COMPARING TAXONOMIC DATA

Israël-César LERMAN

RÉSUMÉ – Comparaison de données taxonomiques
La prise en compte fidèle de la structure des données est une des caractéristiques essentielles de la méthode de classification hiérarchique AVL de l’Analyse de la Vraisemblance des Liens. Pour cette méthode les variables descriptives sont interprétées en termes de relations sur l’ensemble des objets. Le cas où ces variables définissent des taxonomies sur l’ensemble des objets devient ces dernières années de plus en plus important en Classification. Néanmoins, les coefficients d’association entre variables ainsi que les indices de similité (resp. dissimilarité) entre objets de la sorte obtenus, ont une portée très générale. Ils peuvent en effet être adaptés à toute méthode de classification.

MOTS CLÉS – Données structurées, Classification, Coefficients d’association, Indices de similité.

SUMMARY – Taking structured data into account is one of the main features of the Likelihood Linkage Analysis (LLA) (Analyse de la Vraisemblance des Liens, AVL) hierarchical classification method. For this methodology the descriptive attributes are represented in terms of relations on the object set. The case where these attributes define taxonomies on the set of objects is becoming these last years more and more important in Classification. Nevertheless, the association coefficients between attributes, respectively the similarity (or dissimilarity) indices between objects, have a very general scope and can be adapted to any classification method.

KEYWORDS – Structured data, Classification, Association coefficients, Similarity indices.

Communicated by B. Leclerc.

---

* Irisa – Université de Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, e-mail : lerman@irisa.fr.
1 INTRODUCTION

We are interested in the form of the data Table 1. \( a^r \) is a categorical variable. A “taxonomic” similarity is assumed to be given on the category set of \( a^r, 1 \leq r \leq p \). Two similarity matrices have to be established, the first on the attribute set \( A \) and the second, on the object set \( O \). Their elaboration follows the principle of the “Analyse de la Vraisemblance des Liens” (AVL) method. Two main aspects distinguish this approach. The former consists of interpreting attributes in terms of relations on the object set \( O \) and the latter, in establishing probabilistic similarity indices with respect to a statistical independence hypothesis between the data units. By its extreme generality this approach allows us to analyze the different ways of quantifying numerically the resemblance notion between objects or attributes. For this purpose, it is of crucial importance to clearly distinguish between establishing an association coefficient on the descriptive attributes and a similarity index on the described objects.

A taxonomic attribute (see the given example \( a^r \) in Figure 1) can be interpreted in terms of a specific case of a weighted symmetrical binary relation [Lerman(1999)].

\[
\begin{array}{cccc}
A^r & a^1 & a^j & a^p \\
O & 2 \\
o_1 & 1 \\
o_2 & 1 \\
o_3 & 1 \\
o_4 & 3 \\
o_5 & 2 \\
o_6 & 3 \\
o_7 & 3 \\
o_8 & 5 \\
o_9 & 4 \\
o_{10} & 4 \\
\end{array}
\]

Table 1

\[
\begin{array}{c}
a^j : O \longrightarrow \{1,2,3,4,5\}
\end{array}
\]

\[
\begin{array}{c}
\text{level or valuation (1)}
\end{array}
\]

\[\begin{array}{c|c}
3 & \lambda_3 \\
2 & \lambda_2 \\
1 & \lambda_1 \\
0 & \lambda_0 \\
\end{array}\]

Figure 1. Taxonomy
Then, we are going first, in Section 2, to treat this general case, in order to build similarity matrices on the attribute set $\mathcal{A}$ and on the object set $\mathcal{O}$. The weighting is in our context and for a given attribute, determined on the category set of the attribute, by a similarity notion provided by the expert knowledge. Attention will be devoted to the important case where this similarity is ordinal. In this case, the concerned attribute is called "preordonnance attribute (or variable)" [Lerman and Peter (1986) Peter(1987), Lerman(1987,1992), OuaI-Allah(1991)]. A taxonomic attribute is also, in some way, a particular structure of a preordonnance attribute. Section 3 is devoted to the case where the attributes are taxonomic. We will consider our construction of similarity matrices (between attributes and between objects) by emphasizing the specificity of the structures determined by the descriptive attributes on the object set $\mathcal{O}$.


