Hierarchical Cluster Analysis of Groups of Individuals: Application to Business Data

Áurea Sousa [1], Helena Bacelar-Nicolau[2], and Osvaldo Silva[3]

[1] Dep. of Math., CEEAplA, University of Azores, Ponta Delgada, Azores, Portugal
(Email: aurea@uac.pt)
[2] Faculty of Psychology, LEAD; ISAMB, CEA; University of Lisbon, Lisboa, Portugal
(Email: hbacelar@fp.ul.pt)
[3] Dep. of Math., CES-UA, University of Azores, Ponta Delgada, Azores, Portugal
(Email: osilva@uac.pt)

Abstract: We present one example, in which the data are issued from a questionnaire in order to find satisfaction typologies (with the services provided by an automobile company) of independent groups of individuals. The Agglomerative Hierarchical Cluster Analysis (*AHCA*) was based on two approaches: one based on a particular case of the generalized weighted affinity coefficient, which deals with classical data, and the other one on the weighted generalized affinity coefficient for the case of symbolic/complex data. Both measures of comparison between elements were combined with classical and probabilistic aggregation criteria. We used the global statistics of levels (*STAT)* to evaluate the quality of the obtained partitions.

Keywords: Hierarchical cluster analysis, Affinity coefficient, Independent groups of individuals, *VL* Methodology, Classical data, Symbolic data.

1 Introduction

Recent computational advances allow us to summarize very large datasets in terms of their underlying concepts, which can only be described by symbolic or complex data. Each entry of a symbolic data table can contain one or several values such as subsets of categories, intervals of the real dataset $\mathcal{R}$, or frequency distributions (e. g., Bacelar-Nicolau, 2000; Bock and Diday, 2000; Bacelar-Nicolau et al., 2009, 2010). A symbolic variable Y with domain (or range or observation space) $\mathcal{Y}$ is a mapping $E \longrightarrow B$ defined on a set E of statistical entities (individuals, classes, objects,…). Depending of the specification of B in terms of $\mathcal{Y}$, symbolic variables can be classified as: classical single-valued, set-valued, interval, multi-valued (categorical or quantitative), and modal (probabilistic) variables. A variable Y is modal with observation space $\mathcal{Y}$ if, for each $a \in E, Y(a) = \pi_a$ is a non-negative measure on $\mathcal{Y}$, such as a frequency distribution, a probability distribution or a weighting (Bock and Diday, 2000). Here, in the case of symbolic data we will focus on Ascendant Hierarchical Cluster Analysis (*AHCA*) of data units described by modal variables. The *VL* methodology (*V* for Validity, *L* for Linkage) is a probabilistic approach for clustering methods, based on the cumulative distribution function of basic similarity coefficients, and the probabilistic aggregation criteria

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

under this methodology resort essentially to probabilistic notions for the definition of the comparative functions (e.g. Lerman 1970, 1981; Nicolau, 1983; Bacelar-Nicolau, 1985, 1987, 1988; Nicolau and Bacelar-Nicolau, 1998). In this work, two classical aggregation criteria, Single Linkage (*SL*) and Complete Linkage (*CL*), as well as three probabilistic aggregation criteria - in the scope of the *VL* methodology- *AVL*, *AV1*, and *AVB*, are used to look for satisfaction typologies of independent groups of individuals in two contexts: classical data and symbolic/complex data. The measures of comparison between elements are based on the affinity coefficient.

Two different approaches of *AHCA* of independent groups of individuals are described in Section 2. In the first one the data units (independent groups of individuals) are described by classical single-valued variables defined on an ordinal scale and a particular case of the generalized weighted affinity coefficient was used. The second one is based on the weighted generalized affinity coefficient for the case of symbolic data. In Section 3 we refer some experimental results from Business area. Section 4 contains some concluding remarks about this work and its results.

