Revista Colombiana de Estadística
ISSN: 0120-1751
revcoles_fcbog@unal.edu.co
Universidad Nacional de Colombia
Colombia

Zamora-Matamoros, Larisa; Díaz-Silvera, Jorge Rey; Portuondo-Mallet, Lariza
Fundamental Concepts on Classification and Statistical Implicative Analysis for Modal Variables
Revista Colombiana de Estadística, vol. 38, núm. 2, julio, 2015, pp. 335-351
Universidad Nacional de Colombia
Bogotá, Colombia

Available in: http://www.redalyc.org/articulo.oa?id=89940050002

How to cite
Complete issue
More information about this article
Journal's homepage in redalyc.org

Scientific Information System
Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal
Non-profit academic project, developed under the open access initiative

Revista Colombiana de Estadística
July 2015, Volume 38, Issue 2, pp. 335 to 351
DOI: http://dx.doi.org/10.15446/rce.v38n2.51664

Fundamental Concepts on Classification and Statistical Implicative Analysis for Modal Variables

Conceptos fundamentales de la clasificación y del análisis estadístico implicativo para variables modales

Larisa Zamora-Matamoros[a], Jorge Rey Díaz-Silvera[b], Lariza Portuondo-Mallet[c]

Department of Mathematics, Faculty of Mathematics and Computer Science, University of Oriente, Santiago de Cuba, Cuba

Abstract

The present work offers some fundamental concepts of the Statistical Implicative Analysis for modal variables and proposes an index to establish the similarity between two modal variables. Expressions to calculate typicality and contribution of the individuals to the classes that are formed in the classification are also presented. This technique is illustrated by two examples, one with binary data, which allows the coincidence between the formulas presented and the existent for the binary case to be shown, and the other for modal data with more than two modalities.

***Key words*:** Classificatory Analysis, Typicality, Contribution, Modal Variables.

Resumen

En el presente trabajo se ofrecen algunos conceptos fundamentales del análisis estadístico implicativo para el caso de variables modales y se propone un índice para establecer la similaridad entre dos variables modales, así como expresiones para el cálculo de la tipicalidad y contribución de los individuos a las clases que se forman en la clasificación. Con el objetivo de ilustrar la técnica presentada, es aplicada a dos juegos de datos, uno binario, el cual permite mostrar numéricamente la coincidencia de las fórmulas presentadas con las existentes para el caso de variables binarias, y otro modal con más de dos modalidades.

***Palabras clave*:** análisis clasificatorio, tipicalidad, contribución, variables modales.

[a]Professor. E-mail: larisa@csd.uo.edu.cu
[b]Professor. E-mail: jdiaz@csd.uo.edu.cu
[c]Professor. E-mail: lportuondo@csd.uo.edu.cu

1. Introduction

Régis Gras and its collaborators have developed the Statistical Implicative Analysis (SIA), which allows association rules in a dataset to be established, crossing individuals and variables.

The initial objective of this method was to answer the question: if an object has a certain quality, could it have some other one? This is the case of binary variables (Gras 2000).

The research developed by Marc Bailleul from 1991 to 1994, (Bailleul & Gras 1994) showed the necessity to extend the concept of statistical implication to non-binary variables, in particular to the modal variables.

Bailleul & Gras (1994) refer to the representation that the professors of mathematics make of their own teaching, for which, the professors have been asked to order a list of significant words depending on their importance. Their elections are no longer binary. The words selected by a professor are arranged in a scale where the top words are the most representative. The question of Bailleul is centered on questions of the type: "if I select a word x with the importance i_x then I select the word y with the importance $i_y \geq i_x$".

For the application of the SIA techniques, CHIC (Classification Hiérarchique Implicative et Cohésitive), (Couturier 2008) has been developed. This is a software of analysis of data that was initially conceived by Regis Gras and it has been successively improved for personal computers by Saddo Ag Almouloud, Harrisson Ratsimba-Rajohn and, in its latest versions, by Raphaël Couturier (Version 5.0), who incorporates the classification, by means of the similarity analysis.

Unlike the techniques of SIA, which are concerned with the cohesive and the implicative analysis, in the specialized literature, expressions for the determination of the similarity index, the significant nodes, the typicality and the contribution of individuals to the classes in the case of modal variables do not appear. For this reason, the present paper proposes a similarity index for the case of modal variables, as well as expressions for the determination of the typicality and contribution of the individuals to the classes formed during the construction of the similarity tree, extending the works of Bailleul & Gras (1994) and Lagrange (1998).

2. Statistical Implicative Analysis for Modal Variables

In this section, a brief summary of the theory developed by Bailleul & Gras (1994) and Lagrange (1998) for the case of modal variables will be presented. It will be the guideline to define a similarity index for these variables.

2.1. The Intensity of Propensity

Lagrange (1998) defines the "propensity" as a non-symmetric association of two variables whose values belong to the interval [0, 1] and introduces the "intensity of propensity" to decide whether or not an observed association is significant. He also proves that in the case of binary variables, the intensity of propensity is similar to the implication intensity.

Similar to the case of binary variables, let us consider a set I formed by n individuals and a set $A = \{a_1, a_2, \ldots, a_p\}$ formed by p modal variables, each one of which can take values in the set $M = \{k \in N : 1 \leq k \leq m\}$.

To transform the m modalities to values in the [0, 1] interval, a weight is assigned to each modality, varying from 1 (the strongest) to 0 by fractions of size $\frac{1}{m-1}$, by means of the function ψ defined by:

$$\begin{aligned} \Psi: \quad & \mathrm{M} \longrightarrow \mathrm{VAL} \\ & \mathrm{k} \longrightarrow \frac{k-1}{m-1} \end{aligned} \tag{1}$$

where $VAL = \left\{0, \frac{1}{m-1}, \frac{2}{m-1}, \ldots, \frac{m-2}{m-1}, 1\right\}$, conserving the notation used by Bailleul & Gras (1994).

Let a and b be two modal variables, Lagrange (1998) defines the non propensity index of a on b as:

$$S = \sum_{x \in I} \frac{\psi_a(x)\,(1 - \psi_b(x))}{n} \tag{2}$$

and to define the intensity of propensity, he considers a set E of size N sufficiently large, of which I is a random part. The random index is defined as:

$$Z = \sum_{x \in I} \frac{A_x\,(1 - B_x)}{n} \tag{3}$$

where $\{A_x\}_{x \in E}$ and $\{B_x\}_{x \in E}$ are sets of random variables with equal distributions, representative of the variables ψ_a and ψ_b respectively, for which the observed non propensity index s is one of its possible values. It is also proved that Z follows approximately a Normal distribution of mean $\mu = E\left[A_x\,(1 - B_x)\right]$ and variance $\sigma^2 = \frac{E\left[A_x\,(1 - B_x)\right]^2}{n}$, and therefore, for sufficiently large n, $\frac{Z - \mu}{\sigma}$ follows approximately a Normal distribution of mean zero and variance one.

