DOI:10.2478/bile-2013-0010

Biometrical Letters
Vol. 49 (2012), No. 2, 135-147

A Global Approach to the Comparison of Clustering Results

Osvaldo Silva[1], Helena Bacelar-Nicolau[2], Fernando C. Nicolau[3]

[1]University of Azores, Department of Mathematics, CMATI, 9501-855-Ponta Delgada, Portugal, osilva@uac.pt
[2]University of Lisbon, Faculty of Psychology, Laboratory of Statistics and Data Analysis 1649-013-Lisboa, Portugal, and DataScience, hbacelar@fp.ul.pt
[3]New University of Lisbon, FCT, Department of Mathematics, 2829-516-Caparica, Portugal, and DataScience, geral@datascience.org

SUMMARY

The discovery of knowledge in the case of Hierarchical Cluster Analysis (HCA) depends on many factors, such as the clustering algorithms applied and the strategies developed in the initial stage of Cluster Analysis. We present a global approach for evaluating the quality of clustering results and making a comparison among different clustering algorithms using the relevant information available (e.g. the stability, isolation and homogeneity of the clusters). In addition, we present a visual method to facilitate evaluation of the quality of the partitions, allowing identification of the similarities and differences between partitions, as well as the behaviour of the elements in the partitions. We illustrate our approach using a complex and heterogeneous dataset (real horse data) taken from the literature. We apply HCA based on the generalized affinity coefficient (similarity coefficient) to the case of complex data (symbolic data), combined with 26 (classic and probabilistic) clustering algorithms. Finally, we discuss the obtained results and the contribution of this approach to gaining better knowledge of the structure of data.

Keywords: Cluster Analysis, VL Methodology, Affinity Coefficient, Comparing Partitions, Cluster Stability and Cluster Validation

1. Introduction

Partition evaluation and comparison of partitions within the scope of Hierarchical Cluster Analysis (HCA) is of great importance, because depending on the comparison coefficients between elements and between clusters and on the strategies developed in the initial stage of the Cluster Analysis, different results can be obtained. The comparison of information from different sources and of the results obtained is a difficult task, particularly when there is no previous knowledge about the data. However, it is important to try to answer questions such as: *i) How to compare partitions obtained using different cluster algorithms or several resamples from the first set of data? ii) Is it possible to join information from several approaches in the decision-making process of choosing the most representative partition?*

Since there are many indicators for the evaluation of clusters, partitions and hierarchical classifications, which may differ substantially in the type of information or in the range of variation of their values, it is useful to take a global indicator which makes it possible to unify and summarize different indexes into a single indicator (Silva et al., 2010; Silva, 2011).

All the difficulties underlying partition evaluation and comparison are the main reason for the development of a global methodology for the evaluation and the validation of the obtained clustering structures, taking into account the use of different indexes and sets of partitions.

Section 2 contains a set of different indexes to evaluate the quality of the clustering structures and global indexes based on linear combinations of some of them. In addition, in Section 2.3 we present a visual approach that allows quick perception of the quality of clustering structures.

2. Methodological Framework

This section provides a methodology for the evaluation of the results of a Cluster Analysis, with particular emphasis on evaluation/validation and comparison of partitions. This methodology is based on a set of indicators for evaluating the results (set of partitions) of different clustering algorithms and the most relevant information available about the data. To assess the quality of the results of a Cluster Analysis, particularly the partition considered the most suitable (the most significant partition), the global approach comprises the following steps:

1) From the original data, different classifications are obtained using several algorithms and the most significant partition (according to several validation indexes of partitions) is noted;
2) Based on the set of partitions obtained by different algorithms, which contain the same number of clusters as the most significant partition, some indexes are calculated, according to the method described in Section 2.1**;**
3) Global indicators (see Section 2.2) are calculated;
4) The visualization method described in Section 2.3 is applied.

