. Surprisingly, the distributions of the 3 indices are bimodal.

Keywords. Latent class, K-means, Rand index, Jaccard index, partitions

1 Introduction

When one observes two partitions of the same set of units, a natural question arises: do the partitions agree or disagree? One of the most popular measures of concordance is the Rand's measure of association (Hubert & Arabie 1985), which is based upon the number of pairs of units, which belong to the same clusters. A natural idea is to decide that the 2 partitions do not differ significantly if the index is larger than a critical value. We thus need to know, even approximately, the distribution of Rand's index under some null hypothesis. Few publications (Idrissi 2000) deal with that problem, and only under the hypothesis of independence. However this hypothesis is unrealistic, and departure from independence does not mean that there exists a strong enough agreement (Saporta 1997).

But the difficulty consists in conceptualising a null hypothesis of "identical" partitions and a procedure to check it.

In this communication we first remind the main properties of Rand's index and its distribution under independence. We propose an alternative to Rand's by using an idea derived from Mac Nemar's test for comparing proportions: here we compare proportions of discordant pairs. Finally we also use Jaccard's index. We simulate similar partitions coming from a common latent class model. Then we split arbitrarily the p variables into two groups and perform a partitioning algorithm on each set with the k-means's method.

2 Measures of agreement between partitions

2.1 Notations

Let V_1 and V_2 be two partitions (or two categorical variables) of n objects with the same number of classes k. If K_1 and K_2 are the disjunctives tables and N the corresponding contingency table with elements n_{ij}, we have: $N = K_1^{'} K_2$

Each partition V is also characterized by the $n \times n$ paired comparison table C with general term c_{ij}:

c_{ij}=1 if i and j are in the same class of V, c_{ij}=0 otherwise

We have $C_1 = K_1 K_1^{'}$ and $C_2 = K_2 K_2^{'}$

The four types of pairs of objects are:

Type 1: pairs belonging to the same class of V_1 and to the same class of V_2
Type 2: pairs belonging to different classes of V_1 but to the same class of V_2
Type 3: pairs belonging to the same class of V_1 but to different classes of V_2
Type 4: pairs belonging to different classes of V_1 and to different classes of V_2

If the respective frequencies of these four cases are : a, b, c, d, we have:

$$a + b + c + d = \frac{n(n-1)}{2} .$$

We note also $A = a + d$ (total number of agreements) and $D = b + c$ (total number of discordances)

2.2 Rand index

The Rand index (similar to Kendall's measure) is the proportion of agreements:

$$R = \frac{2A}{n(n-1)}$$

It may be proved that:

$$A = C_n^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{k} n_{.j}^2 \right]$$

We will not use the index corrected for chance by Hubert and Arabie (1985), but Marcotorchino's modified version (1991) for all n^2 pairs :

$$R = \frac{2\sum\sum n_{ij}^2 - \sum n_{i.}^2 - \sum n_{.j}^2 + n^2}{n^2}$$

which leads to a simple expression based in terms of paired comparisons:

$$R = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(c_{ij}^1 c_{ij}^2 + \bar{c}_{ij}^1 \bar{c}_{ij}^2)}{n^2} \qquad \text{With } \bar{c} = 1 - c$$

Idrissi (2000) used this last formula to study the asymptotic normality of R under the hypothesis of independence. If the k classes are equiprobable, one finds that

$c_{ij}^1 c_{ij}^2 + \bar{c}_{ij}^1 \bar{c}_{ij}^2$ has a Bernoulli distribution with parameter $1-\frac{2}{k}+\frac{2}{k^2}$. Hence:

$$E(R) = 1 - \frac{2}{k} + \frac{2}{k^2}$$

A. Idrissi claims that the Rand index between two categorical variables with k equiprobable modalities follows asymptotically a normal distribution with variance:

$$V(R) = \frac{1}{n^2}(1-\frac{1}{n})(1-\frac{2}{k}+\frac{2}{k^2})(\frac{2}{k}-\frac{2}{k^2})$$

