RAIRO. RECHERCHE OPÉRATIONNELLE

ISRAËL-CÉSAR LERMAN

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

RAIRO. Recherche opérationnelle, tome 33, n° 3 (1999), p. 339-365

<http://www.numdam.org/item?id=RO_1999__33_3_339_0>

RAIRO Operations Research
RAIRO Oper. Res. **33** (1999) 339-365

COMPARING CLASSIFICATION TREE STRUCTURES: A SPECIAL CASE OF COMPARING q-ARY RELATIONS (*)

by Israël-César LERMAN ([1])

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 4-ary relation. The most classical approaches to tree comparison are discussed in the context 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.*

Keywords: Classification tree, relations, mathematical representation, random permutational model.

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 composantes de cette construction fondée sur un modèle permutationnel, quand les structures à comparer sont des arbres de classification.*

Mots clés : 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 (Arabie & Hubert 1992; Guénoche & Monjardet 1987; Hubert 1987;

(*) Received April 1997.

([1]) Irisa, Université de Rennes 1, Campus de Beaulieu, Avenue du Général Leclerc, 35042 Rennes Cedex, France.

Recherche opérationnelle/Operations Research, 0399-0559/99/03/$ 7.00

Lerman 1992; Marcotorchino & Michaud 1979; Messatfa 1990; Régnier 1965). It does intervene crucially on the different levels of a data synthesis process. Thus, descriptive variables 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: (1) associating objects described by relational variables and (2) 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) (see the above indicated references) is devoted to the cases of $q = 1$ or 2. And, in the latter, not enough attention is paid to 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 purpose. However, the generality of this coefficient makes it not accurate enough for the concerned structures. The general method we set up (Lerman 1992), has its origin in the Pearson and 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 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 concerns 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 first is defined by a weighted binary relation, using a ranking function. It can be related – in some sense – to the Spearman approach (1904, 1906). The latter form can be associated with the Kendall approach (1970), and needs the definition of a 4-ary relation on $\mathcal{O}$.

Formal notions, associated with the shape of a classification tree need to be introduced. The presented work is very concerned with combinatorial computing. This paper is devoted to the first part of this work. The second part will be developed in a second paper which is in preparation. At the end of this second part the most general case of comparing q-ary relations for any q is considered.

2. COMPARING NUMERICAL VARIABLES

For this most classical case, the Bravais-Pearson [Bravais (1846), Pearson(1920)] 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 this construction, 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. The variable 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 my approach, is less to be tested; but rather 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 $s(v, w)$ 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 me 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 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 $V^* = \{v^{*j} \mid, 1 \leq j \leq p\}$ 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, I adopt the probabilistic similarity index by means of the equation

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

$1 \leq j < k \leq p$, 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:

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

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

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

A classification tree on an object set $\mathcal{O}$ is a tree associated with an ordered sequence of partitions on $\mathcal{O}$. Let me denote by $(\pi_0, \pi_1, \ldots, \pi_l, \pi_{l+1}, \ldots, \pi_m)$ this sequence. π_{l+1} is deduced from π_l by class agregation, $0 \leq l \leq m-1$. π_0 is the partition for which each class contains a single element of $\mathcal{O}$ and π_m is the partition into one only class grouping all the set $\mathcal{O}$.

This tree is decomposed into $m+1$ levels. Each of them is associated with one of the previous partitions. Thus, the l^{th} level represents the partition $\pi_l, 0 \leq l \leq m$. And the number of nodes situated at the l_{th} level is the number of classes of the π_l partition, $0 \leq l \leq m$. Nodes representing classes of π_l to be agregated in π_{l+1} are arranged in an adjacent manner on the l^{th} level. They are joined by edges to the node of the $(l+1)^{th}$ level representing

the agregation class. In these conditions, the leaves of this tree represent the classes of the π_0 partition and the root, the single class of π_m.

Example:

$$O = \{o_1, o_2, o_3, o_4, o_5\}$$

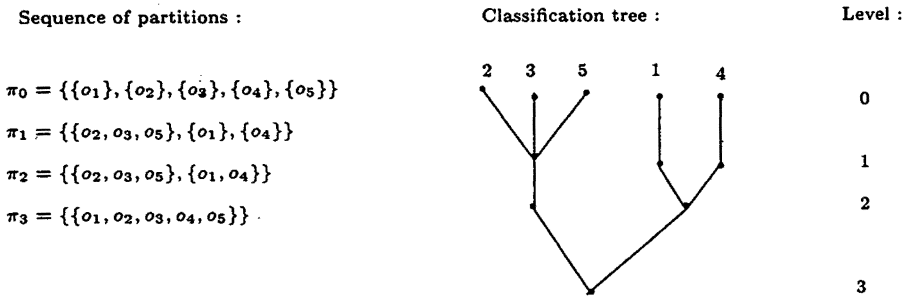

Figure 1.

Such a classification tree is often called a "stratified hierarchy" [Benzecri (1973)] or also a labelled ranked dendrogram [Murtagh (1984)]. It is usually obtained by a hierarchical classification algorithm on the set $\mathcal{O}$, provided by a dissimilarity measure between its subsets.

An "indexed hierarchy" is obtained by replacing the increasing sequence of the level numbers by an increasing sequence of numerical values. With each level is associated one value which generally corresponds to a dissimilarity measure between the joined classes at the concerned level. An indexed hierarchy determines a weighted classification tree. For this purpose one may provide each edge by a numerical value representing the positive difference between the values of the two levels limiting the extremities of the concerned edge.

