Comparison of distance indices between partitions

L. Denoeud[1,2], H. Garreta[3], and A. Guénoche[4]

[1] École nationale supérieure des télécommunications, 46, rue Barrault, 75634 Paris cedex 13 (e-mail: denoeud@infres.enst.fr)
[2] CERMSEM CNRS-UMR 8095, MSE, Université Paris 1 Panthéon-Sorbonne, 106-112, boulevard de l'Hôpital, 75647 Paris cedex 13
[3] Laboratoire d'Informatique Fondamentale, 163, avenue de Luminy, 13009 Marseille (e-mail: garreta@lif.univ-mrs.fr)
[4] Institut de Mathématiques de Luminy, 163, avenue de Luminy, 13009 Marseille (e-mail: guenoche@iml.univ-mrs.fr)

Abstract. In this paper, we compare several distance indices between partitions on the same set. First, we build a set $\mathcal{P}_k(P)$ of partitions close to each others by applying to an initial partition P, k transfers of one element from its class to another. Then we compare the distributions of several indices of distance between partitions of $\mathcal{P}_k(P)$.
Keywords: distance index, partition.

1 Introduction

The comparison of partitions is a central topic in clustering, as well for comparing partitioning algorithms as for classifying nominal variables. The literature abounds in indices defined by multiple authors to compare two partitions P and Q on the same set X. The most used are: the Rand index [Rand, 1971], the Jaccard index and the Rand index corrected for chance by Hubert and Arabie [Hubert and Arabie, 1985]. We also wanted to study the Wallace index [Wallace, 1983] and the normalized index of Lerman [Lerman, 1981]. The comparison of these indices is only interesting (in a practical point of view) if we consider close partitions, which differ randomly one from each others as it is mentioned by Youness and Saporta [Youness and Saporta, 2004]. They generate such partitions according to the *latent class model* [Bartholomew and Knott, 1999] adapted to an euclidian representation of the elements of X. We develop here a more general approach, independent of the representation space of X.

In 1964, Régnier proposed a distance between partitions which fits this type of study [Régnier, 1964]. It is the minimum number of transfers of one element from its class to another (eventually empty) to turn P into Q. We have recently studied this measure [Charon *et al.*, 2005] and called it the *transfer distance*. We compare the distributions of the distance indices above on partitions at k transfers from P. If k is small enough, these partitions are

close to P since they represent only a small percentage α of all the partitions of X. This permits to define the value k_α of the maximum number of transfers allowed, and to build the set $\mathcal{P}_{k_\alpha}(P)$ of random partitions obtained by at most k_α transfers from P.

2 The transfer distance

Let P and Q be two partitions on the set X of n elements with respectively p and q classes ; we will admit that $p \leq q$.

$$P = \{C_1, .., C_p\} \text{ and } Q = \{C'_1, .., C'_q\}.$$

The minimum number of transfers to turn P into Q, denoted $\theta(P, Q)$, is obtained by establishing a bijection between the classes of P and those of Q keeping a maximum number of elements in matching classes, those that don't need to be moved. Consequently, we begin to add $q - p$ empty classes to P, so that P is considered as a partition with q classes.

Let Υ be the mapping from $P \times Q \longrightarrow \mathbb{N}$ which associates to one pair of classes the cardinal of their intersection. Classically, $n_{i,j} = |C_i \cap C'_j|$ and $n_i = |C_i|$ and $n'_j = |C'_j|$ denote the cardinals of the classes. Let Δ be the mapping which associates to each pair of classes (C_i, C'_j) the cardinal of their symmetrical difference, noted $\delta_{i,j}$. We have $\delta(i, j) = n_i + n'_j - 2 \times n_{i,j}$. So we consider the complete bipartite graph $K_{q,q}$ whose vertices are the classes of P and Q, with edges weighted either by Υ or by Δ.

Proposition 1 ([Day, 1981]) *The bijection minimizing the number of transfers between two partitions with q classes P and Q corresponds to a matching of maximum weight w_1 in $K_{q,q}$ weighted by Υ or, equivalently, to a matching of minimum weight w_2 in $K_{q,q}$ weighted by Δ; moreover, $\theta(P, Q) = n - w_1 = \frac{w_2}{2}$.*

Establishing the bipartite graph is in $O(n^2)$. The weighted matching problem in a complete bipartite graph can be solved by an assignment method well known in operational research [Kuhn, 1955], [Kuhn, 1956]. The algorithm has a polynomial complexity in $O(q^3)$. We won't go into further details, given for instance in [Faure *et al.*, 2000]. A computer program (in C) can be requested to the authors. We just develop an example of computation of the transfer distance.

