CLUSTER ANALYSIS USING AFFINITY COEFFICIENT IN ORDER TO IDENTIFY RELIGIOUS BELIEFS PROFILES

Aurea Sousa, PhD
University of the Azores, Department of Mathematics, CEEAplA, and CMATI, Portugal
Fernando C. Nicolau, PhD
New University of Lisbon, Department of Mathematics, Portugal
Helena Bacelar-Nicolau, PhD
University of Lisbon, Faculty of Psychology, Laboratory of Statistics and Data Analysis, Portugal
Osvaldo Silva, PhD
University of the Azores, Department of Mathematics; CES, and CMATI, Portugal

Abstract

We present an application of Ascendant Hierarchical Cluster Analysis (AHCA) to a dataset related to religion, in order to find a typology of religious beliefs profiles of individuals who live on São Miguel island (Azores) according to the frequency they go to the Mass. AHCA was based on the weighted generalized affinity coefficient for symbolic or complex data, and on classical and probabilistic aggregation criteria; the probabilistic ones belong to a parametric family of methods in the scope of the VL methodology. Additionally, we applied some validation measures (based on the values of the proximity matrix and adapted for the case of similarity measures) to evaluate the obtained results (clusters and partitions).

Keywords: Cluster analysis, affinity coefficient, VL methodology, complex data, validation measures

Introduction

With the advent of computers, it is usual to record very large datasets, so it is imperative to summarize these data in terms of their underlying concepts, which can only be described by a more complex type of data, called symbolic data (Diday, 2000). Rows correspond to symbolic objects (data units), whereas columns correspond to symbolic variables, which may take values such as subsets of categories, intervals of real axes, or frequency distributions. They are also called complex data. Each entry of the table can contain just one value or several values (Bacelar-Nicolau, 2000; Bock and Diday, 2000; Doria et al., 2013). Thus, the symbolic data types are generalizations of classical data types.

An important source of symbolic objects is provided by relational databases containing a set of individuals that are distributed into some groups. Moreover, the symbolic objects can be used to define queries from a database and for concept propagation between databases (Bock and Diday, 2000).

A modal variable Y, with domain (or range or observation space) y, defined on a set E={a, b, ...} of objects, is a mapping $Y(a)=(U(a), \pi_a)$, $a \in E$, where π_a is a non-negative measure in y, such as a frequency distribution, a probability or a weight distribution on the domain y and $U(a) \subseteq y$ is the support for π_a in the domain y. If π_a is specified by a histogram, Y is called a histogram variable. Y is a (bar or) diagram variable if the observation space y is finite and π_a is described by a bar diagram (Bock and Diday, 2000). In this paper, we

concentrate on data units described by modal variables and use the weighted generalized affinity coefficient (Bacelar-Nicolau, 2000) as the basis of hierarchical clustering algorithms (classical and probabilistic aggregation criteria) in our approach to this subject.

Section 2 is devoted to the weighted generalized affinity coefficient for the case of modal variables in the field of Symbolic Data Analysis. Some measures of validation to evaluate the quality of the results of an AHCA are referred in Section 3. We present, in Section 4, the main results (best partitions according to some validation measures) obtained with the application of the AHCA to complex data related to religion, in order to investigate the proximity of religious beliefs profiles of individuals who live on São Miguel island according to the frequency they go to the Mass. Finally, Section 5 contains some final considerations about the developed work and the obtained results.

Weighted Generalized Affinity Coefficient for the Case of Modal Data

In the scope of Cluster Analysis, Bacelar-Nicolau (1980, 1988) introduced the affinity coefficient, as a basic similarity coefficient (between the pairs of columns or rows of a data matrix), from the affinity coefficient between two discrete probability distributions proposed by Matusita (1951). Afterward, that coefficient was extended to different types of data, including complex and heterogeneous data (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al., 2009, 2010).

