REVUE DE STATISTIQUE APPLIQUÉE

ISRAËL-CÉSAR LERMAN
PHILIPPE PETER

Indice probabiliste de vraisemblance du lien entre objets quelconques. Analyse comparative entre deux approches

Revue de statistique appliquée, tome 51, n° 1 (2003), p. 5-35
<http://www.numdam.org/item?id=RSA_2003__51_1_5_0>

Rev. Statistique Appliquée, 2003, **LI** (1), 5-35

INDICE PROBABILISTE DE VRAISEMBLANCE DU LIEN ENTRE OBJETS QUELCONQUES ANALYSE COMPARATIVE ENTRE DEUX APPROCHES

Israël-César LERMAN[1], Philippe PETER[2]

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

[2] *École Polytechnique de l'Université de Nantes, Site de la Chantrerie, Rue Christian Pauc - B.P. 50609, 44306 Nantes Cédex 3,* ppeter@ireste.fr

RÉSUMÉ

La méthode AVL (Analyse de la Vraisemblance des Liens) est davantage connue pour la classification de l'ensemble $\mathcal{V}$ des variables descriptives que pour celle d'un ensemble $\mathcal{O}$ d'objets ou $\mathcal{C}$ de catégories décrits au moyen de $\mathcal{V}$. Cependant cette méthode permet avec la même rigueur conceptuelle d'élaborer une classification ascendante hiérarchique sur $\mathcal{O}$ (respectivement sur $\mathcal{C}$) et de fournir des coefficients d'«explication» compte tenu de l'organisation de $\mathcal{V}$. Le problème que nous reprenons ici est celui de la construction d'un indice de similarité probabiliste entre objets décrits par des variables de types quelconques. Plus précisément, il s'agit d'analyser une approche dûe à W.D. Goodall que nous ignorions lors de la conception de nos indices et de la situer par rapport à notre méthode d'élaboration. Cette dernière s'avère essentiellement distincte : plus souple, très générale et tenant compte de la sémantique des variables. Des résultats expérimentaux étayeront la comparaison théorique. Enfin, nous proposerons un logiciel HETAVL qui permet l'établissement d'une matrice d'indices probabilistes entre objets décrits par un ensemble de variables de types hétérogènes et quelconques.

Mots-clés : *Similarité probabiliste, Classification, Données hétérogènes*

ABSTRACT

The LLA (Likelihood Linkage Analysis) (AVL (Analyse de la Vraisemblance des Liens)) classification method is more known to adress the classification problem of a set $\mathcal{V}$ of descriptive variables than that of a set $\mathcal{O}$ of elementary objects or a set $\mathcal{C}$ of categories described by $\mathcal{V}$. In fact, this approach enables to elaborate an ascendant hierarchical classification on the set $\mathcal{O}$, respectively $\mathcal{C}$, with the same conceptual rigour. On the other hand, it provides «explanation» coefficients associating both classifications on descriptive variables and on described object set (respectively, category set). The problem that we revisit here concerns the construction of a probabilistic similarity index between objects described by heteregenous variables. More precisely, we compare the approach proposed by W.D. Goodall with that we consider. Our approach is essentially different. Its flexibility, its extreme generality and its ability to integrate

the variable structure are clearly showed. Experimental results will assess the theoretical comparison. Finally, we propose a software HETAVL which enables the establishment of a table of probabilistic similarity indices between objects described by a set of heteregenous variables.

Keywords : *Probabilistic similarity, Classification, Heteregenous data*

1. Introduction

La donnée est définie par un ensemble d'objets $\mathcal{O}$ décrit par un ensemble $\mathcal{V}$ de variables (on dit encore attributs ou caractères,...) où chaque variable prend sur chaque objet *une* valeur. Le problème général, considéré ici, est celui de l'élaboration d'un indice de similarité entre objets. C'est l'un des problèmes les plus fondamentaux en classification et analyse des données. La conception d'un indice de similarité entre objets dépend de différents facteurs qui sont liés. Citons :

(i) La nature des variables; c'est-à-dire, les structures des échelles descriptives sous jacentes (quantitative-numérique, qualitative-catégorielle où l'ensemble des catégories est ou non muni d'une relation d'un type donné)

(ii) La représentation mathématique des variables compte tenu de leur nature. Il y a à cet égard deux types de représentation : la représentation géométrique qui est particulièrement adaptée au cas où on peut «naturellement» se ramener à des variables numériques et la représentation ensembliste et relationnelle qui est la plus générale et permet de représenter fidèlement les attributs qualitatifs.

(iii) La méthodologie concernée de classification ou d'analyse des données.

Pour ce qui concerne le troisième point et relativement à deux méthodologies A et B, on peut chercher à «transporter» dans B, un indice de similarité conçu au coeur de A. Ce type de recherche est toujours riche d'enseignement. Cependant il faut noter que ce «transport» est d'autant plus aisé que l'algorithmique que suppose B est simple dans son principe. Mais alors, si tel est le cas, une conception au sein de B peut s'avérer plus libre et donc plus riche. L'algorithmique par laquelle nous sommes concernés est celle de la Classification Ascendante Hiérarchique (CAH). Elle est dans son principe de base particulièrement simple. Cette algorithmique sera «habillée» avec la méthode AVL (Analyse de la Vraisemblance du Lien) qui permettra d'intégrer et de comparer les indices que nous voulons étudier ici.

En effet, le premier objet essentiel de cet article est la comparaison de deux indices de similarité probabiliste entre objets décrits par des variables. Le premier est dû à W.D. Goodall [5] et le second, conçu tout à fait indépendamment, correspond à celui proposé dans [15] et [19]. La structure des données où le premier type d'indice peut le mieux se justifier et se comprendre est celui où les attributs sont catégoriels, qualitatifs nominaux. Il s'agit d'ailleurs du cas où la comparaison méthodologique peut la mieux être mise en évidence. Signalons que cette structure de la donnée est également considérée dans [2] où – comme c'est le cas pour notre approche [15-17] – c'est la représentation relationnelle des variables descriptives qui est adoptée et c'est également la décomposition additive de la similarité globale par rapport aux similarités unitaires, variable par variable qui est prise en compte. Toutefois, les indices proposés

dans [2] appelés de « Similarité probabiliste » ne sont pas des probabilités. Ils restent liés à la manière de calculer une partition au moyen d'un programme linéaire sous contrainte en $\{0, 1\}$.

Lorsque le nombre de variables augmente et lorsque les structures des échelles descriptives deviennent de plus en plus riches, l'indice de Goodall, surtout exprimé dans toute sa rigueur, montre vite ses limites en termes de complexité et d'interprétation. Notre approche s'avère plus souple, très générale en permettant de tenir compte finement de la sémantique des variables.

Au paragraphe II nous rappellerons le principe général de l'AVL. Au paragraphe III nous montrerons comment AVL traite la classification des objets dans le cas de référence où les variables sont toutes numériques. C'est au paragraphe IV que le cas qualitatif est étudié. Pour bien cerner le problème on commence par considérer le cas d'une seule variable avant d'envisager le cas général de plusieurs variables. D'autre part, nous commencerons par expliciter l'approche de Goodall avant de rappeler la nôtre. Enfin nous préciserons des aspects applicatifs liés à l'AVL. Un exemple illustratif de traitement de données, par chacun des deux indices, via la méthode AVL, est présenté au paragraphe V. La prise en compte des autres types de variables, en même temps qu'un programme de calcul d'indices de similarité probabiliste entre objets décrits par un mélange de variables de types quelconques (programme HETAVL), sont exprimés au paragraphe VI. Nous y évoquerons également la dualité entre la classification des objets et celle des variables.

2. Principe de l'AVL

La méthode AVL (Analyse de la Vraisemblance des Liens) constitue une approche très générale dans la représentation des données ou connaissances et l'évaluation numérique des ressemblances mutuelles entre variables descriptives et entre objets, respectivement catégories (concepts), décrits. L'algorithmique qui a le plus supporté cette approche est celle de la CAH (Classification Ascendante Hiérarchique). Mais d'autres algorithmiques telles que celle de la classification non hiérarchique [27] ou celle du positionnement multidimensionnel (« multidimensional scaling ») [4, 24, 25] ont pu être expérimentées avec beaucoup d'intérêt.

Deux caractéristiques fondamentales distinguent l'AVL. D'une part, les variables (attributs) de description sont interprétées en termes de relations sur l'ensemble $\mathcal{O}$ des objets et d'autre part, les indices de similarité et les critères d'association se réfèrent à une échelle de probabilité. Ils reflètent les degrés d'invraisemblance (ou d'étonnement) de la grandeur d'indices formellement « bien » conçus et pouvant avoir une nature combinatoire ou métrique.

La méthode AVL est davantage connue pour la classification de l'ensemble $\mathcal{V}$ des variables descriptives (Daudé [3], Lerman [10, 9, 11, 17, 18], Nicolaü et Bacelar-Nicolaü [1, 26, 28], Ouali-Allah [29]) que pour la classification de l'ensemble $\mathcal{O}$ des objets (Lerman [15], Lerman et Peter [19, 20]). Cependant, cette méthode permet avec la même rigueur conceptuelle d'élaborer une classification ascendante hiérarchique sur l'ensemble $\mathcal{O}$ des objets (*cf.* les dernières références ci-dessus). S'agissant ici d'un développement de ce dernier aspect, nous allons illustrer le principe de l'AVL

dans un cas simple et intuitif où il y a une seule variable numérique et où donc il s'agit d'un nuage de points rectiligne et donc, unidimensionnel.

Il importe d'avoir pleinement conscience que l'exigence première d'une CAH est de nature *ordinale*. Partant de la partition discrète où chaque classe comprend un seul élément de $\mathcal{O}$ il s'agit d'établir un ordre des agrégations qui fait apparaître des classes et des sous classes composantes qui sont cohérentes et qui définissent différents niveaux de généralisation. Pour établir un ordre qu'on souhaiterait le plus judicieux, des agrégations, un *critère numérique est nécessaire* pour sélectionner à chaque pas la «meilleure» agrégation ou, plus généralement, les également «meilleures» agrégations. Ce critère est défini à partir d'une notion de dissimilarité entre parties disjointes de l'ensemble à classifier. Il y a à cet égard différents types d'indices dépendant de la représentation mathématique des unités de données. Tous ces indices dépendent, chacun d'une certaine manière de la définition d'un indice de dissimilarité entre éléments de l'ensemble à classifier. Considérons dans le cadre de notre illustration, l'indice ici naturel, défini par la distance euclidienne ordinaire et qu'on notera d. Considérons d'autre part, entre deux classes (ou parties disjointes) $Cl1$ et $Cl2$ un des indices de dissimilarité les plus classiques appelé le «lien simple» («single linkage») que nous notons $\delta(Cl1, Cl2)$:

$$\delta(Cl1, Cl2) = min\{d(x, y)|(x, y) \in Cl1 \times Cl2\} \tag{1}$$

Considérons alors la situation illustrative annoncée, matérialisée par la figure suivante :

FIGURE 1

Relativement à la problématique de la CAH, la question posée est le choix de la première à retenir, parmi les deux fusions :

$C \leftarrow C1 \vee C2$ ou $D \leftarrow D1 \vee D2$.

