Biometrical Letters
Vol. 47 (2010), No. 1, 45-56

Weighted Generalised Affinity Coefficient in Cluster Analysis of Complex Data of Interval Type

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

[1]Department of Mathematics, University of Azores, 9501-855-Ponta Delgada, Portugal,
e-mail: aurea@uac.pt; osilva@uac.pt
[2]Department of Mathematics, FCT, New University of Lisbon, 2829-516-Caparica, Portugal,
e-mail: fcnicolau@gmail.com
[3]Laboratory of Statistics and Data Analysis, FPCE, University of Lisbon, 1649-013-Lisboa,
Portugal, e-mail: hbacelar@fpce.ul.pt

SUMMARY

Complex Data Analysis is a relatively new field that provides a range of methods for analysing complex/symbolic data, and can be defined as the extension of standard data analysis to more complex data tables. There are two steps in Complex or Symbolic Data Analysis: i) knowledge extraction from large databases as in Data Mining; and ii) application of new tools to the extracted knowledge in order to extend Data Mining to Knowledge Mining. The weighted generalised affinity coefficient appears to be an appropriate resemblance measure between elements (statistical data units or variables) in cases where we deal with complex data from large databases. In this work we apply two different processes to determine values of the weighted generalised affinity coefficient in the case where we are dealing with data units described by variables whose values are intervals of the real axis.

We present one example concerned with real data (with a known structure) in the field of Biometry, in which objects are described by variables whose values are intervals, in order to illustrate the effectiveness of Ascendant Hierarchical Cluster Analysis based on the weighted generalised affinity coefficient and classical and/or probabilistic aggregation criteria. In this example, we applied a method of validation to identify the best partitions.

Key words: Cluster Analysis, VL Methodology, Weighted Generalised Affinity Coefficient, Symbolic Data, Measures of Validation.

1. Introduction

Classical data analysis starts with a given number *n* of individuals (often termed objects, cases, etc.) characterized by *p* variables Y_1, …, Y_p. Each variable Y_j

takes values in an observation space y_j of possible levels, alternatives or numbers, and for each individual *k* the variable Y_j takes just one single value $Y_j(k)$ from $\mathsf{y}_{j.}$ which can be denoted by $x_{kj.}$

With the development of computer technology it is usual to record huge sets of data in large databases, so it is important to summarize these data in terms of their underlying concepts. These concepts can only be described by a more complex type of data, called symbolic data (Bock and Diday, 2000).

In a symbolic data table the rows correspond to symbolic objects and the columns correspond to symbolic variables, which can take values as a single quantitative value, a single categorical value, a set of values or categories (multivalued variable), an interval, or a set of values with associated weights. Thus symbolic data tables may describe heterogeneous data, and their cells may contain data of different types that can be weighted and linked by logical rules and taxonomies (Bock and Diday, 2000). Thus, formally, a symbolic variable Y with domain (or range or observation space) y is a mapping $E \longrightarrow \mathsf{B}$ defined on a set E of statistical entities (individuals, classes, objects, …). Depending on the specification of B in terms of y, symbolic variables can be classified as (Bock and Diday, 2000):

(i). classical single-valued if $\mathsf{B}=\mathsf{y}$.

(ii). set-valued if $Y(a) \subseteq \mathsf{y}$ is a subset of y.

(iii). interval if, for all $a \in E$, $Y(a) = [\alpha, \beta]$ is an interval of y, in the order established on y.

(iv). multi-valued (categorical or quantitative) if set-valued with $Y(a) \subseteq \mathsf{y}$ and $|Y(a) < \infty|$, $\forall \ a \in E$.

(v). modal (probabilistic) with observation space y if, for each $a \in$ E, $Y(a)=\pi_a$ is a non-negative measure on y, such as a frequency distribution, a probability distribution or a weighting. 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.

Symbolic Data Analysis (*SDA*) is a relatively new field that provides a range of methods for analysing symbolic data, and can be defined as the extension of standard data analysis to symbolic data tables (Bock and Diday, 2000). There are two steps in Symbolic Data Analysis (SDA): i) knowledge extraction from large data bases as in Data Mining; and ii) application of new tools to the extracted knowledge in order to extend Data Mining to Knowledge Mining. The symbolic objects allow us to make a mathematical modelling of concepts and are used as input and as an explanatory output of an SDA.