Example 1 *We consider the two partitions $P = (1, 2, 3|4, 5, 6|7, 8)$ and $Q = (1, 3, 5, 6|2, 7|4|8)$. The two following tables correspond to the intersections and to the symmetrical differences of the classes of P and Q. Two extreme matchings are edited in bold. Each one gives $\theta(P, Q) = 4$.*

To the maximum weighted matching in the table Υ corresponds the series of 4 transfers: $(1, 2, 3|4, 5, 6|7, 8) \rightarrow (1, 3|4, 5, 6|2, 7, 8) \rightarrow (1, 3, 5|4, 6|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7, 8) \rightarrow (1, 3, 5, 6|4|2, 7|8)$.

Υ	1,3,5,6	2,7	4	8	Δ	1,3,5,6	2,7	4	8
1,2,3	**2**	1	0	0		3	**3**	4	4
4,5,6	2	0	**1**	0		**3**	5	2	4
7,8	0	**1**	0	1		6	2	3	**1**
$\emptyset$	0	0	0	**0**		4	2	**1**	1

To the minimum weighted matching in the table Delta corresponds another optimal series: $(1,2,3|4,5,6|7,8) \rightarrow (1,2,3,7|4,5,6|8) \rightarrow (2,3,7|1,4,5,6|8) \rightarrow (2,7|1,3,4,5,6|8) \rightarrow (2,7|1,3,5,6|8|4)$.

3 Close partitions in terms of transfers

We note $\mathcal{P}_n$ the set of partitions on a set of n elements and $\mathcal{P}_k(P)$ the set of partitions at k transfers from P and $\mathcal{P}_{\leq k}(P)$ the set of partitions at at most k transfers from P.

$$\mathcal{P}_k(P) = \{Q \in \mathcal{P}_n \text{ such that } \theta(P,Q) = k\}$$

$$\mathcal{P}_{\leq k}(P) = \{Q \in \mathcal{P}_n \text{ such that } \theta(P,Q) \leq k\} = \bigcup_{0 \leq i \leq k} \mathcal{P}_i(P)$$

Statistically, we consider that a partition Q is close to P at threshold α if the probability of observing a partition closer to P than $\theta(P,Q)$ is lower than or equal to α. The matter is then to know how many partitions are within a k radius from P. For $k = 0$, there is just one partition, P itself, otherwise θ would'nt be a distance. We can easily enumerate $\mathcal{P}_1(P)$, but for larger k it becomes difficult. We call *critical value* of the partition P, at threshold α, the greatest number of transfers k_α such as

$$\frac{|\mathcal{P}_{\leq k_\alpha}(P)|}{|\mathcal{P}_n|} \leq \alpha.$$

While $n \leq 12$, we can enumerate all the partitions in $\mathcal{P}_n$ and we compute $|\mathcal{P}_k(P)|$. For that, we use the procedure NexEqu in [Nijenhuis and Wilf, 1978]. Each partition is coded by the vector of the class number to which each element belongs. The algorithm builds the next partition for the lexicographic order on this code, starting from the partition with a single class.

For $n > 12$, there are too many partitions to realize an exhaustive enumeration. Then we select at random a large number of partitions, to be compared to P to estimate $|\mathcal{P}_{\leq k}(P)|/|\mathcal{P}_n|$. To obtain a correct result, the partitions must be equiprobable; the book of Nijenhuis and Wilf provides also such a procedure (RandEqu).

Thus we measure a frequency f in order to estimate a proportion p. We want to approximate $p = 0.1$ for a risk ρ fixed ($\rho = 5\%$) and a gap δ between

f and p judged as acceptable ($\delta = 0.01$). For these values, we can establish the size of the sample E by the classical formula:

$$t(\rho)\sqrt{\frac{f(1-f)}{|E|}} \leq \delta$$

in which $t(\rho)$ is given by the normal distribution of Gauss [Brown *et al.*, 2002]. We obtain that 3600 trials should be carried out, which are quite feasible. We can notice that this number decreases with p (when $p < 0.5$) and it is independent of n.

