Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision ; application à l'identification de la structure secondaire d'une protéine

Israël-César Lerman, Joaquim F. Pinto da Costa

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision ; application à l'identification de la structure secondaire d'une protéine.

Israël César Lerman et Joaquim F. P. da Costa

N 2803

Février 1996

PROGRAMME 3

ISSN 0249-6399

Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision ; application à l'identification de la structure secondaire d'une protéine.

Israël César Lerman* et Joaquim F. P. da Costa**

Programme 3 — Intelligence artificielle, systèmes cognitifs et interaction homme-machine
Projet Repco

Rapport de recherche n° 2803 — Février 1996 — 46 pages

Résumé : Des apports méthodologiques dans l'élaboration des Arbres de Décision binaires nous conduisent à définir une nouvelle méthode ARCADE (ARbre de ClAssification et de DEcision). Cette dernière reprend la célèbre méthode CART de Breiman & al. en y injectant une nouvelle famille de coefficients d'association entre variables qualitatives, jusqu'alors, seulement utilisé dans le cadre de la classification AVL. Cela permet en outre de contester l'affirmation selon laquelle la technique de choix d'un attribut binaire predictif n'a aucune importance et peut même se faire au hasard. La contribution la plus importante concerne la réduction à un faible nombre d'attributs binaires pertinents, de variables qualitatives prédictives à très grand nombre de valeurs (20^4 dans le cadre de notre application). À cette fin, la classification hiérarchique AVL est utilisé de façon adaptée. L'application motrice concerne la prédiction de la structure secondaire des protéines, en utilisant une base formé de 151 protéines globulaires, où la seule information utilisée est celle donnée par ces dernières, résidu par résidu. Les résultats obtenus par une procédure de type "jackknife" ont été $Q_{chain} \simeq 67\%$ et $Q_{total} \simeq 65.8\%$.

Mots-clé : Arbres de décision, attributs qualitatifs à très grand nombre de catégories, coefficients d'association entre variables, prédiction de la structure secondaire d'une protéine.

(Abstract: pto)

∗. IRISA-INRIA-Rennes, lerman@irisa.fr
∗∗. Grupo de Matemática Aplicada, Universidade do Porto, Portugal,jpcosta@ncc.up.pt; IRISA-INRIA-Rennes, costa@irisa.fr

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

Association Coefficients and Variables with very large Number of Categories in Decision Trees. An Application to Protein Secondary Structure Prediction.

Abstract: Methodological developments in the construction of binary decision trees allowed us to define a new method. This method, ARCADE ("ARbre de ClAssification et de DEcision"), takes again the celebrated method CART of Breiman & al. by injecting for the first time a new family of measures of association between categorical variables, which had only been used in the context of AVL automatic classification. This allowed us to contest the claim made by different authors that the coefficient used to select an attribut is irrelevant and can even be random. The most important contribution concerns the reduction of the number of binary splits defined by a categorical variable with a very large number of categories (20^4 in our data) to a few pertinent ones. To achieve this, we have adapted the AVL program of hierarchic classification to our problem. The application concerns the prediction of protein secondary structure. The single residues of 151 globular proteins of known structure was the only information used (no homologous proteins were used). The results, obtained by a "jackknife" procedure, were $Q_{chain} \simeq 67\%$ and $Q_{total} \simeq 65.8\%$.

Key-words: Decision trees, categorical attributes with a very large number of categories, measures of association between variables, protein secondary structure prediction.

Table des matières

1 Introduction générale

Les arbres binaires de décision représentent un outil logique de discrimination de concepts (classes d'objets définies en intension). Les problèmes méthodologiques et d'inférence relatifs à leur elaboration, ont cristallisé les efforts conjoints de deux domaines; l'analyse des données qualitatives et l'apprentissage (branche de l'intelligence artificielle) [Breiman et al., 1984], [Quinlan, 1986], [Nakhaeizadeh, 1994], [Taylor, 1994].

Leur construction s'effectue à partir d'un ensemble $\mathcal{A}$ d'attributs de description dichotomiques (chacun à deux modalités ou catégories), sur la base d'un ensemble $\mathcal{E}$ d'appren-

tissage. La variable "concept à reconnaître" représente un attribut c à K catégories; où on peut imaginer que K n'est pas nécessairement petit.

Sur une large base expérimentale, nous montrons que, contrairement à l'avis exprimé dans [Mingers, 1989], il est crucial que cette construction soit fondée sur un "bon" coefficient d'association entre attributs qualitatifs. Nous aurons quant à nous et ici à comparer à chaque fois, un attribut binaire prédictif a ($a \in \mathcal{A}$) et l'attribut c, présentant K valeurs, à prédire. Cette comparaison est effectuée sur la base de l'ensemble plein $\mathcal{E}$, si on se trouve à la racine de l'arbre ou sur la base d'une partie propre de $\mathcal{E}$, si on se trouve à l'un des nœuds de l'arbre.

La notion de "bon" coefficient est difficile à cerner; et, un des intérêts de ce travail est de présenter une adaptation pour le probléme posé, d'une large palette d'indices, dont certains - conçus dans le contexte de la classification des variables - sont nouveaux dans le domaine. La conception d'un tel coefficient peut se référer à des représentations du tableau de contingence et des conditions qui sont; soit plutôt géométriques et statistiques (Gini, χ^2, Matusita, Bacelar-Nicolau, Twoing), soit plutôt contingentielles et informationnelles (Shannon, Haldane, Light et Margolin) soit enfin plutôt, combinatoires et statistiques (coefficients de Lerman). On peut certes ensuite, interpréter dans un cadre, un coefficient d'association conçu dans un autre. Ainsi, nous réduisons le coefficient de Haldane, Light et Margolin à celui de Gini auquel nous donnons le sens d'une inertie (comme d'ailleurs c'est le cas pour le χ^2); d'autre part, nous rapprochons au mieux l'indice centré de Lerman de celui de Gini. C'est au chapitre 2 que nous développerons les différents types de coefficients.

À cet égard on minimise dans [Breiman et al., p.38, 1984], l'intérêt comparé des différents coefficients - Encore faut-il-nous avons pu le tester - qu'il s'agisse de "bons" coefficients répondant à des critères statistiques. Si alors, les performances globales en termes de prédictivité sont très proches, il n'en est pas de même, des prédictivités respectives, classe par classe. D'autre part, des différences sensibles peuvent apparaître en termes de complexité, mesurée par le nombre de feuilles, de l'arbre de décision obtenu; cette complexité restant de toute façon très raisonnable. On mettra en évidence ces aspects au chapitre 5.

Les variables de description prédictives sont rarement, initialement binaires. Si $\mathcal{S}(v)$ désigne l'échelle des valeurs d'une telle variable v, un aspect fondamental de la recherche, consiste en la manière de construire des bipartitions (partitions à 2 classes) de $\mathcal{S}(v)$, afin d'obtenir des attributs dichotomiques associés à v. Une telle construction doit dépendre de la structure (on dit encore sémantique) de $\mathcal{S}(v)$. Précisèment, ces dernières années, on s'est trouvé concerné par la manière de discrétiser le numèrique [Krzanowsky, 1975], [Van de Merckt, 1993], [Heath et al., 1993], [Müller et Wysotzki, 1994].

Le cas spécifique où se situe notre apport méthodologique est celui où chacune des variables prédictives v est qualitative nominale; mais à très grand nombre de modalités (catégories). En d'autres termes $L(v) = \text{card}[\mathcal{S}(v)]$ est "grand" (et peut dans notre application atteindre 20^4), pour chacune des variables v de l'ensemble $\mathcal{V}$ des variables prédictives.

Si $L = L(v)$ était petit, rien n'aurait empêché le remplacement de la variable initiale v par $(2^{L-1}-1)$ attributs binaires; où une même modalité de l'une de ces dernières variables binaires, correspond à un sous ensemble propre des modalités de la variable initiale.

Référons nous à présent au tableau de contingence à L lignes et K colonnes, établi sur la base de l'ensemble d'apprentissage et croisant une variable v prédictive, à L catégories et la variable à prédire dont les valeurs sont les concepts. Si $K = 2$ et si le critère appartient à une classe dont font partie les indices inertiels (e.g. Gini et χ^2), on peut montrer que l'obtention - par regroupement des catégories - de la meilleure variable binaire associée à v, est de complexité linéaire par rapport à L. Cette démarche qui est considérée et proposée pour Gini dans [Breiman et al., 1984], devient moins claire dés lors que K est supérieur à 2; et cela, même si K reste de l'ordre de quelques unités - La complexité de la solution devient sérieuse si K augmente - Enfin, quoi faire si on veut considérer des critères non inertiels? Le chapitre 3 est réservé à une description de la méthode CART de Breiman, Friedman, Olshen et Stone; et donc, de ces derniers aspects.

Dans ce qui précède, comme dans ce qui suit, on suppose que les contenus des cases du tableau de contingence ont la consistance nécessaire pour la significativité des calculs.

Notre méthode consiste dans sa première étape à réduire par regroupement, l'ensemble des modalités de chacune des variables prédictives - Une même variable v se trouve ainsi remplacée par une macro variable $w = w(v)$, dont chaque modalité est une classe de modalités de la variable d'origine v. Mais il s'agit de créer ces nouvelles modalités synthétiques en accord même avec le principe de discrimination sous jacent à la formation d'un arbre de décision. En effet, nous le faisons à partir d'une classification automatique de l'ensemble des L modalités de la variable d'origine en se basant sur le tableau de contingence LxK, ci-dessus considéré - Le nombre de classes retenu, sera d'une part fonction de la "significativité" des classes et d'autre part, de l'ordre de grandeur de leur nombre; et ce, en relation avec la complexité qu'on peut accepter pour la formation de l'arbre de décision - Un outil parfaitement adapté à cette fin est la méthode de classification hiérarchique de l'Analyse de la Vraisemblance des Liens ($\mathcal{AVL}$) [Lerman,1993] implantée dans le programme CHAVL [Lerman, Peter et Leredde, 1993-1994]. La partition de l'ensemble des L modalités sera obtenue à partir d'un niveau "significatif" de l'arbre des classifications. De la sorte, à chacune des variables v^j de $\mathcal{V}$ [cf. ci-dessus], on associera une macro variable w^j, à partir d'une partition de l'ensemble des modalités de v^j, comme nous venons juste de l'exprimer, $1 \leq j \leq p$.

Et si J est le nombre de modalités d'une macro variable w; on peut envisager de considérer l'ensemble des $(2^{J-1} - 1)$ attributs binaires et dont chacun se trouve associé à une partition en deux classes de l'ensemble des macro-modalités. Cependant, l'arbre hiérarchique détermine une structure ultramétrique de proximité entre les classes qui définissent les macro modalités; et il est du plus grand intérêt d'exploiter cette structure. Nous le faisons, en imposant à l'une des deux modalités de la variable binaire de fusionner des classes conformément à l'arbre hiérarchique. Ainsi, à chacune des feuilles et à chacun des nœuds de l'arbre de classification sur l'ensemble des macro modalités, correspond une variable. Il en résulte une chute vertigineuse de la complexité.

Nous avons pu annoncer que le nombre L de modalités de la variable prédictive pouvait atteindre une très grande valeur. Dans notre cas, où L peut atteindre 20^4, pour chaque variable initiale v, nous avons abouti à ne considérer qu'environ 20 variables binaires; ce qui est considerablement inférieur à L. Mais alors, dans ce cas, si la taille de l'ensemble $\mathcal{E}$ d'apprentissage n'est pas suffisante, le tableau de contingence LxK sera inconsistant puisque le contenu de beaucoup de ses cases sera nul ou trop faible. Dans ce cas nous procédons - mail il faut que cela s'y prête - par approximation, en factorisant l'ensemble M des modalités de la variable prédictive v. Cela sera en effet structurellement naturel pour notre description, puisque v se trouve défini par un mot dont les lettres prennent valeurs dans le même alphabet fini de taille 20. C'est donc une procédure conjointe de factorisation de l'ensemble des valeurs et de classification par la méthode AVL, qui nous a permis, de façon statistiquement signifiante, la réduction annoncée ci-dessus.

Cette technique, qui constitue notre apport méthodologique majeur dans le travail rapporté ici, sera explicité au chapitre 4.

Notre méthode résulte d'une application d'importance en Biologie Moléculaire dans l'analyse des séquences protéiques. Rappelons que la structure. Rappelons que la structure primaire d'une telle séquence est formellement un mot pris dans un alphabet de 20 lettres (représentant les 20 acides aminés) et dont la longueur peut atteindre plusieurs centaines. La structure secondaire peut également être formalisée au moyen d'un mot de même longueur, calé sur le premier; mais dont l'alphabet comprend trois lettres E, H et X, qui seront les noms des trois concepts à discriminer. Il s'agit en effet, à partir d'une *description* de la position donnée d'une lettre du premier mot, de prédire la lettre correspondante du second mot. On comprend dans ces conditions qu'une approche de type arbre de décision puisse être adaptée pour le problème posé où on a $K = 3$. Encore faut il une description pertinente pour ce problème réputé très difficile de reconnaissance. En effet, la prédiction est loin de seulement dépendre de la seule lettre du premier mot dont il s'agit de déterminer la lettre associée (i.e. ayant la même position) dans le second mot. C'est au chapitre 5 que nous préciserons la description

adoptée par des mots de 4 lettres. La base expérimentale pour laquelle la méthode a été testée, est formé de 151 séquences d'une protéïne globulaire (consulter [Colloc'h et al. 1993] pour une description de ces données). Les fréquences relatives des trois classes X, H et E, sont respectivement, 46,6%, 29% et 24,4%. La taille d'une même séquence présente une grande variabilité; d'environ une dizaine de lettres, jusqu'environ 800. À chaque fois, il y en a 151, l'ensemble d'apprentissage est formé de 150 séquences et la prédiction est effectuée sur la séquence restante.

Le programme ARCADE (ARbre de ClAssification et de DEcision) mis au point par Pinto da Costa en 1994 reprend en tout point les aspects du programme CART, pour la validation et l'élagage de l'arbre. Mais il comprend deux points nouveaux qui concernent d'une part, l'introduction d'une nouvelle famille de coefficients issus d'$\mathcal{AVL}$; et d'autre part, la réduction significative de l'espace des variables associées à une même variable prédictive qui présente un grand nombre de modalités.

2 Les différents types de critères et indices d'association

2.1 Introduction

Nous considérons dans ce travail l'introduction de divers critères et indices d'association entre variables qualitatives nominales pour la construction des arbres de décision binaires. Dans le processus de construction de ces arbres, il faut, à chaque nœud t , choisir une variable binaire w pour le segmenter en deux nœuds descendants, t_l et t_r. Il s'agit donc de trier, parmi un ensemble de variables binaires, celle qui est la plus associée (la plus prédictive, la plus liée) avec la variable qualitative à discriminer, qu'on désigne par c .

$$w : O \longrightarrow \{t_l, t_r\} \text{ et } c : O \longrightarrow \{c_1, c_2, ..., c_k, ..., c_K\}.$$

La portée de ces coefficients est générale et concerne le croisement de deux variables qualitatives, ayant chacune un nombre quelconque de modalités ou valeurs. Ainsi, dans le contexte de notre application, ils peuvent également servir à évaluer la qualité globale de la prédiction par rapport à un ensemble test; et ce, en croisant la variable " classe observée" avec celle, "classe prédite". Cette qualité est très généralement mesurée par la proportion (ou pourcentage) de bon classement. Cependant, des biologistes [Rost et Sander 1993] ont introduit un coefficient appelé $Info$, qui est bien un coefficient d'association entre variables qualitatives nominales; et qu'ils utilisent pour juger de la qualité de la prédiction. Nous nous contenterons d'exprimer ce coefficient dans le cadre de notre application et dans le contexte où il a été utilisé (cf. chap. 5).

Tout autant, nous présenterons ici les coefficients qui ont été utilisés pour bâtir l'arbre de décision. Et d'ailleurs, nous limiterons notre présentation au cas d'espèce où la première variable (prédictive) est à deux modalités et où la seconde variable (à prédire) est à K modalités.

Il existe dans la littérature de nombreux critères ou indices d'association. Outre les coefficients les plus classiques d'interprétation géométrique ou informationnelle, construits à partir du tableau de contingence croisant les deux variables qualitatives nominales (consulter [Goodman et Kruskal 1979] pour un excellent répertoire ou encore [Everitt 1977]), nous considérons aussi les critères de comparaisons par paires, où il s'agit de comparer les relations d'equivalence induites par deux variables à modalités (variables qualitatives nominales où encore variables partition). La différence fondamentale entre ces deux approches est que dans le cas des coefficients contingentiels on se situe dans l'espace O des objets élementaires et dans les comparaisons par paires on se situe dans l'espace OxO des couples d'objets.

Par ailleurs, et de façon très intéressante, on peut, quel que soit le niveau de conception O ou OxO, chercher à appréhender la distribution statistique du coefficient dans l'hypothèse d'indépendance entre les deux variables (e.g. cas célèbre du χ^2).

On peut aussi, quel que soit le degré du lien entre les deux variables, mesuré au moyen du coefficient, au niveau de la population mère P (dont O représente un échantillon de taille n), étudier la distribution d'échantillonnage d'un tel coefficient (Goodman & Kruskal 1963, Bacelar-Nicolau 1980, Lerman 1984, Daudé 1992).

Maintenant, très souvent, la base même de la construction d'un coefficient d'association est purement intuitive. Lorsqu'elle se situe au niveau de O , à partir de l'appréhension du tableau de contingence, elle repose généralement sur des considérations statistiques simples. Lorsque cette conception se situe au niveau de OxO, c'est finalement la situation relative des graphes des relations d'équivalences, associées aux deux partitions à comparer, qui est évaluée à partir d'indices cardinaux simples. En effet, chaque graphe est considerée comme un sous ensemble sans spécificité particulière, de OxO.

On pourra chercher à donner une interprétation métrique voire même, géométrique, à un coefficient conçu à partir de considérations ensemblistes ou statistiques (Régnier 1965, Giakoumakis & Monjardet 1987). On peut aussi et en même temps étudier si l'expression du coefficient se linéarise au niveau de l'ensemble des paires (Marcotorchino 1984a,b).

Inversement, partant d'une interprétation géométrique au niveau de O (Saporta 1975, Issa Khalil 1991) géométrique ou métrique au niveau de OxO (divers travaux), on peut proposer des coefficients qu'il s'agit alors d' etudier des points de vue ensembliste et statistique.

Notre propre démarche consiste à partir d'un critère "simple" de comparaison entre les graphes des deux relations d'équivalence respectivement associées aux deux partitions. Cet critère, que nous appelons "brut" et qui représente un cardinal, peut d'ailleurs être proposé à partir de considérations métriques au niveau de OxO . La démarche consiste alors à normaliser cet critère par rapport à une hypothèse de nature combinatoire et statistique, d'absence de liaison (ou d'indépendance) (Lerman 1973, 1981, 1992a, 1992b, Ouali-Allah 1991, Daudé 1992).

Enfin, il est clair que quelle que soit la conception du coefficient d'association, le calcul s'effectue au niveau du tableau de contingence 2xK, croisant la variable prèdictive et celle, à prèdire.

2.2 Les coefficients contingentiels d'interprétation géométrique ou informationnelle.

Dans l'approche contingentielle (developpée par les statisticiens, surtout les anglo-saxons), pour comparer deux variables à modalités, u et v, on forme d'abord un tableau de dimension pxq (p est le nombre de modalités de u et q est le nombre de modalités de v) croisant les deux variables. On désigne par :

- n_{ij} le nombre d'objets possédant la modalité i de u et la modalité j de v ;
- $n_{i.}$ le total de la ligne i ($n_{i.} = \sum_{j=1}^{q} n_{ij}$) ;
- $n_{.j}$ le total de la colonne j ($n_{.j} = \sum_{i=1}^{p} n_{ij}$) ;
- $n_{..}$ ou plus simplement par n le total général ($n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$).

Plusieurs coefficients d'association, basés sur ce tableau de contingence, ont été proposés dans notre siècle: Le critère du Chi-deux, introduit par Pearson en 1904 ; le critère de Belson (1959) (par abus de language on attribue ce critère à Belson, bien qu'il avait déjà été proposé en 1942, sous une forme différente, par Höffding) ; le critére ou indice de Rand, introduit en 1971, pour comparer deux partitions - avec le même nombre de classes - issues d'une classification hiérarchique (Anderberg, en 1973, généralise l'utilisation de ce critère à la comparaison de deux partitions quelconques). On cite encore l'indice de Goodman-Kruskal (1954) et critères derivés (Light et Margolin (1971), Haldane (1940)) ; le critère de la moyenne des interactions, derivé d'une idée de Jordan de 1927 ; le coefficient de corrélation généralisé, proposé par Der Megreditchian, en 1988, etc. D'entre tous ces critères, on a choisi huit, qu'on décrit ci-dessous. D'abord on introduira nos notations.