An important source of symbolic objects is provided by relational databases containing a set of individuals that are distributed into some groups. DB2SO is the part of the Sodas software (Bock and Diday, 2000) which enables a user to build a set of assertions, one assertion for each group of individuals, from data stored in a relational database. The usual interaction between the user and DB2SO includes connection with a database and retrieval of individuals distributed into groups (symbolic objects) by means of a SQL query. On the other hand, the symbolic objects can be used to define queries from a database and for concept propagation between databases (Bock and Diday, 2000).

In the case of data sets of (very) high dimension, one of the possible solutions for their analysis is to find clusters in these data. Cluster Analysis (classical and symbolic) aims to construct an appropriate classification either of the set E of data units or the set Y of variables, from one given data matrix ($N \times p$). As in the classic case, the goal is to obtain homogeneous clusters of objects in a population Ω or E, such that objects of the same cluster present a high similarity and objects of different clusters present more dissimilarities. Some comparison measures between elements, e.g. Gower’s similarity coefficient (Gower, 1971), and the dissimilarity measures of Gowda and Diday (1991) allow us to apply Hierarchical Cluster Analysis to data of mixed types. However, its application is limited to data sets of small dimension. The majority of Cluster Analysis methods are either too complex to be applied to data sets of high dimension (e.g. hierarchical methods), or are implemented only for numerical data (e.g. the k-means method).

In Section 2 we describe two different processes for determining values of the weighted generalised affinity coefficient in the case where we deal with data units described by variables whose values are intervals of the real axis.

In Section 3, we present the results of an example related to real data (with known structure), obtained using the AHCA of a symbolic data matrix whose symbolic objects are described by variables whose values are intervals of the real axis. In this example, we applied a method of validation to identify the best partitions.

2. Weighted Generalised Affinity Coefficient in Symbolic Cluster Analysis

Nicolau and Bacelar-Nicolau (1999) proposed the weighted generalised affinity coefficient for the case in which symbolic objects are described by *p* probability or frequency distribution vectors, or "some other data support which can be applied to this type of description, such as histograms and variables whose values are intervals of the real axis" (Nicolau and Bacelar-Nicolau, 1999; Bacelar-Nicolau, 2000, 2002). In particular, if the data units are described by variables whose values are intervals of the real axis, we can use the following processes:

Process 1: Before calculation of the generalised affinity coefficient (for the case of symbolic data), an appropriate relative frequency distribution is associated with each of the intervals, thus obtaining a codification of the *Symbolic Data Matrix* by discretization of the variables of interval type. It is possible to effect this codification (Sousa, 2005) in the following way:

Let A=[*LinfA*, *LsupA*] and B=[*LinfB*, *LsupB*] be two intervals with ranges, respectively *AmplA* and *AmplB*. The relative frequency distributions associated with the intervals *A* and *B* are obtained in the following way:

i) Determination of I=[inters1, inters2] corresponding to the intersection of the intervals A and B;

ii) Calculation of the frequency distributions of the interval A: Let ax1, ax2 and ax3 be three auxiliary real variables and (freqA1, freqA2, freqA3) the relative frequency distribution associated with the interval A.

```
If (LinfA=inters1) then ax1=0 and ax1=inters1-min(LinfA, LinfB) if not.
freqA1=ax1/amplA
ax2=inters2-inters1
freqA2=ax2/amplA
If (LsupA=inters2) then ax3=0 and ax3=max(LsupA, LsupB)-inters2 if not.
freqA3=ax3/amplA;
```

iii) Calculation of the frequency distribution associated with interval *B*, according to the algorithm presented in ii).

Once we have obtained the frequency distributions corresponding to the intervals *A* and *B*, the affinity coefficient is calculated as described in Nicolau and Bacelar-Nicolau (1999), considering that in the case of no intersection between the two intervals the value of the local affinity coefficient is 0. In the case of coincidence of the two intervals or the intersection being a single point, the value of the local affinity coefficient is 1.