Il y a à cet égard lieu de tenir compte de

(i)

$\delta(C1, C2) = min\{d(x, y)|(x, y) \in C1 \times C2\} >$
$\delta(D1, D2) = min\{d(x, y)|(x, y) \in D1 \times D2\}$;

(ii)

les densités des points dans $D1$ et $D2$ sont notablement supérieures à celles dans $C1$ et $C2$.

Si on adopte le critère du «lien simple» dit encore de la «distance minimale» on sélectionnera d'abord la fusion $D \longleftarrow D1 \vee D2$. Au contraire, le critère de

l'AVL, ici de la vraisemblance du lien maximal, choisira d'abord la fusion entre $C1$ et $C2$. En effet, il jugera par rapport au degré d'invraisemblance ou d'exceptionnalité de la petitesse de $\delta(C1, C2)$ relativement à celle de $\delta(D1, D2)$, eu égard à la comparaison entre le produit des densités des points dans $(C1, C2)$ d'une part, et dans $(D1, D2)$ d'autre part. Plus formellement, on associera dans le cadre d'un modèle aléatoire d'indépendance respectant les densités des points dans les classes, au couple $(C1, C2)$, respectivement $(D1, D2)$, un couple $(C1^*, C2^*)$, respectivement $(D1^*, D2^*)$. La fusion de $C1$ et $C2$ précèdera celle de $D1$ et $D2$ si $\delta(C1, C2)$ est plus invraisemblablement petit; c'est à dire si

$$Pr\{\delta(C1*, C2*) \leqslant \delta(C1, C2)\} < Pr\{\delta(D1*, D2*) \leqslant \delta(D1, D2)\}. \quad (2)$$

Nous pouvons sans difficulté majeure préciser un tel modèle et aller jusqu'au calcul effectif. Cependant, cela nous éloignerait trop de notre propos. Ce qu'il est important de retenir c'est que notre approche rejoint la philosophie de la théorie de l'information; mais relativement à l'évaluation des liaisons observées.

L'analyse de l'exemple ci-dessus permet de distinguer les étapes suivantes :

(i) Définition d'un indice de dissimilarité entre objets décrits tenant compte de la sémantique de l'échelle descriptive qui est la droite réelle.

(ii) Définition d'un indice brut de dissimilarité entre classes correspondant à la distance minimale (« single linkage »).

(iii) Introduction d'un modèle probabiliste d'absence de liaison, tenant compte des tailles des classes.

(iv) Comparaison des degrés d'invraisemblance des petitesses de $\delta(C_1, C_2)$ et $\delta(D_1, D_2)$:

$Comp\{Prob[\delta(C_1^*, C_2^*) \leqslant \delta(C_1, C_2)], Prob[\delta(D_1^*, D_2^*) \leqslant \delta(D_1, D_2)]\}$

(v) Adoption de la règle de la fusion de la paire de classes pour laquelle la probabilité est la plus petite.

Dans la situation ci-dessus – qui peut d'ailleurs être largement généralisée – l'approche « vraisemblance du lien » intervient après l'étape **(ii)**. Nous la faisons en fait intervenir dès l'étape **(i)** en définissant un indice probabiliste de la vraisemblance du lien; ou, ce qui est équivalent, un indice de « dissimilarité informationnelle » qui consiste à prendre la fonction $-Log_2$ de l'indice probabiliste.

Avant d'aborder le cas qui nous intéresse ici où les attributs sont catégoriels, il est instructif de se rendre compte comment AVL traite le cas le plus classique et de référence où les variables sont numériques.

3. AVL dans le cas de la classification d'objets décrits par des variables numériques

La description de l'ensemble $\mathcal{O}$=$\{o_i | 1 \leqslant i \leqslant n\}$ des objets par l'ensemble $\mathcal{V}$=$\{v^j | 1 \leqslant j \leqslant p\}$ des variables descriptives numériques est matérialisée par le

tableau suivant T des données. $x_i^j = v^j(o_i)$ est la mesure de la j-ème variable sur le i-ème objet, $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p$.

$$T = \{x_i^j = v^j(o_i) | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\} \tag{3}$$

Dans la représentation relationnelle des attributs de description, la variable numérique est interprétée comme une relation unaire valuée. Tout en gardant cette interprétation, on peut ici l'enrichir par la représentation «naturelle» de ce type de variable par une forme linéaire coordonnée de l'espace géométrique R^p où l'ensemble des objets est représenté par un nuage de points. En effet, notre perception de l'espace géométrique nous permet dans ce cas de mieux asseoir notre intuition dans la construction d'un indice de similarité entre objets. On suppose et c'est important pour la suite de la démarche que les domaines de valeurs des différentes variables sont homogènes et de même ordre de grandeur.

Dans ces conditions, AVL étant une méthode centrée sur la notion de similarité des formes, le *point de départ* que nous adoptons est l'indice cosinus. Plus précisément, en substituant au tableau T, le tableau

$$\Theta = \{\xi_i^j = \frac{x_i^j}{\sqrt{\sum_{1 \leqslant k \leqslant p} (x_i^k)^2}} \ | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}, \tag{4}$$

on définit alors l'indice de similarité

$$s(i, i') = Cos(o_i, o_{i'}) = \sum_{1 \leqslant j \leqslant p} \xi_i^j \xi_{i'}^j \tag{5}$$

On considère alors les étapes suivantes :

– *Décomposition additive de* $s(i, i')$ *:*

$$s(i, i') = \sum_{1 \leqslant j \leqslant p} s_j(i, i') \quad \text{où} \tag{6}$$

$$s_j(i, i') = \frac{1}{p} - \frac{1}{2}(\xi_i^j - \xi_{i'}^j)^2 \tag{7}$$

Avec une telle définition on a les propriétés suivantes :

(a) $s(i, i') = \sum\{s_j(i, i') | 1 \leqslant j \leqslant p\}$

(b) $s_j(i, i')$ est maximal et égal à 1/p si o_i et $o_{i'}$ sont homothétiques par rapport à l'origine O.

Introduisons à présent l'ensemble $I = \{1, 2, \ldots, i, \ldots, n\}$ qui indexe l'ensemble $\mathcal{O}$ des objets.

– Normalisation statistique de $s_j(i,i')/I \times I$

$$S_j(i,i') = \frac{s_j(i,i') - \mu(s_j)}{\sigma(s_j)} \tag{8}$$

où

$$\mu(s_j) = \frac{1}{n^2} \sum \{s_j(l,l') \quad |(l,l') \in I \times I\}, \tag{9}$$

$$\sigma^2(s_j) = \frac{1}{n^2} \sum \{[s_j(l,l') - \mu(s_j]^2) \quad |(l,l') \in I \times I\} \tag{10}$$

– Indice S *: somme des contributions normalisées.*

$$S(i,i') = \sum_{1 \leqslant j \leqslant p} S_j(i,i') \tag{11}$$

Nous considérons à présent l'ensemble $P_2(I)$ des paires ou parties à deux éléments de I. La taille de cet ensemble est $n(n-1)/2$.

– Normalisation statistique de S *au niveau de l'ensemble* $P_2(I)$ *des paires d'éléments de* I

$$Q(i,i') = \frac{S(i,i') - moy(S)}{\sqrt{var(S)}} \tag{12}$$

où

$$moy(S) = \frac{2}{n(n-1)} \sum \{S(l,l') | \{l,l'\} \in P_2(I)\} \tag{13}$$

$$var(S) = \frac{2}{n(n-1)} \sum \{[S(l,l') - moy(S)]^2 | \{l,l'\} \in P_2(I)\} \tag{14}$$

– Indice probabiliste de la vraisemblance du lien

$$P(i,i') = \phi[Q(i,i')] \tag{15}$$

où ϕ est la fonction de répartition de la loi normale centrée réduite $N(0,1)$.

– Indice de dissimilarité informationnelle

$$D(i,i') = -Log_2(P(i,i')), \{i,i'\} \in P_2(I). \tag{16}$$

Ce sont des raisons de cohérence logique qui nous conduisent à normaliser $s_j(i,i')$ au niveau de l'ensemble $I \times I$; alors que S est normalisé au niveau de

l'ensemble $P_2(I)$ des paires. On pourra en particulier remarquer qu'en prenant la forme suivante pour $s(i,i')$:

$$s_j(i,i') = (x_i^j - \bar{x}^j)(x_{i'}^j - \bar{x}^j) \tag{17}$$

où $\bar{x}^j$ est la moyenne de la j-ème variable v^j ; alors, la somme S des contributions normalisées n'est autre que le produit scalaire entre les deux vecteurs représentant o_i et $o_{i'}$ dans l'espace de l'Analyse en Composantes Principales normée.

Certes, d'autres choix de la contribution brute de la j-ème variable v^j à la comparaison des deux objets o_i et $o_{i'}$ peuvent être envisagés. On peut par exemple considérer :

$$S_j(i,i') = d_j^2 - (x_i^j - x_{i'}^j)^2 \tag{18}$$

où

$$d_j^2 = max\{(x_i^j - x_{i'}^j)^2 | (i,i') \in I \times I\}, 1 \leqslant j \leqslant p. \tag{19}$$

Un dernier point. La référence à loi normale centrée réduite pour juger du degré d'invraisemblance de la grandeur de S, et donc de Q se justifie d'autant mieux dans le cadre du théorème central limite que le nombre p de variables n'est pas trop petit.

On remarquera que le modèle probabiliste d'absence de liaison n'utilise des distributions empiriques observées des indices de similarité locaux $s_j (1 \leqslant j \leqslant p)$ et global S (*cf.* ci-dessus), que des caractéristiques globales de tendance centrale et de dispersion.

4. AVL dans le cas de la classification d'objets décrits par des variables qualitatives (catégorielles) nominales

4.1. Cas d'une seule variable c^1

4.1.1. Introduction

c^1 est une application de l'ensemble $\mathcal{O}$ des objets dans un ensemble C^1 de $k1$ catégories (on dit encore modalités), associant à chaque objet o, la catégorie qu'il possède :

$c^1 : \mathcal{O} \longrightarrow C^1 = \{c_1^1, \ldots, c_j^1, \ldots, c_{k1}^1\}$

$o \longrightarrow c^1(o) \in C^1$

Nous désignons par

$$T1 = \{\gamma_i^1 = c^1(o_i) \in C^1 | 1 \leqslant i \leqslant n\} \tag{20}$$

le tableau des données qui se réduit à un vecteur colonne.

