Comparing Classification tree Structures: a Special Case of Comparing q-Ary Relations

Israël-César Lerman

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Comparing classification tree structures : a special case of comparing q-ary relations

Israël-César Lerman

N ° 3167

Mai 1997

THÈME 3

ISSN 0249-6399

INRIA
RENNES

Comparing classification tree structures : a special case of comparing q-ary relations

Israël-César Lerman

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet REPCO

Rapport de recherche n°3167 — Mai 1997 — 37 pages

Abstract: Comparing q-ary relations on a set $\mathcal{O}$ of elementary objects is one of the most fundamental problems of classification and combinatorial data analysis. In this paper the specific comparison task that involves classification tree structures (binary or not) is considered in this context. Two mathematical representations are proposed. One is defined in terms of a weighted binary relation; the second uses a four-ary relation. The most classical approaches to tree comparison, are discussed in the contex of a set theoretic representation of these relations. Formal and combinatorial computing aspects of a construction method for a very general family of association coefficients between relations are presented. The main purpose of this article is to specify the components of this construction, based on a permutational procedure, when the structures to be compared are classification trees.

Key-words: Classification tree; Relations; Mathematical representation; Random permutational model.

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 02 99 84 71 00 – Télécopie : (33) 02 99 84 71 71

Comparaison d'arbres de classification : un cas spécifique de la comparaison de relations q-aires

Résumé : La comparaison des relations q-aires sur un ensemble $\mathcal{O}$ d'objets élémentaires, est l'un des problèmes les plus fondamentaux de la Classification et Analyse Combinatoire des Données. La comparaison des structures d'arbres de classification (binaires ou non) est étudiée dans ce contexte en tenant compte de la spécificité de ces structures. Deux représentations mathématiques sont proposées. La première correspond à une relation binaire valuée; et la seconde, à une relation 4-aire. Dans ces conditions, les approches les plus classiques de comparaison d'arbres, sont situées relativement à une représentation ensembliste de ces relations. Nous présentons les aspects du calcul formel et combinatoire, d'une méthode de construction, d'une famille très générale de coefficients d'association entre relations. L'objet principal de cet article consiste à spécifier les composandes de cette construction fondée sur un modèle permutationnel, quand les structures à comparer sont des arbres de classification.

Mots-clé : Arbres de classification; Relations; Représentation mathématique; Modèle aléatoire permutationnel

1 Introduction

Comparing q-ary relations on a set $\mathcal{O}$ of elementary objects is one of the most fundamental problems of Classification and Combinatorial Data Analysis. It does intervene crucially on the different levels of a data synthesis process. Thus, descriptive variable of any type, numerical or categorical (eventually provided by a complex structure on the category set), can be clearly expressed in terms of a relation on $\mathcal{O}$. On the other hand, a data analysis result (classification, hierarchical classification, Euclidean representation, ...) also defines a relation on $\mathcal{O}$. Then a data analysis scheme is viewed as taking into account a collection of relations, to produce a global approximating relation of a predefined type. However, we have to clearly distinguish between the two dual problems : associating objects described by relational variables and associating relations observed on elementary objects or object classes. An ultimate stage makes correspondence between these two kinds of association through a given form of synthesis structure (e.g. hierarchical classification). For any fixed positive integer q, we consider q-ary relations comparison, on the basis of the observation of an object set $\mathcal{O}$. As a matter of fact, a huge literature in Combinatorial Data Analysis (CDA) is devoted to the cases of $q = 1$ or 2 . And, in the latter, not enough attention is paid in order to intimately take into account the specific structure of the compared relations. Thus, the reduction done in the Fowlkes and Mallows (1983) paper, for comparing two classification trees cannot be clearly justified. On the other hand, Baker (1974) uses the Goodman – Kruskal coefficient (1954) for this aim. However, the generality of this coefficient makes it not enough accurate for the concerned structures. The general method we set up (Lerman 1992), has its origin in the K. Pearson and M. G. Kendall contributions. It meets Hubert's work (1987) and makes comprehensive a large family of coefficients. But the approach is more concerned with a view of information theory than with that one of statistical testing of hypotheses. On the other hand, the combinatorial nature of the association problem is emphasized and clearly taken into account.

For reasons of clarity, we first consider the most elementary and classical case of comparing numerical variables ($q = 1$). The main case treated does concern the building of an association coefficient between classification trees. The components of this construction are specified in the framework of our general scheme. For this purpose, two mathematical representations are considered. The former is defined by a weighted binary relation, using a ranking function. It can be related – in some meaning – to the Spearman approach. When, the latter form can be associated with the Kendall approach, and needs the definition of a 4-ary relation on $\mathcal{O}$.

Formal notions, associated with the shape of a classification tree have to be introduced. On the other hand, the presented work is very concerned with combinatorial computing.

At the end of our paper, we will consider the most general case of comparing q-ary relations, for any q.

2 Comparing numerical variables

For this most classical case, the Bravais-Pearson correlation cœfficient is the best known and the best established association cœfficient. It will be obtained at a given level of a general construction scheme of an association cœfficient between relational variables. For the latter, any geometrical or linear mathematical representations are considered. A descriptive numerical variable v observed on a set $\mathcal{O} = \{o_1, o_2, ...o_i, ..., o_n\}$ of objects, is basically a mapping of $\mathcal{O}$ on a numerical scale. v is viewed as an unary weighted (or valued) relation, assigning the weight (value) $v(i) = v(o_i)$ to the i^{th} object. To build a Similarity measure between two descriptive variables v and w , a *raw index* $s(v, w)$ is first introduced, taking into account algebraic conditions. Then and importantly, a *random model of no relation* (or independence) is considered. It associates with the observed variables (v, w) on $\mathcal{O}$, a pair of independent random variables on $\mathcal{O}$, (v^*, w^*). It is fundamental to realize that the reason for this model in our approach, is not to be tested; but to establish a statistically justified similarity measure. In this context, the classical model has a permutational nature. But, it is not the only one which can be considered Lerman (1992).

To the classical raw index

$$s(v, w) = \sum_{1 \leq i \leq n} v(i)\, w(i)\ , \tag{1}$$

the permutational random model will associate the random raw index :

$$s(v^*, w^*) = \sum_{1 \leq i \leq n} v[\sigma(i)]\, w[\tau(i)] \tag{2}$$

where (σ, τ) is an ordered pair of independent random permutations, belonging to $G_n \times G_n$, where G_n is the set – provided by a uniform probability measure – of all permutations on $I = \{1, 2, ..., i, ..., n\}$ $[card(G_n) = n\,!]$.
The exact probability law of $s(v^*, w^*)$ is the same as that of $s(v, w^*)$ [resp. $s(v^*, w)$]. Its limiting form is given, under very general conditions, by the normal distribution [Hàjek & Sidak, 1967].
The centralized and standardized version of $s(v, w)$ is given by :

$$Q(v, w) = \frac{s(v, w) - E[s(v^*, w^*)]}{\sqrt{var\,[s(v^*, w^*)]}}, \tag{3}$$

where E and var respectively denote the mean and variance, is nothing other than – to the multiplicative factor $\sqrt{n-1}$ – the correlation coefficient $\rho(v, w)$ between the descriptive numerical variables v and w :

$$Q(v,w) = \sqrt{(n-1)}\,\rho(v,w) \simeq \sqrt{n}\,\rho(v,w) \tag{4}$$

Then, the correlation cœfficient can be obtained by one of the two following equations :

$$\rho(v,w) = \frac{1}{\sqrt{n}}\,Q(v,w) \tag{5}$$

$$\rho(v,w) = \frac{Q(v,w)}{\sqrt{Q(v,v)\,Q(w,w)}} \tag{6}$$

These equations can be applied in the most general case of comparing q-ary relations.

Now, let us indicate how to establish, for pairwise comparisons of a set $V = \{v^1, v^2, ..., v^j, ..., v^p\}$ of numerical description variables, observed on the object set $\mathcal{O}$, a probabilistic similarity (resp. informational dissimilarity) measure, associated with the table of the Q indices :

$$\{Q(v^j, v^k) \,|\, 1 \leq j < k \leq p\} \tag{7}$$

A globally standardized form of the preceding value table is computed ; namely :

$$\{Q_s(v^j, v^k) \,|\, 1 \leq j < k \leq p\}, \tag{8}$$

with

$$Q_s(v^j, v^k) = \frac{Q(v^j, v^k) - m_e(Q)}{\sqrt{var_e\,(Q)}} \tag{9}$$

where $m_e(Q)$ and $var_e(Q)$ are the empirical mean and variance of the (7) table values.

It has been established by Lerman (1984) and Daudé (1992), under mutual permutational independence hypothesis, associating with V a set $V^* = \{v^{*j} \,/\, 1 \leq j \leq p\}$ of random variables, that the limit distribution of the random coefficient $Q_s(v^{*j}, v^{*q})$ $(\,1 \leq j < k \leq p\,)$ is the normal distribution. Then, we adopt the probabilistic similarity index by means of the equation

$$P_s(v^j, v^k) = \Phi\,[Q_s(v^j, v^k)]\,, \tag{10}$$

$1 \leq j < k \leq p$, where Φ is the normal cumulative distribution function. And in fact, replacing Q [cf.(3)] by Q_s [cf.(9)] makes finely discriminating the probability scale, according to formula (10), for measuring in a relative manner associations between variables.

The Informational Dissimilarity measure $D(v^j, v^k)$ is associated with (10) simply by considering the amount of information which is behind the event of which the probability is $P_s(v^j, v^k)$. Thus, it is given by :

$$D(v^j, v^k) = -\log_2 [P_s(v^j, v^k)] , \tag{11}$$

$1 \leq j < k \leq p$.