2.1. Some quality indexes

Let E={$x_1,\ldots,x_m$} be a set of elements to be classified and $CP=\{P_1, P_2,\ldots,P_t\}$ a set of t partitions, where $P_i=\{c_{i1}, c_{i2}, \ldots, c_{ik}\}$ is a partition containing k clusters. Let $c_{P_i}(h)$, with i=1,...,t and h=1,...,m, be the cluster of P_i which contains h, that is, $c_{P_i}(h) \subset P_i$, $P_i \subset CP$, $h \in c_{P_i}(h)$. The resemblance between these two clusters c_{ix} and c_{jy} from two partitions P_i and P_j, respectively, can be evaluated using the affinity coefficient (e.g. Bacelar-Nicolau, 1988):

$$Af\,(c_{ix},\, c_{jy}) = \frac{card(c_{ix} \cap c_{jy})}{\sqrt{card(c_{ix})card(c_{jy})}} \tag{1}$$

where $card$ represents the number of elements of the cluster under analysis. $Af\,(c_{ix}, c_{jy}) = Af\,(c_{jy}, c_{ix})$ and $0 \leq Af\,(c_{ix}, c_{jy}) \leq 1$. Based on the formula (1), the stability of the element h may be evaluated by:

$$E(h) = \frac{\sum_{i=1}^{t} \sum_{j=i+1}^{t} Af\left(c_{P_i}(h),\, c_{P_j}(h)\right)}{t(t-1)/2} \tag{2}$$

The stability of each element h gives us the notion of the permanence of this element in the classes to which it belongs in each of the partitions under comparison. This index varies between 0 and 1, allowing us to assess the degree of stability of each of the elements to be classified, taking into account the set of t partitions.

The SIL modified index (*SIL (h)*), corresponding to formula (3), is based on the silhouette index of Rousseeuw (Gordon, 1999) and, like the latter, takes into consideration the homogeneity and isolation of each of the elements.

$$SIL(h)=\frac{\frac{1}{n_r-1}\sum_{g\in\{C_r \backslash h\}} s_{hg}-\frac{1}{n-n_r}\sum_{\substack{g\in C_s\\ g\notin C_r}} s_{hg}}{max\left\{\frac{1}{n_r-1}\sum_{g\in\{C_r \backslash h\}} s_{hg},\frac{1}{n-n_r}\sum_{\substack{g\in C_s\\ g\notin C_r}} s_{hg}\right\}} \tag{3}$$

In formula (3), n_r is the number of elements in the cluster Cr and S_{hg} is the index of similarity between the elements *h* and *g*. The first part of the numerator is a measure of the resemblances between the element *h* and all the other elements of the cluster (c_r) to which the element *h* belongs. The second part of the numerator is the average of the resemblances between one element *h* and all other elements which do not belong to the cluster to which the element *h* belongs. This index also varies between 0 and 1 and carries the sense of the magnitude with which an element is inserted into the class to which it belongs.

Let P_i be a partition obtained at the *k* level (cut level) of a dendrogram which corresponds to a stage of the constitution of the partition hierarchy. Let $STAT(P_i)$ be the global statistic of levels (Bacelar-Nicolau, 1980; Lerman, 1981), which measures the information given by the corresponding partition, relative to the initial preordination associated with the applied index of (dis)similarity, being expressed by :

$$STAT(P_i)=\frac{card(w\cap(R\times S))-\frac{card(R\times S)}{2}}{\sqrt{card(RxS)(card(F)+1)/12}} \tag{4}$$

where F is the total number of pairs in the partition, $w=\{(hg,kl)\in F\times F{:}s_{hg}\geq s_{kl}\}$ represents the graph of the initial preordination defined in $F\times F$, and R and S are the sets of pairs respectively assembled and separated in the partition under analysis. To make it possible to compare the partitions obtained from different algorithms and/or from different resamples of the initial set of data, we also take into consideration the following index:

$$STATnor(P_i) = \frac{STAT(P_i)}{\underset{1 \le j \le t}{Max}\left\{STAT\left(P_j\right)\right\}} \tag{5}$$

The index STATnor(Pi) corresponds to an overall normalization of the statistical levels, so as to vary between 0 and 1.