Considering a set of N symbolic data units described by p modal variables, $Y_1,...,Y_p$, the so called weighted generalized affinity coefficient (extension of the affinity coefficient to the case of symbolic data) $a(k,k')$ between a pair of statistical data units k, k' (k, k'=1,…,N), is defined as follows (Bacelar-Nicolau, 2000, 2002; Nicolau and Bacelar-Nicolau, 1999):

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

where: aff (k,k';j) is the generalized local affinity between k and k' over the j-th variable, m_j denotes the number of modalities of the j-th variable; $x_{kj\ell}$ is a absolute frequency or a relative frequency (real non-negative value) of individuals (in unit k) which share category ℓ of variable 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 π_j are weights such that $0 \leq \pi_j \leq 1$, $\Sigma \pi_j = 1$. The weighted generalized affinity coefficient, $a(k,k')$, takes values in the interval [0,1] and satisfies a set of proprieties which characterize affinity measurement as a robust similarity coefficient (e.g. Bacelar-Nicolau, 2000, 2002). The weighted generalized affinity coefficient appears to be an appropriate resemblance measure between elements (symbolic data units or symbolic variables) in cases where we are dealing with complex data from large databases.

A suitable adaptation of formula (1) may be considered if real or frequency negative values appear, and in that case the meaning of $x_{kj\ell}$ depends on the type of j-th variable. Bacelar-Nicolau et al. (2009, 2010) demonstrated that the weighted generalized affinity coefficient is appropriated when mixed and complex variables types are present in a database and the same coefficient works for those variables types. However, here the analyzed dataset contains only modal variables.

Given a similarity matrix, a dataset can be classified through classical aggregation criteria or probabilistic ones. The probabilistic approach of AHCA, named VL methodology (V for Validity, L for Linkage) is a set of agglomerative hierarchical clustering methods, based on the cumulative distribution function of basic similarity coefficients (Lerman, 1970; Bacelar-Nicolau, 1988; Nicolau and Bacelar-Nicolau, 1998).

Validation in Ascendant Hierarchical Cluster Analysis (AHCA) of Complex Data

The several comparative coefficients between elements and aggregation criteria raise pertinent questions as to identify: i) the best measure or the best criterion to use, ii) the most significant partition resulting from a classification algorithm, and iii) whether the clusters obtained reflect the real structure of the data.

Measures of validation based on the values of the proximity matrix between elements, such as, for instance, the global statistics of levels (STAT) (Bacelar-Nicolau, 1980; Lerman, 1970), the P(I2mod, ∑) measure, and the γ index, proposed by Goodman and Kruskal (1954), can be used, even in the case of symbolic data (see, Sousa et al. 2013). In addition, to determine the appropriate number of clusters, we used other two measures (adapted for the case of similarity measures) defined as follows:

The Sil measure (Sousa, 2005) is based on the Silhouette plots (Rousseuw, 1987), and if the ith object belongs to cluster C_r, which contains n_r (≥ 2) is defined by:

$$Sil(i) = \frac{\frac{1}{n_r - 1}\sum_{j \in C_r \wedge j \neq i} s_{ij} - \frac{1}{N - n_r}\sum_{j \in \Omega \setminus C_r} s_{ij}}{max\left\{\frac{1}{n_r - 1}\sum_{j \in C_r \wedge j \neq i} s_{ij}, \frac{1}{N - n_r}\sum_{j \in \Omega \setminus C_r} s_{ij}\right\}}, \qquad (2)$$

with -1≤ Sil(i) ≤ 1, where $N - n_r$ is the number of elements that do not belong to cluster C_r.

This measure takes into consideration the average of the similarities between an element i belonging to cluster C_r and all other elements that do not belong to this cluster. The determination of the average of the Sil(i) values for all objects i belonging to each cluster and for the c clusters may be useful. We can also use the transformation of the Sil(i) values defined by Sil*(i)= (1+Sil(i))/2, in order to obtain values between 0 and 1.

U Statistics (Mann and Whitney, 1947) provide relevant test statistics for assessing the adequacy of a cluster, combining the concepts of its compactness and isolation. Let:

$$U_{ijk\ell} = \begin{cases} 0 \quad se \quad s_{ij} > s_{k\ell} \\ 1/2 \quad se \quad s_{ij} = s_{k\ell} \\ 1 \quad se \quad s_{ij} < s_{k\ell} \end{cases}, \qquad (3)$$

where s_{ij} are values of the similarity matrix between pairs of elements of the set to classify.

