Analyse de la Vraisemblance des Liens Relationnels : *Une méthodologie d'analyse classificatoire des données*

Israël-César Lerman

Université de Rennes 1-IRISA Projet *Symbiose*,
Campus Universitaire de Beaulieu, 35042 Rennes Cedex
lerman@irisa.fr
http://www.irisa.fr/symbiose/

Résumé. La méthodologie de classification des données par l'Analyse de la Vraisemblance des Liens Relationnels a pris naissance vers la fin des années soixante. Elle s'est très largement développée. De nombreux chercheurs et praticiens ont pris part à son développement. De nombreuses applications d'envergure provenant des domaines les plus divers (Bioinformatique, Informatique, Sciences sociales, Traitement d'images, Traitement de langues naturelles, ...) ont validé cette méthodologie. Elle s'adresse à n'importe quel type mathématicologique du tableau de description des données. L'objet de notre article est de présenter de façon illustrée les principes fondamentaux de cette approche. Ces principes se situent d'une part au niveau de la représentation de la description et d'autre part, au niveau de l'évaluation quantifiée des ressemblances entre les structures mathématiques à comparer. Dans notre cas la représentation de la description sera ensembliste et relationnelle et la ressemblance sera évaluée au moyen d'une similarité qui se réfère à une échelle de probabilité établie par rapport à une hypothèse statistique d'absence de liaison. Le texte ci-dessous se veut le reflet de ma présentation orale aux "$3^{\text{-èmes}}$ Journées Thématiques Apprentissage Artificiel et Fouille des Données", 8-9 avril 2008.

1 Préambule

Mon texte s'apparente à une introduction. Il se veut le reflet exact de ma présentation orale aux "$3^{\text{-èmes}}$ Journées Thématiques Apprentissage Artificiel et Fouille des Données", 8-9 avril 2008. Ce texte ne correspond pas à la forme classique d'un article.

Il s'agit pour nous de présenter une méthodologie d'analyse classificatoire des données sur laquelle nous avons travaillé de très longues années. De nombreuses collaborations ont contribué de façon importante à sa mise au point. Cette méthode n'est pas vraiment connue dans ses différentes facettes ; alors qu'elle repose sur deux principes fondamentaux. Le premier consiste en la représentation ensembliste et relationnelle des variables (attributs) de description de l'ensemble des objets ou individus et le second, en l'évaluation probabiliste de la similarité entre les structures mathématiques à comparer (Indices de la *vraisemblance des liens*). Ces principes sont concrétisés quelle que soit la complexité de la description des données. Il peut s'agir

d'associer des descripteurs (resp., des objets), des classes de descripteurs (resp., des classes d'objets), il peut aussi s'agir de reconnaître des agrégats "intéressants" qui se distinguent, ... Dans ma présentation, je chercherai autant que possible à insister sur les principes en m'appuyant sur des exemples.

Il est fondamental d'avoir à l'esprit qu'une méthodologie d'*Analyse des Données* se conçoit *toujours* par rapport à une algorithmique d'un type donné. À cet égard, les concepteurs de méthodes de classification ont vite tendance à vouloir faire oublier l'algorithmique qui a servi de nid de fécondation aux méthodes construites pour parler de "théorie générale". Certes, il est très important et très intéressant de vouloir transposer les acquis méthodologiques dans le cadre d'autres algorithmiques ; et d'ailleurs, c'est une richesse de l'approche initialement proposée. Dans notre cas l'algorithmique en question est fournie par la *Classification Ascendante Hiérarchique*, la CAH, comme on dit.

Il sera donc juste de rappeler le principe et les ingrédients de cette algorithmique pour montrer comment ils se déclinent dans le cadre de *l'Analyse de la Vraisemblance des Liens Relationnels* $AVLR$. Le plan de notre présentation est :

1. Préambule
2. Introduction : principes généraux et exemples
3. Le principe algorithmique et les éléments constitutifs d'une *CAH*
4. La méthode de l'*Analyse de la Vraisemblance des Liens Relationnels* (Une présentation à partir d'exemples)
5. Les grands types logico-mathématiques de données
6. Logiciels et Applications
7. Références

2 Introduction : principes généraux et exemples

2.1 Le paradigme de la CAH, schéma d'arbre et exemples

Le paradigme de la CAH remonte à loin ...Michel Adanson, dans son *Histoire naturelle du Sénégal* (Bauche, Paris, 1757) écrivait :

"Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances ... Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore afin d'en faire un tout dont les parties soient unies et liées intimement."

Je pourrais peut-être ajouter : *"liées intimement"* sur le plan de la cohérence.

Le schéma d'arbre, tel que celui présenté ci-dessous (voir FIG. 1) sur un ensemble E de 10 éléments, donne précisément une image de la suite des *rapprochements* telle que décrite par Michel Adanson ; en "petites" familles puis ces dernières en de plus "grosses". Le processus de rapprochement est parfaitement conforme à la construction ascendante hiérarchique d'un

arbre des classifications. Le niveau 0 est celui des feuilles de l'arbre. Chaque feuille représente - j'allais dire un élément - plus exactement, une partie comprenant exactement un élément. Entre les niveaux 0 et 1 on a les rapprochements suivants : $\{g\}$ et $\{h\}$, $\{f\}$ et $\{i\}$ et $\{c\}$ et $\{e\}$. Entre les niveaux 1 et 2, on a d'une part le rapprochement entre $\{b\}$ et $\{g, h\}$ et d'autre part, celui entre $\{a\}$, $\{j\}$ et $\{c, e\}$. Ce dernier rapprochement donne lieu au noeud $\{a, c, e, j\}$. La racine de l'arbre correspond à l'ensemble plein E. Après une étape donnée on peut distinguer la notion de *classification* (partition) produite à un niveau donné. Ainsi, en est-il de la partition $\{\{b, g, h\}, \{d\}, \{f, i\}, \{a, c, e, j\}\}$ qu'on peut noter π_2, produite au niveau 2 de l'arbre.

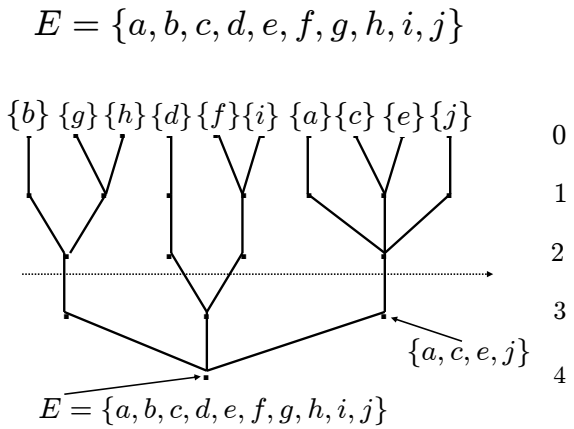

FIG. 1 – *Un exemple d'arbre de classification.*

Rentrons maintenant dans le vif du sujet et considérons une illustration réelle issue d'un travail de thèse citée dans les références (M. Ouali-Allah). Il s'agit d'une partie d'une enquête d'opinion réalisée en 1989 par l'institut d'Agoramétrie qui dépendait du consortium CEA-EDF. L'objet de l'enquête globale portait sur les conflits qui pouvaient alors agiter l'opinion publique. Pour la partie concernée de l'enquête où un échantillon de 500 individus ont été intérrogés, 19 questions, chacune associée à un homme politique, de la forme suivante, ont été introduites :

"*Souhaitez-vous voir jouer un rôle important à tel homme politique*"

Les codes des résultats à la question posée sont :

1. OUI
2. NON
3. SANS RÉPONSE

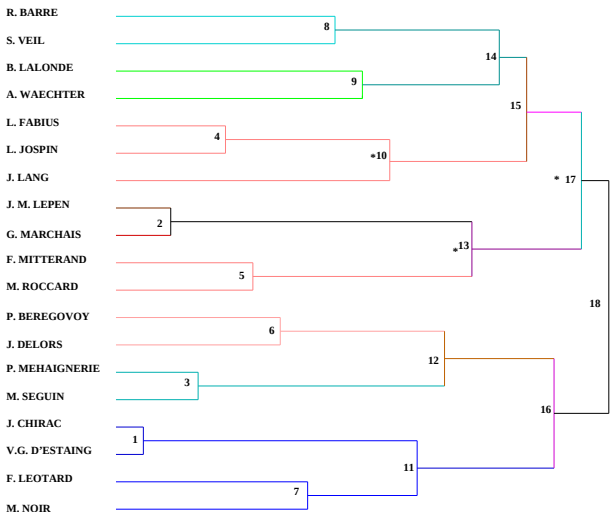

FIG. 2 – *Un premier arbre organisant l'opinion politique dans la France de 1989.*

Chaque question définit ainsi une variable qualitative à 3 valeurs (on dit encore modalités ou catégories). Le plus classique consiste à ne considérer aucune structure sur l'ensemble des valeurs de la variable. Il s'agit dans ce cas d'une variable qualitative *nominale*. Une interprétation plus riche résulte d'une similarité ordinale, pouvant être posée par l'expert, sur l'ensemble des valeurs de la variable. Il s'agira alors d'une variable qualitative *préordonnance*. Nous avons retenu le préordre total suivant sur l'ensemble des paires de valeurs :

$$\{1,2\} < \{2,3\} < \{1,3\} < \{1,1\} \sim \{2,2\} \sim \{3,3\} \quad (1)$$

où les catégories "OUI" et "NON" sont les plus dissemblables et où, la ressemblance entre une catégorie et elle même est maximale quelle que soit la catégorie. On comprendra également la position des autres paires de catégories.

C'est la même méthode, celle précisément de l'$AVLR$ qui a été appliquée pour chacun des deux codages. Dans le premier résultat, où le codage classique a été utilisé, la droite et la gauche politiques ont certes une certaine influence dans les regroupements ; mais, elles ne se séparent pas vraiment. La tendance qui domine dans les associations ou séparations est celle d'une certaine rigueur, voire de sectarisme, face à une certaine ouverture, autour du patriotisme allant jusqu'au nationalisme, d'attachement à la terre, ... Ainsi voit-on des associations telles que "Lepen-Marchais", "Mitterand-Rocard", "Méhaignerie-Séguin", "Beregovoy-Delors", "Veil-Barre", "Lalonde-Waechter", ...

Considérons à présent le second résultat où le codage en termes de *préordonnance* a été utilisé. Dans ce cas la droite et la gauche se séparent clairement. D'autre part, au sein de chacune de ces deux formations politiques on distingue la vieille et traditionnelle génération et celle montante. Ainsi, pour la droite, la génération la plus classique est représentée par {J.Chirac,

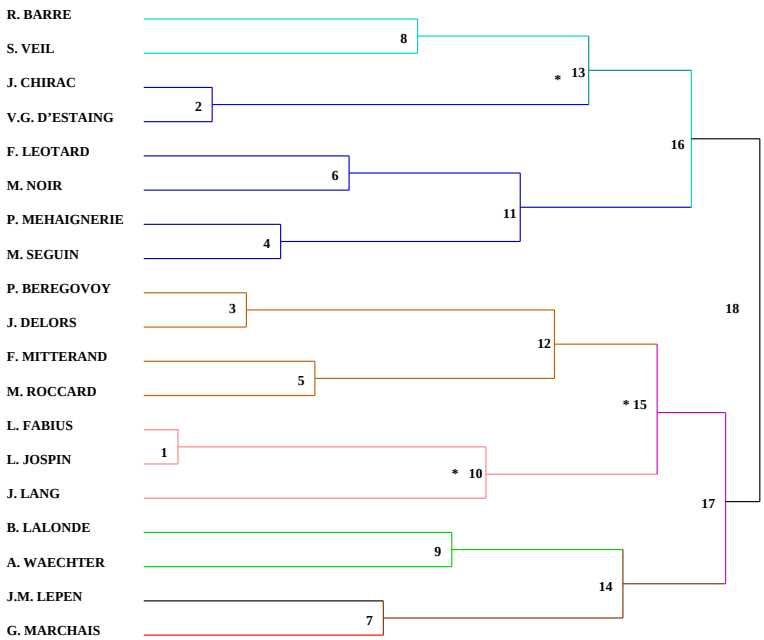

FIG. 3 – *Un deuxième arbre organisant l'opinion politique dans la France de 1989.*

V.G. D'estaing, R. Barre, S. Veil} ; celle plus jeune et avec un accent nouveau, par {F. Léotard, M. Noir, M. Séguin, P. Méhaignerie}. Pour la gauche, la génération classique est représentée par {F. Mitterand, M. Roccard, P. Beregovoy, J. Delors} et celle, jeune et montante par {L. Fabius, L. Jospin, J. Lang}. Une dernière classe {J.M. Lepen, G. Marchais, B. Lalonde, J. Waechter} regroupe les aspects les plus extrêmes du paysage politique français où le nationalisme et l'attachement à la terre sont présents.

