Variable Selection in Cluster Analysis: An Approach Based on a New Index

Isabella Morlini and Sergio Zani

Abstract In cluster analysis, the inclusion of unnecessary variables may mask the true group structure. For the selection of the best subset of variables, we suggest the use of two overall indices. The first index is a distance between two hierarchical clusterings and the second one is a similarity index obtained as the complement to one of the previous distance. Both criteria can be used for measuring the similarity between clusterings obtained with different subsets of variables. An application with a real data set regarding the economic welfare of the Italian Regions shows the benefits gained with the suggested procedure.

1 Introduction

In cluster analysis, the inclusion of 'noisy' variables may mask the recovery of the true underlying structure. In the literature, various procedures aimed at determining the best subset of variables have been proposed, both in the context of model-based and not-model-based clustering (Fowlkes et al., 1988; Gnanadesikan et al., 1995; Montanari and Lizzani, 2001; Tadesse et al., 2005; Raftery and Dean, 2006; Fraiman et al., 2008; Steinley and Brusco, 2008). In this paper we propose a new approach, based on an overall index measuring the distance between two hierarchical clusterings. This criterion is novel since it is applied directly to the whole hierarchies and may be thought of as a generalization of the measures used

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

S. Zani
Department of Economics, University of Parma, Via Kennedy 6, 43100, Parma, Italy
e-mail: sergio.zani@unipr.it

A. Giusti et al. (eds.), *Classification and Data Mining*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-642-28894-4_9,

for comparing two partitions (Rand, 1971; Fowlkes and Mallows, 1983; Hubert and Arabie, 1985). The paper is organized as follows: in Sect. 2 we define the index, we present its properties and its decomposition with reference to each stage of the hierarchy; in Sect. 3 we consider the similarity index obtained as the complement to one of the suggested distance and we deal with the adjustment for agreement due to chance; in Sect. 4 we describe the use of the index for measuring the similarity between clusterings obtained with different subsets of variables, following a forward and a backward approach; in Sect. 5 we present results on a real data set.

2 The Index and Its Properties

Suppose we have two hierarchical clusterings of the same number of objects, n. Let us consider the $N = n(n-1)/2$ pairs of objects and let us define, for each non trivial partition in k groups ($k = 2, \ldots, n-1$), a binary variable X_k with values $x_{ik} = 1$ if objects in pair $i (i = 1, \ldots, N)$ are classified in the same cluster in partition in k groups and $x_{ik} = 0$ otherwise. A binary $(N \times (n-2))$ matrix $\mathbf{X}_g$ for each clustering g $(g = 1, 2)$ may be derived, in which the columns are the binary variables X_k. A global measure of dissimilarity between the two clusterings may be defined as follows:

$$Z = \frac{\| \mathbf{X}_1 - \mathbf{X}_2 \|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|}, \tag{1}$$

where $\| \mathbf{A} \| = \sum_i \sum_k \| a_{ik} \|$ is the L_1 norm of the matrix $\mathbf{A}$. In (1) the matrices involved take only binary values and the L_1 norm is equal to the square of the L_2 norm. The derivation of Z uses the Rand's idea of considering the N object pairs. However, Z is a new index since it is applied to a whole hierarchy and not only to a single partition. Z has the following properties.

- It is bounded in [0,1]. $Z = 0$ iff the two hierarchical clusterings are identical and $Z = 1$ when the clusterings have the maximum degree of dissimilarity, that is when for each partition in k groups and for each i, objects in pair i are in the same group in clustering 1 and in different groups in clustering 2 (or vice versa).
- It is a distance, since it satisfies the conditions of non negativity, identity, symmetry and triangular inequality (Zani, 1986).
- The complement to 1 of Z is a similarity measure, since it satisfies the conditions of non negativity, normalization and symmetry.
- It does not depend on the group labels since it refers to pairs of objects.
- It may be decomposed in $(n-2)$ parts related to each pair of partitions in k groups since:

$$Z = \sum_k Z_k = \sum_k \sum_i \frac{|x_{1ik} - x_{2ik}|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|}. \tag{2}$$

  The plot of Z_k versus k shows the distance between the two clusterings at each stage of the procedure.

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	$Q_k - T_k$	$U_k = N + T_k - P_k - Q_k$	$N - P_k$
Sum	Q_k	$N - Q_k$	$N = n(n-1)/2$

3 The Complement of the Index

Consider the quantities in the (2×2) contingency table showing the cluster membership of the object pairs in each of the two partitions (Table 1).