This process is generalized and can be applied for pairwise mutual comparison of q-ary relations, for any q. We shall now consider the case of interest in this paper, which concerns association cœfficients between classification trees.

3 Comparing classification trees

3.1 Mathematical representation of a classification tree

We shall only be interested here in labeled trees. However, generalizations can easily be considered for weighted trees by replacing the discrete relation associated with the tree on the object set $\mathcal{O}$, by a weighted one.

Many methods are limited to comparison of binary trees. The given justification argues that it is always possible to associate with a nonbinary tree, a binary one, compatible with it. Nevertheless, multiple agregation at a given level of a classification tree may occur very often in real cases (Lerman (1989), Jovicic (1996)).This is specially, when large data sets are described by qualitative variables, for which the total number of categories is not big enough with respect to the size of the set $\mathcal{O}$.

The number of binary trees compatible with a non binary one becomes considerably large. Let us define the type of the transformation from the l^{th} level tree to the following one, by a sequence of integers $(c_1, c_2, ..., c_q, ..., c_r)$ for which, respectively, $c_1, c_2, ..., c_q, ..., c_r$ classes of the l^{th} partition level π_l , are agregated in the following partition level π_{l+1} . By recalling that the number of binary trees on a set of c elements, is given by

$$\beta(c) = (c-1)\,!\,c\,!\,/\,2^{c-1} ,$$

we obtain the following number of compatible binary decompositions of the transition from the levels l to $(l+1)$:

$$(\prod_{1 \leq q \leq r} \beta(c_q)) \times \frac{d\,!}{d_1\,!\,d_2\,!\,\times\,\ldots\,\times\,d_r\,!}\,, \tag{12}$$

where $d_q = c_q - 1$ $(1 \leq q \leq r)$ and where $d = d_1 + d_2 + \cdots + d_q + \cdots + d_r$.

Let $\mathbb{P}$ be the set of all unordered object pairs

$$\mathbb{P} = \{\{x, y\} \mid x \in \mathcal{O}\,, y \in \mathcal{O}\,,\, x \neq y\} \tag{13}$$

A faithful mathematical representation that we have adopted for a labeled tree is given by the notion of an " ultrametric preordonnance " [Lerman (1970)].

Denoting by

$$(\pi_0\,,\,\pi_1\,,\,\ldots\,\pi_{l-1}\,,\,\pi_l\,,\,\ldots\,,\,\pi_m) \tag{14}$$

the partition sequence associated with levels of an ω tree, the ultrametric preordonnance $UP(\omega)$ is a total preorder on $\mathbb{P}$ given by

$$R(\pi_1) < R(\pi_2) - R(\pi_1) < \cdots < R(\pi_l) - R(\pi_{l-1}) < \cdots < \mathbb{P} - R(\pi_{m-1})\,, \tag{15}$$

where $R(\pi_l)$ is the set of all unordered object pairs joined by the partition π_l , $1 \leq l \leq m$; otherwise, $R(\pi_{l-1}) \subset R(\pi_l)$ and $R(\pi_l) - R(\pi_{l-1})$ – which indicates a set difference – is the set of all unordered object pairs agregated for the first time at the l^{th} level, $1 \leq l \leq m$. Finally note that $R(\pi_0) = \emptyset$ and $R(\pi_m) = \mathbb{P}$.

Example (Figure 1) :

$UP(\omega) : 14 \sim 23 < 15 \sim 26 \sim 36 \sim 45 < 12 \sim 13 \sim 16 \sim 24 \sim 25 \sim 34 \sim 35 \sim 46 \sim 56\,;$

where $\{i, j\}$ has been denoted by ij , $1 \leq i < j \leq 6$.

Now, for purposes of comparison, we have to code the total preorder $UP(\omega)$. The first approach consists of definition a weighting (numerical valuation) on $\mathbb{P}$ corresponding to a ranking function. Mostly if not always, the ranking function is given by the level function l_ω ; for which $l_\omega(i, j)$ is the first level where i and j are joined in the same class, $1 \leq i < j \leq n$.

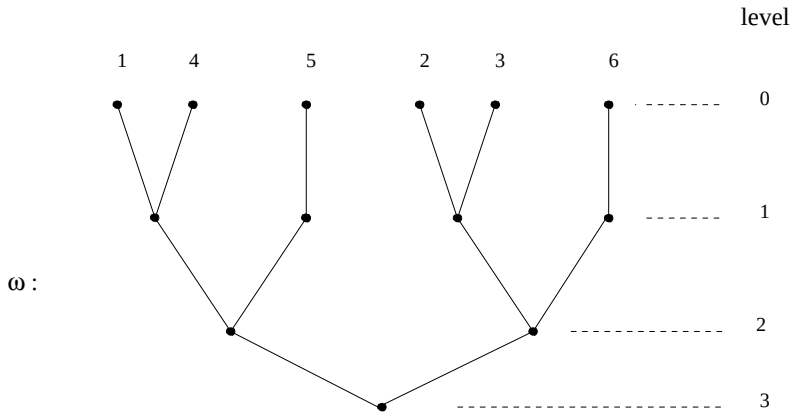


Figure 1:

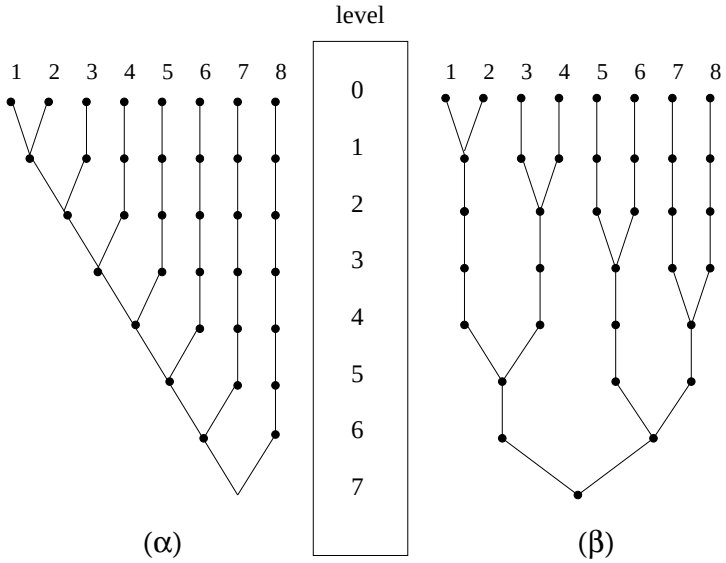


Figure 2:

Here, we do suggest to use the " mean rank " function which respects faithfully ties included in the total preorder and which captures more accurately the tree shape. As an example, consider the two trees α and β : For the associated level functions, whe have

$$l_\alpha(1,8) = l_\beta(1,8) = 7\,;$$

however, the number of agregated element pairs, strictly before $\{1, 8\}$ is $7(7-1)/2 = 21$ for α; when it is $2 \times [4(4-1)/2] = 12$ for β.

Now, denoting by λ_ω the mean rank function coding the total preorder $UP(\omega)$, we have

$$\lambda_\alpha(1, 8) = 21 + [(7+1)/2] = 25$$

whereas,

$$\lambda_\beta(1, 8) = 12 + [(16+1)/2] = 20.5$$

And, the following implicit normalization holds, whatever is the total preorder $UP(\omega)$; and then, whatever is the ω tree shape :

$$\sum\{\ \lambda_\omega(x, y) \mid \{x, y\} \in \mathbb{P}\} = p(p+1)/2\ , \tag{16}$$

where $p = card(\mathbb{P}) = n(n-1)/2$

Notice also that for the β structure

$$\lambda_\beta(1, 8) = \lambda_\beta(4, 5) = 20.5\ ,$$

whereas for the α structure

$$\lambda_\alpha(4, 5) = 3 + (3+1)/2 = 5 < \lambda_\alpha(1, 8) = 25\ .$$

The second mathematical coding proposed here for a tree ω is given by the indicator function of a structured subset $R(\omega)$ of $\mathbb{P} \times \mathbb{P}$. $R(\omega)$ does strictly and faithfully represent ω . For a concise expression of $R(\omega)$, let us, without restriction, designate by $\{1, 2, \ldots, i, \ldots, n\}$ the object set $\mathcal{O}$.

Thus $\mathbb{P}$ can be expressed by

$$\mathbb{P} = \{(i, j) \mid 1 \leq i < j \leq n\}\ . \tag{17}$$

With these notations

$$R(\omega) = \{[(i, j)\ ,\ (i', j')] \mid [(i, j)\ ,\ (i', j')] \in \mathbb{P} \times \mathbb{P}\ and\ l_\omega(i, j) < l_\omega(i', j')\}\ , \tag{18}$$

where l_ω is the level function defined by the ω tree.

We may, without ambiguity, also denote by ω the indicator function of $R(\omega)$. Therefore, ω is defined as follows :

$$\omega((i,j),(i',j')) = \left\{ \begin{array}{ll} 1 & if\ l_\omega(i,j) < l_\omega(i',j')\ , \\ 0 & if\ not\ , \end{array} \right. \tag{19}$$

for every $((i,j),(i',j')) \in \mathbb{P} \times \mathbb{P}$.