La comparaison de ces deux résultats illustre toute l'importance de la prise en compte d'une structure relationnelle au niveau de l'ensemble des valeurs d'une même variable descriptive.

2.2 Le tableau des données et son analyse par la classification

2.2.1 Le tableau des données

Dans la très grande majorité des cas, mais pas toujours, les données peuvent être représentées au moyen d'un tableau (voir FIG. 4).

La description peut concerner un ensemble $\mathcal{O}$ d'objets élémentaires ou bien, un ensemble $\mathcal{C}$ de catégories (on dit encore "concepts"). Les colonnes du tableau des données sont indexées par l'ensemble $\mathcal{A}$ des attributs (variables) de description. En supposant que le nombre d'attributs est p et que le nombre d'objets (resp., catégories) est n, on peut noter :

$$\mathcal{A} = \{a^j | 1 \leq j \leq p\}$$

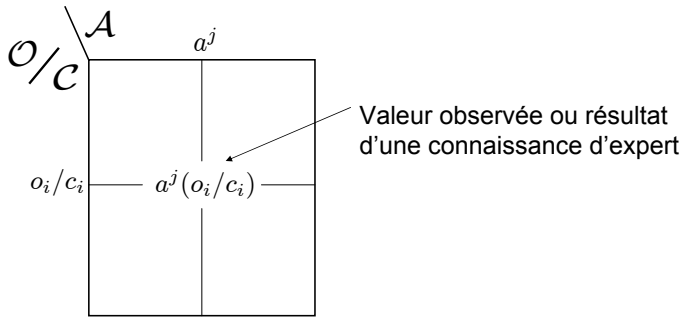

FIG. 4 – *Le tableau des données.*

$$\mathcal{O} = \{o_i | 1 \leq i \leq n\}\ , \mathcal{C} = \{c_i | 1 \leq i \leq n\}$$

Si la description concerne un ensemble $\mathcal{O}$ d'objets (resp., un ensemble $\mathcal{C}$ de catégories), la i-ème ligne du tableau est indexée par l'objet élémentaire o_i (resp., par la catégorie c_i), $1 \leq i \leq n$. De toute façon, la j-ème colonne du tableau est indexée par l'attribut a^j, $1 \leq j \leq p$. D'autre part, quel que soit l'ensemble décrit (un ensemble d'objets ou un ensemble de catégories), à l'intersection de la i-ème ligne et de la j-ème colonne se trouve une valeur observée ou une connaissance d'expert de l'attribut a^j sur l'objet o_i s'il s'agit d'une description d'objets, ou sur la catégorie c_i, s'il s'agit d'une description de catégories.

Dans le cas des données ci-dessus, il s'agit de la description d'un ensemble $\mathcal{O}$ d'objets élémentaires qui est défini par l'échantillon formé des 500 individus. D'autre part, comme cela a déjà été exprimé, on dispose de 19 variables qualitatives. La valeur $a^j(o_i)$ est représentée par le code de la réponse du i-ème individu à la j-ème question, $1 \leq i \leq 500$, $1 \leq j \leq 19$. Comme nous l'avons vu, ce code est 1, 2 ou 3.

Il est important de noter que le tableau des données ne peut en aucun cas contenir l'information concernant la structure relationnelle dont se trouve munie l'ensemble des valeurs d'une même variable descriptive. Ainsi dans l'exemple ci-dessus, on précise la structure *préordonnance* qui est supposée sur l'ensemble des valeurs d'une variable qualitative préordonnance donnée. D'ailleurs, même dans le cas d'une variable qualitative nominale, le calcul tient compte de l'absence de structure sur l'ensemble des modalités de la variable.

Les données de connaissance qui sont parfois appelées *"symboliques"* peuvent parfaite-

ment rentrer dans le moule précédent et être traitées efficacement par l'$AVLR$. D'ailleurs, nous présenterons en section 5 notre expression formelle des grands types logico-mathématiques de données.

2.2.2 Analyse et réorganisation par la classification hiérarchique du tableau des données

Les classifications proposées ci-dessus sont des classifications hiérarchiques d'ensembles de variables. Il est clair que - pour chacun des deux codages proposés ci-dessus (qualitatif *nominal* et qualitatif *préordonnance*) - nous aurions pu poursuivre par une classification de l'ensemble des 500 individus. Peut-être que nous pouvons évoquer ici la réalisation de ce type de classification dans le cas d'un codage pouvant être apparenté à celui *"préordonnance"*. Il s'agissait d'un ensemble d'objets définis par des séquences protéiques. Une séquence protéique est une suite d'acides aminés et peut être représentée comme un mot dans un alphabet à 20 lettres. La description relationnelle suppose précisément un graphe valué complet sur l'ensemble des 20 lettres.

Une étape ultime - nous y reviendrons - consistera à situer des classes de variables par rapport à des classes d'objets. En effet, la réorganisation des lignes et des colonnes d'un tableau de données conformément à un couple d'arbres de classification (issus de la CAH), le premier sur l'ensemble des attributs descriptifs et le second sur l'ensemble des objets, est très riche d'enseignement sur le plan de l'interprétation des tendances comportementales. Considérons à cet égard une illustration schématique très simplifiée où on considère un très petit tableau de données de présence-absence, décrivant un ensemble de 5 objets au moyen se 4 attributs booléens (voir FIG. 5). Le croisement des deux arbres de classification (sur l'ensemble $\mathcal{A}$ des attributs de description et sur l'ensemble $\mathcal{O}$ des objets décrits) donne la partie droite de la figure, où une case noircie indique que l'attribut est à $VRAI$. La reconnaissance de niveaux intéressants (nous disons "significatifs") sur chacun des deux arbres permet de situer des classes cohérentes d'objets par rapport à des classes cohérentes de variables. Nous avons repéré sur chacun des deux arbres un niveau "significatif". Il s'agit du deuxième (resp., troisième) niveau pour l'arbre de classification sur $\mathcal{A}$ (resp., $\mathcal{O}$). Ces niveaux sont indiqués sur la partie droite de la figure. La partition déterminée sur l'ensemble $\mathcal{A}$ est $\pi = \{\{a^1, a^2\}, \{a^3, a^4\}\}$, celle déterminée sur l'ensemble $\mathcal{O}$ est $\chi = \{\{o_1, o_2, o_4\}, \{o_3, o_5\}\}$. Si la classe $\{o_1, o_2, o_4\}$ est en correspondance avec la classe $\{a^1, a^2\}$, celle $\{o_3, o_5\}$, l'est avec la classe $\{a^3, a^4\}$. Pour ce qui est de la première correspondance on constatera que l'association la plus forte est relative à la sous classe $\{o_1, o_2\}$. Pour ce qui est de la deuxième correspondance, on peut aussi constater que l'association la plus forte concerne la sous-classe $\{o_3\}$.

Il y a bien sûr de nombreuses méthodes de classification conjointe des lignes et des colonnes d'un tableau de données booléennes ou numériques, où on cherche à optimiser au mieux un critère objectif. Cependant, il faut être conscient que la *philosophie* qui préside au croisement de deux arbres condensés à leurs niveaux les plus "significatifs" *n'est pas vraiment la même*. Deux questions importantes de complexité subsistent : la première de nature statistique et la seconde de nature formelle. Pour la première, il s'agit de savoir comment gérer un tel croisement en cas d'un "gros" tableau de données. D'autre part, comment gérer le cas où les

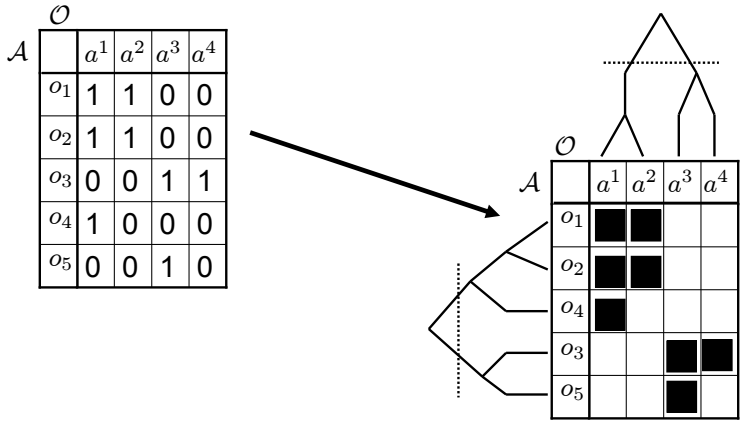

FIG. 5 – *La source (le tableau des données) et le but (deux arbres de classification croisés).*

variables descriptives sont complexes ; par exemple le cas où l'ensemble des valeurs d'une même variable est muni d'une structure relationnelle.

3 Le principe algorithmique et les éléments d'une CAH

3.1 Une expression formalisée de la démarche d'Adanson

La donnée est un couple de la forme (E, δ) où E est l'ensemble à organiser par la Classification Ascendante Hiérarchique. Nous avons déjà vu que E peut être un ensemble $\mathcal{A}$ d'attributs (de variables) de description, un ensemble $\mathcal{O}$ d'objets élémentaires ou un ensemble $\mathcal{C}$ de catégories. δ correspond à une notion de dissimilarité quantifiée numériquement entre parties disjointes de E.

L'état initial se trouve défini par l'ensemble des parties singletons de E qu'on peut noter $P_0 = \{\{e\}|e \in E\}$. P_0 détermine une partition en classes à un élément de E. La progression de l'algorithme est définie comme suit :

D'un niveau au suivant de l'arbre agréger les paires de classes
"les plus proches" au sens de δ

Dans cette expression on ne suppose pas nécessairement qu'il y a une seule fusion binaire de classes qui est opérée en passant d'un niveau de l'arbre au suivant. D'ailleurs, dans le schéma de la figure 1, trois fusions binaires de classes sont opérées "en même temps" pour passer du niveau 1 au niveau 2.

La condition d'arrêt est définie par l'obtention d'une seule classe qui regroupe tous les éléments de E.

On comprend dans ces conditions que la question fondamentale est de savoir comment élaborer l'indice numérique δ compte tenu de la nature de l'ensemble à organiser et de sa description. Il est clair que la première étape doit consister en la définition d'un indice de dissimilarité d entre éléments de E. Cette définition conduit à un tableau carré de dissimilarité de la forme :

$$\{d(x,y)|(x,y) \in E \times E\} \tag{2}$$

d est une fonction numérique à valeurs positives qui, dans le cas classique, est symétrique : $d(x,y) = d(y,x)$ pour tout $(x,y) \in E \times E$, pour laquelle on a : $d(x,x) = 0$ pour tout $x \in E$.

Maintenant, comment définir d compte tenu de la nature de E et des caractéristiques mathématiques et statistiques du tableau des données (voir ci-dessus). D'autre part, comment opérer le passage *crucial* entre l'indice d entre éléments de E et celui δ entre parties disjointes de E.