Since $\| \mathbf{X}_1 \| = \sum_k Q_k$ and $\| \mathbf{X}_2 \| = \sum_k P_k$, the complement to 1 of Z is:

$$S = 1 - Z = \frac{2\sum_k T_k}{\sum_k Q_k + \sum_k P_k}. \tag{3}$$

Also the similarity index S may be decomposed in $(n-2)$ parts V_k related to each pair of partitions in k groups:

$$S = \sum_k V_k = \sum_k \frac{2T_k}{\sum_k Q_k + \sum_k P_k}. \tag{4}$$

The components V_k, however, are not similarity indices for each k since they assume values < 1 even if the two partitions in k groups are identical. For this reason, we consider the complement to 1 of each Z_k in order to obtain a single similarity index for each pair of partitions:

$$S_k = 1 - Z_k = \frac{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j - P_k - Q_k + 2T_k}{\sum_{j=2}^{n-1} P_j + \sum_{j=2}^{n-1} Q_j} = \frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2T_k}{\sum_j P_j + \sum_j Q_j}. \tag{5}$$

A similarity index between two partitions may be adjusted for agreement due to chance (Hubert and Arabie, 1985; Albatineh et al., 2006; Warrens, 2008). With reference to formula (5) the adjusted similarity index AS_k has the form:

$$AS_k = \frac{S_k - E(S_k)}{max(S_k) - E(S_k)}. \tag{6}$$

Under the hypothesis of independence of the two partitions, the expectation of T_k in Table 1 $E(T_k) = P_k Q_k / N$. Therefore, the expectation of S_k is given by:

$$E(S_k) = \frac{\sum_{j \neq k} P_j + \sum_{j \neq k} Q_j + 2P_k Q_k / N}{\sum_j P_j + \sum_j Q_j}. \tag{7}$$

Considering $max(S_k) = 1$ and simplifying terms we obtain:

$$AS_k = \frac{2T_k - 2P_k Q_k / N}{P_k + Q_k - 2P_k Q_k / N}. \tag{8}$$

The adjusted Rand index for two partitions in k groups is given by Warrens (2008):

$$AR_k = \frac{2(NT_k - P_k Q_k)}{N(P_k + Q_k) - 2P_k Q_k}, \tag{9}$$

and so AS_k is equal to the Adjusted Rand Index.

4 Criteria for Variable Selection

Indexes Z and S can be used for variable selection in cluster analysis (Fowlkes et al., 1988; Fraiman et al., 2008; Steinley and Brusco, 2008). The inclusion of 'noisy' variables can actually degrade the ability of the clustering procedures to recover the true underlying structure (Friedman and Meulman, 2004). For a set of p variables and a certain clustering method, we suggest three different approaches, suitable for data sets with tens of variables. Variable selection in data sets containing hundreds or thousands of variables (like gene expression data) is not considered in this paper.

First we may obtain the p one dimensional clusterings with reference to each single variable and then compute the $p \times p$ similarity matrix $\mathbf{S}$. The pairs of variables reflecting the same underlying structure show high similarity. On the contrary, the noisy variables should present a similarity with the other variables near to the expected value for chance agreement. We may select a subset of variables that best explains the classification into homogeneous groups. These variables help us to better understand the multivariate structure and suggest a dimension reduction that can be used in a new data set for the same problem (Tadesse et al., 2005).

Next we may find the similarities between clusterings obtained with subsets of variables (regarding, for example, different features). This approach is helpful in showing aspects that lead to similar partitions and subsets of variables that, on the contrary, lead to different clusterings.

A third way to proceed consists in finding the similarities between the 'master' clustering obtained by considering all the variables and the clusterings obtained by eliminating each single variable in turn, in order to highlight the 'marginal' contribution of each variable to the master structure.

5 An Application to a Real Data Set

We consider the 20 Italian regions and the following 9 variables measuring different aspects of the economic wealth: $X_1 =$ activity rate, $X_2 =$ unemployment rate, $X_3 =$ youth unemployment rate, $X_4 =$ family average income, $X_5 =$ family median income, $X_6 =$ income Gini concentration index, $X_7 = \%$ of poor families, $X_8 = \%$ of people dissatisfied for their economic conditions, $X_9 = \%$ of families with inadequate income. We standardize variables to zero mean and unit variance before

