A Validation Methodology in Hierarchical Clustering

Fernanda Sousa and Jorge Tendeiro

Faculdade de Engenharia/CEC, Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, PORTUGAL
(e-mail: fcsousa@fe.up.pt, jorgetendeiro@net.sapo.pt)

Abstract. This paper presents a validation methodology in ascending hierarchical clustering. The objects in validation are clustering hierarchies, and simulation is used. Under certain conditions, this methodology allows us to evaluate the quality of hierarchical structures, its robustness and fiability, according to the data structure. The effect of the application of a given criterion on some kind of structures is also analyzed.

Keywords: Cluster Analysis, Hierarchical Clustering, Robustness, Validation.

1 Introduction

The use of clustering methods has progressively increased. On one hand, there are several computer programs which include these methodologies; on the other, there are big data sets which need to be studied (and summarised). Since nowadays it is quite easy and not very expensive to have big databases, it is essential to find tools in order to extract relevant information.

Generally speaking, the main goal of clustering is to define partitions or hierarchies of partitions, over a set of two-by-two comparable elements, that respect the resemblence between them in a predefined optimal manner. The elements to classify may be objects or variables of a data set.

This work belongs to the ascending hierarchical clustering (A.H.C.) field, whose usual output is a succession of partitions whose classes are partially ordered by inclusion. This methodology begins with the most refined partition (with singleton classes); in each stage the most resembled classes are gathered together according to a predefined criterion. The most common graphical result drawn out is named a classification tree or dendrogram. Choosing an element of the succession of partitions we get a division of the elements in clusters, as well as the history of the formation of each class.

Although clustering is a powerful tool in analysing data, we need to assure that the division into several clusters suggested by the algorithm does not distort the structure of the initial data. In other words, the relations between the elements to classify cannot lead to artificial clusters without real meaning. The need and importance of a next stage for the attainment of results in a clustering method is unquestionable. This stage, here named as validation, consists of questioning over the (preliminary) results obtained, in order to the achieve final results or conclusions.

Section two of this paper is dedicated to clustering validation. In section three we present a methodology based on Monte Carlo simulation which allows us to evaluate the quality and robustness of a hierarchic structure, and to give us information on the quality of adjustment to data structure. This methodology also allows us to analyse the effect that the use of a given clustering criterion can have when applied to a specific kind of structure. Section four presents an application of the proposed methodology. Finally, in section five we lay out conclusions, as well as some perspectives of developments.

2 Validation in clustering

The application of a clustering algorithm to a data set leads to a partition or a hierarchy of partitions over the set of elements in classification. After an accurate interpretation, this result will give us information on the relations between the elements in classification.

A clustering method requires choices, and it is well known that these choices affect the result of the process of clustering. In other words, different choices may lead to different classifications. This fact creates a new problem: whether to decide which choice is to result into the best clustering. We admit that each method has its own underlying structure model, which can be optimized in each situation. Also, a clustering method always produces a partition or a hierarchy of partitions, inducing a structure on the data. It seems reasonable to question the existence of structure on the initial data and, if that is the case, if there's a close relation between the initial and final structures. These and other questions really do justify the existence of a stage of validation in clustering results before its interpretation.

Several authors have studied different approaches to clustering validation. We can mention (among others) Bock's investigations [Bock, 1985], [Bock, 1996], which insert clustering models into a probabilistic context, assuming that the observed data is a sample of a structured multivariate population. [Gordon, 1994], [Gordon, 1996] and [Milligan, 1996] (and other references indicated in these works) appeal to empirical, descriptive or exploratory tools in analysing the quality of the clusters obtained.

It is usual to apply several clustering methods to a data set with the goal of choosing one of them or a new one determined from the data set. This issue includes the comparison of clustering trees or dendrograms ([Lapointe and Legendre, 1990], [Lapointe and Legendre, 1995]) and the consensus theory [Barthélemy *et al.*, 1986] and other references in [Gordon, 1999]. Another kind of issue is trying to understand the quality and stability of the results obtained from a clustering process. Here validation may also be considered under different perspectives: we may wish to validate a single cluster, a partition, or a hierarchy. The validation of a single cluster was studied (among others) by [Gordon, 1994] and [Bel Mufti, 1998]. For references about research on validation of partitions see [Hubert, 1987] and [Gordon, 1999]. Due to its complexity, hierarchies validation research has less references. Note that validation of hierarchies often appears connected to validation of partitions, since hierarchies are successions of partitions.

3 Validation methodology in A.H.C.

An A.H.C. algorithm has underlying two important choices: an index, comparison function between pairs of elements of the set to classify, and the comparison function between clusters associated to the aggregation criterion. It is assumed that the result of an A.H.C. depends on the data set as well as on these choices. Each method tends to adjust its structure model in each situation; effectiveness depends on the type and structure intensity of the data. When analysing a structure obtained by a clustering method, it is important to evaluate which part of it is due to the criterion used.

