Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2008, Article ID 125797, 13 pages
doi:10.1155/2008/125797

Research Article

Clustering Objects Described by Juxtaposition of Binary Data Tables

Amar Rebbouh

MSTD Laboratory, Faculty of Mathematics, University of Science and Technology-Houari Boumediene (USTHB), El Alia, BP 32, Algiers 16011, Algeria

Correspondence should be addressed to Amar Rebbouh, arebbouh@yahoo.fr

Received 3 April 2008; Revised 14 August 2008; Accepted 28 October 2008

Recommended by Khosrow Moshirvaziri

This paper seeks to develop an allocation of 0/1 data matrices to physical systems upon a Kullback-Leibler distance between probability distributions. The distributions are estimated from the contents of the data matrices. We discuss an ascending hierarchical classification method, a numerical example and mention an application with survey data concerning the level of development of the departments of a given territory of a country.

1. Introduction

The automatic classification of the components of a structure of multiple tables remains a vast field of research and investigation. The components are matrix objects, and the difficulty lies in the definition and the delicate choice of an index of distance between these objects (see Lerman and Tallur [1]). If the matrix objects are tables of measurements of the same dimension, we introduced in Rebbouh [2] an index of distance based on the inner product of Hilbert Schmidt and built a classification algorithm of k-means type. In this paper, we are interested by the case when matrix objects are tables gathering the description of the individuals by nominal variables. These objects are transformed into complete disjunctive tables containing 0/1 data (see Agresti [3]). It is thus a particular structure of multiple data tables frequently encountered in practice each time one is led to carry out several observations on the individuals whom one wishes to classify (see [4, 5]). We quote for example the case when we wish to classify administrative departments according to indices measuring the level of economic idE and human development idH which are weighted averages calculated starting from measurements of selected parameters. Each department i gathers a number L_i of subregions classified as rural or urban. Each department is thus described by a matrix with 2 columns and L_i lines of positive numbers ranging between 0 and 1. But the fact that

the values of the indices do not have the same sense according to the geographical position of the subregion, and its urban or rural specificity led the specialists to be interested in the membership of the subregion to quintile intervals. Thus, each department will be described by a matrix object with 10 columns and L_i lines. This matrix object is thus a juxtaposition of 2 tables of the same dimension containing only one 1 on each line which corresponds to the class where the subregion is affected and the remainder are 0. The use of conventional statistical techniques to analyze this kind of data requires a reduction step. Several criteria of good reduction exist. The criterion that gives results easily usable and interpretable is undoubtedly the least squares criterion (see, e.g., [6]). These summarize each table of data describing each object for each variable in a vector or in subspace. Several mathematical problems arise at this stage:

(1) the choice of the value which summarizes the various observations of the individual for each variable, do we take the most frequent value or another value, for example an interval [7]; why this choice? and which links exist between variables,

(2) the second problem concerns the use of *homogeneity analysis* or *multiple correspondence analysis (MCA)* to reduce the data tables. We make an *MCA* for each of the n data tables describing, respectively, the n individuals. We get n factorial axis systems. To compare elements of the structure, we must seek a common system or compromise system of the n ones. This issue concerns other mathematical discipline such as differential geometry (see [8]). The proposed criteria for the choice of the common space are hardly justified (see Bouroche [9]). This problem is not fully resolved,

(3) the problem of the number of observations that may vary from one individual to another. We can use the following procedure to complete the tables. We assume that $L_i > 1$, for all $i = 1, \ldots, n$, L_i is the number of observation of the individual ω_i and define L as the least common multiple of L_i. Hence, there exists r_i such that $L = L_i \times r_i$. Now, we duplicate each table T, r_i times, we obtain a new table T_i^* of dimension $L \times d$, d is the number of variables. But if L_i is large, the least common multiple becomes large itself, and the procedure leads to the structure of large tables. Moreover, this completion removes any chronological appearance of data. This cannot be a good process of completion, and it seems more reasonable to carry out the classification without the process of completion.

