Representation of Concept Description by Multivalued Taxonomic Preordonance Variables

Israël-César Lerman[1] and Philippe Peter[2]

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

[2] Lina, École Polytechnique de l'Université de Nantes
- La Chantrerie - BP 50609 - 44306 Nantes Cédex, France,
Philippe.Peter@polytech.univ-nantes.fr.

Abstract. Mathematical representation of complex data knowledge is one of the most important problems in Classification and Data Mining. In this contribution we present an original and very general formalization of various types of knowledge. The specific data are endowed with biological descriptions of phlebotomine sandfly species. Relative to a descriptive categorical variable, subsets of categories values have to be distinguished. On the other hand, hierarchical dependencies between the descriptive variables, associated with the mother → daughter relation, have to be taken into account. Additionally, an ordinal similarity function on the modality set of each categorical variable. The knowledge description is formalized by means of a new type of descriptor that we call "Taxonomic preordonance variable with multiple choice". Probabilistic similarity index between concepts described by such variables can be built.

1 Introduction

An early work (Lerman and Peter (1988), Lerman and Peter (1989)) is revisited here in a clearer, more synthetic and more accurate manner. In order to build similarity indices between complex descriptions, a mathematical representation of structured data by a knowledge expert is needed. This subject is becoming more and more important in Classification and Data Mining (Batagelj (1989), Bock and Diday (2000), Lerman (2000), Pennerath and Napoli (2006)). This work results from a collaboration with the late Jacques Lebbe.

This collaboration took place when Diday introduced the general idea of logical knowledge data analysis that he called "symbolic" data analysis Diday (1989). In this case and for a description of an objects set by attributes, the attribute value on a given object is not necessarily reduced to a single element of the scale associated with the concerned attribute. In other words the description system (attribute, single value) is left and substituted by the system (attribute, knowledge value). For example let us consider a knowledge value of a categorical attribute on a given object; this can be endowed with a logical formula on the category set, satisfied by the described object. Another example can be given by a probability distribution over the category set expressing uncertainty for the attributed value. In our subsequent development we consider only qualitative descriptions. A categorical attribute will also be called qualitative variable and a category value is expressed in terms of modality of the concerned qualitative variable.

Often, in “symbolic” data analysis papers qualitative data analysis is improperly interpreted as belonging to the “symbolic” domain. For this reason we prefer to speak in terms of “knowledge” data analysis. Furthermore, the notion of a classical data table which crosses an object set with an attribute set, is neglected and even rejected for knowledge description in (Diday (1989)). However, some evolution can be noticed in (Billard and Diday (2003)). From the begining (1988) the general notion of data table has played a fundamental part in our approach of knowledge data analysis. The only distinction considered is defined by the difference in nature of the cell content corresponding to the value of a descriptive variable on a given object. Semantic data relative to the scale associated with the value set of a given attribute can be recorded separately. On the other hand, logical relationships between descriptive variables have to be integrated in order to build the most synthetic attributes. In this paper we will be concerned by this type of construction leading to a very general and multivalued structured attribute called “taxonomic preordonance variable with multiple choice”.

This type of descriptive variable or “descriptor” has been obtained by a formalization of the expert knowledge of the biological descriptions of phlebotomine sandflies of French Guiana (Lebbe et al. (1987)). Descriptions are very complex. Each species is a class of specimens and its description must represent not only a prototype, but all possible variations in the species. Thus, the description by a qualitative variable of a given species, requires - most often - a subset of possible modalities. For sake of generality, we assume that the value of a given variable on a given species is defined by a probability distribution on a collection of modality subsets of this variable. Moreover, descriptive attributes are related by the mother → daughter relation; that is to say, if (v^0, v^1) is a such ordered pair of variables, v^1 is only defined when v^0 takes some of its values. Finally, we assume an ordinal similarity function on the modality set of each variable. A mathematical coding of this function in terms of a binary weighted relation is given in Section 3. In order to address the problem of conceptual knowledge description, Section 2 introduces the general notion of qualitative variable with multiple choice. The mother → daughter relations among the descriptive attributes lead to taxonomic variables organizing the initial qualitative variables (Section 4). By combining this structuration with local ordinal similarities, established on the respective modality sets of the different qualitative variables, we obtain the "taxonomic preordonance variable". Its construction and its mathematical coding are discussed in Section 5. In our description the components of a taxonomic preordonance variable are qualitative attributes with multiple choice. By integrating this descriptive property, "taxonomic preordonance variable with multiple choice" is derived. Section 6 is devoted to clarify the value set of a such variable. A similarity index between two concepts (or classes) described by this variable, has been proposed in (Lerman and Peter (1988), Lerman and Peter (1989)). Relative to a description by many taxonomic preordonance variables with multiple choice, a statistical normalization process was considered in order to establish a probabilistic similarity index. The latter is employed in the LLA (Likelihood of the Linkage Analysis) hierarchical classification method (Lerman (1993), Lerman and Peter (1988), Lerman and Peter (1989)). For concision reasons, these last aspects cannot be reported in this paper.