Process 2: Another process consists of working directly with the intervals, in the following way: Let *a* and *b* be two unconstrained Boolean symbolic *objects* (i.e. two symbolic Boolean objects with no logical dependencies between the variables), defined in the following way:

$$a = [Y_1 \in A_1] \wedge [Y_2 \in A_2] \wedge \cdots \wedge [Y_p \in A_p];$$

$$b = [Y_1 \in B_1] \wedge [Y_2 \in B_2] \wedge \cdots \wedge [Y_p \in B_p],$$

where each variable Y_j takes values in the domain y_j, and A_j, B_j are subsets of y_j.

De Carvalho (1994, 1996, 1998a, 1998b) suggests the calculation of the comparison function of each variable, Y_j, on the basis of the agreement and disagreement indexes presented in Table 1, where we have defined for each subset $V_j \subseteq y_j$:

$$\mu(\mathrm{V_j})=\begin{cases}|\mathrm{V_j}|, & \text{that is the cardinal of } \mathrm{V_j}, \\ & \text{if } \mathrm{Y_j} \text{ is integer, nominal or ordinal} \\ |\bar{\mathrm{v}}_\mathrm{j}-\underline{\mathrm{v}}_\mathrm{j}|, & \text{that is the range of } \mathrm{V_j}, \\ & \text{if } \mathrm{Y_j} \text{ is continuous and } \mathrm{V_j}=[\underline{\mathrm{v}}_\mathrm{j},\bar{\mathrm{v}}_\mathrm{j}] \text{ an interval}\end{cases}$$

and c(V$_j$)= y_j-V$_j$ is the complementary set of V$_j$ in the domain y_j. In particular, in this work we assume that the values of each variable are intervals of the real axis.

Table 1. De Carvalho's agreement/disagreement indices

	Agreement	Disagreement	Total
Agreement	$\alpha=\mu(A_j\cap B_j)$	$\beta=\mu(A_j\cap c(B_j))$	$\mu(A_j)$
Disagreement	$\chi=\mu(c(A_j)\cap B_j)$	$\delta=\mu(c(A_j)\cap c(B_j))$	$\mu(c(A_j))$
Total	$\mu(B_j)$	$\mu(c(B_j))$	$\mu(\mathrm{y}_j)$

Table 2. Five possible comparison functions

s_i	Comparison function	Range	Property	*= 0 for*	*=1 for*
s_1	$\frac{\alpha}{\alpha+\beta+\chi}$	*[0, 1]*	Metric	$A_k\cap B_k=\phi$	$A_k=B_k$
s_2	$\frac{2\alpha}{2\alpha+\beta+\chi}$	*[0, 1]*	Semi metric	$A_k\cap B_k=\phi$	$A_k=B_k$
s_3	$\frac{\alpha}{\alpha+2(\beta+\chi)}$	*[0, 1]*	Metric	$A_k\cap B_k=\phi$	$A_k=B_k$
s_4	$\frac{1}{2}\left[\frac{\alpha}{\alpha+\beta}+\frac{\alpha}{\alpha+\chi}\right]$	*[0, 1]*	Semi metric	$A_k\cap B_k=\phi$	$A_k=B_k$
s_5	$\frac{\alpha}{\sqrt{(\alpha+\beta)(\alpha+\chi)}}$	*[0, 1]*	Semi metric	$A_k\cap B_k=\phi$	$A_k=B_k$

De Carvalho has proposed the comparison functions presented in Table 2, which are an extension of the similarity measures defined for classical binary variables. Note, in particular, that the function s_5 is the Ochiai coefficient (Anderberg, 1973), which coincides with the affinity coefficient in the case of binary variables (Nicolau and Bacelar-Nicolau, 1999; Bacelar-Nicolau, 2002).

