Simulation Experiments for Similarity Indexes Between Two Hierarchical Clusterings

Isabella Morlini

Abstract In this paper we report results of a series of simulation experiments aimed at comparing the behavior of different similarity indexes proposed in the literature for comparing two hierarchical clusterings on the basis of the whole dendrograms. Simulations are carried out over different experimental conditions.

1 Introduction

Morlini and Zani (2012) have proposed a new dissimilarity index for comparing two hierarchical clusterings on the basis of the whole dendrograms. They have presented and discussed its basic properties and have shown that the index can be decomposed into contributions pertaining to each stage of the hierarchies. Then, they have obtained a similarity index S as the complement to one of the suggested distance and have shown that its single components S_k obtained at each stage k of the hierarchies can be related to the measure B_k suggested by Fowlkes and Mallows (1983) and to the Rand index R_k. In this paper, we report results of a series of simulation experiments aimed at comparing the behavior of these new indexes with other well-established similarity measures, over different experimental conditions. The first set of simulations is aimed at determining the behavior of the indexes when the clusterings being compared are unrelated. The second set tries to investigate the robustness to different levels of noise. The paper is organized as follows. In Sect. 2 we report the indexes recently proposed in Morlini and Zani (2012) and the similarity indexes used as benchmarks in the simulation studies. We also illustrate some of the properties of these indexes, together with theirs limitations and the

I. Morlini (✉)
Department of Economics, University of Modena and Reggio Emilia, Via Berengario 51,
41100 Modena, Italy
e-mail: isabella.morlini@unimore.it

P. Giudici et al. (eds.), *Statistical Models for Data Analysis*, Studies in Classification,
Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-00032-9_28,

implied assumptions underlying them. In Sects. 3 and 4 we report results obtained in the simulations. In Sect. 5 we give some concluding remarks.

2 The Indexes

Consider two hierarchical clusterings (or dendrograms) of the same number of objects, n. For measuring the agreement between two non trivial partitions in k clusters ($k = 2, \dots, n-1$) at a certain stage of the procedure, an important class of similarity indexes is based on the quantities T_k, U_k, P_k and Q_k reported in Table 1. This table is a (2×2) contingency table, showing the cluster membership of the $N = n(n-1)/2$ object pairs in each of the two partitions. Among the indexes defined on counting the object pairs on which the two partitions agree or disagree, the most popular ones are perhaps the Rand index:

$$R_k = \frac{N - P_k - Q_k + 2T_k}{N - U_k}, \tag{1}$$

and the criterion B_k suggested by Fowlkes and Mallows (1983):

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}. \tag{2}$$

The simple matching coefficient, formulated in terms of the quantities in Table 1, is equivalent to the Rand index, while the Jaccard coefficient is $J_k = T_k/(N - U_k$. In Morlini and Zani (2012) we have proposed the following new measure S_k:

$$S_k = \frac{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j - P_k - Q_k + 2T_k}{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j}. \tag{3}$$

The complement to one of S_k, $Z_k = 1 - S_k$, is a metric bounded in [0,1]. This metric takes value 0 if and only if the two clusterings in k groups are identical and value 1 when the two clusterings have the maximum degree of dissimilarity, that is when for each partition in k groups and for each pair i, objects in pair i are in the same group in clustering 1 and in two different groups in clustering 2 (or vice versa). The statistics B_k, J_k and S_k may be thought of as resulting from two different methods of scaling T_k to lie in the unit interval. In these indexes the pairs U_k, which are not joined in either of the two clusterings, are not considered as indicative of similarity. On the contrary, in the Rand index the counts U_k are considered as indicative of similarity. With many clusters U_k must necessarily be large and the inclusion of this count makes R_k tending to 1, for large k. How the treatment of the pairs U_k may influence so much the values of R_k, for different k, is illustrated in Wallace (1983). R_k and S_k may be related to distance measures defined on Table 1, like the

Table 1 Contingency table of the cluster membership of the N object pairs

	Second clustering ($g = 2$)		
First clustering ($g = 1$)	Pairs in the same cluster	Pairs in different clusters	Sum
Pairs in the same cluster	T_k	$P_k - T_k$	P_k
Pairs in different clusters	$Q_k - T_k$	$U_k = N - T_k - P_k - Q_k + 2T_k$	$N - P_k$
Sum	Q_k	$N - Q_k$	$N = n(n-1)/2$