2 Qualitative variable with multiple choice

As mentioned above the data which have motivated this work are knowledge biological descriptions of species of phlebotomine sandflies of French Guiana (Lebbe et al. (1987)). Let us consider the 33rd variable of this description: "Aspect of individual duct". Its modalities are:

1. Smooth non-sclerotized
2. Smooth sclerotized
3. Transversely striated or annulated
4. With small prominent tubercles

The knowledge description of a given species (e.g. Lutzomyi carvalhoi) can be expressed as follows: "Specimens of this species have the value 1 and others of the same species have the value 3".

In these conditions, the value of the qualitative variable with multiple choice is defined by the modality subset $\{1, 3\}$, or equivalently by the conjunction 1&3. Thus, a qualitative variable with multiple choice is directly deduced from an ordinary qualitative (categorical) variable, for concept (one may also say class) description. For this, a given value is then defined in terms of a modality subset of the initial variable, or equivalently, in terms of a modality conjunction.

More formally, let us consider a universe $\mathcal{U}$ of elementary units (the whole set of phlebotomine sandflies specimens in our case) and suppose defined on $\mathcal{U}$ a partition where a distinct concept is associated with each of its classes. Let us denote by $\mathcal{C}$ the set of concepts or classes (the set of species in our case). Now, let us consider a classical qualitative (categorical) variable v defined on $\mathcal{U}$. For u belonging to $\mathcal{U}$, $v(u)$ is a single value of the modality set of v. Now, J coding this modality set, assume a collection of subsets of J:

$$\mathcal{P}_v(J) = \{J_1, J_2, ..., J_i, ..., J_k\} \tag{1}$$

so that for each concept c of $\mathcal{C}$, only one subset J_i of modality values can be met in c. The qualitative variable with multiple choice deduced from v and which we denote by $v_{\mathcal{C}}$, is defined as a mapping of $\mathcal{C}$ onto $\mathcal{P}_v(J)$

$$v_{\mathcal{C}} : \mathcal{C} \longrightarrow \mathcal{P}_v(J) \tag{2}$$

For generality reasons, we will consider a higher description level introducing a probability distribution $\{p_i \mid 1 \leq i \leq k\}$ on $\mathcal{P}_v(J)$. Therefore the $v_{\mathcal{C}}$ value can be written as follows:

$$(J_1, p_1)\&...\&(J_i, p_i)\&...\&(J_k, p_k) \tag{3}$$

or, more explicitly:

$$(\&\{j \mid j \in J_1\}, p_1)\&...\&(\&\{j \mid j \in J_i\}, p_i)\&...\&(\&\{j \mid j \in J_k\}, p_k) \tag{4}$$

This type of description introduces uncertainty in the concept recognition or can be associated with a partition of $\mathcal{C}$ in higher concepts (genus in our case) which can be described by (3). In this richer case the descriptive variable can be expressed in terms of probabilistic qualitative variable with multiple choice.

Because of the generalized data table formalization, the included value in the entry situated at the intersection of the c row and the $v_{\mathcal{C}}$ column is given by expression (3) or by that (4).