In our methodological framework we can also use information about the values of the fusion coefficient for each of the obtained dendrograms.

2.2. Global indexes

In this subsection some indexes used to evaluate the overall quality of the results of a Cluster Analysis are defined and analysed, taking into account the most important properties such as stability, isolation and homogeneity of classes. It is intended that the results presented below assess each of the elements to be classified, as a whole, as well as each class and each of the partitions under consideration. These indexes are based on the results presented in Subsection 2.1.

Let h be one of the m elements to be classified; let c be one of the clusters of a partition into k clusters; let P be a one of the t partitions to be evaluated, and let CP be the set of partitions. Assuming that q indexes are relevant for assessing the quality of the results of a Cluster Analysis, one can set a global indicator, ranging between 0 and 1, to assess whether an element/ cluster/partition is better, worse or similar to another element/cluster/partition and how much it is so.

Let the index $j=1,\ldots,q$ denote the components to be used, which vary between 0 and 1, and U be the set of three cases: $U=h$ or $U=c$ or $U=P$, depending on whether one is evaluating an element, a cluster or a partition.

Based on this assessment, the overall index *Glob_Ind*(*U*) can be defined by the following formula:

$$Glob\_Ind(U) = \sum_{j=1}^{q} \alpha_j S_j(U), \tag{6}$$

where the weights $\alpha_1,\ldots,\alpha_q$ are all nonnegative and their sum is equal to unity, i.e. $\alpha_j \geq 0$, with $j=1,\ldots,q$, and $\sum_{j=1}^{q} \alpha_j = 1$. Thus the overall indicator *Glob_Ind*(*U*) corresponds to a convex linear combination of the various component indexes $S_j(U)$, for $j=1,\ldots,q$, all of which measure similarity/agreement, with $S_j(U)$ and q defined appropriately, depending on $U=h$ or $U=c$ or $U=P$, as described above.

An indicator of this nature seeks to compare the performances of the elements/clusters/partitions provided by the different algorithms applied, using information from the indexes which are considered to be the most relevant (i.e. component indexes). Each of the component indexes takes values that may initially vary on very different scales, and need to be transformed so as vary between 0 and 1, so that they can serve as a basis for calculating the overall index *Glob_Ind*(*U*), with the possibility of assigning different weights to each of the component indexes $S_j(U)$, $j=1,\ldots,q$.

For each element h in the set of t partitions, the overall indicator can be defined by following expression:

$$Glob\_Ind\ (h) = \ \alpha_1 \times S_1(h) + \alpha_2 \times S_2(h)\ , \tag{7}$$

with $S_1(\mathrm{h}) = E(h)$ and $S_2(h) = t^{-1} \sum_{i=1}^{t} sil_{Pi(h)}$, where $E(h)$ is given by formula (2) and $Sil_{Pi(h)}$ is the value of $Sil(h)$ relative to partition P_i, obtained by formula (3). The weights α_k, $k=1, 2$ may be equal or different according to the objectives set beforehand.

The overall indicator of a cluster c of the set of t partitions is given by:

$$Glob\_Ind\ (c) = \ \alpha_1 \times S_1(c) + \alpha_2 \times S_2(c)\ , \tag{8}$$

with $S_1(\mathrm{c}) = E(c)$ and $S_2(\mathrm{c}) = Sil(c)$, where $E(c)$ and $Sil(c)$ are respectively the mean values of $E(h)$ and $Sil(h)$ with $h \in c$.