2 AHCA of independent groups of subjects

From the affinity coefficient between two discrete probability distributions proposed by Matusita (1951) as the basic similarity measure for comparing two probability laws of the same type, Bacelar-Nicolau (1980, 1988) introduced the affinity coefficient, as a basic similarity coefficient between pairs of variables or of subjects in cluster analysis context (corresponding to pairs of columns or rows of a data matrix). Later on she extended that coefficient to different types of data, including complex or symbolic data and variables of mixed types (heterogeneous data), possibly with different weights (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010). The extension of the affinity coefficient for the case of symbolic data is called weighted generalized affinity coefficient. In the present work, we use two different approaches of *AHCA* of independent groups of individuals based on two different generalized approaches for the affinity coefficient.

Approach 1: particular case of the weighted generalized affinity coefficient

In this approach, the data are initially represented in G tables (one table for each one of the independent groups of individuals), containing, respectively, N_1, N_2,…, N_G, individuals described by p identical variables defined on an ordinal scale. Later, G new tables, each one containing the same number $n = min\{N_1, N_2, \ldots, N_G\}$ of individuals (selected from a stratified random sampling) have to be obtained from the initial corresponding tables. Each new table corresponds to a (n×p) data table, and x_{ihj} (i=1,…, n, h=1,..., G, j=1,..., p) is the value of the individual i, belonging to the table T_h (abbreviated, h), in the j-*th* variable (see Table 1). Then, the total scores of each independent group of individuals in each variable are computed as follows, where $x_{\bullet hj} = \sum_{i=1}^{n} x_{ihj}$ (i=1,…, n, h=1,..., G, j=1,..., p) is the total score of the group h in the variable j (sum in the column j of T_h):

Table 1. G new tables (same number $n = min\{N_1, N_2, \ldots, N_G\}$ of subjects)

	T_1 (Group 1)				T_G (Group G)
Ind. i	V_1 ⋯ V_p		*Ind. i*	V_1 ⋯ V_P
1	$x_{111} \cdots x_{11p}$	⋯	1	$x_{1G1} \cdots x_{1Gp}$
2	$x_{211} \cdots x_{21p}$	⋯	2	$x_{2G1} \cdots x_{2Gp}$
⋮	⋮	⋯	⋮	⋮
n	$x_{n11} \cdots x_{n1p}$	⋯	*n*	$x_{nG1} \cdots x_{nGp}$
Total	$x_{\bullet 11} \cdots x_{\bullet 1p}$		*Total*	$x_{\bullet G1} \cdots x_{\bullet Gp}$

The computation of the affinity coefficient between the groups h and h', with h, h'=1, …, G, and $h \neq h'$, is based on the following data matrix (Table 2), and in the formula (1):

Table 2. Classical data matrix (approach 1)

	V_1 ⋯ V_p
Group 1	$x_{\bullet 11} \cdots x_{\bullet 1p}$
Group 2	$x_{\bullet 21} \cdots x_{\bullet 2p}$
⋮	⋮
Group G	$x_{\bullet G1} \cdots x_{\bullet Gp}$

$$a(h,h') = \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{x_{\bullet hj}}{x_{\bullet h\bullet}} \cdot \frac{x_{\bullet h'j}}{x_{\bullet h'\bullet}}}\,, \qquad (1)$$

where $x_{\bullet h\bullet} = \sum_{j=1}^{p} x_{\bullet hj}$ (respectively, $x_{\bullet h'\bullet} = \sum_{j=1}^{p} x_{\bullet h'j}$) is the total score of the group h, in the p variables [sum in the row h (respectively, h') of Table 2]:

Approach 2: weighted generalized affinity coefficient (case of modal data)
Given a set of N data units (typically groups of individuals: symbolic data units) described by p modal variables, $Y_1,\ldots,Y_p$ (each variable may have a different number of "modalities"), the weighted generalized affinity coefficient between the data units k and k' is given by:

$$a(k,k') = \sum_{j=1}^{p}\pi_j\, aff(k,k';j) = \sum_{j=1}^{p}\pi_j \sum_{\ell=1}^{m_j}\sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}} \cdot \frac{x_{k'j\ell}}{x_{k'j\bullet}}} \qquad (2)$$

where: $aff(k,k';j)$ is the generalized local affinity between k and k' over the j-th variable, m_j is the number of modalities of the j-th variable; $x_{kj\ell}$ is the number of individuals (in the unit k) which share the category ℓ of Y_j; $x_{kj\bullet} = \sum_{\ell=1}^{m_j} x_{kj\ell}$,

$x_{k'j\bullet} = \sum_{\ell=1}^{m_j} x_{k'j\ell}$, and the weights, π_j, verify the condition : $\pi_j \geq 0$ *and* $\sum \pi_j = 1$ (see Table 3).

Either the local affinities or the whole weighted generalized affinity coefficient, take values in the interval [0,1] and satisfy a set of proprieties which characterize affinity measurement as a robust similarity coefficient (e.g., Bacelar-Nicolau, 2002; Bacelar-Nicolau et al., 2009). The coefficient associated to the first approach is a particular case of the coefficient associated to this second approach.

Table 3. Symbolic data matrix $\underline{X}$ with integer frequency distributions

	⋯	Y_j	⋯	$Y_{j'}$	⋯
⋮	⋯	⋯	⋯	⋯	⋯
k	⋯	$\left(x_{kj1},\cdots,x_{kjm_j}\right)$	⋯	$\left(x_{kj'1},\cdots,x_{kj'm_{j'}}\right)$	⋯
⋮	⋯	⋯	⋯	⋯	⋯
k'	⋯	$\left(x_{k'j1},\cdots,x_{k'jm_j}\right)$	⋯	$\left(x_{k'j'1},\cdots,x_{k'j'm_{j'}}\right)$	⋯
⋮	⋯	⋯	⋯	⋯	⋯

This approach is appropriated when we deal with large datasets.

3 Experimental results based on business data

Data were collected using a questionnaire applied to 450 customers in order to evaluate the satisfaction (latent variable) with the services provided by an automobile company, based on 18 component variables, which are described in Sousa et al. (2014). The variables (items) are measured in a scale with ordered modalities (1- *very dissatisfied (VD)*, 2- *generally dissatisfied (GD)*, 3- *neither satisfied nor dissatisfied (NSND)*, 4- *generally satisfied (GS)* and 5- *very satisfied (VS)*). The respondents are distributed by 11 professional occupations (O1- *Doctors, architects and engineers*; O2- *Teachers*; O3-*Businessmen*; O4-*Salesmen*; O5-*Employees of banks and insurance companies*; O6-*Military and police*; O7- *Administrative and similar*; O8- *Employees of the civil construction*; O9- *Employees of the commerce and industry*; O10- *Employees of hotels and restaurants*; O11- *Employees of other services*. The numbers of individuals in each modality of the variable "*Professional occupation*", with 11 modalities, are respectively 45, 40, 79, 42, 38, 40, 35, 34, 51, 24, 22.

The clustering of the 11 professional occupations was based on two approaches (see Section 2). The measures of comparison between elements were combined with two classical aggregation criteria, *Single Linkage* (*SL*) and *Complete Linkage* (*CL*), and three probabilistic aggregation criteria, *AVL*, *AV1*, and *AVB.* In the present work, the validation of the results is based on the global statistics of levels (*STAT*), as proposed by Lerman (1970, 1981) and Bacelar-Nicolau (1980, 1985), in both paradigms (classical and symbolic data).