To overcome all these difficulties with the proposed solutions which are not rigorously justified, we introduce a formalism borrowed from the theory of signal and communication see (Shannon [10]) and which is used to classify the elements of the data structure [11]. Our approach is based on simple statistical tools and on techniques used in information theory (physical system, entropy, conditional entropy, etc.) and requires the introduction of the concept of discrete physical systems as models for the observations of each individual for the variables which describe them. If we consider that an observation is a realization of a random vector, it appears reasonable to consider that each value of the variable or the random vector represents a state of the system which can be characterized by its frequency or its probability. If the variable is discrete, the number of states is finite, each state will be measured by its frequency or its probability. This approach gives a new version and in the same time an explanation of the distance introduced by Kullback [12]. This index makes it possible to build an indexed hierarchy on the elements of the structure and can be used if the matrix objects do not have the same dimension.

In Section 2, we introduce an adapted formalism and the notion of physical random system as a model of description of the objects. We define in Section 3 a distance between the elements of the structure. The numerical example and an application are presented in Section 4. Concluding remarks are made in Section 5.

2. Adapted formalism

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite set of n elementary objects, $\{V^1, \ldots, V^d\}$ be d discrete variables defined over Ω and taking a finite number of values in $D_1, \ldots, D_d$, respectively, $D_j = \{m_1^j, \ldots, m_{r_j}^j\}$ and m_t^j is the tth modality or value taken by V^j. We suppose that the observations of the individual ω_i for the variable V^j are given in the table

$$E_i^j = \begin{bmatrix} V^j(1) \\ V^j(2) \\ \vdots \\ V^j(l) \\ \vdots \\ V^j(L_i) \end{bmatrix}^{V^j}, \quad l = 1, \ldots, L_i, \tag{2.1}$$

and L_i represents the number of the observations of the individual ω_i, $V^j(l) = m_t^j$ if the lth observation of the individual ω_i for the the variable V^j is m_t^j where $t = 1, \ldots, r_j$. E_i^j is the vector with L_i components corresponding to the different observations of ω_i for V^j.

The structure of a juxtaposition of categorical data tables is

$$E = [E_1, \ldots, E_n] \quad \text{with } E_i = [E_i^1, \ldots, E_i^d], \tag{2.2}$$

E_i is a matrix of order $L_i \times d$. For the sake of simplicity, we transform each vector E_i^j in a $0/1$ data matrix Δ_i^j:

$$\Delta_i^j = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ l \\ \vdots \\ L_i \end{array} \begin{array}{c} \begin{array}{ccccccc} m_1^j & m_2^j & \cdots & m_t^j & \cdots & m_{r_j}^j \end{array} \\ \begin{bmatrix} 0 & 1 & \cdots & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ & & \cdots & (Z_i^j)_{lt} & \cdots & \\ \vdots & & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \end{array}, \tag{2.3}$$

$$(Z_i^j)_{lt} = \begin{cases} 1, & \text{if at the } l\text{th observation } \omega_i \text{ takes the modality } m_t^j, \\ 0, & \text{otherwise.} \end{cases}$$

The structure of a juxtaposition of $0/1$ data tables is

$$\Delta = [\Delta_1, \ldots, \Delta_n] \quad \text{with } \Delta_i = [\Delta_i^1, \ldots, \Delta_i^d], \tag{2.4}$$

Δ_i is a matrix of order $L \times M$ where $M = \sum_{j=1}^{d} r_j$.

$p_{it}^{j} = \Pr(V^{j} = m_{t}^{j})$ is estimated by the relative frequency of the value 1 observed in the tth column of the matrix Δ_{i}^{j}.

Let S_{i}^{j} be the random *single physical system* associated to ω_i for V^j:

$$S_{i}^{j} = \bigwedge_{t=1}^{r} \left\{ (S) \rightsquigarrow m_{t}^{j};\ \Pr\left[(S) \rightsquigarrow m_{t}^{j} \right] = p_{it}^{j} \right\}, \tag{2.5}$$

where the symbol $(S) \rightsquigarrow m_t$ means that the system lies in the state m_t, and $\bigwedge$ is the conjunction between events.