Soit t un nœud de l'arbre de décision et O_t le sous ensemble d'objets sous-tendu. La partition induite par c sur O_t est notée $\{O_{tj}/1 \leq j \leq K\}$. On estime la probabilité de la classe j de c par la proportion relative

$$p_j{}^t = \frac{card(O_{tj})}{card(O_t)}.$$

Comme on a vu dans l'introduction, dans une première étape on remplace toutes les variables par des atributs binaires. Ainsi, pour un nœud t, considérons, pour une variable binaire w, les deux nœuds descendants qu'elle détermine : t_l et t_r.

On forme alors le tableau de contingence croisant w et c:

c / w	1	...	k	...	K	
t_l	$p_1^{t_l}$	...	$p_k^{t_l}$	...	$p_K^{t_l}$	1
t_r	$p_1^{t_r}$	...	$p_k^{t_r}$	...	$p_K^{t_r}$	1
t	p_1^t	...	p_k^t	...	p_K^t	1

où on a installé les proportions conditionnelles

$$p_l{}^t = \frac{card(O_{t_l})}{card(O_t)}\,;\; p_r{}^t = \frac{card(O_{t_r})}{card(O_t)}\,;$$

$$p_k{}^{t_l} = \frac{card(O_k \cap O_{t_l})}{card(O_{t_l})}\,;\; p_k{}^{t_r} = \frac{card(O_k \cap O_{t_r})}{card(O_{t_r})}\,;\; p_k{}^t = \frac{card(O_k \cap O_t)}{card(O_t)}\ .$$

2.2.1 Le Coefficient de Gini

L'idée fondamentale dans la construction d'un arbre de décision binaire est la sélection, dans chaque nœud t, d'un attribut binaire, de façon que les nœuds descendants, t_l et t_r, soient plus purs que le nœud parent, t. On commence donc par définir une mesure d'impureté. Par exemple, l'indice d'impureté de Gini concernant le nœud t est:

$$\varphi(p_1^t, \ldots, p_K^t) = \sum_{1 \leq i \neq j \leq K} p_i^t p_j^t = 1 - \sum_{1 \leq i \leq K} (p_i^t)^2 \tag{1}$$

φ satisfait les conditions:

- $\varphi(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}) =$ maximum ;
- $\varphi(1, 0, \ldots, 0) = \varphi(0, 1, 0, \ldots, 0) = \ldots = \varphi(0, \ldots, 0, 1) = 0$;
- φ est une fonction symétrique des quantités $p_1^t, p_2^t, \ldots, p_K^t$.

C'est à dire, l'impureté est maximale quand toutes les classes sont également mélangées dans le nœud et minimale quand le nœud ne contient qu'une classe. φ correspond à une notion de dissimilarité. φ est d'autant plus grand que la création du nœud t n'est pas discriminante. Pour un nœud t, la règle de Gini consiste en choisir la variable binaire
w - qui coupe t en t_l et t_r - qui maximize la diminuition de l'impureté, donnée par:

$$G(t, w) = \varphi(t) - p_l^t \varphi(t_l) - p_r^t \varphi(t_r) \tag{2}$$

Ce coefficient se met sous la forme:

$$G(t, w) = p_l^t \sum_{j=1}^{K} (p_j^{t_l})^2 + p_r^t \sum_{j=1}^{K} (p_j^{t_r})^2 - \sum_{j=1}^{K} (p_j^t)^2 \tag{3}$$

$$G(t, w) = p_l^t \textstyle\sum_{j=1}^{K} (p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^{K} (p_j^{t_r} - p_j^t)^2 - p_l^t \sum_{j=1}^{K} (p_j^t)^2 - p_r^t \sum_{j=1}^{K} (p_j^t)^2 +$$

$$2p_l^t \sum_{j=1}^K p_j^t p_j^{t_l} + 2p_r^t \sum_{j=1}^K p_j^t p_j^{t_r} - \sum_{j=1}^K (p_j^t)^2 =$$

$$p_l^t \sum_{j=1}^K (p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^K (p_j^{t_r} - p_j^t)^2 - (p_l^t + p_r^t + 1) \sum_{j=1}^K (p_j^t)^2 + 2(\sum_{j=1}^K (p_l^t p_j^{t_l} + p_r^t p_j^{t_r}).p_j^t)$$

$$p_l^t \sum_{j=1}^K (p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^K (p_j^{t_r} - p_j^t)^2 - 2 \sum_{j=1}^K (p_j^t)^2 + 2(\sum_{j=1}^K (p_l^t p_j^{t_l} + p_r^t p_j^{t_r}).p_j^t)$$

Or,

$$p_j^t = p_l^t p_j^{t_l} + p_r^t p_j^{t_r} \tag{4}$$

et par conséquent, les deux dernières termes de $G(t,w)$ s'anullent:

$$G(t,w) = p_l^t \sum_{j=1}^K (p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^K (p_j^{t_r} - p_j^t)^2. \tag{5}$$

Relativement au tableau ci-dessus, si on adopte la représentation géométrique de l'analyse des correspondances de t_l et t_r par leurs profils respectifs, à travers $J = \{1, 2, \ldots, j, \ldots, K\}$, on obtient le nuage formé de deux sommets pesants:

$$\{(\mathrm{p_J^{t_l}}, \mathrm{p_l^t}), (\mathrm{p_J^{t_r}}, \mathrm{p_r^t})\} \tag{6}$$

dont le centre de gravité est le point p_J^t [cf. (4)]. De sorte que le coefficient de Gini, $G(t,w)$ (c.f. (5)), n'est autre que le moment total d'inertie de ce nuage par rapport à la métrique euclidienne ordinaire.

Dans l'expression de $G(t,w)$ (c.f. (5)), la quantité $p_j^{t_r} - p_j^t$ devient (cf. (4)) :

$$p_j^{t_r} - p_j^t = \frac{p_j^t - p_l^t p_j^{t_l}}{p_r^t} - p_j^t = \frac{(1-p_r^t)p_j^t - p_l^t p_j^{t_l}}{p_r^t} = \frac{(p_l^t)p_j^t - p_l^t p_j^{t_l}}{p_r^t}.$$

$$\text{Alors} \quad \mathrm{p_j^{t_r}} - \mathrm{p_j^t} = \frac{\mathrm{p_l^t}}{\mathrm{p_r^t}}(\mathrm{p_j^t} - \mathrm{p_j^{t_l}}) \tag{7}$$

Donc, $G(t,w)$ devient $p_l^t \sum_{j=1}^K (p_j^{t_l} - p_j^t)^2 + p_r^t (\frac{p_l^t}{p_r^t})^2 \sum_{j=1}^K (p_j^t - p_j^{t_l})^2$, c'est à dire,

$$G(t,w) = \frac{p_l^t}{p_r^t} \sum_{1 \le j \le K} (p_j^{t_l} - p_j^t)^2 \qquad (Coefficient\ de\ GINI) \tag{8}$$

On a la nullité de $G(t,w)$ en cas d'indépendance entre la variable prédictive w et la variable à predire, c. Puisque $\forall t, \forall w, G(t,w) \ge 0$, l'impureté diminue toujours, en descendant du nœud t aux nœuds t_l et t_r.

Remarquons enfin que, compte tenu de l'expression du moment total d'inertie en fonction de la somme pondérée des carrés des distances mutuelles entre points du nuage, on a :

$$G(t,w) = p_l^t p_r^t \sum_{1 \le j \le K} (p_j^{t_l} - p_j^{t_r})^2 \tag{9}$$

2.2.2 Le coefficient de Shannon

Considérons maintenant comme mesure d'impureté du nœud t non plus l'indice de Gini mais la quantité d'information de Shannon

$$I(t) = - \sum_{1 \le j \le K} p_j^t log_2(p_j^t) \tag{10}$$

$I(t)$ mesure la quantité d'information nécessaire pour classifier les individus ; si les p_j^t son tous les mêmes, alors l'uncertitude est maximale et l'information nécessaire, $I(t)$, est maximale. l'information est nulle quand il n'y a qu'un p_j^t non nul. On cherche toujours à maximizer la diminution d'impureté (le gain d'information), donnée par

$$S(t,w) = I(t) - p_l^t I(t_l) - p_r^t I(t_r).$$

Après quelques calculs, on obtient

$$S(t,w) = p_l \sum_{1\leq j\leq K} p_j^{t_l} log_2(\frac{p_j^{t_l}}{p_j^t}) + p_r \sum_{1\leq j\leq K} p_j^{t_r} log_2(\frac{p_j^{t_r}}{p_j^t}) \quad (Coefficient\ de\ SHANNON) \qquad (11)$$

2.2.3 Le critère de Twoing

On suppose que le nombre K de classes à prédire est supérieur à 2- Ce critère apparait alors pour les deux stratégies adoptées dans [Breiman et al. 1984, p.103] et permettant de réduire la complexité en contraignant l'espace de recherche- Considérons une variable prédictive candidate ayant L modalités. La première stratégie consiste à binariser cette variable sur la base de l'indice de Gini, en l'optimisant par rapport à toutes les bipartitions (partitions en 2 classes) de l'ensemble des K classes à prédire. On verra au chapitre 3 que l'ordre de la complexité est alors $(2^{K-1}-1)O(L)$.

La deuxième stratégie consiste à considérer toutes les variables binaires issues de la variable prédictive (il y en a $2^{L-1}-1$) et, respectivement, à évaluer chacune d'entre elles, par rapport à la variable binaire à prédire w (résultant d'une bipartition) qui lui est la mieux associée. Les auteurs précités démontrent que cette bipartition doit être choisie comme suit :

$$C_1(w) = \{j : p_j^{t_l} \geq p_j^{t_r}\} \quad \text{et} \quad C_2(w) = \{j : p_j^{t_l} < p_j^{t_r}\};$$

et, on montre que le critère optimisé prend la forme dite de "Twoing":

$$T(t,w) = \frac{p_l^t p_r^t}{4}[\sum_{1\leq j\leq K} |p_j^{t_l} - p_j^{t_r}|]^2 \qquad (Critère\ de\ TWOING) \qquad (12)$$

(Le terme entre parenthèses représente la distance "city-block" entre $p_J^{t_l}$ et $p_J^{t_r}$. Le terme $p_l^t p_r^t/4$ privilégie les divisions avec $p_l^t \approx p_r^t$).

Le critère de Gini et celui de Twoing sont ainsi équivalents pour un nombre K de classes à prédire égal à 2. Cependant, nous retiendrons la forme directe (12) du critère pour le tester dans le cas où le nombre K de classes est supérieur à 2.

2.2.4 Le critère du Chi-deux

L'expression du critère du Chi-deux

$$\chi^2 = \sum_{\mathrm{i}=1}^{\mathrm{p}} \sum_{\mathrm{j}=1}^{\mathrm{q}} \frac{\mathrm{n}[\mathrm{n_{ij}} - \frac{\mathrm{n_{i.}\,n_{.j}}}{\mathrm{n}}]^2}{\mathrm{n_{i.}\,n_{.j}}}, \qquad (13)$$

proposé par [Pearson 1904], représente une distance entre deux tableaux de contingence: celui qu'on a observé et le tableau qu'on obtiendrait sous hypothèse d'indépendance entre les deux variables à modalités, u et v. Cette distance est fonction de p, q et N. Il en existe deux versions "indices d'association", indépendantes de p, q et de N: L'indice de Tchuprov, $\tau = \chi^2/n\sqrt{(p-1)(q-1)}$ et la version donnée par Cramer (1946), $\chi^2_{norm} = \chi^2/(n.min[(p-1),(q-1)])$. On a la nullité de τ et χ^2_{norm} en cas d'indépendance entre les deux variables et la valeur maximale (c'est à dire, 1), dans le cas d'association complète.

Malgré toute sa popularité, le Chi-deux n'est pas toujours consideré par les specialistes comme une bonne mesure d'association. Néanmoins, l'existence d'un test d'indépendance entre variables, basé sur la statistique du χ^2, justifie pleinement son utilisation.

Dans notre cas, où il s'agit de choisir une variable binaire w pour couper le nœud t de l'arbre de décision, on va trouver l'expression de χ^2 d'une façon indirècte et géométrique.

Relativement au nuage de deux sommets pesants(cf.(6)), considéré ci-dessus, si maintenant on considère le moment total d'inertie par rapport à la métrique du χ^2, par exemple, on obtient,

$$p_l^t d^2_{\chi^2}(p_J^{t_l}, p_J^t) + p_r^t d^2_{\chi^2}(p_J^{t_r}, p_J^t) = p_l^t \sum_{j=1}^K \frac{1}{p_j^t}(p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^K \frac{1}{p_j^t}(p_j^{t_r} - p_j^t)^2 =_{(cf.(\ 7))}$$

$$p_l^t \sum_{j=1}^K \frac{1}{p_j^t}(p_j^{t_l} - p_j^t)^2 + p_r^t (\frac{p_l^t}{p_r^t})^2 \sum_{j=1}^K \frac{1}{p_j^t}(p_j^t - p_j^{t_l})^2 =$$

$$\frac{p_l^t}{p_r^t} \sum_{j=1}^K \frac{1}{p_j^t}(p_j^{t_l} - p_j^t)^2. \tag{14}$$

Cette expression n'est autre que χ^2/n_t, où χ^2 est le coefficient usuel du chi-deux pour le tableau ci-dessus, et $n_t = card(O_t)$.

En effet, dans l'expression (13), pour le nœud t de l'arbre de décision binaire $(p = 2, q = K)$ et pour tout $j, 1 \leq j \leq K$, on a:

$$\frac{\mathrm{n^t[n^t_{1j} - \frac{n^t_{1.} n^t_{.j}}{n^t}]^2}}{\mathrm{n^t_{1.} n^t_{.j}}} + \frac{\mathrm{n^t[n^t_{2j} - \frac{n^t_{2.} n^t_{.j}}{n^t}]^2}}{\mathrm{n^t_{2.} n^t_{.j}}}. \tag{15}$$

La deuxième quantité se met sous la forme

$$\frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{(n^t_{2.} n^t_{.j}/n^t)},$$

après avoir remplacé au numérateur n^t_{2j} par $n^t_{.j} - n^t_{1j}$ et $n^t_{2.}$ par $n^t - n^t_{1.}$.

L'expression (15) devient

$$\frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{n^t_{.j}} (\frac{n^t}{n^t_{1.}} + \frac{n^t}{n^t_{2.}}).$$

Par conséquent, dans ce cas,

$$\chi^2 = (\frac{n^t}{n^t_{1.}} + \frac{n^t}{n^t_{2.}}) \sum_{j=1}^K \frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{n^t_{.j}} = (1/p_{1.} + 1/p_{2.}) \sum_{j=1}^K \frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{n^t n^t_{.j}/n^t} =$$

$$(1/p_{1.} + 1/p_{2.}) \sum_{j=1}^K \frac{1}{p_{.j}} \frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{n^t};$$

$$\frac{\chi^2}{n^t} = (1/p_{1.} + 1/p_{2.}) \sum_{j=1}^K \frac{1}{p_{.j}} \frac{(n^t_{1j} - (n^t_{1.} n^t_{.j}/n^t))^2}{n^{t2}} =$$

$$(1/p_{1.} + 1/p_{2.}) \sum_{j=1}^K \frac{1}{p_{.j}} (p_{1j} - p_{1.} p_{.j})^2 =$$

$$(1/p_{1.} + 1/p_{2.}) \sum_{j=1}^K \frac{p^2_{1.}}{p_{.j}} \frac{(p_{1j} - p_{1.} p_{.j})^2}{p^2_{1.}} = (p_{1.} + \frac{p^2_{1.}}{p_{2.}}) \sum_{j=1}^K \frac{1}{p_{.j}} (p_j^1 - p_{.j})^2 =$$

$$\frac{p_{1.}}{p_{2.}} \sum_{j=1}^K \frac{1}{p_{.j}} (p_j^1 - p_{.j})^2.$$

Si on introduit nos notations, on retrouve l'expression (14),

$$\chi^2/n_t = \frac{p_l^t}{p_r^t} \sum_{1 \leq j \leq K} \frac{1}{p_j^t}(p_j^{t_l} - p_j^t)^2 \qquad (Coefficient\ du\ \chi^2/n_t) \tag{16}$$

Terminons en signalant que compte tenu de l'expression inertielle du critère, on a:

$$\chi^2/n_t = p_l^t p_r^t \sum_{1 \leq j \leq K} \frac{1}{p_j^t}(p_j^{t_l} - p_j^{t_r})^2 \tag{17}$$

2.2.5 Le critère de Hellinger

Comme c'est le cas pour les indices de Gini, Twoing et χ^2, ce critère est défini au moyen d'une distance entre les deux distributions empiriques de probabilité $\{p_j^{t_l}/1 \leq j \leq K\}$ et $\{p_j^{t_r}/1 \leq j \leq K\}$, qu'on représente dans la portion positive de la sphère unité de l'espace $\mathbb{R}^K$. Plus précisèment les deux précédentes distributions sont respectivement représentés par les deux points suivants :

$$(\sqrt{p_1^{t_l}}, \sqrt{p_2^{t_l}}, \ldots, \sqrt{p_K^{t_l}}) \ et \ (\sqrt{p_1^{t_r}}, \sqrt{p_2^{t_r}}, \ldots, \sqrt{p_K^{t_r}}). \tag{18}$$

Ainsi, en adoptant la métrique euclidienne ordinaire, on a pour l'expression de ce critère:

$$H(t_l, t_r) = \sum_{j=1}^{K} (\sqrt{p_j^{t_l}} - \sqrt{p_j^{t_r}})^2 = 2(1 - A(t_l, t_r)); \tag{19}$$

où

$$A(t_l, t_r) = \sum_{j=1}^{K} \sqrt{p_j^{t_l} p_j^{t_r}} \tag{20}$$

Et ce coefficient $A(t_l, t_r)$, sur lequel nous reviendrons ci-dessous, est précisèment celui introduit par Matusita ([Matusita 1955]).

Nous allons à présent considérer, relativement aux représentations euclidiennes qui ont fait des indices de Gini, du χ^2 et de Hellinger, des distances au carré, entre les points de $\mathbb{R}^K$ représentant les deux distributions $\{p_j^{t_l}/1 \leq j \leq K\}$ et $\{p_j^{t_r}/1 \leq j \leq K\}$, des coefficients de "similarité" (on dit encore "association"). Ces derniers auront à chaque fois, le sens d'un cosinus. Cet indice peut alors être développé par une approche statistique.

2.2.6 Les indices cosinus associés à la distance de Gini et du χ^2 .

Vu que c'est la métrique euclidienne ordinaire qui est sous jacente à la distance de Gini; on a, en plaçant l'origine de l'espace au centre de gravité du nuage (6), l'indice cosinus associé:

$$Cos_G(p_J^{t_l}, p_J^{t_r}) = \frac{\sum_{j=1}^{K} (p_j^{t_l} - p_j^t)(p_j^{t_r} - p_j^t)}{\sqrt{[\sum_{j=1}^{K} (p_j^{t_l} - p_j^t)^2][\sum_{j=1}^{K} (p_j^{t_r} - p_j^t)^2]}} \tag{21}$$

Cet indice détermine un coefficient d'affinité entre les deux distributions et que nous pourrons noter $AG(p_J^{t_l}, p_J^{t_r})$ (G pour Gini).

De façon analogue, l'indice cosinus associé à la distance du χ^2 s'écrit:

$$Cos_{\chi^2}(p_J^{t_l}, p_J^{t_r}) = \frac{\sum_{j=1}^{K} (p_j^{t_l} - p_j^t)(p_j^{t_r} - p_j^t)/p_j^t}{\sqrt{[\sum_{j=1}^{K} (p_j^{t_l} - p_j^t)^2/p_j^t][\sum_{j=1}^{K} (p_j^{t_r} - p_j^t)^2/p_j^t]}} \tag{22}$$

Cet indice détermine également un coefficient d'affinité entre les deux distributions et que nous pourrons noter $\mathrm{A}\chi^2(p_J^{t_l}, p_J^{t_r})$.

On aurait certes pu rapporter l'indice cosinus à l'origine O de l'espace $\mathbb{R}^K$; et on aurait alors obtenu deux autres expressions d'un coefficient d'affinité qu'on pourrait noter $AG_O(p_J^{t_l}, p_J^{t_r})$ et $\mathrm{A}\chi^2_O(p_J^{t_l}, p_J^{t_r})$.

Les expressions ci-dessus, (21) et (22), peuvent naturellement - à partir de considérations baricentriques - être simplifiées. L'analyse de ce type de coefficients est consideré dans [Lerman et Peter (1985)]. On y montre notamment que l'indice cosinus dans l'espace $\mathbb{R}^K$ est identique au coefficient de corrèlation entre t_l et t_r, regardés comme des variables, pour le nuage dual (tel consideré en analyse des correspondances), où il sera question ici de K sommets dans $\mathbb{R}^2$. Ce coefficient de corrélation avait été introduit indépendamment [Lerman & Tallur (1980), Tallur (1988)].