Si o_i et $o_{i'}$ sont deux objets de l'ensemble $\mathcal{O}$ et si γ_i^1 et $\gamma_{i'}^1$ sont les catégories respectivement possédées par o_i et $o_{i'}$:

$$\gamma_i^1 = c^1(o_i) \quad \text{et} \quad \gamma_{i'}^1 = c^1(o_{i'}), \tag{21}$$

comparer les objets o_i et $o_{i'}$ revient à comparer les catégories γ_i^1 et $\gamma_{i'}^1$. On comprend aisément qu'il s'agit d'une situation essentiellement distincte de celle où on aurait une variable quantitative numérique v^1. On aurait en effet dans ce cas à comparer deux nombres réels $x_i^1 = v^1(o_i)$ et $x_{i'}^1 = v^1(o_{i'})$; ce que permet la sémantique du corps des nombres réels.

Il y a donc pour notre problème la nécessité absolue d'une table d'indices *numériques* d'association sur le produit cartésien $C^1 \times C^1$, définissant ainsi une valuation que nous noterons λ sur $C^1 \times C^1$.

On considère en général la valuation suivante :

$$\gamma(c_h^1, c_j^1) = \begin{cases} 1 & \text{pour} \quad 1 \leqslant h = j \leqslant k1; \\ 0 & \text{pour} \quad 1 \leqslant h \neq j \leqslant k1. \end{cases}$$

En d'autres termes $\gamma(c_h^1, c_j^1)$ n'est autre que le symbole δ_{hj} de Kronecker.

Cependant, on peut également considérer tout couple (α, β) de nombres réels pour lesquels $\alpha > \beta$ et considérer le tableau suivant définissant la valuation γ associée, où -à titre d'illustration- $k1$ a été pris égal à 6. Compte tenu de la symétrie, seule la moitié au dessus de la diagonale de ce tableau est spécifiée.

	c_1^1	c_2^1	c_3^1	c_4^1	c_5^1	c_6^1
c_1^1	α	β	β	β	β	β
c_2^1		α	β	β	β	β
c_3^1			α	β	β	β
c_4^1				α	β	β
c_5^1					α	β
c_6^1						α

valuation $\lambda\,(\alpha,\beta)$

FIGURE 2

Introduisons ici la partition π^1 induite par l'attribut c^1, qu'on notera comme suit :

$$\pi^1 = \{\mathcal{O}_1^1, \ldots, \mathcal{O}_j^1, \ldots, \mathcal{O}_{k1}^1\}$$

où $\mathcal{O}_j^1 = (c^1)^{-1}(c_j^1), 1 \leqslant j \leqslant k1.$

est l'ensemble des objets possédant la modalité c_j^1.

L'ensemble des paires d'objets réunis par π^1, est défini par :

$$R(\pi^1) = \sum\{P_2(\mathcal{O}_j^1 | 1 \leqslant j \leqslant k1\} \quad \text{(somme ensembliste)} \tag{22}$$

où $P_2(\mathcal{O}_j^1)$ est l'ensemble des parties à deux éléments de $\mathcal{O}_j^1$.

L'ensemble des paires d'objets séparés par π^1 :

$$S(\pi^1) = \sum\{\mathcal{O}_h^1 * \mathcal{O}_j^1 | 1 \leqslant h < j \leqslant k1\} \tag{23}$$

où $\mathcal{O}_h^1 * \mathcal{O}_j^1$ est l'ensemble des parties à deux éléments dont l'une des composantes appartient à $\mathcal{O}_h^1$ et l'autre à $\mathcal{O}_j^1$.

En désignant par n_j la cardinalité de $\mathcal{O}_j^1, 1 \leqslant j \leqslant k_1$, on a :

$$r^1 = |R(\pi^1)| = \sum_{1 \leqslant j \leqslant k1} n_j(n_j - 1)/2, \tag{24}$$

$$s^1 = |S(\pi^1)| = \sum_{1 \leqslant h < j \leqslant k1} n_h \times n_j \tag{25}$$

4.1.2. L'indice à prendre en compte dans le cas de l'AVL

Ici, où on n'a qu'une seule variable, on se base sur la distribution empirique de l'indice brut, défini à partir de la valuation γ, sur l'ensemble $P_2(\mathcal{O})$ des paires d'objets. On considère le modèle aléatoire consistant à choisir une paire $\{o_{i*}, o_{i'*}\}$ uniformément au hasard dans $P_2(\mathcal{O})$. Si o_i et $o_{i'}$ sont deux objets donnés, le degré d'invraisemblance de la force de la liaison est défini par la petitesse de :

$$\mathcal{T}(i, i') = Pr\{s(i*, i'*) \geqslant s(i, i')\} \tag{26}$$

$$\text{où } s(i, i') = \lambda[c^1(o_i), c^1(o_{i'})] \text{et} \tag{27}$$

$$s(i*, i'*) = \lambda[c^1(o_{i*}), c^1(o_{i'*})] \tag{28}$$

L'indice de vraisemblance du lien est alors défini par

$$P(i, i') = \Pr\{s(i*, i'*) < s(i, i')\} \tag{29}$$

$$= 0 \text{ si } s(i, i') = \beta \tag{30}$$

$$= \frac{s^1}{r^1 + s^1} \text{ si } s(i, i') = \alpha \tag{31}$$

Maintenant, si on considère la normalisation statistique de l'indice brut pour une variable, par rapport à $I \times I$, telle qu'envisagée dans le cas numérique (*cf.* § III), on obtient le résultat suivant :

Propriété. — Quel que soit le système de valeurs α et β ($\beta < \alpha$), la normalisation statistique (centrage et réduction) de l'indice par rapport à $I \times I$, conduit aux valeurs suivantes :

$$\beta_r = -(p^1/q^1) \tag{32}$$

$$\alpha_r = (q^1/p^1)^{1/2} \tag{33}$$

$$\text{où } p^1 = \sum_j n_j^2/n^2 \quad (\text{resp. } q^1 = \sum_{(h,j)} n_h n_j/n^2) \tag{34}$$

est la proportion de couples d'objets réunis (resp. séparés) par π^1.

Ce résultat ne devient intéressant que dans le cas de plusieurs variables, dans l'optique «somme des contributions normalisées».

4.1.3. L'indice de W.D. Goodall

L'indice de W.D. Goodall [5] correspond bien à une valuation particulière sur $C^1 \times C^1$ qui – comme c'est le cas de l'AVL ci-dessus – a une nature probabiliste et se trouve fondée sur la distribution empirique de la variable. Nous noterons λ_G cette valuation. Elle se présente, dans le cas illustratif où $k1 = 6$, sous la forme suivante, dont les cases vides peuvent être complétées par symétrie avec des zéros :

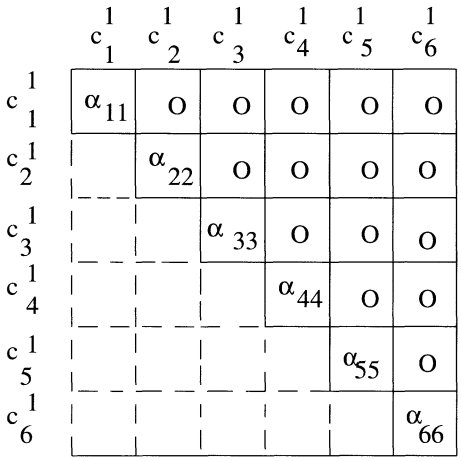


FIGURE 3

Par différence avec le tableau de la figure 2, on se rend compte que l'association entre une catégorie et elle même n'est pas constante. On la pose d'autant plus forte que la fréquence relative d'observation de la catégorie est petite. Très précisément

$$\alpha_{jj} = \frac{1}{p_j} \text{ où} \tag{35}$$

$$p_j = \frac{n_j(n_j - 1)}{n(n-1)} \tag{36}$$

est la proportion de couples d'objets distincts possédant la j-ème modalité, $1 \leqslant j \leqslant k1$.

Ce choix qui conduit à considérer que deux objets possédant une même catégorie se ressemblent d'autant plus que la catégorie a été rarement observée au niveau de l'échantillon, n'est pas sans prêter à discussion. On peut en effet se poser la question de savoir dans quelle mesure la STATISTIQUE informationnelle rejoint la SÉMANTIQUE résultant de la connaissance de l'expert. A cet égard Sneath et Sokal [31] écrivent :

«Goodall claims that the working taxonomist prefers to enhance the weights of rarer characteristics, but this does not seem to us to be a generally accepted dictum of taxonomic practice».

Le reste de la construction est conforme à AVL. Désignons ici par P_G l'indice probabiliste et par s_G l'indice brut qui prend ses valeurs dans la table de la figure 3. On a

$$P_G(i,i') = \Pr\{s_G(i*,i'*) < s_G(i,i')\} \tag{37}$$
$$= 0 \text{ si } s_G(i,i') = \lambda_G[c^1(o_i), c^1(o_{i'})] = 0 \tag{38}$$
$$= 1 - \sum\{p_j | \alpha_{jj} \geqslant s_G(i,i')\} \text{ sinon} \tag{39}$$

4.1.4. Enrichissement de l'échelle descriptive et indice AVL *associé*

C'est à partir de considérations purement sémantiques faisant appel à la connaissance de l'expert que nous proposons d'enrichir l'échelle descriptive sous-tendant l'ensemble des valeurs de l'attribut qualitatif. Il s'agit d'enrichir la valuation λ où seules deux valeurs numériques α et β sont prévues (*cf.* Figure 2) par une valuation λ_L pouvant comprendre jusqu'à $k1(k1+1)/2$ valeurs, comme c'est le cas de la Figure 4 suivante où $k1 = 6$.

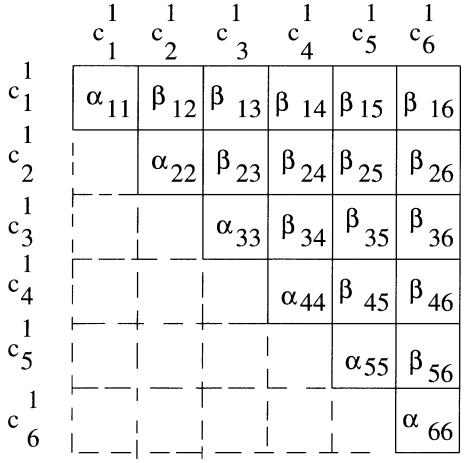


FIGURE 4

α_{jj} représente l'évaluation numérique de la similarité entre la catégorie c_j^1 et elle même, $1 \leqslant j \leqslant k1$; alors que β_{hj} représente celle, entre les catégories c_h^1 et $c_j^1, 1 \leqslant h \leqslant j \leqslant k1$. On peut clairement demander

$$\min\{\alpha_{jj} | 1 \leqslant j \leqslant k1\} > \max\{\beta_{hj} | 1 \leqslant h < j \leqslant k1\} \tag{40}$$

De toute façon, une manière réaliste d'obtenir une telle valuation λ_L est ordinale. Introduisons l'ensemble D suivant des couples de catégories

$$D = \{(c_h^1, c_j^1) | 1 \leqslant h \leqslant j \leqslant k1\} \tag{41}$$

Partant de D, on demandera à l'expert de sélectionner à chaque pas, dans l'ensemble des couples restants, l'ensemble des couples de catégories également les plus similaires. Plus formellement, il s'agit de l'algorithmique suivante :

Initialisation $E \longleftarrow D; l \leftarrow 1$
jqa $E = \phi$ faire
sélectionner dans E l'ensemble PP des couples de catégories les plus proches;
$E(l) \leftarrow PP; l \leftarrow l + 1;$
$E \longleftarrow E - PP$
finfaire

On définit ainsi un préordre total sur D, conforme à la perception de la ressemblance entre catégories que possède l'expert du domaine. Ainsi, la variable qualitative nominale est transformée en variable qualitative *préordonnance*. On codera *numériquement* ce préordre total au moyen de la fonction «rang moyen» . Ainsi, la valuation λ_L, permettant de charger un tableau tel que celui de la figure 4, pourra être spécifiée.

