Clustering Validation in the Context of Hierarchical Cluster Analysis: an Empirical Study

Osvaldo Silva, Áurea Sousa, and Helena Bacelar-Nicolau

Abstract The evaluation of clustering structures is a crucial step in cluster analysis. This study presents the main results of the hierarchical cluster analysis of variables concerning a real dataset in the context of Higher Education. The goal of this research is to find a typology of some relevant items taking into account both the homogeneity and the isolation of the clusters.Two similarity measures, namely the standard affinity coefficient and Spearman's correlation coefficient, were used, and combined with three probabilistic (*AVL*, *AVB* and *AV1*) aggregation criteria, from a parametric family in the scope of the *VL* (Validity Link) methodology. The best partitions were selected based on some validation indices, namely the global *STAT* levels statistics and the measures P(I2, Σ) and γ, adapted to the case of similarity coefficients. In order to evaluate the clusters and identify their most representative elements, the Mann and Whitney *U* statistics and the silhouette plot were also used.

Keywords: clustering validation, affinity coefficient, Spearman correlation coefficient, *VL* methodology

Osvaldo Silva (✉)
Universidade dos Açores and CICSNOVA.UAc, Rua da Mãe de Deus, 9500-321, Portugal, e-mail: osvaldo.dl.silva@uac.pt

Áurea Sousa
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, Portugal,
e-mail: aurea.st.sousa@uac.pt

Helena Bacelar-Nicolau
Universidade de Lisboa (UL) Faculdade de Psicologia and Institute of Environmental Health (ISAMB/FM-UL), Portugal, e-mail: hbacelar@psicologia.ulisboa.pt

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_37

1 Introduction

Cluster analysis or unsupervised classification usually concerns exploratory multivariate data analysis methods and techniques for grouping either a set of data units or an associated set of descriptive variables in such a way that elements in the same group (cluster) are more similar to each other than elements in different clusters [6]. Therefore, it is important to validate the results obtained, bearing in mind that, in an ideal situation, the clusters should be internally homogeneous and externally well separated or isolated. Thus, according to Silva et al. ([15], p. 136), there are some important questions, such as: "i) How to compare partitions obtained using different cluster algorithms? ii) Is it possible to join information from several approaches in the decision-making process of choosing the most representative partition?"

This paper presents the main results of a hierarchical cluster analysis of variables concerning a real dataset in the field of Higher Education, in order to find a typology taking into account relevant validation measures. Two similarity measures (standard affinity coefficient and Spearman's correlation coefficient) were used, and combined with a parametric family aggregation criteria in the scope of the *VL* methodology (e.g., [10, 11, 17]).

With regard to the validation of clustering structures, some validation indices were used for the evaluation of partitions and the clusters that integrate them, which are referred to in Section 2. The main results are presented and discussed in Section 3. Section 4 contains some final remarks.

2 Data and Methods

Data were obtained from a questionnaire administered to three hundred and fifty students who were attending Higher Education in a public university, after their informed consent. The questionnaire contains, among others, eleven questions related to academic life and the respective courses.

Several algorithms of hierarchical cluster analysis of variables were applied on the data matrix. The variables (items) are: T1-Participation, T2-Interest, T3-Expectations, T4-Accomplishment, T5-Job Outlook, T6- Teachers' Professional Competence, T7-Distribution of Curricular Units, T8- Number of weekly hours of lessons, T9-Number of hours of daily study, T10-School Outcomes and T11-Assessment Methods, which were evaluated based on a Likert scale from 1 to 5 (1-Totally disagree, 2- Partially disagree, 3- Neither disagree nor agree, 4- Partially agree, 5- Totally agree).

The Ascendant Hierarchical Cluster Analysis (AHCA) was based on the standard affinity coefficient [1, 17] and Spearman's correlation coefficient. In this paper both measures of comparison were combined with three probabilistic aggregation criteria (*AVL*, *AVB* and *AV1*), issued from the *VL* parametric family. This methodology, in the scope of Cluster Analysis, uses probabilistic comparison functions, between pairs of elements, which correspond to random variables following a unit uniform distribu-

tion. Besides, this approach considers probabilistic aggregation criteria, which can be interpreted as distribution functions of statistics of independent random variables, that are i.i.d. uniform on [0, 1] (e.g., [17]).