3 Preordonance structure on the modality set of a qualitative variable. Representation

A “preordonance” qualitative variable is a qualitative (categorical) variable whose modality set is endowed with an ordinal similarity. Formally, a preordonance is a total preorder (ranking with ties) on the set of unordered (or ordered) modality pairs. By denoting $J = \{1, 2, ..., j, ..., m\}$ the modality codes of the concerned variable, the total preorder is defined on the following set:

$$J^{\{2\}} = \{(j, h) \mid 1 \leq j \leq h \leq m\} \tag{5}$$

(Lerman and Peter (1985), Lerman (1987), Ouali-Allah (1991), Lerman (2000), Lerman and Peter (2003)). This total preorder is established by the expert knowledge by going from the highest ordinal similarity pairs to the lowest ones. For two pairs (j, h) and (j', h'), two cases have to be considered: either

$$(j, h) > (j', h') \tag{6}$$

or

$$(j, h) \sim (j', h') \tag{7}$$

In the first case j and h are assumed, without loss of generality, to be more similar than j' and h' and in the second case, j and h are assumed to be equally similar as j' and h'.

Let us consider the above example of the previous section ("Aspect of individual duct") where $J = \{1, 2, 3, 4\}$. By going from the most similar modality pair to the least similar one, the submitted preordonance by the expert is the following:

$$11 \sim 22 \sim 33 \sim 44 > 12 \sim 13 \sim 23 > 14 \sim 24 \sim 34 \tag{8}$$

where jh represents the pair $\{j, h\}, 1 \leq j \leq h \leq 4$.

The total preorder on $J^{\{2\}}$ is coded by means of the "mean rank function" given by the table:

$$\{r_{jh} \mid 1 \leq j \leq h \leq m\} \tag{9}$$

where the rank r_{jh} is computed with the following equation:

$$r_{jh} = l_1 + l_2 + ... + l_{p-1} + \frac{1}{2} \times (l_p + 1) \tag{10}$$

where l_q denotes the q^{th} class size of the total preorder on $J^{\{2\}}$ according to an increasing order and where jh belongs to the p^{th} class.

Then, in our example, the above table (9) becomes in our example:

$$\{8.5, 5, 5, 5, 2, 8.5, 5, 2, 8.5, 2, 8.5\} \tag{11}$$

4 Taxonomic variable organizing a set of dependent variables. Representation

Let us begin by an example and consider the variables 1, 18, 19 and 20 of Lebbe et al. (1987) that we denote v^1, v^{21}, v^{31}, and v^{32}, respectively. v^1 is the "Sex" attribute, v^{21} is defined by the "Number of style spines", v^{31} indicates the "Distribution of 4 style spines" and v^{32}, the "Distribution of 5 style spines". The value sets of these variables are:

{1: male, 2: female}, {1, 2, 3, 4, 5}, {1, 2, 3, 4, 5, 6} and {1, 2, 3, 4, 5}, respectively, where each integer code is associated with a modality value. We obtain the following taxonomic structure:

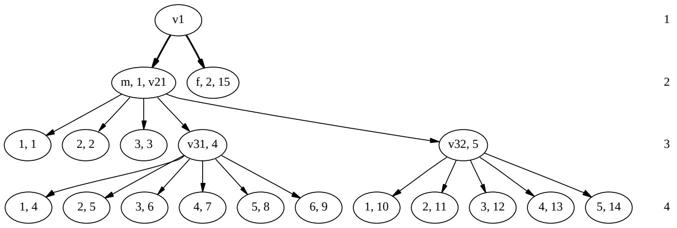

Fig. 1. Taxonomic variable

Clearly, the variable v^{21} is defined only when the v^1 value is 1. On the other hand, the variables v^{31}, and v^{32} are defined only when the values of v^{21} are 4 and 5, respectively. The mother variable of v^{31} and v^{32} is v^{21}.

More generally, a taxonomic variable denoted ω, organizing a set of logically dependent variables consists of a sequence of collections of qualitative variables of the following form:

$$\begin{aligned}\omega = (\{v^1\}, \{v^{21}, v^{22}, ..., v^{2k_2}\}, ..., \{v^{p1}, v^{p2}, ..., v^{pk_p}\}, \\ ..., \{v^{q1}, v^{q2}, ..., v^{qk_q}\})\end{aligned} \tag{12}$$

The first collection is necessarily reduced to one element: the variable v^1. This corresponds to the root of the tree representing the taxonomic variable ω. The variables $v^{p1}, v^{p2}, ..., v^{pk_p}$ are represented at the p^{th} level of this tree built in a descendant fashion. The set of variables $\{v^{p1}, v^{p2}, ..., v^{pk_p}\}$ can be divided into disjoint subsets (classes) according to mother variable concerned. More precisely, if $\{v^{pi}, ..., v^{pi'}\}$ $(i' > i)$ denotes a such subset, two of its elements v^{pj} and $v^{pj'}$ are characterized by the same mother variable $v^{(p-1)h}$. They are respectively defined on two distinct subsets of the described objects (specimens of phlebotomine sandflies in our case), where each subset is defined by one modality of $v^{(p-1)h}$.