3.2 Comparing classification trees : the classical solutions.

Most methods only take into account the comparison of binary trees. The well–known Fowlkes and Mallows approach (1983) associates with a pair of binary trees

$$\left. \begin{array}{l} \alpha = (\pi_0, \pi_1, \ldots, \pi_l, \ldots, \pi_{n-2}, \pi_{n-1}) \\ \beta = (\chi_0, \chi_1, \ldots, \chi_l, \ldots, \chi_{n-2}, \chi_{n-1}) \end{array} \right\} \tag{20}$$

a sequence of similarity indices

$$(B_l | 1 \leq l \leq n-2) \tag{21}$$

where B_l compares the partitions π_l and χ_l, obtained at the l^{th} level of the trees α and β, $1 \leq l \leq n-2$.

According to our previous notations [see (13) and (15)], we associate with a partition π of $\mathcal{O}$, a bipartition of $\mathbb{P}$ denoted by $[R(\pi), S(\pi)]$ where $R(\pi)$ [resp. $S(\pi)$] is the subset of $\mathbb{P}$ comprising the object pairs joined (resp. separated) by π. In these conditions, B_l is nothing other than an association coefficient between two bipartitions of $\mathbb{P}$, respectively associated with π_l and χ_l; namely :

$$[R(\pi_l), S(\pi_l)]\ and\ [R(\chi_l), S(\chi_l)]\ . \tag{22}$$

The specific cœfficient considered by Fowlkes and Mallows (1983) can be written as follows :

$$B_l = \frac{card[R(\pi_l) \cap R(\chi_l)]}{\sqrt{card[R(\pi_l)] \times card[R(\chi_l)]}}\ , \tag{23}$$

where $card$ designates the cardinality.

Notice that this coefficient has exactly the same structure as the one of Ochiai (1957), defined with respect to another type of representation set. Obviously, every similarity index comparing sets of subsets, can be used as B_l. In this way, comparison between the trees α and β is based on the sequence of numerical values (21).

Even in the restricted framework of comparing binary classification trees, two main and related criticisms remain. Why have we to only compare the pairs of partitions having respectively the same level in both trees ? Indeed, disconnection is made by this technique between the different level partitions of a same tree. The second criticism is about producing a global coefficient $B(\alpha, \beta)$ summarizing the sequence (21) by means of a non arbitrary function f :

$$B(\alpha, \beta) = f(B_l | 1 \leq l \leq n-1) \ . \tag{24}$$

The well–known Goodman and Kruskal cœfficient (1954) gives a global comparison of two total preorders on a finite set. And then, it can be used for comparing ultrametric preordonnances associated with trees [see (19)] since an ultrametric preordonnance is a specific total preorder on $\mathbb{P}$ [see (13)].

In order to clearly set up the nature of this comparison, let us introduce the following sets associated with $UP(\alpha)$ and $UP(\beta)$; and assume the general case where α and β are not necessarily binary trees :

$$C_l = R(\pi_l) - R(\pi_{l-1}) \ , \ 1 \leq l \leq m \ , [resp.\ D_p = R(\chi_p) - R(\chi_{p-1}) \ , \ 1 \leq p \leq q] \ , \tag{25}$$

where $\alpha = (\pi_0, \pi_1, \ldots, \pi_l, \ldots, \pi_{m-1}, \pi_m) [resp. \beta = (\chi_0, \chi_1, \ldots, \chi_p, \ldots, \chi_{q-1}, \chi_q)]$, by using the symbol of set sum,

$$\begin{array}{rcl} E_1 & = & \sum_{l<l'} \ C_l \times C_{l'} \ (resp.\ F_1 = \sum_{p<p'} \ D_p \times D_{p'}) \\ E_2 & = & \sum_l \ C_l^{[2]} \ (resp. F_2 = \sum_p D_p^{[2]}), \end{array}$$

where $X^{[2]} = \{(x,y) / x \in X, \ y \in X, \ x \neq y\}$, and

$$E_3 = \sum_{l>l'} \ C_l \times C_{l'} \ (resp.\ F_3 = \sum_{p>p'} \ D_p \times D_{p'}) \ . \tag{26}$$

Clearly, we have

$$\mathbb{P} \times \mathbb{P} = E_1 + E_2 + E_3 = F_1 + F_2 + F_3 \ . \tag{27}$$

The following decomposition is of the same type that as considered by Giakoumakis & Monjardet (1987) :

$$\mathbb{P} \times \mathbb{P} = \sum_{1 \leq i \leq 3} \sum_{1 \leq j \leq 3} E_i \cap F_j \ . \tag{28}$$

By introducing the following cardinalities

$$\{s_{ij} = card\,(E_i \cap F_j) \mid 1 \leq i \leq 3 \ , \ 1 \leq j \leq 3\} \ , \tag{29}$$

the following identities hold :

$$s_{11} = s_{33} \ , \ s_{12} = s_{32} \ , \ s_{13} = s_{31} \ and \ s_{21} = s_{23} \ . \tag{30}$$

And therefore, the Goodman and Kruskal cœefficient can be written

$$\gamma = \frac{s_{11} - s_{13}}{s_{11} + s_{13}} \tag{31}$$

It has the same mathematical meaning as the Hamann (1961) similarity index, defined at the level of the representation set $\mathcal{O}$. More precisely, if $\mathcal{O}(a)$ and $\mathcal{O}(b)$ are the two subsets to be associated, the latter index can be put in the following form

$$\eta = \frac{(s+t) - (u+v)}{(s+t) + (u+v)} \ , \tag{32}$$

where $s = card(\mathcal{O}(a) \cap \mathcal{O}(b))$, $u = card(\mathcal{O}(a) \cap \mathcal{O}(\tilde{b}))$, $v = card(\mathcal{O}(\tilde{a}) \cap \mathcal{O}(b))$ and $t = card(\mathcal{O}(\tilde{a}) \cap \mathcal{O}(\tilde{b}))$; $\mathcal{O}(\tilde{a})$ $[resp.\ \mathcal{O}(\tilde{b})]$ being the complementary subset of $\mathcal{O}(a)$ $[resp.\ \mathcal{O}(b)]$.

The Yule coefficient (1912) of which the expression is

$$Y = \frac{st - uv}{st + uv} \tag{33}$$

has also the same structure as the Goodman and Kruskal one. It is defined at the $\mathcal{O} \times \mathcal{O}$ level and does compare two total preorders on $\mathcal{O}$, into two classes each : $\mathcal{O}(\tilde{a}) < \mathcal{O}(a)$ for the former and $\mathcal{O}(\tilde{b}) < \mathcal{O}(b)$ for the latter.

It can be established [Lerman (1992)], in case of comparing two total preorders on an object set $\mathcal{O}$, that the γ numerator is a centralized index, as for the numerator of (3). An adequate mathematical representation and independence hypothesis have to be considered in this latter context. This is much more easier than the concerned here, where the total preorders are established on $\mathbb{P}$ [see (13)] and deduced from tree comparisons [see (15)]. Now, the justification of the γ denominator is to make this coefficient included between 0 and 1, where the latter value is reached only when any strict inversion between both total preorders exists.

The last and commonly used cœfficient we want to mention is the classical correlation cœfficient between the two level functions l_α and l_β, respectively associated with binary trees α and β [see (20)] [Sokal and Rohlf (1962)]. Namely,

$$\gamma(\alpha,\beta) = \frac{\sum\{[l_\alpha(ij) - \bar{l}_\alpha][l_\beta(ij) - \bar{l}_\beta] \mid ij \in \mathbb{P}\}}{\sqrt{(\sum_{ij}[l_\alpha(ij) - \bar{l}_\alpha]^2)(\sum_{ij}[l_\beta(ij) - \bar{l}_\beta]^2)}} \tag{34}$$

where ij $(1 \leq i < j \leq n)$ codes an element of the set $\mathbb{P}$ of unordered object pairs and where

$$\bar{l}_\omega = \frac{2}{n(n-1)} \sum\{l_\omega(ij) \mid ij \in \mathbb{P}\} \ , \tag{35}$$

for $\omega = \alpha$ *or* β.

l_α $(resp.\ l_\beta)$ function on $\mathbb{P}$ stands for the ultrametric dissimilarity directly defined by the tree α $(resp.\ \beta)$. We have suggested in the preceding section to replace the level function l_ω of an ω tree by the mean rank function λ_ω associated with the total preordonnance $UP(\omega)$. This proposition is done in order to take more intimately into account the shape of the tree, whatever is the number of its levels; and, at the same time, for normalization purpose. As a matter of fact, the common mean of λ_α and λ_β is $(p+1)/2$ [see (16)].

A correlation cœfficient like $\gamma(\alpha,\beta)$ [see (34)] is considered by Lapointe and Legendre (1995), with eventually replacement of the level function l_ω $(resp.\ l_\beta)$ by an ultrametric height function. The point of view developed in this paper is that of testing independence hypotheses. The considered random model comprises the permutational one (see §4.1 below). For the latter and relative to an ω tree, the valuation of a pair $\{i,j\}$ is implicitly given by

$$\mu_\omega(i\,j) = \frac{l_\omega(i\,j) - \bar{l}_\omega}{\sqrt{\sum_{i'\,j'} [l_\omega(i',j') - \bar{l}_\omega]^2}} \tag{36}$$

$1 \leq i < j \leq n$. Here, the mean and variance over $\mathbb{P}$ of the function μ_ω, are respectively 0 and 1/p.

Only simulations of the random permutational model, are taken into account in the mentioned work. Normal distribution could have been envisaged, in order to approximate the distribution of the correlation coefficient between trees (Daudé 1992).

4 Permutational approach for comparing classification trees

As said above (see section 2), the general principle considered here is the same as that one used for comparing numerical variables, viewed as unary relations. The new situations are provided by the specificity of the relations to be compared and by the manner in which these relations are mathematically represented.