Example 2 *For $n = 12$, there are $|\mathcal{P}_{12}| = 4213597$ partitions that can be compared to P in order to establish the distribution of $|\mathcal{P}_k(P)|$ according to k. For $P = \{1,2,3,4|5,6,7|8,9|10,11|12\}$, as for all the partitions with classes having the same cardinality, the number of partitions at $0,\ldots,8$ transfers from P are respectively* 1, 57, 1429, 20275, 171736, 825558, 1871661, 1262358, 60522 *and 0 beyond. The cumulated proportions in % are respectively* 0.0, 0.0, 0.0, 0.5, 4.6, 24.2, 68.6, 99.6, *and* 100. *For $\alpha = .1$ the critical value is* 4*; indeed there are just* 4.6% *of the partitions that are at most at* 4 *transfers from P, while for* 5 *transfers, there are* 24.2%. *The cumulated frequencies computed from P and 5000 random partitions are:* $0.0, 0.0, 0.1, 0.5, 4.4, 23.9, 68.7, 98.3$ *and* 100. *Thus the critical value computed by sampling is also equal to* 4.

4 Indices of proximity between partitions

The comparison of partitions is based on the pairs of elements of X. Two elements x and y can be joined together or separated in P and Q. The two partitions agree on (x, y) if these elements are simultaneously joined or separated in P and Q. On the other hand there is a disagreement if x and y are joined in one of them and separated in the other. Let r be the number of pairs simultaneously joined together, s the number of pairs simultaneously separated, an u (resp. v) the number of pairs joined (resp. separated) in P and separated (resp. joined) in Q.

According to the previous notations, we have $r = \sum_{i,j} \frac{n_{i,j}(n_{i,j}-1)}{2}$. Equivalent formulas for s, u and v appear in several papers. We will note $\pi(P)$ the set of joined pairs in P, that is to say $|\pi(P)| = \sum_{i=1,p} \frac{n_i(n_i-1)}{2}$.

4.1 The Rand index

The Rand index [Rand, 1971], noted R, is simply the percentage of pairs for which there is an agreement. It belongs to $[0, 1]$ and $1 - R(P, Q)$ is the symmetrical difference distance between $\pi(P)$ and $\pi(Q)$.

$$R(P,Q) = \frac{r+s}{n(n-1)/2}$$

4.2 The Jaccard index

In the Rand index, the pairs simultaneously joined or separated are counted in the same way. However, partitions are often interpreted as classes of joined elements, the separations being the consequences of this clustering. We use then the Jaccard index (1908), noted J, which does not take into account the s simultaneous separations:

$$J(P,Q) = \frac{r}{r+u+v}$$

4.3 The corrected Rand index

In their paper of 1985 [Hubert and Arabie, 1985], they noticed that the Rand index is not *corrected for chance* that is equal to zero for random partitions having the same number of objects in each class. They introduced the corrected Rand index, whose expectation is equal to zero, noted here HA, in homage to the authors.

The corrected Rand index is based on three values: the number r of common joined pairs in P and Q, the expected value $Exp(r)$ and the maximum value $Max(r)$ of this index, among the partitions of the same type as P and Q. It leads to the formula

$$HA(P,Q) = \frac{r - Exp(r)}{Max(r) - Exp(r)}$$

with $Exp(r) = \frac{|\pi(P)|\times|\pi(Q)|}{n(n-1)/2}$ and $Max(r) = \frac{1}{2}(|\pi(P)| + |\pi(Q)|)$. This maximum value is questionable since the number of common joined pairs is necessarily bounded by $\inf\{|\pi(P)|, |\pi(Q)|\}$, but $Max(r)$ insures that the maximum value of HA is 1 when the two partitions are identical. On the other hand this index can take negative values.

4.4 The Wallace index

This index is very natural, it's the number of joined pairs common to P and Q divided by the number of possible pairs [Wallace, 1983]. This last quantity depends on the partition of reference and, if we don't want to favour neither P nor Q, the geometrical average is used.

$$W(P,Q) = \frac{r}{sqrt(|\pi(P)| \times |\pi(Q)|)}$$

.