3.2 Une expression formelle plus générale et formule de réactualisation

La donnée est maintenant un triplet de la forme (E, μ_E, d) où, avec les mêmes notations que ci-dessus, $\mu_E = \{\mu_x | x \in E\}$ se trouve défini par un système de poids ponctuels affectés aux éléments de E. La dissimilarité d sur E, représentée toujours au moyen du tableau $\{d(x,y)|(x,y) \in E \times E\}$, fait généralement intervenir ces poids.

Il s'agit maintenant pour pouvoir construire une CAH de passer du triplet précédent à un triplet de la forme : $(\mathcal{P}(E), \mu_\mathcal{P}, \delta)$ où $\mathcal{P}(E)$ désigne l'ensemble des parties de E et où $\mu_\mathcal{P}$ désigne l'extension additive de μ_E sur $\mathcal{P}(E)$. Plus précisément,

$$\forall Z \in \mathcal{P}(E)\ ,\ \mu_Z = \sum_{z \in Z} \mu_z \tag{3}$$

Comme ci-dessus, δ définit une notion de dissimilarité sur $\mathcal{P}(E)$; mais dépendant de $\mu_\mathcal{P}$. δ se présente sous la forme de l'application suivante :

$$\delta : \Big(\mathcal{P}(E) \times \mathcal{P}(E)\ ,\ \mu_\mathcal{P}\Big) \longrightarrow \mathcal{R}_+ \tag{4}$$

où $\mathcal{R}_+$ désigne l'ensemble des nombres réels positifs. Plus précisément et de façon assez générale - pour couvrir l'ensemble de tous les critères de fusion de paires de classes qui ont pu être proposés - la fonction δ peut s'écrire :

$$\begin{aligned} &\forall (X,Y) \in \mathcal{P}(E) \times \mathcal{P}(E)\ , \\ &\delta(X,Y) = f[\{d(x,y)|(x,y) \in (X \cup Y) \times (X \cup Y)\}\ ,\ \{\mu_z | z \in X \cup Y\}] \end{aligned} \tag{5}$$

où f est une fonction numérique à valeurs positives dépendant de deux arguments. Le premier est le tableau des dissimilarités sur l'union $X \cup Y$ des deux classes X et Y à comparer et le second, est le système des masses ponctuelles sur $X \cup Y$. Cette équation exprime que les classes formées déjà en dehors de $X \cup Y$ n'interviennent pas dans la comparaison entre X et Y.

On peut voir la construction ascendante hiérarchique d'un arbre de classification comme

l'évolution d'un système. Si k est un niveau donné de l'arbre, nous caractérisons l'état du système par le couple (T_k, μ^k), où T_k est la matrice des dissimilarités δ entre classes déjà formées au niveau k et où μ^k est la mesure (telle que $\mu_{\mathcal{P}}$ ci-dessus) sur l'ensemble de ces classes. L'état initial (T_0, μ^0) est défini par la matrice T_0 des indices δ entre classes singletons (comprenant chacune exactement un élément) et par le système initial de poids μ^0 où chaque classe singleton $\{x\}$ est affectée par le poids de l'élément x concerné, $x \in E$. En désignant par l le dernier niveau de l'arbre ($l = 4$ dans l'exemple de la figure 1), les états du système pour $0 \leq k \leq l-1$ ont à être considérés. Dans ces conditions, la formule suivante appelée "formule de réactualisation" est de la première importance :

$$(T_{k+1}, \mu^{k+1}) = \varphi(T_k, \mu^k) \tag{6}$$

où φ est une fonction à déterminer compte tenu de l'expression spécifique de l'indice δ choisi de comparaison entre classes (critère de fusion). Une telle équation dite aussi formule de "récurrence" permet après la reconnaissance des paires de classes également les plus proches - et donc à fusionner - de déterminer l'état du système au niveau $k+1$, sachant son état au niveau k, $0 \leq k \leq l-1$. Cette formule est très utilisée pour une forme d'implémentation de l'algorithme de classification ascendante hiérarchique.

4 L'Analyse de la Vraisemblance des Liens Relationnels (*Une présentation à partir d'exemples*)

4.1 Caractéristiques principales de la méthode et principe de l'indice de fusion entre classes

Trois caractéristiques fondamentales conditionnent cette méthodologie :

1. Représentation ensembliste puis relationnelle des variables (attributs) de description
2. Notion très générale de similarité se référant à une échelle probabiliste établie par rapport à une *hypothèse d'absence de liaison* entre les variables (attributs) de description "*respectant*" les distributions marginales observées des différentes variables
3. Guidage statistique dans l'interprétation de l'arbre de classification, conduisant à la détermination - au moyen d'un critère objectif - d'agrégats "intéressants" et de partitions (classifications) "intéressantes"

La représentation ensembliste et relationnelle des variables descriptives entraîne une extrême généralité de l'approche par rapport aux structures des données les plus complexes. Nous le mentionnerons plus explicitement dans la section traitant des grands types mathématico-logiques de données. La notion de similarité probabiliste va rejoindre la philosophie de la théorie de l'information ; mais où, les évènements observés sont des valeurs d'indices de similarité entre les structures de données à comparer. Les partitions intéressantes vont se produire aux niveaux que nous dirons "significatifs" - dans un sens que nous préciserons - de l'arbre de classification. Les agrégats "intéressants" vont correspondre à des noeuds "significatifs" - dans un sens que nous préciserons - de l'arbre de classification.

Il y a une très forte cohérence dans la suite des différentes étapes de la méthode de classification ascendante hiérarchique de l'$AVLR$. Nous allons commencer par considérer une articulation fondamentale concernant le critère (indice) de fusion des paires de classes. Nous allons en donner le principe dans le cadre d'une illustration géométrique très simple alors que dans notre cas, comme nous venons de l'exprimer, c'est une représentation ensembliste et relationnelle des variables descriptives qui prévaut. Il s'agit dans l'exemple, de la classification d'un nuage de points unidimensionnel (voir FIG. 6). Ayons à l'esprit que l'exigence première d'une *CAH* consiste à établir un ordre des agrégations et que c'est un critère *numérique* qui permet de sélectionner à chaque étape la "meilleure" agrégation (ou les également "meilleures" agrégations). C'est de ce critère numérique dont il s'agit. Commençons par poser un indice δ classique et naturel ici, défini par la distance *minimale* et imaginons qu'à une étape donnée de l'algorithme on ait à choisir parmi deux agrégations candidates : $\mathcal{C}_1$ et $\mathcal{C}_2$ d'une part et $\mathcal{D}_1$ et $\mathcal{D}_2$ d'autre part. En se limitant au critère défini, compte tenu de ce que $\delta(\mathcal{D}_1, \mathcal{D}_2)$ est strictement inférieur à $\delta(\mathcal{C}_1, \mathcal{C}_2)$, la première agrégation à choisir est celle de $\mathcal{D}_1$ et $\mathcal{D}_2$. Cependant, on peut remarquer que les classes $\mathcal{D}_1$ et $\mathcal{D}_2$ sont sensiblement *plus denses* que celles $\mathcal{C}_1$ et $\mathcal{C}_2$. Or, il est "normal" que les deux extrémités mutuellement les plus voisines soient plus proches pour deux classes *fortement denses* $\mathcal{D}_1$ et $\mathcal{D}_2$ que pour deux classes *faiblement denses* $\mathcal{C}_1$ et $\mathcal{C}_2$. Dans ces conditions, pour le critère de l'*AVLR*, la fusion choisie est celle de $\mathcal{C}_1$ et $\mathcal{C}_2$. Ainsi, c'est l'*exceptionnalité* de la petitesse de δ qui guide vers le "meilleur" choix. On se rend compte qu'ainsi, nous rejoignons la philosophie de la théorie de l'information au niveau des relations observées.

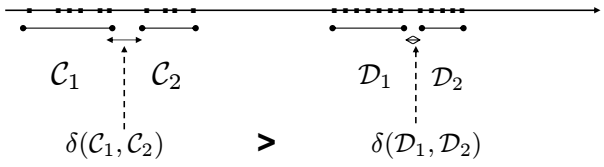

FIG. 6 – *Principe de l'agrégation AVL.*

Considérons à présent une deuxième illustration géométrique supposant un nuage de points planaire (voir FIG. 7). $\mathcal{C}_1$ et $\mathcal{C}_2$ sont deux classes *fortement denses* ; alors que $\mathcal{D}_1$ et $\mathcal{D}_2$ sont deux classes *faiblement denses*. Pour le critère de la distance minimale δ on a clairement

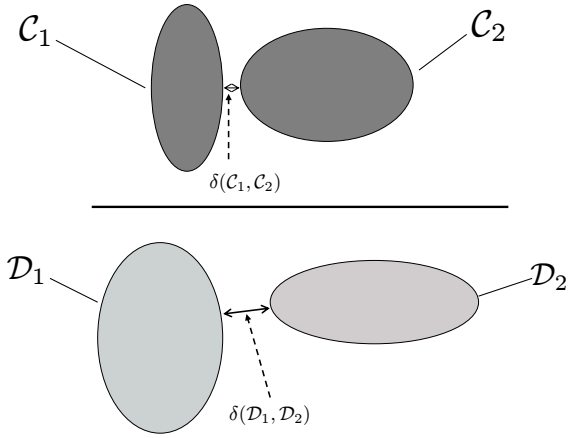

FIG. 7 – *Comparaison entre deux dissimilarités entre classes.*

$\delta(\mathcal{C}_1, \mathcal{C}_2) < \delta(\mathcal{D}_1, \mathcal{D}_2)$. Avec le même raisonnement intuitif que ci-dessus, entre les deux fusions ; celle de $\mathcal{C}_1$ et $\mathcal{C}_2$ ou celle de $\mathcal{D}_1$ et $\mathcal{D}_2$, $AVLR$ choisit celle de $\mathcal{D}_1$ et $\mathcal{D}_2$.

Profitons de signaler ici que l'effet bien connu de "chaînage" du critère de la distance minimale, disparaît dès lors qu'on lui substitue la vraisemblance de la petitesse de la valeur de cette distance minimale. D'autre part et enfin, il est clair qu'on peut prendre d'autres indices de base (nous disons "bruts") que la distance minimale et suivre la même démarche.

4.2 Les comparaisons par paires dans la méthode $AVLR$ en cas de données booléennes

4.2.1 Représentation ensembliste de l'ensemble des attributs de description

Nous noterons de façon conforme à ci-dessus par $\mathcal{A}$ l'ensemble des attributs booléens de description et par $\mathcal{O}$ l'ensemble des objets décrits. Nous représentons un attribut booléen a ($a \in \mathcal{A}$) par le sous ensemble $\mathcal{O}(a)$ des objets où il est à $VRAI$ (où il est présent) (voir FIG. 8). Ainsi, l'ensemble de représentation est l'ensemble des parties de l'ensemble $\mathcal{O}$ des objets. C'est le cas de représentation le plus simple puisque l'attribut booléen définit une variable relationnelle unaire. Signalons ici qu'un autre cas très important de variable relationnelle unaire est fourni par la variable quantitative qui elle, définit une valuation sur l'ensemble $\mathcal{O}$ des objets.

Il est clair que s'il s'agit d'attributs de même type dont l'ensemble des valeurs peut être structuré de façon plus ou moins complexe, la représentation fidèle d'un même attribut ne peut plus se faire au niveau de $\mathcal{O}$. Il faudra travailler à un niveau supérieur : l'ensemble des paires d'objets ou celui des couples de paires d'objets, ou ...

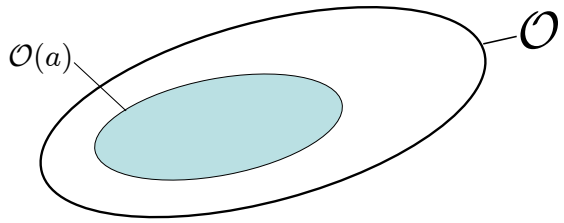

FIG. 8 – *Représentation d'un attribut booléen.*

La figure 9 donne l'exemple de la représentation de l'ensemble des attributs de description dans le cas d'un petit tableau de données booléennes (10 objets et 8 attributs). Il s'agit de façon générale d'un échantillon d'éléments dans l'ensemble des parties de $\mathcal{O}$.

