Biocybernetics and Biomedical Engineering
2009, Volume 29, Number 2, pp. 9–18

Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets

HELENA BACELAR-NICOLAU[1,*], FERNANDO NICOLAU[2], ÁUREA SOUSA[3], LEONOR BACELAR-NICOLAU[4]

[1] *Universidade de Lisboa, FPCE, Laboratório de Estatística e Análise de Dados, Lisboa, Portugal*
[2] *Universidade Nova de Lisboa, FCT, Departamento de Matemática, Caparica, Portugal*
[3] *Universidade dos Açores, Departamento de Matemática, Ponta Delgada, Açores, Portugal*
[4] *Universidade de Lisboa, FML, Laboratório de Biomatemática & Instituto de Medicina Preventiva, Lisboa, Portugal*

Cluster analysis or classification usually concerns a set of exploratory multivariate data analysis methods and techniques for finding a clustering structure on a dataset. That may refer either to groups of statistical data units or to groups of variables. In this work we deal with a generalization of this paradigm concerning clustering of complex data described by three different types of variables, frequently present in a three-way context. We obtain compatible versions of the same affinity coefficient for measuring similarity between statistical data units described by those three types of variables. A global generalized similarity coefficient is analyzed for such kind of mixed data, often arising in data mining or knowledge mining.

K e y w o r d s: cluster analysis, different type variables, similarity coefficient, three-way data

1. Introduction

Cluster analysis or classification usually concerns a set of exploratory multivariate data analysis methods and techniques for finding a cluster structure on a dataset. This may refer to grouping either statistical data units or variables. Traditional clustering methods usually work with a set of subjects as statistical data units described by a set

* Correspondence to: Helena Bacelar-Nicolau, Laboratory of Statistics and Data Analysis, Faculty of Psychology and Education, University of Lisbon, Alameda da Universidade 1649-013 Lisboa, Portugal, email: hbacelar@fpce.ul.pt
Received 10 October 2008; Accepted 05 January 2009

of homogeneous (that is, of the same type) variables. In previous work [1, 2, 3] we have extended this situation to the case where data units represent some kind of data sets (data units of second order or more) and variables may be of different types. In fact we may refer to three steps of generalization of the traditional paradigm that appear to be particularly useful (but not only) when large data bases are used, for instance in a data mining extended context: 1 – classification/clustering of complex or "three-way" data instead of the most common "two-way" data approach; 2 – clustering models where prior knowledge of data structure allows to some probabilistic framework as a tool of extracting (new) knowledge from clustering structures; 3 – comparison and clustering of hierarchical clustering models, based on the affinity of associated parameter profiles, inside an adaptive family of aggregation criteria.

The first two steps allow us to mine knowledge from large data bases, by building clustering models to explain data structure. The third one allows us to go into knowledge extraction from the models (somehow extending data mining to knowledge mining).

The present work has mainly to do with the above first two steps of generalization, when clustering of three-way data units described by heterogeneous variables is concerned. In fact, in large data bases, we are very often confronted with (large) matrices where data units are described by a heterogeneous set of variables. Therefore the question arises of how we should measure the similarity between statistical data units in a coherent way, if different types of variables are involved. Traditionally partial similarity coefficients for each type of variables are computed, and then a convex linear combination of those similarities gives a global similarity between data units. Such procedure must be performed in a consistent way, combining comparable similarity coefficients in a valid global similarity. In two-way data matrices a well known coefficient for comparing subjects described by different types of variables was proposed in 1971 by Gower.

So far we have been using the affinity coefficient for that purpose, either in two-way or in three-way data cases. Here we generalize that procedure to cases where three types of heterogeneous variables often arising in a three-way context are concerned. We use the same way of measuring similarity for those three types of variables, based on comparable versions of the affinity coefficient. Then a consistent weighted linear combination of the partial or local affinities provides a global generalized affinity coefficient between statistical data units.