2 COMPARING RELATIONAL DATA

2.1 Introduction

Consider for illustration the data shown in Table 1. For this type of data we have to assume that for each categorical variable $a^r$, a symmetrical weighted binary relation is given on its value set $C^r, 1 \leq r \leq p$. By denoting $m_r$ the cardinality of the value set, this relation can be standardized into a table with $m_r(m_r + 1)/2$ entries. Thus, in our illustration, for the indicated attribute $a^r$, this table has the following form: where $\alpha_{lk}^r(l \geq k)$ is a numerical value which measures the

<table>
<thead>
<tr>
<th>$\alpha_{11}^r$</th>
<th>$\alpha_{21}^r$</th>
<th>$\alpha_{22}^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_{31}^r$</td>
<td>$\alpha_{32}^r$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{41}^r$</td>
<td>$\alpha_{42}^r$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{51}^r$</td>
<td>$\alpha_{52}^r$</td>
</tr>
</tbody>
</table>

Table 2. Representation of $\alpha^r$

association (similarity) between the $l^{th}$ and the $k^{th}$ categories of the attribute $a^r(1 \leq r \leq p)$. As mentionned above, the numbers $\alpha_{lk}^r(l \geq k)$ are assumed to be directly given by or derived from the expert knowledge. This can be formalized by a mapping

$$\alpha^r : \{(l,k) | 1 \leq k \leq l \leq m_r\} \rightarrow \mathbb{R}_+$$

(2)
$1 \leq r \leq p$, where $\mathcal{R}_+$ denotes the real positive numbers.

2.2 Comparing two relational attributes (variables)

Let us designate by $a$ and $b$ the two relational attributes to be compared on the basis of their observation on a set $\mathcal{O}$ of objects. We denote by $m_a$ and $m_b$ the respective numbers of categories of $a$ and $b$ and then

$$\alpha = \{\alpha(l,k)|1 \leq k \leq l \leq m_a\} \quad (3)$$

and

$$\beta = \{\beta(q,p)|1 \leq p \leq q \leq m_b\} \quad (4)$$

will indicate the two valuations (weightings) (see (2) above) associated with the category sets of $a$ and $b$. Now, if $n = \text{card}(\mathcal{O})$, the set $\mathcal{O}$ can be coded by $I = \{1,2,\ldots,i,\ldots,n\}$. Hence, by introducing the set $\mathcal{P}$ of unordered objects pairs

$$\mathcal{P} = \{\{x,y\}|x \in \mathcal{O}, y \in \mathcal{O}, x \neq y\}, \quad (5)$$

we code $\mathcal{P}$ with

$$J = \{(i,j)|1 \leq i < j \leq n\} \quad (6)$$

Then, we consider the extension on $J$ of the $\alpha$ and $\beta$ valuations:

$$(\forall(i,j) \in J), \alpha(i,j) = \alpha\{\max[a(i), a(j)], \min[a(i), a(j)]\} \quad (7)$$

$$(\forall(i,j) \in J), \beta(i,j) = \beta\{\max[b(i), b(j)], \min[b(i), b(j)]\} \quad (8)$$

More explicitly, $a(i)$ and $a(j)$ are the integer codes associated with the categories possessed by $i$ and $j$, respectively. And then, the right member of equation (7) is directly given by an entry of the table representing $\alpha$ (see for example Table 2). Equation (8) can be understood in the same way with respect to the table representing $\beta$.

In our approach, the elaboration of an association coefficient between two relational variables (attributes) follows the four following steps:

(i) Definition of a raw association coefficient:

This is simply the inner product on $\mathcal{P}$ of $\alpha$ and $\beta$, namely

$$s(a,b) = \sum\{\alpha(i,j)\beta(i,j)|(i,j) \in J\} \quad (9)$$

It is of interest to realize that we can obtain a compact expression for $s(a,b)$ by introducing the contingency table

$$\{n_{kp}|1 \leq k \leq m_a, 1 \leq p \leq m_b\} \quad (10)$$
which crosses the two categorical variables \( a \) and \( b \).

More precisely, \( n_{kp} \) is the number of objects whose values for \( a \) and \( b \) are the categories coded by \( k \) and \( p \), respectively, \( 1 \leq k \leq m_a, 1 \leq p \leq m_b \). As a matter of fact \( s(a, b) \) can be written as follows

\[
s(a, b) = \sum_{(k, p)} (\binom{n_{kp}}{2}) \alpha(k, k) \beta(p, p) + \sum_{p < q} \sum_{k \leq l} n_{kp} n_{lq} \alpha(l, k) \beta(q, p)
\]

(11)

where \( (\binom{n_{kp}}{2}) = n_{kp}(n_{kp} - 1)/2, 1 \leq k \leq l \leq m_a, 1 \leq p < q \leq m_b \).