From this point on, our approach differs from that of De Carvalho, essentially because instead of working with the corresponding dissimilarity function, d_i, for example, through the transformation: $d_i = 1 - s_5$, we work with the similarity function, in this case, the measure **s**$_5$. Thus, the similarity measure between the symbolic objects *a* and *b* is given by:

$$s(a,b) = \sum_{j=1}^{p} w_j s_5\left(A_j, B_j\right).$$

Let *E= {1, 2,..., N}* be a set of *N* data units described by a set $\{Y_1,..., Y_j,..., Y_p\}$ of *p* symbolic variables. The data units can be either simple elements (e.g. subjects, individuals) or subsets of objects in some population (e.g., subsamples of a sample, classes of a partition, subgroups of the population) (Bacelar-Nicolau, 2000).

Given a proximity matrix between symbolic objects, classifications of them can be obtained using classical agglomerative algorithms (Diday, 1988; Gowda and Diday, 1991, 1992), like Single Linkage (*SL)* and Complete Linkage (*CL).* In this approach, having obtained a proximity matrix between the elements of *E*, the classification is obtained without attention to the fact that the data are symbolic (Gordon, 1999). The clustering of symbolic data can be also based on probabilistic algorithms. In particular, the probabilistic approach of AHCA, called VL methodology (V for *Validity*, L for *Linkage*), can be used (Bacelar-Nicolau, 2002; Sousa, 2005). The α_R coefficient combined with the AVL, AV1, AVB and AVM methods (Bacelar-Nicolau, 1988; Nicolau, 1980; Nicolau and Bacelar-Nicolau, 1998) are good examples of this approach.

3. Example: “Abalone Data”

The data set analyzed is called “*Abalone data*” and contains 4177 cases of marine crustaceans, which are described by means of the nine attributes (Malerba et al., 2001) listed in Table 3. Initially, using the *DB2S0 facility available in the SODAS software,* nine boolean symbolic objects were generated, each of which corresponds to an interval of values for the number of rings of the crustaceans: *A (1-3), B (4-6), C (7-9), D (10-12), E (13-15), F (16-18), G (19-21), H (22-24), I (25-29).* For each of the groups we considered 7 symbolic variables *("Length", "Diam.", "Height", "Whole", "Shucked", "Viscera", "Shell")* of interval type (see Table 4).

Table 3. Attributes of the "Abalone" data set

Attribute Name	Data Type	Unit	Description
Sex	Nominal		M. F. I. (Infant)
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	Perpendicular to length
Height	Continuous	mm	Measured with meat in shell
Whole weight	Continuous	grams	Weight of the whole abalone
Shucked weight	Continuous	grams	Weight of the meat
Viscera weight	Continuous	grams	Gut weight after bleeding
Shell weight	Continuous	grams	Weight of the dried shell
Rings	Integer		Number of rings

Table 4. Symbolic data matrix

	Sex	*Length*	*Diam.*	*Height*	*Whole*	*Shucked*	*Viscera*	*Shell*
A	M(0.18), I(0.82)	[0.08: 0.24]	[0.05: 0.17]	[0.01: 0.06]	[0.00: 0.07]	[0.00: 0.03]	[0.00: 0.01]	[0.00: 0.02]
B	M(0.10), F(0.05), I(0.85)	[0.13: 0.66]	[0.09: 0.47]	[0.00: 0.18]	[0.01: 1.37]	[0.00: 0.64]	[0.00: 0.29]	[0.00: 0.35]
C	M(0.32), F(0.25), I(0.43)	[0.20: 0.75]	[0.16: 0.58]	[0.00: 1.13]	[0.04: 2.33]	[0.02: 1.25]	[0.01: 0.54]	[0.02: 0.56]
D	M(0.46), F(0.41), I(0.13)	[0.29: 0.78]	[0.22: 0.63]	[0.06: 0.51]	[0.12: 2.78]	[0.04: 1.49]	[0.02: 0.76]	[0.04: 0.73]
E	M(0.46), F(0.43), I(0.11)	[0.32: 0.81]	[0.25: 0.65]	[0.08: 0.25]	[0.16: 2.55]	[0.06: 1.35]	[0.03: 0.57]	[0.05: 0.80]
F	M(0.44), F(0.45), I(0.11)	[0.40: 0.77]	[0.31: 0.60]	[0.10: 0.24]	[0.35: 2.83]	[0.11: 1.15]	[0.06: 0.48]	[0.12: 1.00]
G	M(0.46), F(0.47), I(0.07)	[0.45: 0.74]	[0.35: 0.59]	[0.12: 0.23]	[0.41: 2.13]	[0.11: 0.87]	[0.07: 0.49]	[0.16: 0.85]
H	M(0.41), F(0.59)	[0.45: 0.80]	[0.38: 0.63]	[0.14: 0.22]	[0.64: 2.53]	[0.16: 0.93]	[0.11: 0.59]	[0.24: 0.71]
I	M(0.40), F(0.60)	[0.55: 0.70]	[0.47: 0.58]	[0.18: 0.22]	[1.06: 2.18]	[0.32: 0.75]	[0.19: 0.39]	[0.38: 0.88]