In the multidimensional case, the associated multiple random physical system S is

$$S = \bigwedge_{l_1=1}^{r_1} \cdots \bigwedge_{l_d=1}^{r_d} \left\{ (S) \rightsquigarrow \left(m_{l_1}^{1}, \ldots, m_{l_d}^{d} \right);\ \Pr\left[(S) \rightsquigarrow \left(m_{l_1}^{1}, \ldots, m_{l_d}^{d} \right) = p_{l_1,\ldots,l_d} \right] \right\}, \tag{2.6}$$

where

$$\sum_{l_1=1}^{r_1} \cdots \sum_{l_d=1}^{r_d} p_{l_1,\ldots,l_d} = 1. \tag{2.7}$$

The multiple random physical system associated to the marginal distributions is

$$\widehat{S} = \widetilde{\bigwedge_{j=1}^{d}} S_j, \tag{2.8}$$

where $\widetilde{\bigwedge}$ is the conjunction between single physical systems, and $\{S_j,\ j = 1, \ldots, d\}$ are the single random physical systems given by

$$\forall j = 1, \ldots, d; \quad S_j = \bigwedge_{t=1}^{r_j} \left[(S) \rightsquigarrow m_{t}^{j};\ \Pr\left[(S) \rightsquigarrow m_{t}^{j} \right] = p_{t}^{j} \right], \tag{2.9}$$

$$p_{l}^{j} = \Pr\left[(S) \rightsquigarrow m_{l}^{j} \right] = \sum_{l_1=1}^{r_1} \cdots \sum_{l_{j-1}=1}^{r_{j-1}} \sum_{l_{j+1}=1}^{r_{j+1}} \cdots \sum_{l_d=1}^{r_d} p_{l_1,\ldots,l_d}. \tag{2.10}$$

3. Distance between multiple random physical systems

3.1. Entropy as a measure of uncertainty of states of a physical system

For measuring the degree of uncertainty of states of a physical system or a discrete random variable, we use the entropy which is a special characteristic and is widely used in information theory.

3.1.1. Shannon's [10] formula for the entropy

The entropy of the system is the positive quantity:

$$H(S) = -\sum_{t=1}^{r} p_t \log_2(p_t). \tag{3.1}$$

The function H has some elementary properties which justify its use as a characteristic for measuring the uncertainty of a system.

(1) If one of the states is certain ($\exists l \in \{1,\dots,r\}$ such that $p_t = \Pr[(S) \rightsquigarrow m_t] = 1$), then $H(S) = 0$.

(2) The entropy of a physical system with a finite number of states $(m_1,\dots,m_r)$ is maximal if all its states are equiprobable: for all $t \in \{1,\dots,r\}$; $p_t = \Pr[(S) \rightsquigarrow m_t] = 1/r$. We have also $0 \le H(S) \le \log_2(r)$.

The characteristic of the entropy function expresses the fact that probability distribution with the maximum of entropy is the more biased and the more consistent with the information specified by the constraints [10].

3.2. Entropy of a multiple random physical system

Let S be a multiple random physical system given by (2.6). If the single physical systems $(S_j;\ j = 1,\dots,d)$ given by (2.9) are independent, then

$$H(S) = \sum_{j=1}^{d} H(S_j). \tag{3.2}$$

The conditional random physical system $S_1/[(S_2) \rightsquigarrow m_l^2]$ is given by

$$S_1/\left[(S_2) \rightsquigarrow m_l^2\right] = \bigwedge_{j=1}^{r_1} \left[(S_1) \rightsquigarrow m_j^1/(S_2) \rightsquigarrow m_l^2;\ \Pr\left[(S_1) \rightsquigarrow m_j^1/(S_2) \rightsquigarrow m_l^2\right] = p_{j/l}\right], \tag{3.3}$$

where $p_{j/l}$ is a conditional probability.