We shall only be interested here in labelled and unweighted classification trees. However, generalizations can easily be considered for weighted classification trees by replacing the discrete relation associated with the tree on the object set $\mathcal{O}$, by a weighted one, More precisely, this discrete relation will be defined (see below) at the level of the set -denoted $\mathbb{P}$- of all unordered object pairs of $\mathcal{O}$; and, the general principle for representing and treating a weighted relation of the same type as the discrete one, consists of replacing counting elements of $\mathbb{P}$ by summing the weights of these elements.

Many methods are limited to the comparison of binary trees. The given justification argues that it is always possible to associate to a nonbinary tree,

a compatible binary one. Nevertheless, multiple agregation at a given level of a classification tree may occur very often in real cases (Lerman 1989, Jovicic 1996).This is especially likely 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 $\mathcal{O}$. In such cases, the number of binary trees compatible with a nonbinary one becomes considerably large. Let me 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 [Lerman (1970, 1981), Frank and Svensson (1981)]

$$\beta(c) = (c-1)\,!\,c\,!\,/\,2^{c-1}, \tag{12}$$

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

$$\left(\prod_{1\leq q\leq r} \beta(c_q)\right) \times \frac{d\,!}{d_1\,!\,d_2\,!\,\times\,\ldots\,\times\,d_r\,!}\,, \tag{13}$$

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

$$\mathbb{P} = \{\{x, y\} \mid x \in \mathcal{O}\,, y \in \mathcal{O}\,, x \neq y\}. \tag{14}$$

A faithful mathematical representation that we have adopted for a labelled tree is given by the notion of an "ultrametric preordonnance" [Lerman (1970)], which is defined by a specific total preorder on $\mathbb{P}$ (see below). This form of ordinal similarity has been very recently extended to some general types of trees [Guénoche (1998)].

Denoting by

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

the partition sequence associated with the 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 (Fig. 2):

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

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

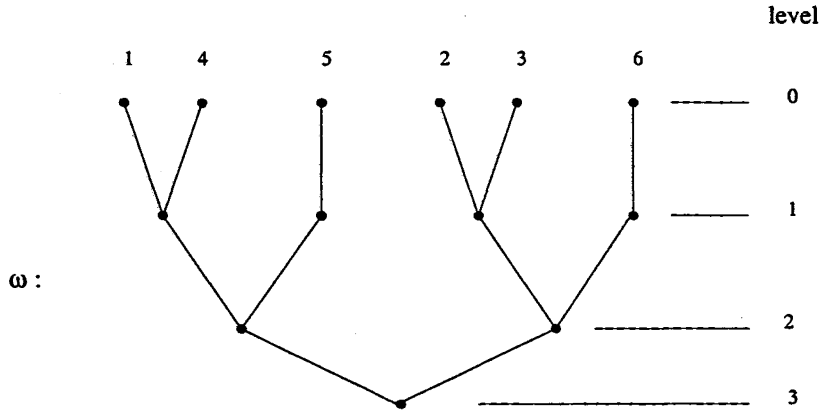

Figure 2.

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

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 (see below). As an example, consider the two trees α and β (Fig. 3).

For the associated level functions we have the following matrices (see Figs. 4 and 5).

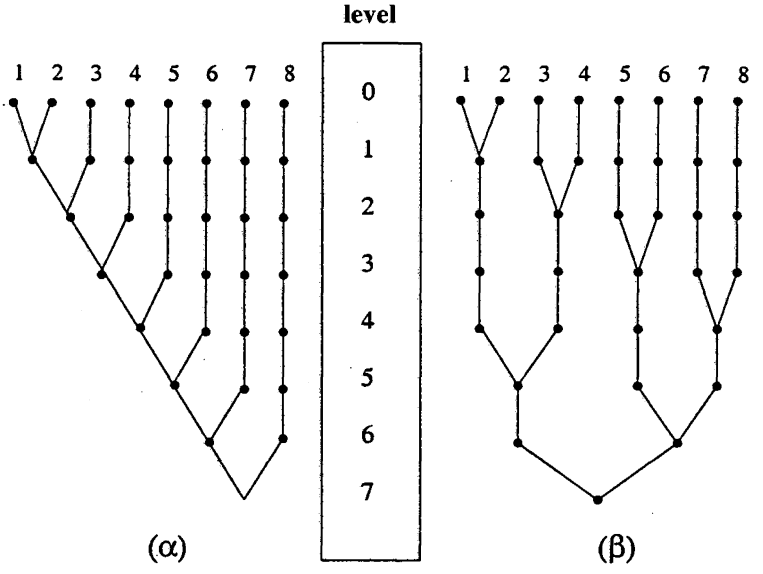

Figure 3.

1							
2	1						
3	2	2					
4	3	3	3				
5	4	4	4	4			
6	5	5	5	5	5		
7	6	6	6	6	6	6	
8	7	7	7	7	7	7	7
	1	2	3	4	5	6	7	8

Figure 4. – Matrix of l_α.