We consider that for each cluster C of size r (Gordon, 1999):

$W \equiv \{(i, j): i, j \in C, i < j\}$ is the set of $r(r-1)/2$ within-cluster pairs, and

$B \equiv \{(k, \ell): k \in C, \ell \notin C\}$ is the set of $r(n-r)$ between-cluster pairs.

The global U index, U_G, is defined by:

$$U_G \equiv \sum_{(i,j) \in W} \sum_{(k,\ell) \in B} u_{ijk\ell}. \qquad (4)$$

The local U index, U_L, is defined by:

$$U_L \equiv \sum_{i \in C} \sum_{j \in C \setminus \{i\}} \sum_{k \notin C} u_{ijik}. \qquad (5)$$

The "best" cluster is the one that presents the smallest value of these indexes. In the case of a cluster-L* we have U_G=0 and in the case of a ball cluster we have U_L=0 (Gordon, 1999)

In a methodological framework and in order to evaluate the obtained partitions, the values of STAT, DIF, P(I2mod∑) and γ indexes (for each partition) were calculated. In

addition, the values of the Sil* index and of the U statistics were calculated for the clusters of the most significant partitions (according to the previous indexes).

Application to real data: A questionnaire related to religious beliefs

A questionnaire was used in order to investigate the proximity of religious beliefs profiles of individuals who live on São Miguel island (Azores) according to the frequency of their visits to the Mass. The initial classical data matrix (517 x 10) is constituted by 517 respondents (individuals) and 10 statements (items) corresponding to 10 categorical variables (V1 – God is one, but in three persons, V2 – Christ is God, V3- Christ performed authentic miracles, V4 – The Pope is never wrong when he speaks of the truths of faith, V5 – Something exists after death, V6 – Christ saved us by dying for our sins, V7 – The devil exists, V8 – The good are rewarded and the bad are punished in the afterlife, V9 – The sacrament of confession forgives our sins, V10 – Everyone is born with original sin), each of them, with four not ordered modalities (Believe (BEL), D- Doubt (DOUBT), Don't Believe (D_BEL), Don't know/no response (NR)).

Individuals were distributed into eight groups by a SQL query according to the frequency that the individuals of each group go to the Mass: "Never", "Rarely", "on Important Dates or Celebrations (IDC)", "only for Weddings, Baptisms or Funerals (WBF)", "Once or Twice a month (OT)", "every Sunday (S)", "every Sunday and during the Week (SW)", "when they Feel it's Necessary (FN)" . The symbolic data table (see Table 1) describes a set of eight symbolic objects (the rows) by a set of ten modal variables**.** The data units "Never", "Rarely", "IDC ", "WBF", "OT" "S", "SW and "FN" contain, respectively, 14, 60, 29, 62, 41, 228, 14 and 69 individuals and each entry of Table 1 contains a frequency distribution.

Table 1. Symbolic Data Matrix

	V1	V2	…
Never	BEL(0.29), DOUBT(0.21), D_BEL(0.29), NR(0.21)	BEL(0.36), DOUBT(0.14), D_BEL(0.36), NR(0.14)	…
Rarely	BEL(0.63), DOUBT(0.20), D_BEL(0.06), NR(0.08)	BEL(0.68), DOUBT(0.13), D_BEL(0.10), NR(0.08)	…
IDC	BEL(0.72), DOUBT(0.17), D_BEL(0.03), NR(0.07)	BEL(0.69), DOUBT(0.03), D_BEL(0.14), NR(0.14)	…
WBF	BEL(0.58), DOUBT(0.19), D_BEL(0.13), NR(0.10)	BEL(0.61), DOUBT(0.15), D_BEL(0.11), NR(0.13)	…
OT	BEL(0.85), DOUBT(0.05), NR(0.10)	BEL(0.80), DOUBT(0.10), D_BEL(0.02), NR(0.07)	…
S	BEL(0.88), DOUBT(0.05), D_BEL(0.01), NR(0.06)	BEL(0.90), DOUBT(0.03), D_BEL(0.01), NR(0.06)	…
SW	BEL(0.93), DOUBT(0.07)	BEL(0.71), DOUBT(0.07), D_BEL(0.07), NR(0.14)	…
FN	BEL(0.75), DOUBT(0.06), D_BEL(0.09), NR(0.10)	BEL(0.86), DOUBT(0.04), D_BEL(0.04), NR(0.06)	…