Maintenant, précisons l'indice probabiliste de la vraisemblance du lien qui se fonde sur la distribution empirique de l'indice

$$s_L(i, i') = \lambda_L[c(o_i), c(o_{i'})] \tag{42}$$

sur l'ensemble $P_2(I)$ des paires d'éléments de $I = \{1, 2, \ldots, i, \ldots, n\}$ qui code l'ensemble $\mathcal{O}$ des objets. Relativement à une paire d'objets $\{i_1, i_2\}$, cet indice s'écrit

$$P_L(i_1, i_2) = \Pr\{s_L(i*, i'*) < s_L(i_1, i_2)\} \tag{43}$$

où $\{i*, i'*\}$ est une paire aléatoire prise uniformément au hasard dans $P_2(I)$.

L'indice $P_L(i_1, i_2)$ représente la proportion de couples d'objets distincts pour lesquels la valeur de s_L est strictement inférieure à celle $s_L(i_1, i_2)$.

4.1.5. Arbre de classification associé à l'indice

Il peut paraître curieux de vouloir associer à une seule variable catégorielle C^1 qui définit une partition $\pi^1 = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_j, \ldots, \mathcal{O}_{k1}\}$ (*cf.* § 4.1.1), un arbre de classification obtenu au moyen d'une CAH. La construction d'un tel arbre ne prend son sens que si on considère une description par plus d'une seule variable. Néanmoins, une telle construction, associée aux diverses valuations λ, λ_G ou λ_L va permettre de se rendre compte comment les classes de la partition ci-dessus se situent dans l'arbre de classification.

(i) valuation λ

En considérant la similarité sur l'ensemble des objets induite par λ, on se rend compte que l'arbre de classification sur $\mathcal{O}$ issu d'une CAH, comprend trois niveaux. Le premier est celui des feuilles dont chacune représente une classe singleton comprenant un seul objet. Le second est celui comprenant $k1$ noeuds dont chacun sous-tend l'une des classes $\mathcal{O}_j$. Le troisième comprend le seul noeud racine qui fusionne les $k1$ noeuds.

(ii) valuation λ_G

Ici le nombre de niveaux dépend du nombre de valeurs distinctes α_{jj}, $1 \leqslant j \leqslant k1$. Si $h1$ est le nombre de valeurs distinctes ($h1 \leqslant k1$), alors le nombre de niveaux est $2 + h1$. Comme ci-dessus, le premier niveau est celui des feuilles dont chacune représente une classe singleton. Ensuite apparaissent, dans l'ordre décroissant des valeurs de α_{jj}, $h1$ niveaux. Pour un niveau donné, les noeuds, soustendent chacun l'une des classes $\mathcal{O}_j$, où les différentes classes correspondent à une même valeur de α_{jj}. Plus précisément, relativement aux catégories concernées par ces classes, l'autosimilarité d'une catégorie est constante. Le dernier niveau comprend la racine de l'arbre où les différents noeuds (correspondant chacun à l'une des classes $\mathcal{O}_j$) se rejoignent.

(iii) valuation λ_L

La forme de la section commençante de l'arbre jusqu'à l'apparition des noeuds dont chacun recouvre une classe $\mathcal{O}_j (1 \leqslant j \leqslant k1)$, est la même que dans le cas **(ii)**. Cependant, l'ordre d'apparition des classes n'est pas le même. En effet, les valeurs α_{jj} de la valuation λ_L (voir Figure 4) ne sont pas les mêmes que celles de la valuation λ_G (voir Figure 3) (*cf.*§ 4.1.3 et 4.1.4). D'autre part, il y a un ordre des agrégations des différents noeuds formés qui dépend de la valuation λ_L (voir le dessus de la diagonale principale dans le tableau de la figure 4) ainsi que du critère de fusion des classes adapté. A cet égard, un critère de la vraisemblance du lien maximal [8, 11, 26] s'avèrerait – compte tenu de sa souplesse – particulièrement adpaté.

Nous allons à présent considérer le cas général d'une description par plusieurs variables qualitatives (catégorielles).

4.2. Cas de plusieurs variables

On désigne par $\mathcal{V}$ l'ensemble des p variables qualitatives.

$\mathcal{V}$=$\{c^1, c^2, \ldots, c^j, \ldots, c^p\}$

L'application que définit la j-ème variable qualitative c^j est notée comme suit :

$c^j : \mathcal{O} \longrightarrow C^j = \{c_1^j, \ldots, c_h^j, \ldots, c_{kj}^j\}$

$o \longrightarrow c^j(o) \in C^j$

Dans ces conditions, on notera comme suit le tableau T des données :

$$T = \{c_i^j = c^j(o_i) \in C^j | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\} \tag{44}$$

On considère une Hypothèse d'Absence de Liaison (HAL) ou d'indépendance où à la suite des caractères observés on associe une suite de caractères aléatoires indépendants :

$$(c^1, c^2, \ldots, c^j, \ldots, c^p) \longrightarrow (c^{1*}, c^{2*}, \ldots, c^{j*}, \ldots, c^{p*}) : \tag{45}$$

Relativement à un même attribut aléatoire c^{j*} le modèle est permutationnel, $1 \leqslant j \leqslant p$.

4.2.1. L'indice de W.D. Goodall

On commence à établir pour chaque j, une valuation λ_G^j qui est relative à la variable c^j et de même nature que celle ci-dessus λ_G, illustrée par la Figure 3 ci-dessus, $1 \leqslant j \leqslant p$.

On considère ensuite, relativement au modèle aléatoire, l'indice du degré d'invraisemblance de la grandeur de la quantité $\lambda_G^j(c_g^j, c_h^j)$ où le couple de valeurs (c_g^j, c_h^j) a pu être observé chez un couple d'objets $(o_i, o_{i'})$:

$$P_{gh}^j = Pr\{\lambda_G^j[c^j(o_{i*}), c^j(o_{i'*})] \geqslant \lambda_G^j(c_g^j, c_h^j)]\} \tag{46}$$

où $(o_{i*}, o_{i'*})$ est un couple d'objets distincts aléatoire et où $1 \leqslant g \leqslant h \leqslant kj; kj$ étant le nombre de modalités de $c^j, 1 \leqslant j \leqslant p$.

On considère la distribution de la matrice λ_G, illustrée par la Figure 3 ci-dessus sur l'ensemble des couples d'objets distincts. L'indice P_{gh}^j représente très exactement la proportion de couples d'objets distincts $(o_i', o_{i'}')$ tels que

$$\lambda_G^j[c^j(o_i'), c^j(o_{i'}')] \geqslant \lambda_G^j(c_g^j, c_h^j) \tag{47}$$

Il peut y avoir pour j donné, jusqu'à $k_j + 1$ valeurs possibles de P_{gh}^j ; en effet, il y a une seule valeur pour $g \neq h$ et au plus k_j valeurs pour $g = h$ (voir Figure 3).

On peut ici rappeler que s'il n'y avait que le seul attribut c^j, l'indice de similarité probabiliste de Goodall, pour un couple d'objets, $(o_i, o_{i'})$ tel que $c^j(o_i) = c_g^j$ et $c^j(o_{i'}) = c_h^j$, s'exprime par

$$Q_{gh}^j = 1 - P_{gh}^j \tag{48}$$

Considérons à présent le cas de la suite des p attributs $(c^1, c^2, \ldots, c^j, \ldots, c^p)$. Relativement à un couple d'objets $(o_i, o_{i'})$, désignons par $(c_g^j | 1 \leqslant j \leqslant p)$ la suite des valeurs des différents attributs sur l'objet o_i et par $(c_h^j | 1 \leqslant j \leqslant p)$, celle, correspondante, sur $o_{i'}$. Considérons alors la fonction

$$\{(c_g^j, c_h^j) | 1 \leqslant j \leqslant p\} \rightarrow \Pi_{1 \leqslant j \leqslant p} P_{gh}^j \tag{49}$$

La suite $(c_g^j | 1 \leqslant j \leqslant p)$ est d'autant plus associée à la suite $(c_h^j) | 1 \leqslant j \leqslant p)$ que le second membre est *petit*.

Le second membre de (49) représente dans l'hypothèse d'indépendance entre attributs et de façon jointe, le degré d'invraisemblance de valeurs aussi grandes $\{\lambda_G^j(c_g^j, c_h^j) | 1 \leqslant j \leqslant p\}$, observées sur le couple d'objets $(o_i, o_{i'})$. Un indice de similarité probabiliste qu'on peut proposer ici, en se situant par rapport à l'hypothèse d'indépendance, est

$$P_1(o_i, o_{i'}) = 1 - \Pi_{1 \leqslant j \leqslant p} P_{gh}^j \tag{50}$$

C'est cet indice que nous utiliserons pour comparer le codage probabiliste de la similarité dans l'esprit de Goodall à celui que nous proposons dans le cadre de la méthode AVL.

L'approche précise de Goodall s'avère beaucoup plus complexe. Elle est fondée sur la distribution *empirique* de l'indice qui fait le second membre de (50) sur l'ensemble des couples d'objets distincts. Compte tenu de ce que nous venons d'exprimer ci-dessus il y a *a priori* et au maximum $\prod\{(k_j + 1) | 1 \leqslant j \leqslant p\}$ valeurs distinctes de cet indice; qu'il y a lieu de calculer et de trier. Il y a lieu ensuite de calculer quelle est la proportion de couples d'objets distincts pour lesquels chacune des valeurs est atteinte. Ceci, afin de calculer l'indice probabiliste suivant sur un couple d'objets $(o_i, o_{i'})$:

$$P_2(o_i, o_{i'}) = \Pr\{\Pi_{1 \leqslant j \leqslant p} P^j_{(c^j(o_i^*), c^j(o_{i'}^*))} > \Pi_{1 \leqslant j \leqslant p} P^j_{(c^j(o_i), c^j(o_{i'}))}\} \tag{51}$$

où $(o_i^*, o_{i'}^*)$ est un couple aléatoire d'objets distincts pris uniformément au hasard.