Let me now formally introduce the coding of an ultrametric preordonnance UP_ω associated with an ω tree [as in (15)], by means of the "mean rank" function. This defines a valuation of the set $\mathbb{P}$ of unordered object pairs:

$$\Lambda_\omega = \{\lambda_\omega(i,j) \mid 1 \leq i < j \leq n\}. \tag{16}$$

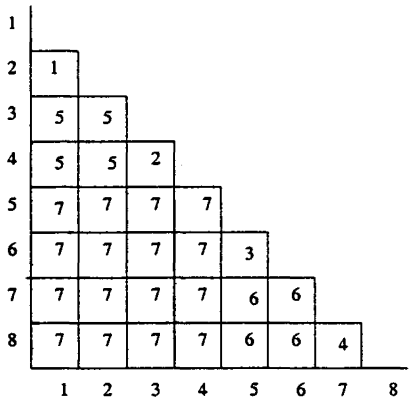

Figure 5. – Matrix of l_β.

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

$$\lambda_\omega(i,j) = card[R(\pi_{l-1}] + \frac{1}{2}\,[\text{card}(R(\pi_l) - R(\pi_{l-1}) + 1],\ 1 \leq l \leq m. \quad (17)$$

With this equation we obtain the matrices associated with λ_α and λ_β. For this purpose we need to establish the preordonnances $UP(\alpha)$ and $UP(\beta)$ according to the above example (see $UP(\omega)$ associated with the tree ω given in Fig. 2).

As an example, consider the calculation of $\lambda_\beta(5,8)$. By denoting π_l^β the partition of the l^{th} level of the β tree ($0 \leq l \leq 7$), we have

$$\begin{aligned}\lambda_\beta(5,8) &= card[R(\pi_5^\beta)] + \frac{1}{2}[card(R(\pi_6^\beta) - R(\pi_5^\beta)) + 1] \\ &= 8 + \frac{1}{2}[(12-8)+1] = 10.5.\end{aligned}$$

Notice that l_α and l_β give the same value for $\{1,8\}$:

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

but we have

$$\lambda_\alpha(1,8) = 25 > \lambda_\beta(1,8) = 20.5.$$

Also

$$l_\alpha(6,7) = l_\beta(6,7) = 6$$

1								
2	1							
3	2.5	2.5						
4	5	5	5					
5	8.5	8.5	8.5	8.5				
6	13	13	13	13	13			
7	18.5	18.5	18.5	18.5	18.5	18.5		
8	25	25	25	25	25	25	25	
	1	2	3	4	5	6	7	8

Figure 6. – Matrix of λ_α.

1								
2	1							
3	6.5	6.5						
4	6.5	6.5	2					
5	20.5	20.5	20.5	20.5				
6	20.5	20.5	20.5	20.5	3			
7	20.5	20.5	20.5	20.5	10.5	10.5		
8	20.5	20.5	20.5	20.5	10.5	10.5	4	
	1	2	3	4	5	6	7	8

Figure 7. – Matrix of λ_β.

and

$$\lambda_\alpha(6,7) = 18.5 > \lambda_\beta(6,7) = 10.5.$$

Even more

$$l_\alpha(1,5) = 4 < l_\beta(1,4) = 5$$

but

$$\lambda_\alpha(1,5) = 8.5 > \lambda_\beta(1,4) = 6.5.$$

Finally, take into account that the following normalization condition holds, whatever is the total preorder $UP(\omega)$; and then, whatever is the ω tree shape:

$$\Sigma\{\lambda_\omega(x,y) \mid \{x,y\} \in \mathbb{P}\} = p(p+1)/2 \tag{18}$$

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

Consider now the case of binary labelled and ranked trees. For a given ω such tree the corrected version of the Colless imbalance score I_ω (see in Mooers 1995), can be written as follows

$$I_\omega = \frac{2}{(n-1)(n-2)} \sum_{in \in \mathcal{J}_n(\omega)} |r_{in} - s_{in}| \tag{19}$$

where $\mathcal{J}_n(\omega)$ is the set of the internal nodes of ω (comprising the root), n is the number of tips, and $r_i n$ and $s_i n$ are the number of tips on the right and left sides of the bifurcation from in. For the above trees α and β (see Fig. 3) we have:

$$I_\alpha = 1 \text{ and } I_\beta = 0.$$

One may also associate with this mentionned ω binary tree the following score derived from the definition of the mean rank function λ_ω on $\mathbb{P}$:

$$\lambda(\omega) = \frac{\max\{\lambda_\omega(x,y) | \{x,y\} \in \mathbb{P}\}}{[(n^2 - 2n + 2)/2]}. \tag{20}$$

This maximum value is clearly reached for the element pairs joined at the last level of ω. The denominator expression is obtained from the most imbalanced tree (as above for α) for which at each level, one single object is aggregated to the formed class.

$$\lambda(\alpha) = 1 \text{ and } \lambda(\beta) = 0.82.$$

Therefore, the λ function on the binary labelled and ranked trees, reflects an imbalancy measure. And, it is envisageable to derive from this function an imbalance score which lies in the [0,1] interval and which is highly correlated with the Colless index.

Consider now a general (not necessarily binary) classification tree (labelled and ranked) ω. We are going now to give a very general and more explicit equation [than (17)] for the mean rank function λ, on the set of unordered object pairs denoted $\mathbb{P}$ (see above). For this purpose, we need to introduce

a notion of "indexed type of a classification tree". This concept captures entirely the tree shape. It corresponds to the sequence of the partition types, associated with the decreasing sequence of the level tree.