Table 2 Values of S between pair of clusterings of the Italian regions data set

	Euclidean distance					Manhattan distance				
Method	Average	Complete	Single	Ward	Centroid	Average	Complete	Single	Ward	Centroid
Average	1	0.80	0.90	0.76	0.96	0.96	0.81	0.88	0.79	0.95
Complete	0.80	1	0.73	0.72	0.78	0.83	0.98	0.72	0.82	0.78
Single	0.90	0.73	1	0.71	0.93	0.87	0.73	0.92	0.72	0.90
Ward	0.76	0.72	0.71	1	0.75	0.78	0.73	0.68	0.79	0.74
Centroid	0.96	0.78	0.93	0.75	1	0.92	0.79	0.88	0.77	0.95
Average	0.96	0.83	0.87	0.78	0.92	1	0.84	0.87	0.83	0.94
Complete	0.81	0.98	0.73	0.73	0.79	0.84	1	0.72	0.82	0.78
Single	0.88	0.72	0.92	0.68	0.88	0.87	0.72	1	0.72	0.93
Ward	0.79	0.82	0.72	0.79	0.77	0.83	0.82	0.72	1	0.78
Centroid	0.95	0.78	0.90	0.74	0.95	0.94	0.78	0.93	0.78	1

applying hierarchical cluster analysis with different distances and different methods. We compute the S index for each pair of clusterings. Results, reported in Table 2, show that, in general, clustering remains stable varying distances or methods or both (all pairwise similarity indexes take values greater than 0.7). The fact that the clustering does not change appreciably leads to the evidence that the topologies of the trees are natural and are not simply artifacts of the algorithms. Analyzing the values of the pairwise similarities, we note that the Ward and the single linkage seem to behave a little bit differently from the other methods, while the complete linkage, the average linkage and the centroid method seem to be more similar to each other. The global measure of similarity S may be decomposed in parts related to each partition in $k = 2, \ldots, 18$ groups. As an example, Table 3 presents the values of S_k and AS_k for two pairs of clusterings. This table shows the reason why the second couple has a slightly less similarity. In these two dendrograms, 12 partitions (among the 18 ones) are exactly the same while for the first two dendrograms the identical partitions are 13. In order to determine the 'true' number of clusters, we may count the couples of clusterings in which each partition in k groups is identical. From counts reported in Table 4 we see that the partition in 2 groups remains identical in 36 clusterings. Only partition in 18 clusters has a larger count. This may be taken as evidence that partition in two groups comes naturally from data and is not driven by the algorithm. In this partition, northern and central regions are separated from southern regions.

Table 5 reports the values of S between the clustering obtained considering all variables ($\{X_i\}_{i=1,\ldots,9}$) (in the following we will refer to this tree as the overall tree) and the clusterings obtained eliminating each variable in turn. In the table, the column or row header $\{X_i\}_{i\neq j}$ indicates the subset of variables without X_j. For example, $\{X_i\}_{i\neq 1}$ is the subset $\{X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9\}$. The Euclidean distance and the average method are used for obtaining partitions.

If we eliminate X_9, the clustering remains identical. This means that X_9 has no 'marginal' contribution to the overall clusterings, given the other variables. X_8 is the variable which seems to have the major marginal influence to the overall

Table 3 Values of S_k and AS_k for two pairs of clusterings of the Italian regions data set

	Number k of clusters																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
	Similarity between partitions obtained with Euclidean distance and the average method and partitions obtained with Euclidean distance and the centroid method																		
S_k	1	1	1	1	0.98	0.98	0.98	1	0.98	1	1	1	1	1	0.98	1	1	1	
AS_k	1	1	1	1	0.86	0.75	0.86	1	0.81	1	1	1	1	1	0.66	1	1	1	
	Similarity between partitions obtained with Euclidean distance and the average method and partitions obtained with Manhattan distance and the centroid method																		
S_k	1	1	1	1	0.98	0.97	0.98	1	1	1	1	0.98	1	1	0.98	1	1	0.98	
AS_k	1	1	1	1	0.86	0.73	0.86	1	1	1	1	0.57	1	1	0.66	1	1	0.00	

Table 4 Counts of pairs of clusterings in which each partition in k groups is identical

k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
n. of pairs	36	29	8	8	7	3	1	6	6	8	12	9	17	29	21	16	45	20

Table 5 Values of S for couples of clusterings obtained with different subsets of variables