Lagrange (1998) defines the propensity coefficient and the intensity of propensity of a on b as follows:

- Propensity coefficient:

$$\widetilde{q}\left(a,\bar{b}\right)=\frac{\sum_{x\in I}\frac{\psi_a\left(x\right)\left(1-\psi_b\left(x\right)\right)}{n}-m_a\left(1-m_b\right)}{\sqrt{\frac{\left(s_a^2+m_a^2\right)\left(s_b^2+\left(1-m_b\right)^2\right)}{n}}} \tag{4}$$

- Intensity of propensity:

$$\phi\left(-\widetilde{q}\left(a,\bar{b}\right)\right)=1-P\left(\frac{Z-m_a\left(1-m_b\right)}{\sqrt{\frac{\left(s_a^2+m_a^2\right)\left(s_b^2+\left(1-m_b\right)^2\right)}{n}}}<\widetilde{q}\left(a,\bar{b}\right)\right) \tag{5}$$

where:

- m_a (resp. m_b) is the empirical mean of ψ_a (resp. ψ_b),
- s_a^2 (resp. s_b^2) is the variance of ψ_a (resp. ψ_b), and
- ϕ is the standard normal distribution function.

For a risk level α $(0<\alpha<1)$, propensity of a on b will be assumed if $\phi\left(-\widetilde{q}\left(a,\bar{b}\right)\right)\geq 1-\alpha$.

2.2. Contribution of the Individuals

To define the typicality and the contribution of the individuals to a class or a path for the binary case, the approach defined by Ratsimba-Rajohn (1992) has been used, which presents the notion of respect of an implication among binary variables, in the following way:

$$\varphi_x\left(a,\bar{b}\right)=1 \text{ if } \left[a\left(x\right)=1,\ a\left(x\right)=0\right] \text{ and } b\left(x\right)=1,$$
$$\varphi_x\left(a,\bar{b}\right)=0 \text{ if } a\left(x\right)=1 \text{ and } b\left(x\right)=0,$$
$$\varphi_x\left(a,\bar{b}\right)=p \text{ if } a\left(x\right)=0 \text{ and } b\left(x\right)=0,$$

where $\varphi_x\left(a,\bar{b}\right)$ is the intensity of the implication of a on b for the individual x and p is a number between 0 and 1.

For modal variables, Bailleul & Gras (1994) only analyze the contribution of the individuals to a class or a path. They proceed in two steps:

First. The set of the possible values that the couples of variables can take is arranged according to the weights that have been assigned to each one of the modalities.

In this classification, they divide the couples of variables in two groups:

G1: The couples (a, b) that verify: $\psi_a(x) = \dfrac{m-j}{m-1}$ and $\psi_b(x) = \dfrac{i-j}{m-1}$, with $1 \leq i \leq (m-1)$ and $1 \leq j \leq i$, which contradict the implication $a \Rightarrow b$. The rank assigned to these couples of variables is given through the expression $rg(\psi_a(x), \psi_b(x)) = \dfrac{i(i-1)}{2} + j$.

G2: The couples (a, b) that respect the implication $a \Rightarrow b$ and verify $\psi_a(x) = \dfrac{m-i}{m-1}$ and $\psi_b(x) = \dfrac{m-i+j}{m-1}$, with $m \geq i \geq 1$, and decrement -1, and $0 \leq j \leq (i-1)$. The rank assigned to these couples of variables is given by the expression $rg(\psi_a(x), \psi_b(x)) = m^2 - \dfrac{i(i-1)}{2} + j + 1$.

Second. The value that measures the respect or the contradiction of the implication is related to the implication intensity.

Thus, the grade of adhesion of the individual x to the implication $a \Rightarrow b$ is defined as:

$$deg_{a,b}(x) = \frac{rg(\psi_a(x), \psi_b(x)) - 1}{m^2 - 1} \varphi\left(a, \overline{b}\right) \tag{6}$$

From this grade of adhesion, Bailleul & Gras (1994) define the distance d of an individual x to the class C formed by g sub-classes, for which the g generic couples $(r_1, r_2, \ldots, r_g)$ and the g generic implications $(\varphi_1, \varphi_2, \ldots, \varphi_g)$ have been defined as (see Gras & Kuntz 2007):

$$d(x, C) = \left[\frac{1}{g} \sum_{i=1}^{g} \frac{(\varphi_i - deg_{r_i}(x))^2}{1 - \varphi_i}\right]^{1/2} \tag{7}$$

and the contribution of an individual x to the class C as:

$$\gamma(x, C) = 1 - \frac{d(x, C)}{d(x_n, C)} \tag{8}$$

where x_n is the neutral individual, that is, the individual that assigns the value 0 to all the variables of the g generic couples r_i.

3. Classificatory Analysis for Modal Variables

In this section, an index to quantify the similarity among modal variables, the typicality and the contribution of the individuals to a class C is presented.

3.1. Similarity Index

Following the method to define the propensity coefficient, let us denote a set E of size N, with sufficiently large N, from which a random subset I of size n is selected, and a set $A = \{a_1, a_2, \ldots, a_p\}$ formed by p modal variables, each one of which can take values in the set $M = \{k \in N : 1 \leq k \leq m\}$. Specific weights are

assigned to each variable, from 1 (the strongest) to 0, in size fractions $\frac{1}{m-1}$ by means of the function ψ defined in (1).

Let $\{A_x\}_{x\in E}$ and $\{B_x\}_{x\in E}$ be two groups of random variables with equal distributions, representatives of the variables a and b with weights ψ_a and ψ_b respectively. Under the assumption of non a priori relationship between a and b, the $2N$ variables $\{A_x\}$, $\{B_x\}$ are independents and the N events $\{x \in I\}$ are independents with an occurrence probability equal to $\frac{n}{N}$.

Let us define the random variable:

$$\sum_{x\in I} \frac{A_x B_x}{n} \tag{9}$$

for which $\sum_{x\in I} \frac{\psi_a(x)\,\psi_b(x)}{n}$ is one of their possible values. This random variable follows approximately a Normal distribution of mean $\mu = E[A_x]\,E[B_x]$ and variance $\sigma^2 = \left(Var(A_x) + [E(A_x)]^2\right)\left(Var(B_x) + [E(B_x)]^2\right)$ and therefore, for n sufficiently large,

$$\frac{\sum_{x\in I} \frac{A_x B_x}{n} - \mu}{\sigma} \tag{10}$$

follows approximately a Normal distribution of mean zero and variance one.

The proof of this result is similar to the one presented by Lagrange (1998) for the random variable $\sum_{x\in I} \frac{A_x(1-B_x)}{n}$, therefore it doesn't show up in this article.