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

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

More generally, for the following tree δ (Fig. 9), the indexed type is

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

where, in the figure, we have indicated by $N_{i1\ i_2 \ldots i_k}$ the object class of which the cardinality is $n_{i1\ i_2 \ldots i_k}$.

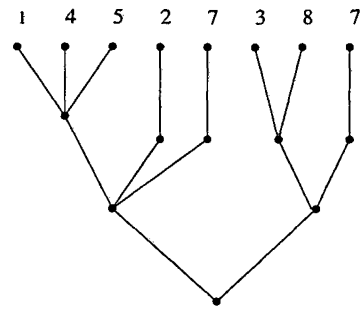

Figure 8. – Tree γ.

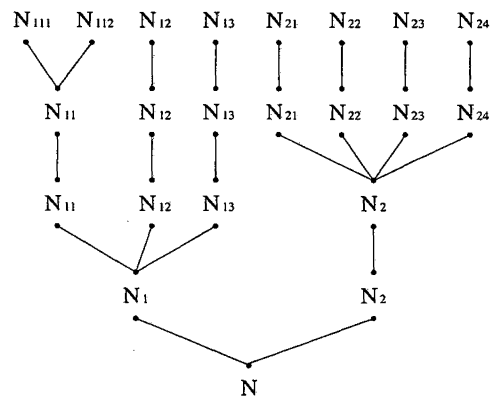

Figure 9. – Tree δ.

More precisely consider the following 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. This subscript can be written $i_1 i_2 \ldots i_{k-1} i_k$ and indicates that $N_{i1}\ i_2 \ldots i_{k-1} i_k$ is the i_k^{th} subclass –from the left to the right– of the class $N_{i1} i_2 \ldots i_{k-1}$ (for example, we have $N_{112} \subset N_{11} \subset N_1 \subset N$). In a given partition type, the subscripts are lexicographically ordered from the left to the right [see above $\tau(b)$].

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

$$p_k = \sum_{i_1 \ldots i_{k-1}} \sum_{\{i_k, i'_k\}} n_{i1} \ldots i_{k-1}, i_k\ n_{i1} \ldots i_{k-1} i'_k \tag{22}$$

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

$$\begin{aligned} &(\forall (i,j) \in \mathbb{P});\\ &l\omega(i,j) = k \Leftrightarrow\\ &\lambda_\omega(i,j) = p_1 + p_2 + \ldots + p_{k-1} + \frac{1}{2}(p_k + 1). \end{aligned} \tag{23}$$

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 $R(\omega)$ (see example below), let us, without restriction, designate by $\{1, 2, \ldots, i, \ldots, n\}$ the object set $\mathcal{O}$.

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

$$\mathbb{P} = \{(i,j) \mid 1 \leq i < j \leq n\}. \tag{24}$$

With these notations

$$\begin{aligned} R(\omega) = \{[(i,j)\ ,\ (i',j')] \,|\, &[(i,j)\ ,\ (i',j')]\\ &\in \mathbb{P} \times \mathbb{P}\ and\ l_\omega(i,j) < l_\omega(i',j')\}\,, \end{aligned} \tag{25}$$

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

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

$$\omega((i,j),(i',j')) = \begin{cases} 1 & if\ l_\omega(i,j) < l_\omega(i',j')\ , \\ 0 & if\ not\ , \end{cases} \tag{26}$$

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

Example: Consider the classification ω tree given in Figure 2 and represented by the total preorder $UP(\omega)$ (see above after (15)). In this example $n = 6$ and $card(\mathbb{P}) = n(n-1)/2 = 15$. The cardinality sequence of the preorder classes $UP(\omega)$ is (2,4,9), therefore

$$card[R(\omega)] = 2 \times 4 \ + \ 2 \times 9 \ + \ 4 \times 9 = 62.$$

	12	13	14	15	16	23	24	25	26	34	35	36	45	46	56
12	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
13	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
16	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
25	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
26	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
34	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
35	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
36	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
45	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
46	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0
56	0	0	1	1	0	1	0	0	1	0	0	1	1	0	0

Figure 10. – Matrix of the indicator function ω of $R(\omega)$.

More explicitely, by denoting $ij(i < j)$ the pair $\{i,j\}, 1 \leq i < j \leq 6$, we have

$$R(\omega) = \{(14,15),(14,26),(14,36),(14,45),(23,15),(23,26),(23,36),$$
$$(23,45),(14,12),...,(14,56),(23,12),...,(23,56),(15,12),...,(15,56),$$
$$(26,12),...,(26,56),(36,12),...,(36,56),(45,12),...,(45,56)\}.$$

The indicator function ω is given by the matrix represented in Figure 10 (see above).

In this figure, the first and the second components of the ordered pair $(ij,i'j'), 1 \leq i < j \leq 6$ and $1 \leq i' < j' \leq 6$, have respectively to be read horizontally and vertically.

On the other hand, by reordering the rows and the columns of the previous matrix according to the total preorder $UP(\omega)$, we obtain a decomposition of this matrix into rectangular blocks filled with the common value 1 or, exclusively, 0. The blocks situated below the diagonal blocks uniformely contain the value 1 and the others, the value 0.

This mathematical representation of a classification tree ω by $R(\omega)$ (see (25)) will be exploited in second paper (see Sect. 1) which is in preparation.