En d'autres termes, on commence par calculer un indice probabiliste (celui (49)); mais, dans le cadre d'une hypothèse d'indépendance mutuelle entre attributs. C'est la fonction de répartition empirique de cet indice qui donne lieu à l'indice probabiliste final. Une expression plus explicite peut être trouvée dans [5]. Certes, la complexité demeure polynomiale. Mais, elle est trop élevée; surtout à l'époque actuelle du «Data Mining» où il y a lieu de brasser des masses énormes de données.

4.2.2. L'indice adopté dans AVL

Pour chaque attribut $c^j, 1 \leqslant j \leqslant p$, une valuation λ_L^j telle que celle de la figure 4 ci-dessus, est établie à partir de la connaissance du domaine. Cette valuation numérique (voir § 4.1.4) permet la comparaison de chaque couple de catégories

$(c_g^j, c_h^j), 1 \leqslant g, h \leqslant kj$. Dans ces conditions, on peut définir la contribution de l'attribut c^j à l'indice brut de comparaison entre deux objets o_i et $o_{i'}$ sous la forme :

$$s_j(i, i') = s[c^j(o_i), c^j(o_{i'})] = \lambda_L^j[c^j(o_i), c^j(o_{i'})] \tag{52}$$

Le reste de la construction est alors en tout point analogue à ce qui a été proposé dans le cas numérique (*cf.* § 3). On peut en rappeler les différentes étapes. Pour tout $\{i, i'\}$ de $P_2(I)$

- $S_j(i, i')$: contribution statistiquement normalisée sur $I \times I$ de $s_j(i, i')$;
- $S(i, i')$: somme des contributions normalisées;
- $Q(i, i')$: normalisation de $S(i, i')$ sur l'ensemble $P_2(I)$ des parties à deux éléments de I;
- $P(i, i') = \Phi[Q(i, i')]$ où Φ est la fonction de répartition de la loi normale centrée et réduite.
- On peut associer à cet indice probabiliste, l'indice de «dissimilarité informationnelle» :

$$\mathcal{D}(i, i') = -Log_2(P(i, i')) \tag{53}$$

4.3. Différences conceptuelles entre les deux indices

A - Indice de Goodall

(i) La préordonnance valuée sur l'ensemble $\{c_g^j | 1 \leqslant g \leqslant kj\}$ des catégories de c^j est telle que deux catégories distinctes ont une ressemblance nulle; et, une même catégorie se trouve d'autant plus associée à elle même qu'elle est rare.

(ii) L'indice probabiliste entre objets est fondé sur la distribution empirique de la suite des variables qualitatives $(c^1, c^2, \ldots, c^j, \ldots, c^p)$. Le calcul exact est par trop complexe. Un indice de substitution peut être considéré. Il est calculé en tenant compte de la distribution empirique de chacune des variables prises séparément; mais se trouve calculé dans l'hypothèse d'indépendance entre ces variables. Il s'agit de l'indice (50).

B - Indice proposé dans la méthode AVL

(i) Seules des considérations sémantiques interviennent pour établir la préordonnance. Et, point très important, la ressemblance entre deux catégories différentes n'est pas nulle et varie entre une paire de catégories et une autre.

(ii) L'indice probabiliste suit un indice défini par la normalisation d'une somme de contributions normalisées variable par variable. Ainsi, il ne tient pas compte de la distribution empirique d'une même variable qu'à travers la moyenne et la variance

sur $I \times I$ d'un indice numérique de similarité entre catégories. Il se justifie dans l'hypothèse d'indépendance à partir du théorème central limite.

4.4. Réalisation de la méthode AVL

Nous avons déjà exprimé au paragraphe 2 le principe général de AVL et son extrême généralité par rapport à la structure mathématique du tableau des données. Dans l'usage de l'algorithmique de la CAH (Classification Ascendante Hiérarchique) la méthode propose une famille très générale de comparaison de classes dès lors qu'on a défini un indice de similarité probabiliste entre éléments [8, 11, 26, 28]. Nous nous restreignons quant à nous à la famille de critères que nous appelons de la « Vraisemblance du Lien Maximal » . Dans ce cas l'indice de comparaison entre deux classes $C1$ et $C2$ prend la forme suivante :

$$Pr\{\delta(C1*, C2*) \leqslant \delta(C1, C2)\} \tag{54}$$

où

$$\delta(C1, C2) = min\{D(i, i') \mid (i, i') \in C1 \times C2\} \tag{55}$$

où $D(i, i')$ a la forme (53) et où $(C1*, C2*)$ est un couple de classes aléatoires associé à $(C1, C2)$ dans une hypothèse probabiliste d'indépendance mutuelle entre les variables de description. L'expression calcul est en fait relativement simple. On pourra s'en rendre compte et suivre le traitement pas à pas et à la main d'un petit exemple dans [22].

La méthode comprend en son sein l'évaluation de la suite des niveaux de l'arbre des classifications au moyen de la distribution observée sur cette suite d'un critère statistique de nature non paramétrique [13]. Pour un niveau donné, ce dernier mesure l'adéquation entre la partition produite et la similarité sur l'ensemble des paires d'éléments de l'ensemble organisé. Ce critère est normalisé dans le sens où il s'agit d'une variable aléatoire normale centrée et réduite si au lieu de la partition observée on avait une partition aléatoire ayant des classes de mêmes cardinaux. Ce critère est appelé dans la sortie du programme CHAVL [22] et dans le contexte d'utilisation que nous venons d'exprimer « STATISTIQUE GLOBALE » des niveaux. Les tableaux 2 et 3 suivants donnent des illustrations de la distribution de cette statistique. Un maximum local indique une partition qui définit un état d'équilibre dans la synthèse. Nous appelons « significative » une telle partition. Pour ce qui est du tableau 2, il s'agit clairement de la partition de niveau 17; mais aussi de celle du niveau 20. Pour ce qui est maintenant du tableau 3, il s'agit des niveaux 19 et 21. La « STATISTIQUE LOCALE » des niveaux correspond au taux d'accroissement de celle globale entre le niveau concerné et celui qui le précède. La statistique locale juge l'apport des noeuds qui viennent de se former à un niveau donné. Les tableaux 2 et 3 en donnent des illustrations. La marge gauche de ces tableaux indique les maxima locaux de la distribution de la statistique locale le long de la suite des niveaux. Ils indiquent ce que nous appelons les « noeuds significatifs » de l'arbre des classifications. Une expérience éprouvée montre que ces derniers correspondent à des niveaux d'achèvement de classes. Ils sont ponctués dans la représentation graphique de l'arbre par la présence d'une *.

Nous nous proposons maintenant de comparer les comportements sur un exemple illustratif, d'une part, de la forme simplifiée de l'indice de Goodall [voir (50)], d'autre part, de l'indice que nous proposons dans le cadre de notre méthode, $P(i, i')$ ou d'ailleurs $D(i, i')$ [voir (53)].

5. Un exemple de traitement

Nous avons voulu comparer sur un petit exemple réel via la méthode AVL les comportements de l'indice de Goodall et celui que nous proposons dans le cadre de cette approche et qui prend pour base une préordonnance sur l'ensemble des catégories. À cet égard, nous avons pris un petit tableau de données issu de [6] et qui correspond à un extrait des scrutins de l'ONU en 1968 (*cf.* Tableau 1).

TABLEAU 1
Scrutins de l'ONU en 1968

USA	3	3	1	IREL	3	2	1
CANA	3	2	1	NETH	3	3	1
CUBA	1	4	3	BELG	3	3	1
HAIT	4	1	1	LUXE	3	3	1
DOMI	1	1	1	FRAN	3	2	3
JAMA	3	1	1	SPAI	3	2	1
TRIN	3	1	1	PORT	4	3	2
BARB	4	1	2	POLA	1	1	3
MEXI	3	1	1	AUST	3	2	2
GUAT	3	1	1	HUNG	1	1	3
HOND	3	2	1	CZEC	1	1	3
EL S	3	2	1	ITAL	3	2	1
NICA	3	2	1	MALT	4	4	1
COST	3	1	1	ALBA	1	4	3
PANA	3	1	1	YUGO	1	1	3
COLU	4	2	1	GREE	3	1	1
VENE	3	1	1	CYPR	3	1	1
GUYA	4	1	1	BULG	1	1	3
EQUA	3	4	2	ROMA	1	1	3
PERU	3	1	1	USSR	1	1	3
BRAZ	3	2	1	UKRA	1	1	3
BOLI	4	1	1	BYEL	1	1	3
PARA	3	2	1	FINL	2	2	3
CHIL	3	1	1	SUED	3	2	3
ARGE	3	1	1	NORW	3	2	3
URUG	3	4	1	DENM	3	2	3
U.K.	3	3	1	ICEL	3	2	1

À chacun des scrutins considérés nous associons une variable qualitative nominale dont l'ensemble des valeurs est l'ensemble des votes possibles. On a ainsi l'ensemble $\{c^1, c^2, c^3\}$ des trois attributs qualitatifs. L'ensemble des valeurs

possibles d'un même attribut c^j est {oui, non, abstention, absence}. Nous codons respectivement, simplement pour des raisons de représentation en machine, par 1,2,3 et 4, les quatre modalités de réponse : oui, non, abstention et absence. Dans le cadre de notre approche, nous avons besoin d'établir une préordonnance sur l'ensemble des valeurs de l'attribut. À partir de considérations intuitives liées à la consistance de la réponse donnée, nous proposons la préordonnance suivante pour chacun des trois attributs c^1, c^2 et c^3 :

$$12 < 13 < 14 < 24 < 23 < 34 < 44 < 33 < 11 < 22 \tag{56}$$

où $lm(l \leqslant m)$ désigne le couple de catégories (l, m) et où on suppose que cette similarité ordinale est symétrique : la ressemblance entre l et m est la même que celle entre m et l. Cette préordonnance peut bien sûr être rediscutée. Dans notre cas, il s'agit d'un ordre total et strict sur l'ensemble $\{(l, m)|1 \leqslant l \leqslant m \leqslant 4\}$ auquel on peut associer la fonction rang qui prend l'ensemble des valeurs depuis 1 jusqu'à 10 :1 pour (1,2), 2 pour (2,3),. . ., 10 pour (2,2).