4.5 The normalized Lerman index

The Lerman index(denoted ICL) is the difference between the number of simultaneously joined pairs and its expectation, divided by its standard deviation [Lerman, 1988].

$$ICL(P,Q) = \frac{r - Exp(r)}{\sqrt{Var(r)}}$$

These two values are computed on the set of pairs of partitions having the same types as P and Q; they are defined according to the cardinals of the classes. The expected value of r already appears in the formula given by Hubert and Arabie and its variance $Var(r)$ is given by:

$$\frac{V_1(P)V_1(Q)}{2n(n-1)} + \frac{V_2(P)V_2(Q)}{n(n-1)(n-2)} + \frac{V_3(P)V_3(Q)}{4n(n-1)(n-2)(n-3)} - [\frac{V_1(P)V_1(Q)}{2n(n-1)}]^2$$

where $V_1(P) = \sum_{i=1,p} n_i(n_i-1)$, $V_2(P) = \sum_{i=1,p} n_i(n_i-1)(n_i-2)$ and

$$V_3(P) = [\sum_{i=1,p} n_i(n_i-1)]^2 - 2\sum_{i=1,p} n_i(n_i-1)(2n_i-3)],$$

with similar expressions for $V_1(Q)$, $V_2(Q)$ and $V_3(Q)$, in which the sums are computed on q classes and the n_i are replaced by n'_i.

The index value is not defined when $Var(r) = 0$, that is when one of the partitions has a single class or n singletons. As for the HA index, it can be negative, but it is not upper bounded. Finally, Lerman proposes a normalized index defined as a correlation coefficient given by the formula:

$$ILN(P,Q) = \frac{ICL(P,Q)}{\sqrt{ICL(P,P) \times ICL(Q,Q)}}$$

5 Comparison of indices

Let P be a partition on X with p classes, defined by its type, that is to say by the cardinal of its classes. When $n = |X| \leq 12$, we enumerate the sets $\mathcal{P}_k(P)$, then we evaluate the minimum and maximum values of each index above between P and any Q belonging to $\mathcal{P}_k(P)$. The table 1 contains the results for $P = (1,2,3,4,5|6,7,8,9,10)$. The partitions being at at most 3 transfers represent 1.7% of the 115975 partitions on 10 elements.

One can observe that, for each index, the maximum value obtained for partitions at 5 transfers are greater than the minimum value obtained for 2 transfers. Moreover the minimum values at 3 transfers are very small and don't reflect the closeness of these partitions and P. Finally, for the normalized Lerman index, the maximum values do not decrease with k and the closest partition from P is at 4 transfers.

Nb. of transfers	1	2	3	4	5	6	7	8
Nb. of partitions	20	225	1720	9112	31361	54490	17500	1546
J min	.64	.43	.32	.22	.15	.08	.04	0.0
J max	.80	.70	.60	.50	.44	.21	.10	0.0
R min	.80	.64	.53	.47	.44	.44	.44	.44
R max	.91	.87	.82	.78	.69	.64	.60	.56
HA min	.60	.28	.06	-.08	-.12	-.17	-.19	-.22
HA max	.82	.72	.63	.53	.32	.22	.11	0.0
W min	.78	.60	.49	.37	.28	.16	.09	0.0
W max	.89	.84	.77	.71	.67	.45	.32	0.0
ILN min	.61	.28	.06	-.08	-.20	-.20	-.23	-.32
ILN max	.86	.84	.95	1.15	.67	.39	.25	-.14

Table 1. Distribution of the number of partitions at k transfers from P and extreme values of the distance indices

.

In the case $n > 12$, we cannot enumerate $\mathcal{P}_n$ anymore. Then, in order to compare very close partitions in the neighborhood of a given partition P,

- we compute by sampling the critical number of transfers $k_{5\%}$;
- we build a set $\mathcal{Q}_k(P)$ of 100 partitions Q randomly selected such as $\theta(P,Q) \leq k$, with $k \leq k_{5\%}$;
- we compare all the partitions of $\mathcal{Q}_k(P)$ two by two and measure the average value and the standard deviation of each studied index.

The partitions close to P are obtained by selecting recursively at random one element; if this element is not alone in its class, its new class number is selected between 1 and $p+1$, and the number of classes is updated. Here, we restrict our study at the single partition of 100 elements spread in 5 balanced classes of 20 elements each. The critical value at 5% is 83, that is to say that only 5% of the partitions with 100 elements are at less at 83 transfers from the balanced partition with 5 classes.

The figure 1 represents the computed averages and standard deviations of each index for $k \in [5; k_\alpha]$, with a step of 5.