Let A and B be two clusters with cardinals, respectively, α and β, and let γ_{xy} be a similarity measure between pairs of elements, $x, y \in E$ (set of elements to classify). Concerning the family I of *AVL* methods (e.g., *SL*, *AV1*, *AVB*, and *AVL*), the comparison functions between clusters can be summarized by the following conjoined formula:

$$\Gamma(A, B) = (p_{AB})^{g(\alpha,\beta)} \tag{1}$$

where $\alpha = Card\ A$, $\beta = Card\ B$, $p_{AB} = max[\gamma_{ab} : (a, b) \in (A \times B]$, with $1 \leq g(\alpha, \beta) \leq \alpha\beta$, and γ_{xy}, establishing a bridge between *SL* and *AVL* methods which have a braking effect on the formation of chains. For example, $g(\alpha, \beta) = 1$ for *SL*, $g(\alpha, \beta)=(\alpha + \beta)/2$ for *AV1*, $g(\alpha, \beta)=\sqrt{\alpha\beta}$ for *AVB*, and $g(\alpha, \beta) = \alpha\beta$ for *AVL* (see [3, 17]).

The application of the two measures of comparison between elements (Spearman correlation coefficient and standard affinity coefficient), combined with the aforementioned aggregation criteria, aims to find a typology of items corresponding to the best partition among the best partitions obtained by the several algorithms, in order to verify if there are any substantial changes in the results. Therefore, some validation indices based on the values of the corresponding proximity matrices were used, namely the global levels statistics (*STAT*) [1, 10, 11] and the indices P(I2mod, Σ) and γ [8], adapted to this type of matrices [16], so that the choice of the best partition is judicious and based on the desirable properties (e.g., isolation and homogeneity of the clusters). Concerning the best partitions, the respective clusters and the identification of their most representative elements were based on appropriate adaptations of the Mann and Whitney *U* statistics [8] and of the silhouette plots [14] to the case of similarity measures.

Each level of a dendrogram corresponds to a stage in the constitution of the partitions hierarchy. Therefore, the study of the most relevant partition(s) is strictly related to the choice of the best cut-off levels (e.g., [6, 5])

According to Bacelar Nicolau [1, 2], the global levels statistics (*STAT*) values must be calculated for each of the $k = 1, nivmax$ levels of the corresponding dendrograms, designating them by $STAT(k)$. At each level k, $STAT(k)$ is the global statistics that measures the total information given by the pre-order associated to the corresponding partition, in relation to the initial pre-order associated with the similarity or dissimilarity measure. A "significant" level is considered to be one that corresponds to a partition for which the global statistics undergoes a significant increase in relation to the information provided by neighbouring levels, that is, a local maximum of the differences $DIF(k) = STAT(k) - STAT(k - 1)$, $k = 1, nivmax$.

2.1 Adaptation of the P (I2, Σ)

To evaluate the partitions, an appropriate adaptation of the index P (I2, Σ) [8] for the case of similarity measures was used, given by the following formula:

$$P(I2mod, \Sigma) = \frac{1}{c} \sum_{r=1}^{c} \frac{\sum_{i \in C_r} \sum_{j \notin C_r} s_{ij}}{n_r \times (N - n_r)} \tag{2}$$

where c is the number of clusters of the partition and s_{ij} is the value of the similarity measure between the element i belonging to cluster C_r and the element j belonging to another cluster. This index takes into account the number of clusters and the number of elements in each of the clusters and evaluates the isolation of clusters belonging to a given partition.

