Publications Internes de l'IRISA
ISSN : 2102-6327
PI 1988 – Janvier 2012

Facets of the set theoretic representation of categorical data

Israël-César LERMAN*

Abstract: There are three basic notions in *Data Analysis* : object, category and descriptive attribute. Two description cases are of concern ; describing an object set $\mathcal{O}$ or describing a category set $\mathcal{C}$. For the former case the set theoretic representation of a given attribute is defined with respect to $\mathcal{O}$ and for the latter, it is defined within each category of $\mathcal{C}$. For these representations a given descriptive attribute is interpreted in terms of a relation with a given arity defined on the described set. We propose in this work a unique principle for the set theoretic representation of a descriptive attribute of different types : boolean, numerical, nominal categorical, ordinal categorical, preordonance categorical, taxonomic categorical and taxonomic preordonance categorical. The formal representations of these different types are explicited and compared in both cases : for an object set description and for a category set description. Then these representations are applied in interesting real examples.

Key-words: Categorical attribute, Relational description, Category set, Preordonances and Taxonomies

---

Aspects de la représentation ensembliste des données catégorielles

Résumé : *Trois notions basiques interviennent en* Analyse des Données *: la notion d'objet, celle de catégorie et celle d'attribut descriptif (on dit encore variable descriptive). La description peut concerner un ensemble d'objets ou un ensemble de catégories. Dans le premier cas, l'attribut est représenté par rapport à l'ensemble des objets qu'il décrit et dans le second cas, il est représenté au sein de chacune des catégorie. L'attribut est interprété dans les termes d'une relation d'arité donnée sur l'ensemble des objets qu'il décrit. Nous proposons dans ce travail de synthèse un principe unique de représentation ensembliste des attributs de description de différents types : booléen, numérique, catégoriel (on dit aussi qualitatif) nominal, catégoriel ordinal, catégoriel préordonnance, catégoriel taxonomique et catégoriel préordonnance taxonomique. Nous explicitons et nous comparons les représentations formelles de ces différents types d'attributs ; d'une part, dans le cas de la description d'un ensemble d'objets et d'autre part, dans le cas de la description d'un ensemble de catégories.*

Mots clés : *Attribut catégoriel, Description relationnelle, Ensemble de catégories, Préordonnances et Taxonomies*

---

* Irisa - Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cédex — lerman@irisa.fr

1 Introduction

We show in this report how the set theoretic representation of descriptive attributes allows to cover a very large scope of data description in *Data Analysis*. For this representation the attributes are interpreted in terms of relations on the described object set. For a given descriptive attribute, the arity of the associated relation is defined by the structure endowing the value set of this attribute. A typology of the descriptive attributes based on the mentioned structure is proposed. In this typology the following types are defined : boolean, numerical, nominal categorical, ordinal categorical, preordonance categorical, taxonomic categorical and taxonomic preordonance categorical. We will study below these different types and we will mutually compare them. The categorical attributes will play a very important part in our analysis. A value of a categorical attribute is called "category". Conditions of exclusivity and exhaustivity are required for the different categories of a categorical attribute (see Section 3.1). A category is determined by the intuitive notion of "concept".

Formal definition, mathematically expressed, of a "concept" in real life is very difficult Sutcliffe (1992) [45]. It depends on a knowledge domain and on recognition techniques. For example, let us consider the cirrhosis concept defined in the hepatho-biliary pathology. For this pathology, we assume a universe $\mathcal{U}$ of liver ill persons. $\mathcal{U}$ is a real or hypothetical finite set and each of its elements defines an elementary and indivisible object interesting the domain studied. On a given element of $\mathcal{U}$, the concept may be $TRUE$ or $FALSE$. A series of clinical tests are necessary in order to recognize if a given person is a cirrhosis ill. A concept is defined in "intension" (one can say in "comprehension") from a precise description using primitive concepts. It is detected by the domain expert for two reasons : first, its frequency is not negligible ; and second, its consequences are globally comparable on all the objects where it is observed.

Let us denote by γ such a concept, for example the "liver cirrhosis". γ is defined with respect to the universe $\mathcal{U}$ of the ill liver persons. γ is represented by the subset of $\mathcal{U}$, that we denote by $\mathcal{U}(\gamma)$, whose disease is cirrhosis. Equivalently, a boolean attribute corresponds bijectively to the concept γ. This attribute, denoted by a_γ, is a mapping of $\mathcal{U}$ onto the set $\{0,1\}$ comprising two codes 0 and 1 :

$$a_\gamma : \mathcal{U} \to \{0,1\} \tag{1}$$

For a given u in $\mathcal{U}$, $a_\gamma(u) = 1$ (resp., 0) if γ is $TRUE$ (resp., $FALSE$) on u. In these conditions, the above subset $\mathcal{U}(\gamma)$ of $\mathcal{U}$ can be expressed as follows :

$$\mathcal{U}(\gamma) = a_\gamma^{-1}(1) \tag{2}$$

where $a_\gamma^{-1}(1)$ is the reciprocal image of 1.
$\mathcal{U}(\gamma)$ is called "extension" of the concept γ at the level of $\mathcal{U}$.

The introduced concept notion is boolean. However, its definition may require non-boolean attributes. Consider the following example taken in the framework of an epidemiological survey concerning a male adult population living in a given area and aged 18-60 years. The concept "smoking more than 20 cigarettes a day" is a boolean concept. Nevertheless, its definition requires a non-boolean attribute defined by "counting". The scale value of this attribute is the set of integer numbers.

The notion of a value scale associated with a descriptive attribute is preliminary to define descriptions in *Data Analysis* Suppes and Zinnes (1963) [44]. A given descriptive attribute a is defined at the level of a universe $\mathcal{U}$ of objects concerned by a set of concepts (see the above example where the concepts are those of the hepatho-biliary pathology). By denoting $\mathcal{E}$ the value scale of a, a is mathematically interpreted as a mapping of $\mathcal{U}$ in $\mathcal{E}$, associating with each element u of $\mathcal{U}$, a unique value in $\mathcal{E}$ denoted by $a(u)$.

$$\begin{aligned} a &: \mathcal{U} \to \mathcal{E} \\ u &\mapsto a(u) \end{aligned} \tag{3}$$

Generally, in *Data Analysis* and in *Machine Learning* the whole set $\mathcal{U}$ is not available. We only dispose a finite set $\mathcal{O}$ of objects representing $\mathcal{U}$. Mostly, $\mathcal{O}$ is a subset of $\mathcal{U}$, $\mathcal{O} \subset \mathcal{U}$. The observed results of a data analysis at the level of $\mathcal{O}$ are infered at the level of $\mathcal{U}$. For this purpose, $\mathcal{O}$ has to be a statistical representative sample of $\mathcal{U}$. Usually, this condition is statisfied, and specially in *Data Mining* where the size of $\mathcal{O}$ is generally very large (many millions or even more) Therefore, for the following, our reference object set will be $\mathcal{O}$. Thus, the above attribute a will be regarded as a mapping of $\mathcal{O}$ into $\mathcal{E}$:

$$a : \mathcal{O} \to \mathcal{E}$$
$$u \mapsto a(u) \tag{4}$$

Now, let us define the scale $\mathcal{E}$ under the form $\mathcal{E} = \{e_1, e_2, ..., e_h, ..., e_k\}$ where e_h is one possible value of the attribute a, $1 \leq h \leq k$. The meaning of the studied field enables to define a structure on the value set $\mathcal{E}$. This structure induces a mathematical relation on $\mathcal{O}$. In Sections 2 and 3 we shall detail the different types of descriptive attributes and their formal representations at the level of the object set $\mathcal{O}$. The main type of a descriptive attribute is defined from the arity of the relation on $\mathcal{O}$, induced by the structure of the value set $\mathcal{E}$. We distinguish three main types I, II and III. For I the induced relation on $\mathcal{O}$ is unary. It comprises the boolean attribute and the numerical one (Sections 2.1 and 2.2). For II, the relation is binary, non-valuated or valuated. This type includes the nominal categorical attribute (Section 3.1), the ordinal categorical attribute (Section 3.2), the ranking attribute (Section 3.3) and the categorical attribute valuated by a numerical similarity (Section 3.4). These two types (I and II) cover a large range of formal descriptions in combinatorial data analysis (see Sections 2, 3, 5 and 6). Type III is defined when the attribute scale induces a binary relation - generally defined by a ranking - on the set $P_2(\mathcal{O})$ (resp., $\mathcal{O} \times \mathcal{O}$) of unordered (resp., ordered) pairs of elements of $\mathcal{O}$. Then, the *preordonance categorical attribute*, the *taxonomic attribute* and the *taxonomic preordonance attribute* are presented in Sections 4.1, 4.2 and 4.3, respectively. In Section 4.4 preordonance representations of the different types of descriptive attributes are proposed. Section 4 is devoted for defining the representation of the different types of attributes when describing a set of categories instead of a set of objects.

Relational representation of descriptive attributes is clearly emphasized in our work. One of its specificities consists in highlighting the set theoretic representation sustaining the relational one Lerman (1970) [17], (1973) [18], (1981) [19], (1992) [21], [22] and (2009) [27]. In our work we were very influenced by the M.G. Kendall work Kendall (1948) [12]. Several authors in *Combinatorial Data Analysis* interpret, implicitly or explicitly, the representation of a descriptive attribute of an object set $\mathcal{O}$ in terms of of a binary relation on $\mathcal{O}$. This is generally done for a given type of descriptive attribute [e.g. nominal (resp., ordinal) categorical attribute], in relation with a specific method to be developed Guénoche and Monjardet (1987) [10], Hubert (1987) [11], Marcotorchino and Michaud (1979) [36], Marcotorchino (2009) [35], Giakoumakis and Monjardet (1987) [8], Régnier (1965) [40]. In the following, we wish to present a general framework independent of a given methodology, in which the different types of decriptive attributes and their formal representations will be expressed.

2 Representation of the attributes of type I

As mentioned above, an attribute of type I induces a unary relation on $\mathcal{O}$. Such an attribute can be called an "incidence" attribute. For this type we distinguish exactly the "boolean" attribute and the "numerical" one.

2.1 The boolean attribute

This attribute is also called a "presence-absence" attribute. It has been already considered above in Section 1. As previously, let us denote it by a. Formally, a is a mapping of $\mathcal{O}$ onto the set $\{FALSE, TRUE\}$. Mostly, in *Data Analysis*, $FALSE$ and $TRUE$ are coded by the integer numbers 0 and 1, respectively. For the following mapping diagram, also considered above in a different context,

$$a : \mathcal{O} \to \{0, 1\}$$
$$x \mapsto a(x) \tag{5}$$

$a(x)$ denotes the value of a on x, $a(x) = 1$(resp., 0) if a is $TRUE$ (resp., $FALSE$) for x.
a is represented by the subset $\mathcal{O}(a)$ of $\mathcal{O}$ constituted by the objects for which a is $TRUE$:

$$\mathcal{O}(a) = a^{-1}(1) \tag{6}$$

where a^{-1} denotes the reciprocal mapping of a.
Now, let us introduce the cardinalities of $\mathcal{O}$ and $\mathcal{O}(a)$ that we denote by n and $n(a) : n = card(\mathcal{O})$ and $n(a) = card(\mathcal{O}(a))$. Thus, the proportion or relative frequency $p(a)$ of objects for which a is $TRUE$ is defined by $p(a) = n(a)/n$.
The negated boolean attribute $\bar{a}$ is defined by

$$(\forall x \in \mathcal{O}), \bar{a}(x) = TRUE \text{ if and only if } a(x) = FALSE \tag{7}$$

Clearly, with $\bar{a}$, is associated $\mathcal{O}(\bar{a})$, which is the complementary subset of $\mathcal{O}(a)$ in $\mathcal{O}$. We can also define $n(\bar{a}) = card(\mathcal{O}(\bar{a}))$ and
$p(\bar{a}) = n(\bar{a})/n$. Trivially, we have $p(a) + p(\bar{a}) = 1$.

The couple $\{a, \bar{a}\}$ is the value set of a binary catégorical attribute. Let us denote this attribute by α. The empirical distribution of α is defined by the ordered pair $[p(a), p(\bar{a})]$.

Thus, with logically independent boolean attributes [1], binary categorical attributes can be associated. Conversely, with categorical binary attributes, boolean attributes can be associated by retaining for each binary categorical attribute, one of its two possible values. Generally and for significance statistical reasons, the retained value is the least frequent among the both values.

Boolean attributes occur very frequently in database descriptions. For the above example in Section 1 where $\mathcal{O}$ is defined by a sample of liver ill persons, the concept of liver cirrhosis specifies a boolean attribute.

2.2 The numerical attribute

The value scale of a numerical attribute is the set $\mathbb{R}$ of real numbers. Mostly, in *Data Analysis*, numerical descriptions are considered for measuring quantities (e.g. weight, size, ...). In these conditions, we shall assume a positive scale including 0 for the numerical attribute. In any case, this does not restrict the generality. Indeed, by representing geometrically $\mathbb{R}$ with an horizontal axis directed from left to right, the scale origin can be moved to left in order to make positive the observed values of the numerical attribute on the object set $\mathcal{O}$. Let us denote by v the numerical attribute. v is a mapping of $\mathcal{O}$ on the set that we denots by $\mathbb{R}_+$ of real positive numbers

$$v : \mathcal{O} \to \mathbb{R}_+ \tag{8}$$

associating with each object o of $\mathcal{O}$, a positive real number $v(o)$

$$(\forall o \in \mathcal{O})\ o \mapsto v(o) \tag{9}$$

v is interpreted as a valuated unary relation :

$$\{v(o) | o \in \mathcal{O}\} \tag{10}$$

Thereby, boolean attribute and numerical attribute are considered at the same relational level. Indeed, their extensions are represented in a set theoretic way at the level of the object set $\mathcal{O}$: a subset of $\mathcal{O}$ for the boolean attribute, and a numerical valuation on $\mathcal{O}$ for the numerical attribute.

Taking into account the accuracy of measurement, the scale of decimal positive numbers is largely sufficient for defining a numerical attribute in data analysis. Therefore, by denoting $\mathbb{D}_+$ the latter scale, $\mathbb{D}_+$ can be substituted for $\mathbb{R}_+$ in the right member of 8.

As mentioned above, generally and more specially in *Data Mining* the size of the object set $\mathcal{O}$ is very large. Consequently, the size of the reached values by v on $\mathcal{O}$ is much smaller than the cardinality of $\mathcal{O}$. Let us designate by

$$[x_{(1)}, x_{(2)}, ..., x_{(l)}, ..., x_{(m)}] \tag{11}$$

the increasing ordered sequence of the reached values by v on $\mathcal{O}$ $(m < n)$ and introduce the subset $\mathcal{O}_l = v^{-1}(x_{(l)})$ constituted by all objects whose v value is $x_{(l)}$, $1 \leq l \leq m$. Denote $n_l = card[\mathcal{O}_l]$, $1 \leq l \leq m$. The empirical distribution of v on $\mathcal{O}$ is defined by the sequence

[1] (i.e. any attribute can be derived from the others)

$$\{(x_{(l)}, f_l)\} \tag{12}$$

where $f_l = n_l/n$, f_l defines the relative frequency (proportion) of objects whose v value is x_l, $1 \leq l \leq m$.
As an example, let us consider the following increasing sequence of a numerical attribute v on a set of 10 objects

$$(1.5, 2.3, 3.4, 3.4, 3.4, 5.1, 7.2, 7.2, 8.5, 9.0)$$

We suppose from the accuracy measurement of v, only one decimal after the point. The above distribution (12) becomes

$$[(1.5, 0.1), (2.3, 0.1), (3.4, 0.3), (5.1, 0.1), (7.2, 0.2), (8.5, 0.1), (9.0, 0.1)] \tag{13}$$

Thus, there are $m = 7$ distinct values.
In the geometrical methods of data analysis, a numerical attribute is represented by a *linear form*. This gives the projection measurement on a linear axis endowed with an origin and a unit vector. To fix idea, we can assume an horizontal axis. In these conditions, a given object o is represented by a point of the axis whose abscissa is $v(o)$. By considering the sequence (11) of the v values, n_l distinct objects are represented by the same point that we denote by M_l, whose abscissa is x_l, $1 \leq l \leq m$. A graphical representation, called "histogram" is obtained by drawing from each point M_l, $1 \leq l \leq m$, an ascendant vertical segment whose length is proportional to the relative frequency f_l, $1 \leq l \leq m$. Thus we obtain for the distribution (13) (see Figure 2.1)

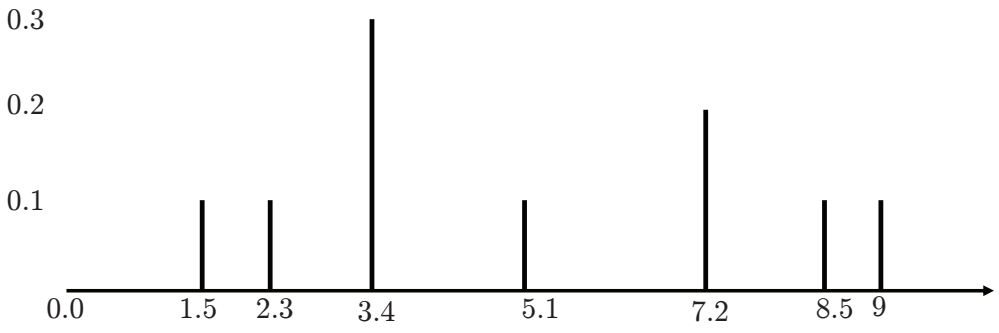

FIG. 1 – Histogram

In terms of factorial analysis the sequence

$$\{(M_l, f_l)\} \tag{14}$$

defines one dimensional cloud of points where the point M_l is endowed with the positive numerical value f_l, the latter being interpreted as a weight, $1 \leq l \leq m$.