The AHCA of the eight symbolic data units is based on the weighted generalized affinity coefficient (Nicolau and Bacelar-Nicolau, 1999; Bacelar-Nicolau, 2000, 2002) with equal weights $(\pi_j = 1/p)$. We used four aggregation criteria, one of them classical, Single Linkage (SL), and three probabilistic, AV1, AVB, and AVL (Lerman, 1981; Bacelar-Nicolau, 1988; Nicolau, 1980, 1983; Nicolau and Bacelar-Nicolau, 1998).

Table 2. The Most Significant Partitions

	The most significant partitions	**Indexes**
SL / AV1 / AVB	Lev. 5 - {WBF, FN, IDC, Rarely, S, OT};{SW}; {Never}	STAT= 4.4914 γ=1.000
	Lev. 6- {WBF, FN, IDC, Rarely, S, OT, SW}; {Never}	P(I2mod, ∑)=0.8728
AVL	Lev. 4 - {WBF, FN, IDC, Rarely};{Never}, {S, OT}; {SW}	STAT= 3.7935 P(I2mod, ∑)=0.8995 γ=0.9728

Table 2 presents the results corresponding to the most significant partitions provided by the four aggregation criteria, according to the validation indexes in the last column of this table. Table 3 presents the values of the Sil* index and the values of U statistics for the partitions presented in Table 2.

The STAT, γ indexes and the U statistics allow us to conclude that the most significant level (the best cut-off level) corresponds to a partition into three clusters given by the methods SL, AV1, and AVB: {WBF, FN, IDC, Rarely, S, OT} {SW} {Never} (see Tables 2 and 3). The first cluster contains the individuals that seldom go to the Mass and the individuals that go to the Mass with some frequency. The second cluster contains the individuals that go to the Mass every Sunday and during the week. Finally, the third cluster contains the individuals who never go to the Mass.

Table 3. U statistics and values of the Sil* index

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Partition UL/UG Sil*	{WBF, FN, IDC, Rarely, S, OT} 0/0 0.6028437	{SW} 0/0 ---	{Never} 0/0 ---	
Partition UL/UG Sil*	{WBF, FN, IDC, Rarely, S, OT, SW} 4/21 0.5666344	{Never} 0/0 ---		
Partition UL/UG Sil*	{WBF, FN, IDC, Rarely} 1/2 0.3818336	{Never} 0/0 ---	{S,OT} 0/0 0.3196724	{SW} 0/0 ---

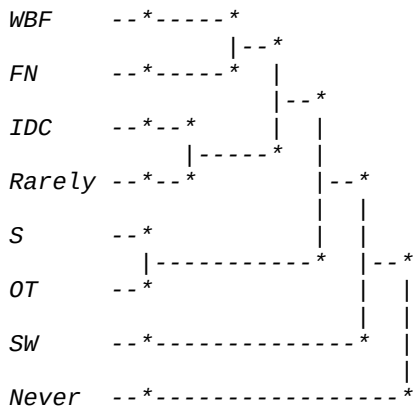

Figure 1. Dendrogram obtained by the AV1/AVB methods

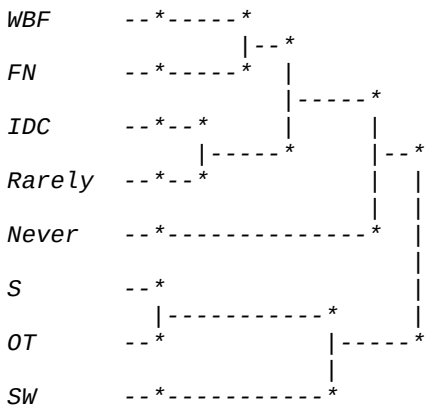

Figure 2. Dendrogram obtained by the AVL method

Figures 1 and 2 show the dendrograms associated with the AV1/AVB and AVL methods, respectively.