2.2.7 Le coefficient d'affinité de Matusita et ceux de Bacelar-Nicolau.

Le coefficient d'affinité a été proposé par K. Matusita en 1955. Dans le domaine de l'analyse classificatoire, Matusita a utilisé ce coefficient pour évaluer une procédure de classification dans le cas multinormal et aussi pour choisir le nombre de classes dans une classification automatique ; [Bacelar Nicolau (1980,1982 a-b)] et [Costa Nicolau 1985] ont traité l'affinité comme une similarité du type probabiliste, en analyse classificatoire aussi.

On donne ici l'expression du coefficient dans notre cas, où il s'agit de minimizer l'affinité entre les deux nœuds descendants t_l et t_r (pour une expression plus général, consulter, par exemple, [Bacelar Nicolau 1988]):

$$A(t_l, t_r) = \sum_{j=1}^{K} \sqrt{p_j^{t_l}.p_j^{t_r}}. \qquad (Coefficient\ de\ MATUSITA) \tag{23}$$

Ce coefficient, qu'on désigne par coefficient "brut" d'affinité, est un produit scalaire entre les vecteurs $(\sqrt{p_1^{t_l}}, \sqrt{p_2^{t_l}}, \ldots, \sqrt{p_K^{t_l}})$ et $(\sqrt{p_1^{t_r}}, \sqrt{p_2^{t_r}}, \ldots, \sqrt{p_K^{t_r}})$.

- $A(t_l, t_l) = A(t_r, t_r) = 1$ (affinité 1 pour des profils identiques)
- $A(t_l, t_r) = A(t_r, t_l)$ (symétrie)
- $0 \leq A(t_l, t_r) \leq 1$ (0 pour des vecteurs orthogonaux)

Le coefficient d'affinité $A(t_l, t_r)$ de Matusita est, comme nous l'avons dit (cf. §???) directement associé à la distance de Hellinger. Pour la représentation que suppose cette distance euclidienne ordinaire (cf. (18)), il s'agit du cosinus rapporté à l'origine, entre les deux vecteurs d'extremités respectives, les deux sommets de $\mathbb{R}^K$ définis en (18). On procéde alors à deux types de normalisation statistique de $A(t_l, t_r)$. La première est non-paramètrique, à caractère permutationnel et se réfère au théorème de Wald et Wolfovitz (1944). Elle est consideré dans notre formalisation [Lerman 1976, 1981, 1992a] dés lors qu'il s'agit de comparer des variables numériques quantitatives. Le premier (resp. second) vecteur de (18), est alors regardé comme la suite des valeurs d'une première variable v (resp. seconde variable w) [Bacelar-Nicolau 1981]. Le coefficient obtenu est alors au facteur multiplicatif $\sqrt{K-1}$ près, le coefficient de corrélation entre les deux variables numériques. On obtient alors

$$\mathrm{B(t_l, t_r)} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{K}} (\sqrt{\mathrm{p_j^{t_l}}} - \mathrm{q(t_l)})(\sqrt{\mathrm{p_j^{t_r}}} - \mathrm{q(t_r)})}{\sqrt{[\sum_{\mathrm{j=1}}^{\mathrm{K}} (\sqrt{\mathrm{p_j^{t_l}}} - \mathrm{q(t_l)})^2][\sum_{\mathrm{j=1}}^{\mathrm{K}} (\sqrt{\mathrm{p_j^{t_r}}} - \mathrm{q(t_r)})^2]}} \tag{24}$$

où $q(t_l) = \frac{1}{K} \sum_{j=1}^{K} \sqrt{p_j^{t_l}}$ et $q(t_r) = \frac{1}{K} \sum_{j=1}^{K} \sqrt{p_j^{t_r}}$.

Dans la conception du second coefficient, on se situe dans un contexte de Statistique inférentielle. Aux profils $(p_1^{t_l}, p_2^{t_l}, \ldots, p_K^{t_l})$ et $(p_1^{t_r}, p_2^{t_r}, \ldots, p_K^{t_r})$ correspondent les vecteurs de fréquences $(n_{11}, n_{12}, \ldots, n_{1K})$ et $(n_{21}, n_{22}, \ldots, n_{2K})$, qu'on considére comme deux observations des vecteurs aléatoires, $(N_{11}, N_{12}, \ldots, N_{1K})$ et $(N_{21}, N_{22}, \ldots, N_{2K})$. Or, ces deux vecteurs ont une distribution multinomial, ce qui nous permet de centrer et réduire le coefficient "brut" d'affinité $A(t_l, t_r)$, en utilisant la méthode-δ (consulter Tiago de Oliveira (1982) pour un excellent développement de cette méthode). L'expression du coefficient d'affinité centré et réduit par la méthode δ([Bacelar-Nicolau 1988]),

$$A^*(t_l, t_r) = \frac{A(t_l,t_r) - \rho(t_l,t_r)}{\sqrt{1-\rho^2(t_l,t_r)}} \frac{2\sqrt{\gamma}}{\sqrt{a^2+a'^2}}$$

"mesure" la distance observée normalisée entre le coefficient "brut" d'affinité $A(t_l, t_r)$ et sa valeur théorique à la convergence, $\rho(t_l, t_r)$. L'idée consiste alors à prendre pour $\rho(t_l, t_r)$ une valeur de référence,

qui peut par exemple dans notre cas correspondre à la valeur du coefficient $A(R_l, R_r)$ à la racine R de l'arbre de décision. Ainsi, de cette façon, on évalue combien $A(t_l, t_r)$ s'écarte de ce qu'était sa valeur à la racine de l'arbre - Mais, il importe de pouvoir descendre de la racine de l'arbre au moyen du seul coefficient A (cf. (23)). Dans ces conditions, la forme du coefficient que nous proposons est :

$$\mathrm{A^*(t_l, t_r)} = \frac{\mathrm{A(t_l, t_r) - A(R_l, R_r)}}{\sqrt{1 - \mathrm{A^2(R_l, R_r)}}} \frac{2\sqrt{\gamma}}{\sqrt{\mathrm{a_l^2 + a_r^2}}} \tag{25}$$

$\gamma = min(n_l, n_r)$, $a_l^2 = \frac{\gamma}{n_l}$ et $a_r^2 = \frac{\gamma}{n_r}$. R est la racine de l'arbre de décision et par conséquent $A(R_l, R_r)(= \sum_{j=1}^{K} \sqrt{p_j^{R_l}.p_j^{R_r}})$ représente l'affinité des profils t_l et t_r à la racine.

Il y a lieu de souligner que ce type de traitement peut être considéré pour n'importe lequel des coefficients et qui peut déjà être normalisé, soit à partir de considérations géométriques, soit à partir de considérations de statistique non paramétrique.

2.2.8 L'indice de Goodman-Kruskal et le critère de Haldane, Light et Margolin

L'indice proposé en 1954 par L. Goodman et W. Kruskal dissymétrise la relation entre les deux variables à comparer u et v. Il s'agit de prédire, pour un individu tiré au hasard de la population, à quelle modalité m_v de la variable v il appartient, en supposant connue la modalité m_u de la variable u qui lui correspond. Si on veut estimer m_v sans tenir compte de m_u, alors la proportion de prédictions correctes sera $\sum_{v=1}^{q}(\frac{n_{.v}}{n})^2$. Si on tient compte de la modalité de u, m_u, à laquelle l'individu appartient, alors la proportion de prédictions correctes sera $\sum_{u=1}^{p}\sum_{v=1}^{q}\frac{n_{uv}^2}{nn_{u.}}$. La différence relative entre les proportions de prédictions incorrectes dans les deux cas, qu'on désigne par τ_b, constitue l'indice d'association de Goodman-Kruskal entre les deux variables u et v:

$$\tau_b = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q}\frac{n_{uv}^2}{nn_{u.}} - \sum_{v=1}^{q}(\frac{n_{.v}}{n})^2}{1 - \sum_{v=1}^{q}(\frac{n_{.v}}{n})^2}.$$

$0 \leq \tau_b \leq 1$. $\tau_b = 0$ en cas d'indépendance et $\tau_b = 1$ en cas d'association compléte, c'est à dire, quand les variables u et v sont identiques (se réporter à [Goodman & Kruskal 1979] ou [Marcotorchino 1984a] pour une discussion de cette indice).

[Lauro et D'Ambra 1984], montrent l'intérêt croissant de τ_b, en le comparant avec le coefficient $\frac{\chi^2}{n}$:

- τ_b, contrairement à $\frac{\chi^2}{n}$, a une limite supérieure définie ;
- τ_b est moins sensible que $\frac{\chi^2}{n}$ quand les marges de v sont très assymétriques ;
- τ_b augmente, contrairement à $\frac{\chi^2}{n}$, si la variabilité de v diminue ;
- $\frac{\chi^2}{n}$ est trés sensible pour de petites fréquences théoriques.

[Lauro et D'Ambra 1984] le considérent donc comme un bon coefficient de prédictabilité, ce qui nous interesse car, dans notre cas, il s'agit de prédire la variable
$c : O \longrightarrow \{c_1, c_2, ..., c_k, ..., c_K\}$. En fait, ce qui nous intéresse c'est le numérateur de τ_b, une fois que le dénominateur est constant, parce qu'il s'agit de choisir une variable binaire w qui soit la plus prédictive par rapport à la variable c (le numérateur de τ_b, multiplié par 2, constitue le critère de Haldane, Light et Margolin). L'expression du numérateur de τ_b dans notre cas ($p = 2$ et $q = K$) devient, pour le nœud t de l'arbre de décision:

$$\sum_{i=1}^{2}\sum_{j=1}^{K}\frac{(n_{ij}^t)^2}{n^t n_{i.}^t} - \sum_{j=1}^{K}(\frac{n_{.j}^t}{n^t})^2 = \sum_{i=1}^{2}\sum_{j=1}^{K}(\frac{n_{ij}^t}{n_{i.}^t})^2(\frac{n_{i.}^t}{n^t}) - \sum_{j=1}^{K}(\frac{n_{.j}^t}{n^t})^2 =$$

$$\sum_{j=1}^{K}\left(\frac{n_{1j}^t}{n_{1.}^t}\right)^2\left(\frac{n_1^t}{n^t}\right)+\sum_{j=1}^{K}\left(\frac{n_{2j}^t}{n_{2.}^t}\right)^2\left(\frac{n_2^t}{n^t}\right)-\sum_{j=1}^{K}\left(\frac{n_j^t}{n^t}\right)^2=$$

$$p_l^t\sum_{j=1}^{K}(p_j^{t_l})^2+p_r^t\sum_{j=1}^{K}(p_j^{t_r})^2-\sum_{j=1}^{K}(p_j^t)^2 \tag{26}$$

Ce coefficient est exactement le coefficient de Gini (cf. 3), décrit ci-dessus. On trouve alors que, pour des tableaux de contingence 2xK le coefficient de Gini n'est autre que le coefficient de Goodman-Kruskal,Haldane Light et Margolin.

2.3 Les coefficients relationnels ayant une base combinatoire et statistique.

On utilise ici le concept de comparaisons par paires entre deux variables qualitatives nominales u et v. Contrairement aux coefficients précedents, définis sur des tableaux de contingence, les critères qu'on va décrire ici sont définis sur les relations que les deux variables à comparer induisent sur l'ensemble O d'objets élémentaires.

Le concept de comparaisons par paires (introduit en 1785 par A. de Condorcet), négligé pendant longtemps, recouvre, a partir des années 1930, son importance avec les travaux de M. G. Kendall (1938), B. Babington-Smith (1939,1940), P. Moran (1947), H. A. David (1964), S. Régnier (1965), I. C. Lerman (1973,1981,1992a,b). On cite, par exemple, le critère de l'association simple, introduit en 1964 par Zahn, pour mesurer la distance entre une relation binaire quelconque et une relation d'équivalence ; le critère de l'écart à l'indépendance, inspiré dans le critère $\frac{ad-bc}{N}$ proposé par M. G. Kendall en 1970 pour un tableaux de contingence 2x2 ; les versions en paires du critère de Belson et du critère de Light et Margolin, ...etc. [Marcotorchino 1984b] présente une discussion très intéressante des avantages de la comparaison par paires . On va se limiter ici aux critères proposés par I. C. Lerman (consulter [Lerman 1992a,b] pour une description excellent des critères s, Q_1 et R exposés ci-dessous)

Notre objectif, dans la construction d'un arbre de décision binaire, c'est de comparer la variable binaire

$$w : O \longrightarrow \{t_l, t_r\}$$

avec la variable à prédire

$$c : O \longrightarrow \{c_1, c_2, ..., c_k, ..., c_K\},$$

pour choisir la variable binaire la plus prédictive. Le contexte est donc celui de la comparaison de deux variables qualitatives, qu'on désigne d'une façon générale par u et v, à partir de leur observation empirique sur un ensemble O d'objets élémentaires. Ces deux variables se présentent comme deux applications de l'ensemble O dans un ensemble C de codes ou catégories. L'ensemble C peut être muni d'une structure comme, par exemple, un préordre total ou partiel ; cependant, d'abord on ne considére que des variables qualitatives nominales où C ne se trouve muni d'aucune structure.

2.3.1 Le critère "brut" s (non centré ni réduit)

La variable u (resp. v) induit sur l'ensemble O une partition π (resp. χ), qui dèfinit une relation déquivalence ; pour comparer u et v, nous ne retenons que cette relation sur l'ensemble des objets élémentaires.

$$\pi = \{E_i / 1 \leq i \leq h\}; \qquad \chi = \{F_j / 1 \leq j \leq k\} \tag{27}$$

On peut représenter ces partitions par l'ensemble des paires d'objets réunis:

$$R(\pi) = \cup\{P_2(E_i)/1 \leq i \leq h\} \quad et \quad R(\chi) = \cup\{P_2(F_j)/1 \leq j \leq k\} \tag{28}$$

Ainsi, une mesure d'association entre les deux variables, surgit naturellement comme l'ensemble des paires d'objets qui sont réunis par les deux partitions ; soit:

$$s(\pi,\chi) = card[R(\pi) \cap R(\chi)] = card[R(\pi \wedge \chi)] = card[\cup\{P_2(E_i \cap F_j)/1 \leq i \leq h, 1 \leq j \leq K\}] \quad (29)$$

où $\pi \wedge \chi$ représente le croisement des deux partitions. Par conséquent,

$$s(\pi,\chi) = \sum\{\frac{n_{ij}(n_{ij}-1)}{2}/1 \leq i \leq h, 1 \leq j \leq K\} \quad (Critère\ s\ de\ Lerman) \quad (30)$$

où $n_{ij} = card[E_i \cap F_j], 1 \leq i \leq h, 1 \leq j \leq K$

2.3.2 Le critère Q_1 (centré et réduit)

On procéde ici à la normalisation du critère "brut", s, par rapport à une hypothèse d'absence de liaison entre les deux variables. Désignos par

$$t(\pi) = (m_1, m_2, \ldots, m_i, \ldots, m_h)$$

le type de la partition π ; c'est-à-dire, la suite des cardinaux de ses classes ($m_i = card(E_i)$); et par $t(\chi) = (n_1, n_2, \ldots, n_j, \ldots, n_k)$ le type de la partition χ . Pour introduire cette démarche de normalisation statistique il y a lieu d'associer à la partition π - définie par la variable u - l'ensemble P(n,t(π)) de toutes les partitions de même type que π. De façon analogue, on associe à χ l'ensemble P(n,t(χ)) de toutes les partitions de même type que χ. On a

$$card[P(n,t(\pi))] = \frac{n!}{m_1!\ldots m_i!\ldots m_h!} \text{ et } card[P(n,t(\chi))] = \frac{n!}{n_1!\ldots n_j!\ldots n_k!}$$

On définit maintenant la variable aléatoire $S = s(\pi^*, \chi^*)$ où $\pi^*(resp.\chi^*)$ est un élement pris avec une probabilité uniforme dans l'ensemble $P(n,t(\pi))$ (resp. $P(n,t(\chi))$). D'ailleurs, comme π^* et χ^* sont deux partitions aléatoires indépendantes, nous disons que (π^*, χ^*) est associé à (π,χ) dans une "hypothése d'absence de liaison"(h.a.l.).

La moyenne et la variance de S ont été obtenues par [Lerman 1973,1981]:

$$E(s(\pi^*,\chi^*)) = \lambda\mu \quad ; \quad var(s(\pi^*,\chi^*)) = \lambda\mu + \rho\sigma + \theta\xi - \lambda^2\mu^2 \quad (31)$$

où

$$\lambda = \sum_{i=1}^{h} \frac{m_i(m_i-1)}{\sqrt{2n(n-1)}} ; \rho = \sum_{i=1}^{h} \frac{m_i(m_i-1)(m_i-2)}{\sqrt{n(n-1)(n-2)}} ;$$

$$\theta = \frac{[\sum_{i=1}^{h} m_i(m_i-1)]^2 - 2\sum_{i=1}^{h} m_i(m_i-1)(2m_i-3)}{2\sqrt{n(n-1)(n-2)(n-3)}}$$

et

$$\mu = \sum_{j=1}^{k} \frac{n_j(n_j-1)}{\sqrt{2n(n-1)}} ; \sigma = \sum_{j=1}^{k} \frac{n_j(n_j-1)(n_j-2)}{\sqrt{n(n-1)(n-2)}} ;$$

$$\xi = \frac{[\sum_{j=1}^{k} n_j(n_j-1)]^2 - 2\sum_{j=1}^{k} n_j(n_j-1)(2n_j-3)}{2\sqrt{n(n-1)(n-2)(n-3)}}$$

L'indice qui nous interésse, celui centré et réduit, est:

$$Q_1(\pi,\chi) = \frac{s(\pi,\chi) - E(s(\pi^*,\chi^*))}{\sqrt{var(s(\pi^*,\chi^*))}} \qquad (Critère\ Q_1\ de\ Lerman) \quad (32)$$

Quelques auteurs, tels que L. Hubert [Hubert 1983], considèrent que la réduction de l'indice centré, $s(\pi,\chi) - E(s(\pi^*,\chi^*))$, doit avoir la même forme que celle proposée par M. G. Kendall en 1970, c'est à dire :

$$\frac{s(\pi,\chi) - E(s(\pi^*,\chi^*))}{Max[s(\pi^*,\chi^*) - E(s(\pi^*,\chi^*))]}.$$

Le probléme difficile de découvrir $Max[s(\pi^*, \chi^*)]$ dans le cas de deux variables partition a été étudié par I. C. Lerman [Lerman 1987,1988] et repris par H. Messatfa [Messatfa 1990] en utilisant des méthodes de programmation linéaire. On se limite ici à la normalisation statistique proposée par Lerman (critère Q_1).

[Lerman 1992b,p.85] présente une forme limite de Q_1, quand n tend vers l'infini. Avec nos notations et dans notre cas (2xK), l'expression de l'indice centré,

$$s(\pi, \chi) - E(s(\pi^*, \chi^*)),$$

dans sa forme limite se met sous la forme:

$$\sum_{1 \leq j \leq K} p_{lj}^2 + \sum_{1 \leq j \leq K} p_{rj}^2 - (p_{l.}^2 + p_{r.}^2)(\sum_{1 \leq j \leq K} p_{.j}^2) \tag{33}$$

Essayons une interprétation géométrique de l'indice centré. Le centre de gravité des sommets p_J^l et p_J^r, est le point, $p_{l.}p_J^l + p_{r.}p_J^r$, dont la j-ème coordonnée est:

$$p_{lj} + p_{rj} = p_{.j}$$

C'est donc le point qu'on notera $p_{.J}$

Décomposons la somme (33) ci-dessus :

$$p_{l.}^2[\sum_{1 \leq j \leq K} (p_j^l)^2 - \sum_{1 \leq j \leq K} p_{.j}^2] + p_{r.}^2[\sum_{1 \leq j \leq K} (p_j^r)^2 - \sum_{1 \leq j \leq K} p_{.j}^2] = \tag{34}$$

$$p_{l.}^2[\|p_J^l\|^2 - \|p_{.J}\|^2] + p_{r.}^2[\|p_J^r\|^2 - \|p_{.J}\|^2] \tag{35}$$

Remplaçons, dans cette expression (35), $p_{l.}^2$ et $p_{r.}^2$ par, respectivement, $p_{l.}$ et $p_{r.}$. On obtient

$$p_{l.}[\|p_J^l\|^2 - \|p_{.J}\|^2] + p_{r.}[\|p_J^r\|^2 - \|p_{.J}\|^2] =$$

$$p_{l.}(\|p_J^l\|^2 + \|p_{.J}\|^2 - 2 < p_J^l, p_{.J} >) + p_{r.}(\|p_J^r\|^2 + \|p_{.J}\|^2 - 2 < p_J^r, p_{.J} >), \tag{36}$$

car $p_{l.}p_J^l + p_{r.}p_J^r = p_{.J}$. L'expression (36) devient alors:

$$p_{l.}\|p_J^l - p_{.J}\|^2 + p_{r.}\|p_J^r - p_{.J}\|^2. \tag{37}$$

L'indice centré est donc de même nature que l'inertie du nuage

$$\{(p_J^l, p_l), (p_J^r, p_r)\}$$

par rapport à la métrique euclidienne ordinaire. Or cette inértie représente précisèment l'indice de Gini! (cf. (5)).