This structure was exploited in image scalar quantization Ghazzali (1992) [7], Ghazzali, Léger and Lerman (1994) [6].In this application the concerned numerical attribute v is the luminance for which a scale of 256 grey levels, from 0 to 255, is established : 0 for the black and 255 for the blank. Here, the object set $\mathcal{O}$ is defined by the image pixels. Thus, for a squared image comprising 512 rows and 512 columns, the object set $\mathcal{O}$ includes $n = 512 \times 512 = 262144$ elements. The one dimensional cloud (14) has at most 256 points. It can be put in the following form :

$$\{(l, f_l)|0 \leq l \leq 255\} \tag{15}$$

where $f_l = n_l/n$ is the proportion of image pixels whose luminance is l, $0 \leq l \leq 255$.

2.3 Defining a categorical attribute from a numerical one

As seen above, in the framework of the attributes of general type I, corresponding to a unary relation on the object set $\mathcal{O}$, we distinguish two cases (see Sections 1 and 2). The first one is defined by the boolean attribute. The scale value of the latter is the poorest one. The second one is defined by the numerical attribute. Its scale value is the richest one.

There are several reasons for transforming a numerical attribute into a categorical one. One important reason may be to make homogeneous the description of the object set $\mathcal{O}$. Indeed, imagine that the vast majority of the description is provided by categorical attributes. To fix idea, we assume a small number of categories by categorical attribute. In addition, suppose that few descriptive attributes are numerical. In these conditions it might be appropriate to transform the numerical attributes into categorical ones (see the following Sections 3.1 and 3.2). Another important reason related to the above first one, consists in retaining from the numerical value $v(o)$ of v on the object o, what can be significant in this value. In fact, knowing the exact value $v(o)$ might be less interesting than knowing that $v(o)$ is located inside a given interval. Indeed, in order to categorize a numerical attribute, its interval variation is divided into subintervals.

Precisely, regarding the sequence $[x_{(1)}, x_{(2)}, ..., x_{(l)}, ..., x_{(m)}]$ (see equation (11)), let us denote here by a, $x_{(1)}$ and by b, $x_{(m)} + \epsilon$, where ϵ is an arbitrary small positive number. Therefore, the attribute v takes its values in the interval $[a, b[= \{x|x \in \mathbb{R}_+, a \leq x < b$, where, as above, $\mathbb{R}_+$ denotes the real positive numbers. An increasing sequence, denoted by σ :

$$\sigma = (y_0 = a, y_1, ..., y_k, y_{k+1}, ..., y_l = b) \tag{16}$$

defines a subdivision of the interval $[a, b[$ into a sequence of l subintervals

$$\left\{[y_k, y_{k+1}[|0 \leq k \leq l-1\right\} \tag{17}$$

With each subinterval $[y_k, y_{k+1}[$, $0 \leq k \leq l-1$, is associated a category of a categorical attribute (see Section 3). By denoting c_k the category representing the interval $[y_k, y_{k+1}[$, $0 \leq k \leq l-1$, c_k is defined by the following boolean attribute

$$(\forall o \in \mathcal{O}), c_k(o) = 1(\text{resp. } 0) \text{ iff } v(o) \in [y_k, y_{k+1}[(\text{resp. } v(o) \notin [y_k, y_{k+1}[) \tag{18}$$

Now, the fundamental question is

"How to define the subdivision σ ?"

There are a large variety of methods depending each on a given type of purpose. The simplest and the most direct one consists in defining σ by dividing the entire interval $[a, b[$ into subintervals with the same length. By denoting L the length of the interval $[a, b[: L = b - a$, we have

$$y_{k+1} = y_k + \frac{L}{l} \tag{19}$$

$0 \leq k \leq l-1$. Notice that for a given practical application this technique requires an adequate value of l. On the other hand, for this technique, the statistical distribution of the attribute v on the object set $\mathcal{O}$ is not taken into account at all.

Another simple technique consists in interval division with as equal frequencies as possible. A given interval will be written $[x_i, x_{i'}[$ where $i' > i$ and where $n_i + ... + n_{i'}$ is adjusted to be as close as possible the ratio n/l. For this solution, in order to be consistent with the statistical distribution of the attribute v, we have also to determine in a non-arbitrary way an adequate value of l.

None of both previous techniques respect intimately the heterogeneity of the statistical distribution of v. Nevertheless, it might exist subintervals of $[a, b[$ such that two consecutive and near values of these, have comparable and high frequencies. Discovering such subintervals leads to the definition of "significant" categorical attribute (see equation (18)). Thus in Figure 2 we recognize intuitively five subintervals.

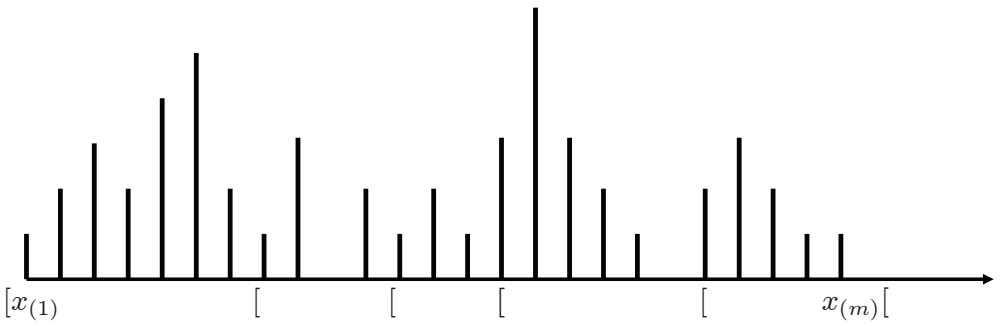

FIG. 2 – Histogram-2

These subintervals correspond to high concentration zones. Precisely, clustering methods enable the discovering of such subintervals. In these latter methods an objective criterion is optimized, locally or globally.

The interest of the well known Fisher method Fisher(1958) [5] is to maximize *globally* the inter-cluster inertia in case where the number of clusters is fixed. Each cluster is represented by an interval of the sought subdivision. The method is based on a dynamic programming algorithm.

Lafaye (1979) [14] analyzed and made effective a graphical method that we suggested the general idea. Let us consider the number of observations in a *small* interval window having a fixed length. This number defines a kind of "local density". By moving the window from left to right along the variation range of v, we can determine the stable minima of this local density. This stability is obtained by varying in a suitable fashion the length of the window interval. The stable local minima define the cut points of the sought subdivision. For this method the number of the subdivision intervals is not fixed. It is a resulting of the employed algorithm. This method shown to be particularily effective for the treatment of small samples Kerjean(1978) [13]. In his paper [14] Lafaye gives a brief and interesting synthesis on the dicretization of numerical attributes, With this respect, see also Rabaseda et al. (1995) [39].

In vector quantization (see the above example at the end of Section 2, considered in scalar quantization) Ghazzali (1992) [7], Ghazzali, Léger and Lerman (1994) [6] hierarchical ascendant (or agglomerative) clustering was applied. Two methods were employed : the Likelihood Linkage Analysis method Lerman (1993) [23] and the classical Ward method Ward (1963) [49]. In this case and for coding compression purpose the order of the number of clusters is given (for example 16). Preliminary,

it is established that each step of the agregation algorithm joins two consecutive intervals of (11) obtained previously as clusters in the hierarchical process. A method of detecting the most consistent levels of a classification tree (see Chapters 4 of [19] and [23]) allows to obtain the wished subdivision.

A non-hierarchical dynamic and adaptative clustering method in the spirit of the K-means one, can also be applied in order to determine a subdivision σ. The method built by A. Schroeder (1976) [41] is based on a statistical criterion having a probabilistic nature : *the likelihood classifying criterion.*

For this criterion the observed cloud of points - one dimensional in our case - is considered as provided from a mixing of probability laws, having distinct modes and distinct probability occurences, respectively, Symons (1981) [46]. Celeux and Govaert (1993) [3] employ a stochastic algorithm for separating the entire cloud into subclouds corresponding each to one of the probability laws. This separation enables to detect an ordered sequence of intervals dominated each by one of the probability laws. And then, a subdivision σ of the variation range of v can be deduced. This approach may lead to fine and sophisticated techniques. However, these assume probabilistic conditions on the shapes of the probability laws, difficult to validate in general. On the other hand, these methods may seem conceptually too heavy for the submitted problem of discretization of an observed numerical attribute.

3 Representation of the attributes of type II

As mentioned in Section 1, an attribute of type II induces a binary relation on $\mathcal{O}$. In this section we shall consider three sub-types of attributes. The nominal categorical attribute, the ordinal categorical attribute and the categorical attribute valuated by a numerical similarity.

3.1 The nominal categorical attribute

Let $\mathcal{C} = \{c_1, c_2, ..., c_h, ..., c_k\}$ be the value set of a categorical attribute c. No structure is assumed on $\mathcal{C}$. The attribute c is defined by a mapping

$$c : \mathcal{O} \rightarrow \{c_1, c_2, ..., c_h, ..., c_k\} \tag{20}$$

of $\mathcal{O}$ on $\mathcal{C}$, such that for x in $\mathcal{O}$, $c(x) = c_h$ if and only if x possesses the value c_h, $1 \leq h \leq k$. $c_1, c_2, ..., c_h, ...$ and c_k are called the categories of the categorical attribute c. The value set $\mathcal{C}$ is assumed *exclusive* and *exhaustive*. That is to say : each object possesses necessarily and exactly one and only one categorical value. The attribute c induces a partition on $\mathcal{O}$ that we denote by

$$\pi = \{\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_h, ..., \mathcal{O}_k\} \tag{21}$$

where $\mathcal{O}_h$ is defined by $c^{-1}(c_h)$ that is the reciprocal image of c_h, $1 \leq h \leq k$. It is easier and without loss of generality to assume this partition defined with labelled classes. Different expressions will be used for expressing $\mathcal{O}_h$: $\mathcal{O}_h$ is the set of objects where c_h is $TRUE$, having the category c_h, belonging to the category c_h, ..., $1 \leq h \leq k$.

With the partition π we associate the sequence $(n_1, n_2, ..., n_h, ..., n_k)$ of its cardinality classes ; $n_h = card(\mathcal{O}_h)$, $1 \leq h \leq k$. This sequence defines the type of the partition π. We also associate the sequence of proportions or relative frequencies $(f_1, f_2, ..., f_h, ..., f_k)$, where
$f_h = n_h/n$. This sequence defines the statistical distribution of c on $\mathcal{O}$.

The mapping (20) can be viewed as the representation of the nominal categorical attribute c at the level of the object set $\mathcal{O}$. It assigns to each elementary object o its value $c(o)$, which is a nominal code. It is important to realize that this representation does not allow to compare directly two different nominal categorical attributes by comparing directly their respective values. Indeed, given two nominal categorical attributes c and c' we cannot compare directly the statistical distributions of c and c' on $\mathcal{O}$. This comparison requires to cross the two respective partitions induced by c and c'.

There are different alternatives to carry out this comparison (see Chapter 4 of [1] and Chapter 2 of [19]). That we adopt requires a higher representation level. By noticing that representing a categorical attribute c is equivalent to represent the associated partition π, we define the following binary relation P_π :

$$(\forall (x, y) \in \mathcal{O} \times \mathcal{O}), xP_\pi y \text{ iff } \exists h, 1 \leq h \leq k, \text{ s.t. } x \in \mathcal{O}_h \text{ and } y \in \mathcal{O}_h \tag{22}$$

This binary relation is an equivalence relation (reflexive, symmetrical and transitive relation). Its symmetry and reflexive properties allow a representation at the level of the set, that we designate by P

$$P = \{\{x, y\} | x \in \mathcal{O}, y \in \mathcal{O}, x \neq y\} \tag{23}$$

of unordered object pairs. P is exactly the set $P_2(\mathcal{O})$ of all subsets with 2 elements of $\mathcal{O}$. Another notation we consider for $P_2(\mathcal{O})$ is $\mathcal{O}^{\{2\}}$. Two related representations denoted by $R(\pi)$ and $S(\pi)$ can be considered :

$$R(\pi) = \big\{\{x, y\} \in P_2(\mathcal{O}) | \exists h, 1 \leq h \leq k, x \in \mathcal{O}_h \text{ and } y \in \mathcal{O}_h\big\} \tag{24}$$

and

$$S(\pi) = \big\{\{x, y\} \in P_2(\mathcal{O}) | \exists g \neq h, 1 \leq g, h \leq k, x \in \mathcal{O}_g \text{ and } y \in \mathcal{O}_h\big\} \tag{25}$$

With $R(\pi)$ and $S(\pi)$ are naturally associated their indicator functions that we denote by ρ_π and σ_π, respectively :

$$(\forall \{x, y\} \in P), \rho_\pi(\{x, y\}) = 1(resp., 0) \text{ iff } \{x, y\} \in R(\pi)(resp., \notin R(\pi)) \tag{26}$$

and

$$(\forall \{x, y\} \in P), \sigma_\pi(\{x, y\}) = 1(resp., 0) \text{ iff } \{x, y\} \in S(\pi)(resp., \notin S(\pi)) \tag{27}$$

The 1 and 0 values represent the logical values $TRUE$ and $FALSE$ respectively. The interpretation where 1 and 0 are integer values can also be considered. In the latter case

$$(\forall \{x, y\} \in P), \rho_\pi(\{x, y\}) + \sigma_\pi(\{x, y\}) = 1 \tag{28}$$

Indeed, $\{R(\pi), S(\pi)\}$ defines a bi-partition of P, that is to say a partition of P into two classes. $R(\pi)$ [resp.,$S(\pi)$] is the set of joined (resp. separated) distinct object pairs, by π.

The expressions of $R(\pi)$ and $S(\pi)$ with respect to the partition classes (see equation (21)) are respectively

$$R(\pi) = \sum_{1 \leq h \leq k} P_2(\mathcal{O}_h)$$

and

$$S(\pi) = \sum_{1 \leq g < h \leq k} \mathcal{O}_g \star \mathcal{O}_h \tag{29}$$

In these expressions the sign Σ means a union of disjoint subsets. $P_2(\mathcal{O}_h)$ is the set of unordered pairs of elements of $\mathcal{O}_h$, or equivalently, the set of all 2-subsets of $\mathcal{O}_h$, $1 \leq h \leq k$. $\mathcal{O}_g \star \mathcal{O}_h$ designates the set of all unordered pairs $\{x, y\}$ such that $x \in \mathcal{O}_g$ and $y \in \mathcal{O}_h$, $1 \leq g \neq h \leq k$. We have

$$card[R(\pi)] = \sum_{1 \leq h \leq k} n_h \times (n_h - 1)/2$$

and

$$card[S(\pi)] = \sum_{1 \leq g < h \leq k} n_g \times n_h \tag{30}$$

The following formula can be verified

$$card[R(\pi)] + card[S(\pi)] = n \times (n-1)/2 \tag{31}$$

where the left member is the cardinality of $P = P_2(\mathcal{O})$ (see equation (23)).

Now, let us consider an example given in Goodman and Kruskal (1954) [9] and where the origin of the data is specified. The population is defined by white Protestant married couples living in Indianapolis, married in 1927, 1928, or 1929. Thus, each elementary element u of our universe $\mathcal{U}$ is defined by such a married couple. The data sampling leads to an object set $\mathcal{O}$ comprising 1438 elements. One categorical attribute c defined is the "highest level of formal education of wife". Three categorical values are considered. We take them as following

- c_1 = "less than three years high school" ;
- c_2 = "3 or 4 years high school" ;
- c_3 = "one year college or more".

The partition π (see equation (21)) includes here three classes $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$, where $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$ are the subsets of married couples whose formal education of wife are c_1, c_2 and c_3, respectively. We have according to the above notations, $n_1 = card[\mathcal{O}_1] = 591$, $n_2 = card[\mathcal{O}_2] = 608$ and $n_3 = card[\mathcal{O}_3] = 239$.

$P_2(\mathcal{O})$ is the set of unordered pairs of distinct married couples. $card[P_2(\mathcal{O})] = 1438 \times 1437/2 = 1033203$.

$P_2(\mathcal{O}_h)$ is the set of unordered pairs of distinct married couples whose formal education of wife is c_h, $1 \leq h \leq 3$.