(ii) Associating with the raw similarity index \( s(a, b) \) a random variable, under an independence hypothesis.

The most classical random model is the permutational one. For this model consider the set \( G_n \) of all permutations on \( I = \{1, 2, \ldots, i, \ldots, n\} \) \([\text{card}(G_n) = n!]\), provided by a uniform probability measure. The random raw index \( s(a^*, b^*) \) is defined from a pair \( (\sigma, \tau) \) of random independent elements taken in \( G_n \). Explicitly we have

\[
s(a^*, b^*) = \sum\{\alpha[\sigma(i), \sigma(j)]\beta[\tau(i), \tau(j)]\mid (i, j) \in J\}.
\]

(12)

One can see that the probability distribution of \( s(a^*, b^*) \) is the same as that of \( s(a^*, b) \) where the permutation \( \tau \) is fixed, \( s(a, b^*) \), respectively, where the permutation \( \sigma \) is deleted. This property is not general. For instance, it does not hold for two other basic random models presented in (Lerman 1992).

(iii) Computing the mean and the variance of the random raw index \( s(\alpha^*, \beta^*) \).

We have the following results (Lerman 1992, Ouali-Allah 1991):

\[
\mathcal{E}[s(a^*, b^*)] = \left(\frac{\alpha}{\beta}\right) \mu_\alpha \mu_\beta
\]

(13)

where \( \mathcal{E} \) denotes the mean and where \( \mu_\alpha \) and \( \mu_\beta \) are, respectively, the means of the \( \alpha \) and \( \beta \) valuations over \( J : \mu_\omega = (\sum\{\omega(i, j)\mid (i, j) \in J\})/\binom{n}{2} \) with \( \omega = \alpha \) or \( \beta \).

\[
\text{var}[s(a^*, b^*)] = \sqrt{n(n - 1)][V_\alpha V_\beta + \frac{1}{2n}(W_\alpha - 2V_\alpha)(W_\beta - 2V_\beta)]
\]

(14)

where var denotes the variance and where

\[
V_\omega = A_\omega - \mu_\omega^2 \text{ and } W_\omega = B_\omega - \mu_\omega^2
\]

where

\[
A_\omega = \frac{1}{n(n - 1)^2} \sum_i [\sum_{j \neq i} \omega(i, j)]^2
\]

and

\[
B_\omega = \frac{1}{n(n - 1)} \sum_{i \neq j} \omega^2(i, j)
\]
with \( \omega = \alpha \) or \( \beta \) and where we set \( \omega(j, i) = \omega(i, j) \) for all \((i, j)\) belonging to \( J \).

(iv) Defining two fundamental distinct types of correlation coefficient

The former which reflects significance in the similarity has the general following form

\[
\rho_s(a, b) = \frac{1}{\sqrt{n}} \times \frac{s(a, b) - \mathcal{E}[s(a^*, b^*)]}{\sqrt{\text{var}[s(a^*, b^*)]}}
\]

it can be more explicitly written with the above notations

\[
\rho_s(a, b) = \frac{2}{m} \times \frac{s(a, b) - m \mu_{a\beta}}{\sqrt{V_{a}V_{\beta} + \frac{1}{2n}(W_{a} - 2V_{a})(W_{\beta} - 2V_{\beta})}}
\]

where \( m = \binom{n}{2} \).

The latter type of correlation coefficient is derived from the former one by means of the following equation

\[
\rho_t(a, b) = \frac{\rho_s(a, b)}{\sqrt{\rho_s(a, a)\rho_s(b, b)}}
\]

It is very difficult to distinguish what part of the resemblance between the two relational attributes is more or less captured by the one or the other of coefficients \( \rho_s(a, b) \) and \( \rho_t(a, b) \). Great experience is needed for this purpose. Anyway, these two coefficients become identical in case of comparing two weightings on \( I \). In this classical case of two numerical variables, the Bravais-Pearson correlation coefficient is obtained.


\[
Q(a, b) = \frac{s(a, b) - \mathcal{E}[s(a^*, b^*)]}{\sqrt{\text{var}[s(a^*, b^*)]}}
\]

has been considered with a view to independence statistical testing hypotheses. But, in our approach we only consider the similarity evaluation aspect by means of the coefficient \( \rho_s(a, b) \).