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

$$\left.\begin{aligned} \alpha &= (\pi_0, \pi_1, \ldots, \pi_l, \ldots, \pi_{n-2}, \pi_{n-1}) \\ \beta &= (\chi_0, \chi_1, \ldots, \chi_l, \ldots, \chi_{n-2}, \chi_{n-1}) \end{aligned}\right\} \tag{27}$$

a sequence of similarity indices

$$(B_l | 1 \leq l \leq n-2) \tag{28}$$

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 (14) 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:

$$[R(\pi_l), S(\pi_l)] \; and \; [R(\chi_l), S(\chi_l)] \, . \tag{29}$$

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

$$B_l = \frac{card[R(\pi_l) \cap R(\chi_l)]}{\sqrt{card[R(\pi_l)] \times card[R(\chi_l)]}} \, , \tag{30}$$

where $card$ designates the cardinality.

Notice that this coefficient has exactly the same structure as the one proposed by 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 (28).

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 (28) by means of a non arbitrary function f:

$$B(\alpha, \beta) = f(B_l | 1 \leq l \leq n-1) \, . \tag{31}$$

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 (26)) since an ultrametric preordonnance is a specific total preorder on $\mathbb{P}$ (see (15)). As mentionned above this data structure has been studied for general types of trees called the X-trees [Guénoche (1998)].

In order to clearly set up the nature of this comparison, let me 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:

$$\begin{aligned} &C_l = R(\pi_l) - R(\pi_{l-1}) \; , \; 1 \leq l \leq m \; , \\ &[resp. \; D_p = R(\chi_p) - R(\chi_{p-1}) \; , \; 1 \leq p \leq q] \; , \end{aligned} \tag{32}$$

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,

$$E_1 = \sum_{l<l'} C_l \times C_{l'} \left(resp.\ F_1 = \sum_{p<p'} D_p \times D_{p'} \right)$$

$$E_2 = \sum_{l} C_l^{[2]} \left(resp.\ F_2 = \sum_{p} D_p^{[2]} \right),$$

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

$$E_3 = \sum_{l>l'} C_l \times C_{l'} \left(resp.\ F_3 = \sum_{p>p'} D_p \times D_{p'} \right). \tag{33}$$

Clearly, we have

$$\mathbb{P} \times \mathbb{P} = E_1 + E_2 + E_3 = F_1 + F_2 + F_3. \tag{34}$$

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

$$\mathbb{P} \times \mathbb{P} = \sum_{1 \leq i \leq 3} \sum_{1 \leq j \leq 3} E_i \cap F_j. \tag{35}$$

By introducing the following cardinalities

$$\{s_{ij} = card\,(E_i \cap F_j) \mid 1 \leq i \leq 3,\ 1 \leq j \leq 3\}, \tag{36}$$

the following identities hold:

$$s_{11} = s_{33},\ s_{12} = s_{32},\ s_{13} = s_{31}\ and\ s_{21} = s_{23}. \tag{37}$$

And therefore, the Goodman and Kruskal cœefficient can be written

$$\gamma = \frac{s_{11} - s_{13}}{s_{11} + s_{13}}. \tag{38}$$

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

$$\eta = \frac{(s+t)-(u+v)}{(s+t)+(u+v)}\,, \tag{39}$$

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

$$Y = \frac{st-uv}{st+uv} \tag{40}$$

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 situation concerned here, where the total preorders are established on $\mathbb{P}$ (see (14)) 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 cophenetic correlation cœfficient between the two level functions l_α and l_β, respectively associated with binary trees α and β [see (27)] [Sokal and Rohlf (1962)]. Namely,

$$\gamma(\alpha,\beta) = \frac{\sum\{[l_\alpha(ij)-\bar{l}_\alpha][l_\beta(ij)-\bar{l}_\beta] \mid ij \in \mathbb{P}\}}{\sqrt{(\sum_{ij}[l_\alpha(ij)-\bar{l}_\alpha]^2)(\sum_{ij}[l_\beta(ij)-\bar{l}_\beta]^2)}} \tag{41}$$

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

$$\bar{l}_\omega = \frac{2}{n(n-1)} \sum\{l_\omega(ij) \mid ij \in \mathbb{P}\}\,, \tag{42}$$

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 (18)).

A correlation cœfficient like $\gamma(\alpha, \beta)$ (see (41)) is considered by Lapointe and Legendre (1990, 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 Sect. 4 below). For the latter and relative to an ω tree, the valuation of a pair $\{i, j\}$ is implicitly given by

$$\mu_\omega(i\,j) = \frac{l_\omega(i\,j) - \bar{l}_\omega}{\sqrt{\sum_{i'\,j'} [l_\omega(i', j') - \bar{l}_\omega]^2}} \tag{43}$$

$1 \leq i < j \leq n$. 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. A normal or some other specific distribution could also have been envisaged, in order to approximate the distribution of the correlation coefficient between trees (Daudé 1992).

4. PERMUTATIONAL APPROACH FOR COMPARING CLASSIFICATION TREES: THE FIRST COMPARISON METHOD

As said above (see Sect. 2), the general principle considered here is the same as the 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.

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 (see (16, 17) and (23) §3.1):

$$\wedge_\omega = \{\lambda_\omega\,(i, j) \mid 1 \leq i < j \leq n\}. \tag{44}$$

And recall that relative to (15), if ij belongs to $R(\pi_l) - R(\pi_{l-1})$, we have

$$\lambda_\omega(ij) = card[R(\pi_{l-1})] + \frac{1}{2} \times [card(R(\pi_l) - R(\pi_{l-1})) + 1] , \qquad (45)$$

$1 \leq l \leq m$.