In this section we present a validation proceeding in A.H.C., whose main purpose is to help to understand some of the following questions:

- Does the data really have a clustering structure? In the affirmative case, does the hierarchy obtained reveal that structure?
- How can we choose the level of a hierarchy which gives the best partition?
- After applying several aggregation criteria or indexes to a data set, how can we decide which one is the best? Does it even exist?

The methodology here presented allows us to provide information about the problems in comparing indexes, hierarchies, indexes and hierarchies, and the effects that random perturbations have on them. This procedure intents to evaluate the quality of the final result using effectiveness and stability of a given A.H.C. method, supporting the results interpretation and helping the choices to be made. The methodology developed uses two main tools: the comparison of clustering structures and random generation of dendrograms.

3.1 Comparison of clustering structures

Let E be a set of m elements to classify, and F the set of the subsets of E with two distinct elements ($\text{card}(F) = \binom{m}{2} := M$):

$$F = \big\{\{x, y\} : x, y \in E, x \neq y\big\}.$$

Consider $\gamma : E \times E \longrightarrow \mathbb{R}_0^+$, the comparison function between pairs of elements of E, and the function $h : E \times E \longrightarrow \mathbb{R}_0^+$ which associates to each element $(x, y) \in E \times E$ the index of aggregation of the smaller cluster that contains simultaneously x and y. A function γ can be associated to a vector of dimension M that contains information about the structure of the data, and a hierarchy H can also be associated to a vector of dimension M that contains information about the clustering structure. Our goal is to compare those kind of structures. In this work we adopted an ordinal approach to do this comparison, associating preordenations to the various structures. We can define a (total) preordenation over E defining a (total) preorder over F.

The choice of a comparison function of the elements of E, γ, defines a total preorder over F; in fact, if γ is a dissimilarity we just consider

$$\forall\,(\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} \gamma(x,y) \leqslant \gamma(z,t). \tag{1}$$

This total preorder over F is the total preordenation over E associated to γ. If γ is injective (which happens often in practice), this preorder is, in fact, an order.

A hierarchy H over the elements of E always defines a total preordenation over F; in fact, we have the following relation:

$$\forall\,(\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} h(x,y) \leqslant h(z,t). \tag{2}$$

This total preorder over F is the total preordenation over E associated to H.

Given a partition π of E consisting of k classes $E_1, E_2, \ldots, E_k$, we can define a partition ξ of F in two classes:

- $R(\pi) = \big\{\{x,y\} \in F : x,y \in E_i \text{ for some } i = 1,2,\ldots,k\big\}$;
- $S(\pi) = \big\{\{x,y\} \in F : x \in E_i, y \in E_j, i \neq j\big\}$.

It is easy to verify that [Lerman, 1981] ξ defines a (non total) preordenation over E. Alternatively, we can specify a total preorder over F associated to ξ as follows:

$$\forall\,(\{x,y\},\{z,t\}) \in F \times F : \{x,y\} \leqslant \{z,t\} \underset{\text{def.}}{\Longleftrightarrow} \begin{cases} \{x,y\},\{z,t\} \in R(\pi) \\ \underline{\text{or}} \\ \{x,y\},\{z,t\} \in S(\pi) \\ \underline{\text{or}} \\ (\{x,y\},\{z,t\}) \in R(\pi) \times S(\pi) \end{cases} \tag{3}$$

So, due to the relations (1), (2) and (3) we conclude that instead of comparing comparison functions, partitions or hierarchies of partitions we can compare the corresponding preordenations (with the same length).

There are several coefficients which allow us to compare two preordenations. The results included in this paper were obtained using the Goodman-Kruskal coefficient:

$$T_{GK} = \frac{C-D}{C+D}, \tag{4}$$

where C and D are, respectively, the number of positive and negative agreements between both preordenations. All the methodology that is going to be described can easily be applied to another coefficient.

Assintotic results on the distributions of these coefficients are not adequate in this context. The main problem is that the deduction of such distributions is based on independence between preordenations, which cannot be verified here in practice. In fact, preordenations that result from relations (1), (2) and (3) may not be independent if, for example, they come out of the application of different clustering processes over a common data set. In these situations, independence is many times what the researcher does not want, because the goal is to prove that there is information shared by the outcoming of several results. Moreover, preordenations related to clustering processes have restrictions imposed by the ultrametric property: property that verify ultrametric matrices associated to clustering structures. By this we mean that not all preordenations can be the outcome of a clustering process.

For the described reasons, it becomes necessary to deduce proper distributions for the comparison coefficients. It is not feasible to deduce the exact distribution, because the number of distinct dendrograms of order m increases very rapidly ($d(m) = \frac{m!(m-1)!}{2^{m-1}}$).