4.1 First comparison method.

The ultrametric preordonnance UP_ω associated with an ω tree [as in (15)] is here coded by means of the "mean rank" valuation on the set $\mathbb{P}$ of unordered object pairs :

$$\wedge_\omega = \{\lambda_\omega\,(i,j) \mid 1 \leq i < j \leq n\} \tag{37}$$

More precisely and relative to (15), if ij belongs to $R(\pi_l) - R(\pi_{l-1})$, we have

$$\lambda_\omega(i\,j) = card[R(\pi_{l-1})] + \frac{1}{2} \times [card(R(\pi_l) - R(\pi_{l-1})) + 1]\,, \tag{38}$$

$1 \leq l \leq m$.

The random permutational model associates with ω, $\omega^* = \sigma(\omega)$ – by relabeling the leaves of ω – and then, with $\wedge_\omega$

$$\wedge_{\omega^*} = \{\lambda_{\omega^*}(i,j) = \lambda_\omega[\sigma(i),\sigma(j)] \mid 1 \leq i < j \leq n\}\,, \tag{39}$$

where σ is a random element in the set G_n – provided by an uniform probability measure – of all permutations on $I = \{1, 2, \ldots, i, \ldots, n\}$. In the previous notations,

$$\lambda_\omega[\sigma(i), \sigma(j)] = \lambda[min(\sigma(i), \sigma(j)), max(\sigma(i), \sigma(j))] \ , \tag{40}$$

$1 \leq i < j \leq n$.

Now, for the comparison between two trees α and β , we may introduce the raw index :

$$s(\alpha, \beta) = \sum_{\mathbb{P}} \lambda_\alpha(i, j)\ \lambda_\beta(i, j) \tag{41}$$

and associate with it the random raw index $s(\alpha^*, \beta^*)$, where $\alpha^* = \sigma(\alpha)$ and $\beta^* = \sigma(\beta)$ are independent. As a matter of fact, the distribution function of $s(\alpha^*, \beta^*)$ is the same as that of $s(\alpha, \beta^*)$ $[resp.\ s(\alpha^*, \beta)]$. Clearly,

$$s(\alpha, \beta^*) = \sum_{\mathbb{P}} \lambda_\alpha(i, j)\ \lambda_\beta\ [\tau(i), \tau(j)] \ , \tag{42}$$

where τ is a random element in the set G_n of all permutations on $I = \{1, 2, \ldots, i, \ldots, n\}$ (see above).

Here, we recognize a permutational random index which appeared in the statistical and data analysis literature in different contexts (Daniels 1944 ; Mantel 1967 ; Lecalvé 1976 ; Lerman 1977, 1992 ; Hubert 1983, 1987). An interesting interpretation of the standardized statistical version of this cœfficient is given in (Ouali-Allah 1991).

As for comparing numerical variables (see section 2), cœfficients as (3), (5) and (6) can be defined and mathematically computed. The reason is because equation as (4) remains valid, whatever the arity of the relations to be compared is, Lerman (1992). But here, expressions as (5) and (6) are not equivalent. More precisely, by writing, as for expression (3),

$$Q(\alpha, \beta) = \frac{s(\alpha, \beta) - E[s(\alpha^*, \beta^*)]}{\sqrt{var[s(\alpha^*, \beta^*)]}} \ , \tag{43}$$

the cœfficient

$$r(\alpha,\beta) = \frac{Q(\alpha,\beta)}{\sqrt{Q(\alpha,\alpha)\ Q(\beta,\beta)}}\ , \tag{44}$$

is nothing other than the usual correlation $Corr_p(\lambda_\alpha, \lambda_\beta)$ between λ_α and λ_β valuations over $\mathbb{P}$.

But, the limit form of

$$\rho(\alpha,\beta) = \frac{1}{\sqrt{n}}\ Q(\alpha,\beta) \tag{45}$$

has, essentially, different nature than $r(\alpha,\beta)$. It can be written as following

$$\rho(\alpha,\beta) = \frac{[s(\alpha,\beta) - p(p+1)^2\ /\ 2]}{\sqrt{[2A_\alpha - (p+1)^2][2A_\beta - (p+1)^2]}}\ , \tag{46}$$

where $p = n(n-1/2$ and

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

with $\omega = \alpha\ or\ \beta$.

The latter expression (46) is deduced from more general expressions Lerman (1987) in (1992), Ouali-Allah (1991). On the other hand, an exact mathematical form for $\rho(\alpha,\beta)$ is available.

Obviously, the tree shapes of α and β intervene intimately in $s(\alpha,\beta)$, A_α and A_β. The tree shapes will also, implicitly, play an important part in the second proposed method.

4.2 Second comparison method

We adopt here the strict mathematical representation (coding) of the ultrametric preordonnance associated with an ω tree, given by the subset $R(\omega)$ of $\mathbb{P}\times\mathbb{P}$ see (18). Recall that ω designates also the indicator function of $R(\omega)$ see (19).

In these conditions, the raw similarity index associated with the comparison of two trees α and β, has the following expression

$$s'(\alpha,\beta) = \sum\{\alpha(\{i,j\},\{i',j'\})\beta(\{i,j\},\{i',j'\}) \mid (\{i,j\},\{i',j'\}) \in J\times J\}\ , \tag{47}$$

where $J = \{\{i,j\} \mid 1 \leq i \neq j \leq n\}$ is the set of all unordered element pairs of $I = \{1,2,\ldots,i,\ldots,n\}$. J codes $\mathbb{P}$.

As before and according to general property, $s'(\alpha,\beta^*)$, $s'(\alpha^*,\beta)$ and $s'(\alpha^*,\beta^*)$ – where α^* and β^* are independent random trees – are equivalent versions of the random raw index. Then, let us consider

$$\begin{array}{rcl} s'(\alpha,\beta) & = & \sum \alpha(\{i,j\}\ ,\ \{i',j'\})\ \beta\ (\{\tau(i),\tau(j)\}\ ,\ \{\tau(i'),\tau(j')\}) \\ & & \mid (\{i,j\}\ ,\ \{i',j'\}) \in J \times J\}\,, \end{array} \tag{48}$$

where – as usual – τ is a random permutation in the set G_n of all permutations on I, equally distributed.

In order to obtain the standardized index

$$Q'(\alpha,\beta) = \frac{s'(\alpha,\beta) - E[s'(\alpha^*,\beta^*)]}{\sqrt{var[s'(\alpha^*,\beta^*)]}}\,, \tag{49}$$

we have to compute the exact values of $E[s'(\alpha,\beta^*)]$ and $E[(s'(\alpha,\beta^*))^2]$; where – as usual – E designates the mathematical expectation.

4.2.1 Computing of $E[s'(\alpha^*,\beta^*)]$

Equivalently, consider computing the mathematical expectation of $s'(\alpha,\beta^*)$. For the latter, it is necessary to decompose $J \times J$ as follows :

$$J \times J = \triangle + G + H \qquad (set\ sum) \tag{50}$$

where

$$\left\{ \begin{array}{l} \triangle = \{(\{i,j\},\{i,j\})\}\,, \\ G = \{(\{i,j\},\{i,k\})\}\ and \\ H = \{(\{i,j\},\{k,l\})\}\,, \end{array} \right. \tag{51}$$

In these expressions, distinct symbols indicate distinct elements of I; and we have the following equations :

$$\left\{ \begin{array}{l} card(\triangle) = p = n(n-1)/2 \\ card(G) = n(n-1)(n-2) \\ card(H) = n(n-1)(n-2)(n-3)/4 \end{array} \right. \tag{52}$$

Therefore, $E[s'(\alpha, \beta^*)]$ can be written :

$$+ \begin{array}{l} \sum_G \alpha(\{i,j\},\{i,k\})\, \beta(\{\tau(i),\tau(j)\},\{\tau(i),\tau(k)\}) \\ \sum_H \alpha(\{i,j\},\{k,l\})\, \beta(\{\tau(i),\tau(j)\},\{\tau(k),\tau(l)\}) \; ; \end{array} \tag{53}$$

because, the sum over $\triangle$ vanishes.

By denoting $n^{[x]} = n(n-1) \times \ldots \times (n-x+1)$, for an integer x, the following result can be established :

Theorem 1

$$E[s'(\alpha^*, \beta^*)] = + \begin{array}{l} n^{[3]}\, \pi_\alpha(G)\, \pi_\beta(G) \\ \frac{n^{[4]}}{4}\, \pi_\alpha(H)\, \pi_\beta(H) \; , \end{array} \tag{54}$$

where $\pi_\omega(G)$ [resp. $\pi_\omega(H)$ is the proportion of $G-$ elements $(\{i,j\},\{i,k\})$ [resp. $H-$ elements $(\{i,j\},\{k,l\})$] for which i and j are joined strictly before i and k (resp. k and l) in the ω *tree* ; $\omega = \alpha$ *or* β.

We may qualify a G(resp. H) $-$ element, as an "attested" ω $G(resp.\ H)-$ element; if the latter is counted in the above $\pi_\omega(G)$ $[resp.\ \pi_\omega(H)]$ proportion. Hence, the problem arises to have a method for determining the number of attested ω G $(resp.\ H)$ elements. These numbers depend strongly on the ω tree shape ($\omega = \alpha$ *or* β). They can be denoted $n_\omega(G)$ and $n_\omega(H)$; and then, obviously, we have :

$$\begin{array}{c} \pi_\omega(G) = \frac{n_\omega(G)}{n(n-1)(n-2)} \quad and \\ \pi_\omega(H) = \frac{4\, n_\omega(H)}{n(n-1)(n-2)(n-3)} \end{array} \tag{55}$$

Clearly, each subtree of ω (Figure 3) does increment $n_\omega(G)$ two unities; one for $(\{i,j\},\{i,k\})$ and one for $(\{i,j\},\{j,k\})$. Then twice the number of such ω subtrees gives $n_\omega(G)$.