The overall indicator for each of the partitions, $P_i \subset CP$, i=1,…,t, is defined by the following expression:

$$Glob\_Ind(P_i) = \alpha_1 \times S_1(P_i) + \alpha_2 \times S_2(P_i) + \alpha_3 \times S_3(P_i) + \alpha_4 \times S_4(P_i), \tag{9}$$

with $S_1(P_i) = E(P_i) = (t-1)^{-1}\sum_{j=1,\, j\neq i}^{t} IC(P_i,P_j)$, in which $IC(P_i, P_j) = m^{-1}\sum_{h=1}^{m} Af(c_{P_i(h)}, c_{P_j(h)})$; $S_2(\mathrm{P_i}) = Sil(\mathrm{P_i})$, which corresponds to the mean values of $Sil(c)$ with $c \in P_i$; $S_3(\mathrm{P_i}) = CF(\mathrm{P_i})$, in which CF($P_i$) designates the fusion coefficient of partition P_i and $S_4(\mathrm{P_i}) = STATnor(\mathrm{P_i})$, which is obtained from formulae (4) and (5).

Formula (9) takes into account a set of characteristics that help us to compare the partitions for each of the methods through a multi-focal perspective, thus providing a more comprehensive comparison between them and the choice of the most appropriate method for the data being analysed. Note that in this formula other partition quality indexes IQ_P may be used; simple (for example the index of silhouettes – SIL), or complex, based on a combination of values of several quality indexes, as long as these values have been previously standardized.

The choice of the component indexes $S_j(U)$, j=1,…,q, their transformations and appropriate weights may play a relevant role in the quality of the overall indicator, since the weights reflect the importance attached to each index in assessing the overall quality of the element/cluster/partition being evaluated. The judgments implicit in the choice of weights should be clear and understandable, and it is important to assess to what extent they influence the results. This overall indicator is part of the overall methodology, which aims to evaluate the results of a Cluster Analysis from a multidimensional perspective, that is, taking into account various relevant factors that require consideration.

2.3. Visual method

In this section a visualization method that uses a graduation of patterns is explained, in order to provide a quick and global perception of the quality of the results. Based on this method, regardless of how the partitions were obtained (resampling methods and/or different algorithms), some information can be extracted which provides more details about the similarities/differences between the partitions and the behaviour of the elements in the partitions under analysis. The similarities/differences between partitions can be observed whether considering the whole partition, as when looking at the level of classes, or observing the colour patterns of the visualization scheme.

Partitions are regarded as similar if in each of them the same colours are linked to the same elements. Homogeneity regarding a particular pattern in different partitions indicates a similar behaviour, while a mixture of patterns indicates different behaviour. If the partitions come from different algorithms, visualization can help distinguish the most appropriate algorithm to identify each of the clusters.

This method is specially useful if we do not have any information on the classification, and we can get a quick perception of the quality of results obtained for each element of each cluster and each partition on each of the t methods used (or r resamples) and also the set of partitions obtained (by different methods or resampling) based only on information intrinsic to the data and methods to be used in the Cluster Analysis.

Variation interval	Graduation of colours	Quality degree
[0, 0.20[		Very weak
[0.20, 0.40[		Weak
[0.40, 0.60[		Reasonable
[0.60, 0.80[		Good
[0.80, 1]		Very good

Figure 1. A visual representation based on a grading scale of patterns

The values for the indicators $Glob\_Ind(h)$, $Glob\_Ind(c)$ and $Glob\_Ind(P_i)$ corresponding respectively to formulas (7), (8) and (9) of Subsection 2.2 provide an indication of the quality of the results that facilitates comparison between the various partitions obtained and the choice of the most appropriate partition (Silva et al., 2010). It is considered appropriate to use the following scale to interpret the values of these global indicators: [0, 0.20[– *Very weak*; [0,20; 0.40[– *Weak*; [0.40, 0.60[– *Reasonable*; [0.60, 0.80[– *Good*; and [0.80,1] – *Very good*. In a visual representation (see Figure 1) we can use a grading scale of patterns to assist the interpretation of the global indicator values.