The entropy of this system is

$$H\left(S_1/\left[(S_2) \rightsquigarrow m_l^2\right]\right) = -\sum_{j=1}^{r_1} p_{j/l} \log_2(p_{j/l}). \tag{3.4}$$

The multiple random physical system (S_1/S_2) is written by

$$(S_1/S_2) = \bigwedge_{l=1}^{r_2} \left[\bigwedge_{t=1}^{r_1} \left[(S_1) \rightsquigarrow m_t^1/(S_2) \rightsquigarrow m_l^2;\ \Pr\left[(S_1) \rightsquigarrow m_t^1/(S_2) \rightsquigarrow m_l^2\right] = p_{t/l}\right]\right], \tag{3.5}$$

which implies

$$H(S_1/S_2) = -\sum_{l=1}^{r_2}\left[\sum_{j=1}^{r_1} p_{j/l}\log_2(p_{j/l})\right]. \tag{3.6}$$

Hence,

$$H(S) = H(S_1) + H(S_2/S_1) + H\left(S_3/S_1\widetilde{\bigwedge}S_2\right) + \cdots + H\left(S_d/S_1\widetilde{\bigwedge}S_2\widetilde{\bigwedge}\cdots\widetilde{\bigwedge}S_{d-1}\right). \tag{3.7}$$

The quantity

$$K(P,Q) = -\sum_{i=1}^{r} p_i\log_2(q_i/p_i) \tag{3.8}$$

is nonnegative. We have

$$K(P,Q) = K(Q,P) \Longleftrightarrow P = Q \text{ almost surely.} \tag{3.9}$$

It is clear that $K(\cdot,\cdot)$ is not a symmetric function, thus it is not a distance in the classical sense but characterizes (from a statistical point of view) the deviation between the distributions P and Q. It should be noted that $K(P,Q) + K(Q,P)$ is symmetrical.

Kullback [12] explains that the quantity $K(P,Q)$ evaluates the average lost information if we use the distribution P while the actual distribution is Q.

Let S_{Π_d} be a set of random physical systems with $\prod_{j=1}^{d} r_j$ states

$$S \in S_{\Pi_d} \Longrightarrow S = \bigwedge_{l_1=1}^{r_1}\cdots\bigwedge_{l_d=1}^{r_d}\left[(S) \rightsquigarrow (m_{l_1},\ldots,m_{l_{d_i}}); p_{l_1\cdots l_d}\right]. \tag{3.10}$$

Let dist be the application defined by

$$\text{dist: } S_{\Pi_d}\times S_{\Pi_d} \longrightarrow \mathbb{R}_+$$

$$\left[(S_1),(S_2)\right] \longrightarrow \text{dist}(S_1,S_2) = \left[K_d(P^1,P^2) + K_d(P^2,P^1)\right] - \left[H(S_1) + H(S_2)\right]. \tag{3.11}$$

P^1 and P^2 are the multivariate distributions of order d governing, respectively, the random physical systems S_1 and S_2. K_d is defined as follows:

$$K_d(P^1,P^2) = -\sum_{I_1}^{r_1}\cdots\sum_{I_d}^{r_d} p^{(1)}_{I_1\cdots I_d}\log_2(p^{(2)}_{I_1\cdots I_d}). \tag{3.12}$$

dist verifies

(1) $\text{dist}(S_1, S_2) \geq 0$,

(2) $\text{dist}(S_1, S_2) = 0 \Leftrightarrow S_1 = S_2$,

(3) $\text{dist}(S_1, S_2) = \text{dist}(S_2, S_1)$ (symmetry).

We admit that $\text{dist}(S_1, S_2) = 0 \Leftrightarrow S_1 = S_2 \Leftrightarrow \omega_1 = \omega_2$.

dist measures the similarity between physical systems. The smaller the value of dist is, the larger the uncertainty of the systems is. dist represents the lost quantity of average information if we use the distribution P^1 (P^2) to manage the system while the other distribution is true. dist is nothing else than the Kullback-Leibler distance between the multivariate distributions P^1 and P^2. Indeed, the Kullback-Leibler distance between P^1 and P^2 is given by

$$Ku(P^1, P^2) = \sum_{I_1}^{r_1} \cdots \sum_{I_d}^{r_d} (p^{(1)}_{I_1 \cdots I_d} - p^{(2)}_{I_1 \cdots I_d}) \log_2 (p^{(1)}_{I_1 \cdots I_d} / p^{(2)}_{I_1 \cdots I_d}). \tag{3.13}$$

Developing this expression will give dist.