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

$$\wedge_{\omega^*} = \{\lambda_{\omega^*}(i,j) = \lambda_\omega[\sigma(i), \sigma(j)] \mid 1 \leq i < j \leq n\} , \qquad (46)$$

where σ is a random element in the set G_n – provided by a uniform probability measure – of all permutations on $I = \{1, 2, \ldots, i, \ldots, n\} \ldots$ In the random tree model, the general structure of the classification tree (form and levels) is preserved. Consider the decomposition of the leave set into a sequence $(L_1, L_2, \ldots, L_i, \ldots, L_k)$ of disjoint subsets, such as the elements of L_i, for each $i = 1, \ldots, k$, are joined directly together at a given node. As an example, for the following tree on 8 objects.

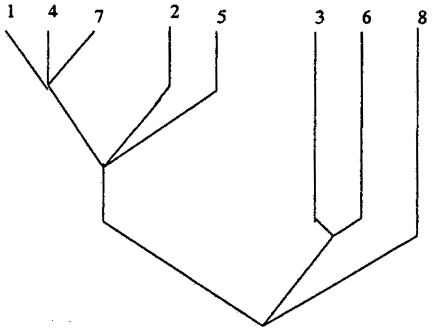

Figure 11. – $k = 4$ and $(L_1, L_2, L_3, L_4) = (\{1,4,7\}, \{2,5\}, \{3,6\}, \{8\})$.

Now, if we denote by $(n_1, n_2, \ldots, n_i, \ldots, n_k)$ the cardinality sequence of $(L_1, L_2, \ldots, L_i, \ldots, L_k)$, the uniformely probability space associated with the concerned classification ω tree, includes

$$\frac{n!}{n_1! n_2! \ldots n_i! \ldots n_k!}$$

elements. And then, by the permutational model, each element of the space tree is retrieved $n_1! n_2! \ldots n_i! \ldots n_k!$ times.

In the previous notations, note that we have exactly

$$\lambda_\omega[\sigma(i),\sigma(j)] = \lambda[\min(\sigma(i),\sigma(j)),\max(\sigma(i),\sigma(j))]\,, \tag{47}$$

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

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

$$s(\alpha,\beta) = \sum_{\mathbb{P}} \lambda_\alpha(i,j)\,\lambda_\beta(i,j) \tag{48}$$

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,

$$s(\alpha,\beta^*) = \sum_{\mathbb{P}} \lambda_\alpha(i,j)\,\lambda_\beta\,[\tau(i),\tau(j)]\,, \tag{49}$$

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 Sect. 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),

$$Q(\alpha,\beta) = \frac{s(\alpha,\beta) - E[s(\alpha^*,\beta^*)]}{\sqrt{var[s(\alpha^*,\beta^*)]}}\,, \tag{50}$$

the cœfficient

$$r(\alpha,\beta) = \frac{Q(\alpha,\beta)}{\sqrt{Q(\alpha,\alpha)\,Q(\beta,\beta)}}\,, \tag{51}$$

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

$$\rho(\alpha,\beta) = \frac{1}{\sqrt{n}}\, Q(\alpha,\beta) \tag{52}$$

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

$$\rho(\alpha,\beta) = \frac{s(\alpha,\beta) - p(p+1)^2}{\sqrt{V_\alpha V_\beta + \frac{1}{2n}\left[\left(\frac{p^2-1}{12} - 2V_\alpha\right)\left(\frac{p^2-1}{12} - 2V_\beta\right)\right]}} \tag{53}$$

where $p = n(n-1)/2$, $V_\omega = A_\omega - \left(\frac{p+1}{2}\right)^2$ and

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

with $\omega = \alpha$ *or* β.

The latter expression (53) is deduced from more general expressions Lerman (1987) in (1992), Ouali-Allah (1991).

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 (paper in preparation).

We conclude this section by showing in an implicit statistical manner the relevancy of the ultrametric preordonnance coding by means of the "mean rank" function λ instead of the level function l (see Sect. 3.1).

Consider two versions of the random standardized coefficient Q (see (50) for the general expression), respectively associated with the two functions λ and l, and that we denote by $Q_\lambda(\alpha,\beta*)$ and $Q_l(\alpha,\beta*)$. In $Q_\lambda(\alpha,\beta*)$, the random rough index $s_\lambda(\alpha,\beta*)$ is given by equation (49). To obtain $Q_l(\alpha,\beta*)$, we have to replace the λ valuation by the l valuation on the set $\mathbb{P}$ of unordered object pairs.