2.2 Goodman and Kruskal Index (γ)

The γ index, proposed by Goodman and Kruskal [7], has been widely used in cluster validation [9]. Comparisons are developed between all within-cluster similarities, s_{ij} and all between-cluster similarities s_{kl} [18]. A comparison is judged concordant (respectively discordant) if s_{ij} is strictly greater (respectively, smaller) than s_{kl}. The γ index is defined by:

$$\gamma = (S_+ - S_-)/(S_+ + S_-), \tag{3}$$

where S_+ (or S_-) is the number of concordant (respectively, discordant) comparisons. This index is a global stopping rule and it evaluates the fit of the partition in c clusters based on the homogeneity (high similarity between the elements within the clusters) and the isolation (low similarity of the elements between the clusters) of the clusters. Note that the higher the value of this index, the better is the adjustment of that partition.

The use of *STAT*, γ and P(I2mod, Σ) indices can help identifying the most significant levels of a dendrogram, taking into account both the homogeneity and the isolation of the clusters [15].

2.3 U Statistics (Mann and Whitney)

U statistics [12] are relevant for assessing the suitability of a cluster, combining the concepts of compactness and isolation. Thus, the "best" cluster is the one with the lowest values of global U-index, U_G, and local U-index, U_L [8]. In the present paper we used an appropriate adaptation of these indices to the case of similarity measures (for details, see [19]). Moreover, the clusters considered "ideal" are those for which U_G and U_L both take the value zero. Mann and Whitney's U statistics are useful in

decision making, in situations of uncertainty, both for the evaluation of the clusters and partitions.

2.4 Silhouette Plots

We also used an appropriate adaptation of the silhouette plots [14], which allows the assessment of compactness and relative isolation of clusters. The adaptation of this measure for the case of similarity measures, $Sil(i)$, considers the average of the similarities between an element i belonging to cluster C_r , which contains $n_r (\geq 2)$ elements, and all other elements that do not belong to this cluster (see [19]). The values of this measure $\{Sil(i) : i \in C_r\}$ lie between -1 and $+1$, with "values near $+1$ indicating that element strongly belongs to the cluster in which it has been placed" ([8], p. 205). In the case of a singleton cluster, $Sil(i)$ assumes the value zero [8] in the corresponding algorithm.

3 Results and Discussion

The best partitions provided by the dendrograms are shown in Table 1.

Table 1 The best partitions concerning the dendrograms.

Coefficient	Method	The best partition	Validation indices
Affinity	*AVL*	(T1, T3, T4, T5 ,T6, T7, T8, T10, T11), (T2, T9)	STAT=5.1301 γ= 0.8589 P(I2mod,Σ)=0.2077
	AV1/AVB	(T1, T3, T4 , T5, T6, T7, T8, T10, T11), (T2), (T9)	STAT=5.3453 γ= 0.8830 P(I2mod,Σ)=0.2049
Spearman	*AVL*	(T3, T4 ,T2 , T9) (T7, T11, T8), (T6, T10), (T1), (T5)	STAT=4.0152 γ= 0.8178 P(I2mod,Σ)=0.3896
	AV1/AVB	(T3, T4 ,T2 , T9, T6) (T7, T11, T8), (T1, T10), (T5)	STAT=4.05751 γ= 0.7317 P(I2mod,Σ)=0.38177

Figure 1 shows the dendrograms obtained, respectively, by the standard affinity coefficient (left side) and Spearman's correlation coefficient (right side), both combined with the *AVL* method.