Le critère $Q_1(\pi, \chi)$ est d'autant plus grand que les partitions π ou χ tendent à être en classes de même taille. Si on veut un critère insensible à cet effet de taille, et ne mettant en évidence que la similarité des formes, on pourra prendre

$$R(\pi, \chi) = \frac{Q_1(\pi, \chi)}{\sqrt{Q_1(\pi, \pi)Q_1(\chi, \chi)}} \qquad (Critère\ R\ de\ Lerman) \tag{38}$$

On remarque que, dans des conditions très générales, ce critère tend vers le critère de K. Pearson dans le cas de deux atributs booléans et quand n est assez grand.

2.3.3 Le critère φ

On considére maintenant le problème de développer un critère qui privilégie les classes minoritaires; c'est-à-dire, les concepts qui sont rares. Ainsi, supposons que la variable qualitative à discriminer a trois classes - E, H et X - et que la variable prédictive est binaire:

	E	H	X	
B	(1,1)	(1,2)	(1,3)	m_1
C	(2,1)	(2,2)	(2,3)	m_2
	n_1	n_2	n_3	O

À chaque case (i,j) (i=1,2 ; j=1,2,3) nous associons un couple de partitions dont une seule classe n'est pas réduite à un singleton. Ainsi, à la première case, (B,E), on associe

$$\pi_1 = \{B, \{\{x\}/x \in O - B\}\} \text{ et } \chi_1 = \{E, \{\{y\}/y \in O - E\}\},$$

et on affecte la valeur du critère d'association entre les deux partitions:

$$s_{11} = s(\pi_1, \chi_1) = n_{11}(n_{11} - 1)/2$$

Pour le centrage et réduction de s_{11} on obtient, conformément à l'equation (31), les expressions suivantes:

$$\lambda_1 = m_1(m_1 - 1)/\sqrt{2n(n-1)}\,;\ \rho_1 = m_1(m_1 - 1)(m_1 - 2)/\sqrt{n(n-1)(n-2)}\,;$$

$$\theta_1 = \{[m_1(m_1 - 1)]^2 - 2m_1(m_1 - 1)(2m_1 - 3)\}/\sqrt{n(n-1)(n-2)(n-3)}.$$

Le numérateur de θ_1 devient:

$m_1(m_1 - 1)[m_1(m_1 - 1) - 2(2m_1 - 3)] = m_1(m_1 - 1)[m_1^2 - 5m_1 + 6] = m_1(m_1 - 1)[(m_1 - 2)(m_1 - 3)]$,

et donc $\theta_1 = m_1(m_1 - 1)(m_1 - 2)(m_1 - 3)/\sqrt{n(n-1)(n-2)(n-3)}$

$$\text{Alors} \qquad \mathrm{Q}(1,1) = \frac{\mathrm{s}_{11} - \lambda_1\mu_1}{\sqrt{\lambda_1\mu_1 + \rho_1\sigma_1 + \theta_1\xi_1 - \lambda_1^2\mu_1^2}} \tag{39}$$

où μ_1, σ_1 et ξ_1 ont la même forme que λ_1, ρ_1 et θ_1.

Le principe du critère est le suivant: La qualité de la prédiction est d'autant meilleure, qu'on prédit bien ce qui est rare. Posons dans ces conditions $p_j = n_j/n, j = 1, 2, 3$; et soit la famille suivante de critères:

$$\varphi(p_1)\{max[Q(1,1), Q(2,1)]\} + \varphi(p_2)\{max[Q(1,2), Q(2,2)]\} + \varphi(p_3)\{max[Q(1,3), Q(2,3)]\} \tag{40}$$

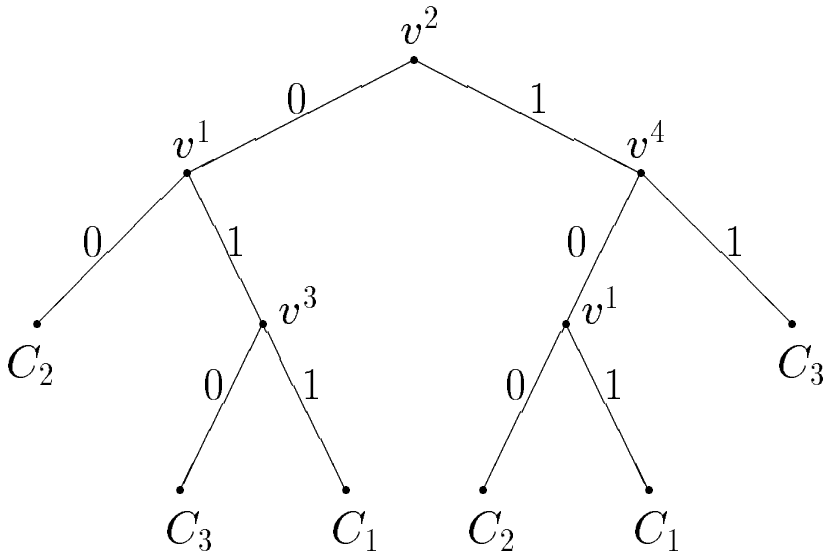
(Critère φ de Lerman)

φ est une fonction décroissante de p. On peut prendre pour $\varphi(p)$, $-Log_2(p)$ ou bien $1/p$, ou bien $1/p^\alpha$ où α est un paramétre réel compris entre 0 et 1.

3 Rappel de la méthode CART

3.1 Introduction

Dans les années 1963 et 1964 , les travaux de J. A. Sonquist et J. N. Morgan sur les arbres de régression ont été le point de départ pour les développements des techniques de segmentation ou de discrimination par arbre, initialisés par Messenger et Mandell (1972) et Morgan et Messenger (1973). Ces arbres de décision sont aujourd'hui très populaires, parce qu'il s'agit d'une méthode logique de discrimination qui requière peu d'hypothèses, qui sélectionne les variables au fur à mesure que l'arbre se construit (ce qui permet le traitement des cas où les variables descriptives sont nombreuses) et de plus, les règles de discrimination produites, qui sont très simples à utiliser et même à interpréter, ont à la fois un pouvoir explicatif et décisionnel. Dans ce travail, nous ne considérons que des arbres de décision binaires bien qu'on ait l'intention d'étendre les apports méthodologiques introduits ici au cas non binaire. On présente, comme illustration, l'arbre de décision suivant, où $\mathcal{V} = \{ v^m \ / \ 1 \leq m \leq 4 \}$ désigne l'ensemble des attributs binaires descriptifs et $\mathcal{C} = \{ C_j \ / \ 1 \leq j \leq 3 \}$ celui des classes (ou groupes) à discriminer.



$$(\neg v^2 \wedge v^1 \wedge v^3) \vee (v^2 \wedge \neg v^4 \wedge v^1) \longrightarrow C_1$$
$$(\neg v^2 \wedge \neg v^1) \vee (v^2 \wedge \neg v^4 \wedge \neg v^1) \longrightarrow C_2$$
$$(\neg v^2 \wedge v^1 \wedge \neg v^3) \vee (v^2 \wedge v^4) \longrightarrow C_3$$

FIG. 1 - Exemple illustratif d'un arbre de décision et des régles de discriminantion logiques correspondantes.

Généralement, la plupart des variables prédictives ne sont pas binaires. Dans ces conditions, pour la construction d'un arbre de décision binaire, il faut d'abord "binariser" toutes les variables descriptives de l'ensemble $\mathcal{V}$. Dans notre cas, $\mathcal{V}$ ne comprend que des variables (on dit encore attributs) qualitatives nominales présentant plus que 2 modalités (catégories) et ainsi, a priori, l'interrogation dans un nœud t, de l'une de ces variables - comprenant L modalités - conduit à la remplacer par $(2^{L-1} - 1)$ attributs binaires. Cette croissance exponentielle du nombre d'attributs binaires devient très vite prohibitive, en particulier dans nos données où L peut atteindre un ordre de 20^4. Si $K = 2$ (K est le nombre de concepts à discriminer) alors, dépendant du critère choisi, on peut parfois réduire la complexité

pour trouver la variable binaire la plus prédictive, au moyen d'un algorithme de complexité $O(L)$ (méthode CART de [Breiman et al., 1984]). Toutefois, la complexité augmente exponentiellement pour K supérieur à 2 et, d'ailleurs, pour la plupart des critères que nous considérons dans ce travail, cette réduction devient inapplicable même pour $K = 2$.

3.2 Rappel du principe général de la méthode CART, en cas où les attributs sont binaires.

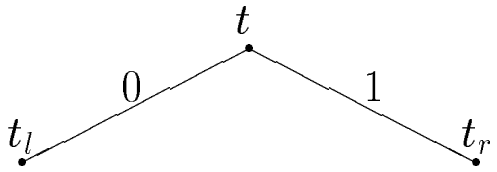
La méthode de discrimination par arbre binaire CART ("Classification And Regression Trees"), proposée par Breiman, Friedman, Olshen et Stone (1984) inclut des solutions qui répondent aux critiques les plus importantes faîtes aux arbres de décision, en particulier en ce qui concerne la règle d'arrêt de division d'un nœud qui alors, n'est plus considerée. Bien que cette méthode soit très générale, on se limite ici à faire un rappel des solutions apportées par CART dans le cas ou les attributs de description sont déjà binaires. C'est ensuite que nous décrirerons comment CART traite les attributs qualitatifs nominaux, c'est à dire, la façon de les binariser.

Désignons par $\mathcal{O}$= $\{o_i/1 \leq i \leq n\}$ l'ensemble d'apprentissage et par $\mathcal{V}$= $\{v^j/1 \leq j \leq p\}$ celui des variables ou attributs de description, qui seront les variables prédictives et que, encore une fois, nous supposons ici, toutes binaires. Ainsi, un même élément o_i de $\mathcal{O}$ se trouve décrit par la suite des valeurs des différents variables sur o_i :

$$[v^1(o_i), \ldots, v^j(o_i), \ldots, v^p(o_i)], 1 \leq i \leq n.$$

Désignons par E l'ensemble des états de description possibles des p variables binaires explicatives ($E = \{e_1, e_2, \ldots, e_{2^p}\}$ et peut être identifié avec le cube logique $\{0,1\}^p$). Une règle de discrimination, qui est une fonction de E dans $\{1, 2, \ldots, K\}$, permet d'affecter chaque état de E à une des classes $C_1, C_2, \ldots, C_K$.

La construction d'un arbre de décision binaire s'effectue d'une façon recursive à travers de divisions successives de l'ensemble d'états E (qui correspond à la racine de l'arbre) en sous-ensembles correspondants aux nœuds descendants. Cette construction s'effectue sur la base de l'ensemble d'apprentissage $\mathcal{E}$ par segmentations successives. Ainsi, la première division ségmente la racine de l'arbre en deux nœuds descendants. Pour effectuer la division d'un nœud t, il faut choisir, parmi les p attributs binaires de l'ensemble $\mathcal{V}$, celui qui maximise la diminuition d'impureté entre le nœud parent t et ses descendants, t_l et t_r.



Il y a divers critères de choix de l'attribut binaire (voir chapitre antérieur). La méthode CART commence pour utiliser l'indice de Gini par mesurer l'impureté d'un nœud t:

$$\varphi(p_1^t, \ldots, p_K^t) = \sum_{1 \leq i \neq j \leq K} p_i^t p_j^t = 1 - \sum_{1 \leq i \leq K} (p_i^t)^2$$

$p_1^t, \ldots, p_K^t$ sont les probabilités de chacune des K classes dans le nœud t. φ est maximale quand toutes les classes son également mélangés ($\varphi(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ = maximum) et minimale quand le nœud ne contient qu'une classe ($\varphi(1, 0, \ldots, 0) = \varphi(0, 1, 0, \ldots, 0) = \ldots = \varphi(0, \ldots, 0, 1) = 0$). Le coefficient utilisé par CART consiste alors en choisir l'atribut binaire w qui maximize la diminuition d'impureté, donné par (voir chapitre antérieur):

$$G(t, w) = \varphi(t) - p_l^t \varphi(t_l) - p_r^t \varphi(t_r) \quad (G(t, w) \geq 0, \forall t, w).$$

Il y a de très bonnes raisons en faveur du coefficient de Gini et qui seront exposées, d'une façon géométrique, dans la section suivante. D'autres coefficients peuvent néanmoins s'avérer également et différemment intéressants. De toutes façons, l'utilisation d'autres critères dans notre méthode *ARCADE* nous a permis de contester l'affirmation selon laquelle le critère de choix d'un atribut binaire n'a pas d'importance et peut même se faire au hasard (Mingers (1989)).

À chaque nœud t d'un arbre T il est possible d'attribuer une des K classes ou groupes. Désignons par $\gamma(i/j)$ le coût de classement d'une observation du groupe C_j dans le groupe C_i. Alors, pour $i, j = 1, \ldots, K$, $\gamma(i/j) \geq 0$ et $\gamma(i/i) = 0$. La classe choisie, i, doit être celle qui minimise le coût de classement $c(t)$, qui est ainsi défini par:

$$c(t) = \min_i \sum_j \gamma(i/j) p_j^t$$

Soît $\tilde{T}$ l'ensemble des nœuds terminaux de T et $p(t)$ la probabilité d'appartenance au nœud t. Alors, le coût global de mauvais classement de l'arbre T est

$$C(T) = \sum_{t \in \tilde{T}} C(t) = \sum_{t \in \tilde{T}} c(t) p(t).$$

Le plus souvent, $\gamma(i/j)$ est constant pour $i \neq j$ et alors, la classe choisie est celle qui maximise la probabilité p_i^t. En utilisant la règle de Bayes, on a

$$p_i^t = p(i/t) = \frac{p(i,t)}{p(t)} = p(i) \frac{p(t/i)}{p(t)}$$

$p(i)$ désigne la probabilité a priori d'appartennace au groupe i. Dans notre cas, le schéma d'échantillonage permet d'estimer $p(i)$ par N_i/N où N_i désigne le nombre total d'observations de la classe i et N le nombre total d'observations. On peut estimer $p(t)$ par N_t/N (N_t désigne le nombre total d'observations de l'ensemble d'apprentissage dans le nœud t) et $p(t/i)$ par N_i^t/N_i (N_i^t est le nombre total d'observations de la classe i dans le nœud t). Alors, $p_i^t = p(i).p(t/i)/p(t)$ est estimé par

$$\frac{N_i}{N} \frac{N_i^t/N_i}{N_t/N} = \frac{N_i^t}{N_t}$$

et ainsi dans ce cas, la classe choisie est la classe majoritaire.

Un autre point important de la méthode CART, peut-être le plus important, concerne l'introduction d'une nouvelle stratégie pour arriver à l'arbre de décision final. En effet, avant CART, les méthodes de construction d'arbres de décision utilisaient une règle d'arrêt pour terminer la division d'un nœud t. Cette règle consistait à déclarer un nœud comme terminal si la diminuition d'impureté maximale était inférieur à un seuil pré-fixé. Il y a eu divers développements de cette règle, mais les résultats n'étaient jamais satisfaisants, parce que, dépendant du seuil pré-fixé , l'arbre final était soit très grand, soit petit et en plus, il était impossible de fixer un seuil efficace pour tous les nœuds. Breiman et al.(1984) ont beaucoup repensé ce problème et ont proposé de construire un arbre trop grand T_{max} sans utiliser aucun seuil. Ainsi, un nœud est declaré comme terminal s'il est pur ou s'il n'y a pas de divisions admissibles pour ce nœud; parfois si le nœud ne contient qu'un nombre très faible d'observations (généralement entre 1 et 5), on peut le déclarer comme terminal pour ne pas augmenter inutilement le temps de calcul. Comme l'arbre T_{max} a été construit de façon a réduire successivement le coût de mauvais classement de l'ensemble d'apprentissage, alors aucun sous-arbre de T_{max} a un coût de mauvais classement inférieur à celui de T_{max}. Toutefois, certaines divisions de cet arbre, en particulier les dernières, peuvent être peu pertinentes et ainsi il peut y avoir des sous-arbres plus fiables parce que leur coût n'est pas sous-estimé. Breiman et al. ont proposé d'élager successivement l'arbre T_{max} afin d'obtenir une séquence de sous-arbres emboîtés:

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}$$

($\{t_1\}$ désigne l'arbre minimal ne contenant qu'un nœud, la racine). Chacun des sous-arbres T_i est obtenu à partir de T_{i-1} en élaguant une de ses branches. Pour décider quelle branche élaguer, ont utilise une mesure de coût-complexité d'un arbre T, défini par

$$C_\alpha(T) = C(T) + \alpha|\tilde{T}| = \sum_{t\in\tilde{T}}[C(t)+\alpha] = \sum_{t\in\tilde{T}} C_\alpha(t)$$

où $|\tilde{T}|$ représente le nombre de nœuds terminaux de l'arbre T. $C_\alpha(T)$ représente le coût global de l'arbre T auquel est ajouté une pénalité égale à α pour chaque nœud terminal. $C_\alpha(t) = C(t) + \alpha$ est la mesure de coût-complexité du nœud t. On peut étendre cette mesure à une branche T^t dont le nœud racine est t:

$$C_\alpha(T^t) = C(T^t) + \alpha|\tilde{T^t}|$$

Pour chaque valeur de α, on note $T(\alpha)$ le plus petit sous-arbre de T_{max} qui minimise la mesure de coût-complexité. Plus α est petit, plus la pénalité due à un nombre élevé de nœuds terminaux est faible, et ainsi, plus l'arbre $T(\alpha)$ a un nombre élevé de nœuds terminaux. Au fur et à mesure que α augmente, l'arbre $T(\alpha)$ devient de plus en plus petit et pour α suffisament grand, $T(\alpha)$ se réduit à la racine de l'arbre T_{max}. [Breiman et al. 1984] ont démontré que l'ensemble des sous arbres de T_{max} qui minimisent la mesure de coût-complexité pour les divers valeurs de α, consiste dans une séquence $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$ de sous-arbres emboîtés, et qui peut être obtenue d'une façon récursive:

- Commençons par poser $\alpha_1 = 0$. Comme il n'y a pas de pénalité, alors parmi tous les sous-arbre T de T_{max}, celui qui minimise $C_0(T)$ est l'arbre T_{max}. Cependant, comme l'indice de Gini vérifie $G(t,w) \geq 0$, alors il peut y avoir des divisions qui ne font pas décroître le coût. Soient alors t_l et t_r deux nœuds terminaux avec le même nœud parent t. Si $C(t) = C(t_l) + C(t_r)$ on coupe la branche issue du nœud t. On poursuit jusqu'à ce que $C(t) > C(t_l) + C(t_r)$ pour tous les nœuds non terminaux t, obtenant ainsi le premier arbre de la séquence, T_1.

- Maintenant, il y a lieu d' obtenir l'arbre T_2 à partir de l'arbre T_1. Soît t un nœud intermédiaire de T_1 et T_1^t le sous arbre issu de t. Alors, $C_\alpha(t) = C(t) + \alpha$ et $C_\alpha(T_1^t) = C(T_1^t) + \alpha|\tilde{T_1^t}|$. Si $\alpha = 0$ alors $C_\alpha(t) > C_\alpha(T_1^t)$. L'égalité est ontenue pour $\alpha = (\mathrm{C(t) - C(T_1^t)})/(|\tilde{\mathrm{T}}_1^{\mathrm{t}}| - 1)$. On définit alors la fonction
$$g_1(t) = \frac{\mathrm{C(t) - C(T_1^t)}}{|\tilde{\mathrm{T}}_1^{\mathrm{t}}| - 1}, \qquad \forall\ t\ \in\ T_1 - \tilde{T}_1$$
et on détermine le nœud (ou les nœuds) qui minimisent cette fonction. L'arbre T_2 est obtenu en élaguant tous les sous arbres issus des nœuds t de $T_1 - \tilde{T}_1$ minimisant cette fonction $g_1(t)$.

- Depuis, la procédure se poursuit de façon similaire. À chaque pas k, on commence par définir la fonction
$$g_k(t) = \frac{\mathrm{C(t) - C(T_k^t)}}{|\tilde{\mathrm{T}}_{\mathbf{k}}^{\mathbf{t}}| - 1} \qquad \forall\ t\ \in\ T_k - \tilde{T}_k$$
et on élague les branches issues des nœuds qui minimisent g_k.