The similarity index of a on b is defined as:

$$\widetilde{s}(a,b) = P\left(\frac{\sum_{x\in I} \frac{A_x B_x}{n} - m_a m_b}{\sqrt{\frac{\left(s_a^2 + m_a^2\right)\left(s_b^2 + m_b^2\right)}{n}}} < \frac{\sum_{x\in I} \frac{\psi_a(x)\,\psi_b(x)}{n} - m_a m_b}{\sqrt{\frac{\left(s_a^2 + m_a^2\right)\left(s_b^2 + m_b^2\right)}{n}}}\right) \tag{11}$$

where:

- m_a (m_b) is the empirical mean of ψ_a (ψ_b),
- s_a^2 (s_b^2) is the variance of ψ_a (ψ_b),

which coincides with the similarity index defined for the binary case.

Let us prove this. For the binary case, there are just two modalities, this is, $M = \{1, 2\}$ and the unique assignable weights are 0 and 1, therefore:

- $\sum_{x\in I} \frac{A_x B_x}{n} = \frac{1}{n} Card(A \cap B)$

- $\sum_{x\in I} \frac{\psi_a(x)\,\psi_b(x)}{n} = \sum_{x\in I} \frac{a(x)\,b(x)}{n} = \frac{n_{a\wedge b}}{n}$, where:
  $n_{a\wedge b} = \{x \in I : \psi_a(x) = 1 \wedge \psi_b(x) = 1\} = \{x \in I : a(x) = 1 \wedge b(x) = 1\}$
- $m_a = \frac{1}{n}\sum_{x\in I} \psi_a(x) = \frac{n_a}{n}$, where:
  $n_a = \{x \in I : \psi_a(x) = 1\} = \{x \in I : a(x) = 1\}$
- $s_a^2 = \widehat{\sigma}^2(\psi_a) = \frac{1}{n}\left[\sum_{x\in I} \psi_a^2(x) - nm_a^2\right] = m_a - m_a^2$,
  therefore, $s_a^2 + m_a^2 = m_a$, and

$$\widetilde{s}(a,b) = P\left(\frac{\frac{Card(A\bigcap B)}{n} - \frac{n_a n_b}{n^2}}{\sqrt{\frac{n_a n_b}{n^3}}} < \frac{\frac{n_{a\wedge b}}{n} - \frac{n_a n_b}{n^2}}{\sqrt{\frac{n_a n_b}{n^3}}}\right)$$

$$= P\left(\frac{Card(A\bigcap B) - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}} < \frac{n_{a\wedge b} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}\right) = s(a,b).$$

3.2. Significant Nodes

For the determination of the significant nodes, a similar procedure to that of the binary case is used, but to define the initial and global pre-order on $A \times A$, the index $\widetilde{s}(a,b)$ is used, this is, $G_{\widetilde{s}}(\Omega) = \{((a,b),(c,d)) : \widetilde{s}(a,b) < \widetilde{s}(c,d)\}$. For more details, see (Zamora, Gregori & Orús 2009).

3.3. Typicality and Contribution to a Class C

The couple (a,b) so that:

$$\widetilde{s}(a,b) \geq \widetilde{s}(i,j), \quad \forall i \in A \text{ and } j \in B,$$

is called the generic couple of the class $C : A \sim B$. The number $\widetilde{s}(a,b)$ is called generic similarity of C.

If the class C has g subclasses (C included), then there will be g generic similarities $(\widetilde{s}_1, \widetilde{s}_2, \ldots, \widetilde{s}_g)$.

Based on the idea presented by Bailleul & Gras (1994), the grade of agreement between two modal variables a and b is defined. For that purpose:

First. The set of the possible values that the couples of variables can take is arranged, according to the weights that have been assigned to each modality.

The couples of this ordered set are divided into two groups:

G1: The couples (a,b) that verify the similarity $a \sim b$ or for which $a(x) = b(x)$.

G2: The couples (a, b) that don't verify the similarity.

To establish a classification to the levels of similarity or non similarity, ranks are assigned to the couples of values that the couples of variables (a, b) can take for a given individual x, like it is shown below.

Ranks are placed on a $m \times m$ matrix (see Table 1 for the case $m = 4$), where the rank assigned to the cell (i, j), corresponding to the i-th modality of the variable a and the j-th modality of the variable b, is given by the expression:

$$Rg_{i,j} = \begin{cases} (m - j + 1) + \dfrac{(i-1)(2m-i)}{2}, & for\ i \neq j \\ \dfrac{m(m-1)}{2} + 1, & for\ i = j \end{cases} \tag{12}$$

with $1 \leq i \leq (m-1)$ and j decreasing from m to $(i+1)$ with decrement (-1). It is also defined that $Rg_{i,j} = Rg_{j,i}$.

Table 1: Assigned ranges.

i/j	1	2	3	4
1	7	3	2	1
2		7	5	4
3			7	6
4				7

From this table we can appreciate that:

- The couple $(1, 4)$ corresponding to an individual for which $\psi_a(x) = 0$ and $\psi_b(x) = 1$ receives the lower rank, 1.
- For the rest of the cases, the cell receives a rank between 1 and 7, according to the level of discrepancy or disagreement between $a(x)$ and $b(x)$.

Second. The value that measures the respect or the contradiction to the full agreement is linked to the similarity index.

For that, the grade of agreement of the individual x to the relationship $a \sim b$ is defined as:

$$deg_{a,b}(x) = \frac{Rg_{i,j} - 1}{\dfrac{m(m-1)}{2}} \widetilde{s}(a, b) \tag{13}$$

where $a(x) = i$ and $b(x) = j$.

The typical subject is defined as the subject whose grade of agreement coincides with the value of the generic similarity for all the generic couples, this is, $deg_{r_i}(x) = \widetilde{s}_i\ \forall r_i$.

The distance of an individual x to the class C formed by g sub-classes is then defined as:

$$d(x, C) = \left[\frac{1}{g} \sum_{i=1}^{g} \frac{\left(\widetilde{s}_i - deg_{r_i}(x)\right)^2}{1 - \widetilde{s}_i} \right]^{1/2} \tag{14}$$

where $(r_1, r_2, \ldots, r_g)$ are the g generic couples and $(\widetilde{s}_1, \widetilde{s}_2, \ldots, \widetilde{s}_g)$ represents the g generic similarities.

The typicality is defined by the fact that certain individuals are representative of the population's behavior; this is, with a similarity index close to that of the formed class. The typicality of an individual x to the class C is calculated by the expression:

$$\gamma(x, C) = 1 - \frac{d(x, C)}{\max\limits_{y \in I} d(y, C)} \tag{15}$$

Let us analyze the two extreme cases:

- If x is a typical subject, it should satisfy that $deg_{r_i} = \widetilde{s}_i$ for all generic couples r_i, which implies that $Rg_{r_i} = \dfrac{m(m-1)}{2} + 1$, that is, $a(x) = b(x)$ for all the generic couples $r_i = (a, b)$, and therefore, $d(x, C) = 0$ and $\gamma(x, C) = 1$.
- If x is in more disagreement with C, $d(x, C) = \max\limits_{y \in I} d(y, C)$, which means that for all generic couples r_i, $deg_{r_i}(x) = 0$, that is, the rank that is assigned to all generic couples is 1, which occurs if for all generic couples $r_i = (a, b)$, $a(x) = 1$ and $b(x) = m$, or $a(x) = m$ and $b(x) = 1$. However, if $d(x, C) = \max\limits_{y \in I} d(y, C)$, then $\gamma(x, C) = 0$.