- $card[P_2(\mathcal{O}_1)] = 591 \times 590/2 = 174345$;
- $card[P_2(\mathcal{O}_2)] = 608 \times 607/2 = 184528$;
- $card[P_2(\mathcal{O}_3)] = 239 \times 238/2 = 28441$.

Equations (29) and (30) give

$$card[R(\pi)] = 174345 + 184528 + 28441 = 387314$$

$\mathcal{O}_g \star \mathcal{O}_h$ is the set of unordered pairs of married couples whose formal education of wives are c_g and c_h, respectively, $1 \leq g < h \leq 3$.

- $card[\mathcal{O}_1 \star \mathcal{O}_2] = card(\mathcal{O}_1) \times card(\mathcal{O}_2) = 591 \times 608 = 359328$;
- $card[\mathcal{O}_1 \star \mathcal{O}_3] = card(\mathcal{O}_1) \times card(\mathcal{O}_3) = 591 \times 239 = 141249$;
- $card[\mathcal{O}_2 \star \mathcal{O}_3] = card(\mathcal{O}_2) \times card(\mathcal{O}_3) = 608 \times 239 = 145312$.

Equations (29) and (30) give

$$card[S(\pi)] = 359328 + 141249 + 145312 = 645889$$

We immediatly verify the equation (31) :

$$387314 + 645889 = 1033203$$

3.2 The ordinal categorical attribute

As in Goodman and Kruskal (1954) [9], we assume in the above example a total (i.e ; linear) order on the category set $\{c_1, c_2, c_3\}$ for the categorical attribute c defined by the "Highest level of formal education of wife". We adopt the total order :

$$c_1 < c_2 < c_3$$

More generally, by considering the formalism introduced in Section 3.1 (see equation (20)), for an ordinal version of the categorical attribute c, the category set $\mathcal{C}$ is provided by a strict total order ; that is to say, a ranking on $\mathcal{C}$. By supposing h the rank of the category c_h, $1 \leq h \leq k$, the latter total order is defined by

$$c_1 < c_2 < ... < c_h < ... < c_k \tag{32}$$

This structure induces a total preorder on the object set $\mathcal{O}$ that we denote by ω. By defining, as in Section 3.1, $\mathcal{O}_h = c^{-1}(c_h)$, $1 \leq h \leq k$, we have

$$\mathcal{O}_1 < \mathcal{O}_2 < ... < \mathcal{O}_h < ... < \mathcal{O}_k \tag{33}$$

and that means : for any (x, y) in the cartesian product $\mathcal{O}_g \times \mathcal{O}_h$, $c(x) < c(y)$ if and only if $g < h$, $1 \leq g < h \leq k$. The above total order on the set of classes $\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_h, ..., \mathcal{O}_k$ is called the *quotient order* associated with ω. Given a total preorder on $\mathcal{O}$ consists of given a partition on $\mathcal{O}$ and, additionally, a total order on its classes. This mathematical expression formalizes the notion of a "ranking with ties" given in Kendall (1948) [12]. Now, the ordered sequence of the cardinal classes $(n_1, n_2, ..., n_h, ..., n_k)$, denoted as in Section 3.1, is called here *composition* of the total preorder ω. The statistical distribution of c on $\mathcal{O}$ has the same meaning as for the nominal case (Section 3.1). It is defined by $(f_1, f_2, ..., f_h, ..., f_k)$, where $f_h = n_h/n$, $1 \leq h \leq k$. Nevertheless, the labelling of the different classes satisfies here the linear order (33).

As for the nominal categorical attribute and in spite of the linear order defined on the category set $\mathcal{C} = \{c_1, c_2, ..., c_h, ..., c_k\}$ the formal valuation on $\mathcal{O}$ defined by (20) does not allow the comparison between ordinal categorical attributes (see Sections 2.3.2, 2.3.4 and 2.4). As discussed in Section 3.3 a ranking attribute defining a linear order on the object set can be interpreted as a very specific case of an ordinal categorical attribute. However, the ranking function has a clear numerical interpretation and this enables one version of the comparison between two ranking attributes defining respectively two specific numerical valuations on the object set $\mathcal{O}$.

In order to make possible the comparison between two ordinal categorical attributes in the most general case (see Chapter 2 of [19]) let us define the following binary relation R_ω associated with the ordinal categorical attribute c :

$$(\forall((x, y) \in \mathcal{O} \times \mathcal{O})), xR_\omega y \text{ iff } c(x) < c(y) \tag{34}$$

Because of no symmetry and no reflexivity of R_ω, we consider the following representation set

$$C = \{(x, y) | x \in \mathcal{O}, y \in \mathcal{O}, x \neq y\} \tag{35}$$

In other words, C is the set of all ordered pairs of distinct objects from $\mathcal{O}$. Another notation we consider for C is $C^{[2]}$. In C, R_ω is represented by the following subset

$$R_\omega = \{(x, y) \in C | xR_\omega y\} \tag{36}$$

More explicitly R_ω can be put in the following form

$$R(\omega) = \sum_{1 \leq g < h \leq k} \mathcal{O}_g \times \mathcal{O}_h \tag{37}$$

where Σ designates a union of disjoint subsets and where $\mathcal{O}_g \times \mathcal{O}_h$ is the cartesian product of $\mathcal{O}_g$ and $\mathcal{O}_h$.
C can be decomposed as follows

$$C = R(\omega) + E(\omega) + S(\omega) \tag{38}$$

where

$$E(\omega) = \sum_{1 \leq h \leq k} \mathcal{O}_h^{[2]}$$
$$\text{and } S(\omega) = \sum_{1 \leq g < h \leq k} \mathcal{O}_h \times \mathcal{O}_g \tag{39}$$

The cardinalities of the three components of c (see equation (38)) are, respectively

$$
\begin{aligned}
card[R(\omega)] &= \sum_{1\leq g<h\leq k} n_g \times n_h \\
card(E(\omega)) &= \sum_{1\leq h\leq k} n_h \times (n_h - 1) \\
card(S(\omega)) &= \sum_{1\leq g<h\leq k} n_h \times n_g
\end{aligned} \tag{40}
$$

Clearly,

$$
card[R(\omega)] = card(S(\omega)) \tag{41}
$$

on the other hand, we can verify the formula

$$
card[R(\omega)] + card(E(\omega)) + card(S(\omega)) = n \times (n-1) \tag{42}
$$

Let us retake the above example in its ordinal version : $c_1 < c_2 < c_3$. We have

$$
\begin{aligned}
card[R(\omega)] &= 591 \times 608 + 591 \times 239 + 608 \times 239 \\
&= 359328 + 141249 + 145312 = 645889 \\
card(E(\omega)) &= 591 \times 590 + 608 \times 607 + 239 \times 238 \\
&= 348690 + 369056 + 56882 = 774628
\end{aligned}
$$

In these conditions, we verify the equation (42) where the right member is equal to $1438 \times 1437 = 2066406$.
With $R(\omega)$ is naturally associated its boolean indicator function that we denote by ρ_ω :

$$
(\forall (x,y) \in C), \rho_\omega(x,y) = 1(resp.,0) \text{ iff } (x,y) \in R(\omega)(resp., \notin R(\omega)) \tag{43}
$$

As above (see what follows the equations (26) and (27)), 1 and 0 are interpreted as the logical values $TRUE$ and $FALSE$, respectively. However, the interpretation where 1 and 0 are numerical scorings can be taken into account. In his work Kendall (1948) [12] uses a scoring numerical function with two values $+1$ and -1 in order to code a total order (ranking) $R(\omega)$ on an object set $\mathcal{O}$, coded itself by means of a given ranking R. More clearly, for the R ordered pair (x,y) such that xRy, the scoring value is $+1$ (resp.,-1) if $xR_\omega y$ (resp., $yR_\omega x$). The value $+1$ (resp., -1) indicates that $(x,y) \in R(\omega)$ (rep., $(y,x) \in R(\omega)$) (see equation (36)).Thus, our set theoretic representation is equivalent to the Kendall coding of a ranking. However, for our representation the initial ranking R is useless.

The set representation of the ordinal categorical attribute c by $R(\omega)$ (see equation (36)) does not take into account explicitly the set of ordered pairs (x,y) of distinct objects such that x and y belong to the same class $\mathcal{O}_h$ $c(x) = c(y) = h$, $1 \leq h \leq k$. This set is denoted by $E(\omega)$ in equations (38) and (39). However, due to (38) $E(\omega)$ is implicitly taken into account. Indeed, $E(\omega)$ is deduced from $R(\omega)$. In order to weight the part of $E(\omega)$, the attribute c can be represented by the following numerical scoring function denoted $score_\omega$:

$$
score_\omega(x,y) = \begin{cases} 1 & \text{if } (x,y) \in R(\omega) \\ 0.5 & \text{if } (x,y) \in E(\omega) \\ 0 & \text{if } (x,y) \in S(\omega) \end{cases} \tag{44}
$$

Let us now consider the more general case where the category set $\mathcal{C} = \{c_1, c_2, ..., c_h, ..., c_k\}$ is provided by a strict *partial* order. An instructive exercise we propose to the reader consists in taking up again the above set representation formalism for this general case. For our part, we reconsider the above example of the categorical attribute "Highest level of formal education of wife" where we only assume $c_1 < c_2$ and $c_2 < c_3$ for the categorical values (see above). Therefore, c_1 and c_2 are considered no comparable. This categorical scale induces a partial preorder on the object set $\mathcal{O}$, defined here by married couples. As above, denote ω this partial preorder. The expression of $R(\omega)$, $E(\omega)$ and $S(\omega)$ (see equations 37, 38 and 39) become

$$\begin{aligned} R(\omega) &= \mathcal{O}_1 \times \mathcal{O}_3 + \mathcal{O}_2 \times \mathcal{O}_3 \\ E(\omega) &= \mathcal{O}_1^{[2]} + \mathcal{O}_2^{[2]} + \mathcal{O}_1^{[3]} + \mathcal{O}_1 \times \mathcal{O}_2 + \mathcal{O}_2 \times \mathcal{O}_1 \\ S(\omega) &= \mathcal{O}_3 \times \mathcal{O}_1 + \mathcal{O}_3 \times \mathcal{O}_2 \end{aligned} \tag{45}$$

where the sign + indicates a set sum defined by a union of disjoint subsets.
The scoring function defined in (44) is appropriate to code the partial strict order ω.

3.3 The ranking attribute

In his book Kendall (1948) [12] uses the ability in a given subject taught (e.g. mathematics) to rank totally and strictly the described object set $\mathcal{O}$. Thus, a linear order that we denote by ω_l is induced on $\mathcal{O}$. For this two conditions must be satisfied : (i) the cardinality of $\mathcal{O}$ is enough small ; (ii) the scoring function is enough discriminating. The ranking attribute r is a bijective mapping of $\mathcal{O}$ on the set of the first n integer numbers :

$$r : \mathcal{O} \to \{1, 2, ..., i, ..., n\} \tag{46}$$

associating with a given object x of $\mathcal{O}$, its rank defined as follows :

$$(\forall x \in \mathcal{O}), x \mapsto r(x) = card\{y | y \in \mathcal{O} \text{ and } y \leq x \text{ for } \omega_l\} \tag{47}$$

By using the rank function r, the representation of the ranking attribute would have been considered in the framework of descriptive attributes of type I. Indeed, r defines a numerical valuation on the set $\mathcal{O}$, given by the first n integer numbers. In fact, this coding of a linear order leads to the Spearman coefficient Spearman(1904) [42]. But, as expressed in the preceding Section 3.2 the originality of the Kendall representation is the level $\mathcal{O} \times \mathcal{O}$ of its definition. In fact, a ranking attribute can be seen as a very particular case of an ordinal categorical attribute. For this, each class of the induced total order on $\mathcal{O}$ becomes a singleton (i.e. comprises exactly one element) (see equation (33)). The representation set $R(\omega)$ (see equation (36)) becomes

$$R(\omega_l) = \{(x, y) | (x, y) \in \mathcal{O} \times \mathcal{O} \text{ and } r(x) < r(y)\} \tag{48}$$

where $r(x) < r(y)$ is equivalent to $x \leq y$ and not $y \leq x$ for ω_l.
We have

$$card[R(\omega_l)] = \frac{n \times (n-1)}{2} \tag{49}$$

The "mean rank" function is an extension of the rank function (see equation (47)) employed to code a total preorder (ranking with ties) on an object set $\mathcal{O}$. For the total preorder ω considered in (33), the mean rank function, denoted r_m, is defined as follows

$$(\forall h, 1 \leq h \leq k), (\forall x \in \mathcal{O}_h), r_m(x) = n_1 + n_2 + ... + n_{h-1} + \frac{1}{2} \times (n_h + 1) \tag{50}$$

where $n_g = card(O_g)$, $1 \leq g \leq h$. The expression "mean rank" is explained by the fact that the mean rank of an arbitrary element x of $\mathcal{O}_h$, over all the linear orders compatible with the total preorder ω, is given by $r_m(x)$. The following formula is easy verify

$$\sum_{x \in \mathcal{O}} r_m(x) = \frac{n \times (n+1)}{2} \tag{51}$$

In these conditions, a total preorder ω on $\mathcal{O}$ and then, an ordinal categorical attribute describing $\mathcal{O}$ can be represented as a particular numerical valuation on $\mathcal{O}$, given by the mean rank function r_m. However the nature of this representation is very different from the relational one defined by $R(\omega)$ (see equation (36)). Indeed, the latter representation is logical. It gives for a given ordered pair (x, y) in $\mathcal{O} \times \mathcal{O}$, a logical value.

For the above mentioned reasons, a ranking attribute inducing a linear (total and strict) order on $\mathcal{O}$ cannot occur in case of the description of large data sets. However, a realistic case where such a description occurs, concerns the problem of m rankings Kendall (1948) [12]. Let us imagine a few number of objects : $n = card(\mathcal{O})$ is relatively small. $\mathcal{O}$ is for example a set of manufactured products of a given type. We assume a set of m judges giving each his preferences by ranking (without ties) the n objects. Thus, each judge defines a ranking attribute on $\mathcal{O}$. In this type of data there is no restriction on the number m of judges.

3.4 The categorical attribute valuated by a numerical similarity

We consider here the case where the category set $\mathcal{C}$ (see equation (20)) is provided by a numerical similarity. The latter is supposed given *a priori*, for example, by an expert knowledge. Let us denote it by ξ. ξ is defined by a mapping of the cartesian product $\mathcal{C} \times \mathcal{C}$ onto the real numbers. Mostly, the set value of ξ is the *positive* reals that is denoted by $\mathbb{R}_+$:

$$\begin{aligned} \xi : \mathcal{C} \times \mathcal{C} &\rightarrow \mathbb{R}_+ \\ (c_g, c_h) &\mapsto \xi(c_g, c_h) \end{aligned} \tag{52}$$

where $\xi(c_g, c_h)$ is the numerical similarity value between the categories c_g and c_h, $1 \leq g, h \leq k$.
Mostly, ξ is symmetrical :

$$(\forall (g,h), 1 \leq g, h \leq k), \xi(c_g, c_h) = \xi(c_h, c_g) \tag{53}$$

However, real important cases of asymmetrical similarity may occur. This point will be mentioned below. On the other hand we assume

$$((\forall (g,h), 1 \leq g, h \leq k)), min[\xi(c_g, c_g), \xi(c_h, c_h)] > \xi(c_g, c_h) \tag{54}$$

Nevertheless, $\xi(c_h, c_h)$ is not necessarily invariant with respect to h,
$1 \leq h \leq k$. ξ can be figured by the following square matrix

$\mathcal{C}$	c_1	$\cdots$	c_h	$\cdots$	c_k
c_1					
$\vdots$					
c_g			$\xi(c_g, c_h)$		
$\vdots$					
c_k					

Table 1 : Matrix of ξ

Now, let us consider the above example of the categorical attribute "Highest level of formal education of wife", having the three categories denoted c_1, c_2 and c_3 (see Section 3.1). The matrix of ξ might be :

$\mathcal{C}$	c_1	c_2	c_3
c_1	5	2	1
c_2	2	4	3
c_3	1	3	6

Table 2 : Matrix of ξ for the example

In fact, we have determined this matrix of numerical integer numbers from the following similarity ranking on $\mathcal{C} \times \mathcal{C}$

$$(c_1, c_3) < (c_1, c_2) < (c_2, c_3) < (c_2, c_2) < (c_1, c_1) < (c_3, c_3) \tag{55}$$

It may seem surprising that the similarity between a given category with itself is not the same whatever is this category. This can be justified by taking into account the specificity of the concerned category. In our example, we have considered that rarer is a category, more specific it is. Thereby, the category c_3 ("one year college or more") is more specific than c_2 ("3 or 4 years high school"). Intuitively, the resemblance between two different married couples whose common category is c_3, is stronger than that between two different married couples whose common category is c_2.