The data units can be either simple elements (e.g., subjects, individuals) or groups of objects in some population (e.g., subsamples of a sample, classes of a partition, subsets of the population).

We may assume that statistical data units refer to rows of a generalized table, while variables refer to generalized columns or sub-tables, where each column/sub-table may have a different number of "modalities". Therefore in this three-way generalized data table each cell (crossing each data unit with each variable) may contain instead of one, a set of different values – for instance a frequency distribution (histogram), a binary vector or an interval – depending on the variable type.

A methodology to find clustering models or adaptive families of clustering models, based on some (successive) generalizations of the so-called affinity coefficient, has been developed for such kind of data (e.g. [1, 2], [4–7]), in the scope of several national and European research projects on multivariate data analysis and modelling.

The next section presents an overview of the three-way clustering approach based on the weighted generalized affinity coefficient of statistical data units. In Section 3 we extend this coefficient to three way data when mixed type of variables appear, namely in case of histogram, binary and interval-type variables. Section 4 illustrates how the extended coefficient works in a small example issued from the literature of symbolic data analysis [8, 9]. Section 5 presents some conclusions and future developments. So far applications in real data were made in biomedicine, education and marketing.

2. Weighted Generalized Affinity and Asymptotic Permutational Standardized Coefficients

Let D be a set of statistical data units and let V be a set of p variables, as depicted in the previous section. Here we will be concerned with clustering models on the set of data units.

The weighted generalized affinity coefficient $a(k, k')$ between a pair of data units $k, k' \in D$ $(k, k' = 1,...,n)$, may be defined in a three-way context, as the weighted mean of local affinities between k and k' over the *j-th* variable ($j = 1,\ldots, p$), as follows:

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

where: $aff(k, k';j)$ is the local affinity over the *j-th* variable, m_j represents the number of modalities in the *j-th* variable; $x_{kj\ell}$ is a real value whose meaning depends on the type of *j-th* variable (e.g. a discrete variable described by a frequency distribution or histogram, a binary vector or an interval variable) or equivalently on the nature of *j-th* corresponding sub-table; and π_j are weights such that $0 \le \pi_j \le 1$, $\Sigma \pi_j = 1$. Either the local affinities or the whole weighted generalized affinity coefficient, take values in the interval [0, 1] and satisfy the set of main proprieties of a similarity coefficient (e.g. [1, 5]). If *all* m_j are equal, say, to m, and all modalities are associated to the same experimental situation S, for instance space or time, then one gets a three dimensional real-valued matrix or table $X = D \times V \times S = \{x_{ij\ell},\ i = 1,\ldots,n;\ j = 1,...,p;\ \ell = 1,\ldots,m\}$. Note that in this case, a global affinity coefficient of two complete p-dimensional data units $k, k' \in D$, or p-multivariate affinity, might be defined in a similar way over the whole set of pairs (j, j') of variables. In the present work we are dealing with the more generalized table X as described above.

Let us suppose that some prior knowledge on the data base may be taken in account such as statistical reference hypothesis allowing us to compute (*asymptotic*) standardized affinity values and/or the corresponding (*asymptotic*) cumulative distribution function values. Then new similarity coefficients arise – and as a result new probabilistic clustering models (PCM), instead of empirical clustering models may be selected. Thus a reference hypothesis usually stands in such a probabilistic approach not only as a convenient reference point, but also by having a natural interpretation, depending on the type of data and context.