4.2.2 Comparaison entre attributs booléens

En vue de l'organisation par la classification de l'ensemble $\mathcal{A}$ des attributs booléens considérons le problème de la comparaison deux à deux de ces attributs booléens. À cet effet, commençons par considérer une paire $\{a, b\}$ d'attributs faisant partie de $\mathcal{A}$. La représentation ensembliste conduit à une paire de parties de $\mathcal{O}$ de la forme $\{\mathcal{O}(a), \mathcal{O}(b)\}$. La situation relative entre $\mathcal{O}(a)$ et $\mathcal{O}(b)$ est schématisée dans la figure 10 où on introduit les paramètres s, u, v et t. On a avec des notations que l'on comprend :

$$\begin{cases} s = n(a \wedge b) & = card(\mathcal{O}(a) \cap \mathcal{O}(b)) \\ u = n(a \wedge \neg b) & = card(\mathcal{O}(a) \cap \mathcal{O}(\neg b)) \\ v = n(\neg a \wedge b) & = card(\mathcal{O}(\neg a) \cap \mathcal{O}(b)) \\ t = n(\neg a \wedge \neg b) & = card(\mathcal{O}(\neg a) \cap \mathcal{O}(\neg b)) \end{cases}$$

On a $s + u + v + t = n$. De nombreux indices (nous dirons aussi "coefficients") de similarité (nous dirons aussi d'"association") entre deux attributs booléens ont été proposés. Ils se présentent tous respectivement sous la forme, de fonctions des trois paramètres s, u et v, strictement croissantes par rapport à s, symétriques en u et v et décroissantes par rapport à u. Dans ces fonctions on cherche d'une façon ou d'une autre et de façon implicite à neutraliser l'influence dans la ressemblance des tailles $n(a) = card(\mathcal{O}(a))$ et $n(b) = card(\mathcal{O}(b))$. Ainsi en est-il du fameux indice de Jaccard (1908) qui se met sous la forme $s/(s + u + v)$ que nous

	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
o_1	0	0	1	1	0	1	0	1
o_2	0	0	0	1	0	0	1	1
o_3	1	1	0	1	1	0	0	0
o_4	0	0	1	0	0	0	1	1
o_5	1	0	1	0	1	0	0	0
o_6	1	0	0	1	1	0	0	0
o_7	1	0	0	0	1	1	1	0
o_8	0	0	1	1	0	0	1	1
o_9	1	0	0	1	1	1	0	0
o_{10}	1	1	0	1	0	0	1	1

$O(a^1) = \{o_3, o_5, o_6, o_7, o_9, o_{10}\}$
$O(a^2) = \{o_3, o_{10}\}$
$O(a^3) = \{o_1, o_4, o_5, o_8\}$
$O(a^4) = \{o_1, o_2, o_3, o_6, o_8, o_9, o_{10}\}$
$O(a^5) = \{o_3, o_5, o_6, o_7, o_9\}$
$O(a^6) = \{o_1, o_7, o_9\}$
$O(a^7) = \{o_2, o_4, o_7, o_8, o_{10}\}$
$O(a^8) = \{o_1, o_2, o_4, o_8, o_{10}\}$

FIG. 9 – *Représentation de la description d'un tableau de données booléennes.*

considérerons ci-dessous pour notre illustration. Notre approche constructive est différente. Nous commençons bien par le choix initial d'une fonction de similarité. Mais, compte tenu de l'invariance du résultat par rapport à ce choix, nous partirons le plus "naturellement" de l'indice "brut" de ressemblance s qui représente le nombre d'objets où les deux attributs a et b sont à $VRAI$. Il s'agit alors d'évaluer la grandeur relative de s compte tenu du contexte défini "localement" par le couple $(n(a), n(b))$. Cette évaluation se fera par rapport à une *Hypothèse d'Absence de Liaison* ($\mathcal{H}.\mathcal{A}.\mathcal{L}.$) où au couple de parties observées $(\mathcal{O}(a)), \mathcal{O}(b)))$, on associe un couple de parties aléatoires et indépendantes $(\mathcal{O}(a^*)), \mathcal{O}(b^*))$ "respectant" respectivement les cardinalités $n(a)$ et $n(b)$. Différents modèles aléatoires peuvent être considérés. La valeur observée s est alors située par rapport à la distribution probabiliste de l'indice brut aléatoire $S = card(\mathcal{O}(a^*)) \cap \mathcal{O}(b^*))$. Dans ces conditions, la valuation de la "ressemblance" entre a et b est d'autant plus grande que s apparaît "invraisemblablement" grand eu égard à la distribution de S. On introduit ainsi une notion de "vraisemblance" dans la notion de "ressemblance". Dans la figure 11 qui schématise la construction, $P_l(a, b)$ est l'indice probabiliste d'association.

Signalons ici l'indice statistiquement normalisé suivant qui permet le calcul de l'indice probabiliste via la fonction de répartition de la loi normale centrée réduite Φ :

$$\begin{aligned} Q(a,b) &= \big(s - \mathcal{E}(S)\big) / \sqrt{var(S)} \\ P_l(a,b) &= \Phi\big(Q(a,b)\big) \end{aligned} \tag{7}$$

où $\mathcal{E}(S)$ et $var(S)$ désignent l'espérance mathématique et la variance de l'indice brut aléatoire S.

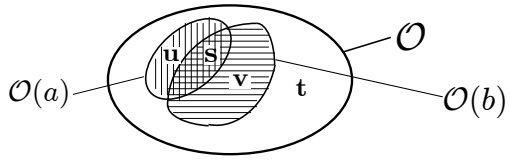

FIG. 10 – *Représentation ensembliste de la comparaison entre deux attributs booléens.*

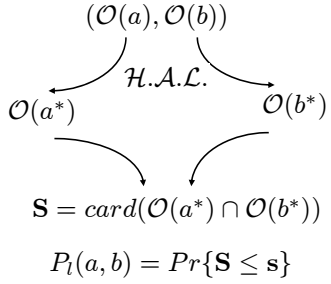

FIG. 11 – *Indice probabiliste local.*

L'indice $P_l(a,b)$ a un caractère *local*, circonscrit aux deux attributs à comparer. Le contexte définitif de comparaison deux à deux pour l'établissement de l'indice probabiliste sur $\mathcal{A}$ sera *global*.

Les coefficients d'association (indices de similarité) que nous considérons ont essentiellement un caractère symétrique. Ainsi, relativement à un couple d'attributs booléens (a,b), l'indice $P_l(a,b)$ cherche à "mesurer" le *degré d'équivalence* entre a est à $VRAI$ et b est à $VRAI$. Régis Gras a eu l'idée d'adapter ce type d'indice de la *Vraisemblance du Lien* pour "mesurer" une relation dissymétrique de la forme "combien a à $VRAI$ implique b à $VRAI$". Il en est résulté un ensemble important de travaux - évoqués dans les références - autour de la recherche dans les données de structures implicatives.

$\mathcal{A}$	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
a^1	6	0.89	0.27	0.58	0.95	0.74	0.38	0.15
a^2	2	2	0.45	0.84	0.74	0.55	0.74	0.74
a^3	1	0	4	0.44	0.38	0.66	0.68	0.88
a^4	4	2	2	7	0.51	0.65	0.51	0.75
a^5	5	1	1	3	5	0.82	0.24	0.06
a^6	2	0	1	2	2	3	0.54	0.54
a^7	2	1	2	3	1	1	5	0.92
a^8	1	1	3	4	0	1	4	5

$\mathbf{s}\backslash P_l$

FIG. 12 – *Tableau des similarités pour le tableau des données booléennes ci-dessus.*

Exemple du tableau précédent des données booléennes Le tableau de la figure 12 est relatif aux données booléennes du tableau de la figure 9. Les valeurs de l'indice brut s se trouvent sous la diagonale principale (au sens large). Ainsi, $s(a^4,a^7)=3$, $s(a^5,a^6)=2$, ... Le j-ème élément de la diagonale tel que $s(a^j,a^j)$ n'est autre que $n(a^j)=card(\mathcal{O}(a^j))$, $1\leq j\leq 8$. Les valeurs d'un indice probabiliste *local* P_l - conformément à un modèle aléatoire d'absence de liaison - sont placées strictement en dessous de la diagonale principale. On a ainsi $P_l(a^4,a^7)=0.51$ et $P_l(a^5,a^6)=0.82$.

On constate qu'il peut exister des inversions entre d'une part l'indice brut s et d'autre part, l'indice probabiliste *local* P_l. On a par exemple :

$$s(a^4,a^7)=3>s(a^5,a^6)=2 \text{ et } P_l(a^4,a^7)=0.51<P_l(a^5,a^6)=0.82$$

Des inversions peuvent même avoir lieu entre l'indice de Jaccard J mentionné ci-dessus et l'indice probabiliste *local* P_l. On a ainsi :

$$J(a^1, a^6) = \frac{2}{7} < J(a^4, a^7) = \frac{3}{9} \text{ et } P_l(a^1, a^6) = 0.74 > P_l(a^4, a^7) = 0.51$$

Réduction *globale* des similarités Le contexte *local* permet avec P_l des comparaisons bien différenciées pourvu que le nombre d'objets ne soit pas "trop grand". Autrement, P_l a tendance à tendre soit vers 1 (resp., 0) en cas d'association positive (resp., négative) par rapport à l'indépendance probabiliste. C'est qu'en fait, pour l'organisation classificatoire de l'ensemble $\mathcal{A}$ des attributs booléens, le problème consiste en la comparaison *relative* deux à deux des éléments de $\mathcal{A}$. C'est pour cette raison que dans le contexte *global* on rapporte la comparaison entre deux attributs aux comparaisons mutuelles deux à deux entre attributs. Dans ces conditions, à la suite $(a^1, a^2, ..., a^j, ..., a^p)$ des attributs observés on associe dans une hypothèse d'absence de liaison respectant de façon marginale la configuration du tableau des données, une suite $(a^{1*}, a^{2*}, ..., a^{j*}, ..., a^{p*})$ d'attributs aléatoires mutuellement indépendants. L'indice $Q(a^j, a^k)$ (cf. (7)) entre deux attributs de $\mathcal{A}$, statistiquement *localement* normalisé, est lui même à nouveau *globalement* normalisé par rapport à sa distribution empirique sur l'ensemble $\mathcal{P}_2(\mathcal{A})$, $1 \leq j < k \leq p$. Ainsi, on substitue à l'indice $Q(a^j, a^k)$, celui $Q_g(a^j, a^k)$ où (voir FIG. 13) où $m_e(Q)$ et $var(Q)$ sont respectivement la moyenne et la variance empirique de $Q(a^j, a^k)$ sur $\mathcal{P}_2(\mathcal{A})$. C'est la loi asymptotiquement normale de $Q_g(a^{j*}, a^{k*})$ qui conduit à l'indice probabiliste *global* $P_g(a^j, a^k)$, $1 \leq j < k \leq p$. La figure 13 donne le schéma général de la procédure.

$$(a^1, a^2, ..., a^j, ..., a^p) \xrightarrow[\mathcal{H}.\mathcal{A}.\mathcal{L}.]{} (a^{1*}, a^{2*}, ..., a^{j*}, ..., a^{p*})$$

$$\mathcal{P}_2(\mathcal{A}) = \{\{a^j, a^k\} | 1 \leq j < k \leq p\}$$

$$Q(a^j, a^k) \longleftarrow Q_g(a^j, a^k) = (Q(a^j, a^k) - moy_e(Q))/\sqrt{var_e(Q)}$$

$$m_e(Q) \text{ et } var_e(Q) \text{ sur } \mathcal{P}_2(\mathcal{A})$$

$$P_g(a^j, a^k) = Pr\{Q_g(a^{j*}, a^{k*}) \leq Q_g(a^j, a^k)\} = \Phi(Q_g(a^j, a^k))$$

FIG. 13 – *Similarité probabiliste globale.*

Cette référence à la loi normale est justifiée dans les références Daudé (1992) et Lerman (1984) de la section 7.2. L'approximation normale devient dans la plupart des cas, très bonne dès lors que la taille n de l'ensemble des objets dépasse l'ordre de la dizaine d'unités.