Table 4- Responses given by the individuals belonging to each cluster (%) - three profiles

V1					**V2**				**V3**				**V4**			
	BEL	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**
C 1	78 %	9%	5%	8 %	81 %	6%	5%	8 %	80 %	8%	2%	10 %	53 %	25%	10%	12 %
C 2	93 %	7%	0%	0 %	71 %	7%	7%	14 %	93 %	7%	0%	0 %	93 %	7%	0%	0 %
C 3	29 %	21%	29%	21 %	36 %	14%	36%	14 %	43 %	7%	29%	21 %	29 %	21%	36%	14 %

V5					**V6**				**V7**				**V8**			
	BEL	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**
C 1	55 %	21%	9%	15 %	79 %	7%	5%	10 %	35 %	26%	23%	16 %	36 %	25%	23%	16 %
C 2	79 %	21%	0%	0 %	93 %	0%	7%	0 %	57 %	21%	14%	7 %	57 %	29%	14%	0 %
C 3	21 %	36%	29%	14 %	43 %	14%	36%	7 %	21 %	14%	64%	0 %	14 %	29%	50%	7 %

	V9				**V10**			
	BEL	**DOUBT**	**D_BEL**	**NR**	**BEL**	**DOUBT**	**D_BEL**	**NR**
C 1	52%	23%	13%	12%	49%	19%	15%	17%
C 2	79%	7%	7%	7%	93%	0%	7%	0%
C 3	14%	29%	57%	0%	21%	21%	36%	21%

Reading the dendrogram associated with the AV1/AVB aggregation criteria from top to bottom (see Figure 1), the most frequent response given by the individuals of clusters 1 and 2 has been "Believe". The 2D Zoom Star, as showed in Figure 3, doesn´t distinguish the clusters 1 and 2, but from the observation of Table 4, it can be seen that there are differences between the profiles associated to these two clusters. These differences also could have been observed from a 3D Zoom Star containing the information associated with Table 4 (bar graphs for each variable). There are more individuals belonging to the cluster 2 who go to the Mass every Sunday and during the week, comparatively to the individuals belonging to the cluster 1, that believe that: "V1 – God is one, but in three persons" (93% versus 78%), "V3- Christ performed authentic miracles" (93% versus 80%), "V4 – The Pope is never wrong when he speaks of the truths of faith" (93% versus 53%), "V5 – Something exists after death" (79% versus 55%), "V6 – Christ saved us by dying for our sins" (93% versus 79%), "V7 – The devil exists" (57% versus 35%), "V8 – The good are rewarded and the bad are punished in the afterlife" (57% versus 36%), V9 – The sacrament of confession forgives our sins" (79% versus 52%), and "V10 – Everyone is born with original sin" (93% versus 49%) (see Figure 3 and Table 4). Note that none of the individuals of the cluster 2 answered "Don't Believe" nor "Don't know/no response" to the statements of the variables V1, V3, V4, and V5.