4. Numerical application

4.1. Procedure to estimate the joint distribution

In the case where all variables involved in the description of the individuals are discrete, we give a procedure taken from classical techniques of factor analysis to estimate the joint distribution and derive the entropy of the multiple physical system.

Let $\Delta_i = [\Delta_i^1, \ldots, \Delta_i^d]$ be a juxtaposition of d 0/1 data tables. For $\omega_i \in \Omega$ fixed, we have

$$p^{(i)}_{l_1 \cdots l_d} = \Pr\left[(S) \rightsquigarrow (m^1_{l_1}, \ldots, m^d_{l_d})\right] = \frac{1}{L_i} N^{(i)}(l_1, \ldots, l_d). \tag{4.1}$$

$N^{(i)}(\cdot)$ is the number of simultaneous occurrences of the modalities $m^1_{l_1}, \ldots, m^d_{l_d}$:

$$\sum_{l_1=1}^{r_1} \cdots \sum_{l_d=1}^{r_d} p^{(i)}_{l_1 \cdots l_d} = 1. \tag{4.2}$$

4.2. Algorithm

We use an algorithm for ascending hierarchical classification [13]. We call points either the objects to be classified or the clusters of objects generated by the algorithm.

Step 1. There are n points to classify (which are the n objects).

Step 2. We find the two points x and y that are closest to one another according to distance dist and clustered in a new artificial point h.

Table 1: Juxtaposition of the disjunctive data tables describing the 6 objects.

		ω_1					ω_2					ω_3				
V^1		V^2			V^1		V^2			V^1		V^2				
m_1^1	m_2^1	m_1^2	m_2^2	m_3^2	m_1^1	m_2^1	m_1^2	m_2^2	m_3^2	m_1^1	m_2^1	m_1^2	m_2^2	m_3^2		
1	0	1	0	0	1	0	0	1	0	1	0	1	0	0		
0	1	0	0	1	0	1	0	0	1	0	1	0	0	1		
1	0	0	1	0	0	1	1	0	0	1	0	0	1	0		
0	1	1	0	0	0	1	1	0	0	0	1	0	1	0		
0	1	1	0	0	1	0	0	1	0	0	1	1	0	0		
1	0	0	0	1	0	1	1	0	0	1	0	0	0	1		
0	1	0	1	0	0	1	0	0	1	1	0	1	0	0		
1	0	0	1	0	1	0	0	1	0	0	1	0	1	0		
1	0	0	1	0	0	1	0	1	0	0	1	1	0	0		
1	0	1	0	0	1	0	0	0	1	1	0	0	1	0		
		ω_4					ω_5					ω_6				
V^1		V^2			V^1		V^2			V^1		V^2				
m_1^1	m_2^1	m_1^2	m_2^2	m_3^2	m_1^1	m_2^1	m_1^2	m_2^2	m_3^2	m_1^1	m_2^1	m_1^2	m_2^2	m_3^2		
1	0	1	0	0	1	0	0	1	0	1	0	1	0	0		
0	1	0	1	0	0	1	1	0	0	1	0	0	0	1		
1	0	0	0	1	1	0	0	0	1	0	1	0	1	0		
0	1	1	0	0	1	0	0	1	0	1	0	1	0	0		
0	1	0	1	0	0	1	0	1	0	0	1	0	0	1		
1	0	0	1	0	1	0	0	1	0	0	1	0	0	1		
1	0	0	1	0	0	1	0	0	1	0	1	1	0	0		
1	0	0	0	1	1	0	0	1	0	1	0	1	0	0		
0	1	0	0	1	1	0	0	1	0	1	0	0	1	0		
1	0	1	0	0	1	0	1	0	0	1	0	1	0	0		

Step 3. We calculate the distances between the new point and the remaining points using the single linkage of Sneath and Sokal [14] D defined by

$$D(\omega, h) = \text{Min}\left\{\text{dist}(\omega, x), \text{dist}(\omega, y)\right\}, \quad \omega \neq x, y. \tag{4.3}$$

We return to Step 1 with only $(n-1)$ points to classify.