On the other hand, each ω subtree of the following forms (Figure 4) intervenes in counting $n_\omega(H)$: once for (b) or (c) ; but three times for (d) or (e). Therefore, $n_\omega(H)$ is the total number of subtrees having the forms (b) or (c) with addition of three times the number

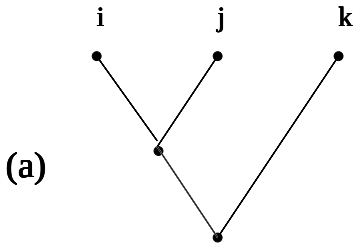
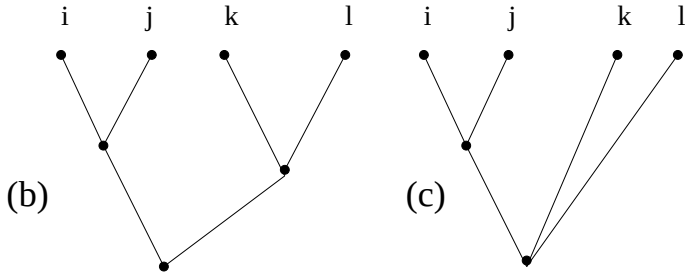


Figure 3:



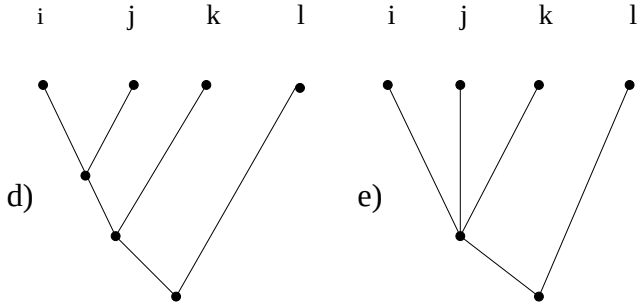


Figure 4:

of trees having the forms (d) or (e).

At the end of this paper we will try to give mathematical definition of a tree shape. In spite of this, it seems very complicated to derive mathematical formula for $n_\omega(G)$ and $n_\omega(H)$. An appropriate solution for this problem is an algorithmic one. The specified algorithm has to enumerate all the ω subtrees of the above forms (a), (b), (c), (d) and (e). The same type of

problem will arise, in much more complicated version, in case of computing the variance of $s'(\alpha^*, \beta^*)$.

4.2.2 Computing of $var[s'(\alpha^*, \beta^*)]$

We have to exactly evaluate $E[(s^v(\alpha^*, \beta^*))^2]$. The direct expression of this mathematical expectation is given by :

$$\begin{array}{l} \sum\{\alpha(\{i,j\},\{i',j'\})\ \alpha(\{i'',j''\},\{i''',j'''\}) \\ \times\ \beta(\{\tau(i),\tau(j)\},\{\tau(i'),\tau(j')\})\ \beta(\{\tau(i''),\tau(j'')\},\{\tau(i'''),\tau(j''')\} \\ |\ ((\{i,j\},\{i',j'\}),(\{i'',j''\},\{i''',j'''\}) \in (J\times J)\times(J\times J)\} \end{array} \tag{56}$$

In order to detect invariance properties, we have to decompose the set $(J \times J) \times (J \times J)$, over which the sum is, according to the structure of
$((\{i,j\},\{i',j'\}),(\{i'',j''\},\{i''',j'''\}))$.
This structure is defined from repetitions of I elements in the couple of couples of unordered element pairs of I. Each structure determines a "configuration". As an example consider the following one, where distinct symbols indicate different elements of I :

$$((\{i,j\},\{k,l\}),(\{i,m\},\{j,m\}))\,, \tag{57}$$

it belongs to $H \times G$ [see(51)]. This configuration defines a class of $H \times G$ which comprises $n(n-1)(n-2)(n-3)(n-4)/2$ elements.

Therefore and first, decompose $(J \times J - \triangle) \times (J \times J - \triangle)$ according to the bipartition of $J \times J - \triangle$ into the two classes G and H :

$$\begin{array}{rcll} (J\times J-\triangle)\times(J\times J-\triangle) & = & (G+H)\times(G+H) & \\ & = & G\times G+G\times H+H\times G+H\times H & (set\ sum) \end{array} \tag{58}$$

and split each of the four classes into subclasses respectively associated with the different configurations. The detail of all the configurations and the number of represented elements for each of them is explicitly given in section 6. The following table gives the number of configurations included in each of the above subsets [see (58)].

set	G × G	G × H	H × G	H × H
number of configurations	34	25	25	26

Table 1

Now, let us designate by $\mathcal{C}$, the set of all configurations. According to the above table, $\mathcal{C}$ comprises 110 elements; and $\mathcal{C}$ can be generated according to the above decomposition into four classes. If c is an element of $\mathcal{C}$ and $C = C(c)$ the associated subset of $(J \times J - \triangle)^2$; by denoting $m(C)$ the cardinality of C, we may express the following property :

Theorem 2

$$E[(s'(\alpha^*, \beta^*))^2] = \sum_{c \in \mathcal{C}} m(C)\ \pi_\alpha(c)\ \pi_\beta(c)\,, \tag{59}$$

where $\pi_\omega(C)$ is the proportion of C-elements $[(\{i,j\},\{i',j'\}),(\{i'',j''\},\{i''',j'''\})]$ for which the first and the third pairs $(\{i,j\}\, and \{i'',j''\})$ are joined strictly before the second and the fourth pairs $(\{i',j'\}\, and\, \{i''',j'''\})$, in the ω tree, $\omega = \alpha\ or\ \beta$.

Consequently, we have to enumerate the set of C-elements for which the stated condition of the above theorem, holds. For this purpose, we have to introduce the notion of a c-compatible type of an ω subtree. The number of leaves of the latter is the number of distinct elements which intervene in the c configuration, it is comprised between 3 and 8; 3 in case of $[(\{i,j\},\{i,k\}),(\{i,j\},\{i,k\})]$ type and 8 in case of $[(\{i,j\},\{k,l\}),(\{p,q\},\{r,s\})]$ type. In the latter and as previously, distinct symbol letters indicate distinct elements of I.

As an example, consider the following c-configuration which belongs to $H \times G$:

$$[(\{i,j\},\{k,l\}),(\{i,m\},\{j,m\})]$$

We are going to illustrate two cases (among others) of compatible trees. For each of them we will give the number of times where the above congiguration c is instanciated.
The first compatible tree which is defined on the set $\{x,y,z,u,v\}$, is the following (Figure 5).

It is easy to see that the subset $\{i,j,m\}$ must be instanciated by $\{x,y,z\}$. On the other hand, the repeated element m, that we can call a pivotal element, is necessarily x or y. And then we have the two following instanciations of c : $((\{y,z\},\{u,v\}),(\{x,y\},\{x,z\}))$ and

$((\{x,z\},\{u,v\}),(\{x,y\},\{y,z\}))$.

The second compatible tree which is also represented on the set $\{x,y,z,u,v\}$, has the following form (Figure 6). It gives rise to eight instanciations of the above c-configuration. To realize that, begin by constituting the right ordered pair of unordered element pairs $(\{i,m\},\{j,m\})$, where m indicates the pivotal element. For this purpose, we have to choose a subset of size three in the set $\{x,y,z,u\}$. Afterwards we have to choose on among two possible elements. As an example, consider the 3-subset $\{x,z,u\}$, the pivotal elements can be x or z. Therefore, the eight instanciations of the configuration c can be expressed as follows :
$((\{y,z\},\{u,v\}),(\{x,y\},\{x,z\}))$

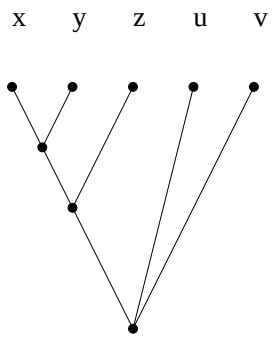


Figure 5:

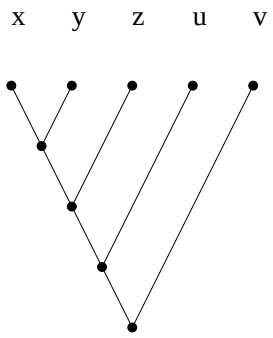


Figure 6:

$((\{x,z\},\{u,v\}),(\{x,y\},\{y,z\}))$
$((\{y,u\},\{z,v\}),(\{x,y\},\{x,u\}))$
$((\{x,u\},\{z,v\}),(\{x,y\},\{y,u\}))$
$((\{z,u\},\{y,v\}),(\{x,z\},\{x,u\}))$
$((\{x,u\},\{y,v\}),(\{x,z\},\{z,u\}))$
$((\{z,u\},\{x,v\}),(\{y,z\},\{y,u\}))$
$((\{y,u\},\{x,v\}),(\{y,z\},\{z,u\}))$

Therefore, for a given configuration c and an ω tree, the general enumeration method can be decomposed as follows :

- Derive all types of c-compatible subtrees.
- For a subtree of a given type, determine how many countable elements of $C(c)$, it does give rise.
- For a given type, determine how many subtrees of this type there are, in the whole ω tree.