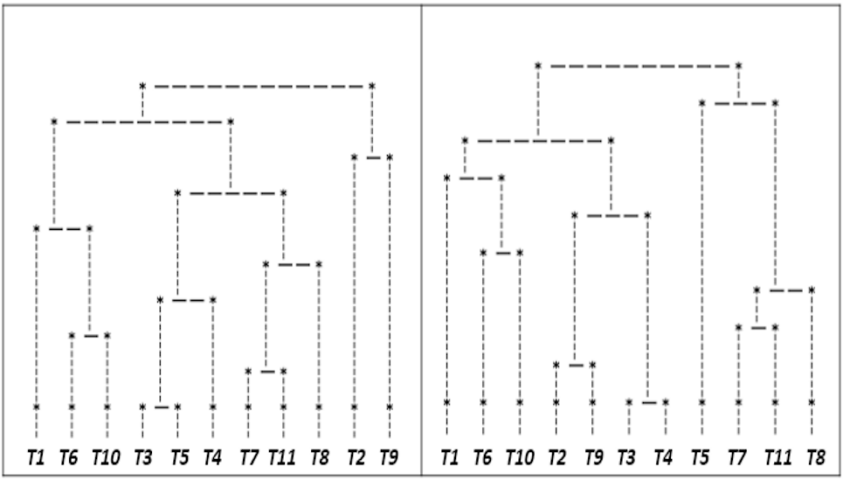

Fig. 1 Dendrograms based on standard affinity coefficient (left side) and Spearman's correlation coefficient (right side) - *AVL*.

The "best" partition obtained using the affinity coefficient and the *AVL* method is the partition into two clusters (level 9 of the aggregation process). The first cluster consists of nine items that highlight the importance of the teachers' professional competence, the structuring/content of the course and the future perspectives in relation to the career opportunities, mostly factors exogenous to the students. The second one is composed by two items (T2 and T9) which emphasize the role of interest in the study of Mathematics.

The algorithms in which the standard affinity coefficient was used are the ones that provided the best partitions and their hierarchies are the ones that remained closest to the initial pre-orders. In fact, in the case of Spearman correlation coefficient the values of *STAT* and γ indices are clearly lower than the previous ones. Moreover, the cluster {T1, T3, T4, T5, T6, T7, T8, T10, T11}, corresponding to the best partition provided by the combination of the standard affinity coefficient with the aggregation criteria *AVL*, *AV1* and *AVB*, presents (U_G =39 and U_L=4, both lower than those obtained for the cluster {T3, T4, T2, T9, T6} (U_G=65 and U_L=26) provided by the Spearman correlation coefficient combined, respectively, with *AV1* and *AVB* methods.

Focusing the attention on the two first partitions of Table 1, the only difference between them is that while the best partition provided by *AV1* and *AVB* methods contains the singletons T2 and T9, the best partition given by *AVL* joins these two singletons in the same cluster. The values of the numerical validation indices shown in Table 1 indicate that the best partition is the one provided by *AV1* and *AVB* methods. This conclusion is reinforced by the observation of the silhouette plot (see Figure 2), which indicates that the cluster joining T2 and T9, given by *AVL* method, includes the elements which have the two lowest values of *Sil* and *Sil* (T2) is negative

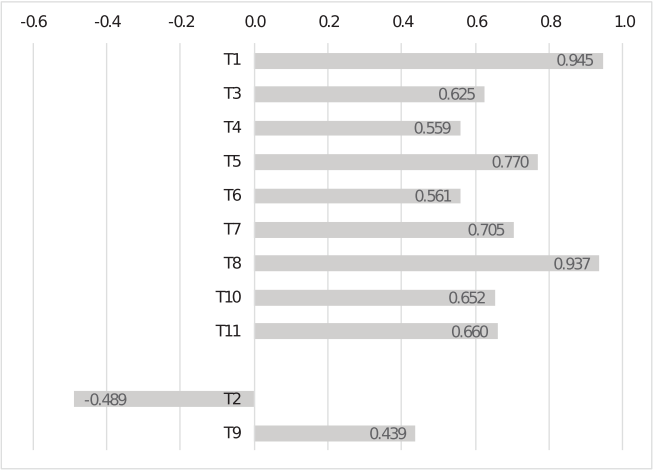

Fig. 2 Silhouette plot - standard affinity coefficient and *AVL* method.

(i.e., T2 does not fit very well in this cluster). Note that the silhouette plot cannot be used for the best partition, since it does not apply for singletons.

4 Final Remarks

This research was useful concerning the identification of relevant partitions of items in the context of Higher Education. In the cases where the affinity and the Spearman correlation coefficients were used, it was concluded that the probabilistic criteria *AV1* and *AVB* showed a higher agreement regarding the hierarchies of partitions obtained than the *AVL* method.