In a three-way clustering probabilistic analysis, a permutational reference hypothesis *R* based on a well known limit theorem of Wald and Wolfowitz (other reference hypothesis have been used based, for instance, on the limit theorem of delta-method), may be applied very often [5]. Then the random variable *aff* (*k, k′; j*) has asymptotic normal distribution, [2, 3, 5], whose asymptotic mean value and variance are as follows:

$$\mu^{*}_{WW}\left(k,k';j\right)=\frac{1}{m_j}\sum_{\ell=1}^{m_j}\sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}}}\quad\sum_{\ell'=1}^{m_j}\sqrt{\frac{x_{k'j\ell'}}{x_{k'j\bullet}}}$$

$$\sigma_{WW}{}^{*2}\left(k,k';j\right)=\frac{1}{m_j-1}\sum_{\ell=1}^{m_j}\left(\sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}}}-\frac{1}{m_j}\sum_{\ell'=1}^{m_j}\sqrt{\frac{x_{kj\ell'}}{x_{kj\bullet}}}\right)^2\times\sum_{\ell=1}^{m_j}\left(\sqrt{\frac{x_{k'j\ell}}{x_{k'j\bullet}}}-\frac{1}{m_j}\sum_{\ell'=1}^{m_j}\sqrt{\frac{x_{k'j\ell'}}{x_{k'j\bullet}}}\right)^2.$$

Therefore, if such reference hypothesis *R* holds, this leads us to a local asymptotic normal coefficient aff_{WW} (*k, k′; j*) that, it is easy to prove, also satisfies the main properties of a similarity coefficient. This coefficient applies in the present work. So, instead of using the basic generalized affinity coefficient $a(k, k')$ between data units $k, k' \in D$ ($k, k' = 1,\ldots,n$), we will use:

$$a_{WW}(k,k')=a^{*}(k,k')=\sum_{j=1}^{p}\pi_j.aff^{*}_{WW}(k,k';j),$$

where $aff^{*}_{WW}(k,k';j)=(aff(k,k';j)-\mu^{*}_{WW}(k,k';j))/\sigma^{*2}_{WW}(k,k';j)$.

The previous results can also bring us to a third coefficient related to affinity measurement, that is a probabilistic coefficient $\alpha_R(k, k')$ between two data units $k, k' \in D$ ($k, k' = 1,\ldots,n$), as follows:

$$a_R(k, k') = P_R(A^{*}(k, k') \le a^{*}(k, k')) \cong \hat{a}_R(k, k') = \Phi(a^{*}(k, k')),$$

where Φ denotes the cumulative distribution function of the standard normal distribution and A^{*} represents an asymptotic standardized random variable. A large value

of the probabilistic coefficient means that the "observed" affinity value is "significantly" larger than one might expected, under the above reference hypothesis. Thus the probabilistic coefficient validates the affinity coefficient between two data units k, k' in a probabilistic scale (e.g. [3, 5], [10, 12]). This led $\alpha_R(k, k')$ to be sometimes roughly called "Validity Link" or VL affinity coefficient. Again it is a similarity coefficient, which takes values in the interval [0, 1]. In the present work we will not use this probabilistic coefficient yet.

3. Clustering of Three-way Data Units Described by Heterogeneous and Complex Variables

Let T be a data matrix for n statistical data units (usually *groups* of individuals) and V a set of p variables of different types. We assume that Y_j is a *discrete* or a *categorical (modal)* variable with m_j ($\ell = 1,...,m_j$) modalities, variable $Y_{j'}$ is a $m_{j'}$ – dimensional binary vector and variable $Y_{j''}$ is an interval variable, where j, j' and j'' belong to $\{1,\ldots,p\}$ [1–3, 7]. Thus the corresponding sub-tables have n rows, and the k-*th* row ($k = 1,\ldots,n$) in each sub-table contains: for Y_j, a frequency distribution ($n_{kj1},...,n_{kjm_j}$), where $n_{kj\ell}$ is the number of individuals in the k-*th* data unit k who share the ℓ-*th* category of the j-*th* variable; for $Y_{j'}$, an *element* $\{0,1\}_k^{m_{j'}}$ *of the* power set $\{0, 1\}^{m_{j'}}$, the whole binary sub-table being an element of $\{0, 1\}^{n\times m_{j'}}$; for $Y_{j''}$, an interval $I_{kj''}$ of the real axis. Thus the dataset may be represented by the following generalized table T:

Table 1. Generalized Data Matrix

$D\backslash V$	...	Y_j	...	$Y_{j'}$	...	$Y_{j''}$	...
$\vdots$	…	...	…		…	…	…
k	…	$(n_{kj1},...,n_{kjm_j})$	…	$\{0,1\}_k^{mj'}$	…	$I_{kj''}$	…
$\vdots$	…	...	…	…	…	…	…
k'	…	$(n_{k'j1},...,n_{k'jm_j})$	…	$\{0,1\}_{k'}^{mj'}$	…	$I_{k'j''}$	…
$\vdots$	…	...	…	…	…	…	…

The generalized local affinity coefficient formula given in the previous section applies for each of the three types of variables and we have:

Discrete and categorical variable: It is easy to see that all the formulae and results mentioned above hold for Y_j variable, replacing $x_{kj\ell}$ ($x_{k'j\ell}$) by the frequency $n_{kj\ell}$ ($n_{k'j\ell}$) of ℓ-*th* category or modality ($\ell = 1,...,m_j$). Hence the local affinity $aff(k, k'; j)$ over the j-*th* variable measures the similarity between the two profile vectors (histograms) associated to the pair of data units $k, k' \in D$ ($k, k' = 1,\ldots,n$) on the j-*th* sub-table.

Binary vector: Let us take now variable $Y_{j'}$. The local affinity *aff* (*k, k′; j′*) may be computed from the 2×2 contingency table associated to the pair (*k, k′*) of rows in the *j′-th* binary sub-table:

Table 2. Table of agreements and disagreements for a binary vector

$k \setminus k'$	Agreement (1)	Disagreement (0)	Total
1	$s_{j'} = \sum_{\ell=1}^{m_{j'}} x_{kj'\ell} x_{k'j'\ell}$	$u_{j'} = \sum_{\ell=1}^{m_{j'}} x_{kj'\ell}(1 - x_{k'j'\ell})$	$s_{j'} + u_{j'} = m_{kj'}$
0	$v_{j'} = \sum_{\ell=1}^{m_{j'}} (1 - x_{kj'\ell}) x_{k'j'\ell}$	$t_{j'} = \sum_{\ell=1}^{m_{j'}} (1 - x_{kj'\ell})(1 - x_{k'j'\ell})$	$v_{j'} + t_{j'} = m_{j'} - m_{kj'}$
Total	$s_{j'} + v_{j'} = m_{k'j'}$	$u_{j'} + t_{j'} = m_{j'} - m_{k'j'}$	$m_{j'}$

where: $s_{j'}$ is the cardinal of positive agreements ($x_{kj'\ell} = x_{k'j'\ell} = 1$); $t_{j'}$ is the cardinal of negative agreements ($x_{kj'\ell} = x_{k'j'\ell} = 0$); $u_{j'}$ and $v_{j'}$ are the cardinals of disagreements (respectively $x_{kj'\ell} = 1$, $x_{k'j'\ell} = 0$ and $x_{kj'\ell} = 0$, $x_{k'j'\ell} = 1$). Therefore the local affinity is defined as in formula (1) above and it gives:

$$aff\left(k, k'; j'\right) = \frac{s_{j'}}{\sqrt{m_{kj'} m_{k'j'}}}, \tag{2}$$

that is the well known Ochiai coefficient for binary data.

Interval-type variable: Let $Y_{j''}$ be an interval variable, associated to a generalized column j'', where each cell *(k,j'')* contains an interval $I_{kj''}$, (k =1,…,n).
Let $I_{j''}$ be the union of the intervals $I_{kj''}$: $I_{j''} = \cup I_{kj''}$ (k =1,…,n).
Let $\{I_{j''\ell} : \ell = 1,\ldots,m_{j''}\}$ be a set of $m_{j''}$ elementary intervals, such that the following properties hold, for $\ell,\ell' = 1,\ldots,m_{j''}$, $\ell \neq \ell'$; $k = 1,...,n$:

1. $I_{j''} = \cup I_{j''\ell}$;
2. $\left|I_{j''\ell} \cap I_{j''\ell'}\right| = 0$;
3. $\left|I_{kj''} \cap I_{j''\ell}\right| = \left|I_{j''\ell}\right|$, if $\left|I_{kj''} \cap I_{j''\ell}\right| \neq 0$; $\left|I_{kj''} \cap I_{j''\ell}\right| = 0$, otherwise;

where $|\ |$ represents the interval range.