There are individuals more responsible than others in the formation of the class. The contribution is defined to evaluate this degree of responsibility. The contribution of an individual x to the class C is given by:

$$\widetilde{\gamma}(x, C) = 1 - \widetilde{d}(x, C) \tag{16}$$

where $\widetilde{d}(x, C) = \left[\dfrac{1}{g}\sum_{i=1}^{g}\left(1 - deg^*_{r_i}(x)\right)^2\right]^{1/2}$ is the distance from the individual x to the class C and $deg^*_{r_i} = \dfrac{Rg_{r_i} - 1}{\dfrac{m(m-1)}{2}}$.

Let us analyze the two extreme cases, the individual of higher contribution and the individual of lowest contribution to the formation of the class:

- If x is the individual of higher contribution to the formation of the class, then it satisfies that $a(x) = b(x)$ and therefore $Rg_{r_i} = \dfrac{m(m-1)}{2} + 1$ for all generic couples $r_i = (a, b)$. Hence $deg^*_{r_i}(x) = 1$, $\forall r_i$, $\Rightarrow \widetilde{d}(x, C) = 0 \ \Rightarrow \widetilde{\gamma}(x, C) = 1$.
- If x is the individual of lowest contribution to the formation of the class, it satisfies that $a(x) = 1$ and $b(x) = m$ or $a(x) = m$ and $b(x) = 1$ and therefore $Rg_{r_i} = 1$ for all generic couples $r_i = (a, b)$. Hence $deg^*_{r_i}(x) = 0$, $\forall r_i$ and $\widetilde{d}(x, C) = 1 \ \Rightarrow \ \widetilde{\gamma}(x, C) = 0$.

4. Illustration of the Technique

The techniques presented have been programmed in language R (R Core Team 2014), in a file that will be denoted as ASIM.R. Two examples are presented, the first with binary data, to show the equivalence of the similarity index proposed with the existent one for the binary case. The second example deals with modal variables with more than two modalities, where the similarity index, the typicality and contribution of the individuals to the classes formed during the construction of the similarity tree are computed.

In the first example, seventeen binary variables are analyzed in a sample of 100 patients, from which 50 suffer from lung cancer and the other 50 don't. Before presenting the results of the application of both systems, CHIC and ASIM, it is necessary to point out that CHIC, for similarity analysis in the binary case, allows the user to choose the distribution (Poisson or Binomial); however, the results that CHIC shows, in both cases, correspond to the Poisson approximation to the Normal distribution. Given this difficulty, it is possible to compare the result shown by the CHIC for the binary case, with the result shown by ASIM for the modal case with two modalities.

The index proposed for the modal case with 2 modalities coincides with the index proposed for the binary case, in this case the random variable $\sum_{x\in I}\frac{A_xB_x}{n}$ follows approximately a Normal distribution of mean $\mu = E[A_x]\,E[B_x] = \dfrac{n_a n_b}{n}$ and variance $\sigma^2 = \left(Var(A_x) + [E(A_x)]^2\right)\left(Var(B_x) + [E(B_x)]^2\right) = \dfrac{n_a n_b}{n}$, distribution that coincides with the Poisson approximation to the Normal distribution.

The Figures 1 and 2 show the similarity indexes at level zero of the hierarchy. The coincidence of the results can be appreciated.

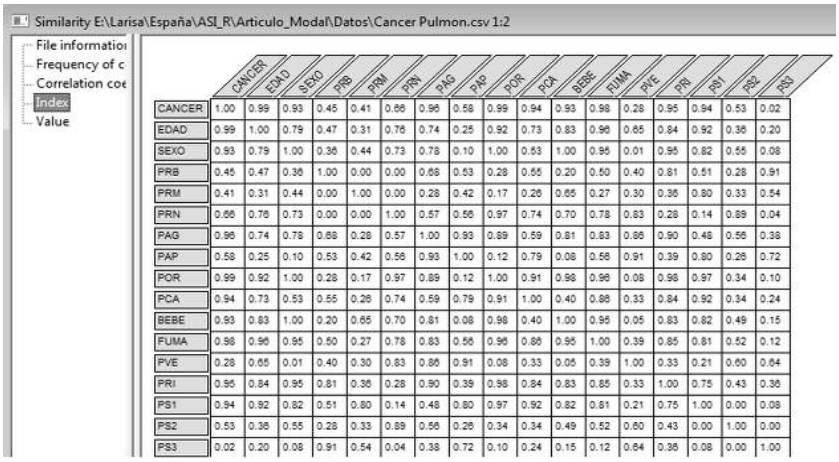
Console E:/Larisa/Larisa_Personales/España/ASI_R/Modal_Agosto_19_2014/