	$\{X_i\}_{i\neq 1}$	$\{X_i\}_{i\neq 2}$	$\{X_i\}_{i\neq 3}$	$\{X_i\}_{i\neq 4}$	$\{X_i\}_{i\neq 5}$	$\{X_i\}_{i\neq 6}$	$\{X_i\}_{i\neq 7}$	$\{X_i\}_{i\neq 8}$	$\{X_i\}_{i\neq 9}$	$\{X_i\}$
$\{X_i\}_{i\neq 1}$	1	0.96	0.83	0.85	0.88	0.86	0.93	0.84	0.89	0.89
$\{X_i\}_{i\neq 2}$	0.96	1	0.86	0.88	0.86	0.87	0.95	0.85	0.91	0.91
$\{X_i\}_{i\neq 3}$	0.83	0.86	1	0.83	0.82	0.87	0.84	0.81	0.89	0.89
$\{X_i\}_{i\neq 4}$	0.85	0.88	0.83	1	0.97	0.86	0.86	0.80	0.89	0.89
$\{X_i\}_{i\neq 5}$	0.88	0.86	0.82	0.97	1	0.85	0.85	0.80	0.89	0.89
$\{X_i\}_{i\neq 6}$	0.86	0.87	0.87	0.86	0.85	1	0.85	0.84	0.94	0.94
$\{X_i\}_{i\neq 7}$	0.93	0.95	0.84	0.86	0.85	0.85	1	0.85	0.88	0.88
$\{X_i\}_{i\neq 8}$	0.84	0.85	0.81	0.80	0.80	0.84	0.85	1	0.86	0.86
$\{X_i\}_{i\neq 9}$	0.89	0.91	0.89	0.89	0.89	0.94	0.88	0.86	1	1
$\{X_i\}$	0.89	0.91	0.89	0.89	0.89	0.94	0.88	0.86	1	1

clustering structure. The value of S between $\{X_i\}_{i\neq 4}$ and $\{X_i\}_{i\neq 5}$ ($S = 0.97$) shows that X_4 and X_5, as one would expect, bring the same marginal contribution. We may also consider the similarities between the clustering recovered by all variables $\{X_i\}$ and the clusterings obtained by using each single variable. The values of S are:

$$S(\{X_i\},\{X_1\}) = 0.74, \quad S(\{X_i\},\{X_2\}) = 0.66, \quad S(\{X_i\},\{X_3\}) = 0.58,$$
$$S(\{X_i\},\{X_4\}) = 0.55, \quad S(\{X_i\},\{X_5\}) = 0.76, \quad S(\{X_i\},\{X_6\}) = 0.69,$$
$$S(\{X_i\},\{X_7\}) = 0.77, \quad S(\{X_i\},\{X_8\}) = 0.52, \quad S(\{X_i\},\{X_9\}) = 0.53.$$

None of the values are particularly high and thus the clustering recovered with all variables seems to derive from a multivariate effect and not to be dominated by the univariate effect of a single variable. As shown in Fig. 1, variables X_1, X_5 and

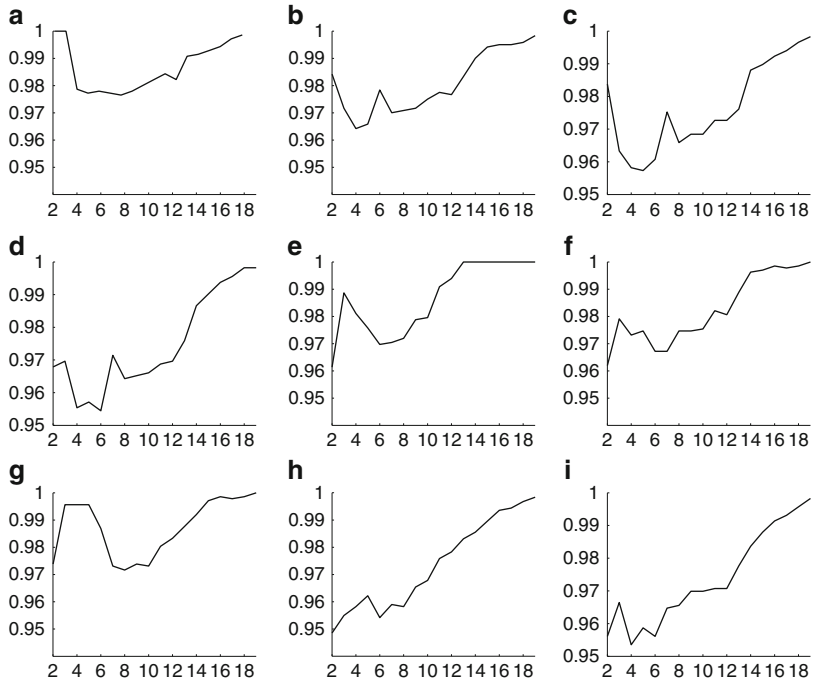

Fig. 1 Values of S_k for the partitions obtained with all variables and the partitions obtained with (**a**) X_1, (**b**) X_2, (**c**) X_3, (**d**) X_4, (**e**) X_5, (**f**) X_6, (**g**) X_7, (**h**) X_8 and (**i**) X_9