Let $x_{kj''\ell}$ be $x_{kj''\ell} = \left|I_{kj''} \cap I_{j''\ell}\right|$.

Then $x_{kj''\ell} = \left|I_{j''\ell}\right|$ if $I_{kj''} \cap I_{j''\ell} = I_{j''\ell}$; $x_{kj''\ell} = 0$, otherwise.

Therefore we also have:

$x_{kj''\bullet} = \left|I_{kj''}\right|$, $x_{k'j''\bullet} = \left|I_{k'j''}\right|$ and $\sum_{\ell=1}^{m_{j''}} \sqrt{x_{kj''\ell} x_{k'j''\ell}} = \left|I_{kj''} \cap I_{k'j''}\right|$.

Hence the local affinity $aff(I_{kj''}, I_{k'j''}) = aff(k, k'; j'')$ is also defined as in formula (1) above and we have, for k =1,…,n; ℓ = 1,...,$m_{j''}$:

$$aff\left(I_{kj''}, I_{k'j''}\right) = \frac{\left|I_{kj''} \cap I_{k'j''}\right|}{\sqrt{\left|I_{kj''}\right| \times \left|I_{k'j''}\right|}} \quad . \tag{3}$$

Consequently the local affinity *aff* (*k, k'; j''*) is a generalized Ochiai coefficient, which may be computed from the generalized 2×2 contingency Table 3, associated to the pair of intervals over the *j''-th* generalized column/sub-table of the table *T*:

Table 3. Table of agreements and disagreements for an interval-variable

$k \setminus k'$	Agreement	Disagreement	Total
Agreement	$s_{j''} = \left\|I_{kj''} \cap I_{k'j''}\right\|$	$u_{j''} = \left\|I_{kj''} \cap I^c_{k'j''}\right\|$	$s_{j''} + u_{j''} = \left\|I_{kj''}\right\|$
Disagreement	$v_{j''} = \left\|I^c_{kj''} \cap I_{k'j''}\right\|$	$t_{j''} = \left\|I^c_{kj''} \cap I^c_{k'j''}\right\|$	$v_{j''} + t_{j''} = \left\|I^c_{kj''}\right\|$
Total	$s_{j''} + v_{j''} = \left\|I_{k'j''}\right\|$	$u_{j''} + t_{j''} = \left\|I^c_{k'j''}\right\|$	$\left\|I_{j''}\right\|$

Here $I^c_{kj''}$ represents the complementary set of $I_{kj''}$ in the domain $I_{j''}$.

Therefore the generalized local affinity coefficient given in the previous section in (1) applies for each of the three types of variables. In case of a binary vector or an interval variable, we obtain the Ochiai coefficient and the generalized Ochiai coefficient, respectively. Moreover the local standardized affinity coefficients may be computed in the same way, for each one of the three types of variables. Then a weighted generalized affinity coefficient $a(k, k')$ between a pair of data units $k, k' \in D$ ($k, k' = 1,\ldots,n$) can also apply and consequently both asymptotic normalized and probabilistic associated coefficients hold as well.

4. Example/Case Study

Since for discrete and categorical variables the formula (1) holds just replacing real values by frequencies the most interesting cases appear to be those where binary and interval variables are simultaneously concerned. Here we use a small example satisfying that requirement.