4.2.3 Comparaison entre objets décrits par des attributs

Il s'agit à présent d'adresser le problème de la classification de l'ensemble $\mathcal{O}$ des objets décrits par des attributs. Pour se fixer les idées on peut imaginer que les attributs sont booléens. Cependant, la procédure est d'une extrême généralité par rapport à la nature des attributs. Ainsi, comment effectuer de façon cohérente les comparaisons mutuelles entre objets pour aboutir au même type d'indice probabiliste que dans le cas de la comparaison entre attributs. La procédure se décompose en la suite des étapes suivantes :

(i) Contribution brute d'un attribut à la comparaison de deux objets Considérons le couple formé par un attribut a^j et une paire d'objets $\{o_i, o_{i'}\}$, $1 \leq j \leq p$, $1 \leq i < i' \leq n$. On définit la contribution "brute" $s^j(o_i, o_{i'})$ de l'attribut a^j à la ressemblance entre les deux objets o_i et $o_{i'}$. Un exemple en cas d'attributs booléens - que nous ne pouvons ici justifier - est donné par :

$$s^j(o_i, o_{i'}) = \frac{1}{p} - \frac{1}{2}(\eta_i^j - \eta_{i'}^j)^2 \tag{8}$$

où

$$\eta_i^j = \epsilon_i^j / \sqrt{(\sum_{1 \leq k \leq p} \epsilon_i^k)} \tag{9}$$

et où $\epsilon_i^j = 1(\text{resp.},0)$ selon que l'attribut a^j est présent (à $VRAI$) (resp., absent (à $FAUX$)) chez l'objet o_i.

(ii) Normalisation statistique de la contribution "brute" de similarité $s^j(o_i, o_{i'})$ est normalisé par rapport à la distribution statistique de la contribution "brute" du j-ème attribut sur l'ensemble $\mathcal{O} \times \mathcal{O}$ des couples d'objets. Il s'agit nommément de la distribution :

$$\{s^j(o_l, o_{l'}) | 1 \leq l, l' \leq n\}$$

La contribution normalisée de l'attribut a^j à la comparaison des deux objets o_i et $o_{i'}$ se met dans ces conditions sous la forme :

$$S^j(o_i, o_{i'}) = \big(s^j(o_i, o_{i'}) - m_e(s^j)\big) / \sqrt{var_e(s^j)} \tag{10}$$

où $m_e(s^j)$ et $var_e(s^j)$ sont la moyenne et la variance de la précédente distribution, $1 \leq i < i' \leq n$.

(iii) Somme des contributions normalisées On définit l'indice totalisant l'ensemble des contributions normalisées sous la forme :

$$S(o_i, o_{i'}) = \sum_{1 \leq j \leq p} S^j(o_i, o_{i'}) \tag{11}$$

$1 \leq i < i' \leq n$. Par rapport au cas dual de la construction d'un indice d'association entre attributs de description, on peut ici considérer qu'on se trouve au même niveau que celui de la définition du coefficient Q (cf. (7)). Il nous reste donc à procéder à la

(iv) Normalisation statistique *globale* de $S(o_i, o_{i'})$ par rapport à l'ensemble $P_2(\mathcal{O})$ des paires d'objets distincts L'indice ainsi normalisé conduit via la fonction de répartition de la loi normale centrée et réduite, à l'indice probabiliste de *vraisemblance du lien*. La référence à la loi normale dans l'hypothèse probabiliste d'absence de liaison mutuelle entre les différents attributs de description, pour la loi de l'indice aléatoire associé à $S(o_i, o_{i'})$ se justifie en faisant appel au théorème central limite. En effet, cette variable aléatoire est formée d'une somme de contributions normalisées aléatoires et indépendantes. Il importe dans ces conditions que le nombre p de variables de descriptions ne soit pas trop petit. Ce qui est le cas le plus souvent. Toutefois, quelle que soit la valeur de p, il n'est pas interdit de vouloir se positionner, pour la loi de l'indice aléatoire globalement normalisé dans l'hypothèse d'absence de liaison, relativement à la loi normale centrée et réduite.

Signalons ici que l'élaboration d'une matrice de similarités probabilistes sur un ensemble $\mathcal{C}$ de catégories décrits par des attributs booléens obéit à la même démarche que ci-dessus.

4.2.4 Cas général

L'établissement d'une matrice d'indices probabilistes de la vraisemblance du lien telle qu'elle a été introduite pour la classification d'un ensemble $\mathcal{A}$ d'attributs, respectivement pour la classification d'un ensemble $\mathcal{O}$ d'objets, dans le cas d'un tableau de données booléennes, a été étendu dans le cas d'un tableau très général de données (voir la sous section 2.2.1 et la section suivante). Cette généralisation mathématico-statistique a été validée sur le plan expérimental par de nombreuses, difficiles et très intéressantes applications. La figure 14 donne le schéma général. L'ensemble E à organiser par la classification peut être soit un ensemble $\mathcal{A}$ de variables descriptives, soit un ensemble $\mathcal{O}$ d'objets décrits (resp., un ensemble $\mathcal{C}$ de catégories décrites). Dans l'un ou l'autre de ces deux cas duaux, on aboutit à une matrice d'indices probabilistes $\{P(x,y)|\{x,y\} \in P_2(E)\}$ telle qu'elle est indiquée dans le schéma. $P_2(E)$ est l'ensemble des paires (parties à deux éléments) de E.

Nous proposons de substituer à la table d'indices probabilistes la table suivante que nous appelons "Matrice des dissimilarités "informationnelles"" :

$$\big\{\Delta(x,y) = -Log_2(P(x,y))|\{x,y\} \in P_2(E)\big\}$$

$-Log_2(P(x,y))$ est la quantité d'information de l'évènement dont la probabilité est $P(x,y)$. On remarquera que lorsque $P(x,y)$ varie entre 1 et 0, $\Delta(x,y)$ varie entre 0 et l'infini. La matrice précédente est ce que la méthode $AVLR$ peut transmettre aux algorithmes qui travaillent avec des dissimilarités.

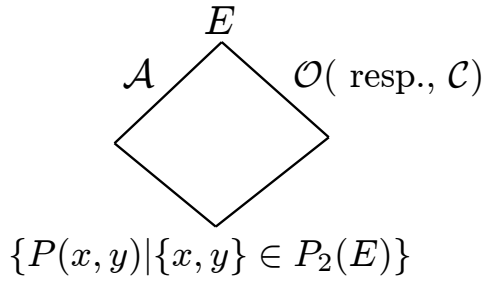

FIG. 14 – *Table de similarités probabilistes quel que soit l'ensemble organisé.*

4.3 Famille de critères de la vraisemblance du lien maximal et construction de l'arbre

Il s'agit maintenant, selon la démarche d'Adanson, de rapprocher des "petites familles" qui représentent dans notre formalisme des parties disjointes, des classes, de l'ensemble à organiser E. Nous avons vu (section 3.2) que cela passait par la définition d'un indice de dissimilarité - ou de façon équivalente - de similarité entre parties disjointes de E. E est maintenant muni d'un indice de similarité probabiliste (section 4.2.4 ci-dessus). Dans ces conditions, soient C et D deux parties disjointes (deux classes) de l'ensemble E. Nous allons partir d'un indice de comparaison qui est défini par le lien "maximal" (on dit encore lien "simple" ou "single linkage") mais dans le contexte de la matrice des indices probabilistes établie. Plus précisément, nous définissons :

$$p(C,D) = max\{P(c,d)|(c,d) \in C \times D\} \tag{12}$$

Conformément au principe de fusion des classes dans l'$AVLR$ (voir section 4.1), nous associons dans une hypothèse d'absence de liaison, au couple de parties (C,D) un couple (C^*,D^*) de parties aléatoires indépendantes et formées respectivement d'éléments mutuellement indépendants. La forme pure du critère "vraisemblance du lien maximal" prend alors la forme :

$$P(C,D) = Pr\{p(C^*,D^*) \leq p(C,D)\} = \big(p(C,D)\big)^{(|C|\times|D|)} \tag{13}$$

où $|C|$ (resp., |D|) désigne le cardinal de C (resp., D). Une famille plus large et paramétrée de critères est définie par l'équation :

$$P_\epsilon(C,D) = \big(p(C,D)\big)^{(|C|\times|D|)^\epsilon} \tag{14}$$

Pour des raisons de discrimination au niveau du calcul numérique ; mais avec un résultat identique on utilise la fonction strictement croissante :

$$S_\epsilon(C,D) = -Log_2[-Log_2(P_\epsilon(C,D))] \tag{15}$$

Nous avions introduit cette famille à l'occasion de travaux avec F. Costa Nicolaü et H. Bacelar-Nicolaü. En effet, F. Costa Nicolaü a cherché à introduire une certaine souplesse en étudiant d'autres fonctions que celles associées à (12) ou (13). Remarquons que pour $\epsilon = 0$, on a le lien simple et que pour $\epsilon = 1$ on a la forme pure du critère de la vraisemblance du lien maximal. Signalons que dès que ϵ se détache de la valeur 0, l'effet bien connu de "chaînage" du "single linkage" disparaît. Une valeur très utilisée pour ϵ est 0.5.

Signalons enfin qu'on dispose clairement d'une formule de réactualisation pour S_ϵ.

4.3.1 Les deux arbres de classification issus de l'exemple

Relativement au tableau de données booléennes de la figure 9, nous avons obtenu par l'application de la méthode les deux arbres des figures 15 et 16. Le premier (FIG. 15) porte sur l'ensemble $\mathcal{A}$ des 8 attributs booléens et le second (FIG. 16) porte sur l'ensemble $\mathcal{O}$ des 10 objets. Dans un sens que nous préciserons ci-dessous, le niveau 5 est distingué dans le premier arbre (sur $\mathcal{A}$), comme étant un niveau "significatif". Le niveau 6 est également intéressant. Dans le second arbre (sur $\mathcal{O}$) c'est le niveau 9 qui est distingué comme "significatif". Nous préciserons également bientôt la notion de noeuds "significatif". Les noeuds significatifs sont marqués par une étoile $*$.

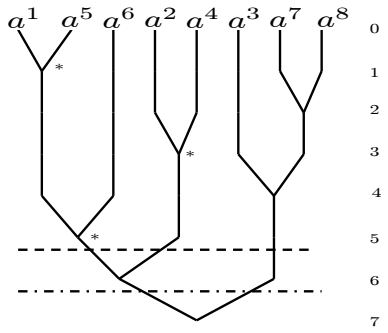

FIG. 15 – *Arbre de classification sur $\mathcal{A}$; niveaux et noeuds "significatifs".*

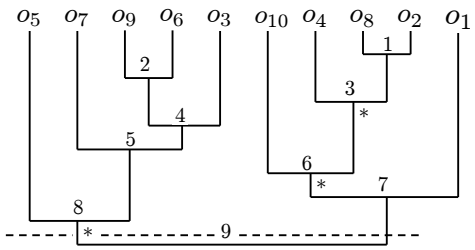

FIG. 16 – *Arbre de classification sur $\mathcal{O}$; niveaux et noeuds "significatifs".*

4.4 Niveaux et Noeuds les plus "significatifs" d'un arbre de classification

Commençons par donner un sens intuitif aux notions d'un niveau "significatif" et d'un noeud "significatif" d'un arbre de classification. Reprenons ici l'image donnée au paragraphe 3.2 de la construction de l'arbre en termes d'un processus de synthèse automatique. Un niveau donné, défini par un état du processus, détermine une partition (classification) sur l'ensemble total E. Un niveau "significatif" correspond à un état d'équilibre dans la synthèse automatique. Les classes obtenues qu'achèvent des noeuds qui sont en dessous (au sens large) du niveau concerné, sont ces familles qui, selon l'expression d'Adanson, forment un "tout dont les parties sont liées intimement", sur le plan de la cohérence, avions nous précisé. La liaison est d'autant plus intime sur le plan de la cohérence que le noeud est "significatif".