Après avoir obtenu la séquence de sous-arbres emboîtés, il faut choisir le sous-arbre optimal. Ceci est fait soit à travers un échantillon test soit à travers l'utilisation d'une procédure de validation croisée.

3.3 La binarisation des attributs qualitatifs dans CART; justification.

Les variables de description prédictives sont rarement, initialement binaires. Si $\mathcal{S}(v)$ désigne l'échelle des valeurs d'une telle variable v, un aspect fondamental de la recherche, consiste en la manière de construire des bipartitions (partitions à 2 classes) de $\mathcal{S}(v)$, afin d'obtenir des attributs dichotomiques

associés à v. Une telle construction doit dépendre de la structure (on dit encore sémantique) de $\mathcal{S}(v)$. Précisèment, ces dernières années, on s'est trouvé concerné par la manière de discrétiser le numérique [Krzanowsky, 1975], [Van de Merckt, 1993], [Heath et al., 1993], [Müller et Wysotzki, 1994].

Nous considérons dans cette section l'étude de la binarisation des variables qualitatives nominales. Chaque variable qualitative v avec $L(v)$ modalités engendre $(2^{L-1}-1)$ attributs binaires, correspondants à toutes les partitions en deux classes non vides de l'ensemble des L modalités. Si $L = L(v)$ était petit, rien n'aurait empêché le remplacement de la variable initiale v par $(2^{L-1}-1)$ attributs binaires; où une même modalité de l'une de ces dernières variables binaires, correspond à un sous ensemble propre des modalités de la variable initiale. Le cas spécifique où se situe notre apport méthodologique est celui où chacune des variables prédictives v est qualitative nominale; mais à très grand nombre de modalités (catégories). En d'autres termes $L(v) = \text{card}[\mathcal{E}(v)]$ est "grand" (et peut dans notre application atteindre 20^4!), pour chacune des variables v de l'ensemble $\mathcal{V}$ des variables prédictives. Dans la section suivante nous décrivons comment notre méthode $\mathcal{ARCADE}$ traite le problème de la binarisation des attributs qualitatifs nominaux à très grand nombre de modalités. Mais ici, nous allons considérer le traitement de la question par CART.

3.3.1 Le cas de deux classes à prédire dans CART.

Dans le cas de la discrimination entre deux classes, la méthode CART utilise un résultat de 1958 de Fisher pour réduire la complexité de recherche de l'attribut binaire le plus prédictif, d'entre tous les $2^{L-1}-1$ attributs binaires, à $O(L)$. Nous démontrons ici cette procédure, d'une façon géométrique, pour les critères de Gini et du χ^2.

Reprenons ici l'indice de Gini [cf. (5) du chapitre 2], pour le cas qui nous intéresse ici où on a 2 classes à prèdire ($K = 2$); mais, une variable prèdictive à L modalités. Nous recommencerons par considérer le cas $L = 2$ pour en donner une représentation géométrique immédiate que nous généraliserons pour L quelconque. Ici donc $L = \{l, r\}$ et $J = \{1, 2\}$ et l'indice s'écrit:

$$p_l d^2(p_J^l, p_J) + p_r d^2(p_J^r, p_J) =$$

$$p_l \sum_{j=1}^m (p_j^l - pj)^2 + p_r \sum_{j=1}^m (p_j^r - pj)^2 = p_l \parallel p_J^l - p_J \parallel^2 + p_r \parallel p_J^r - p_J \parallel^2$$

et comme on a $p_J = p_l p_J^l + p_r p_J^r$, l'expression de l'indice devient

$$p_l \parallel p_J^l \parallel^2 + p_l \parallel p_J \parallel^2 - 2p_l < p_J^l, p_J > + p_r \parallel p_J^r \parallel^2 + p_r \parallel p_J \parallel^2 - 2p_r < p_J^r, p_J >=$$

$$p_l \parallel p_J^l \parallel^2 + p_r \parallel p_J^r \parallel^2 + \parallel p_J \parallel^2 - 2 < p_l p_J^l + p_r p_J^r, p_J >$$

Le dernier terme valant $-2 \parallel p_J \parallel^2$, on obtient pour l'indice de Gini

$$p_l \parallel p_J^l \parallel^2 + p_r \parallel p_J^r \parallel^2 - \parallel p_J \parallel^2$$

En ce qui concerne le critère du χ^2, il s'agit de la même expression; mais par rapport à la métrique du χ^2 [cf. (14), chap. 2].

On est dans le contexte d'un nuage à 2 sommets dans un espace à 2 dimensions a priori. Mais en fait, à une seule dimension, exactement sur le segment de droite (simplexe) défini par: $\{(p, q)/p > 0, q > 0, p + q = 1\}$.

Considérons le tableau:

p_{l1}	p_{l2}	$p_{l.}$
p_{r1}	p_{r2}	$p_{r.}$
$p_{.1}$	$p_{.2}$	1

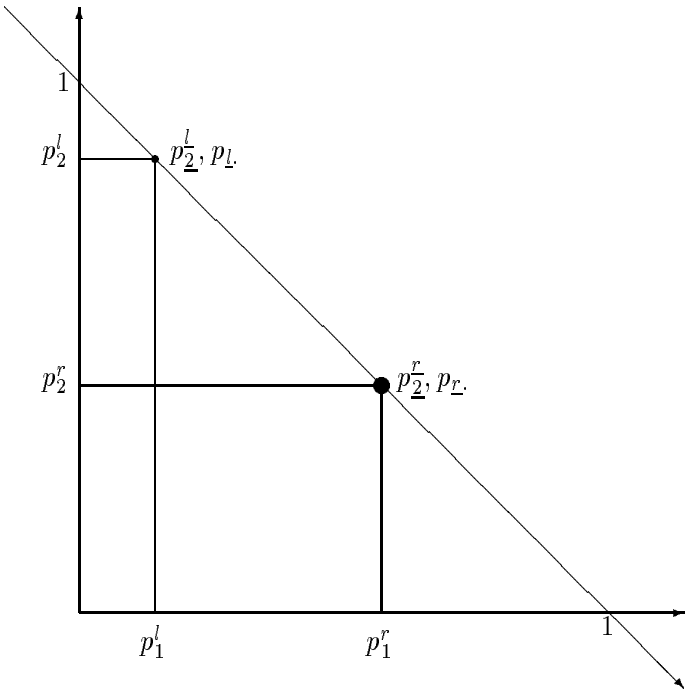


FIG. 2 - Segment de droite contenant les deux sommets pesants $p_{\underline{2}}^l$ et $p_{\underline{2}}^r$.

On a alors les deux sommets pesants:

$$\begin{array}{rcll} p_{\underline{2}}^l & = & (p_1^l, p_2^l); & p_{\underline{l}.} \quad \text{(poids affecté au nœud gauche)} \\ p_{\underline{2}}^r & = & (p_1^r, p_2^r); & p_{\underline{r}.} \quad \text{(poids affecté au nœud droit)} \end{array}$$

C'est l'inertie de ce nuage de deux sommets pesants, selon la métrique euclidienne ordinaire s'il s'agit de l'indice de Gini ou selon la métrique du χ^2, s'il s'agit du critère du χ^2. On peut déterminer, sur l'axe transversal dessiné, parallèle à la deuxième bissectrice, l'abscisse de $p_{\underline{2}}^l$ (resp. $p_{\underline{2}}^r$) en fonction croissante de p_1^l (resp. p_1^r).

Considérons maintenant pour le cas général, le tableau de contingence qui croise la variable qualitative v à L modalités avec la variable à prédire, qui dans ce cas a 2 modalités:

j / l	1	2
1	n_{11}	n_{12}
$\vdots$	$\vdots$	$\vdots$
L	n_{L1}	n_{L2}

$$p_j^l = p(j/x = l) = \frac{p(j,l)}{p(l)} = \frac{\pi(j)p(l/j)}{\sum_j \pi(j)p(l/j)}$$

En ordonnant ces probabilités, on obtient:

$$p(1/l_1) \leq p(1/l_2) \leq \cdots \leq p(1/l_L),$$

et, avec nos notations:

$$p_1^{l_1} \leq p_1^{l_2} \leq \cdots \leq p_1^{l_L}.$$

L'indice de Gini devient ici:

$$\sum_{i=1}^{L} p_{l_i} \sum_{j=1}^{2} (p_j^{l_i} - p_j)^2;$$

il représente l'inertie d'un nuage de points situé sur le segment de droite [(0,1),(1,0)] [cf. Fig. 2]. Il s'agit, par regroupement en 2 classes, de maximiser l'inertie des deux sommets pesants, centres de gravité respectifs des deux classes. Très précisément, il s'agit du moment d'inertie d'un nuage de la forme

$$\left\{ (p_{\underline{2}}^{\underline{l}}, p_{\underline{l}.}), (p_{\underline{2}}^{\underline{r}}, p_{\underline{r}.}) \right\},$$

où $\underline{l}$ représente un sous ensemble propre de modalités et où $\underline{r}$, représente le sous ensemble complémentaire.

Or, a priori, il existe $2^{L-1} - 1$ bipartitions; cependant, le théorème de Fisher [Fisher 1958] nous indique que, compte tenu de la nature du critère, la partition optimale est en intervalles connexes. Il s'agit donc dans notre cas de deux intervalles; l'un commençant et l'autre, finissant. On se trouve ainsi réduit à ne considérer que $(L-1)$ fusions.

De plus, dans notre cas $(J = 2)$, l'indice de Gini s'écrit:

$$\sum_{l=1}^{L} p_l[(p_1^l)^2 + (p_2^l)^2] \quad - \quad [p_1^2 + p_2^2] = \sum_{l=1}^{L} p_l(p_1^l - p_1)^2 + \sum_{l=1}^{L} p_l(p_2^l - p_2)^2$$

Mais, $p_2^l = 1 - p_1^l$ et $p_2 = 1 - p_1$, de sorte que $(p_2^l - p_2)^2 = (p_1^l - p_1)^2$ et alors la deuxième somme de l'equation dernière est identique à la première somme; c'est à dire, l'inertie du nuage projeté sur l'axe horizontal est identique à l'inertie du nuage projeté sur l'axe vertical. Il suffit, dans ces conditions, de maximiser l'inertie du nuage projeté sur l'axe horizontal.

Il suffit donc de considérer les $L-1$ fusions possibles, respectant l'ordre des probabilités ci-dessus; c'est à dire de la forme

$$l_1 \vee l_2 \vee \cdots \vee l_k = \underline{l} \qquad \text{et} \qquad l_{k+1} \vee l_{k+2} \vee \cdots \vee l_L = \underline{r}.$$

où alors,

$$p_1^{\underline{l}} = \frac{p(l_1)p_1^{l_1} + \cdots p(l_k)p_1^{l_k}}{p(l_1) + \cdots + p(l_k)} \qquad \text{et} \qquad p_1^{\underline{r}} = \frac{p(l_{k+1})p_1^{l_{k+1}} + \cdots p(l_L)p_1^{l_L}}{p(l_{k+1}) + \cdots + p(l_L)}$$

On cherchera alors simplement à maximiser l'inertie du nuage suivant des deux points, portés par l'axe horizontal:

$$\left\{ (p_1^{\underline{l}}, p(\underline{l})), (p_1^{\underline{r}}, p(\underline{r})) \right\},$$

La complexité est bien en $O(L)$ et il en est de même si le critère utilisé est celui du χ^2, car seule la métrique change. On a en effet pour la valeur du critère:

$$\sum_{l=1}^{L} p_l[\tfrac{1}{p_1}(p_1^l - p_1)^2] + \sum_{l=1}^{L} p_l[\tfrac{1}{p_2}(p_2^l - p_2)^2]$$

Or,

$$\tfrac{1}{p_2}(p_2^l - p_2)^2 = \tfrac{1}{1-p_1}(p_1^l - p_1)^2,$$

où p_1 et p_2 correspondent aux deux classes à prédire et sont donc fixés. Ainsi, le critère du χ^2 se met sous la forme

$$(\tfrac{1}{p_1} + \tfrac{1}{1-p_1}) \sum_{l=1}^{L} p_l(p_1^l - p_1)^2 = \tfrac{1}{p_1(1-p_1)} \sum_{l=1}^{L} p_l(p_1^l - p_1)^2$$

et se trouve donc ici, parfaitement èquivalent au critère de Gini, qui se met sous la forme:

$$2\sum_{l=1}^{L} p_l(p_1^l - p_1)^2.$$

Donc, l'optimisation du critère du χ^2 s'effectue aussi en $O(L)$. D'autre part, optimiser l'un des critères (Gini ou χ^2) revient à optimiser l'autre. Les deux critères, dans le cas de 2 classes à prédire, sont ainsi équivalents pour detecter la meilleure variable binaire dans l'ensemble des variables binaires associées à une même variable à L modalités.

Cette démarche, que nous démontrons ici d'une façon géométrique, fait partie de la méthode CART de Breiman & al. ([Breiman & al, 1984]).

3.3.2 Le cas de plusieurs classes à prédire dans CART.

Breiman & al. n'ont pas trouvé une généralisation du résultat de Fisher (1958) pour le cas de plusieurs classes. La méthode CART utilise deux approches différentes dans cette situation; que nous avons déjà mentionnées au paragraphe 2.2.3.

Il faut savoir qu'à chaque fois et dans l'étape ultime on a à comparer, sur la base de l'indice de Gini deux bipartitions; l'une résultant d'un attribut prédictif et l'autre, de la variable à prédire. Le critère optimisé peut prendre, pour l'une des approches, une forme particulière dite de "Twoing".

(**i**) Dans la première approche, on associe à la variable à prédire possédant K valeurs, les $2^{K-1}-1$ bipartitions du sous ensemble concerné de l'ensemble d'apprentissage. Par rapport à l'une des bipartitions, la binarisation d'un attribut prédictif à L valeurs, est, on l'a vu, en $O(L)$. On retient l'attribut binaire (ou binarisé) le plus prédictif par rapport à l'ensemble de toutes les binarisations de la variable à prédire; et ce, au sens de l'indice de Gini. Si L_m est le nombre de valeurs du m-ème attribut prédictif, $1 \leq m \leq p$, la complexité prédictif est en

$$\sum_{1\leq \mathrm{m}\leq \mathrm{p}} (2^{\mathrm{K}-1} - 1)\mathrm{O}(\mathrm{L_m});$$

où K est le nombre de classes à prédire.

(**ii**) La deuxième approche (dont nous reprenons l'expression à partir du paragraphe 2.2.1.) consiste à considérer toutes les variables binaires issues d'une même variable prédictive à L modalités (il y en a $2^{L-1}-1$) et, respectivement, à évaluer chacune d'entre elles, par rapport à la binarisation de la variable à prédire w, qui lui est la mieux associée au sens du critère de Gini. Les auteurs de la méthode démontrent que cette binarisation (ou bipartition) doit être choisie telle que:

$$C_1(w) = \{j : p_j^{t_l} \geq p_j^{t_r}\} \quad \text{et} \quad C_2(w) = \{j : p_j^{t_l} < p_j^{t_r}\};$$

le critère optimisé prenant alors la forme dite de "Twoing" [cf. (12) §2.2.1.].

Ici, la complexité de la recherche devient:

$$\sum_{1\leq \mathrm{m}\leq \mathrm{p}} (2^{\mathrm{L_m}-1} - 1)\mathrm{O}(\mathrm{K}).$$

La stratégie qu'on adopte dans ces conditions va dépendre de la comparaison entre les deux types de complexité.

4 La méthode ARCADE

4.1 Introduction

La méthode ARCADE (**AR**bre de **C**l**A**ssification et de **DE**cision) comprend en son sein la méthode CART et notamment, son aspect le plus original qui concerne l'élagage de l'arbre. Elle s'en distingue par un complément. Nous y introduisons en effet, pour la construction de l'arbre, une nouvelle famille de coefficients d'association entre variables qualitatives [Lerman 1970,1981,1992a,1992b], [Bacelar-Nicolau 1980, 1988] issus de la méthode de classification AVL (Analyse de la Vraisemblance des Liens) [Lerman 1970, 1981, 1993], [Bacelar-Nicolau 1980, 1985], [Costa Nicolau 1980]. Cette nouvelle famille permettra de bâtir de façon plus riche, des arbres de décision non binaires; et ce, en tenant compte de la sémantique qui est sous jacente à l'ensemble des valeurs des variables prédictives et de celle, à prédire.

La distinction la plus importante entre la méthode ARCADE et celle CART concerne la manière de binariser les attributs qualitatifs prédictifs. L'efficacité de notre méthode est décisive s'il s'agit d'attributs à ≪grand≫, voire même ≪très grand≫ nombre de valeurs (on dit encore "modalités" ou "catégories"). Soulignons - comme nous l'avons déjà mentionné dans l'introduction - que le nombre de modalités par attribut, pouvait atteindre 20^4; et nous avons abouti dans ce cas, hors formation de l'arbre, à ne considérer qu'une vingtaine d'attributs binaires. Ainsi, même une complexité linéaire par arpport au nombre L de valeurs de l'attribut, devient par trop importante. Et, cette complexité linéaire ne peut être rigoureusement atteinte dans la méthode CART que si la variable à prédire est binaire ($K = 2$) et si le critère de construction de l'arbre est d'un certain type qui comprend les critères inertiels (cf. §3).

Nous avons déjà présenté en introduction (cf. §1) l'idée générale de notre méthode de réduction, qui exploite une synthèse automatique par la classification, de l'ensemble des valeurs de l'attribut prédictif. La base est le tableau de contingence LxK croisant l'attribut prédictif et la variable à prédire. Si le nombre L est ≪trop grand≫, l'aspect classification et un aspect factorisation préalable de l'ensemble des valeurs doivent être combinés selon un schéma que nous préciserons ci-dessous.

La méthode de classification utilisée est ascendante hiérarchique. Il s'agit de la Vraisemblance des Liens ($\mathcal{AVL}$) [Lerman 1970,1981,1993] implantée dans le programme CHAVL [Lerman, Peter et Leredde, 1993-1994]. Cette méthode utilise une forme du critère d'agrégation appelé $\mathcal{AVL}_{0.5}$. Il s'agit de la méthode A.V.B. ($\mathcal{AVL}_{0.5}$) étudié par [Bacelar-Nicolau,1985] et qui fait partie de la famille paramétrique de méthodes proposée par [Costa Nicolau,1980]. Ainsi, nous noterons $\text{CHAVL}_{0.5}$ la méthode programmée mise en œuvre. Cette approche permet le traitement de n'importe quelle structure mathématique du tableau des données et notamment, la classification de l'ensemble des lignes d'un tableau de contingence. D'autre part, dans cette méthode, il y a un repérage automatique des niveaux et des nœuds "significatifs" de l'arbre des classifications. Et, la notion de niveau significatif nous sera très utile pour déterminer ce que nous appellerons les "macro-modalités". Enfin, il est tout à fait pertinent d'adopter une classification hiérarchique; car nous exploiterons la structure d'arbre ultramétrique pour faire sensiblement décroître la complexité.

4.2 La binarisation des attributs prédictifs dans ARCADE

Référons nous de nouveau au tableau de contingence à L lignes et K colonnes, établi sur la base de l'ensemble d'apprentissage, et croisant la variable prédictive v, à L catégories, avec la variable à prédire, dont les valeurs sont les concepts ou classes.

Nous allons supposer deux cas. Le premier est celui où L n'est pas ≪trop grand≫ de sorte que les contenus des cases du tableau de contingence ont la consistance nécessaire pour la significativité des calculs.

Notre méthode consiste dans sa première étape à réduire par regroupement, l'ensemble des modalités de chacune des variables prédictives - Une même variable v se trouve ainsi remplacée par une macro variable $w = w(v)$, dont chaque modalité est une classe de modalités de la variable d'origine v. Mais il s'agit de créer ces nouvelles modalités synthétiques en accord même avec le principe de

discrimination sous jacent à la formation d'un arbre de décision. En effet, nous le faisons à partir d'une classification automatique de l'ensemble des L modalités de la variable d'origine en se basant sur le tableau de contingence LxK, ci-dessus considéré - Le nombre de classes retenu, sera d'une part fonction de la "significativité" des classes et d'autre part, de l'ordre de grandeur de leur nombre; et ce, en relation avec la complexité qu'on peut accepter pour la formation de l'arbre de décision - Signalons pour fixer les idées que, dans notre application, pour $L = 10^3$, nous avons retenu 12 classes.