In the first approach the data were initially represented in *11* tables (one table for each professional occupation), containing, respectively, 45, 40, 79, 42, 38, 40, 35, 34, 51, 24 and 22 subjects, described by *18* identical variables. Then, *11* new tables, composed by n=22 ($n = min\{45, 40, 79, 42, 38, 40, 35, 34, 51, 24, 22\}$) rows (selected from a stratified random sampling) were obtained from the initial corresponding tables (see Table 1). The *AHCA* of the professional occupations was based on a classical data matrix, as Table 2, composed by 11 rows and 18 variables (V_1 to V_{18}). The entry corresponding to the intersection between the *h-th* row and the *j-th* column of this data matrix contains the total scores of the group h (h=1,…,11) in the variable j (j=*1,…,18*). In this approach, the value of the affinity coefficient between the professional occupations h (O_h) and h' ($O_{h'}$) is given by formula (1).

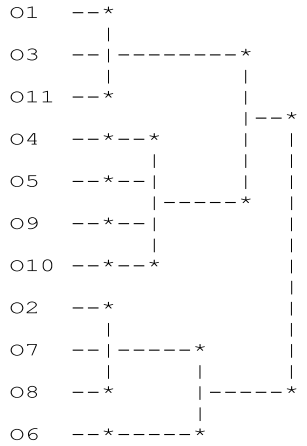

Fig. 1. Dendrogram obtained with *CL, AVL, AV1* and *AVB* (Approach 1)

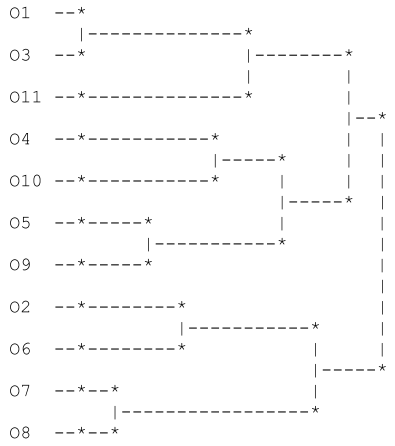

Fig. 2. Dendrogram obtained with *AVL* and *AV1* methods (Approach 2)

The selected partition is the partition into two clusters (*STAT*=5.5222), which was obtained at level 9 by all aggregation criteria (see Figure 1).

In the second approach (case of a symbolic data table for modal variables), from the initial data table (450 × 18), the subjects were distributed into 11 groups according to their professional occupation. The data units, O1 to O11, contain, respectively, 45, 40, 79, 42, 38, 40, 35, 34, 51, 24 and 22 individuals and each entry of the new data table contains a frequency distribution. In fact, the *11* professional occupations correspond to symbolic data units (rows of a symbolic data table as Table 3) described by *18* modal variables (V_1 to V_{18}) .

Figure 2 shows the dendrogram associated with the *AVL* and *AV1* methods. The best partition is the partition into three clusters (*STAT*=5.5372), which was obtained at level 8 by all aggregation criteria.

The clustering results provided by both approaches were compared. Note that at levels 7 and 8 both approaches provide the same partitions (respectively, into two and into three clusters).

Table 4. Responses given by the subjects belonging to each cluster (%)