Table 4 illustrates the data set which is composed of 8 oils and fats (*1–Linseed oil* (*LS*), *2–Perilla oil* (*P*), *3–Cotton seed* (*CS*), *4–Sesame oil* (*S*), *5–Camellia* (*C*), *6–Olive oil* (*O*), *7–Beef Tallow* (*T*), *8–Lard* (*L*)) described in terms of four interval variables and one nominal qualitative feature [8, 9].

The hierarchical clustering agglomerative models for the *eight* complex/symbolic data units (8 oils and fats) were based on the weighted generalized affinity coefficient with equal weights, $\pi_j = 1/p$.

Table 4. Data Matrix (Fats and Oils)

Sample name	Specific_gravity (g/cm^3)	Freezing point (ºC)	Iodine value	Saponification value	Major Fatty Acids
LS	[0.930, 0.935]	[–27, –8]	[170, 204]	[118, 196]	L, Ln, O, P, M
P	[0.930, 0.937]	[– 5, –4]	[192, 208]	[188, 197]	L, Ln, O, P, S
CS	[0.916, 0.918]	[–6, –1]	[99, 113]	[189, 198]	L, O, P, M, S
S	[0.920, 0.926]	[–6, –4]	[104, 116]	[187, 193]	L, O, P, S, A
C	[0.916, 0.917]	[–21, –15]	[80, 82]	[189, 193]	L, O
O	[0.914, 0.919]	[0, 6]	[79, 90]	[187, 196]	L, O, P, S
T	[0.860, 0.870]	[30, 38]	[40, 48]	[190, 199]	O, P, M, S, C
L	[0.858, 0.864]	[22, 32]	[53, 77]	[190, 202]	L, O, P, M, S, Lu

L: Linoleic acid *Ln: Linolenic acid* *O: Oleic acid* *P: Palmitic acid* *M: Myristic acid*
S: Searic acid *A: Arachic acid* *C: Capric acid* *Lu: Lauric acid*

Notice that in order to compute for instance the local affinities between all the sample pairs over the first (generalized) column/interval variable, *Specific_gravity* (g/cm^3), a sub-table with 13 columns corresponding to a set of elementary intervals was computed. Each column of this sub-table contains the ranges of the intersection intervals between each elementary interval and each one of the 8 intervals described in the first (generalized) column above.

The last generalized column of Table 4, representing the *Major Fatty Acids*, may be written as a binary sub-table with nine columns.

Table 5 contains the values of the resulting similarity matrix. Two classical aggregation criteria – single linkage and complete linkage – and three probabilistic aggregation criteria from an adaptive (parametric) family [3, 6, 11] were used – respectively the Algorithms AVL, AVB and AVM. The results were very similar, showing a strong data structure. Figure 1 represents the dendrogram obtained by the AVB method.

Table 5. Similarity Matrix (Fats and Oils): weighted generalized affinity coefficient with equal weights

Sample name	LS	P	CS	S	C	O	T	L
LS	1.000000							
P	0.492318	1.000000						
CS	0.212840	0.427221	1.000000					
S	0.175470	0.437504	0.534230	1.000000				
C	0.284173	0.259825	0.401246	0.289791	1.000000			
O	0.202101	0.356663	0.460931	0.342185	0.449477	1.000000		
T	0.165291	0.275556	0.337778	0.201650	0.163246	0.267497	1.000000	
L	0.185283	0.280774	0.336534	0.216770	0.202073	0.278769	0.467265	1.000000

The dendrogram of Fig. 1 is a good illustration for chemical properties of the given fats and oils. It is known that both elements of each one of the sample pairs (LS, P), (CS, S), (C, O) have similar properties: LS and P are used for painting; CS and S for food; and C and O for cosmetics. On the other hand the pair (T, L) has animal origin. In addition Cluster {CS, S, C, O} has been found in other statistical approaches as well, particularly in [8, 9].