Dans le cas de la construction de l'indice de Goodall, à chacun des attributs c^1, c^2 et c^3 se trouve associé une préordonnance comprenant une première classe à gauche formée des couples de valeurs distinctes de l'attribut. Strictement à droite de cette classe, la suite des couples de valeurs identiques se trouve ordonnée conformément à la suite décroissante de la fréquence de présence. Ainsi on a :

Pour c^1 :

$$12 \sim 13 \sim 14 \sim 23 \sim 24 \sim 34 < 33 < 11 < 44 < 22 \tag{57}$$

où les valeurs de α_{jj}, $1 \leqslant j \leqslant 4$ (voir § 4.1.3) sont :

$$\alpha_{33} = 2.55, \alpha_{11} = 21.68, \alpha_{44} = 68.14 \quad \text{et} \quad \alpha_{22} = \infty.$$

D'autre part, les préordonnances associées à c^2 et c^3 sont respectivement :

$$12 \sim 13 \sim 14 \sim 23 \sim 24 \sim 34 < 11 < 22 < 33 < 44 \tag{58}$$

$$12 \sim 13 \sim 14 \sim 23 \sim 24 \sim 34 < 11 < 33 < 22 < 44 \tag{59}$$

On a en effet respectivement pour c^2 et c^3 :

$$\alpha_{11} = 4.40, \alpha_{22} = 10.52, \alpha_{33} = 95.42 \quad \text{et} \quad \alpha_{44} = 143.06,$$

$$\alpha_{11} = 2.55, \alpha_{33} = 119.19, \alpha_{22} = 238.66 \quad \text{et} \quad \alpha_{44} = \infty.$$

La valuation associée à un couple (l, l) de catégories est une fonction décroissante de la proportion d'objets possédant la catégorie l.

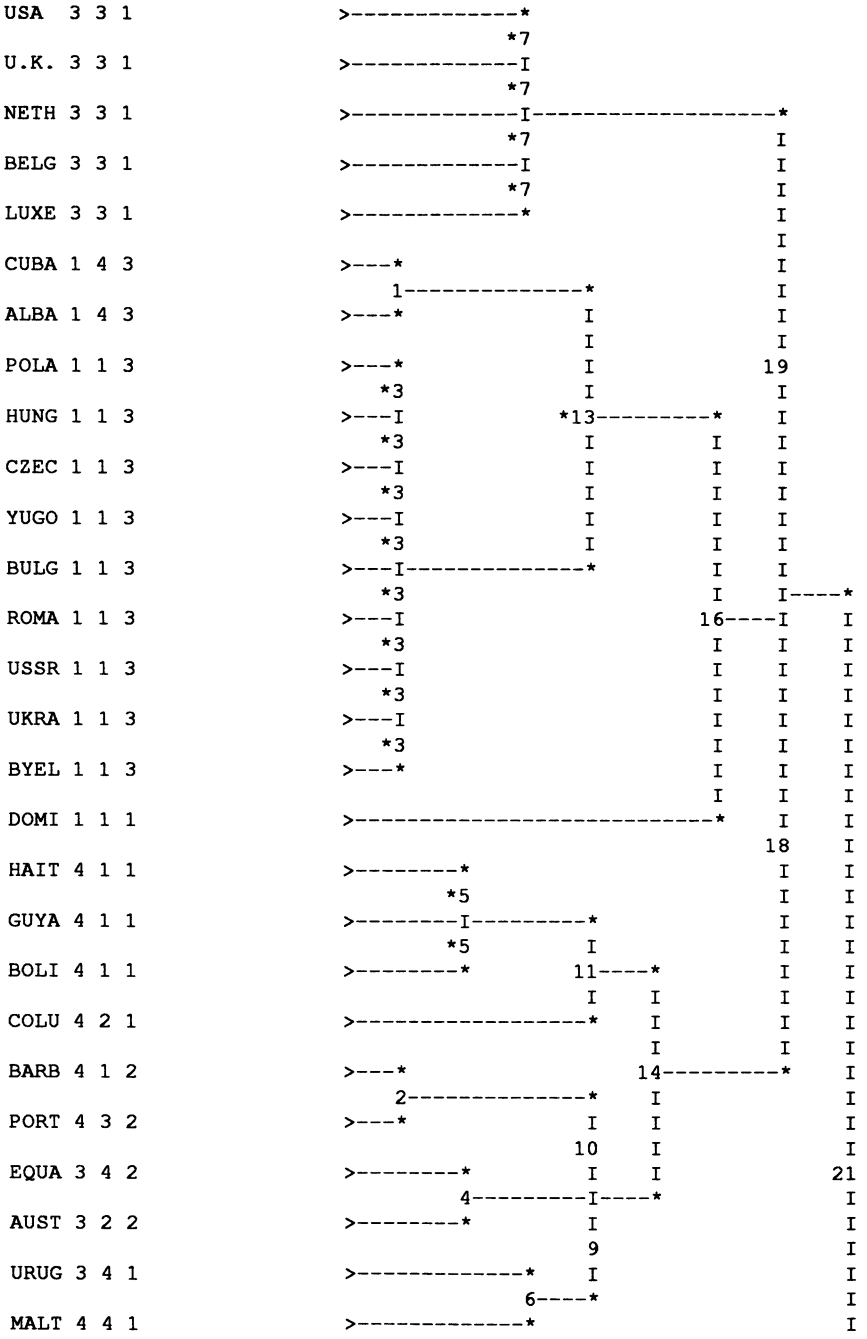


FIGURE 5 (début)
Arbre obtenu par l'AVL(G)

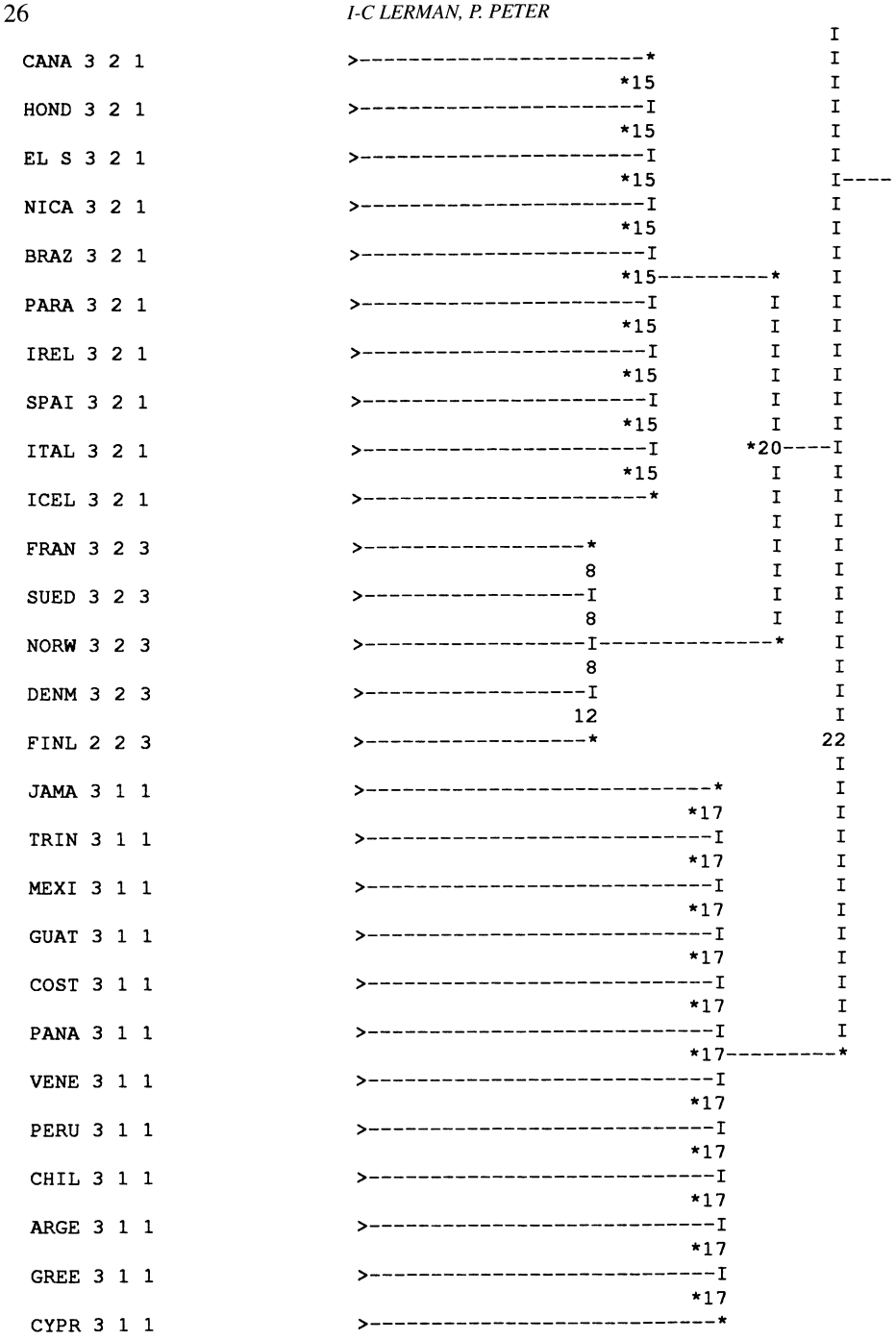


FIGURE 5 (fin)
Arbre obtenu par l'AVL(G)

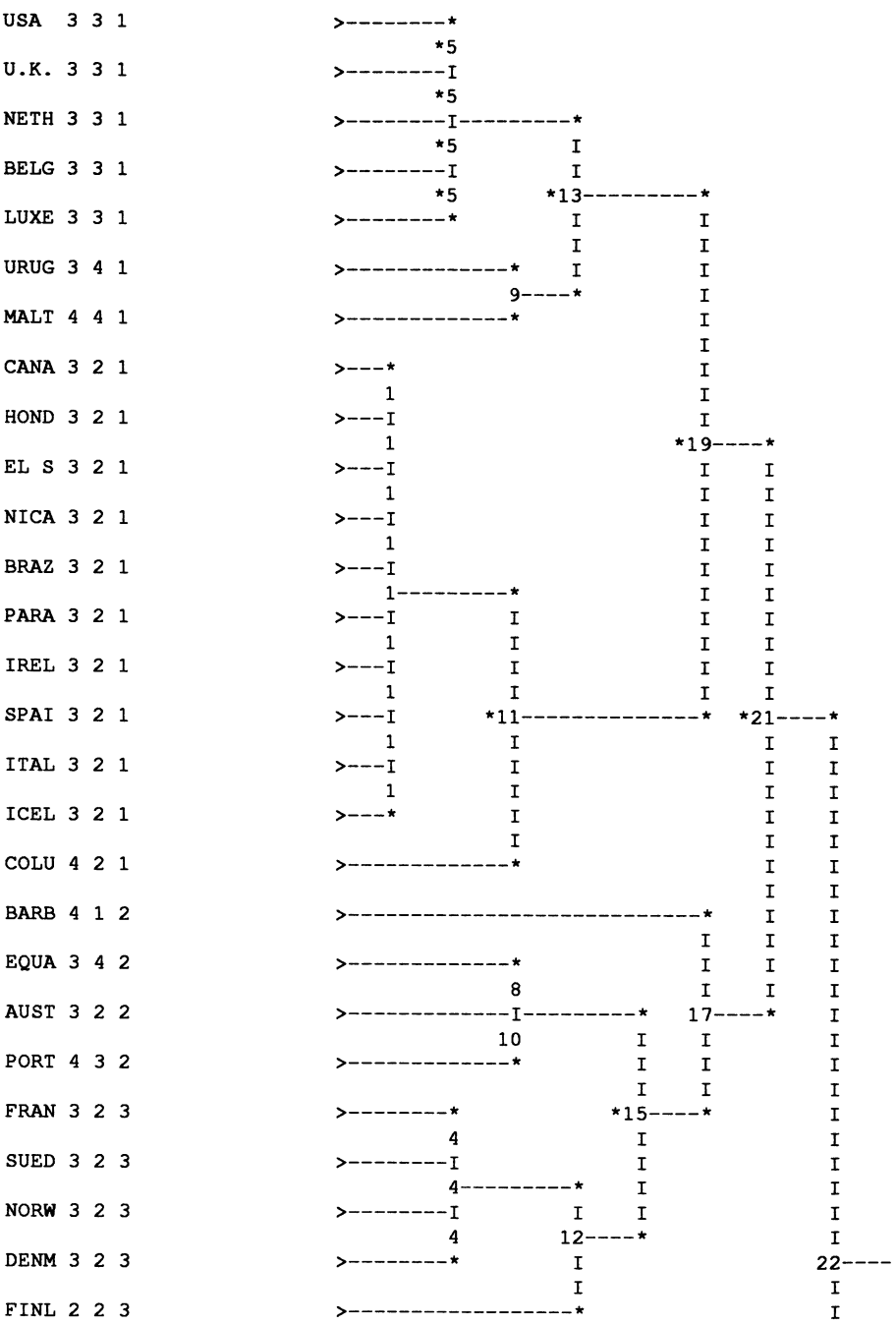


FIGURE 6 (début)
Arbre obtenu par l'AVL(L)