In the above example ω is instanciated as follows:

$$\omega = (\{v^1\}, \{v^{21}\}, \{v^{31}, v^{32}\}) \tag{13}$$

The structure associated with this variable is represented in terms of a "ultrametric preordonance" (Lerman (1970), Lerman (2000)) on the set of taxonomy leaves (in the above example the cardinality of this set is 15). By considering a decreasing construction of the taxonomic tree from the root to the leaves, this total preorder on the set of unordered leaf pairs is such that, the higher the rank of a given pair, the lower the first node which underlies the two concerned leaves. Thus, in the above example, the pair $\{6, 8\}$ has the same rank as that of $\{10, 12\}$. The latter is greater than that of $\{7, 12\}$, which is equal to the $\{2, 3\}$ rank and so on ...

Now, let us denote by L the set of the taxonomy leaves, a ranking function r coding the total preorder defined by the utrametric preordonance is characterized by the following equation:

$$(\forall\{x, y, z\} \in P_3(L)), r(x, z) \geq min(r(x, y), r(y, z)) \tag{14}$$

where $P_3(L)$designates the set of all 3-subsets of L.

As in the general case (see Section 3), we adopt the notion of “mean rank” for the ranking function. Mathematical formula can be derived, relating the tree shape with the mean rank function (Lerman and Peter (1988)). The highest rank is assigned to the elements of the preorder class constituted by the the pairs having the form: $\{x, x\}$, $x \in L$. In these conditions, the taxonomic variable is interpreted as a particular case of a preordonance variable.

5 Taxonomic preordonance variable. Representation

Let us reconsider here the above ordinal similarity structure endowed with a taxonomic variable ω organizing a set of logically dependent qualitative variables. We further assume that the modality set $\mathcal{M}^{pi}$ of a given qualitative variable v^{pi} (see 12) is endowed with a total preordonance (see Section 3), $1 \leq i \leq k_p, 1 \leq p \leq q$. These preordonances are locally defined variable by variable, 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 in the above Section 4 and those we have just mentioned.

Such a preordonance is built step by step, decreasingly, with respect to the resemblance between terminal modalities corresponding to the taxonomy leaves. The general principle consists in refining the ultrametric preordonance associated with the taxonomy by means of the preordonances locally defined on the modality sets of the different variables.

More clearly, let us begin by the above example (see Figure 1) and consider the leaf sets $A = \{4, 5, 6, 7, 8, 9\}$ and $B = \{10, 11, 12, 13, 14\}$ associated with the modality sets of the variables v^{31} and v^{32}, respectively. By denoting $P_2(A)$ (resp., $P_2(B)$) the unordered pairs from A (resp., B), $P_2(A) \cup P_2(B)$ determines a unique class of the total preorder defined by the tree structure. This class comprises all the element pairs joined at the level 3 of the taxonomy. Preordonance structures on the modality sets of the variables v^{31} and v^{32} provide total preorders on $P_2(A)$ and $P_2(B)$, respectively. These, can be represented by rank functions. Specifically, one may consider the mean rank functions r_A and r_B defined on $P_2(A)$ and $P_2(B)$, respectively. In these conditions, a ranking function $r_{A\cup B}$ on $P_2(A) \cup P_2(B)$ is deduced from r_A and r_B as follows:

$$r_{A\cup B} : P_2(A) \cup P_2(B) \longrightarrow Val(r_A) \cup Val(r_B) \tag{15}$$

where $Val(r_A)$ (resp., $Val(r_B)$) is the value set of r_A (resp., r_B). Consequently, $r_{A\cup B}(\{x, y\})$ is defined by $r_A(\{x, y\})$ if $\{x, y\} \in P_2(A)$ and by $r_B(\{x, y\})$ if $\{x, y\} \in P_2(B)$.