Hamming distance H_k (Mirkin 1996) and the $Z_k = 1 - S_k$ distance (Morlini and Zani 2012). It can be shown that the numerator of Z_k is equal to $N(1-R_k)$ (Morlini and Zani 2012) and $H_k = 2N(1-R_k)$ (Meila 2007). Since the values R_k and S_k are not well spread out over the interval [0,1] for large k, it may be convenient to correct the indexes for association due to chance and to consider the measure (Hubert and Arabie 1985; Albatineh et al. 2006):

$$AS_k = \frac{S_k - E(S_k)}{1 - E(S_k).} \tag{4}$$

It is interesting to note that the adjusted S_k obtained with (4) is equivalent to the Adjusted Rand index (Hubert and Arabie 1985). Indeed, the expectation $E(T_k)$, assuming statistical independence under the binomial distribution for the contingency table showing the cluster membership of the object pairs (Table 1) is (Fowlkes and Mallows 1983; Hubert and Arabie 1985):

$$E(T_k) = P_k Q_k / N \tag{5}$$

Using (5), the expectation $E(S_k)$ is:

$$E(S_k) = \frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2P_k Q_k / N}{\sum_k P_k + \sum_k Q_k} \tag{6}$$

Using (6) in (4), after some algebraic simplification we obtain:

$$AS_k = \frac{2T_k - 2P_k Q_k / N}{P_k + Q_k - 2P_k Q_k / N} \tag{7}$$

which is the same expression of the Adjusted Rand Index.

The most innovative index proposed in Morlini and Zani (2012) is a global measure of similarity which considers simultaneously all the k stages in the dendrograms. In the literature, the only measure that has been presented for measuring the agreement between two whole dendrograms is the γ coefficient of Baker (1974). This criterion is defined as the rank correlation coefficient between stages at which pairs of objects combine in the dendrograms and thus it ranges over the interval $[-1, 1]$ and it is not a similarity index. The global measure of agreement proposed in Morlini and Zani (2012) is:

$$S = \frac{2\sum_k T_k}{\sum_k Q_k + \sum_k P_k}. \tag{8}$$

S does not depend on the number k and thus preserves comparability across clusterings. It has some desirable properties not pertaining to γ. It is a similarity index. Therefore, in a sample of G dendrograms $u_g \in U$, $g = 1, \ldots, G$ it is a function $S(u_g, u_{g'}) = S_{gg'}$ from $U \times U$ into $\mathbf{R}$ with the following characteristics:

- $S_{gg'} \geq 0$ for each $u_g, u_{g'} \in U$ (non negativity).
- $S_{gg} = 1$, for each $u_g \in U$ (normalization).
- $S_{gg'} = S_{g'g}$, for each $u_g, u_{g'} \in U$ (symmetry).

The further additivity property $S_{gg'} = \sum_k Vgg'_k = \sum_k \frac{2T_k}{\sum_k Q_k + \sum_k P_k}$ permits to decompose the value of the index into contributions pertaining to each stage k of the dendrograms. This makes the values of S more interpretable and comparable.

3 Simulation Experiments: Unrelated Clusterings

For the first study we generate two data sets according to the following steps:

1. For each data set, the sample size is $n = 50$ and the number of variables is $p = 5$.
2. The 50 elements in each set are generated from a multivariate standard normal distribution with a correlation matrix consisting of equal off-diagonal elements ρ_1 (in the first set) and ρ_2 (in the second set). ρ_1 and ρ_2 are chosen randomly in the set [–0.9, –0.8, …, 0.8, 0.9].
3. We repeat steps 1. and 2. 5,000 times. Each time we perform a hierarchical clustering for the two sets with the Euclidean distance and the average linkage and we compute the indexes S_k, R_k, AS_k, B_k, S and the γ coefficient.

The two sets are generated independently and the agreements between clusterings are only due to chance. Since the range of the indices is different, and in these simulations γ takes negative values, we obtain new values of γ, which we call γ^*, lying in the interval [0, 1], with the transformation $\gamma^* = (\gamma + 1)\backslash 2$. Left panel of Fig. 1 shows the boxplots of the values of S (left) and γ^* (right). The median and the mean values of γ^* are approximately 0.5. The boxplots show that S performs better than γ^*, since the median and mean value of S are nearly 0.23 and the index has fewer outliers. In the right panel are reported the mean values of B_k, R_k and AS_K, for $k = 2, \ldots, 49$. With $k = 2$, R_k and B_k have a similar value. Then, the plot shows the tendency of R_k to increase with k and rapidly approaching 1 and the opposite tendency for B_k to decrease with k and assuming values close to 0 for large k. AS_k performs best, showing average values always close to zero, regardless of k.