Two abalones with the same number of rings should also present similar values for the attributes listed in Table 3. On the basis of this assumption, we hope that "*the degree of dissimilarity between crustaceans computed on the independent attributes will actually be proportional to the dissimilarity in the dependent attribute (i.e., difference in the number of rings*)" (Malerba et al., 2001). This property is called "*monotonic increasing dissimilarity*" (shortly, *MID property*).

After determining the frequency distributions corresponding to the intervals of the initial matrix of data, according to process 1, we used the weighted generalised affinity coefficient (Nicolau and Bacelar-Nicolau, 1999; Bacelar-Nicolau, 2002) with $p_{jj'}=1/p$ *if* $j=j'$ *and* $p_{jj'}=0$ *if* $j\neq j'$(Nicolau and Bacelar-Nicolau, 1999). This measure of comparison between elements has been combined with classical, SL and CL, and probabilistic aggregation criteria, AVL, AV1 and AVB (Bacelar-Nicolau, 1988; Nicolau, 1980; Nicolau and Bacelar-Nicolau, 1998).

Figure 2 contains the graph of the values of the indexes *STAT*, *DIF*, *P(I2mod,∑), GK and GKr (Sousa, 2005)* for the partitions provided by the *AV1* method. Figure 1 presents the corresponding dendrogram.

```
                         levels    1  to  8
A                       --*--------------------*
                                               |
B                       --*-----------------*  |
                                            |  |
C                       --*--------*        |  |
                                   |--*     |  |
D                       --*--*     |  |     |  |
                             |-----*  |     |--*
E                       --*--*        |--*  |
                                      |  |  |
F                       --*           |  |  |
                          |-----*     |  |  |
G                       --*     |-----*  |--*
                                |        |
H                       --*-----*        |
                                         |
I                       --*--------------*
```