Il y a lieu maintenant de préciser comment objectivement, nous déterminons les niveaux et les noeuds les plus "significatifs" d'un arbre de classification. La base est un critère qui se présente comme un coefficient d'association obéissant au principe de la démarche $AVLR$, entre une partition et une similarité sur E. Ce coefficient a une nature combinatoire et non paramétrique. Le point de départ consiste en la représentation de la partition et de la similarité sur E. Elles seront vues comme deux relations binaires symétriques sur E. La représentation se fait au niveau de l'ensemble F des paires ou parties à deux éléments de E :

$$F = P_2(E) = \big\{\{x, y\} | x \in E, y \in E, x \neq y\big\}$$

Pour notre problème de détection de partitions intéressantes, posons :

$$\{\pi_1, \pi_2, ..., \pi_k, ..., \pi_l\}$$

la suite des partitions produites aux différents niveaux de l'arbre des classifications. Indiquons π_k sous la forme :

$$\pi_k = \{E_{k1}, E_{k2}, ..., E_{ki}, ..., E_{kc_k}\}$$

Nous représentons π_k au niveau de F par le sous ensemble des paires qu'elle réunit ; nommément :

$$R(\pi_k) = \sum_{1 \leq i \leq c_k} P_2(E_{ki}),$$

où la somme d'ensembles indique une réunion d'ensembles disjoints.

Prenons l'exemple du premier arbre de la figure 15 où $E = \mathcal{A}$. On a la partition du niveau 5, $\pi_5 = \big\{\{1,5,6\},\{2,4\},\{3,7,8\}\big\}$ qu'on représente comme suit :

$$R(\pi_5) = \big\{\{1,5\},\{1,6\},\{5,6\},\{2,4\},\{3,7\},\{3,8\},\{7,8\}\big\}$$

Précisons maintenant la construction du coefficient d'association entre une partition π et une similarité Q sur l'ensemble E. Nous proposons ici de prendre une similarité numérique. Cependant, il existe une version du critère où une similarité ordinale (préordonnance sur E) est prise en compte. La similarité numérique considérée dans nos programmes est celle Q_g pour la classification de l'ensemble $\mathcal{A}$ des variables descriptives et celle duale correspondante, obtenue après (11), pour la classification de l'ensemble $\mathcal{O}$ des objets.

Le point de départ consiste en la définition de l'indice "*brut*"

$$s(Q, \pi) = \sum_{\{x,y\} \in R(\pi)} Q(x, y) \tag{16}$$

qui représente la somme des similarités des paires réunies par la partition π. Introduisons ici l'ensemble $\mathcal{P}(n; t(\pi))$ qui est l'ensemble des partitions de E de même type $t(\pi)$ (i.e. ayant la même famille de cardinaux de classes) que la partition donnée π. L'*hypothèse d'absence de liaison* associe à la partition observée π une partition aléatoire π^* qui représente un élément aléatoire dans l'ensemble $\mathcal{P}(n; t(\pi))$ muni d'une probabilité uniformément répartie. Dans ces conditions, l'indice *brut aléatoire* $s(Q, \pi^*)$ suit asymptotiquement (l'approximation est excellente) une loi normale dont la moyenne M et la variance V sont calculées mathématiquement. Le critère (coefficient d'association) est alors calculé par la formule :

$$C(Q, \pi) = \frac{s(Q, \pi) - M}{\sqrt{V}} \tag{17}$$

Lorsque ce critère est appliqué à la suite décroissante en finesse, de l'arbre de classification, nous l'appelons "*Statistique globale des niveaux*". Précisément, nous considérons la distribution suivante de la *Statistique globale des niveaux* :

$$\big\{C(Q, \pi_i) | 1 \leq i \leq l\big\} \tag{18}$$

De la sorte un niveau "significatif" correspond à un maximum local de cette distribution. C'est de cette façon que nous avons repéré le niveau 5 de l'arbre de classification sur $\mathcal{A}$ de la figure 15, ainsi que celui 9 de la figure 16 sur $\mathcal{O}$. Nous avons pu exprimer que le niveau 6 du premier arbre restait intéressant, parce que la valeur du critère $C(Q, \pi)$ gardait une certaine force. Le graphique de gauche de la figure 17 donne dans un cas réel, le diagramme de la distribution le long de la suite des niveaux, de la *Statistique globale des niveaux*. Chaque bâtonnet

vertical est associé à un niveau. Sa hauteur est proportionnelle à la valeur de cette statistique. Les niveaux "significatifs" sont marqués à la base par une $*$. Ces niveaux permettent, pour un ordre donné du nombre de classes de faire le bon choix d'une partition de l'ensemble organisé.

Maintenant, relativement à deux niveaux consécutifs $i-1$ et i, on associe le taux d'accroissement *Statistique globale des niveaux* ; c'est-à-dire :

$$\tau(Q,\pi_i) = C(Q,\pi_i) - C(Q,\pi_{i-1})\ ,\ 1 \leq i \leq l,$$

$\tau(Q,\pi_i)$ définit la *Statistique locale* attachée au niveau i. L'importance de ce taux témoigne du "gain en cohérence" obtenu en passant du niveau $i-1$ au niveau i. On constate d'ailleurs dans la pratique expérimentale que ce taux augmente graduellement dès lors qu'une classe en cours de formation se confirme. À l'achèvement à un niveau de synthèse d'une classe, la valeur de ce taux retombe. Il est donc naturel de considérer la distribution le long de la suite des niveaux de la *Statistique locale des niveaux* :

$$\big\{\tau(Q,\pi_i) | 1 \leq i \leq l\big\} \tag{19}$$

Dans ces conditions, un noeud "significatif" est défini par un maximum local de cette distribution. Le graphique de droite de la figure 17 donne la dsitribution de la *Statistique locale des niveaux* associée à celle *globale* figurée à gauche. L'amplitude du bâtonnet vertical pour le niveau i est proportionelle à la valeur $\tau(Q,\pi_i)$. Les noeuds "significatifs" sont marqués à la base par une $*$. Les noeuds et niveaux "significatifs" permettent une interprétation dynamique et ascendante (des feuilles vers la racine) de l'arbre de classification.

4.5 Croisement entre deux classifications duales

Nous en sommes maintenant - dans le cadre de l'$AVLR$ - au niveau du paragraphe 2.2.2 et en particulier de la figure 5 de l'introduction générale. Considérons dans le cadre de l'exemple traité le couple d'arbres de classification sur $\mathcal{A}$ et sur $\mathcal{O}$ (voir FIG. 15 et FIG. 16). Retenons la partition la plus "significative" sur $\mathcal{A}$ de niveau 5. Cette dernière est en 3 classes et peut être notée $\{A_1, A_2, A_3\}$. De même, retenons la partition la plus "significative" sur $\mathcal{O}$ de niveau 9, elle est en deux classes et peut être notée $\{C_1, C_2\}$. Le croisement de ces deux partitions donne le tableau de la figure 18.

Ce tableau est obtenu par permutation des lignes et des colonnes du tableau initial des données (voir FIG. 9), conformément au couple de classifications retenues. Les cases noircies sont celles qui contiennent la valeur logique 1 (attribut à $VRAI$). On constate que la classe C_1 se réfère dans une forte mesure à la classe A_3 et s'oppose à la classe A_1. La référence à la classe A_2 est plus partielle. Au contraire, la classe C_2 se réfère à la classe A_1 et s'oppose à la classe A_3. Là encore, C_2 se réfère de façon partielle mais différente de celle de C_1, à A_2. On peut certes considérer la partition de niveau 6 de l'arbre sur $\mathcal{A}$ où la partition obtenue est $\{A_1, A_2 \cup A_3\}$. Dans ce cas, C_1 se réfère à $A_2 \cup A_3$ et C_2 se réfère à A_1 ; mais, l'opposition de C_2 par rapport à $A_2 \cup A_3$ est moins nette que celle de C_2 par rapport à A_3.

Reprenons une des deux questions posées à la fin du paragraphe 2.2.2 à savoir : s'il s'agit d'un "grand" tableau de données booléennes, comment faire ? À cet égard, nous avons bien une solution qui a été de nombreuses fois expérimentée avec beaucoup d'intérêt. Elle relève

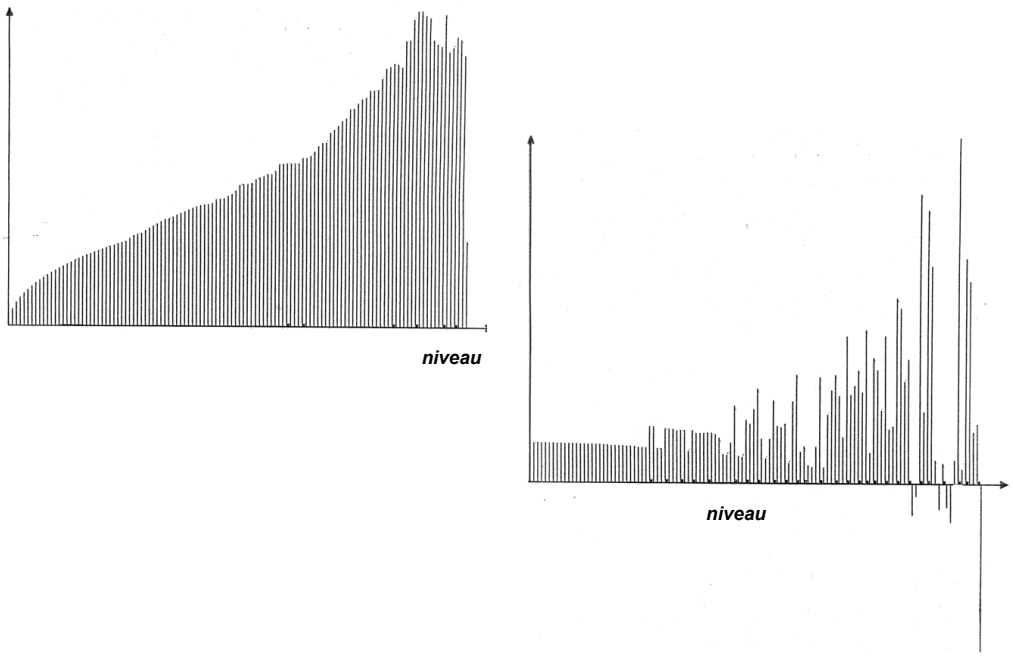

FIG. 17 – *Distributions des statistiques "globale" (en haut à gauche) et "locale" (en bas et à droite) des niveaux.*

d'un développement que nous avons effectué de la théorie du χ^2 attaché à un tableau de contingence. Ce développement nous permet de croiser deux classes d'attributs booléens A et B qui sont *logiquement* indépendants ; mais, bien sûr, *statistiquement* dépendants. Plus précisément, $A \cap B = \emptyset$, d'autre part, dans $A \cup B$, il n'y a pas un attribut et sa négation. La notion d'attribut booléen à $VRAI$ chez un objet donné x est remplacée, relativement à un ensemble A d'attributs booléens (*logiquement* indépendants), par la proportion dans A d'attributs booléens à $VRAI$ chez x. Nous noterons $\Phi_A(x)$ cette proportion. Elle s'écrit :

$$\Phi_A(x) = \frac{1}{c(A)} \sum_{a \in A} \Phi_a(x)$$

où $\Phi_a(x) = 1($ resp. $0)$ si l'attribut a est à $VRAI$ (resp., $FAUX$) chez x. D'autre part, $c(A)$ désigne le cardinal de A.