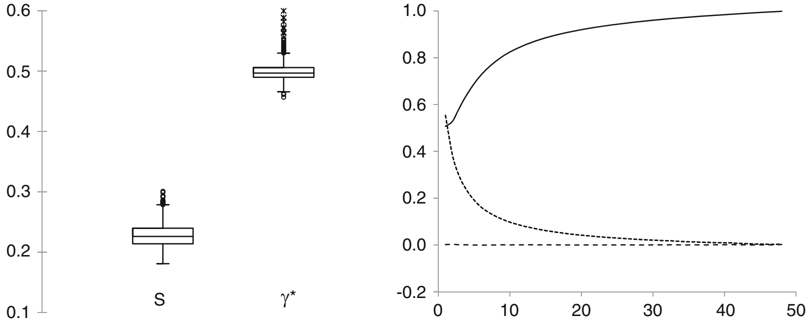

Fig. 1 Results for 5,000 pairs of unrelated samples. *Left panel*: boxplots of S (*left*) and γ^* (*right*). *Right panel*: plots of the mean values of R_k (*solid line*), B_k (*dotted line*) and AS_k (*dashed line*)

Further simulations show that the behavior of all indexes in the case of two unrelated clusterings is robust with respect to the choice of the distance or to the choice of the linkage and also with respect to the size n of the data and to the number of variables p. In several simulations carried out considering the Manhattan distance, different linkages and different values of n (from 50 to 100) and for p (from 2 to 10), boxplots for S and γ^* and plots of B_k, R_k, AS_K are similar to those reported in Fig. 1.

4 Simulation Experiments: Robustness to Noise

In this section simulations are aimed at evaluating the robustness to noise. The first data set is generated as in previous section, setting the sample size $n = 50$, the number of variables $p = 5$ and generating 50 elements from a multivariate standard normal distribution with a correlation matrix consisting of equal elements ρ_1 chosen randomly in the set $[-0.9, -0.8, \ldots, 0.8, 0.9]$. The second data set is obtained by adding to all variables a random normal noise with mean zero and variance σ_e^2. We consider the values $\sigma_e^2 = 0.04, 0.16, 0.36$. Hierarchical clusterings of each data set are carried out using the Euclidean distance and the complete, the single, the average linkages and the Ward method. Since the second data set is just the first one with added noise, indexes should indicate a great similarity between clusterings and the similarity should increase with decrease in σ_e^2. In these simulations γ assumes only positive values, therefore we consider γ instead of the normalized index γ^*. Figures 2 and 3 report the results obtained with $\sigma_e^2 = 0.04$, the single and the complete linkage methods. Results obtained with the average linkage and the Ward methods, not reported for lack of space, are available upon request. For all linkages, the values of S do not exceed 0.9 but are never smaller than 0.4 (for the single linkage, the minimum value obtained in the 5,000 runs is 0.6). On the contrary, γ assumes values greater than 0.9 and close to one but, on the other hand, presents several values smaller than 0.4. If we take the median values for comparing the

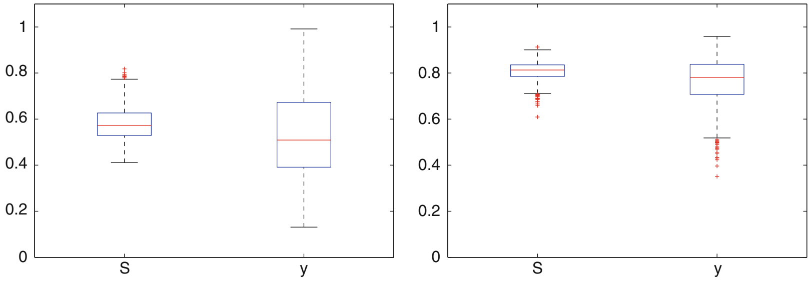

Fig. 2 Boxplots of S and γ using the complete linkage (*left*) and the single linkage (*right*)