Figure 1. Dendrogram obtained with AV1

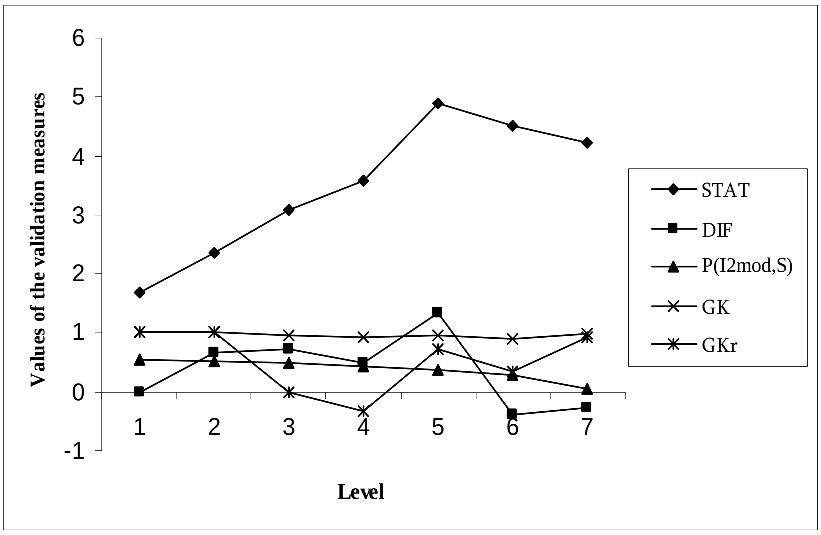

Figure 2. Values of some indexes of validation - AV1

Based on the *GK*, *GKr* and *P(I2mod)* indexes, the most significant partition is the partition into two clusters: {*A*}; {*B*, *C*, *F*, *G*, *E*, *H*, *D*, *I*}, one of them containing the youngest crustaceans and the other the remaining crustaceans. On the other hand, the *STAT and* DIF indexes point to a partition into four clusters: {*A*}; {*B*}; {*F*, *G*, *E*, *H*, *D*, *C*}; {*I*}, where the groups A and B, of younger crustaceans, and the group I, of older crustaceans, remain isolated (see Figures 1 and 2). After that, the obtained results were compared with those obtained using process 2, and the conclusions were identical.

4. Conclusion

The example presented allowed us to illustrate the application of the weighted generalised affinity coefficient and the extension of *VL* methodology to classification of this type of data.The weighted generalised affinity coefficient for the case of symbolic data and the *VL* methodology was able to reproduce well the properties of the symbolic data analysed.

The weighted generalised affinity coefficient is an appropriate resemblance measure between elements when there is some degree of overlapping between intervals. The only limitation of this coefficient occurs in the case where a great number of symbolic objects are described by variables of interval type in which intervals that are compared either do not intersect or have intersection equal to a single value, causing many values of the local affinity coefficient to be equal to zero.

The validation measures used also proved useful in determination of the appropriate number of clusters.

Acknowledgments

This research was partially supported by FCT/POCTI/ and POCTI/FEDER, within the scope of the CEAUL Research Project on Applied Multivariate Data Analysis and Modelling.

References

Anderberg M.R. (1973): Cluster Analysis for Applications. Academic Press, New York.

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: 181–186.

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

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

Bock H.H.; Diday E. (Eds.) (2000): Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer.

De Carvalho F.A.T. (1994): Proximity coefficients between Boolean Symbolic Objects. In: New Approaches in Classification and Data Analysis, Diday, E. et al. (Eds.), Springer, Berlin: 387–394.

De Carvalho F.A.T. (1996): Histogrammes et Indices de Proximité en Analyse de Données Symboliques. Actes de l'École d'Été sur l'Analyse des Données Symboliques. LISE-CEREMADE, Université de Paris IX-Dauphine, Paris: 101-127.

De Carvalho F.A.T. (1998a): Extension based Proximities between Constrained Boolean Symbolic Objects. In: C. Hayashi et al. (Eds.): Data Science, Classification and Related Methods, Proc. of IFCS-96, Springer, Berlin: 370–378.

De Carvalho F.A.T. (1998b): New Metrics for Constrained Boolean Symbolic Objects. Proc. KESDA'98. Eurostat, Luxembourg.

Diday E. (1988): The Symbolic Approach in Clustering and Related Methods of Data Analysis: The Basic Choices. In: H.H. Bock (Ed.): Classification and Related Methods of Data Analysis. Proc. IFCS-87, North Holland, Amsterdam: 673–684.

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

Gowda K.C., Diday E. (1991): Symbolic Clustering Using a New Dissimilarity Measure. In: Pattern Recognition 24(6): 567–578.

Gowda K.C., Diday E. (1992): Symbolic Clustering Using a New Similarity Measure. IEEE Transactions on Systems, Man and Cybernetics 22(2): 368–378.

Gower J.C. (1971): A General Coefficient of Similarity and Some of its Properties, BioMetrics 27: 857–874.

Malerba D., Esposito F., Gioviale V., Tamma V. (2001): Comparing Dissimilarity Measures in Symbolic Data Analysis. Proceedings of the Joint Conferences on "New Techniques and Technologies for Statistics" and "Exchange of Technology and Know-how" (ETK-NTTS'01): 473–481.

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

Nicolau F., 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.

Nicolau F., Bacelar-Nicolau H. (1999): 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, Lisbon, Portugal: 155–158.

Sousa A. (2005): Contribuições à Metodologia VL e Índices de Validação para Dados de Natureza Complexa. Tese de Doutoramento, Universidade dos Açores, Ponta Delgada.