The valuation ξ (see equation (52)) induces a complete valuated symmetrical graph, without loops, on the object set $\mathcal{O}$. Explicitly

$$(\forall (x, y) \in \mathcal{O}^{[2]}), \xi(x, y) = \xi[c(x), c(y)] \tag{56}$$

This graph is decomposed according to the partition π (see equation (21)) as following

$$(\forall (x, y) \in \mathcal{O}_g \times \mathcal{O}_h), \xi(x, y) = \xi(c_g, c_h)$$

and

$$(\forall (x, y) \in \mathcal{O}_h^{[2]}, \xi(x, y) = \xi(c_h, c_h) \tag{57}$$

$1 \leq g \neq h \leq k$.

Now, let us illustrate this graph in the framework of the above example

$$\begin{array}{ll}
(\forall (x, y) \in \mathcal{O}_1 \times \mathcal{O}_2 + \mathcal{O}_2 \times \mathcal{O}_1) & \xi(x, y) = 2 \\
(\forall (x, y) \in \mathcal{O}_1 \times \mathcal{O}_3 + \mathcal{O}_3 \times \mathcal{O}_1) & \xi(x, y) = 1 \\
(\forall (x, y) \in \mathcal{O}_2 \times \mathcal{O}_3 + \mathcal{O}_3 \times \mathcal{O}_2) & \xi(x, y) = 3 \\
(\forall (x, y) \in \mathcal{O}_1^{[2]}) & \xi(x, y) = 5 \\
(\forall (x, y) \in \mathcal{O}_2^{[2]}) & \xi(x, y) = 4 \\
(\forall (x, y) \in \mathcal{O}_3^{[2]}) & \xi(x, y) = 6
\end{array}$$

Let us end this development by noticing that a nominal categorical attribute can be interpreted in terms of a very particular similarity categorical attribute as following

$$\begin{array}{r}
(\forall h, 1 \leq h \leq k), \xi(c_h, c_h) = 1 \\
(\forall (g, h), 1 \leq g \neq h \leq k), \xi(c_g, c_h) = 0
\end{array} \tag{58}$$

Now, for an asymmetrical similarity providing the category set of a categorical attribute, equation (53) does not hold. On the other hand, $\xi(c_h, c_h)$, $1 \leq h \leq k$, may not be defined. Besides, equations (56) and (57) are still valid. As an example, consider the traffic of cellular call phones between the towns of France. "Town of France" defines a categorical attribute. For a given ordered pair (A, B) of towns of France, let us consider the number, $\nu(A, B)$, of cellular call phones emitted from A to B. The ν function defines an asymmetrical similarity on the set of the towns of France (the category set in our example). Indeed, generally $\nu(A, B) \neq \nu(B, A)$. The object set could be the set of antennae of cellular telephone located in the different towns.

3.5 The valuated binary relation attribute

The definition of this attribute is directly given at the level of the cartesian product $\mathcal{O} \times \mathcal{O}$ where $\mathcal{O}$ is the object set. Let us denote by B the binary relation defined by a given attribute. B is represented by the following subset of the cartesian product $\mathcal{O} \times \mathcal{O}$:

$$R(b) = \{(x, y) | (x, y) \in \mathcal{O} \times \mathcal{O} \text{ and } xBy\} \tag{59}$$

Additionally, a valuation v on $R(b)$ is considered. v is a mapping of $R(b)$ on the value scale of v, denoted S in the following expression

$$v : R(b) \to S \tag{60}$$

Generally S is the positive reals. The valuated binary relation can be represented by a valuated graph :

$$\{\big((x, y), v(x, y)\big) | (x, y) \in R(b)\} \tag{61}$$

The valuated binary relation attribute is a generalization of all the above attributes considered in this Section 3. However, it does not correspond to a categorical attribute. And, the specificity of the categorical structure is very important for comparing attributes or building similarity indices between described objects or between categories (see Capter 3). The attribute type considered here occurs frequently in communication problems.

4 Representation of the attributes of type III

As mentioned in the introduction (Section 1) a categorical attribute is of type III if the similarity structure of its category set
$\mathcal{C} = \{c_1, c_2, ..., c_h, ..., c_k\}$ induces a binary relation on the set - denoted above by $\mathcal{O}^{\{2\}}$ (see equation (23)) - of unordered pairs of distinct objects, or on the set denoted above by $\mathcal{O}^{[2]}$ (see equation (35)) of ordered pairs of distinct. Specific ordinal similarity structures on $\mathcal{C}$ occur importantly for describing data Lerman and Peter (1986) [29], Lerman and Peter (2007) [32], Ouali-Allah (1991) [37] and Peter (1987) [38]. Each of them defines a specific total preorder on $\mathcal{O}^{\{2\}}$ (ranking with ties). We shall distinguish below three versions of such a categorical attribute : "preordonance attribute", "taxonomic attribute" and "taxonomic preordonance attribute". A set theoretic representation at the level of the cartesian product $\mathcal{O}^{\{2\}} \times \mathcal{O}^{\{2\}}$ will be first given. Next, a representation by means of an adequate numerical valuation on $\mathcal{O}^{\{2\}}$ will be proposed. The latter is easier and more efficient to handle for comparing data described by such a categorical attribute. As just mentioned, for the below presented categorical attributes, the induced total preorder is defined at the level of the set $\mathcal{O}^{\{2\}}$ of unordered object pairs. For this case, the ordinal similarity structure on $\mathcal{C}$ is symmetrical with respect to the categories to be compared. Nevertheless, extension can be envisaged for the case where the induced total preorder is defined at the level of $\mathcal{O}^{[2]}$ of ordered object pairs. For this case the ordinal similarity on $\mathcal{C}$ is asymmetrical.

4.1 The preordonance categorical attribute

A "preordonance" categorical attribute is a categorical attribute whose category set $\mathcal{C}$ is provided by an ordinal similarity, called preordonance on $\mathcal{C}$. Formally, as said above, a preordonance on $\mathcal{C}$ is a total preorder on a specific set of category pairs of $\mathcal{C}$. For the categorical attributes to be introduced where the ordinal similarity is symmetrical [2], the $\mathcal{C}$ category pairs to be considered is $\{(c_g, c_h) | 1 \leq g \leq h \leq k\}$. By denoting $K = \{1, 2, ..., h, ..., k\}$ the category codes of the concerned attribute, the total preorder is defined on the following set

$$K_2 = \{(g, h) | 1 \leq g \leq h \leq k\} \tag{62}$$

[2] The ordinal similarity between c_g and c_h is the same as that between c_h and c_g, $1 \leq g \leq h \leq k$.

Lerman and Peter (1985) [29], Lerman(1987) [20], Ouali-Allah (1991) [37], Lerman (2000) [25], Lerman and Peter (2003) [31]. In fact, we have used already and implicitly such a categorical attribute in the above example of "Highest level of formal education of wife" (see equation (55)).

Let us now give an example concerning the problem of a database management of real estate advertisements Peter (1987) [38]. One categorical attribute defined in this database is "subject of the transaction". Its categories are : $c_1 = house$, $c_2 = villa$, $c_3 = apartment$, $c_4 = studio\ apartment$, $c_5 = room$, $c_6 = garage$ and $c_7 = piece\ of\ land$. By denoting gh the category pair (c_g, c_h), $1 \leq g \leq h \leq k$, the proposed preordonance is :

$$\begin{aligned} 14 \sim 15 \sim 16 \sim 17 \sim 24 \sim 25 \sim 26 \sim 27 \sim \\ 35 \sim 36 \sim 37 \sim 46 \sim 47 \sim 56 \sim 57 \sim 67 < \\ 23 \sim 34 < 13 < 45 < 12 < \\ 11 \sim 22 \sim 33 \sim 44 \sim 55 \sim 66 \sim 77 \end{aligned} \tag{63}$$

where the symbols $\sim$ and $<$ mean "equivalent" and "strictly lower than", respectively.

There are in all $k \times (k+1)/2 = 7 \times (7+1)/2 = 28$ pairs. This total preorder comprises 6 classes. Their respective cardinalities (*composition* of the total preorder) are 16, 2, 1, 1, 1 and 7.

Generally, the expert establishes the total preorder on K_2 (preordonance on $\mathcal{C}$), recursively, by sorting the pairs at each step the most similar among pairs of K_2 not yet sorted.

The preordonance categorical attribute has played an important part in our formalization work of data description (see the above references). Let us mention that this type of attribute appeared also independently with a different expression and in a very different context in Chah (1985) [4].

The preordonance categorical attribute concept does not require a notion of metrical difference between categories. Nevertheless, psychometric researchers consider a numerical scale measurement called "ordered metric scale" in which the differences between categories are defined and ordered Stevens (1951) [43].

Let us now define the set theoretic representation of a preordonance categorical attribute. Denote by $\omega(K_2)$ the total preorder on (K_2) (see equation (62)) expressing this preordonance attribute.

$$(L_1, L_2, ..., L_q, ..., L_r) \tag{64}$$

will designate the ordered sequence of the class preorder. L_1 (resp., L_r) comprises the most dissimilar (resp., similar) category pairs. To fix idea and for consistent reasons, we can assume that the last classes are defined on the subset $\{(h,h)|1 \leq h \leq k\}$. However, this point is out of our representation problem. The basic representation level is the cartesian product $K_2 \times K_2$. This representation has the same nature as that given in Section 3.2 for a total preorder on an object set (see equation(38)). But clearly, the context is different. In the latter, we define the following set sums of cartesian products

$$R(\omega(K_2)) = \sum_{1 \leq p < q \leq r} L_p \times L_q \tag{65}$$

$$P(\omega(K_2)) = \sum_{1 \leq p \leq r} L_p \times L_p \tag{66}$$

$$S(\omega(K_2)) = \sum_{1 \leq p < q \leq r} L_q \times L_p \tag{67}$$

$$\begin{aligned} \mathcal{I}_m(L_p \times L_q) &= (\sum_{(e,f) \in L_p} \mathcal{O}_e \star \mathcal{O}_f)(\sum_{(g,h) \in L_q} \mathcal{O}_g \star \mathcal{O}_h) \\ \mathcal{I}_m(L_p \times L_p) &= (\sum_{(g,h) \in L_p} \mathcal{O}_g \star \mathcal{O}_h)^2 \end{aligned} \tag{68}$$

$1 \leq p < q \leq r$, where as above the sums indicate union of disjoint subsets and where $\mathcal{O}_g \star \mathcal{O}_h$ is defined by the set of distinct unordered pairs whose components belong to $\mathcal{O}_g$ and $\mathcal{O}_h$, respectively, $1 \leq g \leq h \leq k$.

In order to avoid too big combinatorial complexity we code the total preorder $\omega(K_2)$ with the "mean rank function" r_m (see equation (50) expressed in another context). Referring to equation (64), denote by l_q the cardinality of L_q, $1 \leq q \leq r$. Therefore, for a given pair (g,h) belonging to L_q we get

$$r_m(g,h) = \sum_{1 \leq p < q} l_p + \frac{l_q + 1}{2} \tag{69}$$

Thus,we can verify the equation

$$\sum_{1 \leq q \leq r} l_q = \frac{k \times (k+1)}{2}$$

For example, for the total preorder (63), the sequence of the values of r_m is

$$\frac{16+1}{2} = 8.5 \ , \ 16 + \frac{2+1}{2} = 17.5 \ , \ 19 \ , \ 20 \ , \ 21 \ , \ 21 + \frac{7+1}{2} = 25$$

We verify that the sum of the r_m ranks is equal to $28 \times 29/2 = 406$.

Thus, a preordonance categorical attribute is coded as a specific categorical attribute valuated by a numerical similarity (see Section 3.4), given by the mean rank function.

Let us denote by R_q the right member of (69). R_q is the common value of the mean rank function r_m on the pairs (g,h) belonging to the q^{th} preorder class L_q (see (64)). The empirical statistical distribution of R_q on $\mathcal{O} \star \mathcal{O}$ is given by

$$\{(R_q, \sum_{(g,h) \in L_q} (n_g \times n_h / n^{[2]}))|1 \leq q \leq r\} \tag{70}$$

where $n^{[2]} = n \times (n-1)/2$.

We have considered above the case of symmetrical ordinal similarity for which the ordinal similarity for the ordered pair (c_g, c_h) is the same as that for (c_h, c_g), $1 \leq g \neq h \leq k$. As mentioned above, there might be data where the ordinal similarity is asymmetrical Lerman and Guillaume (2011) [28], Lerman and Kuntz (2011) [16]. In this case the total preorder comparing categories has to be established on

$$H_2 = \{(g,h)|1 \leq g,h \leq k\} \tag{71}$$

instead of on K_2 (see equation (62)).

In these conditions, an analogous development as above (see equations (62) to (70)) has to be setup. In this, the non-ordered object pairs $\mathcal{O}^{\{2\}}$ has to be replaced by the ordered object pairs $\mathcal{O}^{[2]}$. We leave this development to the reader.

4.2 The taxonomic categorical attribute

Let us begin with an example taken from data we have processed Lerman and Peter (1988, 2007) [30], [32]. These data are provided by biological descriptions of phlebotomine sandfly species of French Guiana Lebbe et al. (1987) [15]. Descriptions are very complex. Relative to a descriptive categorical attribute and for subsets of category values, hierarchical logical dependencies associated with the *mother* $\rightarrow$ *daughter* relation, have to be taken into account. Consider the attributes 1, 18, 19 and 20 defined in this database [15] and retaken in [30]. We denote them by a^1, a^{21}, a^{31} and a^{32}, respectively. a^1 is the "Sex" attribute, a^{21} is defined by the "Number of style spines", a^{31} indicates the "Distribution of 4 style spines" and a^{32}, the "Distribution of 5 style spines". The code category sets of these attributes are : $\{1 : male \, 2 : female\}$, $\{1,2,3,4,5\}$, $\{1,2,3,4,5,6\}$ and $\{1,2,3,4,5\}$, respectively. We obtain the following taxonomic structure organizing the different attributes according to the *mother* $\rightarrow$ *daughter* relation. Then, with a categorical value of a given attribute a daughter attribute might be associated. For example, the attribute a^{21} is associated with the value 1 of the attribute a^1 (see Figure 2.3).

Clearly, the attribute a^{21} is defined only when the a^1 value is 1. It is defined on the subset of objects whose a^1 value is 1 (male phlebotomine sandflies in our example). On the other hand, the attributes a^{31} and a^{32} are defined only when the values of a^{21} are 4 or 5. a^{31} (resp., a^{32}) is defined on the subset of objects where the value of a^1 is 1 and where the value of a^{21} is 4 (resp., 5). The common mother of the attributes a^{31} and a^{32} is a^{21}.

More generally, a taxonomic categorical attribute denoted τ, is defined by an organization of logically dependent attributes. It consists of a sequence of collections of categorical attributes of the following form :

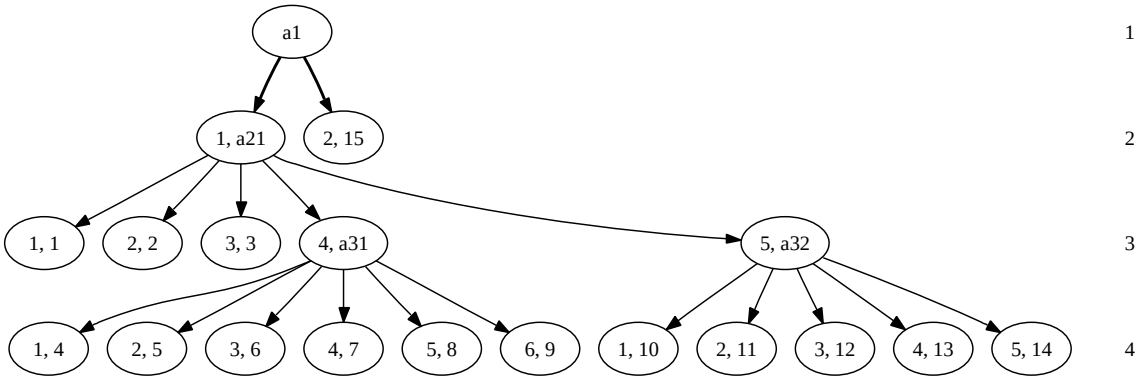

FIG. 3 – Taxonomic attribute

$$\tau = (\{a^1\}, \{a^{21}, a^{22}, ..., a^{2k_2}\}, ..., \{a^{p1}, a^{p2}, ..., a^{pk_p}\}, \\ ..., \{a^{q1}, a^{q2}, ..., a^{qk_q}\}) \tag{72}$$

In the above example τ is instanciated as follows :

$$\tau = (\{a^1\}, \{a^{21}\}, \{a^{31}, a^{32}\})$$

Now, let us imagine two categorical attributes a^{41} and a^{42} defined respectively for $a^{31} = 4$ and for $a^{32} = 5$. By denoting A^{41} and A^{42} the respective category sets of a^{41} and a^{42}, A^{41} and A^{42} will be placed at the same 5^{th} level of the above hierarchical structure (see Figure 2.3). A^{41} and A^{42} will be instanciations of the above $\{a^{pi}, ..., a^{pi'}\}$ $(i' > i)$ subset of categories. Two categorical elements of A^{41} (resp., A^{42}) would have the same mother a^{41} (resp., a^{42}).