Simulations of the respective probability distributions of $Q_\lambda(\alpha,\beta*)$ and $Q_l(\alpha,\beta*)$ have been performed on the basis of 5000 independent random permutations. This has been done with the help of the computer program established by Rouxel in his algorithmic work (Rouxel 1997) following my theoretical research (Lerman 1997). Many experiments have been achieved to study the influence of the shapes of the compared trees. All of them lead us to the same general remark which follows from the observation of the two

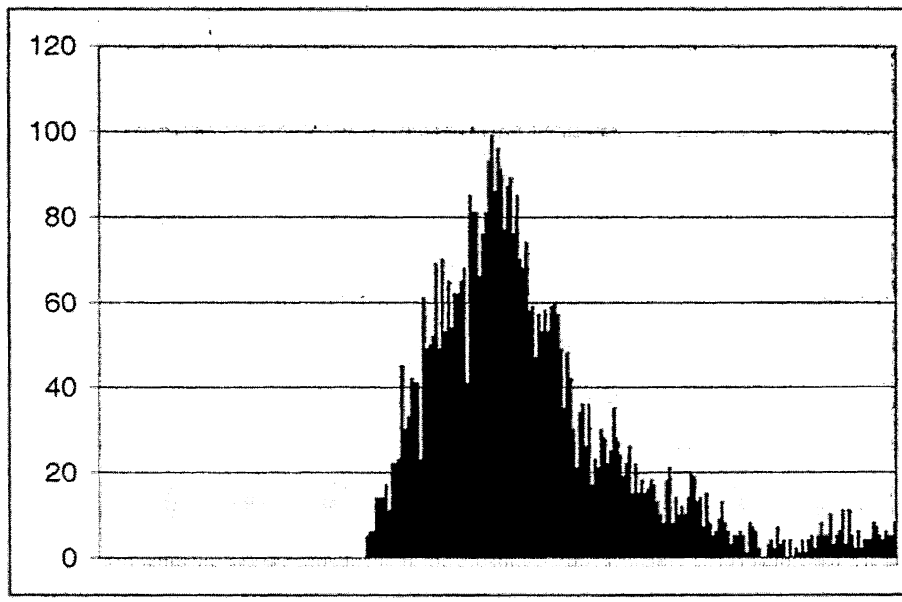

Figure 12. – Empirical distribution of $Q_\lambda(\alpha, \beta^*)$: 2 distinct tree shapes on 14 objets, 5 000 independent random permutations.

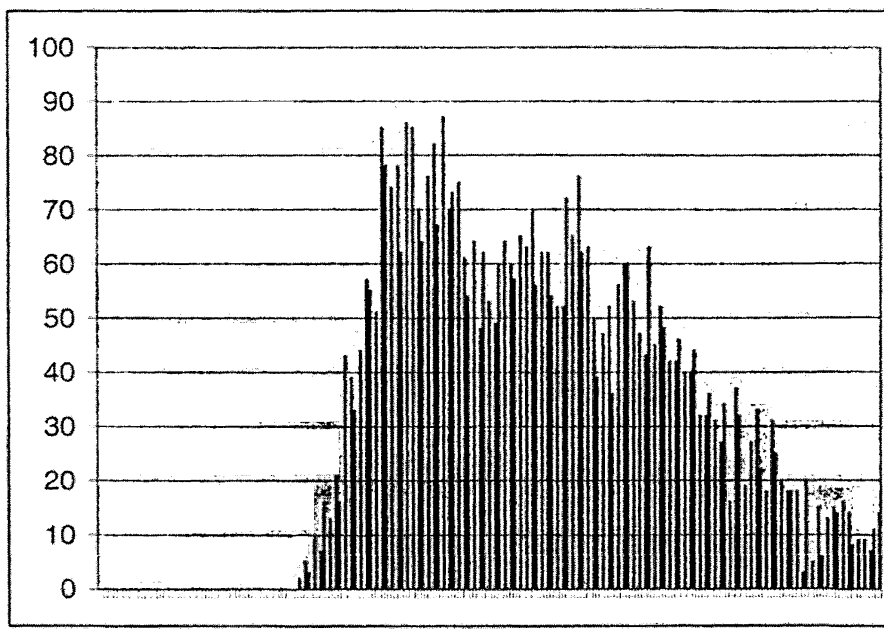

Figure 13. – Empirical distribution of $Q_l(\alpha, \beta^*)$: 2 distinct tree shapes on 14 objets, 5 000 independent random permutations.

following distributions (see Figs. 12 and 13). These distributions concern the comparison between the same two trees α and β having distinct structures (shapes) on 14 elements. These structures have been drawn at random. Figures 12 and 13 show the empirical distributions, based on independent simulations, of respectively $Q_\lambda(\alpha, \beta*)$ and $Q_l(\alpha, \beta*)$.

The comparison between Figures 12 and 13 shows clearly that the probability distribution of $Q_\lambda(\alpha, \beta*)$ is much more significatively concentrated than that of $Q_l(\alpha, \beta*)$. In other words, the distribution of the latter $[Q_l(\alpha, \beta*)]$ is somewhat degenerated with respect to that of $Q_\lambda(\alpha, \beta*)$.

The announced second paper which is in preparation (see Sect. 1) is devoted to the analysis of the same permutational approach for comparing classification trees in case where the second mathematical coding is adopted (see (25) and (26) in Sect. 3.1). In these conditions, the mathematical expectation and variance of the associated raw index cannot be computed analytically. However clear mathematical and combinatorial equations are needed in order to resolve exactly the computing problem. Specific and elaborated algorithmic research using recursivity procedure enables to resolve this problem in a very competitive time (Rouxel 1997). Therefore, we will be able to simulate the probability distribution of the standardized random coefficient that we can denote by $Q_4(\alpha, \beta*)$. This notation is adopted because the mathematical representation considered for a classification tree (labelled ranked dendrogram) is a 4-ary relation; or, more precisely, a binary relation (total preorder) on the set $\mathbb{P}$ of unordered object pairs (see (43) for the general expression of $Q(\alpha, \beta)$).