Let us consider one more example for which the number of elements of $C(c)$ associated with a c-compatible ω subtree, is more than only one.

Relative to the following c-configuration, belonging to $H \times H$:

$$[(\{i,j\},\{k,l\}),(\{i,p\},\{q,r\})]\,,$$

the following subtree is c-compatible (Figure 7) : We focus here on the pairs $\{i,j\}$ and

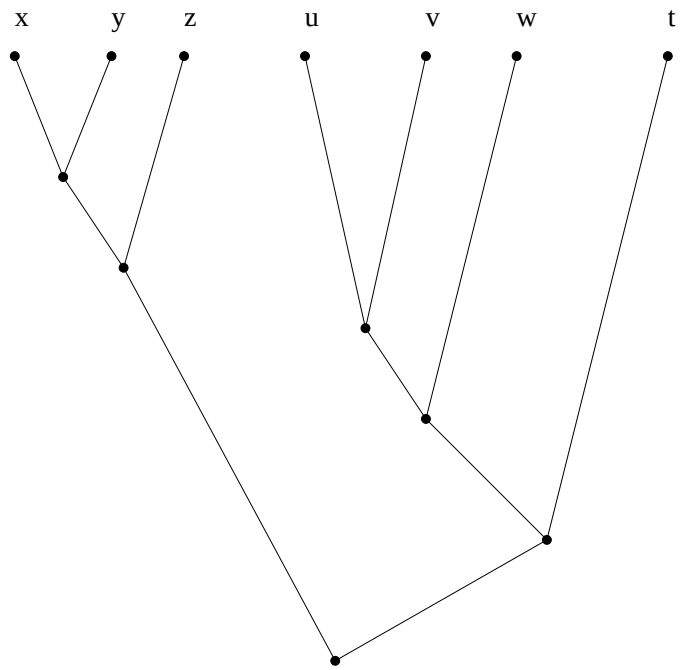


Figure 7:

$\{i,p\}$ which respectively are the first components of both ordered pairs of object pairs $(\{i,j\},\{k,l\})$ and $(\{i,p\},\{q,r\})$. The subset $\{i,j,p\}$ has necessarily an empty intersection with the subset $\{u,v,w,t\}$. Because if not, it would be impossible to constitute $\{k,l\}$ or $\{q,r\}$ with the conditions $\{i,j\} < \{k,l\}$ and $\{i,p\} < \{q,r\}$,
according to the tree structure. Therefore $\{i,j,p\}$ is identical to $\{x,y,z\}$. In these conditions, there are six possibilities for forming $(\{i,j\},\{i,p\})$; namely :
$(\{x,y\},\{x,z\}),(\{x,y\},\{y,z\}),(\{x,z\},\{x,y\}),(\{x,z\},\{y,z\}),(\{y,z\},\{x,y\})$ and $(\{y,z\},\{x,z\})$.
For each possibility, there are $2 \times \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 12$ choices for forming $(\{k,l\},\{q,r\})$, where necessarily $\{k,l,q,r\} = \{u,v,w,t\}$.
Then in all, there are 72 instanciations of the above configuration, from the above tree.

Now, let us denote by $T_\omega(c)$ the set of all ω subtrees types compatible with the c configuration. If $t_\omega(c)$ is a given element of $T_\omega(c)$, we may designate by $n[t_\omega(c)]$ the number of times for which the type $t_\omega(c)$ is instanciated in the whole ω tree. For a given instanciation, $l[t_\omega(c)]$ indicates the number of distinct replications of the c-configuration, which can be obtained in a compatible way, from a given $t_\omega(c)$ subtree. In these conditions, the cardinal – that we denote by $m(\omega, G)$ – which defines the numerator of the ratio $\pi_\omega(c)$, can be put in the following form :

$$\sum_{t_\omega(c) \in T_\omega(c)} n[t_\omega(c)] \times l[t_\omega(c)] \tag{60}$$

Hence, we may state the subsequent property

Property 1: Relative to a given configuration c, the proportion of c-elements compatible with an $\omega - tree$ can be expressed by

$$\pi_\omega(C) = \frac{m(\omega, C)}{m(C)} = \frac{\sum\{n[t_\omega(c)] \times l[t_\omega(c)] \;/\; t_\omega(c) \in T_\omega(c)\}}{m(C)} \tag{61}$$

where the different components of this equation are specified above.

Mathematical expression for $m(C)$ can be provided without great difficulty. However tractable analytical formula for $m(\omega, C)$ depending on the ω tree shape, seems to be very hypothetical to obtain. And that, even characterization is provided in order to capture formally the ω tree shape. For this purpose we may introduce a notion of " indexed type of a classification tree". It does correspond to the sequence of the partition types, associated with the level tree decreasing sequence.

Let us begin by giving an example before more formal definition. For the following tree a (Figure 8), the indexed type is

$$\tau(a) = [8, (5, 3), (3, 1, 1, 2, 1), (1, 1, 1, 1, 1, 1, 1, 1)]$$

More generally, for the following tree b (Figure 9), The indexed type is

$$\begin{aligned} \tau(b) \quad = \quad & [n, (n_1, n_2), (n_{11}, n_{12}, n_{13}, n_2), \\ & (n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{24}), \\ & (n_{111}, n_{112}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{24})], \end{aligned}$$

where, in the figure, we have indicated by $N_{i_1 i_2 \ldots i_k}$ the object class of which the cardinality is $n_{i_1 i_2 \ldots i_k}$.

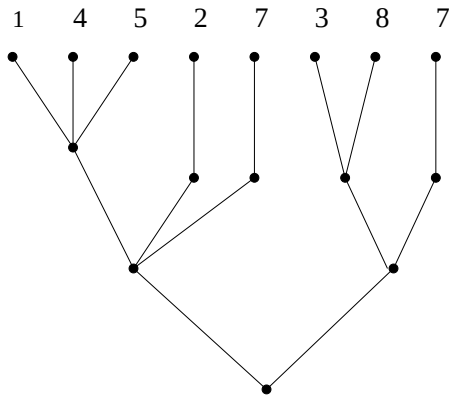


Figure 8:

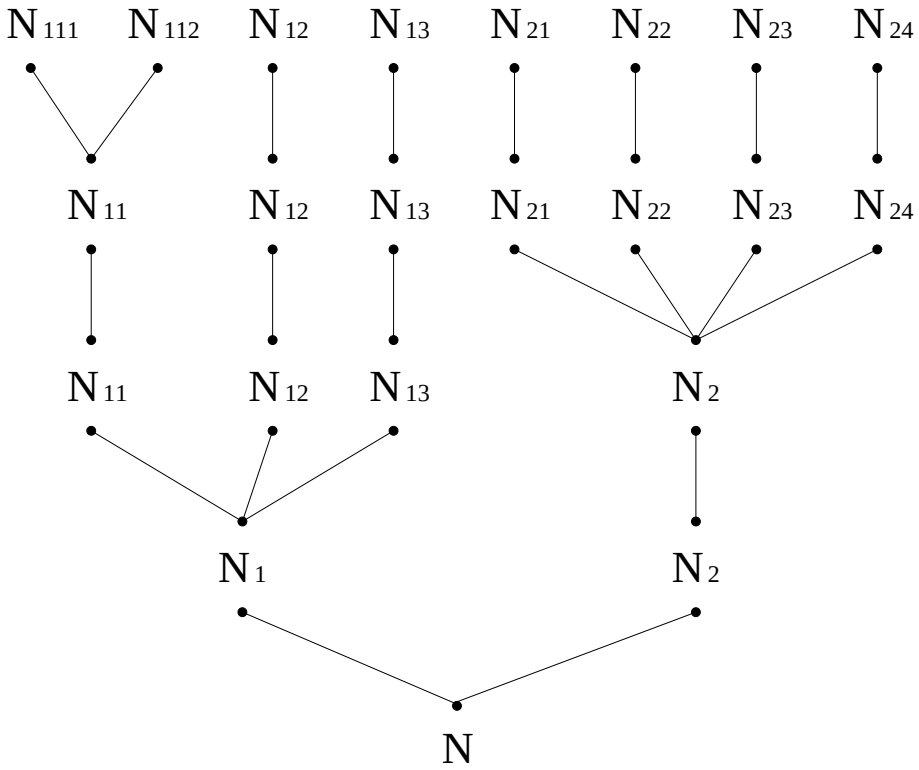


Figure 9:

More precisely we have the next definition.

Definition. The indexed type $\tau(\omega)$ of a classification tree ω is the sequence of the partition types, associated with the decreasing sequence of the tree levels. In each partition type, the subscript of a class cardinal indicates the increasing sequence of its superclasses, ordered by inclusion (112 indicates 11, 1, $\emptyset$; and $N_{112} \subset N_{11} \subset N_1 \subset N$). This subscript can be written $i_1 i_2 \ldots i_{k-1} i_k$ and indicates that $N_{i_1 i_2 \ldots i_{k-1} i_k}$ is the i_k^{th} subclass – from the left to the right – of the class $N_{i_1 i_2 \ldots i_{k-1}}$. In a given partition type, the subscripts are lexicographically ordered from the left to the right [see above $\tau(b)$).

Property 2 : The cardinal $m(\omega, C)$ is a function of the couple $[c, \tau(\omega)]$.

This property is clear to be seen. However, the induced function is very complicated to be set up. The solution we propose is an algorithmic one. It must follow the general scheme pictured above. Experimental work will be considered in near future.