Associons à l'ensemble A d'attributs l'ensemble $\bar{A}$ des attributs obtenus en remplaçant chaque attribut de A par sa négation. On a dans ces conditions :

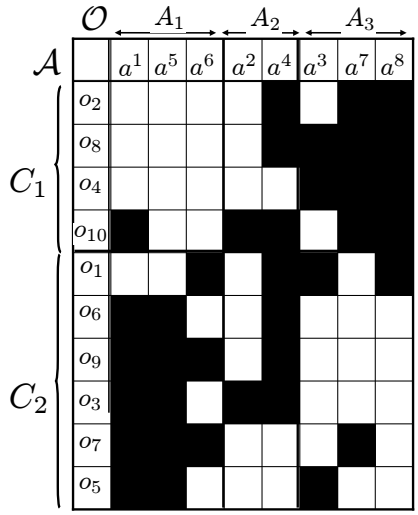

FIG. 18 – *Croisement entre deux partitions significatives duales.*

$$\Phi_{\bar{A}}(x) = \frac{1}{c(A)} \sum_{a \in A} \Phi_{\bar{a}}(x)$$

On a alors l'équation de "cohérence" :

$$\Phi_A(x) + \Phi_{\bar{A}}(x) = 1 \tag{20}$$

Relativement au couple (A, B) de classes d'attributs booléens tel que nous venons de le décrire, la démarche de l'$AVLR$ est toujours la même. Nous introduisons l'indice *brut* $\nu(A, B)$ suivant :

$$\nu(A, B) = \nu(A \wedge B) = \sum_{x \in \mathcal{O}} \Phi_A(x) \times \Phi_B(x) \tag{21}$$

Maintenant, on considère l'association

$$(A, B) \longrightarrow (A^*, B^*)$$

où au couple (A, B) de classes d'attributs booléens on associe un couple (A^*, B^*) de classes d'attributs booléens mutuellements indépendants dans une hypothèse d'absence de liaison adéquate. C'est l'indice centré et réduit

$$\chi(A, B) = \frac{1}{\sqrt{n-1}} \big(\nu(A \wedge B) - \mathcal{E}[\nu(A^* \wedge B^*)]\big) / \sqrt{var[\nu(A^* \wedge B^*)]} \tag{22}$$

qui définit l'intensité de l'association entre les deux classes d'attributs A et B.

Relativement à l'exemple ci-dessus introduisons les attributs booléens c^1 et c^2, où c^1 (resp., c^2) définit la fonction indicatrice de la classe C^1 (rep., C^2). c^2 est l'attribut booléen opposé à c^1. Le tableau des coefficients

$$\{\chi(c^i, A_j)|i = 1, 2, 1 \leq j \leq 3\}$$

est donné dans le tableau de la figure 19. Ce tableau synthétise bien les conclusions que nous avons déjà apportées.

De tels coefficients peuvent aisément être étendus dans le cas où les attributs sont quantitatifs. Cependant la méthode s'applique dans le cas très général de variables relationnelles, conformément au schéma de la figue 20. La question se pose alors de savoir comment faire correspondre les deux arbres de classification duaux sur $\mathcal{A}$ et $\mathcal{O}$ ($Croiser(CAHA, CAHB)$). D'autre part, dans la mesure où il peut être réalisé sous une certaine forme, quel type d'interprétation peut-on tirer d'un tel croisement. À cet égard, des coefficients que nous avons expérimentés avec beaucoup d'intérêt entre d'une part une variable relationnelle quelconque et d'autre part, une classification sur $\mathcal{O}$ ou une classe de $\mathcal{O}$, joueront certainement un rôle important.

	A_1	A_2	A_3
c^1	-0.833	$+0.143$	$+0.833$
c^2	$+0.833$	-0.143	-0.833

FIG. 19 – *Croisement statistique entre deux partitions duales.*

5 Les grands types mathématico-logiques de données

Selon notre point de vue les grands types mathématico-logiques de données se formalisent au moyen de deux systèmes que nous noterons $\mathcal{T}$ et $\mathcal{S}$. Nous allons les présenter ci-dessous en cherchant à les illustrer.

$$E \leftarrow \mathcal{A};$$
$$CAHA \leftarrow E.CAH;$$
$$E \leftarrow \mathcal{O}(\text{ resp., } E \leftarrow \mathcal{C});$$
$$CAHB \leftarrow E.CAH;$$
$$Croiser(CAHA, CAHB)$$

FIG. 20 – *Schéma général de la biclassification ascendante hiérarchique.*

5.1 Le système $\mathcal{T}$

$\mathcal{T}$ est un système de Tarsky. Il se présente sous la forme :

$$\mathcal{T} = \big(\mathcal{O}; R_1, R_2, ..., R_j, ..., R_p\big) \tag{23}$$

où $\mathcal{O}$ est un ensemble d'objets élémentaires et où $R_1, R_2, ..., R_j, ..., R_p$ sont p relations d'arités respectives quelconques. En analyse des données la relation R_j se trouve définie par le j-ème attribut a^j, $1 \leq j \leq p$. Commençons par nous référer au tableau des données du paragraphe 2.2.1 et donnons quelques exemples typiques. Dans le cas d'un attribut *booléen* a^j, la relation R_j est unaire et l'attribut est représenté au niveau de $\mathcal{O}$ de façon ensembliste (voir paragraphe 4.2.1). R_j est toujours une relation unaire mais valuée (valuation sur $\mathcal{O}$) dans le cas d'un attribut quantitatif-numérique. Dans le cas où l'attribut a^j est qualitatif *nominal* R_j est une relation binaire symétrique qu'on représente par un sous ensemble au niveau de l'ensemble $P_2(O)$ des parties à deux éléments de $\mathcal{O}$. Si a^j est un attribut qualitatif *ordinal* la représentation ensembliste est au niveau du produit cartésien $\mathcal{O} \times \mathcal{O}$. Si a^j est un attribut qualitatif *préordonnance*, la représentation est au niveau de l'ensemble des couples de couples $(\mathcal{O})^2 \times (\mathcal{O})^2$ ou, si l'on veut, de $\mathcal{O}^4$. Cependant, un codage en termes d'une *valuation* sur $\mathcal{O}^2$ est le plus souvent proposé pour des raisons de simplification de la complexité. La variable *taxinomique* est un cas particulier de la variable *préordonnance*. Une variable qualitative a^j dont l'ensemble des valeurs, codé par $\{1, 2, ..., i, ..., v_j\}$, est muni d'*un graphe valué de similarité*, peut être interprétée, sous un certain point de vue, comme une généralisation d'une variable qualitative *préordonnance*. Une telle variable définit un graphe valué sur l'ensemble $\mathcal{O}$ des objets. Une telle variable peut être représentée par la j-ème colonne du tableau des données. Cette dernière comprendra des codes issus de $\{1, 2, ..., i, ..., v_j\}$. Il importe dans ces

conditions, d'avoir séparément la matrice donnant le graphe valué sur $\{1, 2, ..., i, ..., v_j\}$.

Un cas d'importance est celui où a^j définit directement un graphe valué sur l'ensemble $\mathcal{O}$ des objets. Le support en termes d'une colonne du tableau des données pour représenter a^j, ne peut plus alors se conçevoir. Néanmoins, le problème bien connu de la classification d'un ensemble de graphes valués sur un ensemble d'objets, peut être abordé, dans le cadre de l'$AVLR$, comme la classification d'un ensemble de variables relationnelles.

Maintenant, il existe des situations où le tableau des données ne peut pas directement représenter l'information descriptive. Il s'agit notamment des données de type "séquences" de longueurs inégales (e.g. séquences génétiques). On ne pourra passer à la description au moyen d'un tableau de données qu'après avoir défini le même ensemble de variables de description pour l'ensemble des séquences. C'est ce que font la grande majorité des méthodes.

Un dernier point concerne le mélange des types de variables dans le cadre de la classification $AVLR$. Il s'agit plus précisément du cas où les diverses relations $R_j, 1 \leq j \leq p$ ne sont pas de même arité. La classification de l'ensemble $\mathcal{O}$ des objets est formellement indifférente à ce mélange. Toutefois, il importe que les différentes variables correspondent au même ordre de finesse structurelle et statistique. Il est ainsi difficile de traiter sur le même plan un attribut booléen et un attribut taxinomique très structuré. Pour ce qui est de la classification $AVLR$ des variables de description, il importe qu'elles soient toutes d'un même type. À cet égard et dans les cas classiques et assez généralement, où les variables induisent sur $\mathcal{O}$ des relations soit unaires, soit binaires, on peut de façon naturelle déterminer un codage uniforme en termes de variables *préordonnances* ou *graphes valués* (thèse M. Ouali-Allah).

5.2 Le système $\mathcal{S}$

Le système $\mathcal{S}$ se présente sous la forme :

$$\mathcal{S} = \big(\mathcal{C}; \text{dist}_{\mathcal{C}}(R_1), \text{dist}_{\mathcal{C}}(R_2), ..., \text{dist}_{\mathcal{C}}(R_j), ..., \text{dist}_{\mathcal{C}}(R_p)\big) \tag{24}$$

$\mathcal{C}$ est un ensemble de catégories (on dit encore "concepts"). R_1, R_2, ..., R_j, ... et R_p sont comme ci-dessus, p relations d'arités respectives quelconques, sur $\mathcal{O}$. $\text{dist}_{\mathcal{C}}(R_j)$ représente la famille des distributions statistiques de R_j sur chacune des catégories c de $\mathcal{C}$.

Pour se rendre compte du degré de généralité de ce système, considérons quelques types de données classiques qui s'y inscrivent. Un *tableau de contingence* constitue un cas très particulier de ce système. Il se met sous la forme :

$$\big(\mathcal{C}; \text{dist}_{\mathcal{C}}(R)\big)$$

où R est une relation d'équivalence associée à une variable qualitative *nominale*. Maintenant, une juxtaposition "horizontale" de tableaux de contingence s'exprime au moyen de :

$$\big(\mathcal{C}; \text{dist}_{\mathcal{C}}(R_1), \text{dist}_{\mathcal{C}}(R_2), ..., \text{dist}_{\mathcal{C}}(R_j), ..., \text{dist}_{\mathcal{C}}(R_p)\big)$$

où R_1, R_2, ..., R_j, ... et R_p sont p relations d'équivalence respectivement associées à p variables qualitatives *nominales*. Plus généralement, les données de type "histogrammes" où

les pieds d'un même histogramme sont munis d'une relation, correspondent à ce système.

Signalons pour terminer que cette formalisation mathématico-logique des données en les deux systèmes $\mathcal{T}$ et $\mathcal{S}$ - qui nous était apparue vers la fin des années 80 - intègre bien les données dites "symboliques" et que nous préférons appeler "données de connaissace" ("knowledge data", en anglais). $AVLVR$ traite efficacement ces données.

6 Logiciels et Applications

6.1 Les programmes CHAVLh, AVARE et LLAhclust

Nous allons mentionner succintement les produits logiciels qui ont été élaborés pour la mise en oeuvre de la méthodologie $AVLR$. Ces programmes ont été écrits en Fortran 77, conformément aux normes rigoureuses de *Modulad* établies par Henri Leredde (maître de conférences à l'Université de Paris Nord).

Le programme le plus important est $CHAVLh$ (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de données hétérogène) (P. Peter, H. Leredde et I-C. Lerman). La dernière version de ce programme a été déposée l'Agence de Protection des Programmes (*APP*) en décembre 2005. Philippe Peter (maître de conférences à l'École Polytechnique de l'Université de Nantes) a été l'artisan de cette dernière version. Il a été installé par le projet *Symbiose* sur la plateforme logicielle de la génopole Ouest. Les structures de données prises en compte sont les suivantes :

1. Numériques-Quantitatives
2. Booléennes
3. Qualitatives nominales
4. Qualitatives ordinales
5. Qualitatives préordonnances ou graphes valués binaires
6. Juxtaposition "horizontale" de tableaux de contingence
7. Tableau de dissimilarités fourni par l'expert

Le programme permet la classification d'un ensemble $\mathcal{O}$ d'objets décrits par des variables toutes d'un même type ou par des variables pouvant être *de types respectifs différents*.