2.3 Comparing objects described by relational variables

According to the above section and at this stage of the elaboration of a similarity coefficient between two attributes, only two columns of the data table (see Table 1) associated with these attributes, are needed for the calculation. On the opposite,
building a similarity index between two objects needs the entire data table. This
construction can be divided into two main steps. The first consists of determining,
for each attribute \( a^r, 1 \leq r \leq p \), a standardized contribution of \( a^r \) to the similarity
between the two objects. All these normalized contributions are added in the
second step.

More precisely, given a pair \((a^r, \{x, y\})\) constituted by a descriptive attribute
\( a^r \) and by a couple of objects \( \{x, y\} \), the raw similarity index between \( x \) and \( y \)
relatively to \( a^r \) is defined as

\[
s^r(x, y) = a^r_{ik}
\]

where \( l = \max[a^r(x), a^r(y)] \) and \( k = \min[a^r(x), a^r(y)] \).

Let us now introduce \( I \times I = \{(i, i')|1 \leq i, i' \leq n\} \) in order to normalize
\( s^r(x, y) \) over \( O \times O \). Then, consider the empirical distribution

\[
\{s^r(o_i, o_{i'})|1 \leq i, i' \leq n\}
\]

and designate by \( \text{mean}(s^r) \) and \( \text{var}(s^r) \) its mean and its variance:

\[
\text{mean}(s^r) = \frac{1}{n^2} \sum\{s^r(o_i, o_{i'})|(i, i') \in I \times I\}
\]

\[
\text{var}(s^r) = \frac{1}{n^2} \sum\{(s^r(o_i, o_{i'}) - \text{mean}(s^r))^2|(i, i') \in I \times I\}
\]

There are different ways of normalizing the raw index \( s^r(x, y) \). That considered in the context of the L.L.A. method consists of standardizing \( s^r(x, y) \) according to the following relation

\[
S^r(x, y) = \frac{s^r(x, y) - \text{mean}(s^r)}{\sqrt{\text{var}(s^r)}}
\]

where \( S^r \) defines the normalized contribution of the \( r^{th} \) attribute to the similarity
between \( x \) and \( y \).

The global similarity between \( x \) and \( y \) is then defined by adding the normalized contributions of the different attributes

\[
S(x, y) = \sum_{1 \leq r \leq p} S^r(x, y)
\]

2.4 The case of ranked relational attributes

We assume here that the similarity knowledge concerning the set of categories of
a given attribute \( a^r \) has only an ordinal nature. More precisely a ranking with ties
(total preorder) is established on the set of category pairs

\[
C^r_2 = \{(l, k)|1 \leq k \leq l \leq m_r\}
\]
of the concerned attribute \( a^r \), in such a way that the higher the rank of a pair, the more similar the components of this pair are. For an example, consider a categorical attribute \( a^r \) with five categories as the one considered above in Section 1. But here, the similarity structure between the categories is given by a complete preorder on \( C_2^r \) (where \( m_r = 5 \)) which can for example be expressed as follows:

\[
(3, 1) \sim (4, 1) \sim (4, 2) < (5, 1) \sim (5, 2) \sim (5, 3) \sim (5, 4) \\
< (2, 1) \sim (3, 2) < (4, 3) < (1, 1) \sim (2, 2) \sim (3, 3) \sim (4, 4) \sim (5, 5)
\] (26)

This ranking is coded here by using the "mean rank" function. Let us recall that for this function, all the elements of a complete preorder class have a unique rank defined as the mean over this class of the rank function associated with any arbitrary strict ranking (without ties), compatible with the concerned complete preorder. Thus for (26) the sequence of ranks associated with the preorder classes is:

\[
(2, 5.5, 9, 13)
\] (27)

In this way a preorder is encoded as a numerically weighted binary relation, where a given weight is a multiple of 0.5. For such an encoding the sum of the weights over \( C_2^r \) [see (25)] is constant and equal to \( M_r (M_r + 1) / 2 \), where \( M_r = m_r (m_r + 1) / 2 \). Therefore, conceptually speaking the problem is reduced here to the previous one where the weighting function \( \alpha^r \) (see Table 2) is defined by the mean rank function defined over \( C_2^r \).