Simulation is the alternative solution, since assintotic distributions do not fit our purposes. To generate empirical distributions we need to be able to generate random clustering structures. At this stage, methods of random generation of dendrograms are extremely useful.

3.2 Random generation of dendrograms or ultrametric matrices

There are some algorithms that allow the random generation of dendrograms (or equivalent structures). Note that the point here is to generate random topologies, labels and aggregation levels; few methods attend at these three features simultaneously. We mention four methods: Double Permutation method [Lapointe and Legendre, 1990]; Uniform generation method [Sousa, 2000]; RA method [Podani, 2000]; Shape Parameter method [Sousa, 2000]. The first three methods are random *sensu* Furnas [Furnas, 1984], in other words, they can generate (for a given order) each possible dendrogram in an equiprobable manner (with probability $\frac{1}{d(m)}$). The Shape Parameter method introduces a coefficient (shape parameter) that, once settled, allows to predict (with some probability) the final shape of the generated dendrogram. This method is a very useful tool for validation in A.H.C., for it is well known that some clustering methods tend to generate particular kinds of trees.

Using one of the methods of random generation of dendrograms we can randomly generate a pair of dendrograms for a given order. By this way, it is simple to deduce empirical distributions for a chosen ordinal comparison coefficient of structures, allowing to give statistical significance to its values.

There is an alternative way in approaching the problem of random generation of clustering structures that is based on the notion of combinatorial structure ([Flajolet *et al.*, 1994] and [Van Cutsem, 1996]).

3.3 Algorithm

We now present a methodologic sequence using Monte Carlo simulation. Our goal is to supply a method that can help us answer some of the questions previously stated.

For a given topologic type of structure of data and for a fixed number of elements to classify, consider the following steps given:

1. Generate a random dendrogram; the associated ultrametric matrix, M_0, will be taken as the (initial) dissimilarity matrix.
2. For each A.H.C. criterion to study: obtain a hierarchy H_0, and compare M_0 with H_0 (comparison $\mathcal{C}^1$).
3. Disturb matrix M_0 by settling a disturbance coefficient; this creates the dissimilarity matrix M_i. Compare M_0 with M_i (comparison $\mathcal{C}^2$).
4. For each A.H.C. criterion to study: obtain a hierarchy H_i, compare M_i with H_i (comparison $\mathcal{C}^3$) and compare H_0 with H_i (comparison $\mathcal{C}^4$).
5. Repeat the steps 3. and 4. a great number of times for the same disturbance coefficient.
6. Repeat the steps 3. to 5. for different values of the disturbance coefficient.

The several comparisons, considered according to section 3.1, try to:

- $\mathcal{C}^1$: Analyse a criterion behaviour when applied to ultrametric data.
- $\mathcal{C}^2$: Control the impact of the disturbance over the associated preordenations.
- $\mathcal{C}^3$: Analyse the ability of a criterion to recover a structure after disturbance.
- $\mathcal{C}^4$: Evaluate if the hierarchical structure maintained, and try to understand what disturbance value is implied in the damaging of the structure.

4 An application

The presented methodology comprehends a diversity of choices to be made in each simulation. We now refer the several options we made in this specific application. The number of elements to classify equal 10. There were considered three types of data structures to generate: predominantly chain type trees, predominantly balanced trees (obtained with the Shape Parameter method), and also trees obtained with Uniform method. Note that both chain and balanced types are very important in classification, either for their association with well known classical methods as for their extreme characteristics. Concerning the A.H.C. methods, we tried to evaluate the performance of a set of methods belonging to classical and probabilistic approaches (the latter is known as VL approach– Validity of the Link due to I.C. Lerman) in which the aggregation criterion of clusters is based on a statistic of central tendency [Sousa, 2000]. The criteria here considered are: Single Linkage (SL), Complete Linkage (CL), Mean Linkage (HMEAN) and Median Linkage

(HMED) (classical approach), Validity of the Mean (AVM [Nicolau, 1980]), Validity of the Median (HVMED) and a method of the VL parametric family AVB proposed by [Bacelar-Nicolau, 1985] (VL approach). The disturbance was carried out adding to each element of M_0 a quantity $\delta(2x-1)$, where x comes from a uniform random variable over $]0,1[$ and δ is the disturbance factor. Values for δ were considered between 0.05 and 0.5. For the comparison of structures it was used the T_{GK} coefficient given by (4).

We now present some conclusions that illustrate how this methodology can give us information.