Indices de similaridad al nivel cero

	CANCER	EDAD	SEXO	PRB	PRM	PRN	PAG	PAP	POR	PCA	BEBE	FUMA	PVE	PRI	PS1	PS2	PS3
CANCER	1.00	0.99	0.93	0.45	0.41	0.66	0.96	0.58	0.99	0.94	0.93	0.98	0.28	0.95	0.94	0.53	0.02
EDAD	0.99	1.00	0.79	0.47	0.31	0.76	0.74	0.25	0.92	0.73	0.83	0.96	0.65	0.84	0.92	0.36	0.20
SEXO	0.93	0.79	1.00	0.36	0.44	0.73	0.78	0.10	1.00	0.53	1.00	0.95	0.01	0.95	0.82	0.55	0.08
PRB	0.45	0.47	0.36	1.00	0.00	0.00	0.68	0.53	0.28	0.55	0.20	0.50	0.40	0.81	0.51	0.28	0.91
PRM	0.41	0.31	0.44	0.00	1.00	0.00	0.28	0.42	0.17	0.26	0.65	0.27	0.30	0.36	0.80	0.33	0.54
PRN	0.66	0.76	0.73	0.00	0.00	1.00	0.57	0.56	0.97	0.74	0.70	0.78	0.83	0.28	0.14	0.89	0.04
PAG	0.96	0.74	0.78	0.68	0.28	0.57	1.00	0.93	0.89	0.59	0.81	0.83	0.86	0.90	0.48	0.56	0.38
PAP	0.58	0.25	0.10	0.53	0.42	0.56	0.93	1.00	0.12	0.79	0.08	0.56	0.91	0.39	0.80	0.26	0.72
POR	0.99	0.92	1.00	0.28	0.17	0.97	0.89	0.12	1.00	0.91	0.98	0.96	0.08	0.98	0.97	0.34	0.10
PCA	0.94	0.73	0.53	0.55	0.26	0.74	0.59	0.79	0.91	1.00	0.40	0.86	0.33	0.84	0.92	0.34	0.24
BEBE	0.93	0.83	1.00	0.20	0.65	0.70	0.81	0.08	0.98	0.40	1.00	0.95	0.05	0.83	0.82	0.49	0.15
FUMA	0.98	0.96	0.95	0.50	0.27	0.78	0.83	0.56	0.96	0.86	0.95	1.00	0.39	0.85	0.81	0.52	0.12
PVE	0.28	0.65	0.01	0.40	0.30	0.83	0.86	0.91	0.08	0.33	0.05	0.39	1.00	0.33	0.21	0.60	0.64
PRI	0.95	0.84	0.95	0.81	0.36	0.28	0.90	0.39	0.98	0.84	0.83	0.85	0.33	1.00	0.75	0.43	0.36
PS1	0.94	0.92	0.82	0.51	0.80	0.14	0.48	0.80	0.97	0.92	0.82	0.81	0.21	0.75	1.00	0.00	0.08
PS2	0.53	0.36	0.55	0.28	0.33	0.89	0.56	0.26	0.34	0.34	0.49	0.52	0.60	0.43	0.00	1.00	0.00
PS3	0.02	0.20	0.08	0.91	0.54	0.04	0.38	0.72	0.10	0.24	0.15	0.12	0.64	0.36	0.08	0.00	1.00

FIGURE 1: Similarity indexes at the level zero of the hierarchy with ASIM.R.

The Figures 3 and 4 show the classes formed during the construction of the similarity tree and its corresponding similarity indexes, obtained by ASIM.R.



	CANCER	EDAD	SEXO	PRB	PRM	PRN	PAG	PAP	POR	PCA	BEBE	FUMA	PVE	PRI	PS1	PS2	PS3
CANCER	1.00	0.99	0.93	0.45	0.41	0.66	0.96	0.58	0.99	0.94	0.93	0.98	0.28	0.95	0.94	0.53	0.02
EDAD	0.99	1.00	0.79	0.47	0.31	0.76	0.74	0.25	0.92	0.73	0.83	0.96	0.65	0.84	0.92	0.36	0.20
SEXO	0.93	0.79	1.00	0.36	0.44	0.73	0.78	0.10	1.00	0.53	1.00	0.95	0.01	0.95	0.82	0.55	0.08
PRB	0.45	0.47	0.36	1.00	0.00	0.00	0.68	0.53	0.28	0.65	0.20	0.50	0.40	0.81	0.51	0.28	0.91
PRM	0.41	0.31	0.44	0.00	1.00	0.00	0.28	0.42	0.17	0.26	0.65	0.27	0.30	0.36	0.80	0.33	0.54
PRN	0.66	0.76	0.73	0.00	0.00	1.00	0.57	0.56	0.97	0.74	0.70	0.78	0.83	0.28	0.14	0.89	0.04
PAG	0.96	0.74	0.78	0.68	0.28	0.57	1.00	0.93	0.89	0.59	0.81	0.83	0.86	0.90	0.48	0.56	0.38
PAP	0.58	0.25	0.10	0.53	0.42	0.56	0.93	1.00	0.12	0.79	0.08	0.56	0.91	0.39	0.80	0.26	0.72
POR	0.99	0.92	1.00	0.28	0.17	0.97	0.89	0.12	1.00	0.91	0.98	0.96	0.08	0.98	0.97	0.34	0.10
PCA	0.94	0.73	0.53	0.65	0.26	0.74	0.59	0.79	0.91	1.00	0.40	0.88	0.33	0.84	0.92	0.34	0.24
BEBE	0.93	0.83	1.00	0.20	0.65	0.70	0.81	0.08	0.98	0.40	1.00	0.95	0.05	0.83	0.82	0.49	0.15
FUMA	0.98	0.96	0.95	0.50	0.27	0.78	0.83	0.56	0.96	0.88	0.95	1.00	0.39	0.85	0.81	0.52	0.12
PVE	0.28	0.65	0.01	0.40	0.30	0.83	0.86	0.91	0.08	0.33	0.05	0.39	1.00	0.33	0.21	0.60	0.64
PRI	0.95	0.84	0.95	0.81	0.36	0.28	0.90	0.39	0.98	0.84	0.83	0.85	0.33	1.00	0.75	0.43	0.36
PS1	0.94	0.92	0.82	0.51	0.80	0.14	0.48	0.80	0.97	0.92	0.82	0.81	0.21	0.75	1.00	0.00	0.08
PS2	0.53	0.36	0.55	0.28	0.33	0.89	0.56	0.26	0.34	0.34	0.49	0.52	0.60	0.43	0.00	1.00	0.00
PS3	0.02	0.20	0.08	0.91	0.54	0.04	0.38	0.72	0.10	0.24	0.15	0.12	0.64	0.36	0.08	0.00	1.00

FIGURE 2: Similarity indexes at the level zero of the hierarchy, obtained by CHIC.

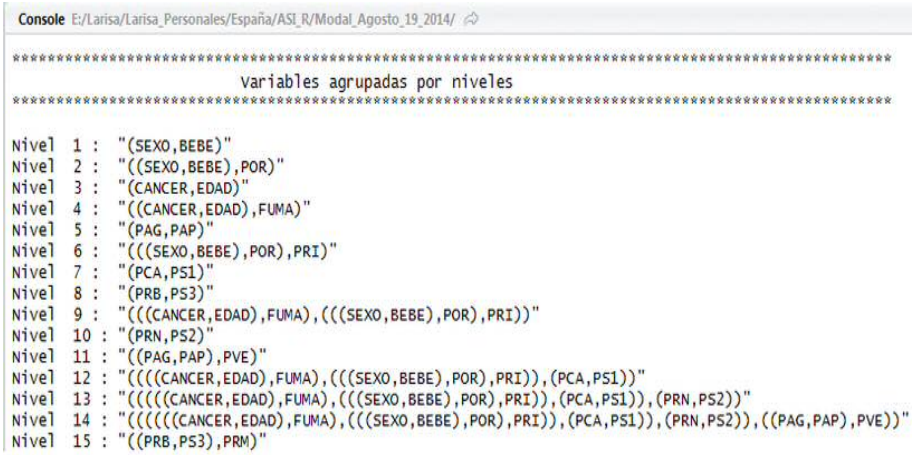


FIGURE 3: Classes obtained by running ASIM.R.

Figure 5 shows the same results, by executing the software CHIC. The coincidence of both results can be appreciated.

Figures 6a and 6b show the similarity trees obtained by means of each program. Again the results are coincident.