Notice that the previous definition gives for the number of object pairs joined at the "first time" at the k^{th} level, the following equation

$$p_k = \sum_{i_1 \ldots i_{k-1}} \sum_{\{i_k, i'_k\}} n_{i_1 \ldots i_{k-1}, i_k} \, n_{i_1 \ldots i_{k-1} i'_k} \tag{62}$$

and then (see § 3.1), we have the following relation between the level function l_ω and the mean rank function λ_ω associated with an ω tree :

$$\begin{array}{l} (\forall (i,j) \in \mathbb{P}) \, ; \\ l\omega(i,j) = k \Leftrightarrow \\ \lambda_\omega(i,j) = p_1 + p_2 + \cdots + p_{k-1} + \frac{1}{2}(p_k + 1) \, . \end{array} \tag{63}$$

5 Comparing q-ary relations and concluding remarks.

Comparison between q-ary relations is outlined in (Lerman 1992). In order to situate the previous development, let us recall the elements of this comparison.

Let $\mathcal{O}^{[q]}$ designate the set of sequences of q objects, mutually distinct. We call such a sequence q-uple, that we indicate by $(i_1, i_2, \ldots, i_q)$, where $\{i_1, i_2, \ldots, i_q\}$ is a q subset of $I = \{1, 2, \ldots, n\}$, the set of labels which codes $\mathcal{O}$. The cardinality of $\mathcal{O}^{[q]}$ is $n(n-1)\ldots(n-q+1)$; and for the comparison of two weighted (valued) q-ary relations, denoted

$$\{\mu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\} \, , \tag{64}$$

$$\{\nu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\}\,; \tag{65}$$

the raw similarity index takes the following form :

$$s(\mu, \nu) = \sum \{\mu_{i_1 i_2 \ldots i_q}\ \nu_{i_1 i_2 \ldots i_q} \mid (i_1, i_2, \ldots, i_q) \in I^{[q]}\} \tag{66}$$

where μ $(resp.\ \nu)$ is a numerical or logical (i.e. binary) valuation.

If μ^* and ν^* are independent random valuations, respectively associated with μ and ν, under the permutational model, then the random indices $s(\mu, \nu^*)$, $s(\mu^* \nu)$ and $s(\mu^*, \nu^*)$ have the same distribution law.
The mathematical expectation and the absolute second moment can be expressed as follows.

$$E[s(\mu, \nu^*)] = n^{[q]}\ \bar{\mu}\ \bar{\nu} \tag{67}$$

where we have denoted by $n^{[q]}$, $n(n-1) \times \cdots \times (n-q+1)$ and where $\bar{\mu}\,(resp.\ \bar{\nu})$ designates the mean of the $\mu\,(resp.\ \nu)$ valuation over $\mathcal{O}^{[q]}$;

$$E[(s(\mu, \nu^*))^2] = \sum_{0 \le r \le q} \sum_{c_r} \frac{1}{n^{[2q-r]}} \Big(\sum_{C(c_r)} \mu_{i^1 \ldots i^q}\ \mu_{j^1 \ldots j^q} \Big) \times \Big(\sum_{C(c_r)} \nu_{i^1 \ldots i^q}\ \nu_{j^1 \ldots j^q} \Big) \tag{68}$$

where c_r is a configuration of $((i_1, i_2, \ldots, i_q), (j_1, j_2, \ldots, j_q))$ for which r components of $(i_1, i_2, \ldots, i_q)$ are repeated in $(j_1, j_2, \ldots, j_q)$. There are

$$\left(\begin{array}{c} q \\ r \end{array} \right)^2 r\,! \tag{69}$$

different configurations c_r. $C(c_r)$ denotes the set of ordered pairs of q-uples $((i_1, i_2, \ldots, i_q), (j_1, j_2, \ldots, j_q))$ having the same configuration c_r. We have

$$\begin{aligned} card[C(c_r)] &= n^{[q]} \times (n-q)^{[q-r]} \\ &= n^{[2q-r]} \end{aligned} \tag{70}$$

The total number of configurations is given by

$$\sum_{0 \leq r \leq q} \left(\begin{array}{c} q \\ r \end{array} \right)^2 r\,! \,, \tag{71}$$

its value for $q = 4$ is 209. This number is much greater than the necessary number of configurations (137, see Table 1) to have to be considered in case of tree comparison. This, because we have taken into account, in the latter case, the specificity of the relations to be associated.

The order of the computational complexity is n^{2q}. This number becomes too large if n is not enough small. For example, for $n = 100$ and $q = 4$, $n^{2q} = 10^8$. However, parallelization computing procedure can be envisaged. Anyway, for our problem of tree classification comparison, one may limit this comparison to its most interesting part, by considering truncated trees. The truncation may consist of deleting the first levels of both trees, starting by a signifiant classification for each of them (Lerman and Ghazzali 1991), having approximatively the same number of classes. This provides a major simplification in determining $m(\omega, C)$ [see equation (61)] by an algorithmic manner. Roughly speaking, the number n is replaced by the number of leaves of the retained tree.

The importance of the scale, with respect to which an association coefficient is established, is not enough emphasized in data analysis literature. It is now admited and mainly evocated in the binary case (Hubert 1983, Messatfa 1990), that the numerator of the association coefficient has to be centralized. The reduction proposed is often based on the maximum of the numerator. This may give rise to very difficult problems of combinatorial optimization (Lerman 1987; Lerman and Peter 1988; Messatfa 1992). In our case and relative to our latter mathematical coding, this leads to the intractable problem of finding the permutation σ which maximizes $s'(\alpha, \beta(\sigma))$ [see (48)]. For statistical reasons and according to likelihood linkage analysis (LLA) classification method (Lerman 1993), we have adopted reduction by means of the standard deviation of $s'(\alpha, \beta^*)$ [see (48)]. And, we have shown the computing problem to be tractable by means of a polynomial algorithmic procedure. The algorithmic research is now under study and will give soon matter to a future paper.

6 Appendix : structural decomposition of $(G+H)\times(G+H)$

We are going here to make explicit the structural decomposition of $(G+H)\times(G+H)$ (see paragraph 4.2.2) and then, to justify the content of table 1. On the other hand, we will give the cardinality associated with each substructure defining a given configuration c of an ordered pair of which each component is an ordered pair of unordered object pairs, such as :

$$((\{x,y\},\{z,t\}),(\{x',y'\},\{z',t'\}))$$

An unordered object pair such as x,y will be denoted here by a word with two letters xy, of which the first letter x precedes lexicographically the second one y.
First recall the general equation (58) :

$$(G+H)\times(G+H)=G\times G+G\times H+H\times G+H\times H \quad (set\ sum)$$

and let us designate by $\mathcal{U}\times\mathcal{V}$ one of the four subsets of the left member of this equation ($\mathcal{U}=G\,or\,H\,and\,\mathcal{V}=G\,or\,H$). Our general decomposition strategy consists of organizing the structure of $\mathcal{V}$ with respect to a given element of $\mathcal{U}$. If (ξ,η) belongs to $\mathcal{U}\times\mathcal{V}$, its configuration $c=c(\xi,\eta)$ is conditionned by the manner in which the objects appearing in ξ are repeated in η. The cardinality of c is the number of elements of $\mathcal{U}\times\mathcal{V}$ covered by the configuration c.

6.1 Decomposition of $\mathcal{U}\times\mathcal{V}=G\times G$

Let $\xi=(xy,xz)$ be a given element of G. Consider the set $\{x,y,z\}$ of the three elements which intervene in the constitution of ξ. We have to consider four general cases according to the number of objects distinct from x,y or z and which intervene in the construction of η. This number can be 0,1,2 or 3 ; and then, the cases will be denoted according to this number. Now we are going to give below the different structures of $\eta=(x'y',x'z')$ and the associated cardinalities of $c=c(\xi,\eta)$.

6.1.1 Structures of η for the case 0

- $(xy,xz),\ card(c)=n(n-1)(n-2)$;
- $(xz,xy),\ card(c)=n(n-1)(n-2)$;
- $(xy,yz),\ card(c)=n(n-1)(n-2)$;
- $(yz,xy),\ card(c)=n(n-1)(n-2)$;
- $(xz,yz),\ card(c)=n(n-1)(n-2)$;
- $(yz,xz),\ card(c)=n(n-1)(n-2)$;

6.1.2 Structures of η for the case 1

- $(xy, xu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xu, xz),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xu, xy),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xz, xu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xy, yu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xz, zu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yu, xy),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zu, xz),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yz, yu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yu, yz),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yz, zu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zu, yz),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xu, yu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yu, xu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xu, zu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zu, xu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yu, zu),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zu, yu),\ card(c) = n(n-1)(n-2)(n-3)$;

6.1.3 Structures of η for the case 2

- $(xu, xv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(xu, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(uv, xv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(yu, yv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(yu, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(uv, yu),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;

- $(zu, zv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(zu, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(uv, zu),\ card(c) = n(n-1)(n-2)(n-3)(n-4)$;

6.1.4 Structures of η for the case 3

- $(uv, uw),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)$.

Finally, the number of distinct configurations is 34. On the other hand one may verify that the sum of the cardinalities, which can be put in the following form

$$6n(n-1)(n-2) + 18n(n-1)(n-2)(n-3) + 9n(n-1)(n-2)(n-3)(n-4) + n(n-1)(n-2)(n-3)(n-4)(n-5),$$

is nothing other than $[n(n-1)(n-2)]^2$ which represents the cardinal of $G \times G$.

6.2 Decomposition of $\mathcal{U} \times \mathcal{V} = G \times H$

As for the preceding section 6.1, $\xi = (xy, xz)$ will designate a given element of G. We also distinguish here four cases according to the number of elements of the set $\{x, y, z\}$ which intervene in the building of the element η belonging to H. Let us denote by case i, the case for which $(3-i)$ elements of $\{x, y, z\}$ are repeated in η.