The construction of a taxonomic attribute must be done in a descendant manner, level by level. The first level 1 is assigned to the root of the taxonomy. Hence, in our example, the root is defined by the categorical attribute "Sex". Now, consider a given node ν defined at the l^{th} level (l integer) and corresponding to a value of a preceding categorical attribute. The categorical attribute a^ν gives rise to a $(l+1)^{th}$ level, on which the values of a^ν are represented by different nodes. Directed descendant arrows join the node ν to the new nodes. Thus, in the example of Figure 2.3, the node corresponding to the value 1 of a^1, where the attribute a^{21} is defined, gives rise to the level 3 on which the nodes corresponding to the values of a^{21} are placed. There are in all 5 values. Moreover, the two sets of nodes associated with the attributes a^{31} and a^{32} are placed at the fourth level,respectively (a^{31} is defined for the value 4 of a^{21} and a^{32} is defined for the value 5 of a^{21}). There are 6 values for a^{31} and 5 values for a^{32}.

Each node of the taxonomic tree representing a taxonomic attribute is labelled. There are two alternatives for this labelling whether the concerned node is terminal or not. For a terminal node - defining a leave of the taxonomic tree - the label is a value of a categorical attribute taking part in the taxonomy. For a non-terminal node the labelling includes two values : the first one is defined by a categorical attribute value and the second one specifies the categorical attribute considered at this node. As an example, consider the level 3 of the taxonomic tree given in Figure 2.3. From left to right, the first three nodes are terminal nodes and the two last nodes are non-terminal nodes.

A lateral ranking of leaves (terminal nodes) from left to right can be considered in the taxonomic structure. This ranking is given in the above pictured tree where there are in all 15 terminal nodes. Consider again the level 3 of the taxonomic tree. The extreme left node is a terminal node. It is labelled by the ordered pair $(1, 1)$. The first component is defined by the first value of the mother categorical attribute a^{21}. The second component is its lateral rank. The terminal node at the extreme left of level 4 is labelled $(1, 4)$. It corresponds to the value coded 1 of the categorical attribute a^{31} involved in its mother node. It is created directly after the terminal node $(3, 3)$ of level 3. Clearly, the lateral ranking depends on drawing options. At the 3 level, the terminal nodes are first placed. This could have been done for the level 2. The chosen alternative is due to legibility reason. Anyway, a precise status could be defined for a consistent and systematic technique of drawing the taxonomic tree associated with a taxonomic attribute. This is left to the reader as an exercise. In the following each terminal node (leave of the taxonomic tree) will be coded by its lateral rank. Notice that each terminal node can be identified with the associated complete chain of the taxonomic tree, starting with the root and ending with it.

Now, we are going to associate bijectively with the taxonomic structure defined by the taxonomic attribute expressed by equation (72) and its graphical representation as given in Figure 2.3, a level labelled classification tree on the set denoted $\mathcal{T}$ of its terminal nodes. This is mathematically defined by a partition chain on $\mathcal{T}$ (see Chapter 0 of [19]). Assume that k levels are comprised in the taxonomic structure ($k = 4$ for our example). The level 1 is the root level and the k level

is determined by the terminal nodes corresponding to attributes introduced at the $(k-1)^{th}$ level. In our example these attributes are a^{31} and a^{32}. Denote $(P_0, ..., P_{j-1}, P_j, ..., P_{k-1})$ the partition chain associated with the taxonomic structure. In order to precise this sequence of partitions, identify each non-terminal node with the subset of terminal nodes deriving from it. As an example, the node $(5, a^{32})$ is identified with the subset $\{10, 11, 12, 13, 14\}$. Besides, the node $(1, a^{21})$ is identified with the subset $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$.

P_0 is the finest partition. Each of its classes is a singleton including exactly one single element. In our example P_0 comprises 15 classes. P_{k-1} is the least fine partition, it includes only one class grouping all the elements. P_j is deduced from P_{j-1} by aggregating nodes (terminal or not) appearing at the the level $k-j+1$ of the taxonomic structure and derived from a same categorical attribute attribute defined for a given node of the level $k-j$. In our example, the partition P_1 is obtained from the partition P_0 by aggregating $\{4\}$, $\{5\}$, $\{6\}$, $\{7\}$, $\{8\}$, $\{9\}$ on one side and by aggregating $\{10\}$, $\{11\}$, $\{12\}$, $\{13\}$, $\{14\}$ on the other side. Thus, we have

$$P_1 = \big\{\{1\}, \{2\}, \{3\}, \{4, 5, 6, 7, 8, 9\}, \{10, 11, 12, 13, 14\}, \{15\}\big\}$$

The partition P_2 into two classes is easily obtained

$$P_2 = \big\{\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11?12, 13, 14\}, \{15\}\big\}$$

In Chapter 0 (Section 4.2) of [19] bijective correspondence has been established between a partition chain on a finite set E and a ultrametric distance on E given by a level notion of the classification tree associated with the partition chain. Here, equivalently, we shall define a bijective correspondence between a taxonomic structure (see equation (72)) and a ultrametric proximity on the set that denoted by $\mathcal{T}$ of its terminal nodes (see Figure 2.3). Let us recall that this ultrametric proximity that we denote p is characterized by the following property :

$$(\forall x, y \text{ and } z \in \mathcal{T}, x \neq y, x \neq z \text{ and } y \neq z), p(x, y) \geq min[p(x, y), p(y, z)] \tag{73}$$

(see Chapter 0 of [19], Section 0.4.1).

Here, the node level is defined by 1 plus the number of branches joining the root of the taxonomic structure to the concerned node. As examples in Figure 2.3, $level\big((4, a^{31})\big) = 3$ and $level\big((5, 8)\big) = 4$. For u and v belonging to $\mathcal{T}$, $p(u, v)$ is defined by the highest node level of the taxonomic structure where u and v are aggregated. Thus, as an example relative to the figured taxonomic tree (see Figure 2.3), $p(10, 14) = 3$ and $p(7, 13) = 2$; $p(10, 14) > p(7, 13)$.

The proof of the above condition (equation (73)) is analogous to that given for the ultrametric distance associated with a partition chain (see Chapter 0 of [19], Section 4.2). Indeed, for a subset $\{x, y, z\}$ of 3 elements of $\mathcal{T}$, suppose $p(x, y) > p(y, z)$: the highest aggregating level of x and y is strictly greater than that of y and z. Therefore, x and y belong to the same node - appearing at the level $p(x, y)$ - excluding z. Otherwise $p(y, z) \leq p(x, y)$. Now, if $p(x, y) = p(y, z)$, the node highest level joining x and y is the same as that joining y and z. Hence and necessarily, the first node is identical to the second one, because both include y. Therefore, $p(x, y) = p(y, z) = p(x, z)$. This ends the proof of (73).

The proximity p function valuates a total preordonance on the set $\mathcal{T}$of the terminal nodes ; that is to say, a total preorder on the set

$$\mathcal{T}_2 = \{(x, y) | x \in \mathcal{T}, y \in \mathcal{T}, 1 \leq x \leq y \leq |\mathcal{T}|\} \tag{74}$$

where $|\mathcal{T}|$ is the cardinality of $\mathcal{T}$ ($|\mathcal{T}| = 15$ in the above example).

This total preorder (ranking with ties) that we denote by $\omega(\mathcal{T}_2)$ is established as follows

$$(\forall (x, y) \text{ and } (u, v) \in \mathcal{T}_2), (x, y) \leq (u, v) \text{ iff } p(x, y) \leq p(u, v) \tag{75}$$

The total preordonance defined on $\mathcal{T}$ is ultrametric in the sense given in Chapter 0 of [19], Section 4.3. More explicitly, if ρ is a ranking function on $\mathcal{T}_2$ compatible with p (i.e. strictly increasing with respect to p), we have

$$(\forall x, y, \text{ and } z \in \mathcal{T}), \rho(x, y) \geq r \text{ and } \rho(y, z) \geq r \Rightarrow \rho(x, z) \geq r \tag{76}$$

where r is an arbitrary positive integer.

We adopt for the ranking function ρ, the mean rank function r_m introduced above (see equations (69) and (70)).

Finally, we represent a taxonomic attribute by a ultrametric preordonance on the set $\mathcal{T}$ of the terminal nodes of the associated taxonomic structure. This preordonance is numerically coded with the mean rank function defined on $\mathcal{T}_2$ (see equations (74) and (75)).

Now, let us ilustrate how this represetation is established in the framework of our example. Consider the diagram of Figure 2.3 from down to top. The root node results from aggregating the two components of all ordered pairs of the form $(i, 15)$, $1 \leq i \leq 14$. Then, the first preorder class of $\omega(\mathcal{T}_2)$ includes 14 pairs. Therefore, the common mean rank of each of them is (14+1)/2=7.5. The node $(1, a^{21})$ is obtained from three types of pair aggregations. These are :

1. $\{(i,j)|1 \leq i < j \leq 3\}$;
2. $\{(i,j)|1 \leq i \leq 3, 10 \leq j \leq 14\}$;
3. $\{(i,j)|1 \leq i \leq 9, 10 \leq j \leq 14\}$.

There are in all $3 + 3 \times 6 + 9 \times 5 = 66$ pairs. These constitute the second class of $\omega(\mathcal{T}_2)$. In these conditions the mean rank of each of these pairs is $14 + (66 + 1)/2 = 47.5$.

The nodes $(4, a^{31})$ and $(5, a^{32})$ are constituted at level 3. The former aggregates the components of the set of pairs $\{(i,j)|4 \leq i < j \leq 9\}$ and the latter those of $\{(i,j)|10 \leq i < j \leq 14\}$. There are in all $((6\times5)/2)+((5\times4)/2) = 25$ pairs. They constitute the third class preorder of $\omega(\mathcal{T}_2)$. Consequently, the common mean rank of these pairs is $14+66+(25+1)/2 = 93$ The fourth class preorder of $\omega(\mathcal{T}_2)$ is defined by all ordered pairs of the form (i, i), $1 \leq i \leq 15$. The common mean rank assigned to each of these pairs is $14 + 66 + 25 + (15 + 1)/2 = 113$.

The expected value of the mean rank sum is easily obtained :

$$14 \times 7.5 + 66 \times 47.5 + 25 \times 93 + 15 \times 113 = 7260$$

.

$\omega(\mathcal{T}_2)$ induces a total preorder on the set $O^{\{2\}}$ of unordered object pairs. In order to explicit this total preorder (ranking with ties), denote $\mathcal{O}(i)$ the $\mathcal{O}$ subset defined by the value i of the taxonomic attribute, $1 \leq i \leq |\mathcal{T}|$. The set $\{\mathcal{O}(i)|1 \leq i \leq |\mathcal{T}|\}$ defines a partition on $\mathcal{O}$. Now, for $i < j$, $1 \leq i < |\mathcal{T}|$, consider the unordered object pairs $\mathcal{O}(i) \star \mathcal{O}(j)$ (see what follows equation(29) for its definition) and substitute in $\omega(\mathcal{T}_2)$, $\mathcal{O}(i) \star \mathcal{O}(j)$ for (i, j). On the other hand, substitute in $\omega(\mathcal{T}_2)$, the unordered object pairs $P_2(\mathcal{O}(i))$ of $\mathcal{O}(i)$ for the ordered pair (i, i), $1 \leq i \leq |\mathcal{T}|$.

4.3 The taxonomic preordonance attribute

Let us reconsider here the ordinal similarity structure provided by a taxonomic attribute τ organizing a set of logically dependent categorical attributes (see equation (72)). We further assume here that the category set denoted $\mathcal{C}(a^{pi})$ of a given categorical attribute a^{pi} is provided by a total preordonance (see Section 4.1), $1 \leq i \leq k_p$, $1 \leq p \leq q$. These preordonances are locally defined, attribute by attribute. They have to be integrated in the taxonomic structure. In these conditions, we have to build a total preordonance on the set of the taxonomy leaves, or, equivalently, on the set of the associated complete chains going from the root to the leaves. This preordonance must take into account both the preordonance defined by the taxonomic structure and those we have just mentioned.

Such preordonance is built step by step, decreasingly according to the taxonomic resemblance between terminal nodes (leaves of the taxonomy). The general principle consists in refining the ultrametric preordonance associated with the taxonomy (see Section 4.2) by means of the preordonances locally defined on the category sets of the different attributes.

In order to clarify the technique, let us begin by illustrating the refinement process in the framework of our example (see Figure 2.3). By going from the deepest level (4 in our case) to the root one, each refinement step concerns the terminal nodes aggregated at the first time at a given level. We begin by ordering the set

$$\Delta(\mathcal{T}) = \{(x, x)|x \in \mathcal{T}\} \tag{77}$$

according to the leaf depth in the taxonomy : in other words, the deeper the leaf the higher the ordinal similarity between the represented category and itself is. Thus, in case of our example (see Figure 2.3) we have

$$\begin{gathered}(4,4) \sim (5,5) \sim (6,6) \sim (7,7) \sim (8,8) \sim (9,9) \\ \sim (10,10) \sim (11,11) \sim (12,12) \sim (13,13) \sim (14,14) \\ > (1,1) \sim (2,2) \sim (3,3) > (15,15)\end{gathered}$$

By considering this ranking in a reverse manner, according to increasing ordinal similarity, the mean ranks assigned to the three preorder classes - that we can distinguish - are 106, 108 and 115, respectively. They replace the common rank 113 (see above). The concerned sum rank is preserved. Indeed, $106 + 3 \times 108 + 11 \times 115 = 15 \times 113 = 1695$.

Now, let us consider the passage from level 4 to level 3. $A = \{4, 5, 6, 7, 8, 9\}$ and $B = \{10, 11, 12, 13, 14\}$ code the category sets of the attributes a^{31} and a^{32}, respectively. A and B are constituted by terminal nodes. According to above notations (see Sections 2.4.1 and 2.4.2), the set $P_2(A)$ (resp., $P_2(B)$) of unordered element pairs is defined by $\{(x, y)|4 \leq x < y \leq 9\}$ (resp., $\{(x, y)|10 \leq x < y \leq 14\}$). $P_2(A) \cup P_2(B)$ determines a unique class of the total preorder defined at the level 3 by the tree structure. This class comprises all the element pairs aggregated at the level 3 of the taxonomy. There are in all $card[P_2(A)] + card[P_2(B)] = 15 + 25$ pairs. Two preordonance structures on the category sets A and B of the attributes a^{31} and a^{32} provide total preorders on $P_2(A)$ and $P_2(B)$, respectively. These two total preorders must be employed consistently in order to define a unique total preorder on the entire set $P_2(A) \cup P_2(B)$. One option for this consists in requiring the expert knowledge for ranking the category pairs of A with respect to those of B. The option we adopt uses the mean rank functions, locally defined on $P_2(A)$ and $P_2(B)$ from the total preordonances on A and on B. These are interpreted as numerical similarity coefficients on A and on B, respectively. Denote r_A and r_B these mean rank functions. The value sum of r_A (rep., r_B) is $(15 \times 16)/2 = 120$ (rep., $(10 \times 11)/2 = 55$). In these conditions, we define a numerical function $r_{A \cup B}$ on $P_2(A) \cup P_2(B)$ directly deduced from r_A and r_B, as follows

$$r_{A \cup B} : P_2(A) \cup P_2(B) \rightarrow Val(r_A) \cup Val(r_B) \tag{78}$$

where $Val(r_A)$ (resp., $Val(r_B)$) is the value set of r_A (resp., r_B) and where $r_{A \cup B}\big((x, y)\big)$ is equal to $r_A\big((x, y)\big)$ [resp., $r_B\big((x, y)\big)$] if $\{x, y\} \in P_2(A)$ [resp., $P_2(B)$].

Therefore, according to the value scale of $r_{A \cup B}$, a total preorder on $P_2(A) \cup P_2(B)$ is established as follows

$$\begin{array}{r} \forall \{x, y\} \text{ and } \{z, t\} \in P_2(A) \cup P_2(B), \\ \{x, y\} \leq \{z, t\} \text{ iff } r_{A \cup B}\big((x, y)\big) \leq r_{A \cup B}\big((z, t)\big) \end{array} \tag{79}$$

This total preorder is substituted for the unique class $P_2(A) \cup P_2(B)$. A new global mean ranking function refining the previous one is established on the set defined by this class. Recall that we had one common value 93 for all of its elements (See Section 4.2). Necessarily the new assigned values belong to the interval $[14 + 66 + 1 + 81, 14 + 66 + 25]$.

The nature of the passage from level 3 to level 2 is somewhat different than the preceding one. All the nodes occupying the level 3 derive from only one mother node occupying the level 2, this being defined by the categorical attribute a^{21}. Recall that the unique preorder class of $\omega(\mathcal{T}_2)$ (see equation (74)) associated with the taxonomy is defined by the set of unordered leave pairs given in equation (77). For this, the mean rank of a given pair of such a preorder class is 47.5. The category set D of a^{21} is defined by $\{1, 2, 3, a^{31}, a^{32}\}$. A total preordonance on D enables to determine a total preorder on the mentioned unique preorder class of $\omega(\mathcal{T}_2)$. In order to built the latter, we have to do the following substitutions

$$\begin{array}{rcl} (\forall x \in \{1, 2, 3\}), (x, 4) & \leftarrow & \{(x, y) \mid y \in A\} \\ (\forall x \in \{1, 2, 3\}), (x, 5) & \leftarrow & \{(x, y) \mid y \in B\} \\ \text{for } \{4, 5\} & \leftarrow & \{(x, y) \mid (x, y) \in A \times B\} \end{array} \tag{80}$$

where the different pairs included in a given class substitution are interpreted as equally similar, A and B have been defined above.