3. Results

The horse data set (http://www.ceremade.dauphine.fr/~touati) is composed of twelve symbolic data units (horses) described by 10 symbolic variables, of which seven are quantitative of the interval type. The coding on the data units is as follows: 1-*ES/R*, 2-*MA/R*, 3-*EN/R*, 4-*AM/R*, 5-*EN/L*, 6-*AM/L*, 7-*ES/L*, 8-*EN/P*, 9-ES/P, 10-AM/P, 11-*ES/D* and 12-*EN/D*. According to Carvalho and Souza (2009), the designations *ES*, *EN*, *AM* and *MA* refer respectively to *Southern Europe* (ES), *Northern Europe* (*EN*), *America* (*AM*) and *Arab World* (*MA*), while the designations *R*, *L*, *P* and *D* refer respectively to *Racehorse* (*R*), *Leisure Horse* (*L*), *Pony* (*P*) and *Draft Horse* (*D*).

Table 1. Values of the indices E(h), SIL(h) and Global_Ind(h) in the set of partitions

Unit of data (h)	*E (h)*	*SIL (h)*	*Global_Ind (h)*
1- ES/R	0.845	0.567	0.706
2-MA/R	0.722	0.838	0.780
3-EN/R	0.845	0.556	0.701
4 - AM/R	0.786	0.166	0.476
5 - EN/L	0.729	0.582	0.656
6 - AM/L	0.729	0.432	0.581
7 - ES/L	0.612	0.591	0.602
8 - EN/P	0.67	0.266	0.468
9 - ES/P	0.824	0.209	0.517
10 - AM/P	0.832	0.286	0.559
11 - ES/D	0.845	0.679	0.762
12 - EN/D	0.845	0.574	0.710

The *HCA* of the 12 units of symbolic data (symbolic objects) was based on the weighted generalized affinity coefficient, with equal weights ($\pi_j=1/p$), centred and reduced by the WW method from Wald and Wolfowitz (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al. 2009, 2010). This coefficient was combined with 26 aggregation criteria, 12 of which were classical (SL, CL,…, AMGT) and 14 probabilistic (AVM, AVmg,…, AVMLD) (Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).

Table 2. Indicators of quality for the partitions P_i, i=1,…,26 and global index

Methods	$E(P_i)$	$SIL(P_i)$	$CF(P_i)$	$STAT\_nor(P_i)$	Global_Ind(Pi)
SL	0.725	0.607	0.897	0.9	0.782
CL	0.806	0.464	0.827	0.994	0.773
AM	0.799	0.464	0.342	0.994	0.650
AMg	0.799	0.464	0.621	0.994	0.720
A Cen	0.772	0.371	0.351	0.981	0.619
A med	0.775	0.397	0.456	0.99	0.655
AMG	0.804	0.371	0.613	0.981	0.692
AMT	0.799	0.464	0.623	0.994	0.720
AMgT	0.799	0.464	0.622	0.994	0.720
A Cen T	0.804	0.371	0.625	0.981	0.695
A Med T	0.797	0.397	0.694	0.99	0.720
AMGT	0.772	0.371	0.746	0.981	0.717
AVM	0.797	0.371	0.781	0.981	0.732
AVmg	0.804	0.371	0.862	0.981	0.754
AV Cen	0.819	0.606	0.859	0.781	0.766
AV med	0.65	0.590	0.966	0.173	0.595
AVMG	0.654	0.419	0.946	0.981	0.750
AVL	0.777	0.388	0.577	0.99	0.683
AVB	0.777	0.388	0.728	1	0.723
AV1	0.819	0.408	0.697	0.85	0.694
AV2	0.777	0.388	0.76	0.99	0.729
AV4	0.777	0.388	0.693	0.99	0.712
AV5	0.819	0.606	0.637	0.85	0.728
AV6	0.746	0.388	0.833	0.998	0.741
AVD	0.705	0.485	0.877	0.738	0.701
AVMLD	0.766	0.408	0.675	0.781	0.658

The indicators *E(h)*, *SIL(h)* and *Global_Ind(h)* presented in Table 1 give detailed information on each of the data units *h*, taking into consideration the cluster to which they belong, how many data units are contained in the cluster with which the data unit *h* is associated, and its degree of homogeneity and isolation.