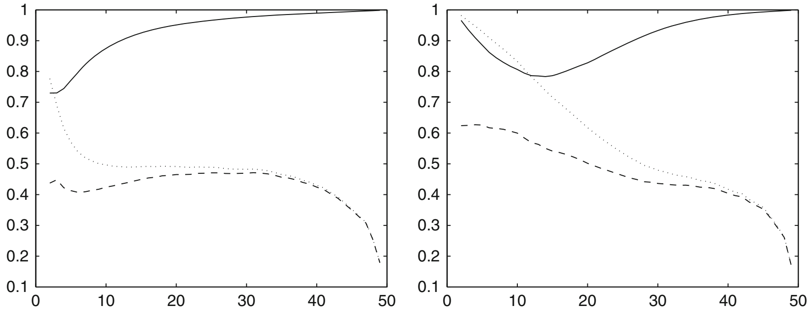

Fig. 3 Plots of the mean values of R_k (*solid line*), B_k (*dotted line*) and AS_k (*dashed line*) using the Euclidean distance and the complete linkage (*left panel*) and the single linkage (*right panel*)

degree of similarity measured by S and γ, we see that S indicates a more marked similarity using the complete and the single linkages. Plots in Fig. 3 show again the opposite tendencies of R_k to approach one and of B_k and AS_k to approach zero as k increases. The plots also show that perturbation affects B_k least for small values of k and greatest for large values of k. This desirable property was just noted in Fowlkes and Mallows (1983). For AS_k this is true for the Ward method, the single and the average linkages, but not for the complete linkage. AS_k shows a relatively more constant pattern with respect to k, without precipitous falloffs. These results show that each index has own desirable properties but also causes for concern and the choice of one index over the others is somehow difficult. That the average values of R_k and B_k are higher in the presence of small perturbation of the sample is reasonable and desirable, but the large values assumed by R_k also in presence of two unrelated clusterings (see Fig. 1) and the greatest variability of B_k across k are causes for concern. For these reasons, a global criterion of similarity like S may be a better choice for measuring the agreement between two hierarchical clusterings.

From Figs. 2 and 3 we may also analyze the stability of the different linkages to small perturbations. Clusterings with the single linkage are less affected by added noise while clusterings recovered by the complete linkage are, in general, less stable.

Figures 4, 5 and 6 show the empirical distribution of S, γ, R_k, B_k, AS_k (with $k = 2, 5, 10, 15, 20, 25, 30, 35, 40, 45$) obtained with $\sigma_e^2 = 0.16$ and $\sigma_e^2 = 0.36$.

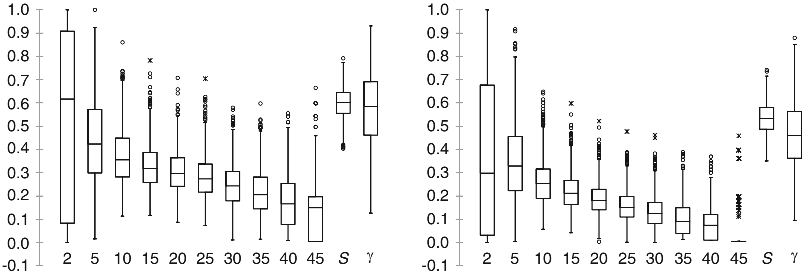

Fig. 4 Boxplots of AS_k ($k = 2, 5, 10, 15, 20, 15, 30, 35, 40, 45$), S and γ. Values are obtained considering pairs of samples where the second one is the first one with added noise with $\sigma_e^2 = 0.16$ (*left panel*), $\sigma_e^2 = 0.36$ (*right panel*)

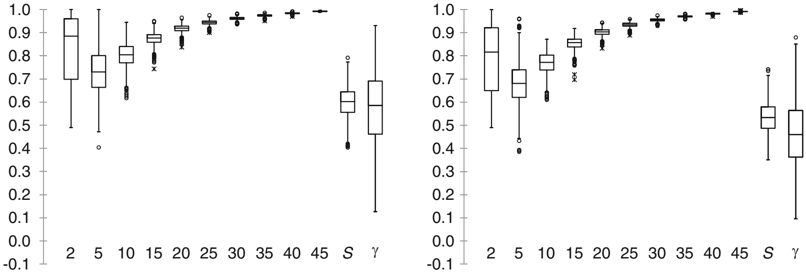

Fig. 5 Boxplots of R_k ($k = 2, 5, 10, 15, 20, 15, 30, 35, 40, 45$), S and γ. Values are obtained considering pairs of samples where the second one is the first one with added noise with $\sigma_e^2 = 0.16$ (*left panel*), $\sigma_e^2 = 0.36$ (*right panel*)