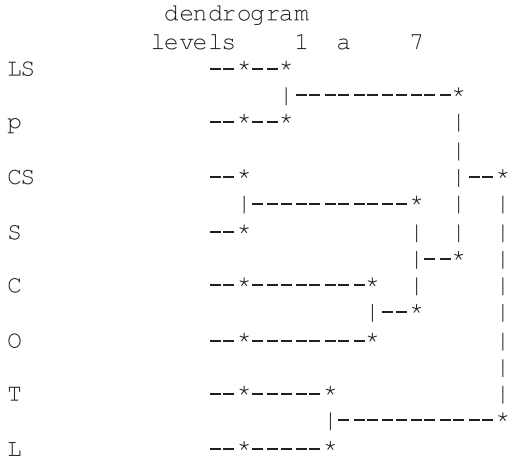

Fig. 1. AVB Dendrogram

5. Conclusion. Future Developments

The weighted generalized affinity coefficient $a(k,k')$ defined as above supports in a consistent way cluster analysis models for statistical data units, when mixed and complex variable types are present in a database. Besides, the asymptotic generalized coefficient $a_{WW}(k,k')$ is often applied instead of $a(k,k')$. Indeed, if V is a set of p independent heterogeneous variables, using $a_{WW}(k,k')$ instead of $a(k,k')$ means doing local standardization accordingly to the different variable types, which in this way all become asymptotic standard normal variables as well as their convex linear combination, $a_W(k,k')$. Furthermore, a probabilistic coefficient of VL kind may be applied in this context. Then empirical or semi-probabilistic clustering models can be built up over the set of statistical data units. Applications are being developed in health sciences, education and management (e.g. [2, 3, 7, 13]).

Acknowledgments

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

References

1. Bock H.H., Diday E. [Eds.]: Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data, Springer, 2000.
2. Bacelar-Nicolau H.: On the generalized affinity coefficient for complex data. Biocybernetics and Biomedical Engineering 2002, 22, 1, 31–42.
3. Nicolau F. C., Bacelar-Nicolau H. et al: Probabilistic models in three way cluster analysis. In: Proceedings of the 56th Session of the International Statistical Institute, Lisbon 2007 (in press), published on the CD Proceedings of ISI 2007.
4. Matusita K.: On the theory of statistical decision functions. Ann. Instit. Stat. Math. 1951, III, 1–30.
5. Bacelar-Nicolau H.: Two probabilistic models for classification of variables in frequency tables – Classification and Related Methods of Data Analysis. In: H. H. Bock [Ed.], Elsevier Sciences Publishers B.V., North Holland, 1988, 181–186.
6. Nicolau F. C., Bacelar-Nicolau H.: Some trends in the classification of variables. In: Hayashi, et al. [eds.], Data Science, Classification and Related Methods, Springer, 1998, 89–98.
7. Sousa A.: Contribuições à Metodologia VL e índices de validação para Dados de Natureza Complexa. PhD Thesis, Univ. Azores 2005.
8. Ichino M.: General metrics for mixed features – The Cartesian Space Theory for Pattern Recognition. IEEE Transactions on Systems, Man and Cybernetics 1988.
9. Ichino M., Yaguchi H.: Generalized Minkowski Metrics for Mixed Feature Type Data Analysis. IEEE Transactions on Systems, Man and Cybernetics, 1994, 24, 4, 698–708.
10. Bacelar-Nicolau H.: On the distribution equivalence in cluster analysis. Proceedings of the NATO ASI on Pattern Recognition Theory and Applications, Springer-Verlag, New York 1987, 73–79.
11. Lerman I.C.: Étude distributionelle de statistiques de proximité entre structures algébriques finies du même type – Application à la Classification Automatique. Cahiers du B.U.R.O., Paris 1972, 19.
12. Lerman I. C.: Classification et Analyse Ordinale des Données. Dunod, Paris 1981.
13. Bacelar-Nicolau L.: Caracterização dos Sistemas de Informação das Organizações com base no modelo de Nolan. Aplicação de modelos de classificação hierárquica aos organismos da Administração Pública. Master Thesis, Univ. Nova de Lisboa 2002.