It is also important that the data should be presented on the same scale, so that they are comparable, in an isolated manner or together. In order to provide an overview of the behaviour of each of the applied methods and each data unit, we used a set of indexes whose values are presented in Tables 1 and 2.

Figure 2 synthesises visually the quality of the obtained results at the level of each element, of each cluster, in each of the 26 methods. The supremacy of cluster 1 was noted, while cluster 3 had the lowest score. The highest score was that assigned to elements 11 and 3, followed by scores for 1 and 12, while the lowest score was that for element 4, followed by those for 8 and 9.

4. Conclusion

The global approach assembles information about the homogeneity, isolation and stability of the elements and of the clusters in the partitions under comparison. At the level of each of the partitions and of the set of partitions this evaluation can be performed in a more detailed way, using the most relevant information available, such as that related to stability, isolation, homogeneity, fusion coefficient or global statistics of levels.

The visualisation approach allows us to perceive, in a quick and detailed way, the resemblances/differences between the partitions and the behaviour of the elements in the partitions under analysis. Thus the comparison of partitions using the developed methodology contributes to more comprehensive knowledge, and has the advantage that all indexes used take values in the interval [0,1]. The comparison of the various obtained qualifying structures, as is done in the methodology, aims to enable a more detailed examination of the results of the Cluster Analysis of a given data set.