Through the running of the above example and from the execution of other datasets, we can say that the execution times between CHIC and ASIM.R are similar.

Next, the results of a modal variable case with five modalities are presented. To accomplish this, the file "DatosMusicaModalII.txt" is used. This file corresponds to the degree of preference of twenty students to fifteen different kinds of music. The degree of preference is assessed by a scale from 1 to 5, 5 being the highest value of preference.

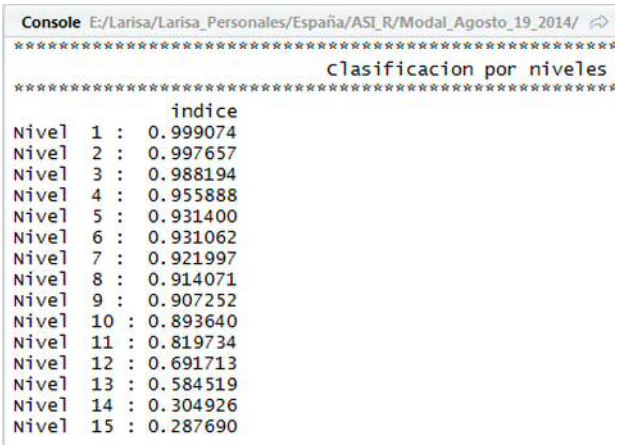


FIGURE 4: Similarity indexes obtained by running ASIM.R.

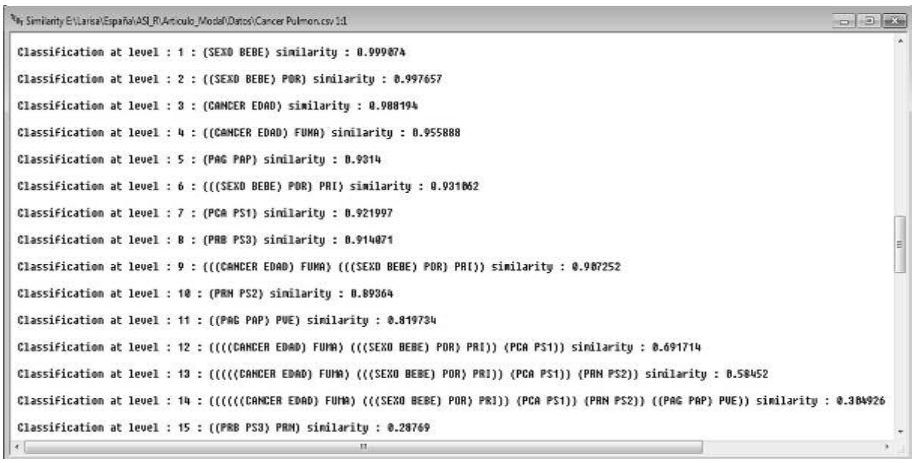
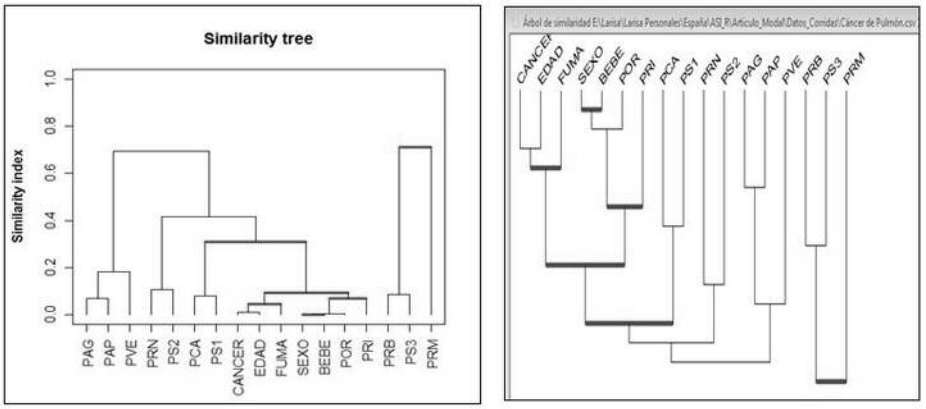


FIGURE 5: Classes and similarity indexes obtained by executing CHIC.

Figures 7, 8 and 9 show the similarity indexes obtained at the level zero of the hierarchy, the grouping of variables in each level of this hierarchy, and the indexes to which they were joined.

The similarity index of the class 1 formed by variables HIP and PUN (see Figures 8 and 9) is 0.934143; it means that HIP and PUN are the most similar music types in the population of which the sample under study has been selected.

The second class is formed by the variables JAZ and HEA. This class also shows a big similarity index (0.922925), but smaller than the previous one (see Figures 8 and 9). Then we can say that the preference for these music types is high, but smaller than for the previous group (HIP and PUN). The other classes are interpreted in a similar way.



(a) By ASIM.R (b) By CHIC

FIGURE 6: Similarity trees.

Console E:/Larisa/Larisa_Personales/España/ASI_R/Modal_Agosto_19_2014/

```
*****************************************************************************
                    Indices de similaridad al nivel cero
*****************************************************************************
     HIP  OPE  MAQ  JAZ  FLA  RUM  HEA  POP  CLA  DAN  REG  HAR  RAP  PUN  SKA
HIP 1.00 0.29 0.56 0.54 0.27 0.63 0.73 0.56 0.75 0.71 0.76 0.56 0.28 0.93 0.60
OPE 0.29 1.00 0.40 0.79 0.78 0.34 0.35 0.40 0.90 0.20 0.33 0.30 0.84 0.26 0.72
MAQ 0.56 0.40 1.00 0.64 0.50 0.43 0.64 0.81 0.40 0.52 0.54 0.73 0.42 0.57 0.61
JAZ 0.54 0.79 0.64 1.00 0.68 0.44 0.92 0.64 0.80 0.36 0.90 0.64 0.38 0.67 0.72
FLA 0.27 0.78 0.50 0.68 1.00 0.46 0.39 0.50 0.61 0.39 0.68 0.50 0.89 0.31 0.43
RUM 0.63 0.34 0.43 0.44 0.46 1.00 0.57 0.43 0.58 0.64 0.59 0.43 0.41 0.65 0.41
HEA 0.73 0.35 0.64 0.92 0.39 0.57 1.00 0.64 0.77 0.49 0.92 0.64 0.22 0.83 0.71
POP 0.56 0.40 0.81 0.64 0.50 0.43 0.64 1.00 0.40 0.52 0.54 0.73 0.42 0.57 0.61
CLA 0.75 0.90 0.40 0.80 0.61 0.58 0.77 0.40 1.00 0.41 0.76 0.30 0.46 0.71 0.51
DAN 0.71 0.20 0.52 0.36 0.39 0.64 0.49 0.52 0.41 1.00 0.52 0.52 0.36 0.73 0.38
REG 0.76 0.33 0.54 0.90 0.68 0.59 0.92 0.54 0.76 0.52 1.00 0.54 0.38 0.79 0.61
HAR 0.56 0.30 0.73 0.64 0.50 0.43 0.64 0.73 0.30 0.52 0.54 1.00 0.42 0.57 0.61
RAP 0.28 0.84 0.42 0.38 0.89 0.41 0.22 0.42 0.46 0.36 0.38 0.42 1.00 0.31 0.63
PUN 0.93 0.26 0.57 0.67 0.31 0.65 0.83 0.57 0.71 0.73 0.79 0.57 0.31 1.00 0.61
SKA 0.60 0.72 0.61 0.72 0.43 0.41 0.71 0.61 0.51 0.38 0.61 0.61 0.63 0.61 1.00
```

FIGURE 7: Similarity indexes at the level zero of the hierarchy.