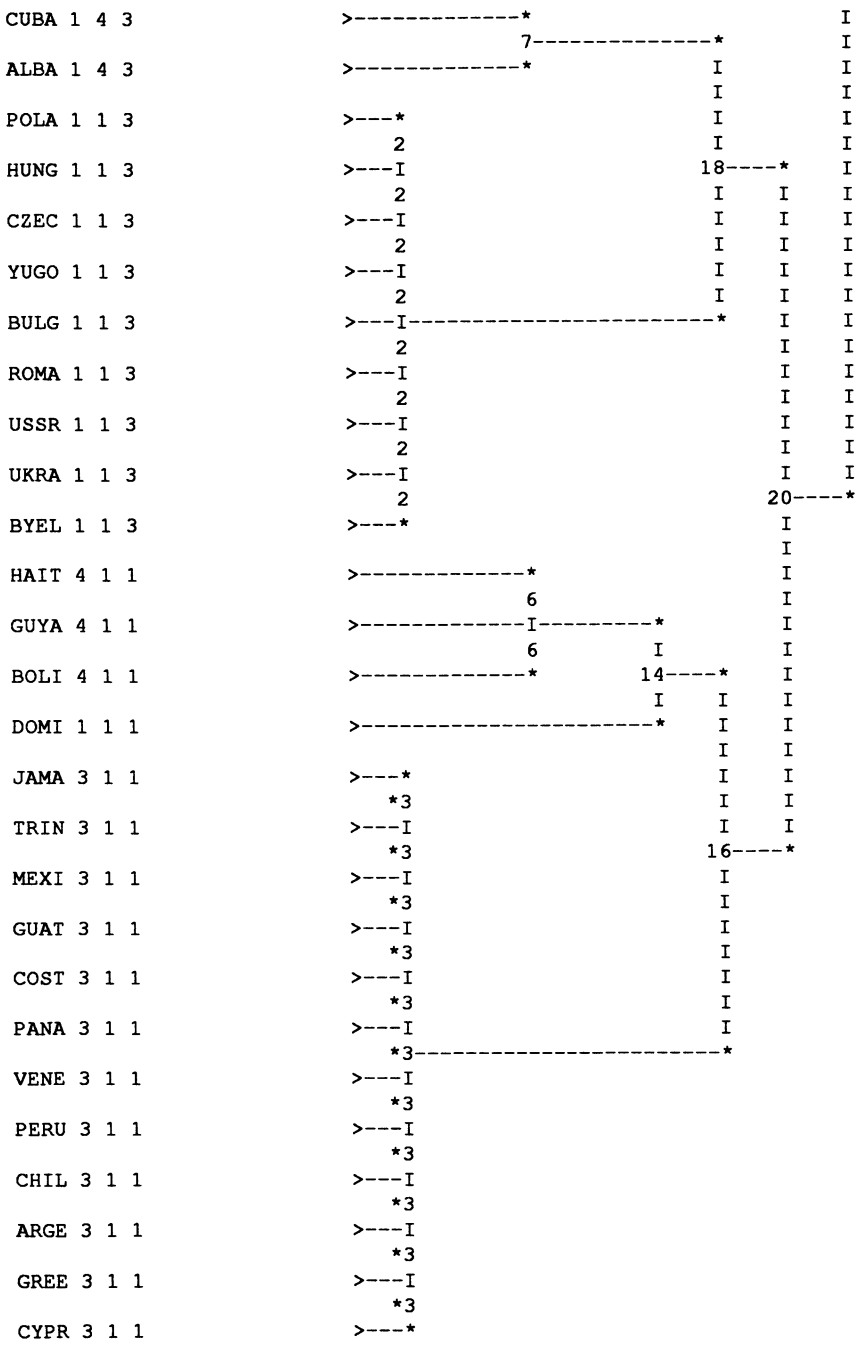


FIGURE 6 (fin)
Arbre obtenu par l'AVL(L)

TABLEAU 2
Statistiques des niveaux (cas de l'AVL(G))

	NIVEAU	STATISTIQUE GLOBALE	STATISTIQUE LOCALE
	1	1.1583	1.1583
	2	1.6427	0.4845
1 MAXIMUM	3	5.4407	3.7980
	4	5.5598	0.1191
2 MAXIMUM	5	5.8958	0.3360
	6	6.0127	0.1169
3 MAXIMUM	7	6.9719	0.9592
	8	7.5698	0.5980
	9	7.5052	-0.0646
	10	7.5282	0.0230
	11	7.8100	0.2817
	12	8.1467	0.3367
4 MAXIMUM	13	8.8135	0.6668
	14	8.0509	-0.7627
5 MAXIMUM	15	11.1676	3.1167
	16	11.6576	0.4900
6 MAXIMUM	17	16.1527	4.4951
	18	6.5586	-9.5941
	19	1.6484	-4.9102
7 MAXIMUM	20	4.1191	2.4707
	21	-2.8096	-6.9287
	22	-2.4414	0.3682

Les figures 5 et 6 donnent respectivement, les arbres de classification AVL respectivement associés à l'indice de Goodall et à celui, conçu dans le cadre de l'AVL et associé à la préordonnance (56). Appelons AVL(G) le premier et AVL(L) le second. Les tables 2 et 3 donnent les distributions des statistiques globale et locale des niveaux pour chacun des deux arbres AVL(G) et AVL(L).

Il est clair d'après ces distributions, que l'arbre de classification AVL(L) associé à la préordonnance (56) que nous avons posée et à la manière de calculer l'indice dans AVL, est notablement plus consistant que celui AVL(G) associé à l'indice de Goodall. En effet, d'une part, la distribution de la statistique globale des niveaux culmine dans AVL(L) avec une valeur notablement plus forte que dans AVL(G) : 26.3 (niveau 19) au lieu de 16.15 (niveau 17), d'autre part, le prochain niveau significatif est beaucoup plus net dans AVL(L) avec une valeur de 20.94 (niveau 21) que dans le cas de AVL(G) avec la valeur de 4.12 (niveau 20). Enfin, davantage les associations de niveaux relativement élevés concernent des classes déjà constitués de plusieurs éléments.

Maintenant, il est difficile d'apporter un commentaire très appuyé sur l'interprétation et la comparaison de deux arbres, sur la base de 3 scrutins seulement. Néanmoins, l'organisation proposée dans AVL(L) semble sensiblement plus cohérente

TABLEAU 3
Statistiques des niveaux (cas de l'AVL(L))

	NIVEAU	STATISTIQUE GLOBALE	STATISTIQUE LOCALE
	1	7.5284	7.5284
	2	10.8348	3.3064
1 MAXIMUM	3	15.6112	4.7763
	4	16.4715	0.8603
2 MAXIMUM	5	17.8499	1.3784
	6	18.3397	0.4898
	7	18.5136	0.1739
	8	18.6684	0.1548
	9	18.8191	0.1507
	10	18.9469	0.1278
3 MAXIMUM	11	19.1849	0.2380
	12	19.3704	0.1855
	13	19.7167	0.3463
	14	20.1709	0.4543
4 MAXIMUM	15	21.4105	1.2396
	16	22.4139	1.0034
5 MAXIMUM	17	23.4604	1.0464
	18	23.8753	0.4149
6 MAXIMUM	19	26.2990	2.4237
	20	19.4604	-6.8386
7 MAXIMUM	21	20.9365	1.4761
	22	-2.8604	-23.7970

que celle de AVL(G) [1]. Appelons ici Bloc un ensemble de pays dont les patrons de réponses sont identiques. Il en est par exemple ainsi de l'ensemble des pays {France, Suède, Norvège, Danemark} ou encore de celui {USA, Grande Bretagne, Hollande, Belgique, Luxembourg}. Il est clair que, compte tenu des relations entre catégories, deux blocs distincts se forment à des niveaux distincts. D'autre part, quel que soit le codage de la similarité entre catégories, deux éléments d'un même bloc se ressemblent davantage que deux éléments de profils différents.

Dans ces conditions on peut remarquer que les Blocs s'imposent beaucoup plus vite dans l'arbre AVL(L) que dans AVL(G). La classe de AVL(L) comprenant le noyau dur de l'occident avec {USA, U.K., NETH} et qui s'achève avec le noeud significatif 21 (voir figure 6) est plus cohérente que celle de AVL(G) contenant ce même noyau et se terminant au noeud 19. En effet, cette dernière finit par inclure le bloc de l'est. Le bloc de l'est où la singularité de la paire {Cuba, Albanie} – qui s'y relie – est reconnue dans les deux cas, rejoint dans AVL(L) un bloc de pays d'Amérique Centrale et du Sud ainsi que la Grèce et Chypre. Ce sont des pays souvent non alignés du tiers monde où la dictature règne. On reconnaît le groupe très cohérent {Haïti, Guyane, Bolivie,

[1] Nous tenons à remercier M. Christian LEBART, directeur du département AES de l'Université de Rennes 2, pour nous avoir aidé à appréhender les résultats.