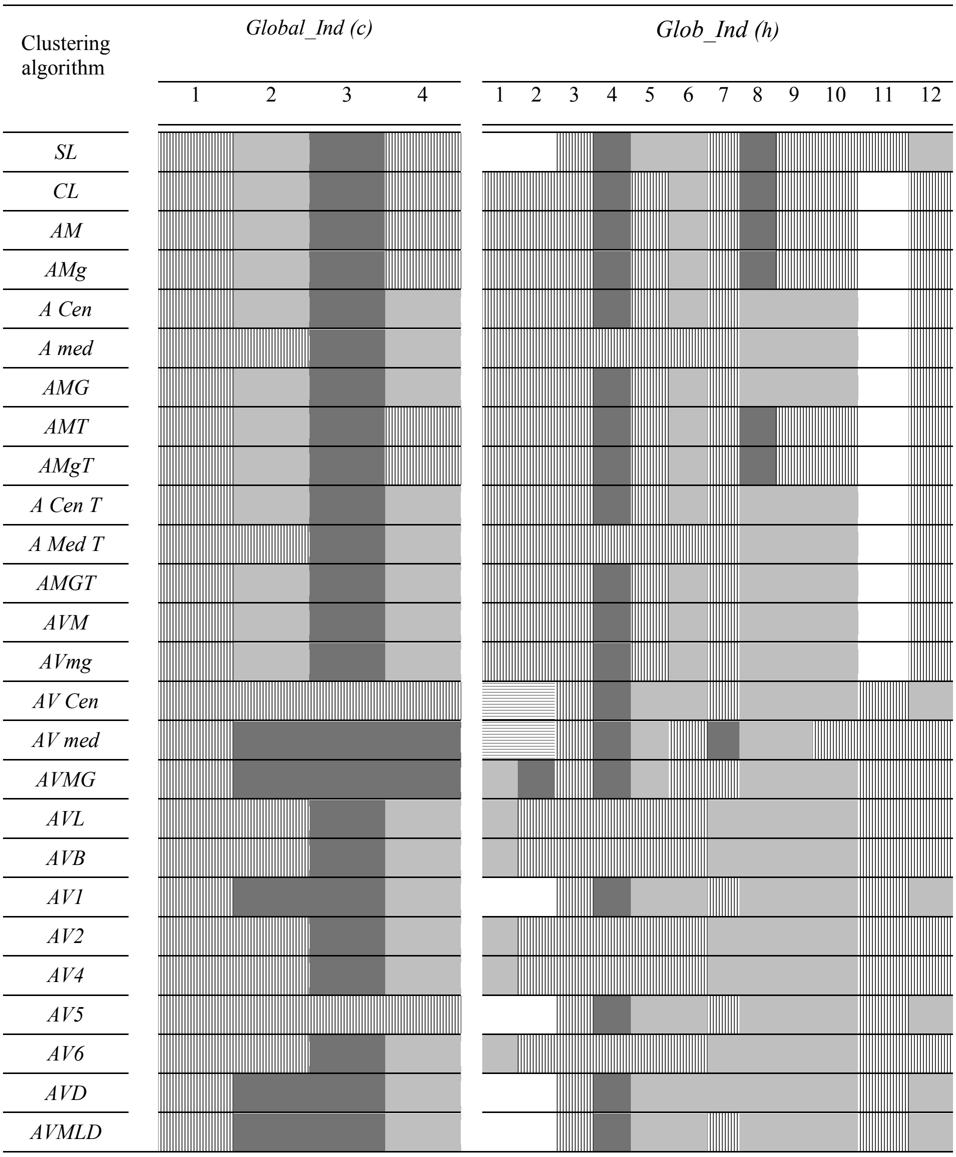

Figure 2. Information from the visualization – quality of the results

REFERENCES

Bacelar-Nicolau H. (1980): Contributions to the Study of Comparison Coefficients in Cluster Analysis, PhD Th. (in Portuguese), Univ. Lisbon.

Bacelar-Nicolau H. (1988): Two Probabilistic Models for Classification of Variables in Frequency Tables. In: Classification and Related Methods of Data Analysis, H.-H. Bock (ed.), North Holland: Elsevier Sciences Publishers B.V.: 181-186.

Bacelar-Nicolau H. (2000): The Affinity Coefficient. In: Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data, H.H. Bock, E. Diday (Eds.), Springer: 160-165.

Bacelar-Nicolau H., Nicolau F.C., Sousa A., Bacelar-Nicolau L. (2009): Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets. Biocybernetics and Biomedical Engineering 29(2): 9-18.

Bacelar-Nicolau H., Nicolau F.C., Sousa A., Bacelar-Nicolau L. (2010): Clustering Complex Heterogeneous Data Using a Probabilistic Approach. Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010), Chania Crete Greece, 8-11 June 2010 – published on the CD Proceedings of SMTDA2010 (electronic publication).

Carvalho F., Souza R. (2009): Unsupervised Pattern Recognition Models for Mixed Feature-Type Symbolic Data. Pattern Recognition Letters 31(5): 430-443.

Gordon A.D. (1999): Classification, 2nd. Chapman &Hall, London.

Lerman I.C. (1981): Classification et Analyse Ordinale des Données. Dunod, Paris, 1981.

Nicolau F.C. (1983): Cluster Analysis and Distribution Function. Meth. Oper. Res. 45: 431-433.

Nicolau F.C., Bacelar-Nicolau H. (1998): Some Trends in the Classification of Variables. In: Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, Y. Baba (Eds.), Springer-Verlag: 89-98.

Silva O., Bacelar-Nicolau H., Nicolau F.C. (2010): Global Approach for Evaluating the Quality of Clustering Results. In: Programme and Abstracts CFE 10 & ERCIM 10 (4th CSDA Intern. Conference on Computational and Financial Econometrics and 3rd Conference of the ERCIM Working Group on Computing and Statistics): 40.

Silva O. (2011): Contributions for Comparing and Evaluating Partitions in Hierarchical Cluster Analysis. PhD. Th. (in Portuguese), Azores University.