Et si J est le nombre de modalités de cette macro variable w; on peut envisager de considérer l'ensemble des $(2^{J-1} - 1)$ attributs binaires et dont chacun se trouve associé à une partition en deux classes de l'ensemble des macro-modalités. Cependant, l'arbre hiérarchique détermine une structure ultramétrique de proximité entre les classes qui définissent les macro modalités ; et il est du plus grand intérêt d'exploiter cette structure. Considérons, par exemple, l'arbre de classification suivant:

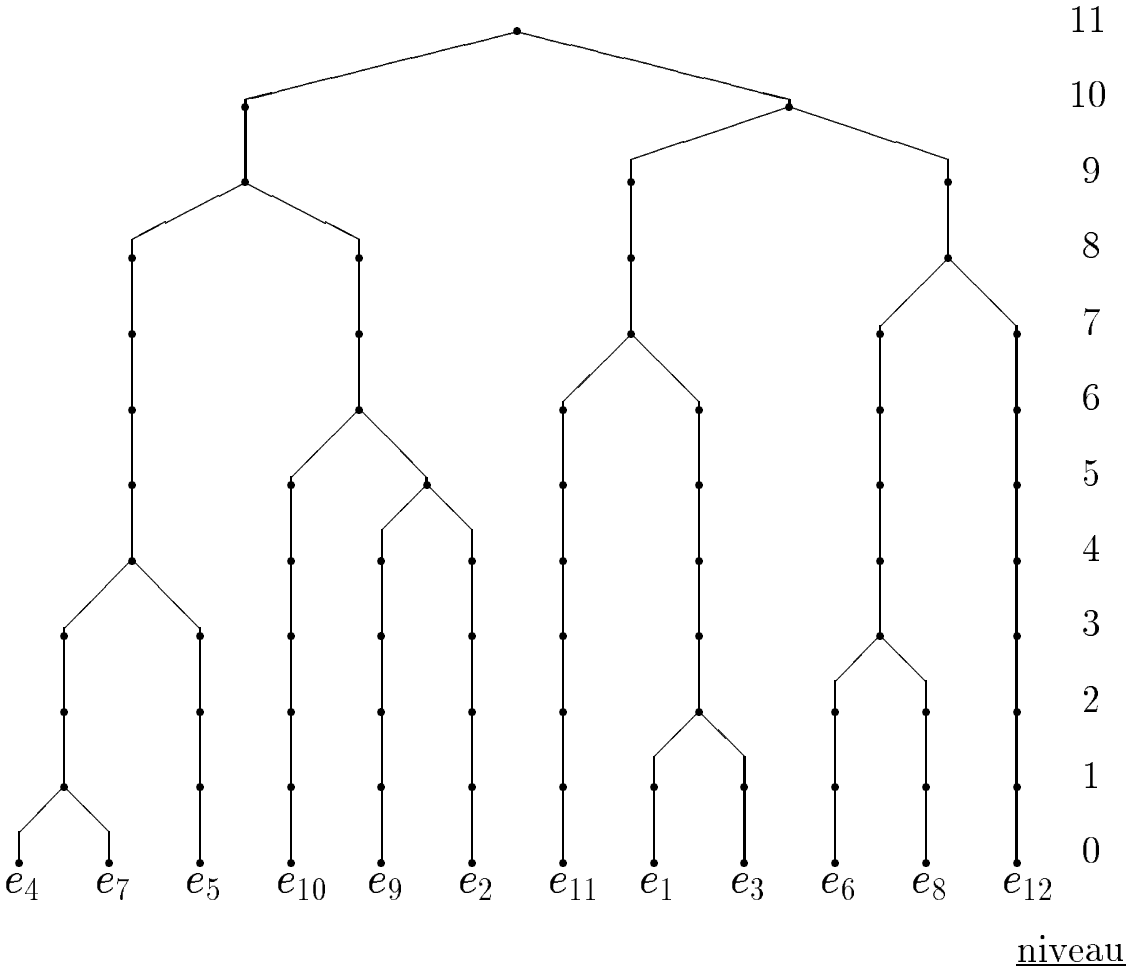


FIG. 3 - Exemple d'un arbre de classification hiérarchique.

L'ensemble des 12 macro-modalités engendre $2^{12-1} - 1 = 2047$ bipartitions (attributs binaires). Cependant, la plupart de ces attributs binaires n'ontpas un pouvoir prédictif, parce qu'ils correspondent à des bipartitions dont chaque classe est constitué par des macro-modalités qui sont éloignées (au sens de la distance ultramétrique) les unes des autres. c'est par exemple le cas de la bipartition

$$(\{e_9, e_3, e_{12}\}, \{e_1, e_2, e_4, e_5, e_6, e_7, e_8, e_{10}, e_{11}\}).$$

Pour éviter de considérer ces attributs non prédictifs, nous imposons à l'une des deux modalités de la variable binaire de fusionner des classes conformément à l'arbre hiérarchique. Ainsi, à chacune

des feuilles et à chacun des nœuds de l'arbre de classification sur l'ensemble des macro modalités, correspond une variable. Il en résulte une chute importante de la complexité. Ainsi, pour $L = 1000$, nous avons abouti à ne considérer qu'environ 20 variables binaires; ce qui est considerablement inférieur à L. Dans le cas de l'arbre hiérarchique précédent, les 21 atributs binaires considérés sont:

a_1 :	$(a_1 = 0)$	$\equiv$	$w \in \{e_1\}$	(niveau 0)
a_2 :	$(a_2 = 0)$	$\equiv$	$w \in \{e_2\}$	(niveau 0)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{12} :	$(a_{12} = 0)$	$\equiv$	$w \in \{e_{12}\}$	(niveau 0)
a_{13} :	$(a_{13} = 0)$	$\equiv$	$w \in \{e_4, e_7\}$	(niveau 1)
a_{14} :	$(a_{14} = 0)$	$\equiv$	$w \in \{e_1, e_3\}$	(niveau 2)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{20} :	$(a_{20} = 0)$	$\equiv$	$w \in \{e_6, e_8, e_{12}\}$	(niveau 8)
a_{21} :	$(a_{21} = 0)$	$\equiv$	$w \in \{e_4, e_7, e_5, e_{10}, e_9, e_2\}$	(niveau 9)

Nous allons maintenant considérer le cas où L est $\ll$trop grand$\gg$ pour pouvoir assurer la consistance statistique nécessaire à la significativité des calculs. En d'autres termes, le tableau de contingence de dimension LxK, devient trop creux. Dans notre cas (cf. §5), L peut atteindre 20^4 pour chacune des variables initiales v. Toutefois, la structure de l'ensemble des valeurs de v est particulière. Ce dernier se présente en effet, comme un produit cartésien d'ensembles; exactement sous la forme A^4, où A représente un alphabet de 20 lettres; la variable v étant définie au moyen d'un mot de 4 lettres.

Dans le cas où L est $\ll$trop grand$\gg$, nous allons considérer deux volets. Le premier est celui où l'ensemble (ou échelle) des valeurs $S(v)$ se factorise naturellement et le second, où ce n'est pas le cas. Dans la première situation $S(v)$ est un produit cartésien d'ensembles et v peut être réduite à un p-uplet $(u_1, u_2, \ldots, u_p)$ de variables élémentaires. Ainsi, dans notre application, où v représente un mot à quatre lettres, v se met sous la forme (u'_1, u'_2, u'_3, u'_4); où u'_j est la lettre occupant la j-ème position.

En désignant par A_j, $1 \leq j \leq p$, l'ensemble des valeurs de u_j, on a

$$\mathcal{S}(v) = A_1 \text{x} A_2 \text{x} \ldots \text{x} A_p$$

Une factorisation équilibrée d'ordre $r (r < p)$ consiste à regrouper le produit des p facteurs précédent en r sous produits connexes; tels que les tailles des différents sous produits, soient aussi égaux que possibles.

Nous nous contenterons de cette définition intuitive qui peut être formalisée et précisée et nous nous limiterons - bien que cela soit aisèment généralisable - à $r = 2$. Ainsi, nous définirons

$$B_1 = A_1 \text{x} A_2 \text{x} \ldots \text{x} A_m$$

et

$$B_2 = A_{m+1} \text{x} A_{m+2} \text{x} \ldots \text{x} A_p;$$

et m, choisi de façon à rendre minimal

$$|\text{card}(B_1)\text{-card}(B_2)|.$$

Ainsi, nous substituons au tableau de contingence global à L lignes et K colonnes:

$$(B_1 \text{x} B_2) \text{x} C,$$

où C est l'ensemble des classes à prédire, le couple de tableaux de contingence:

$$(B_1 \text{x} C, B_2 \text{x} C)$$

Il s'agit d'une approximation dans laquelle on remplace une distribution jointe par un produit de deux distributions marginales; la première étant sur le cube $(B_1 \mathrm{x} B_2)\mathrm{x}C$ et les deux autres, sur $B_1\mathrm{x}C$ et $B_2\mathrm{x}C$, respectivement. Tout se passe comme si on projetait le cube de contingence sur chacune des deux faces que nous venons de mentionner.

On considère dans ces conditions un couple de classifications hiérarchiques $\mathrm{CHAVL}_{0.5}$ sur, respectivement B_1 et B_2, à travers C, sur la base des tableaux de contingence ci-dessus.

Relativement à l'arbre de classification issu de $\mathrm{CHAVL}_{0.5}(B_i)$ $(i = 1, 2)$, on retient un niveau significatif, en un nombre de classes $l(i), i = 1, 2$; de telle sorte que:

(a) $|l(1) - l(2)|$ minimal;

(b) $l(1)\mathrm{x}l(2)$ le plus grand possible, en accord avec un contenu consistant des cases du tableau de contingence à $l(1)\mathrm{x}l(2)$ lignes et à K colonnes que nous préciserons ci-dessous.

Revenons à l'expression de la variable prédictive v sous la forme (u_1, u_2). w indique la variable à prédire qui présente K valeurs. Nous avons déjà indiqué qu'on remplaçait le tableau de contingence (inconsistant) croisant $v = (u_1, u_2)$ avec w, par les deux tableaux de contingence, croisant, respectivement, u_1 avec w et $u2$ avec w. $\mathrm{CHAVL}_{0.5}(B_i)$ conduit à remplacer la variable u_i par une variable synthétique $\overline{u}_i$ dont l'ensemble des valeurs est l'ensemble des $l(i)$ classes de $B_i, i = 1, 2$.

On considère alors le tableau de contingence à $l(1)\mathrm{x}l(2)$ lignes et à K colonnes, croisant l'ensemble des valeurs de $(\overline{u}_1, \overline{u}_2)$ avec l'ensemble des valeurs de la variable w. C'est précisèment la classification hiérarchique des lignes de ce dernier tableau, conformément à $\mathrm{CHAVL}_{0.5}$, qui nous produira les macro-modalités de v; elles mêmes organisées hiérarchiquement de façon ultramétrique (cf. Fig. 3).

Soyons plus précis dans notre formulation. Désignons par $L(i)$ l'ensemble des indices $\{1, 2, \ldots, l(i)\}, i = 1, 2$; et soit,

$$\{\mathrm{B_i(k_i)}/1 \leq \mathrm{k_i} \leq \mathrm{l(i)}\}$$

la partition retenue à un niveau significatif de $\mathrm{CHAVL}_{0.5}(B_i), i = 1, 2$.

Les éléments, organisés par la classification des lignes du tableau de contingence de croisement de $(\overline{u}_1, \overline{u}_2)$ avec w, sont ceux de l'ensemble

$$D = \{B_1(k_1)\mathrm{x}B_2(k_2)/(k_1, k_2) \in L(1)\mathrm{x}L(2)\}$$

Une classe $D(t)$ de la partition obtenue (au niveau significatif retenu de $\mathrm{CHAVL}_{0.5}(D)$) et qu'on écrit

$$\{\mathrm{D(t)}/1 \leq \mathrm{t} \leq \mathrm{s}\},$$

prend la forme:

$$D(t) = \cup\{B_1(k_1)\mathrm{x}B_2(k_2)/(k_1, k_2) \in M_t\}$$

où

$$\{M_t/1 \leq t \leq s\}$$

désigne une partition de $L(1)\mathrm{x}L(2)$.

Chaque $D(t)$ correspondra précisèment à une des macro-modalités retenues. La section finissante de l'arbre hiérarchique (cf. Fig. 3) les organise pour la définition des attributs binaires.

Signalons à titre d'illustration le cadre de notre application, déjà mentionnée et qui sera développée au paragraphe suivant. v représente un mot de quatre lettres prenant toutes valeur dans le même alphabet A de 20 lettres:

$$\mathcal{A} = \{\mathrm{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\};$$

où chaque lettre représente un acide aminé.

Il s'agit de prédire la structure secondaire (cf. §5) qui prend l'une des trois valeurs X, H ou E ("boucle", "hélice α" ou "brin β''). Nous avons donc $K = 3$ classes à prédire.

Conformément à ci-dessus, le tableau de contingence associé au croisement entre v et w, où w est la variable à prédire, prend ici la forme:

v \ w	X	H	E
1(AAAA)			
2(AAAC)			
20^4(YYYY)			

FIG. 4 - Tableau initial de contingence

Ce tableau de contingence concerne 30.000 éléments pour le corpus de données adopté; alors qu'il comporte 20^4x3=480.000 cases.

Par rapport à nos précédents notations, où le mot est noté (u'_1, u'_2, u'_3, u'_4), définissons $u_1 = (u'_1, u'_2)$ et $u_2 = (u'_3, u'_4)$. Les deux tableaux de contingence respectivement associés à (u_1, w) et à (u_2, w) prennent la forme suivante:

u_1 \ w	X	H	E
1(AA)			
2(AC)			
20^2(YY)			

u_2 \ w	X	H	E
1(AA)			
2(AC)			
20^2(YY)			

FIG. 5 - Tableaux dérivés de contingence.

Nous avons retenu une partition significative en environ 31 classes, pour chacun de ces deux derniers tableaux, après application de $\text{CHAVL}_{0.5}$. De sorte, que le tableau définitif soumis à la classification a un nombre de lignes d'environ 31x31=961; alors que le nombre de colonnes est égal à 3.

Lorsqu'on dispose d'une variable descriptive v dont l'ensemble des L modalités (valeurs) ne se met pas sous la forme d'un produit cartésien, la procédure de factorisation peut ne pas paraître sémantiquement naturelle. Néanmoins, elle est techniquement généralisable en cas de nécessité où L est ≪grand≫. Imaginons, ce qui est largement suffisant dans les applications, une factorisation d'ordre 2. Ainsi, l'ensemble M des modalités se trouve mis sous la forme M_1xM_2 ou à tout le moins d'un sous ensemble de M_1xM_2, le plus large qui soit; donc tel que $l(1)$=card(M_1) et $l(2)$=card(M_2) sont choisis de la façon suivante:

(a) $|l(1) - l(2)|$ le plus petit possible

(b) $l(1)$x$l(2) - L$, positif ou nul et le plus petit possible.

Si L est un carré parfait, il y a lieu de prendre $l(1) = l(2) = \sqrt{(}L)$.

Cette procédure est ainsi de pur codage. La variable v est alors considérée comme un couple (u_1, u_2) de variables où M_1 (resp. M_2) est l'ensemble des valeurs de u_1 (resp. u_2). Et, la méthode de réduction se poursuit comme ci-dessus.

Une dernière situation méthodologiquement intéressante est celle où les variables descriptives initiales: v^1,v^2,...,v^m, ...,v^p, n'ont pas chacune nécessairement un grand nombre de modalités; mais où - désirant préserver au mieux la distribution jointe - on introduit des variables multiples avant la binarisation. Une telle variable multiple aura la forme

$$(v^{j_1}, v^{j_2}, \ldots, v^{j_l})$$

où $\{j_1, j_2, \ldots, j_l\}$ est une partie à l éléments de l'ensemble $P = \{1, 2, \ldots, m, \ldots, p\}$ des indices. L'ensemble de ces parties forme une partition $\pi(P)$ de P, qu'on indiquera sous la forme

$$\pi(\mathrm{P}) = \{\mathrm{P_e}/1 \leq \mathrm{e} \leq \mathrm{k}\}.$$

Ainsi, une variable multiple qui correspond à une classe P_e de la précédente partition regroupe l'ensemble suivant de variables:

$$\{v^j / j \in P_e\}$$

Le plus intéressant, toujours pour récolter un maximum d'information de la distribution jointe, consiste à opérer une classification par proximité de l'ensemble $\mathcal{V}$ des variables descriptives, sur la base du tableau des données $\mathcal{O}$x$\mathcal{V}$. La méthode de classification hiérarchique AVL, est parfaitement adaptée à cette fin.

En effet, elle permet de déterminer des nœuds exclusifs et pertinents de l'arbre des classifications, chacun sous tendant une classe de variables [Lerman et Ghazzali 1991] de façon compatible avec le problème de la complexité (nombre de modalités de la variable multiple créée). Une telle opération que nous pourrons appeler de MULTIPLICATION; est en quelque sorte inverse de celle de FACTORISATION ci-dessus introduite.

Finalement, la méthode ARCADE est obtenue selon le schéma suivant:

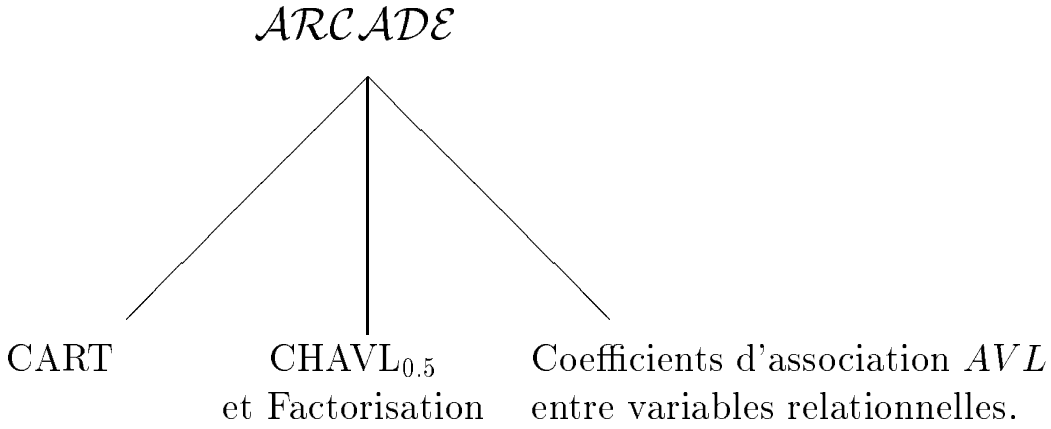


FIG. 3 - Schéma représentant les outils principaux qui font partie de la méthode ARCADE.

5 L'application; prédiction de la structure secondaire d'une protéine

Les développements que nous considérons dans ce travail résultent directement d'une application d'importance en Biologie Moléculaire dans l'analyse des séquences protéiques. La structure primaire

d'une telle séquence est formellement un mot pris dans un alphabet de 20 lettres (représentant les 20 acides aminés) et dont la longueur peut atteindre plusieurs centaines. La structure secondaire peut également être formalisée au moyen d'un mot de même longueur, calé sur le premier; mais dont l'alphabet comprend trois lettres E, H et X, qui seront les noms des trois concepts à discriminer (H fait partie d'une hélice α; E fait partie d'un brin β et X fait partie d'une boucle).

...	T	T	C	C	P	S	I	V	A	R	S	...
...	E	E	E	E	X	X	H	H	H	H	H	...

Le nombre de protéines pour lesquelles on connaît la séquence primaire augmente très rapidement, devant atteindre environ 100.000 d'ici jusqu'à la fin du siècle. Toutefois il n'en est pas de même pour les structures secondaires, qui sont beaucoup plus difficiles à déterminer experimentalement, soit par rayons X, soit par résonnance magnétique. Ainsi, il existe un nombre croissant de séquences d'acides aminés pour lesquelles on ne connaît pas la structure secondaire correspondante et on cherche donc à pouvoir prédire cette structure. Le problème de la prédiction de la structure secondaire n'est pas pénible quand on a des protéines connues et qui sont homologues à celle à prédire. Toutefois, pour environ 85% des nouvelles protéines, il n'y a pas d' homologues et le problème devient très difficile. Il s'agit en effet d'un vieux problème de reconnaissance où le progrès a été lent. La plupart des méthodes de prédiction qui ont été développées pour ce problème cherchent, à partir d'une *description* de la position donnée d'une lettre du premier mot, à prédire la lettre correspondante du second mot. Récemment, Solovyev & Salamov ([Solovyev & Salamov, 1994]) ont développé une méthode qui a un but différent; au lieu de prédire la structure secondaire, résidu par résidu, leur méthode tâche de prédire des hélices et brins entiers.

Parmi la grande variété de méthodes existantes, les plus performantes sont soit celles qui cherchent, dans l'ensemble de protéines de structure secondaire connue, des régions homologues à celle à prédire ("nearest-neighbor methods"), soit les méthodes qui utilisent l'information existante dans l'alignement de protéines homologues, comme par exemple la célèbre méthode PHD de Rost et Sander qui a brisé la barrière des 70%.