Therefore, according to the value scale of $r_{A\cup B}$, a total preorder on $P_2(A) \cup P_2(B)$ is induced. This substitutes the unique class $P_2(A) \cup P_2(B)$.

Let us continue with the above illustrative example. The next preorder class construction is given by the preordonance variable v^{21}. Its modality set C appears at the level 3 of the taxonomy. $P_2(C)$ is endowed with a total preorder. In the latter we have to do the following substitutions:

$$\begin{aligned}
(\forall x \in \{1, 2, 3\}), \{x, 4\} &\leftarrow \{\{x, y\} \mid y \in A\} \\
(\forall x \in \{1, 2, 3\}), \{x, 5\} &\leftarrow \{\{x, y\} \mid y \in B\} \\
\text{for } \{4, 5\} &\leftarrow \{\{x, y\} \mid \{x, y\} \in A \times B\}
\end{aligned} \tag{16}$$

where the different pairs included in a given class substitution are interpreted as equally similar.

Now, let us give a general expression of the construction of a taxonomic preordonance variable. We begin by ordering the set

$$\Delta(L) = \{\{x, x\} \mid x \in L\} \tag{17}$$

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. Thus, in the above example, for these pairs we have

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

(18)

Let us reconsider here the general expression 12 of the taxonomic variable ω. Let us indicate by $\mathcal{M}(v^{qj})$ the modality set of the variable v^{qj}, $1 \leq j \leq k_q$. These modality sets are figured by the deepest leaves of the tree depicting ω. Then the next step of refining the ω taxonomic preordonance consists in introducing the total preorders defined by the preordonance variables v^{qj} on the set of unordered modality pairs of $\mathcal{M}(v^{qj})$, that we denote by $P_2(\mathcal{M}(v^{qj}))$, $1 \leq j \leq k_q$. The unique class

$$\mathcal{P} = \bigcup \{P_2(\mathcal{M}(v^{qj})) \mid 1 \leq j \leq k_q\} \tag{19}$$

is refined according to the mean rank functions defined on the sets $P_2(\mathcal{M}(v^{qj}))$, $1 \leq j \leq k_q$, respectively. The global ranking function $r_{\mathcal{P}}$ on the union $\mathcal{P}$ is defined directly from the partial mean rank functions (see the example above in this Section):

$$(\forall j, 1 \leq j \leq k_q), (\forall \{x, y\} \in P_2(\mathcal{M}(v^{qj}))), r_{\mathcal{P}}(\{x, y\}) = r_{qj}(\{x, y\}) \tag{20}$$

where r_{qj} designates the mean rank function on $P_2(\mathcal{M}(v^{qj}))$ associated with the preordonance variable v^{qj}.

Additionally, the ranking function, that we denote by R_U has to take into account the taxonomic structure. Consequently it can be written as follows:

$$(\forall \{x, y\} \in P_2(\mathcal{M}(v^{qj}))), R_{\mathcal{P}}(\{x, y\}) = r_{qj}(\{x, y\}) + card(L) \tag{21}$$

For all j, $1 \leq j \leq k_q$.

Thus, two modality pairs $\{x, y\}$ and $\{z, t\}$, belonging to two different sets $P_2(\mathcal{M}(v^{qj}))$ and $P_2(\mathcal{M}(v^{qj'}))$ $(j \neq j')$ are compared on the basis of their respective rank functions defined independently on the modality pairs of v^{qj} and on those of $v^{qj'}$. This is consistent with the similarity index construction (Lerman and Peter (1988), Lerman and Peter (1989)).

Now, let us consider the variable set $\{v^{p1}, v^{p2}, ..., v^{pk}\}$ introduced at the p^{th} level of the taxonomy (see 12). The respective modalities of each of these variables arise at the $(p+1)^{th}$ level. $\bigcup \{P_2(\mathcal{M}(v^{pi}) \mid 1 \leq i \leq k_p\}$ determines a unique class of the taxonomic preorder. For a given i $(1 \leq i \leq k_p)$, a total preorder is provided on $P_2(\mathcal{M}(v^{pi})$ by the preordonance variable v^{pi}. This refines the subclass $P_2(\mathcal{M}(v^{pi})$. Moreover, for $\{x, y\}$ belonging to $P_2(\mathcal{M}(v^{pi})$, if x (resp., y) is a node tree from which branches issue, the class of the terminal tree chains passing by x (resp., y) is substituted for x (resp., y) (see 16 in the above example). All the concerned pairs are interpreted as equally similar and the mean rank function value $r_{pi}(\{x, y\})$ deduced from the preordonance variable v^{pi}, is applied to all of these pairs. Denote $\mathcal{M}'(v^{pi}$ the extended value set and r'_{pi} the extended definition of the mean rank function r_{pi} on $P_2(\mathcal{M}'(v^{pi})$. From the set, denoted $\mathcal{R}'_p$, of rank functions