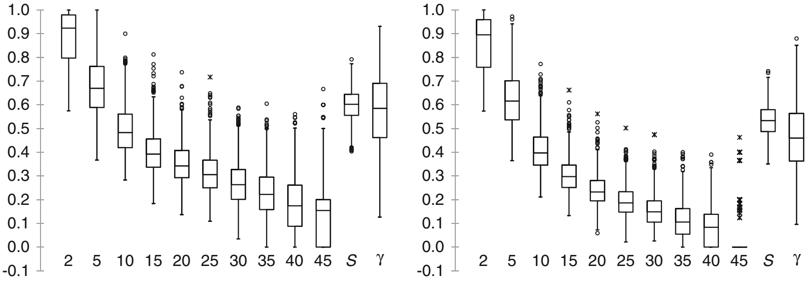

Fig. 6 Boxplots of B_k ($k = 2, 510, 15, 20, 15, 30, 35, 40, 45$), S and γ. Values are obtained considering pairs of samples where the second one is the first one with added noise with $\sigma_e^2 = 0.16$ (*left panel*), $\sigma_e^2 = 0.36$ (*right panel*)

Clusterings are recovered using the Euclidean distance and the average linkage method. The median values of S and γ decrease with increase in σ_e^2. However, this drop is more marked in γ than in S and, for $\sigma_e^2 = 0.36$, the median value of S is substantially higher than the median value of γ. The patterns of the median values of R_k, B_k and AS_k, versus k, do not change across simulations with different σ_e^2. Boxplots show that R_k has a higher variability for small values of k. For $k \geq 30$, R_k is always close to 1 and the values are nearly constant across simulations. The variability of B_k and AS_k, measured by the interquartile range, is more marked for $k = 2$ and $k = 5$.

5 Concluding Remarks

This paper has presented results obtained by simulation studies aimed at comparing the behavior of different similarity indexes used for measuring the agreement between two hierarchical clusterings. In contrast to the well-know criteria like the Rand index and the B_k index of Fowlkes & Mallows, the measure S recently proposed in the literature is not directly concerned with relationship between a single pair of partitions, but depends on the whole set of partitions in the dendrograms. Simulations show that the performances of R_k and B_k strongly depend on the number of groups k. The major drawback of this dependency is that R_k assumes values close to one for large k, even though the two partitions are unrelated. For large k, B_k has improved performances in case of unrelated clusterings but performs worse when the two clusterings are related. There is not a clear best choice between these two competing criteria and thus it is probably meaningless to search for the best criterion. A better goal is to study the behavior of these indexes and their limitations in different experimental conditions. The adjusted version of R_k and S_k, is based on a null model that is reasonable but, nevertheless, artificial. Some authors have expressed concerns at the plausibility of the null model (Meila 2007). However, simulations show that the adjusted version has improved performances and the values of the index are not influenced by k. These results are in agreement with results presented in Albatineh et al. (2006) and Albatineh and Niewiadomska-Bugaj (2011). The new global index S does not depend on k and thus preserves comparability. Simulations show that S has good performances. It takes values close to zero when no clustering structure is present and values close to one when a structure exists.

References

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction fore chance agreement. *Journal of Classification, 23*, 301–313.

Albatineh, A. N., & Niewiadomska-Bugaj, M. (2011). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification, 5*, 179–200.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques. Case I: sensitivity to data errors. *JASA, 69*, 440–445.

Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *JASA, 78*, 553–569.

Hubert, L. J. & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Meila, M. (2007). Comparing clusterings. An information based distance. *Journal of Multivariate Analysis, 98*(5), 873–895.

Mirkin, B. G. (1996). *Mathematical classificationa and clustering*. Dordrecht: Kluwer Academic.

Morlini, I., & Zani, S. (2012). An overall index for comparing hierarchical clusterings. In W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & J. Kunze (Eds.) *Challenges at the interface of data analysis, computer science and optimization* (pp. 29–36). Berlin: Springer.

Wallace, D. L. (1983). Comment on the paper "A method for comparing two hierarchical clusterings". *JASA, 78*, 569–578.