St Domingue}. Dans AVL(G) ce groupe perd St Domingue et raccroche la Colombie pour s'inscrire dans un groupe dont l'homogénéité est difficile à contrôler puisqu'il contient une classe comprenant les trois fourches {Barbada, Portugal}, {Equateur, Autriche} et {Uruguay, Malte}. Toujours dans AVL(G), le groupe européen spécifique {France, Suède, Norvège, Danemark, Finlande} rejoint directement le Bloc {Canada, Honduras, El Salvador, Nicaragua, Brésil, Paraguay, Irlande, Islande, Espagne, Italie}, comprenant des pays de l'Amérique du nord, de celle centrale de celle du sud, ainsi que des pays de l'extrême nord ouest et de l'extrême sud ouest de l'europe. Tandis que dans le cas de AVL(L), le groupe européen que nous venons de mentionner {France, Suède, Norvège, Danemark, Finlande} commence par se mettre dans un moule plus épais comprenant des pays atypiques tels que le Portugal, l'Autriche, l'Equateur et la Barbade. L'ensemble, certes hétérogène, de ces pays, a chacun une personnalité marquée par rapport aux trois tendances que sont celle occidentale comprenant les Etats-Unis et la Grande Bretagne et celles empiétantes des états socialistes et des pays non alignés comprenant le tiers monde. Cet ensemble rejoint un groupe consistant, teinté à la limite par ce qui représente la position occidentale. Cette organisation où d'ailleurs la Colombie rejoint le Bloc comprenant le Paraguay, semble plus cohérente et harmonieuse que celle, fournie dans AVL(G).

6. AVL dans le cas général

6.1. AVL dans le cas de comparaison de variables

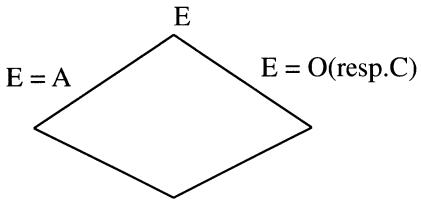
Comme nous l'avons mentionné au paragraphe 2, la méthode AVL a été bâtie en s'adressant d'abord au problème de la classification des variables (Daudé [3], Lerman [9, 11, 16, 17, 18] Nicolaü et Bacelar-Nicolaü [1, 26, 28], Ouali-Allah [29]). C'est dans un deuxième temps que le problème de la classification des objets puis des catégories a été traité (Lerman [15], Lerman et Peter [19, 20], Lerman et Peter [21], Lerman [12, 14], Lerman et Tallur [23]).

Il est important de noter que pour le problème de la classification des variables les différents objets revêtent exactement la même importance. On les considère comme statistiquement indépendants et devant contribuer également à la ressemblance entre variables. Il n'en est pas nécessairement de même pour le problème dual de la comparaison d'objets où la question de la pondération des variables peut être discutée. A cet égard, un des principes posés par les taxinomistes consiste à attribuer *a priori* la même importance aux différentes variables de description. Quant à nous, nous avons adopté le principe de l'égale discrimination, en ayant au préalable défini de façon adéquate ce que doit être la contribution brute d'une variable à la comparaison de deux objets.

Dans ces conditions, quelle que soit la nature de l'ensemble E à être organisé en classes :

$E = \mathcal{A}$: ensemble des variables descriptives ou $E = \mathcal{O}$ (resp. $\mathcal{C}$) : ensemble d'objets (resp. de catégories)

on aboutit à une similarité probabiliste sur E selon le schéma suivant :



Similarité probabiliste sur E.

FIGURE 7

6.2. Un projet de logiciel

Désignons par E.CHAVL l'arbre de classification hiérarchique AVL condensé à un système de noeuds « significatifs » et où les partitions « significatives » sont repérables. Il s'agit de réaliser le méta algorithme défini par le schéma de la figure 7 pour la palette la plus large de structures de la table de données. À cet égard, on suppose établi un ensemble $\mathcal{A}$ de variables descriptives (attributs descriptifs) qui se décompose comme suit :

$$\mathcal{A} = \mathcal{A}_{bl} \cup \mathcal{A}_{nm} \cup \mathcal{A}_{qn} \cup \mathcal{A}_{qo} \cup \mathcal{A}_{qp} \cup \mathcal{A}_{gr},$$

où $\mathcal{A}_{bl}$ est un ensemble d'attributs booléens, $\mathcal{A}_{nm}$ est un ensemble d'attributs numériques, $\mathcal{A}_{qn}$ est un ensemble d'attributs qualitatifs nominaux, $\mathcal{A}_{qo}$ est un ensemble d'attributs qualitatifs ordinaux, $\mathcal{A}_{qp}$ est un ensemble d'attributs qualitatifs préordonnances et $\mathcal{A}_{gr}$ est un ensemble d'attributs graphes valués.

E $\leftarrow \mathcal{A}$;
AVLA $\leftarrow$ E.CHAVL;
E $\leftarrow \mathcal{O}$ (resp. E $\leftarrow C$);
AVLB $\leftarrow$ E.CHAVL;
Croiser (AVLA,AVLB)

FIGURE 8
Le méta-algorithme

Le croisement se fait au moyen de coefficients numériques d'association. Il permet de préciser le rôle d'une feuille ou d'un noeud « significatif » de l'un des arbres par rapport à un système intéressant de noeuds « significatifs » de l'autre arbre.

Les programmes existants conformes aux normes de Modulad (Leredde [7]), Peter, Leredde et Lerman [22], Ouali-Allah [30] sont CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens) et AVARE (Association entre VAriables RElationnelles). Ils permettent de produire une classification hiérarchique de l'ensemble des lignes ou de l'ensemble des colonnes du tableau des données, compte tenu de sa structure mathématique. Cependant, un seul type de variables descriptives est à chaque fois en jeu. Nous voulons ici signaler un dernier développement

matérialisé par un programme nommé HETAVL (HETérogène AVL) où on autorise le mélange de différents types de variables. Ce développement est à inscrire dans le cadre du programme CHAVL [22].

La fonction Croiser(AVLA,AVLB) a été partiellement réalisée. Des modules existent dans le cas où les attributs sont numériques, booléens ainsi que dans le cas où la donnée est une juxtaposition horizontale de tableaux de contingence. À partir de ces acquis, les développements théoriques et de calcul existent pour la compléter.

Remerciements

Nous remercions vivement Pierre Cazes pour sa lecture très attentive qui a conduit à améliorer la présentation et à corriger quelques points.

Références

[1] H. BACELAR-NICOLAÜ (1987) On the distribution equivalence in cluster analysis. In Devijver P.A. and Kitler, J, *Pattern Recognition Theory and Applications*, pages 73-79. Springer-Verlag.

[2] H. BENHADDA and F. MARCOTORCHINO (1998) Introduction à la similarité régularisée en analyse relationnelle. *Revue de Statistique Appliquée*, XIVI(1) :45-69.

[3] F. DAUDÉ (1992) *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*, PhD thesis, Université de Rennes 1.

[4] G. DROUET-D'AUBIGNY (1993) Analyse des proximités et programme de codage multidimensionnel, *La Revue de Modulad*, dec, (12) :1-32.

[5] W.D. GOODALL (1966) A new similarity index based on probability, *Biometrics*, 22(4) :882-90.

[6] M. JAMBU (1978) *Classification automatique pour l'analyse des données*, Dunod.

[7] H. LEREDDE (1991) Normalisation fortran-77. *La revue de modulad*, (7) :5-65.

[8] I.C. LERMAN (1970) Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité. *Mathématique et Sciences Humaines*, (32) :5-15.

[9] I.C. LERMAN (1973) Etude distributionnelle de statistiques de proximité entre structures finies de même type; application à la classification automatique. *Cahiers du BURO*, 19.

[10] I.C. LERMAN (1973) Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie de personnages enfants à travers la littérature enfantine. *Revue de Statistique Appliquée*, XXI(3).

[11] I.C. LERMAN (1981) *Classification et analyse ordinale des données*. Dunod.

[12] I.C. LERMAN (1983) Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de cotingence. *Revue Mathématique et Sciences Humaines*, (83) :5-30.

[13] I.C. LERMAN (1983) Sur la signification des classes issues d'une classification automatique. In J. Felsenstein, editor, *Numerical Taxonomy NATO Series*, pages 179-198. Springer Verlag.

[14] I.C. LERMAN (1984) Analyse classificatoire d'une correspondance multiple, typologie et régression. in E. Diday et al. editor, *Data Analysis and Informatics III*, pages 193-221. North Holland.

[15] I.C. LERMAN. (1987) Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque, application au problème de consensus en classification. *Revue de Statistique Appliquée*, XXXV(2) :49-56.

[16] I.C. LERMAN (1988) Comparing relational variables according to likelihood of the links classification method. in M. Jambu, E. Diday, C. Hayashi and N. Ohsumi, editors, *Proceedings of the Japanese-French Scientific Seminar, March 24-26, 1987*, pages 187-200, Academc Press.

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

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

[19] I.C. LERMAN and Ph. PETER (1985) Elaboration et logiciel d'un indice de similarité entre objets de type quelconque. Application au problème de consensus en classification. Publication Interne 262, IRISA-INRIA.

[20] I.C. LERMAN and Ph. PETER (1986) Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle. In E. Diday et al., editor, *Data Analysis and Informatics IV*, pages 121-136. North Holland.

[21] I.C. LERMAN and Ph. PETER (1989) Classification of concepts described by taxonomic preordonnace variables with multiple choice. Application to the structuration of a species set of phlebotomine. in E. Diday and al., editor, *Data Analysis, Learning Symbolic and Numeric Knowledge*, pages 73-87. Nova Science Publishers.

[22] I.C. LERMAN, Ph. PETER and H. LEREDDE (1993) Principes et calculs de la méthodes implantée dans le programme CHAVL (classification hiérarchique par analyse de la vraisemblance des liens). *La revue de modulad*, (12) :33-70.

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

[24] R.F. NGOUËNET (1993) Une nouvelle famille d'indices de dissimilarité pour la mds. Publication Interne 766, IRISA.

[25] R.F. NGOUËNET (1995) *Analyse géométrique des données de dissimilarité par le multidimensional scaling : une approche parallèle basée sur les algorithmes génétiques.* Application aux séquences biologiques. PhD thesis, Université de Rennes 1.

[26] F. NICOLAÜ (1981) *Critérios de análise classificatória hierárquica baseados na função de distribuição.* PhD thesis, Faculty of Lisbon.

[27] F. NICOLAÜ and P. BRITO (1989) Improvements in NHMEAN method. In E. Diday, editor, *International Symposium on Data Analysis, Learning Symbolic and Numerical Knowledge*. Nova Science.

[28] F.C. NICOLAÜ and H. BACELAR-NICOLAÜ (1998) Some trends in the classification of variables. in C. Hayashi, N. Oshumi, K. Yajima, Y. Tamaka, H.-H. Bock and Y. Baba, editors, *Data Science, Classification and Related Methods*, page 89-98. Springer-Verlag.

[29] M. OUALI-ALLAH (1991) *Analyse en préordonnances des données, applications aux données numériques et symboliques*, PhD thesis, Université de Rennes 1.

[30] M. OUALI-ALLAH (2000) Programme de coefficients d'association entre variables relationnelles. *La revue de modulad*, (25) :63-73.

[31] P.H.A SNEATH and R.R. SOKAL (1973) *Numerical Taxonomy*, W.H. Freeman and company.