$$\mathcal{R}'_p = \{r'_{pi} \mid 1 \leq i \leq k_p\} \tag{22}$$

a unique rank function $r_{\mathcal{P}}$ is induced on

$$\mathcal{P} = \bigcup \{P_2(\mathcal{M}'(v^{pi})) \mid 1 \leq i \leq k_p\} \tag{23}$$

as follows:

$$\forall i, 1 \leq i \leq k_p, \forall \{x, y\} \in P_2(\mathcal{M}'(v^{pi})), r_{\mathcal{P}}(\{x, y\}) = r'_{pi}(\{x, y\}) \tag{24}$$

In these conditions, according to the $r_{\mathcal{P}}$ values, a total preorder on $\mathcal{P}$ is provided. Besides, $r_{\mathcal{P}}$ enables a consistent construction of a similarity index between described objects or concepts (Lerman and Peter (1989)). For this purpose we substitute for $r_{\mathcal{P}}$ a ranking function $R_{\mathcal{P}}$ which takes into account all the leaf pairs preceeding $\mathcal{P}$ in the taxonomic order, strictly. More clearly, by denoting P_{p+1} this set of leaf pairs

$$\forall \{x, y\} \in P_2(\mathcal{M}'(v^{pi})), R_{\mathcal{P}}(\{x, y\}) = r_{\mathcal{P}}(\{x, y\}) + card(P_{p+1}) \tag{25}$$

For all $i, 1 \leq i \leq k_p$.

In the case of the above example we have

$$P_3 = \{\{4, 10\}, \{4, 11\}, ..., \{4, 14\}, \{5, 10\}, \{5, 11\}, ..., \{9, 13\}, \{9, 14\}\} \tag{26}$$

The above ranking function $R_{\mathcal{P}}$ is defined for all leaf pairs joined at p^{th} level (first junction). Each leaf can be associated with a terminal tree chain from the $(p+1)^{th}$ level. In these conditions, a global ranking function R is built from its $R_{\mathcal{P}}$ restrictions.

At the final step, the set of all complete chains of the tree represented by the leaf set, is provided with a total preorder. Consequently, the taxonomic variable is enriched and becomes a "taxonomic preordonance variable", that we code by means of the ranking function R.

6 Taxonomic variable with multiple choice

The descriptive structure of the global variable considered here is defined in the previous Section 4. Nevertheless, the "value" of a given component variable v^{pi} of the taxonomy p level on a given concept is defined by a probabilistically weighted conjunction of conjunctions on the set $J = \mathcal{M}(v^{pi})$ of its modalities (see Formula (4) in Section 2). In Lebbe et al. (1987), only deterministic values are considered and, then, the value of such a variable v^{pi}

on a given concept is defined by a unique conjunction whose terms belong to J (or equivalently, as a subset of J) having the following form

$$\&\{j \mid j \in G\} \tag{27}$$

where G is a subset of J.

Let us begin with an example and imagine that for the above descriptive variables v^1, v^{21}, v^{31} and v^{32}, introduced in Section 4, one has the following values on a given concept (species in our case) c:

$$\begin{aligned} v^1(c) &= 1 \\ v^{21}(c) &= (1\&2, 0.4)\&(2\&3\&4, 0.2)\&(4\&5, 0.4) \\ v^{31}(c) &= (1\&2, 0.4)\&(2\&3, 0.6) \\ v^{32}(c) &= (2\&3\&4, 0.8)\&(3\&5, 0.2) \end{aligned} \tag{28}$$

Denote here by w the taxonomic variable organizing the preceeding variables (see Figure 1). One possible value of w on an element u drawn from c may be: $w(u) = 11\&12$, corresponding to $v^1(u) = 1$ and $v^{21}(u) = 1\&2$. Another possible value of w may be $w(u) = 12\&13\&141\&142$ corresponding to $v^1(u) = 1$, $v^{21}(u) = 2\&3\&4$ and $v^{31}(u) = 1\&2$.