5.1 Les variables prédictives

Notre corpus des données est formé de 151 séquences d'une protéine globulaire qui sont mutuellement non-homologues (consulter [Colloc'h et al. 1993] pour une description de ces données). L'ensemble totalise environ 30.000 résidus qui se répartissent, relativement à la structure à prédire, comme suit: 46,6% pour la classe X, 29% pour celle H et 24,4% pour celle E.

Maintenant, pour décrire une position dans une fenêtre à des fins prédictives; la prédiction est loin de seulement dépendre de la seule lettre du premier mot dont il s'agit de déterminer la lettre associée (X, H ou E) [i.e. ayant la même position (cf. §5.1 ci-dessus)], dans le second mot. Une "bonne" description doit avoir comme support tout un environnement autour de la position à prédire. L'ignorance de cet environnement nous conduit à considérer une fenêtre de longueur $f = 2e + 1$ et centrée sur la position à prédire. C'est dans cette fenêtre que nous introduisons des vecteurs de positions qui, précisément, définiront des variables de description. Chaque vecteur de positions correspondra en fait à un mot de cette fenêtre formé de l lettres adjacentes que nous ferons glisser le long de cette fenêtre. Diverses valeurs de e ($e = 1, 2, 3, 4, 5, 6, 7$ et 8) et diverses valeurs de l (l = 1,2,3 et 4) ont été considérés. On a pu constater que les meilleurs résultats sont obtenus avec une fenêtre de longueur 11; pour les fenêtres plus longues, on constate une insensibilité de la qualité des résultats relativement à la longueur de la fenêtre. D'autre part, qu'il s'agisse de prendre tous les mots à une, deux, trois ou quatre lettres ; ou bien, seulement à quatre lettres, les performances sont tout à fait comparables. De sorte que nous avons terminé pour utiliser seulement les mots de 4 lettres ($l = 4$)et une fenêtre de longueur 11 ($e = 5$). On a ainsi, pour décrire la position centrale de la fenêtre 8 variables-mots,

dont la j-ème correspond au vecteur de positions $(j, j+1, j+2, j+3), 1 \leq j \leq 8$. Si $long(s)$ est la longueur d'une séquence s, les sites dont la position varie entre 6 et $(long(s) - 5)$ ont une description complète par les 8 variables et c'est eux qui interviendront dans la construction de l'arbre de décision. Alors que les 5 premiers (resp. derniers) ont une description incomplète qu'il est aisé de spécifier. Ils n'interviendront pas dans la construction de l'arbre, mais seront prédits au niveau de l'ensemble test.

Comme nous l'avons déjà précisé à la fin du paragraphe 4.2 précédent, chacune des 8 variables prend 20^4 valeurs possibles; alors que - si on imagine 150 séquences de longueur 200 - le nombre d'éléments (de sites) décrits est égal à (200-11+1)x150= 28.500. On se reportera ici au paragraphe 4.2 ci-dessus pour la création d'attributs binaires à partir de ces variables.

5.2 Méthode de construction de l'arbre; résultats et commentaires

5.2.1 Méthode de construction de l'arbre et principe de l'évaluation

Pour un ensemble d'apprentissage donné E, les différents coefficients d'association entre deux variables qualitatives, développés au paragraphe 2, ont été mis en œuvre pour la construction de l'arbre de décision total T_{max} (cf. §3.2). Nous avons ajouté deux procédures. Pour la prmière, on utilise un indice brut dit s de comparaison entre partitions et pour l'autre, le choix de l'attribut binaire pour couper le sous ensemble E sous tendu par un nœud, est choisi de façon aléatoire (procédure RANDOM). Cela, pour montrer la pertinence du choix d'un "bon" coefficient, correctement conçu sur le plan formel et statistiquement normalisé.

La méthode d'élagage de l'arbre et d'évaluation de la qualité de la prédiction (en termes de bonne classification) est celle de CART, utilisant un ensemble test (cf. §3.2 de [Breiman & al. 1984]).

D'autre part, dans notre cas, l'évaluation de la qualité globale de la prédiction est effectuée selon une procédure de type "jacknife".

Plus précisément, désignons par $\mathcal{S} = \{s_i / 1 \leq i \leq 151\}$ l'ensemble de nos séquences protéiques et associons à chaque s_i, l'ensemble complémentaire $E_i = \mathcal{S} - \{s_i\}$, de taille 150. Dans ces conditions et pour i donné, la construction de l'arbre s'effectue sur l'ensemble E_i, ou plus exactement, sur l'ensemble des résidus des séquences de E_i concernés par la description au moyen des 8 variables considérés au paragraphe 5.2. L'arbre i de décision est élagué à partir de l'ensemble test défini par la séquence s_i (c'est à dire, l'ensemble de ses résidus). La qualité de cet arbre est estimée à partir du pourcentage de bonne classification, sur l'ensemble des sites de s_i. Ce pourcentage peut être global ou ne concerner que l'une des classes à prédire (X, H ou E). La qualité globale de notre méthode "Arbre de Décision" pour le problème posé, est estimée par la moyenne de ces pourcentages. Cette moyenne peut être équipondérée ou bien, pondérée par les longueurs des séquences. Dans le premier cas, il s'agira des expressions Q_{chain} et dans le second cas Q_{total}, du tableau de la figure qui suivra.

Bien que l'ensemble test (ici, la séquence protéique laissée de côté) sert à la fois à déterminer - à partir de la meilleure valeur du paramètre d'élagage - l'arbre optimal, ainsi que le taux de mauvaise classification; ce dernier ne se trouve pas sous estimé, conformément aux expériences qui ont été réalisées [Breiman & al. 1984, p.81].

5.2.2 Les différentes mesures de la qualité de la prédiction

Il existe plusieurs façons d'évaluer la qualité d'une prédiction. La plupart des méthodes utilisent des mesures qui comparent la structure observée avec la structure prédite, résidu par résidu. Nous donnons ci-dessous l'expression des mesures les plus utilisées, en commençant par les plus simples, qui estiment la proportion de résidus qui ont été correctement prédits.

Supposons qu'on a une séquence d'acides aminés de longueur N. On peut considerer cette séquence comme un ensemble de N objets, représenté par $S = \{1, 2, \ldots, N\}$, et où l'objet i désigne l'acide

aminé qui est dans la position i de la séquence, pour $i = 1, 2, \ldots, N$. La structure secondaire observée correspondante, qui consiste dans un "mot" de longueur N, calé sur la séquence d'acides aminés, peut être représentée par des attributs booléens. Par exemple, pour la protéine,

T	T	C	C	P	S	I	V	A	R	S
E	*E*	*E*	*E*	*X*	*X*	*H*	*H*	*H*	*H*	*H*

on a $N = 11$ et $S = \{T, T, C, C, P, S, I, V, A, R, S\}$. La structure secondaire observée peut être représentée par les trois attributs booléens (on dit encore "présence-absence") suivants:

$v_E : S \longrightarrow \{0, 1\}$ représente la présence observée pour la classe E
$v_H : S \longrightarrow \{0, 1\}$ représente la présence observée pour la classe H
$v_X : S \longrightarrow \{0, 1\}$ représente la présence observée pour la classe X

De façon analogue on peut définir trois attributs booléens pour représenter la structure prédite par une certaine méthode:

$w_E : S \longrightarrow \{0, 1\}$ représente la présence prédite pour la classe E
$w_H : S \longrightarrow \{0, 1\}$ représente la présence prédite pour la classe H
$w_X : S \longrightarrow \{0, 1\}$ représente la présence prédite pour la classe X

L'attribut v_i (resp. w_j) peut être représenté par le sous-ensemble $S(v_i)$ [resp. $S(w_j)$] des objets de S où v_i (resp. w_j) est présent. Ainsi, pour comparer les deux structures, celle observée avec celle qui a été prédite, on peut se restreindre à comparer ces attributs. D'aprés Lerman (1992a), l'indice "brut" d'association entre les attributs v_i et w_j est

$$s = s(v_i, w_j) = card[S(v_i) \bigcap S(w_j)]$$

On peut [Lerman (1992a)] associer à cet indice, un indice brut aléatoire $s(v_i^*, w_j^*)$, dans le cadre d'une hypothèse d'absence de liaison. La moyenne et la variance de $s(v_i^*, w_j^*)$ permettent de fournir un indice normalisé par centrage et réduction. Précisèment, l'une des formes de l'hypothèse d'absence de liaison conduit au coefficient C_i ci-dessous. Deux autres formes pourraient être considérées.

Pour simplifier les expressions, considérons maintenant le tableau de contingence

$$\overrightarrow{\mathcal{A}} = \{A_{ij} / 1 \leq i, j \leq 3\},$$

Les codes 1,2 et 3 correspondront respectivement à hélice α (notée H ci-dessus), brin β (noté E ci-dessus) et boucle L (noté X ci-dessus). A_{ij} (qui occupe la ligne i et la colonne j) désigne le nombre de résidus qui ont été observés en i et qui ont été prédits en j, $1 \leq i, j \leq 3$. Cette quantité A_{ij} peut s'obtenir en comparant les attributs v_i et w_j: $A_{ij} = card[S(v_i) \bigcap S(w_j)]$.

$$\overrightarrow{a} = (a_1, a_2, a_3)$$

correspond à la marge ligne; ainsi, a_j est le nombre de résidus pour lesquels, la structure secondaire produite est j, $1 \leq j \leq 3$.

$$\overrightarrow{b} = (b_1, b_2, b_3)$$

est le vecteur marge colonne; ainsi, b_i est le nombre de résidus pour lesquels, la structure secondaire observée est i, $1 \leq i \leq 3$.

Dans ces conditions, on introduit:

Q_i^{obs} $=100\text{x}(\frac{card[S(v_i)\bigcap S(w_i)]}{b_i}) = 100\text{x}A_{ii}/b_i$ représente, d'entre tous les résidus observés dans la structure i, le pourcentage de ceux qui ont été correctement prédits; $i = \alpha, \beta, L$.

Q_i^{pred} $=100\text{x}(\frac{card[S(v_i)\bigcap S(w_i)]}{a_i}) = 100\text{x}A_{ii}/a_i$ représente, d'entre tous les résidus prédits dans la structure i, le pourcentage de ceux qui ont été correctement prédits; $i = \alpha, \beta, L$.

Q_{total} $=100\text{x}(\sum_{i=1}^{3} A_{ii}/N)$ représente le pourcentage total de résidus bien classés.

N est la longueur de la chaîne; c.a.d. le nombre de résidus.

Ces pourcentages estiment la performance globale de tous les résidus de l'ensemble des protéines. Si on veut estimer la moyenne de précison d'une seul protéine on peut utiliser:

$$< Q >_{chain} = \frac{1}{N^{chain}} \sum_{c=1}^{N^{chain}} Q_{total}^{c}$$

où N^{chain} est le nombre total de protéines et Q_{total}^{c} la précison de la prédiction pour la protéine (on dit encore chaine) c. L'écart type de $< Q >_{chain}$, σ_{chain}, fourni une estimation de l'intervalle de précision attendu pour une protéine.

B. W. Matthews en 1975 considère une autre mesure pour comparer les deux structures. Ce coefficient n'est en fait rien d'autre que celui introduit et étudié par K. Pearson (1900) au début du siècle. Il s'écrit ici:

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i+u_i)(p_i+o_i)(n_i+u_i)(n_i+o_i)}} \qquad \text{pour i=}\alpha, \beta, L.$$

où,

- p_i est le nombre total de résidus de la structure i correctement prédits ($p_i = card[S(v_i) \bigcap S(w_i)] = A_{ii}$);
- n_i le nombre total de résidus d'une structure autre que i et qui ont été classés dans une structure $j \neq i$ ($n_i = \sum_{j\neq i}^{3} \sum_{k\neq i}^{3} card[S(v_k) \bigcap S(w_j)] = \sum_{j\neq i}^{3} \sum_{k\neq i}^{3} A_{jk}$);
- $u_i = \sum_{j\neq i}^{3} A_{ij}$ et $o_i = \sum_{j\neq i}^{3} A_{ji}$ représentent respectivement le nombre de résidus sousestimés et surestimés.

La mesure Info:

Nous avons bien exprimé au chapitre 2 que la donnée d'un coefficient d'association entre variables qualitatives, permettait d'évaluer - avec le point de vue que suppose un tel coefficient - la qualité de la prédiction, en appliquant ce coefficient au tableau $\overrightarrow{\mathcal{A}}$ ci-dessus défini.

[Rost & Sander, 1993] ont introduit une mesure qu'ils désignent par $Info$ et qui se présente comme le logarithme du rapport entre deux probabilités, P_{obs} et P_{pred}; où P_{obs} estime la probabilité d'observer une chaine particulière de b résidus avec b_i résidus dans la structure i et P_{pred} est la probabilité d'une réalisation particulière du tableau A.

Ainsi,

$$Info = ln\{\frac{P_{pred}}{P_{obs}}\}.$$

L'interprétation du coefficient reste peu claire au niveau de l'article cité. Nous allons dans ces conditions chercher à en donner une vision explicite.

$$Info = ln \left\{ \frac{Pr(\overrightarrow{\mathcal{A}} / \overrightarrow{a})}{Pr(obs/\overrightarrow{b})} \right\} \qquad (*)$$

où $Pr(obs/\overrightarrow{b})$ est la probabilité de la chaîne observée, compte tenu de sa composition (définie par $\overrightarrow{b}$), dans l'hypothèse où toutes les chaînes ont la même probabilité d'apparaître.

$$Pr(\overrightarrow{\mathcal{A}} / \overrightarrow{a}) = \left(\frac{\prod_{i=1}^{3} a_i!}{\prod_{i=1}^{3} \prod_{j=1}^{3} A_{ij}!} \right)^{-1}$$

et

$$Pr(obs/\overrightarrow{b}) = \left(\frac{N!}{\prod_{j=1}^{3} b_j!} \right)^{-1}$$

La quantité sous le signe accolades de (*) peut se mettre sous la forme:

$$\frac{\left(\frac{N!}{\prod_i \prod_j A_{ij}!} \right)^{-1}}{\left(\frac{N!}{\prod_i a_i!} \right)^{-1} \cdot \left(\frac{N!}{\prod_j b_j!} \right)^{-1}}$$

Il s'agit donc du rapport entre la probabilité d'une partition jointe de type $\overrightarrow{\mathcal{A}}$ et le produit entre deux probabilités marginales. La première est celle d'une partition de type $\overrightarrow{a}$ et la seconde, de type $\overrightarrow{b}$. Le produit entre ces deux probabilités marginales correspond au cas de l'indépendance.

Ce rapport est ce que l'on appelle une densité de probabilité de la loi jointe par rapport aux lois marginales.

Info est donc le logarithme de cette densité de probabilité.

Après calcul, et en utilisant l'approximation donnée par la formule de Stirling, Rost & Sander arrivent à l'expression approximée:

$$Info = 1 - \frac{\sum_{i=1}^{3} a_i * ln\ a_i - \sum_{i,j=1}^{3} A_{ij} * ln\ A_{ij}}{N * ln\ N - \sum_{i=1}^{3} b_i * ln\ b_i}$$

Rost & Sander concluent alors que $Info = 0$ si $A_{ij} = N/9$, $i, j = 1, 2, 3$. Nous arrivons à une expression beaucoup plus général: $Info = 0$ en cas d'indépendance entre les deux structures; l'observée et celle qui a été prédite. En effet, s'il y a indépendance entre ces deux structures, alors $A_{ij} = \frac{b_i . a_j}{N}$ et le coefficient $Info$ devient

$$Info = 1 - \frac{\sum_{i=1}^{3} a_i * ln\ a_i - \sum_{i,j=1}^{3} \frac{b_i . a_j}{N} * ln\ \frac{b_i . a_j}{N}}{N * ln\ N - \sum_{i=1}^{3} b_i * ln\ b_i} =$$

$$1 - \frac{\sum_{i=1}^{3} a_i * ln\ a_i - \frac{1}{N} \sum_{i,j=1}^{3} b_i . a_j * (ln\ b_i + ln\ a_j - ln\ N)}{N * ln\ N - \sum_{i=1}^{3} b_i * ln\ b_i} =$$

$$1 - \frac{\sum_{i=1}^{3} a_i * ln\ a_i - \frac{1}{N} \sum_{i,j=1}^{3} b_i . a_j * (ln\ b_i) - \frac{1}{N} \sum_{i,j=1}^{3} b_i . a_j * (ln\ a_j) + \frac{1}{N} \sum_{i,j=1}^{3} b_i . a_j * (ln\ N)}{N * ln\ N - \sum_{i=1}^{3} b_i * ln\ b_i}$$

or

$$\sum_{i,j=1}^{3} b_i . a_j * (ln\ b_i) = \sum_{i=1}^{3} b_i * (ln\ b_i) * (\sum_{j=1}^{3} a_j) = \sum_{i=1}^{3} b_i * (ln\ b_i) * N;$$

de façon analogue

$$\sum_{i,j=1}^{3} b_i.a_j * (ln\ a_j) = \sum_{j=1}^{3} a_j * (ln\ a_j) * N;$$

et

$$\frac{1}{N} \sum_{i,j=1}^{3} b_i.a_j * (ln\ N) = ln\ N * \sum_{i,j=1}^{3} \frac{b_i.a_j}{N} = ln\ N * \sum_{i,j=1}^{3} A_{ij} = N * ln\ N.$$

Finalement, on obtient

$$Info = 1 - \frac{\sum_{i=1}^{3} a_i * ln\ a_i - \sum_{i=1}^{3} b_i * ln\ b_i - \sum_{j=1}^{3} a_j * ln\ a_j + N * ln\ N}{N * ln\ N - \sum_{i=1}^{3} b_i * ln\ b_i} = 0$$

Nous avons ainsi démontré que $Info = 0$ en cas d'indépendance entre la structure observée et la structure prédite.

$Info = 1$ en cas de prédiction parfaite, c'est à dire, si $A_{ij} = 0$ pour $i \neq j$ et $A_{ii} = b_i$, $i, j = 1, 2, 3$.

Cette quantité $Info$ a donc une relation avec la probabilité du tableau A de s'écarter d'un tableau d'indépendance. Une méthode de prédiction qui, par exemple, prédit très bien les boucles et mal les hélices et les brins peut engendrer un grand pourcentage global de résidus bien classés; toutefois, la mesure $Info$ est sensible aux surprédictions et aux sousprédictions, qui font également décroître sa valeur.

On aboutit finalement aux tableaux de résultats suivants:

Les résultats du premier tableau semblent montrer une grande stabilité de comportement entre les coefficients les plus classiques; c'est à dire, les quatre premiers (Gini, χ^2, Shannon et Twoing). En effet, on voit que l'intervalle de variation pour Q_{chain} a une amplitude de 0.7, ce qui montrent leur proximité; toutes les résultats, à l'exception de ceux pour s et $RANDOM$, sont dans l'intervalle [66%, 67%] ce qui montre que, bien que certaines différences soient statistiquement significatives en ce qui concerne Q_{chain}, elles ne sont pas très importantes. Pour Q_{total}, seulement le coefficient de Matusita peut-être inclus dans le même intervalle de variation ([65.3 , 65.8]), bien que Q_1, R et φ sont très proches.

Les coefficients utilisés classiquement (Gini, χ^2, Shannon et Twoing) ont un très bon comportement; et ce, aussi bien pour la qualité globale de la prédiction (Q_{chain} et Q_{total}) que classe par classe [Q_α^{obs}, Q_β^{obs} et Q_L^{obs} (resp. Q_α^{pred}, Q_β^{pred} et Q_L^{pred})]. Quand à Q_α^{obs} ces quatre coefficients donnent les meilleurs résultats, quoique le coefficient φ de Lerman donne des résultats semblables. Il est clair que les nouveaux coefficients ne font pas nécessairement mieux. Cependant - surtout lorsque leur conception statistique est élaborée - ils ne font pas sensiblement plus mal et peuvent faire localement mieux. C'est ainsi que pour le coefficient Q_1, on obtient près de 64% pour Q_α^{pred}; alors que pour Gini, il s'agit de 60%. De même, on obtient avec R, près de 60% pour Q_β^{obs}; alors que le score est de 58% avec Gini. On pourra également noter la très bonne performance du coefficient de Matusita pour Q_β^{obs} et Q_α^{pred}. Les résultats des quatre coefficients classiques en ce qui concerne Q_β^{obs} ne sont pas mauvais; mais les coefficients de Matusita, $Aff.\Delta$, R font mieux; le résultat pour Q_1 (57.8), est aussi bon et φ n'est pas loin. Toutefois, si on veut regarder aussi bien la performance globale que chacune des performances locales, le coefficient φ, avec à peine une performance plus légère, est comparable aux coefficients traditionnels. Ce coefficient, qui a été élaboré de façon à privilégier les classes minoritaires, montre bien son intérêt. Il dépend d'un paramétrage favorisant la prédiction des classes minoritaires, qui sont ici les seules intéressantes, puisqu'elles correspondent aux structures α et β. La recherche doit pouvoir se poursuivre sur le paramétrage le plus adéquat. Comme le plus important c'est de prédire bien les deux classes α et β , on peut peut-être élire les quatre coefficients classiques plus φ, comme étant les "meilleurs" coefficients pour ces données; après Matusita, $Aff.\Delta$, R et Q_1.