6.2.1 Structures of η for the case 0

- $(xy, zt),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zt, xy),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(xz, yt),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yt, xz),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(yz, zt),\ card(c) = n(n-1)(n-2)(n-3)$;
- $(zt, yz),\ card(c) = n(n-1)(n-2)(n-3)$;

6.2.2 Structures of η for the case 1

- $(xy, tu),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(tu, xy),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(xz, tu),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;

- $(tu, xz),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2;$
- $(yz, tu),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2;$
- $(tu, yz),\ card(c) = n(n-1)(n-2)(n-3)(n-4)/2;$
- $(xt, yu),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$
- $(yu, xt),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$
- $(xt, zu),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$
- $(zu, xt),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$
- $(yt, zu),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$
- $(zu, yt),\ card(c) = n(n-1)(n-2)(n-3)(n-4);$

6.2.3 Structures of η for the case 2

- $(xt, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$
- $(uv, xt),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$
- $(yt, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$
- $(uv, yt),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$
- $(zt, uv),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$
- $(uv, zt),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2;$

6.2.4 Structures of η for the case 3

- $(tu, vw),\ card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4.$

One may verify that the sum of the above cardinalities of the 25 categories is equal to $n(n-1)(n-2) \times (1/4)n(n-1)(n-2)(n-3) = card(G \times H)$.
It is obvious that the decomposition of $H \times G$ is structurally analogous to that of $G \times H$.

6.3 Decomposition of $\mathcal{U} \times \mathcal{V} = H \times H$

Let $\xi = (xy, zt)$ be a given element of $\mathcal{U} = H$ and let us designate by $E(\xi)$ the set $\{x, y, z, t\}$ including the four elements which appear in ξ. $D(\xi)$ will indicate the complementary subset of $E(\xi)$ and we have $card(D(\xi)) = n - 4$.
As previously, the structural decomposition of $\mathcal{V} = H$ will be elaborated according to the repetitions of x, y, z or t, in the components of the element $\eta = (x'y', z't')$ which belongs to $\mathcal{V} = H$. But in this situation the respective roles of x and y (resp. z and t) are equivalent. To illustrate this point, consider the two following elements of $H \times H$:

$$((xy, zt), (xz, yt)) \text{ and } ((xy, zt), (yt, xz))$$

and notice that they belong to the same configuration. Indeed, in both cases $x'y'$ (resp. $z't'$) is formed by taking one element from $\{x, y\}$ and one element from $\{z, t\}$.
Thus, we have to introduce three sets $\{x, y\}$, $\{z, t\}$ and $D(\xi)$ of which the cardinalities are 2, 2 and $(n-4)$ and that we respectively label by 1, 2 and 3. Consequently, the characterization of the configuration c of $(\xi, \eta) = ((xy, zt), (x'y', z't'))$ does only depend on the set labels of which x', y', z' and t' are provided. For example, the configuration concerned by the two above elements of $H \times H$ is, for the η definition, $(12, 12)$. Therefore, a given configuration will be specified by an ordered pair of two numbers associated with $(x'y', z't')$. The first (resp. second) number can be $11, 12, 13, 22, 23, 33$. Finally notice that if the same label ($1, 2$ *or* 3) appear more than one time in the definition of the configuration of (ξ, η), the concerned objects are necessarily distinct; precisely because η belongs to H. As above we are going to distinguish five cases according to the number of times where the set labeled 3 intervenes for providing η. Case i is that for which the set 3 intervenes i times, $0 \leq i \leq 4$.

6.3.1 Structures of η for the case 0

- $(11, 22)$, $card(c) = n(n-1)(n-2)(n-3)/4$;
- $(22, 11)$, $card(c) = n(n-1)(n-2)(n-3)/4$;
- $(12, 12)$, $card(c) = n(n-1)(n-2)(n-3)$;

6.3.2 Structures of η for the case 1

- $(11, 23)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(23, 11)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(12, 13)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(13, 12)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(12, 23)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(23, 12)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)$;
- $(22, 13)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;
- $(13, 22)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)/2$;

6.3.3 Structures of η for the case 2

- $(11, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(33, 11)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(13, 13)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(22, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(33, 22)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/8$;
- $(23, 23)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(12, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(33, 12)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)/2$;
- $(13, 23)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)$;
- $(23, 13)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)$;

6.3.4 Structures of η for the case 3

- $(13, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(33, 13)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(23, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;
- $(33, 23)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)/4$;

6.3.5 Structures of η for the case 4

- $(33, 33)$, $card(c) = n(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)(n-7)/16$.

The number of categories is 26 and one may verify that the sum of the above cardinalities is equal to

$$(n(n-1)(n-3)(n-4)/4)^2 = card(H \times H).$$

ACKNOWLEDGEMENT

We are indebted to François Rouxel, PhD Student, for his algorithmic research which leads us to reconsider some structural aspects of enumerative problems.

References

BAKER, F.B. (1974), "Stability of two hierarchical grouping techniques", *Journal of the American Statistical Association*, 69, 440-445.

DANIELS, H.E. (1944), "The relation between measures of correlation in the universe of sample permutations", *Biometrika*, 33, 129-135.

DAUDÉ, F. (1992), *Analyse et Justification de la Notion de Ressemblance dans l'Optique de la Classification Hiérarchique par AVL*, Thèse de l'Université de Rennes I, 24 juin 1992.

FOWLKES, E.B., and MALLOWS, C.L. (1983), "A method for comparing two hierarchical clusterings", *Journal of the American Statistical Association*, 78, 553-584.

GOODMAN, L.A., and KRUSKAL, W.H. (1954), "Measures of association for cross classification", *Journal of the American Statistical Association*, 49, 732-764.

HÁJEK, J., and SIDAK, Z. (1967), *Theory of Rank Tests*, Academic Press, New-York and London.

HAMANN, V. (1961), "Merkmalbestand und verwandtschaft sbeziehungen der farinosae. Ein Beitragzum System der Monokotyledonen", *Willdenowia*, 2, 639-768.

HUBERT, L.J. (1983), "Inference procedures for the evaluation and comparison of proximity matrices", *Numerical Taxonomy*, Ed. J. Felsenstein, NATO ASI Series, Berlin, Springer Verlag.

HUBERT, L.J. (1987), *Assignment Methods in Combinatorial Data Analysis*, Marcel Decker, New-York.

JOVICIC, A. (1996), "Minimal entropy algorithm for solving node problems", IFCS-96, *Data Science Classification and Related Methods*, Abstracts Vol.2, 115-116.

LAPOINTE, F.J., and LEGENDRE, P. (1995), "Comparison tests for dendrograms : A comparative evaluation", *Journal of Classification*, 12, 265-282.

LECALVÉ, G. (1976), "Un indice de similarité pour des variables de types quelconques", *Statistique et Analyse des Données*, 01-02, 39-47.

LERMAN, I.C. (1970), *Les Bases de la Classification Automatique*, Gauthier-Villars, collection Programmation, Paris.

LERMAN, I.C. (1977), "Formal analysis of a general notion of proximity between variables",

Congrès Européen des Statisticiens, Grenoble 1976, *Recent Developments in Statistics*, North Holland, 787-795.

LERMAN, I.C. (1984), "Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées", *Publications de l'Institut de Statistique de l'Université de Paris*, XXIX, fasc. 3-4, 27-57.

LERMAN, I.C. (1987), "Maximisation de l'association entre deux variables qualitatives ordinales", *Revue Mathématiques et Sciences Humaines*, 100, 49-56.

LERMAN, I.C. (1989), "Formules de réactualisation en cas d'agrégations multiples", *RAIRO*, série R.O., vol. 23, 2, 151-163.

LERMAN, I.C. (1992), "Conception et analyse de la forme limite d'une famille de cooefficients statistiques d'association entre variables relationnelles", I and II : *Revue Mathématiques Informatique et Sciences Humaines*; I : 118, 35-522, II : 119, 75-100.

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., and GHAZZALI, N. (1991), "What do we retain from a classification tree ? An experiment in image coding", *Symbolic-Numeric Data Analysis and Learning*, Edit. E. Diday and Y. Lechevallier, Nova Science Publishers, 27-42.

LERMAN, I.C., and PETER, Ph. (1988), "Structure maximale pour la somme des carrés d'une contingence aux marges fixées; une solution algorithmique programmée", *RAIRO* série R.O., vol. 22, 2, 83-136.

MANTEL, N. (1967), "Detection of disease clustering and a generalized regression approach", *Cancer Research*, vol. 2, 2, 209-220.

MESSATFA, H. (1990), *Unification Relationnelle des Critères et Structures Optimales des Tables de Contingence*, Thèse de doctorat de l'Université de Paris 6, 5 mars 1990.

MESSATFA, H. (1992), "An algorithm to maximize the agreement between partitions", *Journal of Classification*, 9, 5-15.

OCHIAI, A. (1957), "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions", *Bull. Jap. Soc. Sci. Fish*, T22, 526-530.

OUALI-ALLAH, M. (1991), *Analyse en Préordonnances des Données Qualitatives, Applications aux Données Numériques et Symboliques*, Thèse de doctorat de l'Université de Rennes

I, 5 décembre 1991.

SOKAL, R.R., and ROHLF, F.J. (1962), "The comparison of dendograms by objective methods", *Taxon*, 11, 33-40.

YULE, G.U. (1912), "On the methods of measuring the association between two attributes", *Journal Royal Statistical Society*, 75, 579-652.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399