This expression is not valid for small k (especially k=2) and only approximately true for large n since the c_{ij} are not independent due to the transitivity constraints

$$c_{ij} = c_{jk} c_{ik}$$

2.3 An adaptation of Mac Nemar's test

Mac Nemar's test is a well known non-parametric test used to check equality of proportions in matching samples:

a	b
c	d

For instance a represents the number of individuals who keep the same favorable opinion before and after a campaign, d the number of individuals who keep the same unfavorable opinion before and after, b and c are the frequency of those who are changing their opinion. The test statistic corresponding to the null hypothesis of equally changing opinions is: $Mc = \frac{b-c}{\sqrt{b+c}}$ and Mc has a normal distribution N(0,1) under H_o for n large.

By using the test for the set of object pairs, we have a new way to measure the agreement between two partitions. It is easy to get :

$$Mc = \frac{\sum_i n_{i.}^2 - \sum_j n_{.j}^2}{2\sqrt{\frac{1}{2}(\sum_i n_{i.}^2 + \sum_j n_{.j}^2) - \sum_i \sum_j n_{ij}^2}}$$

In this case also, the transitivity relations between pairs goes against the assumption of independence between pairs.

2.4 The Jaccard's index

The Jaccard's index is a well-known measure of similarity between objects described by presence-absence attributes, used in cluster analysis. It counts the number of common attributes divided by the number of attributes possessed by at least one of the 2 objects.

Applied to the four types of pairs we have:

$$J = \frac{a}{a+b+c} = \frac{n - \sum_u \sum_v n_{ij}^2}{n + \sum_i \sum_j n_{ij}^2 - \sum_i n_{i.}^2 - \sum_j n_{.j}^2}$$

3 The latent class model

Now we have to define what we mean by "two partitions are close". Our approach consists in saying that the units come from the same common partition, the two observed partitions being noisy realisations. The latent class model is well adapted to this problem for getting partitions and have been used by Green and Krieger (1999) in their consensus partition research. More precisely, we use the latent profile model for numerical variables.

	Latent variables	
Observed Variables	*Qualitative*	*Quantitative*
Qualitative	Latent class	Latent traits
Quantitative	Latent profile	Factor analysis

Figure 1. Latent variables methods (Bartholomew & Knott 1999)

The basic hypothesis is the independency of observed variables conditional to the latent classes:

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

The p_k are the proportion of classes and x is the random vector of observed variables, where the component x_j are independent in each class. Here we use the model only in order to generate data and not to estimate parameters.

For getting "near-identical partitions", we suppose the existence of such a partition for the population according to the latent profile model. Data are generated according to this model, with independent normal components in each class, in other words, a normal mixture model. Then we split arbitrarily the p variables into two sets and perform a partitioning algorithm on each set. The two partitions should differ only at random. We are thus enabled to get simulated sampling distributions of Rand, Mc Nemar or Jaccard's index. Our algorithm has four steps:

1. Generate the sizes n_1, n_2, .., n_k of the clusters according to a multinomial distribution M(n; $p_1 \ldots p_k$)
2. For each cluster, generate n_i values from a random normal vector with p independent components
3. Get 2 partitions of the units according to the first p_1 variables and the last $p-p_1$ variables
4. Compute association measures

4 Empirical results

We applied the previous procedure with 4 equiprobable latent classes, 1000 units and 4 variables. The parameters of the normal distribution are chosen in such a way that $|m_{kj} - m_{k'j}| > 1.5\,\sigma_j$ for every j and k.

The number of iterations N is 1000. We present only one of our simulations (performed with S+ software).