			V1					V2					V3		
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	0%	0%	63%	25%	12%	0%	8%	14%	79%	0%	0%	0%	0%	4%	96%
C2	0%	0%	30%	65%	5%	0%	6%	55%	35%	3%	0%	0%	3%	45%	52%
C3	0%	0%	3%	96%	1%	0%	5%	81%	7%	8%	0%	0%	3%	82%	15%
			V4					V5					V6		
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	0%	0%	76%	20%	4%	8%	13%	55%	23%	2%	0%	8%	9%	79%	5%
C2	0%	0%	41%	57%	2%	4%	16%	69%	11%	0%	0%	6%	48%	41%	5%
C3	0%	0%	9%	91%	0%	1%	13%	80%	7%	0%	0%	5%	80%	7%	9%
				V7				V8					V9		
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	0%	5%	8%	86%	0%	0%	0%	4%	92%	4%	0%	0%	0%	4%	96%
C2	0%	4%	45%	48%	4%	0%	0%	37%	56%	7%	0%	0%	8%	40%	52%
C3	0%	3%	79%	11%	8%	0%	0%	66%	21%	13%	0%	0%	10%	75%	15%
			V10					V11					V12		
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	0%	0%	4%	51%	45%	0%	0%	3%	27%	69%	0%	0%	0%	17%	83%
C2	0%	3%	15%	59%	23%	0%	3%	14%	54%	29%	0%	0%	12%	41%	47%
C3	0%	3%	22%	70%	4%	0%	3%	18%	74%	5%	0%	0%	17%	68%	15%
			V13					V14				V15			
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	4%	16%	47%	27%	5%	0%	0%	14%	86%	0%	0%	0%	0%	15%	85%
C2	4%	16%	62%	15%	3%	0%	3%	45%	45%	6%	0%	0%	12%	43%	45%
C3	4%	9%	73%	11%	2%	0%	3%	79%	7%	11%	0%	0%	17%	73%	9%
			V16					V17					V18		
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C1	0%	0%	75%	16%	9%	0%	0%	75%	12%	13%	3%	17%	51%	23%	5%
C2	0%	0%	39%	59%	3%	0%	0%	39%	57%	5%	2%	18%	69%	9%	2%
C3	0%	0%	7%	93%	1%	0%	0%	7%	93%	1%	0%	13%	77%	10%	0%

The differences between the clustering results appear to be due, in part, to the sampling process associated to the first approach and to the fact that in this approach we work only with the total scores of each independent group of individuals in each variable. Thus, in the remainder text, we will only refer to the best partition provided by the second approach: ***Cluster 1:*** {O1, O3, O11}; ***Cluster 2:*** {O4, O5, O9, O10}; ***Cluster 3:*** {O2, O6, O7, O8}. From the observation of Table 4, it can be seen some of the main differences between the profiles associated to these three clusters.

In a 2D Zoom Star, axes are linked by a line that connects most frequent categorical values of each variable, so it allows us to identify the main characteristics of the objects. Figure 3 shows the 2D Zoom Stars associated to the clusters of the second approach. We observe that, for instance, most respondents included into cluster 3 are: generally satisfied with the aspects associated to variables V1, V3, V4, V9, V10, V11, V12, V15, V16 and V17; and neither satisfied nor dissatisfied with the aspects associated to variables V2, V5, V6, V7, V8, V13, V14 and V18 (see Figure 3 and Table 4).

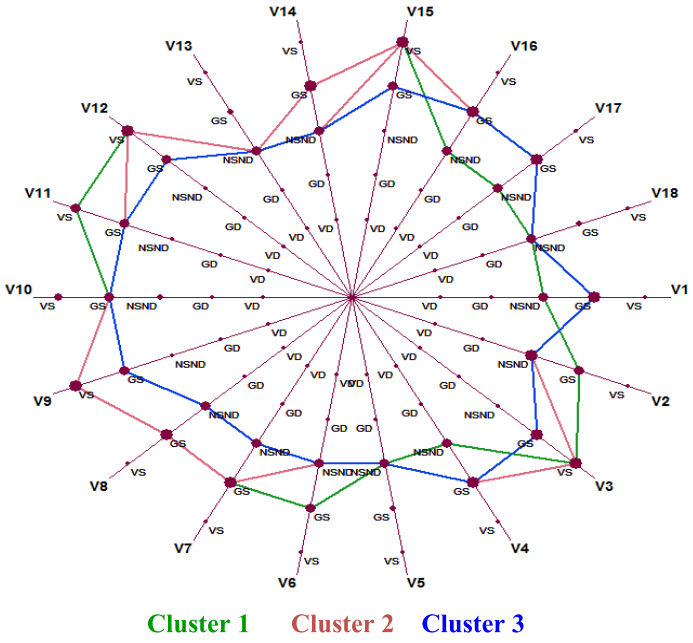

Fig. 3. 2D *Zoom Stars representation for the clusters- Approach 2*

4 Conclusions

The knowledge about the satisfaction profiles is useful, because customers respond better to the Market segmentation strategies which address their specific needs.