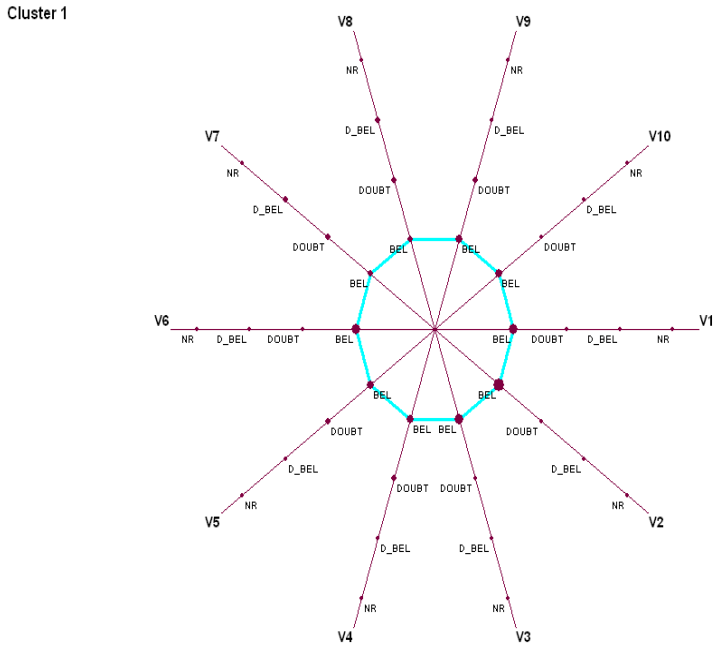

Figure 3. 2D Zoom Star representation for the clusters 1 and 2

The individuals who Never go to the Mass (individuals belonging to the cluster 3, considering the dendrogram associated with the AV1/AVB methods read from top to bottom) have a profile which can be well represented by the 2D Zoom Star showed in Figure 4, where the axes are linked by a line that connects the most frequent values of each variable (the main characteristics of the symbolic objects). Most respondents included into this cluster don't believe that "V7 – The devil exists" (64%), and that "V9 – The sacrament of confession forgives our sins" (57%). About 50% of them don't believe that "V8 – The good are rewarded and the bad are punished in the afterlife". A large proportion of these individuals don't believe that "V4 – The Pope is never wrong when he speaks of the truths of faith" (36%), and that "V10 – Everyone is born with original sin" (36 %). A large proportion of them doubts that "V5 – Something exists after death" (36%). Moreover, 29% of the individuals belonging to the cluster 3 believe that "V1 – God is one, but in three persons"

whereas 29% of these individuals don´t believe in that. 36% of them believe that "V2 – Christ is God" whereas 36% don´t believe in that. Interestingly, a large portion of them believe that "V3- Christ performed authentic miracles" (43%), and that "V6 – Christ saved us by dying for our sins" (43%).

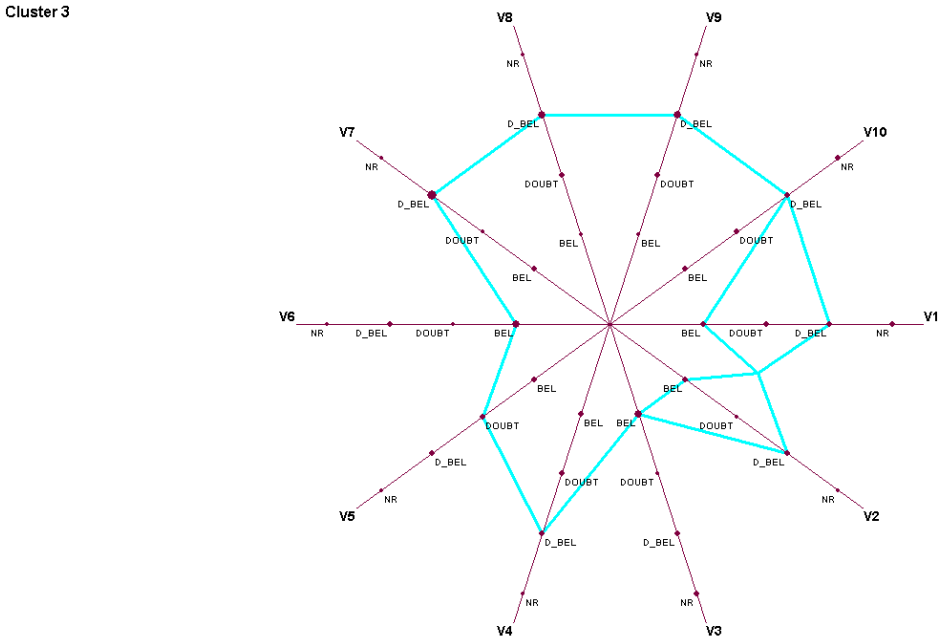

Figure 4. 2D Zoom Star representation for the cluster 3

Note that until the best cut-off level of the AVL method (level four) the hierarchies corresponding to the four obtained dendrograms provide the same four classes, although ordered in a different way (consensus partition): {{WBF, FN}, {IDC, Rarely}}, {S, OT}, {SW}, {Never } (see Figures 1 and 2). Although, the profiles corresponding to the clusters {SW} and {Never} are very different, in the dendrogram associated with the AV1/AVB methods, these two clusters are joined in the latest levels in a chain effect. The subclasses {S, OT} and {IDC, rarely"} were found by all aggregation criteria applied, and the subclass, {WBF, FN} was found by all aggregation criteria except by the SL.