From $\mathcal{C}^1$ comparison we can say that classical methods recover completely the structure of an ultrametric matrix, while VL methods produce hierarchies that can be slightly different. When the dissimilarity matrix differs from the ultrametric structure ($\mathcal{C}^3$ comparison), the methods that give higher values for T_{GK} are HMEAN and HMED, followed by SL. The CL behaviour is similar to SL's for $\delta \leqslant 0.25$, but is different for greater values of δ. In general, HVMED is the VL criterion that works better, but when the data structure approaches chain type we see that AVM and HVMED are equally effective. For balanced structures, AVB seems to be the best method. $\mathcal{C}^4$ comparison allows us to conclude that the stability of the structures produced by some methods strongly depends on the type of data structure. Usually the most stable methods are HMEAN, SL and HMED, and the less stable is CL, followed by AVM. AVM is very stable when trees of chain type are considered, and for balanced trees AVB method has better T_{GK} values. The VL method less influenced by structure is HVMED.

The results obtained let us quantify some known characteristics related to the application of these criteria to real data. In fact, AVM and HVMED methods tend to produce trees of chain type, while AVB tends to produce trees with clusters of similar number of elements (balanced).

5 Conclusion and perspectives

The methodology here presented claims out to be a contribution for the A.H.C. validation subject, and it can be quite general. What was done for hierarchic clustering can easily be adjusted for partitions, too. During experiences, it was necessary to make some choices in specifing some parameters' values. This feature is considered very important. The number of possible combinations of choices is enormous, and the timing of simulation and analyse of results increases dramatically. However, a few new wise options should be tried out, particularly the application to real data.

The validity methodology here presented allows us to say that the behaviour of a clustering method strongly depends on the kind and intensity of the data structure.

The methods of central tendency of classical approach seem to have some common properties that lead to good results. The VL methods, on account of its own approach, can lead to a good performance in particular cases.

References

[Bacelar-Nicolau, 1985]H. Bacelar-Nicolau. The affinity coefficient in cluster analysis. *Methods of Operation Research*, 53:pages 507–512, 1985.

[Barthélemy *et al.*, 1986]J.-P. Barthélemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus classifications. *Journal of Classification*, 3:pages 187–224, 1986.

[Bel Mufti, 1998]G. Bel Mufti. *Validation d'une Classe par Estimation de sa Stabilité, Ph.D. Thesis.* Université Paris IX– Dauphine, Paris, 1998.

[Bock, 1985]H.H. Bock. On some signficance tests in cluster analysis. *Journal of Classification*, 2:pages 77–108, 1985.

[Bock, 1996]H.H. Bock. Probability models and hypotheses testing in partitioning cluster analysis. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 377–453, 1996.

[Flajolet *et al.*, 1994]P. Flajolet, P. Zimmerman, and B. Van Cutsem. A calculus for the random generation of labeled combinatorial structures. *Theoretical Computer Science*, 132:pages 1–35, 1994.

[Furnas, 1984]G.W. Furnas. The generation of random, binary unordered trees. *Journal of Classification*, 1:pages 187–233, 1984.

[Gordon, 1994]A.D. Gordon. Identifying genuine clusters in a classification. *Computational Statistics & Data Analysis*, 18:pages 561–581, 1994.

[Gordon, 1996]A.D. Gordon. Hierarchical classification. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 65–121, 1996.

[Gordon, 1999]A.D. Gordon. *Classification.* Chapman & Hall, London, 2nd edition, 1999.

[Hubert, 1987]L.J. Hubert. *Assignment Methods in Combinatorial Data Analysis.* Marcel Dekker, New York, 1987.

[Lapointe and Legendre, 1990]F.J. Lapointe and P. Legendre. A statistical framework to test the consensus of two nested classifications. *Systematic Zoology*, 39:pages 1–13, 1990.

[Lapointe and Legendre, 1995]F.J. Lapointe and P. Legendre. Comparison tests for dendrograms: A comparative evaluation. *Journal of Classification*, 12:pages 265–282, 1995.

[Lerman, 1981]I.C. Lerman. *Classification et Analyse Ordinale des Données.* Dunod, Paris, 1981.

[Milligan, 1996]G.W. Milligan. Clustering validation: Results and implications for applied analyses. P. Arabie and L.J. Hubert and G. de Soete editors, *Clustering and Classification*, pages 341–375, 1996.

[Nicolau, 1980]F.C. Nicolau. *Critérios de Análise Classificatória Hierárquica Baseados na Função Distribuição, Ph.D. Thesis.* Faculdade de Ciências da Universidade de Lisboa, Lisboa, 1980.

[Podani, 2000]J. Podani. Simulation of random dendrograms and comparison tests: Some comments. *Journal of Classification*, 17:pages 123–142, 2000.

[Sousa, 2000]F. Sousa. *Novas Metodologias e Validação em Classificação Hierárquica Ascendente, Ph.D. Thesis.* Universidade Nova de Lisboa, Lisboa, 2000.

[Van Cutsem, 1996]B. Van Cutsem. Combinatorial structures and structures for classification. *Computational Statistics & Data Analysis*, 23:pages 169–188, 1996.