The probability of the v^{21} value is 0.4 and that of the v^{31} value is $0.2 \times 0.4 = 0.08$. These values are obtained according to computational principle of a conditional probability.

More precisely, denoting $\vee$ the logical disjunction, the w value on a random unit u^* provided from the concept c, can be written:

$$\begin{aligned} w(u^*) = (11\&12, 0.4) \vee (12\&13\&141\&142, 0.08) \\ \vee(12\&13\&142\&143, 0.12) \vee (141\&142\&152\&153\&154, 0.128) \\ \vee(141\&142\&153\&155, 0.032) \vee (142\&143\&152\&153\&154, 0.192) \\ \vee(142\&143\&153\&155, 0.048) \end{aligned} \tag{29}$$

or, by using the coding of the taxonomy leaves with the integers 1 to 15 (see Figure 1),

$$\begin{aligned} w(u^*) = (1\&2, 0.4) \vee (2\&3\&4\&5, 0.08) \\ \vee(2\&3\&5\&6, 0.12) \vee (4\&5\&11\&12\&13, 0.128) \\ \vee(4\&5\&12\&14, 0.032) \vee (5\&6\&11\&12\&13, 0.192) \\ \vee(5\&6\&12\&14, 0.048) \end{aligned} \tag{30}$$

The weight sum is a probability sum and consequently, is equal to 1.

Now, the associated value of w on the concept c, can be put in the following form:

$$\begin{aligned} w(c) = (1\&2, 0.4) \wedge (2\&3\&4\&5, 0.08) \\ \wedge(2\&3\&5\&6, 0.12) \wedge (4\&5\&11\&12\&13, 0.128) \\ \wedge(4\&5\&12\&14, 0.032) \wedge (5\&6\&11\&12\&13, 0.192) \\ \wedge(5\&6\&12\&14, 0.048) \end{aligned} \tag{31}$$

where $\wedge$ is another notation for a conjunction.

Thus $w(c)$ consists of a probability distribution on a collection of leaf subsets of the taxonomic tree or, equivalently, on complete chain subsets of this tree.

The general case can be easily derived from the above illustration. Relative to the modality set $\mathcal{M}(v^{qi})$ of the qualitative variable v^{qi} appearing in the taxonomic variable ω (see 12), let us consider the possible values of v^{qi}. These values, can be put as follows (see Section 2):

$$\{(J_l, p_l) \mid 1 \leq l \leq m_{qi}\} \tag{32}$$

where $\{J_l \mid 1 \leq l \leq m_{qi}\}$ is a collection of m_{qi} modality subsets and where J_l occurs with the probability p_l, $1 \leq l \leq m_{qi}$, $(\sum\{p_l \mid 1 \leq l \leq m_{qi}\} = 1)$.

Now, consider for a given l, a modality x_l belonging to J_l for which a v^{qi} daughter variable $v^{(q+1)j}$ is defined. With its modality set designated by $\mathcal{M}(v^{(q+1)j})$ associate its values in the above form:

$$\{(J'_{l'}, p'_{l'}) \mid 1 \leq l' \leq m_{(q+1)j}\} \tag{33}$$

where $J'_{l'}$ is a modality subset of $\mathcal{M}(v^{(q+1)j})$ occuring in c with the probability $p'_{l'}$, $(\sum\{p'_{l'} \mid 1 \leq l' \leq m_{(q+1)j}\} = 1)$.

The joint probability of J_l and $J'_{l'}$ is obtained according to conditional probability principle by $p_l \times p'_{l'}$. This can also be expressed as follows:

$$Pr(\&\{x_l \& x'_{l'} \mid (x_l, x'_{l'}) \in J_l \times J'_{l'}\}) = p_l \times p'_{l'} \tag{34}$$

Thus $p_l \times p'_{l'}$ is the probability assigned to the conjunction of partial chains of the two elements x_l and $x'_{l'}$ belonging to J_l and $J'_{l'}$, respectively.