There are more complex, but also more accurate mathematical representation of preordonnance attributes (Giakoumakis and Monjardet 1987, Lerman 1992, 2000). For a given attribute \( a^l \), this structure is exactly a binary ordinal relation on \( C_2^r \) [see (25) above] and then, its faithful representation consists of a suitable subset of the cartesian square \( C_2^r \times C_2^r \).

3 COMPARING DATA WITH TAXONOMIC ATTRIBUTES

3.1 Introduction: representation of a taxonomic attribute

Let \( C^r = \{1, 2, \ldots, l, \ldots, m_r\} \) denote the value set of a given taxonomic attribute \( a^r \) (see the above Table 1 and Figure 1 for an illustration). In this case the expert knowledge provides a logical dependency structure between the different categories in a labelled tree form (see Figure 1). This structure, commonly called "dendrogram" (Sneath and Sokall 1973, Gordon 1999), is equivalent to an ordered sequence of partitions on \( C^r \), associated with the increasing sequence of the tree levels.

\[
(\pi_0, \pi_1, \ldots, \pi_{l-1}, \pi_l, \ldots, \pi_L)
\] (28)
where $\pi_0$ is the finest partition for which each class includes one single element (category) and where $\pi_L$ comprises only one class grouping all the elements (categories). $\pi_1$ is deduced from $\pi_{l-1}(1 \leq l \leq L - 1)$ by joining some classes of $\pi_{l-1}$. Let us now designate by $R(\pi_l)$ the set of elements pairs joined together into the classes of the partition $\pi_l, 0 \leq l \leq L$. In our context $R(\pi_1)$ is a subset of $C_2^l$. A faithful mathematical representation for a labelled tree is given by the notion of an “ultrametric preorder” (Lerman 1970a). This is defined here, with notations as follows:

$$R(\pi_L) - R(\pi_{L-1}) < R(\pi_{L-1}) - R(\pi_{L-2}) < \ldots < R(\pi_1) - R(\pi_{l-1})$$
$$< \ldots < R(\pi_2) - R(\pi_1) < R(\pi_1) - R(\pi_0) < R(\pi_0)$$

(29)

In this expression the set difference $R(\pi_l) - R(\pi_{l-1})$ consists of the set of pairs of elements joined together for the first time at the $l$ level of the classification tree. $R(\pi_l) - R(\pi_{l-1})$ is a subset of $C_2^l(1 \leq l \leq L)$ and $R(\pi_L) = C_2^L$. Moreover, $R(\pi_0)$ is the set of pairs having two identical components, namely

$$R(\pi_0) = \{(k,k)|1 \leq k \leq m_r\}$$

(30)

As an example consider the classification given in Figure 1, the above expression (29) becomes

$$2,1 \sim (3,1) \sim (5,1) \sim (4,2) \sim (4,3) \sim (5,4)$$
$$< (3,2) \sim (5,2) \sim (5,3)$$
$$< (4,1) \sim (1,1) \sim (2,2) \sim (3,3) \sim (4,4) \sim (5,5)$$

(31)

Under these conditions, the total preorder (29) can be encoded by means of the mean rank function. The obtained valuation can be denoted in the context of a given taxonomic attribute $a^r$

$$\{\lambda_r(l,k)|1 \leq k \leq m_r\}$$

(32)

This numerical valuation depends strongly on the shape of the classification tree organizing the category set of $a^r$. The formal notion of “indexed type of a classification tree” (Lerman 1999) captures entirely the concept of tree shape. It clearly depends on the cardinalities of the classes underlying the different nodes of the classification tree. The above mean rank valuation [see (32)] can be developed with respect to the indexed type of the taxonomy associated with the taxonomic attribute $a^r$ (see the above reference).