Notice that the new mean rank values in the concerned class of $\omega(\mathcal{T}_2)$ are comprised between $14+1 = 15$ and $14+66 = 80$.

Now, let us give a general expression of the construction of a taxonomic preordonance attribute. For this purpose, we start with the definition (72) of a taxonomic attribute τ. $\{a^{pi}|1 \leq i \leq k_p\}$ is the set of the categories introduced at the p^{th} level in a descendant way from the top to the bottom. Some of these categories define terminal nodes of the τ structure and some others define categorical attributes which are divided at the next $(p+1)^{th}$ level (see level $p = 3$ of the example). Denote $P_2[\mathcal{C}(a^{pi})]$ the set of unordered pairs of the category set $\mathcal{C}(a^{pi})$ of a^{pi}, $1 \leq i \leq k_p$. $\bigcup_i P_2[\mathcal{C}(a^{pi})]$ determines a new unique class - by going from down to top - of the taxonomic preordonance. This unique class is refined by means of the total preordonances defined on $\mathcal{C}(a^{pi})$, $1 \leq i \leq k_p$, respectively.

We begin by ordering the set $\Delta(\mathcal{T}) = \{(x, x)|x \in \mathcal{T}\}$ considered in equation (77), of category pairs of the form (x, x) where x is a terminal node (leaf) of the taxonomic structure (see above for the given example following (77)). Then and recursively, for $p = q$ to $p = 2$, the unique taxonomic class $\bigcup_i P_2[\mathcal{C}(a^{pi})]$ is refined. The following steps are needed for this refinement :

(**i**) Compute the local mean rank rank function r^i_m for the total preorder $P_2[\mathcal{C}(a^{pi})]$, defined from the preordonance attribute a^{pi} ;

(**ii**) Establish a global total preorder on $\bigcup_i P_2[\mathcal{C}(a^{pi})]$ compatible with the respective local mean ranking functions r^i_m, $1 \leq i \leq k_p$, that is to say :

$$(\forall (x,y) \in P_2[\mathcal{C}(a^{pi})], (z,t) \in P_2[\mathcal{C}(a^{pi'})]),$$
$$(x,y) \leq (z,t) \text{ iff } r^i_m(x,y) \leq r^{i'}_m(z,t) \tag{81}$$

(**iii**) Begin by associating with each category set $\mathcal{C}(a^{pi})$ the category set $\mathcal{C}^t(a^{pi})$ corresponding to terminal nodes derived from a^{pi} and then, for any (x,y) belonging to $P_2[\mathcal{C}(a^{pi})]$ consider the subset $A(x)$ [resp., $A(y)$] of $\mathcal{C}^t(a^{pi})$ defined by terminal nodes issued from x (resp., y). In these conditions, the set leaf pairs $A(x) \times A(y)$ is substituted for (x,y) (see (80)). All the concerned pairs are interpreted as equally similar and the mean rank function value $r^i_m(x,y)$ is applied to all of these pairs.

The set $\bigcup_{(x,y)} A(x) \times A(y)$ ($(x,y) \in \bigcup_i P_2[\mathcal{C}(a^{pi})]$) of terminal node pairs constituted a unique class of the total preorder representing the taxonomic attribute. And now, this class is divided into subclasses depending on the respective preordonances defined on the category sets $\mathcal{C}(a^{pi})$ and the local mean rank functions r^i_m calculated for the respective total preorders on $P_2[\mathcal{C}(a^{pi})]$, $1 \leq i \leq k_p$.

When the process ends with $p = 2$, we obtain a total preorder on the entire set of terminal node pairs, comprising the pairs of $\Delta(\mathcal{T})$ (see (77)). This global preordonance is valuated by means of rhe mean rank function according to increasing ordinal similarity (see the above example).

4.4 Coding the different attributes in terms of preordonance or similarity categorical attributes

We have just seen above (Sections 4.2 and 4.3) how the taxonomic categorical attribute can be represented as a specific preordonance categorical attribute (see Section 4.1). By applying for each of them the mean rank function in order to code the associated total preorder on the category pairs (see (70)) we obtain categorical attributes whose value sets are evaluated by specific similarity measures (see Section 3.4), defined by the mean rank functions.

In fact, preordonance coding of the different attributes of types I or II can be considered. And this may contribute to enrich the scale value of the concerned attribute. Let us make clear this representation for the following cases : (i) *boolean* (Section 2.1) ; (ii) *numerical* (Section 2.2) ; (iii) *nominal categorical* (Section 3.1) ; (iv) *ordinal categorical* (Section 3.2) and (v) *ranking* (Section 3.3).

4.4.1 Preordonance coding of the boolean attribute

As expressed in Section 2.1, with a boolean attribute a is associated a binary categorical attribute whose value set being $\{a, \bar{a}\}$, where $\bar{a}$ is the negated boolean attribute of a. Let us index by 1 (resp., 2) the category a (resp., $\bar{a}$) and consider the following set of ordered pairs

$$B_2 = \{(1,2), (1,1), (2,2)\} \tag{82}$$

The preordonance is defined by a total preorder (ranking with ties) on B_2, and for this three alternatives can be envisaged :

1. $(1,2) < (1,1) \sim (2,2)$;
2. $(1,2) < (1,1) < (2,2)$;
3. $(1,2) < (2,2) < (1,1)$.

There is no enrichment of the scale value in the first case. However, this representation is not equivalent to that given in Section 2.1 for comparing categorical attributes (see Chapter 2 of [19]).

Between 2 and 3 the chosen preordonance structure depends on the most significant category. If a (resp., $\bar{a}$) is the most important category according to expert knowledge, then the preordonance 3 (resp., 2) is adopted. For the example given in Section 1 the category a = cirrhosis is more significant than $\bar{a}$ = non-cirrhosis. Therefore, the preordonance 3 is considered.

4.4.2 Preordonance coding of the numerical attribute

Here we consider the presentation given in Section 2.2 frm which the notations are retaken. In fact, the proposed representation is given by a categorical attribute valuated by a numerical similarity. In this case asymmetrical similarity is needed. Denote by w the categorical attribute associated with the numerical attribute v (see Section 2.2). Consider the category set of w indexed by $\{1, 2, ..., i, ..., m\}$. Then , the ordered category pairs is represented by

$$C_2 = \{(i, j) | 1 \leq i, j \leq m\} \tag{83}$$

it includes m^2 elements.

The valuation assigned to the ordered pair (i, j) is
$\xi(i, j) = x_{(j)} - x_{(i)}$, $1 \leq i, j \leq m$ (see Section 2.2). It is asymmetrical and defined by the directed range - from left to right - of the interval $[x_{(i)}, x_{(j)}]$, $1 \leq i, j \leq m$. The empirical distribution of the defined valuation is

$$\{(\xi(i, j), f_i \times f_j) | 1 \leq i, j \leq m\} \tag{84}$$

where f_i is defined in (12), $1 \leq i \leq m$.

As an example, consider the distribution given in (13). We have :

$$(\xi(2, 5), f_2 \times f_5) = ((7.2 - 2.3), 0.1 \times 0.2) = (4.9, 0.02)$$

.

Notice that generally, a 0 value for $\xi(i, j)$ may occur for $i \neq j$, $1 \leq i, j \leq m$. It occurs necessarily for $j = i$, $1 \leq i \leq m$.

4.4.3 Preordonance coding of the nominal categorical attribute

The notations of Section 3.1 are retaken here. $\{1, 2, ..., h, ..., k\}$ indexes the category set of a nominal categorical attribute c. We consider the ordered pairs of categories $K_2 = \{(g, h) | 1 \leq g \leq h \leq k\}$ (see (62)). If $g < h$, (g, h) designates the unordered pair of categories $\{g, h\}$ that we denote by gh (see (63)) ; if $g = h$, (h, h) denoted hh, is considered for the comparison between the category h with itself, $1 \leq g \leq h \leq k$.

In the established preordonance two distinct categories are considered as equally dissimilar. On the other hand, the similarity of a given category with itself is the same whatever is the concerned category. Consequently, the preordonance can be written as follows

$$12 \sim 13 \sim ... \sim 1k \sim 23 \sim 24 \sim ... \sim 2k \sim ... \sim (k-1)k < 11 \sim 22 \sim ... \sim kk \tag{85}$$

There are two classes for this total preorder on K_2. The first one comprises $k \times (k-1)/2$ elements and the second one, k elements. The corresponding mean ranks are $((k^2 - k + 2)/4)$ for one element of the first class and $(k^2 + 1)/2$, for one element of the second class. We verify easily that the mean rank sum is equal to $[(1/2) \times (k \times (k+1)/2)] \times [(k \times (k+1)/2) + 1]$.

Denote by $p = n \times (n-1)/2$ the cardinality of the set $P = \mathcal{O}^{\{2\}}$ of unordered object pairs from the object set $\mathcal{O}$, the empirical statistical distribution of the categorical attribute whose value set is K_2 can be expressed as follows

$$\{(gh, \frac{n_g \times n_h}{p}) | 1 \leq g < h \leq k\} \cup \{hh, \frac{n_h \times (n_h - 1)}{2p} | 1 \leq h \leq k\} \tag{86}$$

(see (21) and following).

By denoting $r(\pi) = card[R(\pi)]$ and $s(\pi) = card[S(\pi)]$ (see (30)), the empirical statistical distribution of the mean rank function is given by

$$\{(\frac{k^2 - k + 2}{4}, \frac{s(\pi)}{p}), (\frac{k^2 + 1}{2}, \frac{r(\pi)}{p})\} \tag{87}$$

Now, we are going to illustrate the above presentation in the framework of the example given in Section 3.1, provided by Goodman and Kruskal (1954) [9]. Three categories c_1, c_2 and c_3 were defined in Section 3.1. The preordonance (85) becomes

$$12 \sim 13 \sim 23 < 11 \sim 22 \sim 33$$

The two mean rank values associated with the two preorder classes are 2 and 5. By recalling that $n_1 = 591$, $n_2 = 608$, $n_3 = 239$ and $p = 1033203$, the empirical statistical distribution (86) is instanciated as follows

$$\begin{aligned}\big\{&(12, \frac{591 \times 608}{1033203} = 0.3478)\ ,\ (13, \frac{591 \times 239}{1033203} = 0.1367),\\ &(23, \frac{608 \times 239}{1033203} = 0.1406)\ ,\ (11, \frac{591 \times 590}{2066406} = 0.1687),\\ &(22, \frac{608 \times 607}{2066406} = 0.1786)\ ,\ (33, \frac{239 \times 238}{2066406} = 0.0275)\big\}\end{aligned}$$

We verify the value 1 for the sum of the relative frequencies. $r(\pi)$ and $s(\pi)$ are equal to 387314 and 645889, respectively (see the end of Section 3.1). Therefore, the empirical statistical distribution (87) of the mean rank function becomes

$$\big\{(2, 0.3749), (5, 0.6251)\big\}$$

4.4.4 Preordonance coding of the ordinal categorical attribute

Recall the notations of Section 3.2. The category set $\{c_h | 1 \leq h \leq k\}$ is indexed by $\{h | 1 \leq h \leq k\}$. The preordonance on the category set is defined by a total preorder on the set $H = \{(g, h) | 1 \leq g, h \leq k\}$ (see (71)) of all ordered category pairs. This will correspond to an asymmetrical ordinal similarity. More precisely, an ordered pair (g, h) is ranked according to the difference $h - g$ between the two integer codes h and g, $1 \leq g, h \leq k$. Therefore, a given class of the total preorder is defined by

$$D_e = \{(g, h) | 1 \leq g, h \leq k, h - g = e\} \tag{88}$$

where $1 - k \leq e \leq k - 1$. Its cardinality is

$$min(k, k + e) - max(1, 1 + e) + 1 \tag{89}$$

where the first term correspond to $g = k$ and the second one to $g = 1$. Therefore, the value of the mean rank function assigned to a current element of D_e is given by

$$\begin{aligned}&\sum_{1-k \leq f \leq e-1} [min(k, k + f) - max(1, 1 + f) + 1]\\ &\quad + \frac{1}{2}[min(k, k + e) - max(1, 1 + e) + 2]\end{aligned} \tag{90}$$

Now, remind that $n \times (n - 1)$ is the cardinality of the set $\mathcal{O}^{[2]}$ of ordered distinct object pairs. For $e \neq 0$, the relative frequency of D_e is given by

$$\sum_{\{(g,h) \in D_e\}} \frac{n_g \times n_h}{n \times (n - 1)} \tag{91}$$

For $e = 0$, this relative frequency becomes

$$\sum_{1 \leq g \leq k} \frac{n_g \times (n_g - 1)}{n \times (n - 1)} \tag{92}$$

(91) and (92) define the empirical statistical distribution of $h-g$, $1 \leq g, h \leq k$.

In the previous treatment we have represented a given category c_h of an ordinal categorical attribute by the integer code h, $1 \leq h \leq k$. And this is arbitrary. However, for comparing two ordered category pairs (c_g, c_h) and $(c_{g'}, c_{h'})$ of this type of attribute, it is intuitively acceptable to consider the ordinal dissimilarity between (c_g and c_h) strictly greater than that between ($c_{g'}$ and $c_{h'}$), if and only if $h-g > h'-g'$. Nevertheless, in case where $h-g = h'-g' = e \neq 0$, assigning the same ordinal dissimilarity to (c_g, c_h) and to $(c_{g'}, c_{h'})$ may seem difficult to admit. One option might consist in refining the class D_e by associating with (g, h) the relative frequency $(n_g - n_h)/n = f_g - f_h$. Thus the ordinal dissimilarity on D_e becomes an increasing function of the latter relative frequency.

Let us now illustrate this coding in our example considered in Section 3.2. The linear order defined for the three categories c_1, c_2 and c_3 is $c_1 < c_2 < c_3$. The total preorder on the set H of all ordered pairs (g, h), $1 \leq g, h \leq 3$, is defined by

$$(3,1) < (2,1) \sim (3,2) < (1,1) \sim (2,2) \sim (3,3) < (1,2) \sim (2,3) < (1,3)$$

The associated mean rank values are 1, 4.5, 5, 7.5 and 9, respectively.

By recalling that $n_1 = 591$, $n_2 = 608$ and $n_3 = 239$, the second and the fourth classes can be refined as follows

$$(2,1) < (3,2) \text{ and } (2,3) < (1,2)$$

Indeed, $n_1 - n_2 < n_2 - n_3$

Let us end this formal presentation by indicating that our focus for comparing ordinal categorical attributes is the set theoretic representation given in Section 3.2 above.

4.4.5 Preordonance coding of the ranking attribute

This coding can be interpreted as a particular case of the previous one. Each category is represented by only one object : $n_h = 1, 1 \leq h \leq k$. For this interpretation $k = n = card(\mathcal{O})$. In these conditions, the above development can be retaken, except the refinement technique. Indeed, for any element (g, h) of D_e (see 91), $n_g = n_h = 1$. But in this case, it is intuitively acceptable to consider as equivalent all the ordered pairs of D_e. Two main approaches for building association coefficients between ranking attributes are derived from the two codings given in Section 3.3 and in this section, respectively (see Chapter 2 of [19], [**?**] and [22]).

5 Attribute representations when describing a set $\mathcal{C}$ of categories

We have seen in the above sections that the basic data for describing an object set $\mathcal{O}$ by an attribute a is an ordered pair (a, o) where o is an element of $\mathcal{O}$. The value of a on o, denoted by $a(o)$ is unique. In this section the description concerns a set Γ categories. Let us designate by $\{C_1, C_2, ..., C_i, ..., C_I\}$ this set of categories :

$$\Gamma = \{C_1, C_2, ..., C_i, ..., C_I\} \tag{93}$$

We assume Γ obtained from a nominal categorical attribute γ defined on a universe Ω of objects. We designate by Ω_i the subset of Ω constituted by the objects where C_i is $TRUE$: $\Omega_i = \gamma^{-1}(C_i)$, $1 \leq i \leq I$. For this description by an attribute a, the basic data is defined here by an ordered pair (a, C), where C belongs to Γ. Relative to a given ordered pair (a, C_i), $1 \leq i \leq I$, the description of C_i by a has necessarily a global and then, a statistical nature. To fix idea, but also because of effective reason, a sample (set learning) $\mathcal{O}_{i.}$ provided from Ω_i is subtituted for C_i, $\mathcal{O}_{i.} \subset \Omega_i$, $1 \leq i \leq I$. Neverthless, cases may occur where the statistical distribution of a on C_i is directly estimated by the expert knowledge, $1 \leq i \leq I$, Lebbe et al. (1987) [15].