We have seen that the empirical probability distribution of $Q_4(\alpha*, \beta)$ is more harmonious than that of $Q_\lambda(\alpha*, \beta)$.

ACKNOWLEDGEMENT

I am very indebted to the anonymous reviewer for this rigorous study of the submitted paper and for his remarks and suggestions which enabled me to give a better presentation and more established justification of my work.

REFERENCES

P. ARABIE and L. J. HUBERT, Combinatorial data analysis, *Annual Review of Psychology*, 1992, *43*, p. 169-203.

F. B. BAKER, Stability of two hierarchical grouping techniques, *J. American Statistical Association*, 1974, *69*, p. 440-445.

J. P. BENZECRI, *L'Analyse des Données*, Tome 1 : *La Taxinomie*, Paris, Dunod, 1973.

A. BRAVAIS, *Analyse mathématique sur les probabilités des erreurs de situation d'un point*, Mémoires de l'Institut de France, 21846, p. 255-332.

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

F. DAUDÉ, *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.

E. B. FOWLKES and C. L. MALLOWS, A method for comparing two hierarchical clusterings, *J. American Statistical Association*, 1983, *78*, p. 553-584.

O. FRANK and K. SVENSSON, On probability distributions of single-linkage dendrograms, *J. Statist. Comput. Simulation*, 1981, *12*, p. 121-131.

L. A. GOODMAN and W. H. KRUSKAL, Measures of association for cross classification, *J. American Statistical Association*, 1954, *49*, p. 732-764.

A. GUÉNOCHE and B. MONJARDET, Méthodes ordinales et combinatoires en analyse des données, *Rev. Mathématiques et Sciences Humaines*, 1987, *25*, p. 5-47.

A. GUÉNOCHE, *Ordinal properties of tree distances* (personnal communication), *Discrete Mathematics*, 1998, *191* (in press).

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

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

L. J. HUBERT, *Inference procedures for the evaluation and comparison of proximity matrices*, Numerical Taxonomy, J. Felsenstein, Ed., NATO ASI Series, Springer Verlag, Berlin, 1983, p. 209-228.

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

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

M. G. KENDALL, *Rank Correlation Methods*, Charles Griffin, fourth edition, 1965.

F. J. LAPOINTE and P. LEGENDRE, Comparison tests for dendrograms: A comparative evaluation, *J. Classification*, 1995, *12*, p. 265-282.

F. J. LAPOINTE and P. LEGENDRE, A statistical framework to test the congruence of two nested classifications, *Systematic Zoology*, 1990, *39*, p. 1-13.

G. LECALVÉ, Un indice de similarité pour des variables de types quelconques, *Statist. Anal. Données*, 1976, *01-02*, p. 39-47.

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

I. C. LERMAN, Formal analysis of a general notion of proximity between variables, Congrès Européen des Statisticiens, Grenoble 1976, *Recent Developments in Statistics, North Holland*, 1977, p. 787-795.

I. C. LERMAN, Classification et Analyse Ordinale des Données, Paris, Dunod, 1981.

I. C. LERMAN, *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, 1984, Fasc. 3-4, p. 27-57.

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

I. C. LERMAN, Formules de réactualisation en cas d'agrégations multiples, *RAIRO Oper. Res.*, 1989, *23*, n°2, p. 151-163.

I. C. LERMAN, *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; 1992, I : *118*, p. 35-522, II : *119*, p. 75-100.

I. C. LERMAN, *Likelihood linkage analysis (LLA) classification method (Around an example treated by hand)*, Biochimie, Elsevier editions, 1993, *75*, p. 379-397.

I. C. LERMAN, *Comparing Classification tree Structures: a Special Case of Comparing q-Ary Relations*, Publication interne 1078 IRISA (April 1997) and Rapport de recherche 3167 INRIA (Mai 1997); 37 pages, 1997.

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

I. C. LERMAN and Ph. PETER, Structure maximale pour la somme des carrés d'une contingence aux marges fixées ; une solution algorithmique programmée, *RAIRO Oper. Res.*, 1988, *22*, p. 83-136.

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

F. MARCOTORCHINO and P. MICHAUD, Optimisation en Analyse Ordinale des Données, Paris, Masson, 1979.

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

H. MESSATFA, An algorithm to maximize the agreement between partitions, *J. Classification*, 1992, *9*, p. 5-15.

F. MURTAGH, Counting dendrograms: A survey, *Discrete Appl. Math.*, 1984, *7*, p. 191-199.

A. OCHIAI, Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Japanese Soc. Sci. Fisheries*, 1957, *22*, p. 526-530.

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

K. PEARSON, Notes on the history of correlation, *Biometrika*, 1920, *13*, p. 25-45.

S. REGNIER, Sur quelques aspects mathématiques des problèmes de la classification automatique, *Internat. Comput. Center Bull.*, 1965, *4*, p. 175-191.

F. ROUXEL, Comparaison d'arbres de classification, rapport de DEA, Informatique et Recherche Opérationnelle, Université Paris VI, 1997.

C. SPEARMAN, The proof and measurement of association between two things, *Amer. J. Psychology*, 1904, *15*, p. 88.

C. SPEARMAN, A footrule for measuring correlation, *British J. Psychology*, 1906, *2*, p. 89.

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

G. U. YULE, On the methods of measuring the association between two attributes, *J. Roy. Statist. Soc.*, 1912, *75*, p. 579-652.