Finally and recursively, the value set of the taxonomic variable ω on c is obtained. This value set consists of a probabilized set of conjunctions of complete chains of the taxonomic tree. Note that each complete chain can be represented by its terminal leaf. Denoting as J the set of all leaves, one can easily see that the probabilized value of the concerned variable has the same general structure as that presented in Section 2 (see 3). More specifically, a given leaf conjunction is concerned by a unique sequence of the initial qualitative variables, totally ordered by the mother → daughter relation.

7 Conclusion

As claimed above, the conceptual notion of a data table remains fundamental in analyzing logical data knowledge. And, the only difference concerns the logical nature of a cell content, describing an object or a concept (class) with respect to a descriptive variable. In case of absence of missing data, the description complexity proceeds from two main causes. The first is associated with the complexity of the relation on the value set of the description endowed with the expert knowledge. The second results from the level knowledge of the description on the entities (objects or concepts) to be clustered according to their similarities. For a concept description by taxonomic preordonance variables with multiple choice, the structural aspects of the value scale have been studied in Sections 3, 4 and 5. Whereas, the formalization of the expert knowledge relative to the values of such descriptive variables on the described concepts, is given in Sections 2 and 6.

A rough similarity index between described concepts has been built in order to minutely take into account the two complexity origins mentioned above (Lerman and Peter (1988), Lerman and Peter (1989)). In case of a description by many multivalued taxonomic preordonance variables, the integration process of the rough similarity indices (taken variable by variable), into the LLA hierarchical classification method (Lerman (1993)), follows a general principle given in (Lerman and Peter (1985), Lerman (1987), Lerman (2000) Lerman and Peter (2003)). This approach is comprised in the hierarchical classification software named CHAVLH (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables Hétérogènes). Significant and interesting results have been obtained in the hierarchical classification of 142 species described by 61 taxonomic preordonance variables with multiple choice (Lerman and Peter (1988)).

Let us end by a general remark: taking into account the expert knowledge in building structured descriptive attributes enables to obtain more synthetic and more robust cluster organization; however, "explaining" the general features of a given "significant" cluster becomes more difficult.

References

BATAGELJ, V. (1989): Similarity Measures Between Structured Objects. In: *Studies in Physical and Theoretical Chemistry.* Elsevier, Amsterdam, Vol. 63, 25–40.

BILLARD, L. and DIDAY, E. (2003): From Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association, Vol. 98, No 462, 470-487.*

BOCK, H. H. and DIDAY, E. (eds) (2000): *Analysis of Symbolic Data: Explotary Methods for Extracting Statistical Information From Complex Data.* Springer-Verlag, Berlin.

DIDAY, E. (1989): Introduction à l'approche symbolique en analyse des données. *Recherche opérationnelle/Operations Research vol. 23, 2, 193-236.*

LEBBE, J., DEDET, J.P. and VIGNES, R. (1987): Identification assistée par ordinateur des phlébotomes de la Guyane Française. *Institut Pasteur de la Guyane Française, Version 1.02.*

LERMAN, I. C. (1970): *Les bases de la classification automatique* Gauthier-Villars, collection "Programmation", Paris.

LERMAN, I. C. (1987): 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.*

LERMAN, I. C. (1993): Likelihood Linkage Analysis (LLA) Classification Method (Around an Example Treated by Hand.) *Biochimie 75, Elsevier editions, 379-397.*

LERMAN, I. C. (2000): Comparing Taxonomic Data. *Math. & Sci. hum.* 38^e *année, 150, 37-51.*

LERMAN, I. C. and PETER, P. (1985): Élaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de consensus en classification. *Publication Interne Irisa n^o 262.*

LERMAN, I. C. and PETER, P. (1988): Classification en présence de variables préordonnances taxonomiques à choix multiple. Application à la structuration des phlébotomes de la Guyane Française. *Publication Interne Irisa n^o 426.*

LERMAN, I. C. and PETER, P. (1989): Classification of Concepts Described by Taxonomic Preordonnance Variables with Multiple Choice. In E. Diday Editor *Data Analysis, Learning symbolic and numerical knowledge* Nova Science Publishers, 73–87.

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

OUALI-ALLAH, M. (1991): *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques*, thèse de doctorat de l'Université de Rennes 1.

PENNERATH, F. and NAPOLI, A. (2006): La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique. In: G. Ritschard and C. Djeraba (Eds.): *Extraction et Gestion des Connaissances.* Cépaduès, Toulouse, France, 517–528.