3.2 Comparing two taxonomic attributes

This comparison is interpreted here as a specific case of comparing two weighted relational attributes (see Section 2.2). Two encodings have to be distinguished. The former can be defined by a numerical criterion of metric or statistical nature,
associated with the different aggregations occurring in the classification tree (see Figure 1 for illustration). The latter has an ordinal nature and is provided from the mean rank function as expressed just above. In both cases we can designate by $\lambda$ the numerical encoding of the classification tree. The function $\lambda$ is defined on the set of category pairs $C_2$ see (25)] which is constant on each class of the total pre-order defined by the ultrametric preordonnance associated with the classification tree see (29)]. More precisely, one may derive

$$\forall (k, j) \in R(\pi_l) - R(\pi_{l-1}), \quad \lambda(k, j) = \lambda_l$$

where $\lambda_l = |R(\pi_L)| - \frac{1}{2}[R(\pi_l) + R(\pi_{l-1}) - 1], \quad 1 \leq l \leq L$

Notice that the taxonomic attribute appears as the particular case of ranked relational attribute (see Section 2.4) where the ordinal representation of the associated classification tree is adopted.

Under these conditions, we may easily adapt both coefficients $\rho_{su}(a, b)$ and $\rho(t, a, b)$ to the comparison of two taxonomic attributes. Compact formulation can be obtained for basic expressions of equation (14). By denoting $n_k$ the cardinality of the class of elementary objects whose value is $k$, relatively to the taxonomic attribute $a', 1 \leq k \leq m_r$, we have

$$\mu_\omega = \frac{1}{n(n-1)} \left( \sum_k n_k (n_k - 1) \lambda_{kk} + \sum_{k \neq l} n_k n_l \lambda_{kl} \right)$$

$$B_\omega = \frac{1}{n(n-1)} \left( \sum_k n_k (n_k - 1) \lambda_{kk}^2 + \sum_{k \neq l} n_k n_l \lambda_{kl}^2 \right)$$

$$A_\omega = \frac{1}{n(n-1)^2} \left( \sum_k n_k [(n_k - 1) \lambda_{kk} + \sum_{\{l \neq k\}} n_l \lambda_{kl}]^2 \right)$$

where $\{\lambda_{lk} | 1 \leq k \leq l \leq m_r\}$ is the mean rank valuation defined on $C_2$ [see (32) and (33) above] and where we set $\lambda_{kl} = \lambda_{lk}$ for all $(l, k), 1 \leq k, l \leq m_j$.

3.3 Comparing two objects described by taxonomic attributes

The similarity index (24) can be easily adapted to the case where the object set is described by taxonomic attributes. For this purpose, we only have to adapt the valuation $\sigma_{lk}$ considered in equation (19). Clearly this valuation has to be defined from (32) which is specified in (33). Let us give here an explicit and condensed expression for the normalized contribution of the rth attribute to the similarity between two given objects $x$ and $y$ for which

$$s^r(x, y) = \lambda_r(l_o, k_o)$$

that is to say, for which the respective categories of $x$ and $y$ are $l_o$ and $k_o$.

$$S^r(x, y) = \frac{\sum_{1 \leq l, k \leq m} \sigma_{lk}(\lambda_r(l_o, k_o) - \lambda_r(l, k))}{\sqrt{\sum_{1 \leq l, k \leq m} \sigma_{lk} \sum_{1 \leq l, k \leq m} \sigma_{lk}(\lambda_r(l', k') - \lambda_r(l, k)))^2}$$

(38)
where by $\sigma_{lk}$ is the proportion $n_l n_k / n^2_2, 1 \leq l, k \leq m_r$, and $n_k$ is as above, $1 \leq k \leq m_r$.

4 PROBABILISTIC SIMILARITY MATRIX FOR AVL (LLA)

We have to deal with two dual classification problems. The first concerns the variable (attribute) classification and the second, the object classification. The former treatment leads to the discovery of dependency classes between attributes, when the objective of the latter consists of management.

Relative to the first problem of variable classification we consider the following table of association coefficients provided by $\rho_s$ or $\rho_t$ [see (16) and (17)] and adapted by means of equations (33) to (36) to the case of taxonomic attributes:

$$\{R(a^i, a^k) | 1 \leq j < k \leq p\}$$ (39)

Similarly and relative to the second problem of object classification we consider the following table of similarity indices

$$\{S(o_i, o_l) | 1 \leq i < l \leq n\}$$ (40)

where $S(o_i, o_l)$ is the sum over the attribute set of $S^j(o_i, o_l)$ [see (38)]. More precisely,

$$S(o_i, o_l) = \sum_{1 \leq j \leq p} S^j(o_i, o_l)$$ (41)

Let $E$ be the set to be classified into a dendrogram structure. $E$ is defined by the set $\mathcal{A}$ of the descriptive attributes or by the set $\mathcal{O}$ of the described objects. Now, consider the associated similarity matrix $E: (39)$ in the case where $\mathcal{E} = \mathcal{A}$, (40) in the case where $\mathcal{E} = \mathcal{O}$, respectively. Since the following steps of the construction of a probabilistic similarity matrix are the same in both cases, let us denote by

$$\{Q(x, y) | \{x, y\} \in P_2(E)\}$$ (42)

the adequate similarity matrix among (39) and (40), where $P_2(E)$ designates the set of unordered element pairs of $E$.