We can see that the indices decrease when k increases since the partitions are less close to each other. The indices of Jaccard, corrected Rand , Wallace, and Lerman have approximately the same behavior: they are high when k is small and decrease near to 0 when $k = k_\alpha$. But they reflect the closeness of partitions only when k is very small. Among these indices the Jaccard index seems to be the most accurate since it has the lowest standard deviation. The Rand index has a different behavior: its values stays above 0,8 whatever is k. Two pairs of partitions at 40 and 90 transfers from each others can have the same value.

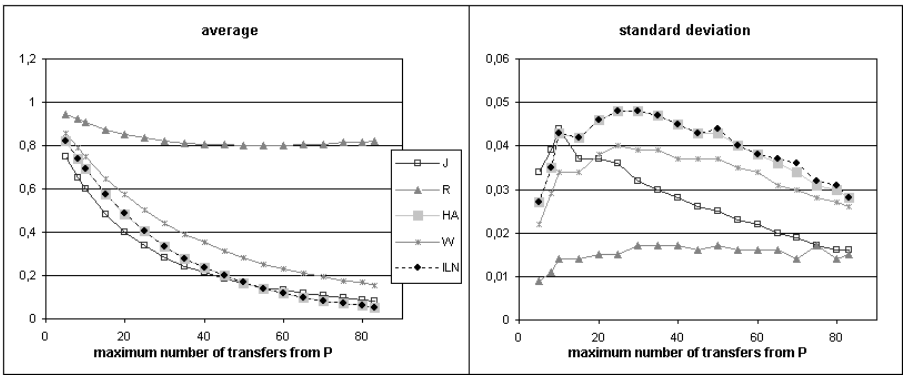

Fig. 1. Average and standard deviation of the distance indices between partitions of $\mathcal{Q}$

We have obtained the same kind of results for other initial partitions, balanced or not. Our conclusion is that the Rand index isn't very satisfying for the comparison of close partitions. Among the others, the Jaccard index seems the best, followed by the Wallace index, because they have the lowest standard deviation. The corrected Rand index and the normalized Lerman index share similar average values but the extreme values of the normalized Lerman index make it less satisfying.

Acknowledgements This work is supported by the CNRS ACI IMP-Bio. We also want to thank B. Fichet (University of Aix-Marseille II) and I.C. Lerman (University of Rennes I) for their help and advices.

References

[Bartholomew and Knott, 1999]D. Bartholomew and M. Knott. *Latent Variables Models and Factor Analysis*. Arnold, London, 1999.

[Brown *et al.*, 2002]L. Brown, T. Cai, and A. DasGupta. Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.*, pages 160–201, 2002.

[Charon *et al.*, 2005]I. Charon, L. Denoeud, A. Guénoche, and Hudry O. Comparing partitions by element transfers. *submitted*, 2005.

[Day, 1981]W. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, pages 269–287, 1981.

[Faure *et al.*, 2000]R. Faure, B. Lemaire, and C. Picouleau. Précis de recherche opérationnelle. *Mathematical Social Sciences*, pages 134–137, 2000.

[Hubert and Arabie, 1985]L. Hubert and P. Arabie. Comparing partitions. *J. of Classification*, pages 193–218, 1985.

[Kuhn, 1955]H.W. Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, pages 83–97, 1955.

[Kuhn, 1956]H.W. Kuhn. Variants on the hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, pages 253–258, 1956.

[Lerman, 1981]I.C. Lerman. *Classification et analyse ordinale des données*. Dunod, Paris, 1981.

[Lerman, 1988]I.C. Lerman. Comparing partitions (mathematical and statistical aspects). In H.H Bock, editor, *Classification and Related Methods of Data Analysis*, pages 121–131, 1988.

[Nijenhuis and Wilf, 1978]A. Nijenhuis and H. Wilf. *Combinatorial algorithms*. Academic Press, New-York, 1978.

[Rand, 1971]W.M. Rand. Objective criteria for the evaluation of clustering methods. *J. of the Am. Stat. Association*, pages 846–850, 1971.

[Régnier, 1964]Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, 1964.

[Wallace, 1983]D.L. Wallace. Comment. *J. of the Am. Stat. Association*, pages 569–579, 1983.

[Youness and Saporta, 2004]G. Youness and G. Saporta. Une méthodologie pour la comparaison des partitions. *Revue de Statistique Appliquée*, pages 97–120, 2004.