Step 4. We again find the two closest points and aggregate them. We calculate the new distances and repeat the process until there is only one point remaining.

In the case of single linkage, the algorithm uses distances in terms of the inequalities between them.

4.3. Numerical example

Consider 6 individuals described by 2 qualitative variables with, respectively, 2 and 3 modalities. 10 observations for each individual, the observations are grouped in Table 1.

Table 2

	(m_1^1, m_1^2)	(m_1^1, m_2^2)	(m_1^1, m_3^2)	(m_2^1, m_1^2)	(m_2^1, m_2^2)	(m_2^1, m_3^2)
F_1	0.2	0.3	0.1	0.2	0.1	0.1
F_2	0.1	0.2	0.1	0.3	0.1	0.1
F_3	0.2	0.2	0.1	0.2	0.2	0.1
F_4	0.2	0.2	0.2	0.1	0.2	0.1
F_5	0.1	0.4	0.2	0.1	0.1	0.1
F_6	0.4	0.1	0.1	0.1	0.1	0.1

Table 3: Entropy of the conditional random physical systems associated to the 6 objects.

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	2,4464	2,6049	2,5219	2,6219	2,6219	2,8219
S_2	2,6049	2,4464	2,6219	2,8219	2,8219	2,9219
S_3	2,6049	2,7049	2,5219	2,6219	2,8219	2,8219
S_4	2,7049	2,8634	2,6219	2,5219	2,7219	2,8219
S_5	2,4879	2,6634	2,6219	2,5219	2,3219	3,0219
S_6	2,6634	2,8634	2,6219	2,6219	3,0219	2,3219

4.3.1. Procedure to build a hierarchy on these objects

The empirical distributions which represent the individuals are given by Table 2.

The program is carried out on this numerical example. We obtain the following results (Table 3).

Step 1. From the similarity matrix, using the single linkage of Sneath (4.3), we obtain

$$\min_{l\neq t;\, l,t=1,\ldots,5}\left\{\operatorname{dist}\left(S_l, S_t\right)\right\} = \operatorname{dist}\left(S_1, S_3\right) = 0,1585. \tag{4.4}$$

Then, the objects ω_1 and ω_3 are aggregated into the artificial object ω_7 which is placed at the last line, and the rows and columns corresponding to the objects ω_1 and ω_3 are removed in the similarity matrix.

Step 2. From the new similarity matrix, we obtain

$$\min_{l\neq t;\, l,t=2,4,5,6}\left\{\operatorname{dist}\left(S_l, S_t\right)\right\} = \operatorname{dist}\left(S_2, S_7\right) = 0,317. \tag{4.5}$$

The objects ω_2 and ω_7 are aggregated into the artificial object ω_8.

Step 3.

$$\min_{l\neq t;\, l,t=4,5,6}\left\{\operatorname{dist}\left(S_l, S_t\right)\right\} = \operatorname{dist}\left(S_4, S_8\right) = 1,268. \tag{4.6}$$

The objects ω_4 and ω_8 are aggregated into the artificial object ω_9.

Step 4.

$$\min_{l\neq t;\ l,t=5,6,9}\{\operatorname{dist}(S_l,S_t)\}=\operatorname{dist}(S_5,S_9)=2,732. \tag{4.7}$$

The objects ω_5 and ω_9 are clustered in the new object ω_{10}. The object ω_6 is aggregated with the object ω_{10} and $\operatorname{dist}(S_6,S_{10})=4,8$.

In Figure 1 it can be seen that two separated classes appear in the graph by simply cutting the hierarchy on the landing above the individual ω_2. In this algorithm, we started by incorporating the two closest objects using the index of distance between corresponding physical systems. The higher the construction of the hierarchy is, the more dubious the obtained states of the mixed system are. The example shows that the index of Kullback-Leibler and the index of aggregation of the minimum bound (single linkage) lead to the construction of a system with a maximum of entropy, and thus lead to a system for which all the states are equiprobable.