For a given pair $(a, \mathcal{O})$ composed by a descriptive attribute a of a fixed type and by an object set $\mathcal{O}$ described by a we have previously (see the above sections) expressed the statistical distribution of a on $\mathcal{O}$. Here, instead of one only pair $(a, \mathcal{O})$, we have a sequence of such a pair, namely

$$\big\{(a, \mathcal{O}_{i.}) | 1 \leq i \leq I\big\} \tag{94}$$

where the different sets $\mathcal{O}_{i.}$ are mutually disjoint.

Let us denote by D_a^i the statistical distribution of a on $\mathcal{O}_{i.}$. On the other hand, define $p_{i.}$ the relative frequency of the objects belonging to $\mathcal{O}_{i.}$ with respect to all the objects belonging to the union of the $\mathcal{O}_{i.}$. Explicitly,

$$p_{i.} = \frac{card(\mathcal{O}_{i.})}{card(\bigcup_{1 \leq i' \leq I} \mathcal{O}_{i.})} \tag{95}$$

$1 \leq i \leq I$.
Under these conditions the data are defined here by the sequence

$$\{(D_a^i, p_{i.})|1 \leq i \leq I\} \tag{96}$$

We have realized previously that the statistical distribution of a descriptive attribute a on an object set $\mathcal{O}$, depends on the nature of the set theoretic representation of the concerned attribute. Consequently and in order to explicit (96) for the different data description cases, we are going to remind and to highlight these statistical distributions. The attributes of type I (boolean and numerical) will be considered in Section 5.2. The nominal and ordinal categorical attributes will be studied in Section 5.3 and finally, the preordonance attributes, in Section 5.4. In Section 5.5 we shall formalize the data table notion by means of two relational systems : The Tarski system $\mathcal{T}$ and a statistical system that we shall designate by $\mathcal{S}$.

5.1 Attributes of type I

Let us refer to the above Section 2. We retake here the same notations. The considered attributes are the boolean attribute (Section 2.1) and the numerical attribute (Section 2.2). For the boolean attribute describing an object set $\mathcal{O}$, the relative frequency (proportion) $p(a) = n(a)/n$ has been defined and then, the sequence of distributions (96) becomes

$$\{(p_a^i, p_{i.})|1 \leq i \leq I)\} \tag{97}$$

where p_a^i is a proportion defined at the level of the object set $\mathcal{O}_{i.}$. More precisely, consider the subset $\mathcal{O}_{i.}(a)$ of $\mathcal{O}_{i.}$ where a is $TRUE$ and denote by $n_{i.}$ the cardinality of $\mathcal{O}_{i.}$ and by $n_{i.}(a)$ the cardinality of $\mathcal{O}_{i.}(a)$, we have $p_a^i = n_{i.}(a)/n_{i.}$, $1 \leq i \leq I$.

Now, let us consider the case of a numerical attribute v whose distribution was expressed in the above equation (12). The distribution sequence can be written

$$\{\{(x_{(l)}^i, f_{(l)}^i)|1 \leq i \leq I\}\} \tag{98}$$

where the distribution $(x_{(l)}^i, f_{(l)}^i)$ is defined at the level of the object set $\mathcal{O}_{i.}$, $1 \leq i \leq I$.

5.2 Nominal or ordinal categorical attributes

5.2.1 The case of the nominal categorical attribute

This case is particularily instructive. Different representation levels have been considered for this type of attribute (see Section 3.1). The most basic one is defined by (20). The latter gives a valuation at the level of the object set $\mathcal{O}$. This valuation assigns to each object o ($o \in \mathcal{O}$) a code representing the value $c(o)$ of the categorical attribute c. According to the notations of Section 3.1, we denote by $\{f_h^i|1 \leq h \leq k\}$ the statistical distribution of c on the set $\mathcal{O}_{i.}$, $1 \leq i \leq I$. In other words, $f_h^i = n_{ih}/n_{i.}$ is the relative frequency (proportion) of objects from $\mathcal{O}_{i.}$ possessing the h^{th} category of the attribute c. As just defined above, $n_{i.} = card(\mathcal{O}_{i.})$. On the other hand, $n_{ih} = card(\mathcal{O}_{i.h})$ where $\mathcal{O}_{i.h} = c^{-1}(h) \cap \mathcal{O}_{i.}$, $1 \leq i \leq I$, $1 \leq h \leq k$. Thus, the distribution sequence (96) can be put in the following form

$$\{(\{f_h^i|1 \leq h \leq k\}, p_{i.})|1 \leq i \leq I\} \tag{99}$$

Let us notice here that this data structure is exactly that addressed by *Correspondence Analysis and related methods* Benzecri (1973) [2] (see also Chapters 2 and 6 of [19]). Using these techniques have meaning only when k and I are large enough. Now, let us consider the relational representation given by the equations (22) to (29). According to the notations of Section 4.4.3, $r(\pi)$ and $s(\pi)$ denote the cardinalities of the sets $R(\pi)$ and $S(\pi)$, respectively (see (30)). The statistical distribution of the indicator function ρ_π (see (26)) can be put in the following form

$$\{(1, \frac{r(\pi)}{p}), (0, \frac{s(\pi)}{p})\} \tag{100}$$

where $p = card(P) = n(n-1)/2$.

Let us denote by π^i the partition on $\mathcal{O}_{i.}$ induced by the categorical attribute c, namely

$$\pi^i = \{\mathcal{O}_{i.h} = c^{-1}(h) \cap \mathcal{O}_{i.} | 1 \leq h \leq k\},$$

$1 \leq i \leq I$.

$r(\pi^i)$ (resp., $s(\pi^i)$) designates the number of object pairs joined (resp., separated) by π^i, $1 \leq i \leq I$. Also, denote by $p^i = n_{i.} \times (n_{i.} - 1)/2$ the number of object pairs of $\mathcal{O}_{i.}$, $1 \leq i \leq I$. With these notations, the sequence of the statistical distributions (96) becomes

$$\{(\{(1, \frac{r(\pi^i)}{p^i}), (0, \frac{s(\pi^i)}{p^i})\}, \frac{p^i}{sump}) | 1 \leq i \leq I\} \tag{101}$$

where $sump = \sum_{1 \leq i' \leq I} p^{i'}$.

Indeed, the weighting $p^i/sump$ associated with the distribution D_c^i is obtained in a relative way from the cardinality of the set $\mathcal{O}_{i.}$ on which D_c^i is defined. However, this statistical distribution - defined by the previous equation - is too global in order to compare finely the different categories of Γ (see (93)), represented, respectively, by the sets $\mathcal{O}_{i.}$, $1 \leq i \leq I$. A discriminant representation has to take into account the decomposition of the sets $R(\pi)$ and $S(\pi)$ (see equations (29) and (30)). In these conditions, the sequence of the statistical distributions (96)

$$\{(\{\{f_{hh}^i | 1 \leq h \leq k\}, \{f_{gh}^i | 1 \leq g < h \leq k\}\}, \frac{p^i}{sump}) | 1 \leq i \leq I\} \tag{102}$$

where $f_{hh}^i = (n_{ih} \times (n_{ih} - 1))/(n \times (n-))$ and
$f_{gh}^i = (n_{ig} \times n_{ih})/(n \times (n-1))$, $1 \leq h \leq k$, $1 \leq g < h \leq k$, $1 \leq i \leq I$. The relative frequencies f_{hh}^i and f_{gh}^i are defined with respect to the set $\mathcal{O}_{i.}^{\{2\}}$ of unordered object pairs.

In Section 4.4.3 a representation in terms of a preordonance attribute was proposed for the nominal categorical attribute. This preordonance was coded from the mean rank function. The statistical distribution of this function (see (87)) refines that (101). We obtain

$$\{\{(\frac{k^2 - k + 2}{4}, \frac{s(\pi^i)}{p^i}), (\frac{k^2 + 1}{2}, \frac{r(\pi^i)}{p^i})\}, \frac{p^i}{sump} | 1 \leq i \leq I\} \tag{103}$$

However, in spite of this refinement, comparing the categories C_i, $1 \leq i \leq I$ of Γ, on the basis of such distributions, remains not discriminant enough. Finally, two statistical relational representations have to be retained for this comparison (99) and (102).

5.2.2 The case of the ordinal categorical attribute

The development of this case follows the same rationale as the preceding one. Assigning to each object of $\mathcal{O}$ a category (see (20)) provides a representation of the categorical attribute c at the level of $\mathcal{O}$. This representation is defined by a valuation on $\mathcal{O}$. In these conditions (99) is retaken exactly. Therefore, the linear order (32) on the category set $\mathcal{C} = \{c_1, c_2, ..., c_h, ..., c_k\}$ cannot be taken into account.

Now, let us consider the relational representation of the categorical attribute c (see equations from (36) to (44)). Denote by $r(\omega^i) = card[R(\omega^i)])$, $e(\omega^i) = card[E(\omega^i)])$ and $s(\omega^i) = card[S(\omega^i)])$ respectively, the cardinalities defined in (40) and restricted to the set $\mathcal{O}_{i.}$. Then, denote $c_{i.}$ the cardinality of $\mathcal{O}_{i.}^{[2]}$ (set of distinct object pairs from $\mathcal{O}_{i.}$), $c_{i.} = n_{i.} \times (n_{i.} - 1)$. By considering the $score_\omega$ function defined in (44), the sequence of the statistical distribution (96) can be written

$$\{(\{(1, \frac{r(\omega^i)}{c_{i.}}), (0.5, \frac{e(\omega^i)}{c_{i.}}), (0, \frac{s(\omega^i)}{c_{i.}})\}, \frac{c_{i.}}{sumc}) | 1 \leq i \leq I\} \tag{104}$$

where $sumc = \sum_{1 \leq i' \leq I} c^{i'}$.

This distribution sequence extends to the ordinal case that (101) considered previously for the nominal case.

Now, let us consider the extension of the (102) expression in the ordinal case. This expression is written here exactly in the same manner, namely

$$\big\{(\{\{f^i_{hh} | 1 \leq h \leq k\}, \{f^i_{gh} | 1 \leq g < h \leq k\}\}, \frac{c^i}{sumc}) | 1 \leq i \leq I\big\} \tag{105}$$

where f^i_{hh} and f^i_{gh} have exactly the same meaning as for the nominal case; but here they derive from counting ordered pairs of distinct objects.

As an example and also as an exercise consider the table Table 3 below provided from Goodman and Kruskal (1954) [9]. These data have been already employed for illustration (see Section 3). Table 3 is a contingency table crossing two classifications. These are associated with two categorical attributes. In Sections 3.1 and 3.2 we have considered the categorical attribute c : "Highest level of formal education of wife". Its values denoted c_1, c_2 and c_3 (see Section 3.1) index the columns of the contingency table (see table 3). The rows of Table 3 are indexed by the values of the categorical attribute "Fertility planning status of couple". These values denoted D, C, B and A, are ranked from the lowest level D to the highest one A. Here, $I = 4$. On the other hand, $card(\mathcal{O}_{1.}) = 379$, $card(\mathcal{O}_{2.}) = 451$, $(\mathcal{O}_{3.}) = 205$ and $card(\mathcal{O}_{4.}) = 403$.

$\mathcal{F} \diagdown \mathcal{L}$	c_1	c_2	c_3	Row totals
D	223	122	34	379
C	168	215	68	451
B	90	80	35	205
A	110	191	102	403
Column totals	591	608	239	1438

Table 3 : Crossing between "Fertility planning" and "Highest level of formal education of wife"

Let us now illustrate the above distribution (104). We suppose as in Section 3.2, the linear order $c_1 < c_2 < c_3$ for the values of the attribute c. For this illustration we are going to give the contribution of $i = 2$ to the (104) expression. We have

$$\begin{aligned}
r(\omega^2) = card[R(\omega^2)] &= 168 \times 215 + 168 \times 68 + 215 \times 68 = 62164 \\
s(\omega^2) = card[S(\omega^2)] &= 215 \times 168 + 68 \times 168 + 68 \times 215 = 62164 \\
e(\omega^2) = card[E(\omega^2)] &= 168 \times 167 + 215 \times 214 + 68 \times 67 = 62164 \\
c^2 = card(\mathcal{O}^{[2]}_{2.}) &= 451 \times 450 = 202950 \\
sumc = c^1 + c^2 + c^3 &= 379 \times 378 + 451 \times 450 + 205 \times 204 = 388032 \\
\frac{c^2}{sumc} &= 0.5230 \\
\frac{r(\omega^2)}{sumc} = 0.3063, \frac{s(\omega^2)}{sumc} &= 0.3063 \text{ and } \frac{e(\omega^2)}{sumc} = 0.3874.
\end{aligned} \tag{106}$$

Thus, the contribution of $i = 2$ to (104) is

$$\big(\{(1, 0.3036), (0.5, 0.3874), (0, 0.3063)\}, 0.5230\big)$$

The statistical distributions (101), (103)or (104) are too global for comparison purpose in case of nominal or ordinal categorical attributes. The most accurate statistical distributions for comparing the categories of Γ ($\Gamma = \{A, B, C, D\}$ in the above example), are (102) for the nominal case and (105) for the ordinal one. Illustrating these distributions in the framework of our example is left for the reader.

5.3 Preordonance or numerical similarity categorical attributes

In Section 4.1 the preordonance categorical attribute was defined. It has been represented from the mean rank function (see (69)) as a categorical attribute valuated by a specific numerical similarity. This type of attribute was presented in Section 3.4. We have seen that the taxonomic categorical attribute (Section 4.3) is represented as a specific preordonance on the the set of the leaves of the taxonomic tree. Consequently, it suffices, without any loss of generality, to give the form of the distribution sequence (96) in case of a categorical attribute valuated by a numerical similarity. By reconsidering the notations of Section 3.4 with respect to the set of ordered distinct object pairs of $\mathcal{O}_{i.}$, $1 \leq i \leq I$, we obtain

$$\{(\{\{f^i_{hh} | 1 \leq h \leq k\}, \{f^i_{gh} | 1 \leq g \neq h \leq k\}\}, \frac{c^i}{sumc}) | 1 \leq i \leq I\} \tag{107}$$

where $f^i_{hh} = (n_{ih} \times (n_{ih} - 1))/(n \times (n-1))$ and
$f^i_{gh} = (n_{ig} \times n_{ih})/(n \times (n-1))$, $1 \leq h \leq k$, $1 \leq g \neq h \leq k$, $1 \leq i \leq I$. c^i and $sumc$ have been defined above. The relative frequencies are computed for distinct object ordered pairs. More precisely, f^i_{hh} is the relative frequency in $\mathcal{O}^{[2]}_{i.}$ whose value is $\xi(c_h, c_h)$, $1 \leq h \leq k$, f^i_{gh} is the relative frequency in $\mathcal{O}^{[2]}_{i.}$ whose value is $\xi(c_g, c_h)$, $1 \leq g \neq h \leq k$.

Let us now illustrate the statistical distribution (106) in the case of our example (see Table 2 and Table 3). Consider $i = 2$. We have $n_{i.} \times (n_{i.} - 1) = 451 \times 450 = 202950$.

$$\begin{aligned}
f^2_{11} = \frac{168 \times 167}{202950} = 0.1382 \ , \ f^2_{22} = \frac{215 \times 214}{202950} = 0.2267 \\
f^2_{33} = \frac{68 \times 67}{202950} = 0.0224 \\
f^2_{12} = \frac{168 \times 215}{202950} = 0.1780 \ , \ f^2_{13} = \frac{168 \times 68}{202950} = 0.0563 \\
f^2_{21} = \frac{215 \times 168}{202950} = 0.1780 \ , \ f^2_{23} = \frac{215 \times 68}{202950} = 0.0720 \\
f^2_{31} = \frac{68 \times 168}{202950} = 0.0563 \ , \ f^2_{32} = \frac{68 \times 215}{202950} = 0.0720
\end{aligned} \tag{108}$$

5.4 The data table : a Tarsky system $\mathcal{T}$ or a statistical system $\mathcal{S}$

In Sections 2, 3 and 4, the description of a set $\mathcal{O}$ of objects by a descriptive attribute a has been defined. Relative to an ordered pair (a, o) constituted by a descriptive attribute a and by an object o, $a(o)$ designates the value of a on o. a is represented by a mapping of $\mathcal{O}$ onto the value scale $\mathcal{E}$. We assume that $\mathcal{E}$ is endowed with a relation r^a. This relation is unary in Section 2 (boolean and numerical attributes), binary in Section 3 (nominal and ordinal categorical attributes, ranking attribute), binary on the set of object pairs in Section 4 (preordonance and taxonomic attributes). Thus, a very large range of attribute description in *Combinatorial Data Analysis* and in *Machine Learning* is covered. In addition, any arity of this relation can be handled Lerman (1999) [24], Lerman and Rouxel (2000) [33]. r^a induces on $\mathcal{O}$ a relation that we denote by R^a. Thus, the description of $\mathcal{O}$ by a can be formalized by the ordered pair $(\mathcal{O}, R^a)$.