Par ailleurs, le programme permet la classification de l'ensemble $\mathcal{V}$ des variables de description, *toutes d'un même type*. Il peut s'agir de variables *numériques*, *booléennes*, *qualitatives nominales* ou *qualitatives ordinales*.

Un deuxième programme important $AVARE$ (Association entre VAriables qualitatives pREordonnance) a été élaboré par M. Ouali-Allah dans le cadre de sa thèse. Il permet la classification $AVLR$ d'un ensemble de variables qualitatives de différents types codées en termes de *préordonnances*. Rappelons à cet égard que la variable *taxinomique* représente un cas très spécifique d'une variable *préordonnance*.

Signalons ici qu'une version ergonomique et donc simplifiée du programme $CHAVLh$ a été implantée (juillet 2007) dans l'environnement logiciel dit **R** (I. Kojadinovic (École Polytechnique de l'Université de Nantes), I.C. Lerman et P. Peter). Bien que ne comportant pas certaines options de $CHAVLh$, cette version reste assez riche. Le programme est intitulé $LLAhclust$ (Likelihood Linkage Analysis hierarchical clustering)

6.2 Quelques applications récentes d'envergure

Nous nous contenterons de mentionner les thématiques dans lesquelles s'inscrivent ces applications et les références des travaux menés dans différents projets ; mais où, à chaque fois, la méthode $AVLR$ et donc le programme $CHAVLh$ a joué un rôle *essentiel*.

6.2.1 Simulation du comportement de l'exécution de "très gros" programmes à partir d'un échantillonnage "représentatif"

T. Lafage ; "*Étude, réalisation et application d'une plate-forme de collecte de traces d'exécution de programmes*", Thèse de doctorat, Université de Rennes 1, Décembre 2000.

6.2.2 Détermination de classes sémantiques dans le *Traitement Automatique de Langues Naturelles*

M. Rossignol ; "*Acquisition automatique de lexiques sémantiques pour la recherche d'information*", Thèse de doctorat, Université de Rennes 1, Octobre 2005.

6.2.3 Correspondance entre profils génotypiques et profils phénotypiques dans l'hémochromatose

I.-C. Lerman ; "Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques", *Revue de Statistique Appliquée*, 2006, vol. 2, pp. 33-63.

6.2.4 Structuration d'une famille spécifique de séquences d'ADN

S. Tempel ; "*Dynamique des hélitrons dans le génôme d'Arabidopsis thaliana : développement de nouvelles stratégies d'analyse des éléments transposables*", Thèse de doctorat, Université de Rennes 1, juin 2007.

6.2.5 Organisation d'espèces de "phlébotomes" pour une description biologique très complexe

I.-C. Lerman, P. Peter ; "Representation of Concept Description by Multivalued Taxonomic Preordonance Variables", in : *Selected Contributions in Data Analysis, and Knowledge organization*, P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (editors), *Studies in Classification, Data Analysis, and Knowledge organization*, Springer, 2007, pp. 271-284.

6.2.6 Segmentation d'images numérisés

I.-C. Lerman, K. Bachar ; "Comparaison de deux critères en Classification Ascendante Hiérarchique sous contrainte de contiguïté. Application en imagerie numérique", *Journal de la Société Française de Statistique* et *Revue de Statistique Appliquée*, tome 149, n° 2, 2008, pp. 45-74.

7 Références importantes pour le fondement et l'élaboration de la méthodologie

Nous organisons ces références par thèmes. À l'intérieur de chacun des thèmes, les références seront ordonnées chronologiquement. Elles concernent nos publications ainsi que celles de chercheurs de notre environnement le plus immédiat.

7.1 Références générales

I.C. Lerman ; *Classification et Analyse Ordinale des Données*, Dunod, Paris, 1981.

I.C. Lerman ; Foundations of the Likelihood Linkage Analysis (LLA) Classification Method, *Stochastic Models and Data Analysis*, vol. 7, ♯ 1, march 1991, pp. 63-76.

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

7.2 Coefficients d'association (de similarité) entre variables relationnelles

I.C. Lerman, R. Gras, H. Rostam ; Élaboration et évaluation d'un indice d'implication pour des données binaires I : *Revue Mathématiques et Sciences*

Humaines, 19ème année, n° 74, 1981, pp. 5-35, II : *Revue Mathématiques et Sciences Humaines*, 19ème année, n° 75, 1981, pp. 5-47.

I.C. Lerman ; Association entre variables qualitatives ordinales nettes ou floues, *Statistique et Analyse des Données*, 1983, vol. 8 n° 7, pp. 41-73.

I.C. Lerman ; Indices d'association partielle entre variables qualitatives nominales, *RAIRO série verte*, vol. 17, n° 3, août 1983, pp. 213-259.

I.C. Lerman ; Indices d'association partielle entre variables qualitatives ordinales, *Publications Institut Statistique de Paris*, XXVIII, fasc. 1, 2, 1983, pp. 7-46.

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 Institut Statistique Universités de Paris*, XXIX, fasc. 3-4, 1984, pp. 27-57.

I.C. Lerman ; Comparing partitions. Mathematical and Statistical aspects, 1ère conférence internationale des fédérations des sociétés de classification. Aix-la-Chapelle, juin 1987, in *Classification and related methods of data analysis*, Edited by H. H. Bock, North Holland, 1988, pp. 121-132.

M. Ouali-Allah ; *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques*, thèse de doctorat de l'Université de Rennes 1, décembre 1991.

F. Daudé ; *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*, thèse de doctorat de l'Université de Rennes 1, juin 1992.

I.C. Lerman ; Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, I : *Mathématique (, Informatique) et Sciences Humaines*, $30^{\text{ième}}$ année, n° 118, 1992, pp. 35-52, II : *Mathématique (, Informatique) et Sciences Humaines*, $30^{\text{ième}}$ année, n° 119, 1992, pp. 75-100, Paris.

I.C. Lerman ; Comparing classification tree structures : a special case of comparing q-ary relations, *RAIRO-Operations Research*, vol. 33, 1999, sept., pp. 339-365.

I.C. Lerman, F. Rouxel ; Comparing classification tree structures : a special case of comparing q-ary relations II, *RAIRO-Operations Research*, vol. 34, 2000, pp. 251-281.

I.C. Lerman ; Comparing taxonomic data, *Revue Mathématiques et Sciences Humaines*, $38^{\text{-ème}}$ année, n° 1510, 2000, pp. 37-51.

I.C. Lerman, J. Azé ; A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link, in *Quality Measures*

in Data Mining. Studies in Computational Intelligence. F. Guillet and H. Hamilton (eds). 2006. Springer. pp. 207-236.

R. Gras, P. Kuntz ; An overview of the Statistical Implicative Analysis (SIA) development, in R. Gras, F. Guillet, F. Spagnolo, E. Suzuki ; *Statistical Implicative Analysis*, Studies in Computational Intelligence, n° 27, Springer, 2008.

7.3 Indices de similarité entre objets décrits par des variables relationnelles de types quelconques

I.C. Lerman et Ph. Peter ; Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle, *Data Analysis and Informatics IV*, North Holland, 1986, pp. 121-136.

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

I.C. Lerman ; 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), 1987, pp. 39-60.

I.C. Lerman, Ph. Peter et J.L. Risler ; Matrices AVL pour la classification et l'alignement de séquences protéiques, *Publication Interne IRISA* n° 866, septembre 1994, *Rapport de Recherche INRIA* n° 2466.

I.C. Lerman, Ph. Peter ; Indice probabiliste de vraisemblance du lien entre objets quelconques : analyse comparative entre deux approches, *Revue de Statistique Appliquée*, volume LI(1), 2003, pp. 5-35.

I.C. Lerman and Ph. Peter ; Representation of Concept Description by Multivalued Taxonomic Preordonance Variables, in *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, P. Brito, P. Bertrand, G. Cucumel and F. de Carvalho, editors, Springer 2007, pp. 271-284.

7.4 Tableaux de contingence

I.C. Lerman et B. Tallur ; Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence, *Revue de Statistique Appliquée*, n° 28, 3, 1980, pp. 5-28.

I.C. Lerman ; Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence, *Mathématique et Sciences Humaines*, 21ème année, n° 83, 1983, pp. 5-30.

I.C. Lerman ; Analyse classificatoire d'une correspondance multiple, typologie et régression, in *Data Analysis and Informatics*, III, E. Diday et al. (editors), North Holland, 1984, pp. 193-221.

B. Tallur ; *Contribution à l'analyse exploratoire de tableaux de contingence par la classification*, thèse de doctorat ès sciences, Université de Rennes 1, septembre 1988.

7.5 Critères de fusion entre classes

I.C. Lerman ; Formules de réactualisation en cas d'agrégations multiples, *RAIRO, série R.O.* vol 23, n° 2, 1989, pp. 151-163.

F. Costa Nicolau and H. Bacelar-Nicolau ; Some trends in the Classification of Variables, in *Data Science, Classification, and Related Methods*, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba Editors, 1998, pp. 89-98.

7.6 Niveaux et noeuds "significatifs"

I.C. Lerman ; Sur la signification des classes issues d'une classification automatique, in *Numerical Taxonomy*, NATO ASI Series, vol. G1. Edited by J. Felsenstein, Springer-Verlag (1983), pp. 179-198.

I.C. Lerman, N. Ghazzali ; What do we retain from a classification tree ? an experiment in image coding, in *Symbolic-Numeric data analysis and learning*, edited by E. Diday and Y. Lechevallier, Nova Science Publishers, Proceedings of the conference of Versailles, september 18-20, 1991, pp. 27-42.

7.7 Croisement de classifications

I.C. Lerman ; Croisement de classifications "floues", *Publications de l'Institut de Statistique des Universités de Paris*, 1979, XXIV, fasc. 1-2, pp. 13-46.

I.C. Lerman, M. Hardouin et T. Chantrel ; Analyse de la situation relative entre deux classifications "floues", Secondes Journées Internationales Analyse des Données et Informatique, in *Data Analysis and Informatics*, North Holland, 1980, pp. 523-533.

I.C. Lerman ; Association entre variables qualitatives ordinales "nettes" ou "floues", *Statistique et Analyse des Données*, n° 7, 1983, pp. 41-73.

I-C. Lerman ; Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques, *Revue de Statistique Appliquée*, 2006, LIV (2), pp. 33-63.

7.8 Logiciels

P. Peter, H. Leredde et I.C. Lerman ; Notice du programme CHAVLh (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables hétérogènes), *Dépôt APP (Agence pour la Protection des Programmes)* IDDN.FR.001.240016.000.S.P.2006.000.20700, Université de Rennes 1, Décembre 2005.

M. Ouali Allah ; Programme pour le calcul de coefficients d'association entre variables relationnelles, *La revue de modulad* n° 25, juin 2000, pp. 63-73.

I. Kojadinovic, I.C. Lerman and P. Peter ; LLAhclust dans R :
http ://cran.rproject.org/src/contrib/Descriptions/LLAhclust.html.

Summary

The edification of the Likelihood Linkage Relational Analysis classification method began at the end of nineteen sixties. From that period this methodology has been extensively developed. Many researchers and practitioners have contributed to its development. Many application works on a large scale, provided by various fields (Bioinformatics, Informatics, Social sciences, Image processing, Natural language processing, ...) have been performed and then, have validated this approach. Any logical or mathematical type of data description can be handled in an accurate fashion. The aim of this paper consists of presenting in an illustrated way the fundamental principles of this methodology. Two conception levels are associated with these principles. The first is defined by the mathematical representation of the data description, while the second level is concerned by the problem of measuring the resemblances between the mathematical structures to be compared. In our case, the description representation will have a set theoretic and relational nature. Besides, quantifying the resemblance will be done by means of a probabilistic similarity. The latter is established with respect to a statistical hypothesis of independence. The following text corresponds to our oral presentation at the "3-èmes Journées Thématiques Apprentissage Artificiel et Fouille des Données", 2008 April 8-9.