X_7 have a peak of the similarity values S_k for $k = 3$. X_1 is in perfect agreement also for $k = 2$ while X_7 for $k = 4, 5$. Variables X_4, X_6 and X_9 have a different S_k pattern, but they also have a peak for $k = 3$. On the contrary, the peak for X_2 and X_3 is for $k = 2$. Thus, in this case, the choice for the 'correct' number of clusters is somehow difficult, since both $k = 2$ and $k = 3$ seem to be good alternative. Figure 1 also shows that variables which have the smaller values in the similarity S, like X_3, X_4, X_5 and X_6, exhibit a less agreement to the overall clusters for small numbers k of groups. The patterns of S_k for these variables display smaller values for $k < 12$. For the other variables, S_k increase less rapidly, with respect to k. Finally, we study the behavior of three subsets of variables, each one related to a specific feature of the economic situation. We consider subset $\{X_1, X_2, X_3\}$, related to the demographic structure, subset $\{X_4, X_5, X_6\}$ related to the income structure and subset $\{X_7, X_8, X_9\}$, related to the relative and the perceived poverty. The similarities between the cluster trees of each subset and of all variables are: $S(\{X_i\}, \{X_{1,2,3}\}) = 0.76$, $S(\{X_i\}, \{X_{4,5,6}\}) = 0.66$, $S(\{X_i\}, \{X_{7,8,9}\}) = 0.78$. The similarities between clusterings of each subsets are: $S(\{X_{4,5,6}\}, \{X_{7,8,9}\}) = 0.59$, $S(\{X_{1,2,3}\}, \{X_{4,5,6}\}) = 0.61$, $S(\{X_{1,2,3}\}, \{X_{7,8,9}\}) = 0.62$. Here again we note that none of the three subsets reveals a clustering very similar to the clustering obtained with all the variables. All the three aspects of the economic health seem equally

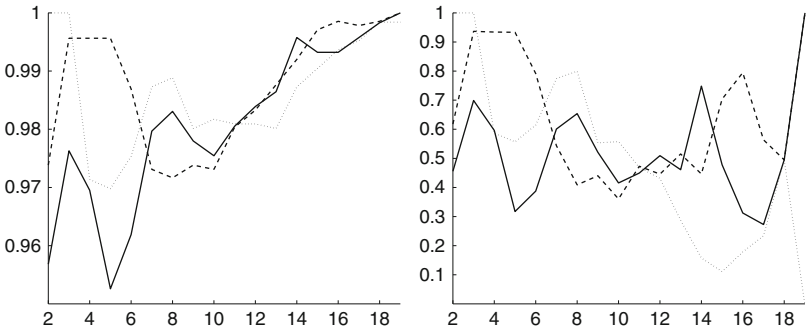

Fig. 2 Plots of S_k (*left*) and AS_k (*right*) between partitions obtained with all variables $\{X_i\}$ and subset $\{X_1, X_2, X_3\}$ (*dotted line*), subset $\{X_4, X_5, X_6\}$ (*solid line*), subset $\{X_7, X_8, X_9\}$ (*dashed line*)

to contribute to the overall clustering. Figure 2 reports the plots of S_k and AS_k. The scales in the Y-axis are different. However, the patterns of S_k and AS_k are nearly identical, for $k \leq 12$. It is a desirable property that the correction for the chance influences the values but not the configuration of the plot for small k. For large k, as one would expect, the correction for chance do influence the patterns of the index and S_k tends to one while AS_k tends to zero. We note that, for example, the configuration in two groups is largely dominated by the demographic structure, while configurations in 3, 4 and 5 clusters are mostly influenced by the perceived poverty.

References

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

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

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification, 5*, 205–228.

Fraiman, R., Justel, A., & Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *JASA, 103*, 1294–1303.

Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subset of attributes. *Journal of the Royal Statistical Society B, 66*, 815–849.

Gnanadesikan, R., Kettering, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification, 12*, 113–136.

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

Montanari, A., & Lizzani, L. (2001). A projection pursuit approach to variable selection. *Computational Statistics and Data Analysis, 35*, 463–473.

Raftery, A. E., & Dean, N. (2006). Variable selection for model based clustering. *JASA, 101*, 168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *JASA, 66*, 846–850.
Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika, 73*, 125–144.
Tadesse, M. G., Sha, N., & Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *JASA, 100*, 602–617.
Warrens, M. J. (2008). On the equivalence of Cohen's Kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25*, 177–183.
Zani, S. (1986). Some measures for the comparison of data matrices. In *Proceedings of the XXXIII Meeting of the Italian Statistical Society* (pp. 157–169), Bari, Italy.