Console E:/Larisa/Larisa_Personales/España/ASI_R/Modal_Agosto_19_2014/

```
*****************************************************************
                    Variables agrupadas por niveles
*****************************************************************

Nivel  1 :  "(HIP,PUN)"
Nivel  2 :  "(JAZ,HEA)"
Nivel  3 :  "(OPE,CLA)"
Nivel  4 :  "(FLA,RAP)"
Nivel  5 :  "((JAZ,HEA),REG)"
Nivel  6 :  "(MAQ,POP)"
Nivel  7 :  "(RUM,DAN)"
Nivel  8 :  "(HAR,SKA)"
Nivel  9 :  "((OPE,CLA),(FLA,RAP))"
Nivel  10 : "((HIP,PUN),((JAZ,HEA),REG))"
Nivel  11 : "((MAQ,POP),(HAR,SKA))"
Nivel  12 : "(((HIP,PUN),((JAZ,HEA),REG)),(RUM,DAN))"
```

FIGURE 8: Classes obtained.

```
Console E:/Larisa/Larisa Personales/España/ASI_R/Modal_Agosto_19_
*********************************************************
                          Clasificacion por niveles
*********************************************************
              indice
Nivel  1 :  0.934143
Nivel  2 :  0.922925
Nivel  3 :  0.900668
Nivel  4 :  0.889664
Nivel  5 :  0.851791
Nivel  6 :  0.814453
Nivel  7 :  0.644702
Nivel  8 :  0.613585
Nivel  9 :  0.497993
Nivel  10 : 0.327310
Nivel  11 : 0.285067
Nivel  12 : 0.042301
```

FIGURE 9: Similarity indexes obtained.

Figure 10 presents the similarity tree for the music data classification shown in Figures 8 and 9.

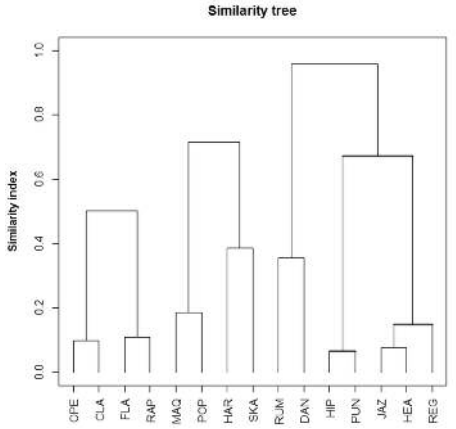


FIGURE 10: Similarity tree.

The typicality of the individuals and the optimal group are shown for each one of the formed classes. In Figure 11 we can see that individuals for which the value of the typicality in a class is 0 exist, for example, the individual 7 in the class 3. These individuals are in more disagreement, this is, the values taken for the variables in each generic couple for these individuals are respectively the minimum (modality 1) and maximum (modality 5) value or vice versa.

On the contrary, there are individuals whose typicality value in a class is 1, for example, the individual 1 of the class 3. The values taken for the variables in each generic couple for these individuals are the same. The degree of agreement among these variables for this individual coincides with the similarity index, being hence the most typical individual.

Figures 13 and 14 show the contributions of the individuals and the optimal group. The results related to the contribution can be interpreted as follows. When the contribution of an individual to a class is 1, for example, individual 1 of class

```
Console E:/Larisa/Larisa Personales/España/ASI_R/Modal_Agosto_19_2014/
*****************************************************************************************************
                    Tipicalidades de los individuos a las clases
*****************************************************************************************************
         Clase[1] Clase[2] Clase[3] Clase[4]  Clase[5] Clase[6] Clase[7] Clase[8]  Clase[9] Clase[10]
Indiv_1   1.00000  1.00000  1.00000        0 1.0000000        1        1        0 0.3110467 0.5715400
Indiv_2   1.00000  1.00000  1.00000        1 1.0000000        1        1        1 1.0000000 0.9330531
Indiv_3   1.00000  1.00000  1.00000        0 1.0000000        1        1        1 0.2309112 0.5715400
Indiv_4   1.00000  1.00000  1.00000        1 1.0000000        1        0        0 1.0000000 0.5715400
Indiv_5   1.00000  1.00000  0.25000        1 1.0000000        1        1        1 0.4527748 0.9330531
Indiv_6   1.00000  1.00000  1.00000        1 1.0000000        1        1        0 0.4637086 0.8928850
Indiv_7   1.00000  1.00000  0.00000        0 1.0000000        1        1        0 0.0000000 1.0000000
Indiv_8   1.00000  0.25000  0.34375        1 0.4696699        1        1        1 0.5211780 0.4696699
Indiv_9   1.00000  1.00000  1.00000        1 0.4696699        1        1        1 0.4637086 0.3799092
Indiv_10  1.00000  1.00000  1.00000        1 1.0000000        1        1        1 0.8659272 1.0000000
```

FIGURE 11: Typicalities of the first 10 individuals to the first 10 classes.

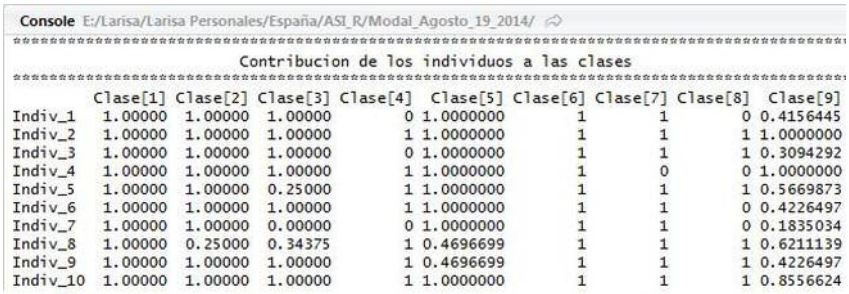
```
Console E:/Larisa/Larisa Personales/España/ASI_R/Modal_Agosto_19_2014/
Grupos optimales de individuos para las tipicalidades:
Optimal Group to Clase[1]: (HIP,PUN)
17 1 2 3 4 5 6 7 8 9 10 12 13 14 15 16 18 19 20
Card :19

Optimal Group to Clase[2]: (JAZ,HEA)
20 1 2 3 4 5 6 7 9 10 11 12 14 16 17 18 19
Card :17

Optimal Group to Clase[3]: (OPE,CLA)
1 2 3 4 6 9 10 11 12 13 14 16 17 18 19 20
Card :16

Optimal Group to Clase[4]: (FLA,RAP)
2 4 5 6 8 9 10 11 12 13 14 15 16 17 18 19
Card :16

Optimal Group to Clase[5]: ((JAZ,HEA),REG)
1 2 3 4 5 6 7 10 11 12 14 16 18 19
Card :14

Optimal Group to Clase[6]: (MAQ,POP)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Card :20
```

FIGURE 12: Optimal group for the first 6 classes.