Conclusion

The example presented allowed us to illustrate the application of the weighted generalized affinity coefficient in the classification of complex or symbolic data units described by modal variables, and the extension of the VL methodology to the classification of this type of data. Taking into consideration the importance of the validation of the results of a Cluster Analysis, we described, in Section 3, the extension of some validation indexes used in the case of classical data matrixes to the case of validation in AHCA of symbolic data.

From the application of the AHCA to the data set under investigation and using the referred validation measures, findings indicate a robust typology of religious beliefs of the individuals of our sample according to the frequency that they go to the Mass. The three selected clusters (represented by new symbolic objects) correspond to distinct profiles of religious beliefs, which were characterized in the last section by a p-multivariate vector of relative frequencies (expressed in %). The applied measures of validation proved usefulness to determine the appropriate cut-off levels of the dendrograms. Moreover, the consensus

partition into four classes (obtained at level four by all applied aggregation criteria) is also conceptually relevant.

References:

Bacelar-Nicolau, H. Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória. Tese de Doutoramento, Universidade de Lisboa, 1980.

Bacelar-Nicolau, H. 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., pp.181-186, 1988.

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. Berlin: Springer-Verlag, pp. 160-165, 2000.

Bacelar-Nicolau, H. On the Generalised Affinity Coefficient for Complex Data. Biocybernetics and Biomedical Engineering 22 (1), pp. 31-42, 2002.

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

Bacelar-Nicolau, H.; Nicolau, F.; Sousa, Á.; Bacelar-Nicolau, L. 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), 2010.

Bock, H.-H. and Diday, E., eds. Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin: Springer-Verlag, 2000.

Diday, E. Symbolic Data Analysis and the Sodas Project: Purpose, History, Perspective. In: Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H.Bock and E.Diday, eds. Springer, pp. 1-23, 2000.

Doria, I.; Sousa, Á.; Bacelar-Nicolau, H.; Le Calvé, G. Comparison of Modal Variables Using Multivariate Analysis. In: Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications. Studies in Theoretical and Applied Statistics, J.L. da Silva, F. Caeiro, I. Natário, and C.A. Braumann, eds. Berlin, Heidelberg: Springer, pp. 363-370, 2013.

Goodman, L. A. and Kruskal, W.H. Measures of Association for Cross-Classifications. Journal of the American Statistical Association, 49, pp. 732-64, 1954.

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

Lerman, I.C. Sur l`Analyse des Données Préalable à une Classification Automatique. Proposition d`une Nouvelle Mesure de Similarité, rapport No. 32, 8e. année, MSH, Paris, 1970.

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

Matusita, K. On the Theory of Statistical Decision Functions. Ann. Instit. Stat. Math., III, pp. 1-30, 1951.

Mann, H. and Whitney, D. On a Test of whether One of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, pp. 50-60, 1947.

Nicolau, F. Critérios de Análise Classificatória Hierárquica baseados na Função de Distribuição. Tese de Doutoramento, FCL, Universidade de Lisboa, 1980.

Nicolau, F. Cluster Analysis and Distribution Function. Methods of Operations Research, 45, pp. 431-433, 1983.

Nicolau, F. and Bacelar-Nicolau, H. 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 and Y. Baba, eds. Springer-Verlag, pp. 89-98, 1998.
Nicolau, F. and Bacelar-Nicolau, H. Clustering Symbolic Objects Associated to Frequency or Probability Laws by the Weighted Affinity Coefficient. In: Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society, H. Bacelar-Nicolau, F. C. Nicolau and Jacques Janssen, eds. INE, Lisboa, Portugal, pp. 155-158, 1999.
Rousseuw, P.J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, pp. 53-65, 1987.
Sousa, Á. Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa. Tese de Doutoramento, Universidade dos Açores, Ponta Delgada, 2005.
Sousa, Á.; Tomás, L.; Silva, O.; Bacelar-Nicolau, H. Symbolic Data Analysis for the Assessment of User Satisfaction: An Application to Reading Rooms Services. Proceedings of 1 st Annual International Interdisciplinary Conference, AIIC 2013, 24-26 April, Portugal, pp. 39-48, European Scientific Journal ESJ June 2013/Special/Edition nº 3, 2013.