Generally, for the description of an object set $\mathcal{O}$ provided from a universe $\mathcal{U}$ of objects (see Section 1) the expert defines a vast set of descriptive attributes. Let us denote by $\mathcal{A} = \{a^j | 1 \leq j \leq p\}$ this set, where p is the number of attributes. Then, the very important notion of a data table T crossing an object set $\mathcal{O}$ with an attribute set $\mathcal{A}$ can be expressed (see Table 4). $\mathcal{O}$ that we denote by $\{o_i | 1 \leq i \leq n\}$ indexes the row set of T whereas, the attribute set $\mathcal{A}$ indexes the column set of T; A given cell is established at the intersection of the i^{th} row and the j^{th} column, it contains the value $a^j(o_i)$, $1 \leq i \leq n\}$, $1 \leq j \leq p$.

$\mathcal{O} \diagdown \mathcal{A}$	a^1	$\cdots$	a^j	$\cdots$	a^p
o_1	$a^1(o_1)$	$\cdots$	$a^j(o_1)$	$\cdots$	$a^p(o_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
o_i	$a^j(o_i)$	$\vdots$	$a^j(o_i)$	$\vdots$	$a^p(o_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
o_n	$a^1(o_n)$	$\cdots$	$a^j(o_n)$	$\cdots$	$a^p(o_n)$

Table 4 : Data table T

As said above, each descriptive attribute induces a relation on the object set $\mathcal{O}$. Let us designate by R^j the relation on $\mathcal{O}$ defined by the a^j attribute. Let us emphasize once again that this relation is induced from the relation which endows the value scale of a^j, $1 \leq j \leq p$. Hence, we associate with the data table T the Tarski system Tarski (1954) [48].

$$\mathcal{T} = \left(\mathcal{O}; R^1, R^2, ..., R^j, ..., R^p\right) \tag{109}$$

where the different relations R^j have not necessarily the same arity, $1 \leq j \leq p$.

In clustering there are two dual problems. The first one which is the most familiar and mostly the only considered consists of organizing by proximity into a classification structure the set $\mathcal{O}$ of objects. The concerned proximity notion is global with respect to the different relations R^j, $1 \leq j \leq p$, by integrating all of them. For a given relation R^j and two objects o_i and $o_{i'}$, the higher the proximity between o_i and $o_{i'}$, the more linked they are with respect to the relation R^j, $1 \leq j \leq p$, $1 \leq i < i' \leq n$.

The second problem is dual of the first one. It consists of organizing by proximity into a classification scheme the set $\mathcal{A}$ of descriptive attributes ; that is to say and according to our formalism, the set $\{R^j | 1 \leq j \leq p\}$ of relations on $\mathcal{O}$, associated with the descriptive attributes, respectively. A fundamental task consists in building a proximity notion between two relations R^j and R^k on $\mathcal{O}$, $1 \leq j \leq p$. A numerical version of this gives rise to the notion of association coefficient between descriptive attributes [21, 22]. A complementary and very instructive clustering analysis consists of organizing the whole set of the categories taking part in the different categorical attributes [34, 47].

The implicit principle of the *Likelihood Linkage Analysis (LLA)* ascendant hierarchical clustering approach Lerman (1993) [23] is to cluster the set $\mathcal{A}$ of descriptive attributes before the set $\mathcal{O}$ of the described objects [27]. An ultimate stage consists of interpreting by crossing techniques each clustering with respect to the other one (see Chapter 3 of [19]), [26]. Condition of same arity for the different relations R^j, $1 \leq j \leq p$ is required in order to cluster them. In practice and mostly, this condition is brought back Lerman (1992) [21] [22], Ouali-Allah (1991) [37]. Moreover, statistical homogeneity in the description by the different attributes a^j, $1 \leq j \leq p$, is also an intuitive necessary condition. More explicitly and for example, for a description by nominal categorical attributes, the number of categories by attribute has to be of the same order.

In the LLA approach the two above requested conditions (same arity of the relations R^j $(1 \leq j \leq p)$ and same statistical homogeneity of the different descriptive attributes) for clustering the set $\mathcal{A}$, are no more required for clustering the object set $\mathcal{O}$ [20, 31].

We have presented above the description of a set $\Gamma = \{C_i | 1 \leq i \leq I\}$ of categories by an attribute interpreted as a relation on a set of objects. We have associated with the category C_i a learning set $\mathcal{O}_{i.}$ composed of objects belonging to the category C_i, $1 \leq i \leq I$. If R denotes the relation endowing the value set of the concerned attribute, we have represented R by its statistical distribution on $\mathcal{O}_{i.}$, $1 \leq i \leq I$. Thus Γ is represented by a sequence of statistical distributions whose general form is defined in (96) (see also (105)). Each distribution is computed for one category. It is weighted according to the relative frequency of the learning set that represents it.

Now, assume a set $\mathcal{A} = \{a^1, a^2, ..., a^j, ..., a^p, \}$ of p relational attributes. As expressed above, these define a sequence of p relations on each of the sets $\mathcal{O}_{i.}$, $1 \leq i \leq I$. We denote by $\{R_i^1, R_i^2, ..., R_i^j, ..., R_i^p\}$ this sequence of relations on $\mathcal{O}_{i.}$, $1 \leq i \leq I$. Notice that for category description, what is retained from R_i^j is its statistical distribution on $\mathcal{O}_{i.}$, $1 \leq j \leq p$, $1 \leq i \leq I$ (see Sections 5.1 to 5.4 when only one set is considered). Thus, we are led to define a system $\mathcal{S}$ as follows :

$$\mathcal{S} = \left(\Gamma; R^1, R^2, ..., R^j, ..., R^p\right) \tag{110}$$

where for each pair (C_i, R^j) $(C_i \in \Gamma)$, the statistical distribution of R^j on C_i is estimated on the basis of a learning set $\mathcal{O}_{i.}$, $1 \leq i \leq I$, $1 \leq j \leq p$. In these conditions, the data table takes the following form

$\Gamma \diagdown \mathcal{A}$	a^1	$\cdots$	a^j	$\cdots$	a^p
C_1	$D^1(a^1)$	$\cdots$	$D^1(a^j)$	$\cdots$	$D^1(a^p)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_i	$D^i(a^j)$	$\vdots$	$D^i(a^j)$	$\vdots$	$D^i(a^p)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_I	$D^I(a^1)$	$\cdots$	$D^I(a^j)$	$\cdots$	$D^I(a^p)$

Table 5 : Data table S

As it was remarked above, the contingency table can be formalized by a system $\mathcal{S}_1 = (\Gamma, R^1)$ where only one relation is considered, the latter being defined by a nominal categorical attribute a^1. R^1 defines a partition on each of the sets $\mathcal{O}_{i.}$, $1 \leq i \leq I$. Designate by $\mathcal{C}^1 = \{c_1^1, ..., c_g^1, ..., c_h^1, ..., c_k^1\}$ the category set of a^1. As said above, this data structure is the main matter of *Correspodence Analysis* Benzecri (1973) [2, 34, 47]. In this approach, basically, the comparison between two categories γ_i and $\gamma_{i'}$ of Γ is evaluated from $\{|f_h^i - f_h^{i'}| 1 \leq h \leq k\}$ where f_h^i (resp., $f_h^{i'}$) is the relative frequency in $\mathcal{O}_{i.}$ (resp., $\mathcal{O}_{i'.}$) of objects having the category h, $1 \leq i < i' \leq I$. Let us suppose now a structure on the category set $\mathcal{C}^1$, given for example by a valuated binary relation (see Section 3.4). This structure cannot in anyway be taken into account in the mentioned approach. On the contrary, clustering approach as that given by the LLA method is able to integrate

$$\left\{(val^1(g,h), f_g^i \times f_h^i) | 1 \leq g < h \leq k\right\}$$

and

$$\left\{(val^1(g,h), f_g^{i'} \times f_h^{i'}) | 1 \leq g < h \leq k\right\} \tag{111}$$

As for the data defined by a Tarski system, it is matter for the data defined by an $\mathcal{S}$ system to organize by clustering the set Γ of categories and the set $\mathcal{A}$ of descriptive attributes. Clustering $\mathcal{A}$ consists of clustering the relation set $\{R^j | 1 \leq j \leq k\}$ on the basis of estimated statistical distributions on the different categories of Γ (see above). On the other hand and particularily for this structure, a fundamental problem for data analysis consists of clustering the whole set of the categories defined by the different attributes a^j, $1 \leq j \leq k$. Let us notice that we obtain the structure of an horizontal juxtaposition of contingency tables Lerman and Tallur (1980) [34], Tallur (1988) [47], where the categorical attributes a^j, $1 \leq j \leq p$, are nominal.

To end, let us emphasize the importance of interpreting in a comparative manner a clustering of Γ, a clustering of $\mathcal{A}$ and also, a clustering of the set of categories taking part in the definition of $\mathcal{A}$. For this purpose specific tools are studied in [26].

Références

[1] M.R. ANDERBERG. *Cluster Analysis for Applications*. Academic Press, 1973.

[2] J.P. BENZÉCRI. *L'analyse des données, tome II*. Dunod, 1973.

[3] G. CELEUX and G. GOVAERT. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, (47) :127–146, 1993.

[4] S. CHAH. Critères de classification sur des donées hétérogènes. *Revue de Statistique Appliquée*, (tome 3 (2)) :19–36, 1985.

[5] W.D. FISHER. On grouping with maximum homogeneity. *Journal of the American Statistical Association*, (53) :4–29, 1958.

[6] A. GHAZZALI, N.and LÉGER and I.C. LERMAN. Rôle de la classification statistique dans la compression du signal image : panorama et une étude spécifique de cas. *La Revue de Modulad*, (14) :51–89, Décembre 1994.

[7] N. GHAZZALI. *Comparaison et réduction d ' arbres de classification, en relation avec des problèmes de quantification en imagerie numérique*. PhD thesis, Université de Rennes 1, mai 1992.

[8] V. GIAKOUMAKIS and B. MONJARDET. Coefficients d ' accord entre deux préordres totaux. *Mathématiques et Sciences Humaines*, (98) :69–87, 1987.

[9] L.A. GOODMAN and W.H. KRUSKAL. Measures of association for cross classifications. *Journal of the American Statistical Association*, (49) :732–764, December 1954.

[10] A. GUÉNOCHE and B. MONJARDET. Méthodes ordinales et combinatoires en analyse des données. *Mathématiques et Sciences Humaines*, (100) :5–47, 1987.

[11] L.J. HUBERT. Assignment methods in combinatorial data analysis. In *Numerical Taxonomy*. Marcel Dekker, New York, 1987.

[12] M.G. KENDALL. *Rank correlation methods*. Charles Griffin, 1970, first edition 1948.

[13] A.M. KERJEAN. *Tentative d'établissement de 100 typologies d'examens biologiques. Contribution à l'établissement du système "A.D.M.". Doctorat d'État*. PhD thesis, Université de Rennes 1, 1978.

[14] J.Y. LAFAYE. Une méthode de discrétisation de variables continues. *Revue de Statistique Appliquée*, (27) :39–53, 1979.

[15] J. LEBBE, J.P. DEDET, and R. VIGNES. Identification assistée par ordinateur des phlébotomes de la Guyane Française. Publication Interne Version 1.02, Institut Pasteur de la Guyane Française, Juillet 1987.

[16] I.-C. LERMAN and P. KUNTZ. Directed binary hierarchies and directed ultrametrics. *Journal of Classification*, (28) :272–296, October 2011.

[17] I.C. LERMAN. *Les bases de la classification automatique*. Gauthier-Villars, 1970.

[18] I.C. LERMAN. Étude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique. *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, (19) :1–52, 1973.

[19] I.C. LERMAN. *Classification et analyse ordinale des données*. Dunod, 1981.

[20] I.C. LERMAN. Construction d ' un indice de similarité entre objets décrits par des variables d ' un type quelconque. application au problème de consensus en classification. *Revue de Statistique Appliquée*, (XXXV (2)) :39–60, 1987.

[21] I.C. LERMAN. Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, i. *Revue Mathématique Informatique et Sciences Humaines*, (118) :35–52, 1992.

[22] I.C. LERMAN. Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, ii. *Revue Mathématique Informatique et Sciences Humaines*, (119) :75–100, 1992.

[23] I.C. LERMAN. Likelihood linkage analysis (lla) classification method (around an example treated by hand). *Biochimie*, (75) :379–397, 1993.

[24] I.C. LERMAN. Comparing classification tree structures : a special case of comparing q-ary relations. *RAIRO-Operations Research*, (33) :339–365, September 1999.

[25] I.C. LERMAN. Comparing taxonomic data. *Revue Mathématiques et Sciences Humaines*, (150) :37–51, 2000.

[26] I.C. LERMAN. Coefficient numérique général de discrimination de classes d ' objets par des variables de types quelconques. application à des données génotypiques. *Revue de Statistique Appliquée*, (LIV(2)) :33–63, 2006.

[27] I.C. LERMAN. Analyse de la vraisemblance des liens relationnels une méthodologie d ' analyse classificatoire des données. In Younès Benani and Emmanuel Viennet, editors, *RNTI A3, Revue des Nouvelles Technologies de l'Information*, pages 93–126. Cèpaduès, 2009.

[28] I.C. LERMAN and S. GUILLAUME. Comparaison entre deux indices pour l'évaluation probabiliste discriminante des règles d ' association. In Ali Khenchaf and Pascal Poncelet, editors, *EGC'2011, RNTI E.20*, pages 647–656. Hermann, 2011.

[29] I.C. LERMAN and Ph. PETER. Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle. In *Data Analysis and Informatics IV*, pages 121–136. North Holland, 1986.

[30] I.C. LERMAN and Ph. PETER. Classification en présence de variables préordonnances taxonomiques à choix multiple. application à la structuration des phlébotomes de la Guyane Française. Publication Interne 426, IRISA-INRIA, Septembre 1988.

[31] I.C. LERMAN and Ph. PETER. Indice probabiliste de vraisemblance du lien entre objets quelconques : analyse comparative entre deux approches. *Revue de Statistique Appliquée*, (LI(1)) :5–35, 2003.

[32] I.C. LERMAN and Ph. PETER. Representation of concept description by multivalued taxonomic preordonance variables. In G. Cucumel P. Brito, P. Bertrand and F. Carvalho, editors, *Selected Contributions in Data Analysis and Classification*, pages 271–284. Springer, 2007.

[33] I.C. LERMAN and F. ROUXEL. Comparing classification tree structures : a special case of comparing q-ary relations ii. *RAIRO-Operations Research*, (34) :251–281, july/September 2000.

[34] I.C. LERMAN and B. TALLUR. Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, (28) :5–28, 1980.

[35] F. MARCOTORCHINO. Essai de typologie structurelle des indices de similarité vectoriels par unification relationnelle. In Younès Benani and Emmanuel Viennet, editors, *RNTI A3, Revue des Nouvelles Technologies de l'Information*, pages 203–318. Cèpaduès, 2009.

[36] F. MARCOTORCHINO and P. MICHAUD. *Optimisation en analyse ordinale des données*. Masson, 1979.

[37] M. OUALI-ALLAH. *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques*. PhD thesis, Université de Rennes 1, décembre 1991.

[38] Ph. PETER. *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d ' informations assistée par ordinateur*. PhD thesis, Université de Rennes 1, mars 1987.

[39] S. RABASEDA, R. RAKOTOMALALA, and M. SEBBAN. Discretization of continuous attributes : a survey of methods. In *Proceedings of the second Annual Joint Conference on Information Sciences*, pages 164–166, 1995.

[40] S. RÉGNIER. Sur quelques aspects mathématiques des problèmes de la classification automatique. *I.C.C. Bulletin*, (4) :175–191, 1965.

[41] A. SCHROEDER. Analyse d ' un mélange de distributions de même type. *Revue de Statistique Appliquée*, (24) :53–62, 1976.

[42] C. SPEARMAN. The proof and measurement of association between two things. *The American Journal of Psychology*, (1) :72–101, Vol. 15 1904.

[43] S.S. STEVENS. Mathematics, measurement and psychophysics. In S.S. Stevens, editor, *Handbook of experimental psychology*, pages 1–49. New York : Wiley, 1951.

[44] P. SUPPES and J.L. ZINNES. Basic measurement theory. In R.R. Bush R.D. Luce and E.H. Galanter, editors, *Handbook of Mathematical Psychology, Vol. 1*, pages 3–76. Wiley, 1951.

[45] J.P. SUTCLIFFE. Concept, class, and category in the tradition of aristotle. In *Categories and concepts : Theoretical news and inductive data analysis*, pages 35–65. Academic Press, 1992.

[46] M.J. SYMONS. Clustering criteria and multivariate normal mixture. *Biometrics*, (37) :35–43, 1981.

[47] B. TALLUR. *Contribution à l'analyse exploratoire de tableaux de contingence par la classification, Doctorat d´État*. PhD thesis, Université de Rennes 1, 1988.

[48] A. TARSKI. Contribution to the theory of models. *Indagationes Mathematicae*, (16) :572–588, 1954.

[49] J.H. WARD. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, (58) :236–244, 1963.