The validation measures *STAT*, γ and P(I2mod, Σ) help us to determine the best cut-off levels of a hierarchy of clusters, taking into account both the homogeneity and the isolation of the clusters. It should also be noted that if there is no absolute consensus between these three measures, the Mann and Whitney *U* statistics and the silhouette plot prove to be very useful, as we have seen with the application of this methodology to evaluate both the clusters and the partitions obtained.

Acknowledgements Funding. This work is financed by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the project «UIDB/04647/2020» of CICS.NOVA – Centro de Ciências Sociais da Universidade Nova de Lisboa.

References

1. Bacelar-Nicolau, H.: Analyse d'un Algorithme de Classification Automatique. Thèse de 3ème Cycle. ISUP, Paris VI (1972)
2. Bacelar-Nicolau, H.: Contributions to the Study of Comparison Coefficients in Cluster Analysis (in Portuguese). Univ. Lisbon (1980)
3. Bacelar-Nicolau, H.: On the distribution equivalence in cluster analysis. In: P. A., Devijver, & J. Kittler (eds.) Pattern Recognition Theory and Applications, NATO ASI Series, Series F. Computer and Systems Sciences, vol. 30, pp. 73-79. Springer - Verlag, New York (1987)
4. Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á., Bacelar-Nicolau, L.: Clustering of variables with a three-way approach for health sciences.Testing, Psychometrics, Methodology in Applied Psychology (TPM) (2014) doi: 10.4473/TPM21.4.56
5. Benzécri, J. P.: Analyse Factorielle des Proximités. Publication de l'Institut de Statistique de l' Universite de Paris (ISUP), XIII et XIV (1965)
6. De La Vega, W.: Techniques de la classification automatique utilisant un índice de ressemblance. Revue Française de Sociologie. **VIII**, 506–520 (1967)
7. Goodman, L. A., Kruskal, W. H.: Measures of association for cross-classifications. Journal of the American Statistical Association. **49**, 732–764 (1954)
8. Gordon, A. D.: Classification, 2nd Ed. Chapman & Hall, London (1999)
9. Hubert, L. J.: Some applications of graph theory to clustering. Psychometrika **39**(3), 283–309 (1974)) doi: 10.1007/BF02291704
10. Lerman, I. C.: Classification et Analyse Ordinale des Données. Dunod, Paris (1981)
11. Lerman, I. C.: Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series: Advanced Information and Knowledge Processing. Springer-Verlag, Boston (2016)
12. Mann, H. B., Whitney, D. R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 50–60 (1947)
13. Nicolau, F. C., Bacelar-Nicolau, H.: Some trends in the classification of variables. In: Hayashi et al. (eds.) Data Science, Classification and Related Methods, pp. 89-98. Springer,Tokyo (1998)
14. Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computation and Applied Mathematics. **20**, 53–65 (1987)
15. Silva, O., Bacelar-Nicolau, H.; Nicolau, F.: A global approach to the comparison of clustering results. Biometrical Letters **49**(2), 135–147 (2013) doi: 10.2478/bile-2013-0010
16. Silva, O., Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á.: Probabilistic approach for comparing partitions. In: Manca, R., McClean, S., Skiadas, C. H.(eds.) New Trends in Stochastic Modeling and Data Analysis, pp. 113-122. ISAST (International Society for the Advancement of Science and Technology), Athens (2015)
17. Sousa, Á., Silva, O., Bacelar-Nicolau, H., Nicolau, F. C.: Distribution of the affinity coefficient between variables based on the Monte Carlo simulation method. Asian Journal of Applied Sciences. **1**(5), 236–245 (2013a)
18. Sousa, Á., Tomás, L., Silva, O., Bacelar-Nicolau, H.: Symbolic data analysis for the assessment of user satisfaction: an application to reading rooms services. European Scientific Journal (ESJ). Special/Edition **3**, 39–48 (2013b)
19. Sousa, Á., Nicolau, F., Bacelar-Nicolau, H., Silva, O.: Cluster analysis using affinity coefficient in order to identify religious beliefs profiles. European Scientific Journal (ESJ). Special/Edition **3**, 252–261 (2014)