If the total number of modalities of the various criteria is large compared with the number of observations, the frequency of choosing a set of modalities becomes small, and a lot of frequencies are zero. The set of modalities whose frequency is zero will be disregarded and does not intervene in the calculation of the distances. This can lead to the impossibility of comparing the systems.

4.3.2. Classification of the six objects after reduction

If each object is described by the highest frequencies "mode," we obtain the following table:

$$\begin{array}{c} \\ \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \\ \omega_6 \end{array}\begin{array}{c} \begin{array}{cc} V^1 & V^2 \end{array} \\ \begin{bmatrix} m_1^1 & m_2^2 \\ m_2^1 & m_3^2 \\ m_1^1 & m_1^2 \\ m_1^1 & m_1^2 \\ m_1^1 & m_2^2 \\ m_1^1 & m_2^2 \end{bmatrix} \end{array}. \tag{4.8}$$

This table contradicts the fact that in our procedure, the objects ω_1 and ω_2 are very close while ω_1 and ω_5 are not the same. This shows that the classification after reduction, for this type of data, can lead to contradictory results.

4.4. Application

The data come from a survey concerning the level of development of n departments $E_1, E_2, \ldots, E_n$ of a country. The aim is to search the less developed subregions in order to establish programs of adequate development. Every department E_i is constituted of L_i subregions $C_1^i, C_2^i, \ldots, C_{L_i}^i$. For every $i=1,\ldots,n$ and $l=1,\ldots,L_i$, we measured the composite economic development index idE and the composite human development index idH. These two composite indices are weighted means of variables measuring the degree of economic

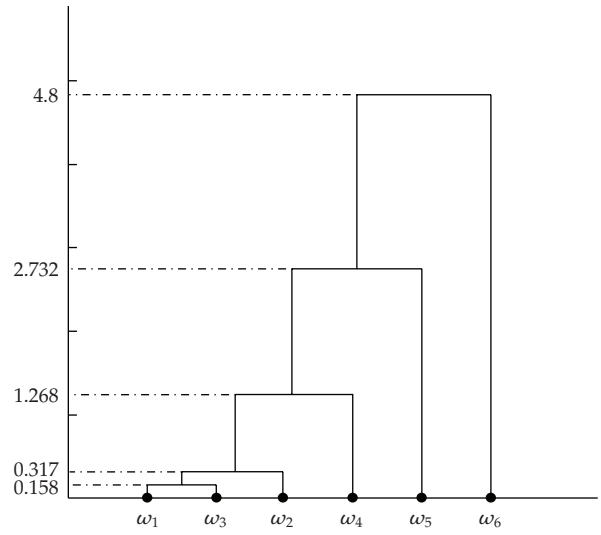

Figure 1: Dendogram of the produced hierarchy.

and human development, developed by experts of the program of development of the United Nations for ranking countries. These indices depend on the geographical situation and on the specificity of the subregions.

For every $i = 1,\dots,n,\ l = 1,\dots,L_i,\ 0 \le idE(C_l^i) \le 1$, and $0 \le idH(C_l^i) \le 1$. The closer to 1 the value of the index is, the more the economic or human development is judged to be satisfactory. However, these indices are not calculated in the same manner. They depend on whether the subregions are classified as farming or urban zone. The ordering of the subregions according to each of the indices do not have sense anymore. The structure of data in entry is for every $i = 1,\dots,n$:

$$\underset{(L_i,2)}{E_i} \longmapsto \begin{array}{c} \\ C_{i1} \\ C_{i2} \\ \vdots \\ C_{iL_i} \end{array} \begin{array}{c} \begin{array}{cc} idE & idH \end{array} \\ \begin{bmatrix} idE(C_{i1}) & idH(C_{i1}) \\ idE(C_{i2}) & idH(C_{i2}) \\ \vdots & \vdots \\ idE(C_{iL_i}) & idH(C_{iL_i}) \end{bmatrix} \end{array}. \tag{4.9}$$