1, the values of the variables for these individuals, in each generic couple, are equal. This individual evaluates in a similar way his preference for the music types that form the vector of generic similarities for the class and therefore it is more responsible in the creation of the class.

If the contribution of an individual to a class is 0, for example, individual 1 of class 4, then its distance to the class is the biggest possible (1). This is the individual that contributes the least to the formation of the class (FLA, RAP), doesn't like the music FLA and assigns the biggest value possible to the preference for the music RAP. It is the individual of lowest contribution for declaring FLA and RAP as similar music types.

For space reasons, only a part of the typicalities, contributions and optimal group are shown.

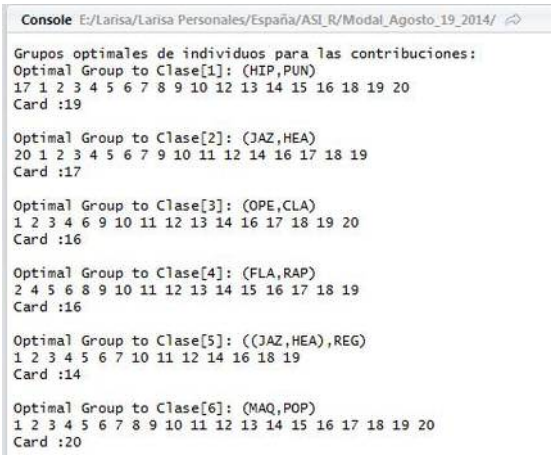


Console E:/Larisa/Larisa Personales/España/ASI_R/Modal_Agosto_19_2014/

Contribucion de los individuos a las clases

	Clase[1]	Clase[2]	Clase[3]	Clase[4]	Clase[5]	Clase[6]	Clase[7]	Clase[8]	Clase[9]
Indiv_1	1.00000	1.00000	1.00000	0	1.0000000	1	1	0	0.4156445
Indiv_2	1.00000	1.00000	1.00000	1	1.0000000	1	1	1	1.0000000
Indiv_3	1.00000	1.00000	1.00000	0	1.0000000	1	1	1	0.3094292
Indiv_4	1.00000	1.00000	1.00000	1	1.0000000	1	0	0	1.0000000
Indiv_5	1.00000	1.00000	0.25000	1	1.0000000	1	1	1	0.5669873
Indiv_6	1.00000	1.00000	1.00000	1	1.0000000	1	1	0	0.4226497
Indiv_7	1.00000	1.00000	0.00000	0	1.0000000	1	1	0	0.1835034
Indiv_8	1.00000	0.25000	0.34375	1	0.4696699	1	1	1	0.6211139
Indiv_9	1.00000	1.00000	1.00000	1	0.4696699	1	1	1	0.4226497
Indiv_10	1.00000	1.00000	1.00000	1	1.0000000	1	1	1	0.8556624

FIGURE 13: Contribution of the first 10 individuals to the first 9 classes.



Console E:/Larisa/Larisa Personales/España/ASI_R/Modal_Agosto_19_2014/

```
Grupos optimales de individuos para las contribuciones:
Optimal Group to Clase[1]: (HIP,PUN)
17 1 2 3 4 5 6 7 8 9 10 12 13 14 15 16 18 19 20
Card :19

Optimal Group to Clase[2]: (JAZ,HEA)
20 1 2 3 4 5 6 7 9 10 11 12 14 16 17 18 19
Card :17

Optimal Group to Clase[3]: (OPE,CLA)
1 2 3 4 6 9 10 11 12 13 14 16 17 18 19 20
Card :16

Optimal Group to Clase[4]: (FLA,RAP)
2 4 5 6 8 9 10 11 12 13 14 15 16 17 18 19
Card :16

Optimal Group to Clase[5]: ((JAZ,HEA),REG)
1 2 3 4 5 6 7 10 11 12 14 16 18 19
Card :14

Optimal Group to Clase[6]: (MAQ,POP)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Card :20
```

FIGURE 14: Optimal group for the first 6 classes.

5. Conclusions

In this work we proposed a similarity index among modal variables, and expressions for the calculation of the typicalities and contributions of the individuals to a class, for these types of variables. It has been demonstrated, theoretically and numerically, that the expression for the calculation of the similarity index for modal variables with two modalities coincides with the expression for the binary case, thus the index proposed is a generalization of the binary case. Following the idea of Bailleul & Gras (1994), we propose an expression to calculate the grade of agreement between two modal variables and from this, expressions are given to calculate the typicalities and contributions of the individuals to the formed classes.

[Received: June 2014 — Accepted: December 2014]

References

Bailleul, M. & Gras, R. (1994), 'L'implication statistique entre variables modales', *Mathématiques et Sciences Humaines* **128**, 47–51.

Couturier, R. (2008), 'CHIC: Cohesive hierarchical implicative classification', *Studies in Computational Intelligence (SCI)* **127**, 41–53.

Gras, R. (2000), 'Les fondements de l'analyse statistique implicative', *Quaderni di Ricerca in Didattica* **9**, 187–208.

Gras, R. & Kuntz, K. (2007), Analyse statistique implicative (ASI), en réponse à des problèmes fondateurs, *in* 'Apports Théoriques à l'Analyse Statistique Implicative et Applications', Université Jaume I, Castellón, pp. 15–40.

Lagrange, J. B. (1998), 'Analyse implicative d'un ensemble de variables numériques; application au traitement d'un questionnaire à résponses modales ordonnées', *Revue de Statistique Appliquée* **46**(1), 71–93.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.
*http://CRAN.R-project.org

Ratsimba-Rajohn, H. (1992), Contribution à l'étude de la hiérarchie implicative, application á l'analyse de la gestión didactique des phénomènes d'ostension et de contradiction, Master Thesis, Université de Rennes, France.

Zamora, L., Gregori, P. & Orús, P. (2009), *Conceptos Fundamentales del Análisis Estadístico Implicativo (ASI) y su Soporte Computacional CHIC*, Departamento de Matemáticas, Universitat Jaume I de Castellón, España.