Two steps remain necessary in order to obtain the probabilistic indices. The first consists of the global normalization of the previous indices [see (42)] with respect to their empirical distribution. More explicitly, we have to substitute the following expression to (42)

$$\{Q_g(x, y) | \{x, y\} \in P_2(E)\}$$ (43)

where

$$Q_g(x, y) = \frac{Q(x, y) - m_e(Q)}{\sqrt{\text{var}_e(Q)}}$$ (44)
for all \(\{x, y\}\) belonging to \(P_2(E)\), where \(m_e(Q)\) and \(\text{var}_e(Q)\) are respectively the mean and the variance of the empirical distribution (42).

The final step consists of computing the probability indices

\[
\{P(x, y)|\{x, y\} \in P_2(E)\}
\]

where

\[
P(x, y) = \phi[Q_s(x, y)]
\]

for all \(\{x, y\}\) belonging to \(P_2(E)\), where \(\phi\) indicates the \(N(0, 1)\) normal cumulative distribution function. The justification of this reference to the normal probability law is given in the context of the Likelihood Linkage Analysis (LLA or AVL : Analyse de la Vraisemblance des Liens) methodology (Bacelar-Nicola 1981, Daudé 1992, Lerman 1981,1984,1987).

Notice that from (45) a dissimilarity index can be derived by applying the decreasing transformation defined by the function \(-\log_2\). More precisely, \(-\log_2P(x,y)\) stands for the information quantity of the event of which the probability is \(P(x,y)\). Thus, the table of “informational dissimilarity indices” can be written as follows

\[
\{-\log_2[P(x, y)]|\{x, y\} \in P_2(E)\}
\]

5 CONCLUSION

In this paper we have shown how Likelihood Linkage Analysis (Analyse de la Vraisemblance des Liens) methodology applies to taxonomic data. For this objective a taxonomic attribute on the object set \(O\) has been encoded as a specific weighted binary relation on \(O\). With this view a large family of attribute structures can be integrated in the same framework (Ouali-Allah 1991).

However this mathematical encoding is not the only possible. A more accurate but more complex coding consists of representing the classification tree associated with a taxonomic attribute as an ordinal binary relation (not weighted) on the set \(\mathcal{P}\) of unordered object pairs (Lerman 1999, Lerman and Rouxel 2000).

A more complex data structure, in which taxonomic attribute has an important part, has been studied by Lerman and Peter (1989). For these data and for a given taxonomic attribute we assume in addition that there exists an ordinal similarity relation on each set of son nodes of any internal node of a dendrogram, whatever is the latter. This preordonnance relation defined node by node is supposed to be compatible with the ordinal similarity relation induced by the taxonomy. In this way, we have been led to introduce for concept description "taxonomic preordonnance variable with multiple choice". Such a variable is obtained by hierarchical organization of preordonnance categorical variables where the value of a given categorical variable on a given concept is specified by a logical formula on the set of its categories [see the above reference].
Let us now end by a general remark. By taking into account faithfully complex descriptions by structured attributes, we obtain more efficient and synthetic classifications. However, it becomes more difficult to interpret the classification results by giving explicitly an important part to the structures defined by the descriptive attributes. On the other hand, computational complexity to perform classifications may increase but remain polynomial.

ACKNOWLEDGEMENT - I am indebted to the anonymous reviewer who helped me to improve the presentation.
This paper is associated with an invited presentation in the workshop “Dealing with Structured Data in Machine Learning and Statistics”, attached to the ECML(European Conference on Machine Learning), Barcelona, May 31, 2000.

BIBLIOGRAPHY


[25] Lerman, I.C., Peter, Ph., “Classification of concepts described by taxonomic preorderance variables with multiple choice. Application to the structuration of


[31] Peter, Ph., *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistée par ordinateur*, thèse de doctorat de l'Université de Rennes 1, 1987.