Class n°1		*Class n°2*		*Class n°3*		*Class n°4*	
X1	N(1.2,1.5)	X1	N(-2,1.5)	X1	N(5,1.5)	X1	N(8,1.5)
X2	N(-10,2.5)	X2	N(0,2.5)	X2	N(-17,2.5)	X2	N(3.8,2.5)
X3	N(6,3.5)	X3	N(12,3.5)	X3	N(13,3.5)	X3	N(-5,3.5)
X4	N(-20,4.5)	X4	N(-12,4.5)	X4	N(0,4.5)	X4	N(7,4.5)

Table 1. The normal mixture model

The following figure shows the spatial repartition of one of the 1000 iterations.

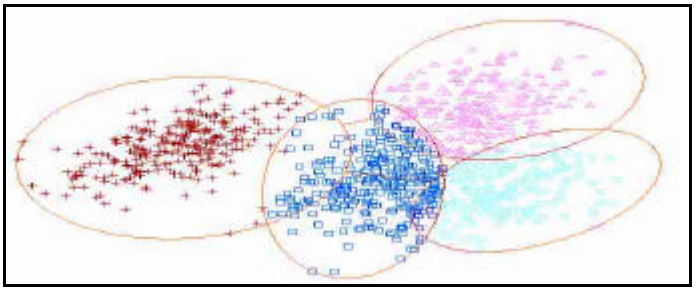

Figure 2. The first two principal components of one of the 1000 samples

Then, we compute 2 partitions with the k-means methods:the first one with X_1 and X_2, the other one with X_3 and X_4; we calculate the association indices 1000 times. Our results show that the distributions of these indices are far from a normal distribution, which is not surprising since the theoretical values should be high (close to 1 for Rand), but they are actually bimodal: this unexpected result has been observed systematically.

We noticed that all the observed Rand's values are over 0.72. Under the hypothesis of independence $E(R) = 0.625$, and with 1000 observations, independence should have been rejected for R>0.626 at 5% risk. 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. However it is not possible to derive universal critical values since the distribution of *R* depends on the number of clusters, on their proportions and separability. An ad hoc bootstrap method may solve this problem.

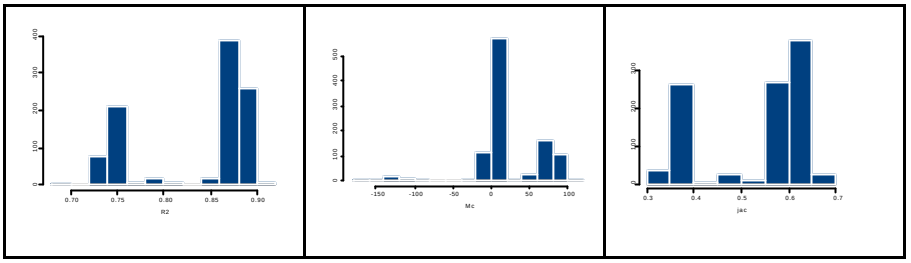

Figure 3. Distributions of Rand, Mac Nemar and Jaccard's indices

5 Discussion

A latent class model has been used to deal with the problem of comparing close partitions and three agreement indices have been studied. The Rand index give the same importance to pairs in the same class, and to pairs in different classes of both partitions, which is arguable. Mac Nemar and Jaccard indices do not have this drawback. The distributions of the three 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: we are studying this point. Finally, one has to add that agreement measures are only one of the many facets of comparing partitions.

References

Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis* , London: Arnold.

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

Hubert, L. & Arabie, P.(1985). Comparing partitions, *Journal of Classification*, **2**, 193-198.

Idrissi, A. (2000). *Contribution à l'unification de Critères d'Association pour Variables Qualitatives*, Ph.D., Paris: Université Pierre et Marie Curie.

Marcotorchino, J.F.& El Ayoubi, N. (1991). Paradigme logique des écritures relationnelles de quelques critéres fondamentaux d'association, *Revue de Statistique Appliquée*, **39**, 2, 25-46.

Saporta, G. (1997). Problèmes posés par la comparaison de classifications dans des enquêtes différentes, in: *Proceedings of the 53rd Session of the International Statistical Institute*.