The structure is not exploitable in this form. It is therefore necessary to transform the tables in a more tractable form. The specialists of the programs of development cut the observations of each index in quintile intervals and affect each of the subregions to the corresponding quintile. We thus determine for the n series of observations of the two indices the various corresponding quintiles:

$$\begin{aligned} idE &\longrightarrow q_{i1}^1, q_{i2}^1, \dots, q_{i5}^1, \\ idH &\longrightarrow q_{i1}^2, q_{i2}^2, \dots, q_{i5}^2. \end{aligned} \tag{4.10}$$

The quintile intervals are

$$I_{i1}^1, I_{i2}^1, \ldots, I_{i5}^1,$$
$$I_{i1}^2, I_{i2}^2, \ldots, I_{i5}^2. \tag{4.11}$$

For every $i = 1, \ldots, n$, the table E_i is transformed into a table of $0/1$ data.

The problem is to build a hierarchy on all departments of the territory in order to observe the level of development of each of the subregions according to the two indices and thus to make comparisons. The observations are summarized in tables $\Delta_1, \Delta_2, \ldots, \Delta_n$ which constitute a structure of juxtaposition of $0/1$ data matrices. These data presented are from a study of 1541 municipalities involved in Algeria. The municipalities are gathered in 48 departments. The departments do not have the same number of municipalities which have not the same specificities: size, rural, urban, and the municipalities do not have the same locations: mountain, plain, costal, and so forth. We have to build typologies of departments according to their economic level and human development according to the United Nations Organization standards.

The result of the study made it possible to gather the great departments (cities) which have large and old universities and the municipalities which have a long existence. Another group emerged which includes enough new departments of the last administrative cutting and develops activities and services of small and middle companies. The other groups are distinguished by great disparities between municipalities in their economic level and human development and according to surface and importance.

5. Conclusion

In this paper, the definition of the entropy is that stated by Shannon [10]. This definition is still used in the theory of signal and information. The suggested formalism gives an explanation and a practical use of the distance of Kullback-Leibler as an index of distance between representative elements of a structure of tables of categorical data. It is possible to extend these results to the case of a structure of data tables of measurements and to adapt an algorithm of classification to the case of functional data.

References

[1] I. C. Lerman and B. Tallur, "Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence," *Revue de Statistique Appliquée*, vol. 28, no. 3, pp. 5–28, 1980.
[2] A. Rebbouh, "Clustering the constitutive elements of measuring tables data structure," *Communications in Statistics: Simulation and Computation*, vol. 35, no. 3, pp. 751–763, 2006.
[3] A. Agresti, *Categorical Data Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.
[4] I. T. Adamson, *Data Structures and Algorithms: A First Course*, Springer, Berlin, Germany, 1996.
[5] J. Beidler, *Data Structures and Algorithms*, Springer, New York, NY, USA, 1997.
[6] T. W. Anderson and J. D. Jeremy, *The New Statistical Analysis of Data*, Springer, New York, NY, USA, 1996.
[7] F. de A. T. de Carvalho, R. M. C. R. de Souza, M. Chavent, and Y. Lechevallier, "Adaptive Hausdorff distances and dynamic clustering of symbolic interval data," *Pattern Recognition Letters*, vol. 27, no. 3, pp. 167–179, 2006.
[8] P. Orlik and H. Terao, *Arrangements of Hyperplanes*, vol. 300 of *Grundlehren der Mathematischen Wissenschaften*, Springer, Berlin, Germany, 1992.

[9] J. M. Bouroche, *Analyse des données ternaires: la double analyse en composantes principales*, M.S. thesis, Université de Paris VI, Paris, France, 1975.
[10] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
[11] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," *Journal of Classification*, vol. 13, no. 2, pp. 195–212, 1996.
[12] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, New York, NY, USA, 1959.
[13] G. N. Lance and W. T . Williams, "A general theory of classification sorting strategies—II: clustering systems," *Computer Journal*, vol. 10, pp. 271–277, 1967.
[14] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, A Series of Books in Biology, W. H. Freeman, San Francisco, Calif, USA, 1973.