In the case of the first approach we loss information because we can´t work with the entire sample but only with a stratified random subsample, and this approach works only with the total scores of each independent group of individuals in each variable (that is, we also loss information about the scores of the groups in the modalities of the variables). Contrary, using the second approach (weighted generalized affinity coefficient, for complex or symbolic objects) it is possible to work with the entire dataset, and with the frequency distributions associated to the scores of each independent group of individuals in the modalities of each variable.
The differences between the clustering results (satisfaction typologies) provided by the two approaches of *AHCA* of independent groups of individuals were due, in part, to the smaller number of individuals of each group when we apply the first approach as a consequence of the sampling process. Nevertheless, we might have opted by inquiring a larger number of individuals in each group, during the planning of the investigation.

References

1. Bacelar-Nicolau, H., Contributions to the Study of Comparison Coefficients in Cluster Analysis, PhD Thesis (in Portuguese), Universidade de Lisboa (1980).
2. Bacelar-Nicolau, H., The affinity coefficient in cluster analysis, Methods of Operations Research, vol. 53, Martin J. Bekmann et al (ed.), Verlag Anton Hain, Munchen, pp. 507-512 (1985).
3. Bacelar-Nicolau, H., On the distribution equivalence in cluster analysis, In Proceedings of the NATO ASI on Pattern Recognition Theory and Applications, Springer - Verlag, New York, pp. 73-79 (1987).
4. Bacelar-Nicolau, H., Two Probabilistic Models for Classification of Variables in Frequency Tables, In: Bock, H.-H. (Eds.), Classification and Related Methods of Data Analysis, North Holland, pp. 181-186 (1988)
5. Bacelar-Nicolau, H., The Affinity Coefficient, In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday (Eds.), Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, pp. 160-165 (2000).
6. Bacelar-Nicolau, H., On the Generalised Affinity Coefficient for Complex Data, *Biocybernetics and Biomedical Engineering*, vol. 22, no. 1, pp. 31-42, (2002).
7. Bacelar-Nicolau, H.; Nicolau, F.C.; Sousa, Á.; Bacelar-Nicolau, L., Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets, *Biocybernetics and Biomedical Engineering*, vol. 29, no. 2, pp. 9-18 (2009).
8. Bacelar-Nicolau, H.; Nicolau, F.C.; Sousa, Á.; Bacelar-Nicolau, L., Clustering Complex Heterogeneous Data Using a Probabilistic Approach, In *Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference* (SMTDA2010), pp. 85-93 (2010) (electronic publication).
9. Bock, H.-H. and Diday, E., Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin (2000).
10. Lerman, I. C., Sur l`Analyse des Données Préalable à une Classification Automatique (Proposition d'une Nouvelle Mesure de Similarité), Rev. Mathématiques et Sciences Humaines, vol. 32, no. 8, pp. 5-15 (1970).

11. Lerman, I. C., Classification et Analyse Ordinale des Données, Dunod, Paris (1981).
12. Matusita, K., On the Theory of Statistical Decision Functions, Ann. Instit. Stat. Math, vol. III, pp. 1-30 (1951).
13. Nicolau, F.C., Cluster Analysis and Distribution Function, Methods of Operations Research, vol. 45, pp. 431-433 (1983).
14. Nicolau, F.C. and Bacelar-Nicolau, H., Some Trends in the Classification of Variables, In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Eds.), Data Science, Classification, and Related Methods. Springer-Verlag, pp. 89-98 (1998).
15. Sousa, Á., Bacelar-Nicolau, H., Silva, O., Cluster Analysis of Business Data. *Asian Journal of Business and Management*, 2(1) 18-26 (2014).