Les résultats médiocres de s et $RANDOM$ étaient nécessaires à notre démonstration statistique de la nécessité d'un "bon" coefficient. Il y a sans doute des différences significatives entre les deux

Coeff.	Q_{chain} (%)	Q_{total} (%)	$Info$ [0,1]	Q_α^{obs} (%)	Q_α^{pred} (%)	C_α [-1,1]	Q_β^{obs} (%)	Q_β^{pred} (%)	C_β [-1,1]	Q_L^{obs} (%)	Q_L^{pred} (%)	C_L [-1,1]
GINI	67	65.8	.174	51	60	.424	57.9	65.8	.474	78.5	68.1	.464
χ^2	66.6	65.5	.171	51.4	59.6	.424	57.9	65	.468	77.6	68.1	.459
Shannon	66.3	65.3	.168	50.9	59.6	.421	58	64.8	.467	77.3	67.7	.451
TWOING	66.7	65.6	.172	50.6	59.4	.419	58.5	65.6	.476	77.8	68	.458
Matusita	66.8	65.4	.169	44.7	63.8	.417	62.7	63.1	.477	77.8	67.1	.445
Aff. W	65.9	64.5	.159	48.1	57.3	.39	54.3	66.2	.457	79.4	66.4	.445
Aff. Δ	66.1	64.7	.160	43.5	63.2	.407	60.7	62.4	.460	78.2	66.3	.437
s	63.1	61.7	.124	42.2	55.5	.346	47.8	63.1	.398	80.5	63.1	.400
Q_1	66.4	65.2	.165	45.1	63.9	.420	57.8	64.2	.461	80.3	66.1	.446
R	66.3	65.2	.166	45.4	62	.408	59.9	63.4	.465	78.9	67.3	.454
φ	66	64.8	.162	50.4	59	.414	56.5	64.3	.454	77.4	67.3	.445
RANDOM	64	62.8	.136	42.9	58.9	.374	53	61.4	.412	79.3	64.7	.418

FIG. 7 - Tableau des performances.

GINI	χ^2	Shannon	TWOING	Matusita	Aff. W	Aff. Δ	s	Q_1	R	φ	RANDOM
295	278	286	241	305	351	327	764	248	282	286	545

FIG. 8 - Tableau des complexités.

coefficients s et $RANDOM$ et les autres dix, ce qui démontre que les affirmations de certains auteurs, en particulier Mingers (1989), manquent de sens. Ces deux coefficients sont néttement inférieurs aux autres.

En rappelant que s est l'indice brut de comparaison entre deux partitions (ici, celle induite par la variable prédictive et celle, par la variable à prédire), on se rend compte du saut important dans la qualité des résultats, lorsqu'on passe de s, aux coefficients Q_1 et R qui sont élaborés par normalisations à partir de cet indice brut.

Quand au tableau des complexités, il est évident que les coefficients les plus classiques, plus les indices Q_1, R et φ, donnent les meuilleurs résultats, suivis par le coefficent de Matusita et les deux affinités; encore ici, le coefficient de choix aléatoire et le coefficient s donnent les résultats les plus mauvais.

5.2.3 Correction de la prédiction

Dans leur article, Rost & Sander définissent un indice RI qu'ils appellent "Reliability Index" et qu'ils utilisent pour filtrer leur prédiction. RI se met sous la forme suivante:

$$RI = [10.(p_{max} - p_{med})]$$

où [ ] indique la partie entière et où p_{max} est la probabilité la plus forte et p_{med}, celle qui vient ensuite, relativement à la structure secondaire à prédire. Ainsi, si, à une feuille de l'arbre, les proportions des trois classes α,β et L sont respectivement 0.9, 0.05 et 0.05; alors RI vaut [10.(0.9-0.05)]=8. Si par ailleurs, ces proportions valent 0.5,0.4 et 0.1; alors RI vaut [10.(0.5-0.4)]=1.

RI peut ainsi être calculé pour chaque feuille de l'arbre de décision. Et alors, pour une nouvelle protéine à prédire, nous pouvons, soit faire sortir pour chaque résidu (compte tenu de son affectation), la valeur de RI; soit en faire une moyenne relativement à l'ensemble des résidus d'une même séquence protéique. Cette valeur, RI, sert donc de mesure de confiance dans la prédiction de la structure secondaire d'une nouvelle protéine.

En ce qui nous concerne, nous exploiterons toute la distribution (et non seulement la probabilité dominante) de la structure secondaire sur une feuille de l'arbre de décision (où atterit un résidu dont il faut prédire la structure secondaire) comme un des éléments de la correction de la prédiction.

Le second facteur de la correction favorise la rareté de la classe; ainsi, relativement aux classes X, H et E (notées ici L,α et β), on introduit les valuations $x = 1/0.466 = 2.146$, $h = 1/0.29 = 3.448$ et $e = 1/0.244 = 4.098$.

Dans le troisième et le plus intéressant facteur de correction, on introduit au niveau de la structure secondaire, l'environnement de la position à prédire et ce, au moyen d'une fenêtre centrée sur la position à prédire. Nous choisissons cette fenêtre de longueur 5. Ce qui interviendra dans la qualité de la prédiction c'est la longueur du segment connexe le plus long formé de la même lettre (X, H ou E).

Plus précisèment, on considère la structure secondaire prédite par l'arbre de décision, conformément - pour chacune des feuilles - à la probabilité dominante. On obtient (voir introduction) une chaîne formée avec l'alphabet $\{X, H, E\}$. C'est par rapport à cette chaîne produite que la correction de la prédiction va s'opérer. Pour comprendre l'intérêt de cette correction, imaginons la chaîne suivante et la

...H [HHXHH] E...

fenêtre de 5 lettres, centrée sur la position prédite en X. Mais; supposons, que pour la feuille concernée de l'arbre, la distribution des trois classes H, E et X soit (0.4,0.1,0.5). On se rend compte que si on remplace X dont la probabilité est 0.5 par H dont la probabilité 0.4 est plus faible, mais diffère de peu; la longueur de la plus grande connexité est maximale et concerne la structure H qui est plus intéressante que X.

Une expérimentation poussée a montré l'intérêt du critère suivant dont la valeur maximale peut décider du changement d'affectation; par rapport - encore une fois - à la séquence produite qui reste la référence:

$$\mathrm{C}(\mathcal{U}) = 10 * \mathrm{p}_{\mathcal{U}}(\mathrm{lcpg}(\mathcal{U}))^2 * \mathrm{u}$$

$p_{\mathcal{U}}$ est la probabilité de $\mathcal{U}(\mathcal{U} = H, E$ où $X)$ pour la feuille concernée de l'arbre de décision; de sorte que $10 * p_{\mathcal{U}}$ est compris entre 0 et 10. $lcpg(\mathcal{U})$ est la longueur de la plus longue connexité, en mettant $\mathcal{U}$. u est l'inverse de la fréquence de $\mathcal{U}$ dans le corpus ($u = h = 3.448$ ou $u = e = 4.098$ ou $u = x = 2.146$).

Ainsi, par rapport à l'exemple ci-dessus

$$C(H) = 10 * 0.4 * 5^2 * 3.448 = 344.80$$

$$C(E) = 10 * 0.1 * 1^2 * 4.098 = 4.098$$

$$C(X) = 10 * 0.5 * 1^2 * 2.146 = 10.730$$

Dans ces conditions, on remplacera X par H.

Cette procédure a permis de passer, pour la prmière case du tableau des performances (en ce qui concerne Q_{chain}), de 67% à 69%; alors que pour la seconde case sur la même ligne (en ce qui concerne Q_{total}), de 65.8% à 68%.

La plupart des bonnes méthodes de prédiction, ont une précision de résidus bien classés qui varie entre 60% et 71%, quoique, pour certaines méthodes, ces valeurs ne donnent pas une estimation correcte, soit parce qu'elles ont été testées dans des ensembles de protéïnes homologues à l'ensemble d'apprentissage, soit parce que le nombre de ces ensembles-tests était insuffisant. L'objectif ultime de toutes ces méthodes qui, depuis deux dizaines d'années, a été d'arriver à 100% de prédictions correctes, commence maintenant à changer. En effet, la prédiction de la structure secondaire d'une protéine constitue un pas intermédiaire pour arriver à la prédiction de la structure tertiaire, en particulier les aspects de la structure tertiaire qui déterminent la fonction de la protéïne. Or, [Rost et al., 1994] ont démontré qu'il y a des familles de protéines de structure tertiaire similaire mais qui différent dans leur strucure secondaire, d'environ 12%. Par conséquent, on ne peut pas s'attendre à prédire la structure secondaire avec plus de précision que la variation naturelle observée dans les familles structurelles. Ainsi, un obectif plus sage ([Rost & Sander, 1994]) est d'aboutir à 88% de résidus bien prédits.

En 1994, Rost, Sander et Schneider définissent une nouvelle mesure qui, au lieu de comparer les deux structures, résidu par résidu, compare les segments de la structure observée avec ceux de la structure prédite. L'objectif final de la prédiction de la structure secondaire n'est pas d'arriver à 100% de précision, mais plutôt de localiser, approximativement, les hélices, brins et boucles, et après, utiliser ces segments dans la prédiction de la structure terciaire. D'autre part, Solovyev & Salamov, en 1994, ont développé une méthode qui utilise l'analyse discriminante linéaire et dont le but est de localiser des segments d'hélices α ou brins β entiers au lieu de prédire la classe (E, H ou X) résidu par résidu.

A vrai dire, pour comparer des méthodes, il est nécessaire de travailler sur le même corpus de données, formé de protéines non homologues. Il importe également que l'ensemble test soit indépendant de l'ensemble d'apprentissage. Enfin, il y a lieu de bien mettre en évidence les informations prises en compte par une méthode donnée. En effet, une méthode exploitant une connaissance riche doit être évaluée avec plus de sévérité qu'une méthode ne prenant pas en compte une telle connaissance.

6 Conclusion et Prospectives

Ce travail présente essentiellement trois innovations :

i) l'introduction dans les arbres de décision binaires de nouvelles mesures de choix d'un attribut;

ii) l'utilisation de variables qualitatives à très grand nombre de modalités et à n'importe quel nombre

de classes en arbres de décision;
iii) application de la méthode développée, ARCADE, au problème de la prédiction de la structure secondaires des protéines.

La motivation initiale qui nous a conduit à la réalisation de ce travail concerne l'influence sur la construction des mesures d'association entre variables qualitatives. En effet, certains auteurs tels que Breiman & al (1984) et surtout Mingers (1989) donnent peu d'importance à la façon de choisir l'attribut binaire pour segmenter le contenu de chaque nœud d'un arbre de décision; Mingers a même écrit que ce choix peut être aléatoire. Ces affirmations nous ont surpris, car au moins dans le domaine de l'analyse classificatoire, les mesures d'association entre variables qualitatives, qui constituent une grande palette, jouent un rôle très important. Nous avons pu démontrer que cela était aussi vrai dans le domaine des arbres de décision; il y a des différences statistiquement significatives entre les résultats des divers coefficients. Nous avons developpé un nouveau coefficient, φ, qui privilégie les concepts qui sont rares. Ainsi et dans le cadre de notre application, nous avons utilisé douze coefficients, la plupart pour la première fois dans le domaine des arbres de décision. Les différentes performances obtenues nous permettent de séparer les coefficients en trois groupes:

- Le groupe des meilleurs mesures d'association, qui comprend les coefficients classiques de Gini, χ^2, Twoing, Shannon et le nouveau coefficient φ; ces coefficients ont un très bon comportement, aussi bien pour la qualité globale de la prédiction que classe par classe. La recherche sur le nouveau coefficient φ doit se poursuivre, en particulier en ce qui concerne l'étude de nouveaux paramétrages favorisant les classes minoritaires.

- Le groupe composé par les coefficients de Matusita, Affinité W, Affinité Δ, Q_1 et R. Ces coefficients, qui sont nouveaux dans les arbres de décision, ne font pas sensiblement plus mal; localement (c.a.d., classe par classe),certains d'entre eux (Q_1, R et Matusita) peuvent même faire mieux

- le groupe formé par le coefficient de choix aléatoire $RANDOM$ et par le coefficient "brut" (non centré ni réduit) s, qui donnent les résultats les plus médiocres. Cela nous a ainsi permis de démontrer la nécessité d'un "bon" coefficient.

On a encore constaté que dans chaque groupe il existe des différences pour la complexité de l'arbre de décision final, mesurée par le nombre de feuilles. Nous avons ainsi abouti à conclure que différents critères engendrent des différences significatives en performance et en complexité.

Nous pensons étendre cette étude au cas des arbres non binaires de décision. Cela va nous permettre en outre, l'introduction de coefficients d'association entre variables relationnelles où on tient compte d'une sémantique (structure) qui est derrière l'ensemble des catégories des variables prédictives et de la variable à prédire. Encore ici, le problème de la prédiction de la structure secondaire des protéines fournira une application privilégié. Nous allons en effet, pouvoir tenir compte d'une structuration de l'ensemble des valeurs des variables prédictives et d'ailleurs aussi de la variable à prédire qui représente ici la structure secondaire. Cette structuration tiendra compte d'une connaissance qu'on peut induire. Mais alors dans ce cas, on ne pourra plus faire appel à des indices, tels que ceux que nous avons utilisé et qui ne permettent que la comparaison entre variables qualitatives nominales. Nous serons conduits à faire intervenir des coefficients d'association entre variables qualitatives relationnelles [Lerman 1992a, 1992b].

La partie la plus importante de la recherche rapportée ici et qui résulte de l'application rencontrée, concerne la gestion des variables qualitatives à très grand nombre de modalités pour la construction d'arbres binaires de décision. Nous avons défini une approche permettant de récolter un ensemble d'attributs binaires synthétiques, en un nombre assez petit; ce qui suppose une réduction vertigineuse de la complexité. Rappelons en effet, que nous avons pu passer d'une variable qualitative à 20^4 modalités, à une vingtaine d'attributs binaires sybthétiques. Cette approche suppose de façon conjointe et selon un schéma précis, une opération de factorisation, des opérations de classification et une opération de

représentation (cf. §4). Nous avons alors pu faire entrevoir que la puissance de cette approche qui nous a résolu notre problème de prédiction, lui permettait de s'étendre dans des situations plus générales. Précisément, une des situations qui pourra nous concerner, consiste en l'adaptation à des arbres non binaires.

Le troisième volet fortement motivant est fourni par l'application elle même, qui est très intéressante. Bien que nous ne pouvons rigoureusement faire des comparaisons (cf. dernier alinéa du paragraphe 5.3.3), nous avons le sentiment d'avoir abouti à de très bons résultats; 69% pour Q_{chain} et 68% pour Q_{total}, avec l'usage de l'indice de Gini. La qualité de ce résultat doit intégrer les éléments suivants d'appréciation:

- les 151 protéines globulaires sont mutuellement non-homologues;
- une réalisation de l'arbre de décision est effectuée sur un ensemble d'apprentissage disjoint de l'ensemble test;
- la prédiction s'effectue résidu par résidu, avec il est vrai, une connaissance modérée sur l'environnement de la position prédite, au niveau de la structure secondaire telle qu'elle a été directement prédite (résidu par résidu);
- la simplicité finale de la méthode ARCADE.

Il y a à notre connaissance peu de méthodes qui font - a priori - mieux que la nôtre. La célèbre méthode PHD de Rost & Sander qui atteint 70.8% sur une famille de protéines globulaires, exploite beaucoup plus que nous le faisons, l'environnement de la lettre prédite dans le déroulement de la structure secondaire. D'autre part, et surtout, ils utilisent pour leur prédiction des séquences homologues; pour une protéine non globulaire ou qui n'a pas d'homologues, la performance baisse très sensiblement. Enfin et c'est lié, la construction de la méthode de Rost & Sander, nous parait assez complexe ... cependant, nous admettons que sur ce dernier aspect, notre point de vue ne soit pas partagé ...

Références

Bacelar-Nicolau, H. (1980): *Contribuições ao estudo dos coeficientes de comparação em análise classificatória.* Phd Thesis, Faculdade de Ciências da Universidade de Lisboa.

Bacelar-Nicolau, H. (1982a): *L'affinité: un coefficient de similarité.*Actes des Journées de Classification. Toulouse/Nancy/Bruxelles, 81-87.

Bacelar-Nicolau, H. (1982b): *Affinité et analyse classificatoire.*Actes des Journées de Classification. Toulouse/Nancy/Bruxelles, 71-80.

Bacelar-Nicolau, H. (1988): Two probabilistic models for classification of variables in frequency tables .*Classification and Related Methods of Data Analysis*, H. H. Bock(Editor) . Elsevier Science Publishers B. V. (North-Holland),pp. 181-186.

Breiman, L., Friedman, J. H., Olshen, A. and Stone, C. J. (1984): *Classification and Regression Trees.* Wadsworth, Belmont.

Colloc'h N.,Etchebest C.,Thoreau E.,Henrissat B. & Mornon J.P. (1993): *Protein Engineering, (1993)*, 6, 377-382.

Costa Nicolau, F. (1985): *Analysis of a non hierarchical clustering method based on VL-similarity.* Meth. Oper. Research 53, pp. 603-610.

Everitt, B. S. (1977): *The Analysis of Contingency Tables.* Chapman and Hall, Londres.

Fisher, W.D. (1958): On grouping for maximum homogeneity. *J. Am. Statist. Assoc.* 53, pp.789-798.

Goodman, L. A. et Kruskal,W. H. (1979): *Measures of Association for Cross Classification.* Springer Verlag, Berlin, New-York.

Heath, D., Kasif, S. and Salzberg, S. (1993): Induction of Oblique Decision Trees. *Proc. of the IJCAI 93*, 13th International Joint C. On A.I., Chambéry, France.

Hubert, L. J. (1983): *Inference procedures for the evaluation and comparison of proximity matrices.* Numerical taxonomy, Ed. J. Felsenstein, NATO ASI Series, Springer Verlag.

Kendall , M. G. (1970): *Rank correlation methods.* Charles Griffin, fourth edition (first edition in 1948).

Krzanowsky, W. J. (1975): Discrimination and Classification Using Both Binary and Continuous Variables. *Journal of The American Statistical Association*, Volume 70, Number 352, pp. 782-790.

Lauro, N. et D'Ambra, L. (1984): *Analyse non Symétrique des correspondances.*Proceedings du 3éme Congrés "International Data Analysis and Informatics", North Holland, Amsterdam.

Lerman, I.C. (1970): Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité. *Revue Mathématiques et Sciences Humaines*, 8-éme année, numéro 32, pp 5-15.

Lerman, I. C. (1973): Étude distributionnelle de statistiques de proximité entre structures finies de même type; application à la classification automatique. *Cahiers du BURO, n^o* 19, Paris, 50p.

Lerman, I.C. (1981): *Classification et analyse ordinale des données.* Paris, Dunod.

Lerman, I. C. (1987): *Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles.* Rapport de recherche n. 702, Inria, Juillet 1987.

Lerman, I. C. (1988): *Structure maximale pour la somme des carrés d'une contingence aux marges fixées; une solution algorithmique programmée.* Rairo, vol. 22, n.2, pp. 83 à 136.

Lerman, I.C. (1992a): Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I. *Rev. Math. Infor. & Sci. Hum.*, 30e année, Paris, n. 118, 1992, pp. 35-52.

Lerman, I.C. (1992b): Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II. *Rev. Math. Infor. & Sci. Hum.*, 30e année, Paris, n. 119, 1992, pp. 75-100.

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

Lerman & Ghazzali 1991 Lerman I. C., Ghazzali, N. (1991): "What do we retain from a classification tree? am experiment in image coding" in Symbolic-Numeric data analysis and learning, edited by INRIA (E. Diday and Y. Lechevallier), Nova Science Publishers, september 1991, pp. 27-42.

Lerman, I. C. et Peter, Ph. (1985): *Elaboration et logiciel d'un indice de similarite entre objets d'un type quelconque. Application au problème du consensus en classification.* Publication Interne Irisa n. 262, Juillet 1985, Rennes.

Lerman, I.C., Peter, Ph. et Leredde, H. (1993-1994): Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens). *La Revue de Modulad, n^o* 12,pp.33-70 et n^o 13, pp.63-90.

Lerman et Pinto Da Costa Lerman, I. C., Pinto Da Costa, J. F. (1995): Methodological developments in Decision Trees. An application to protein secondary structure prediction. To be published in the Proceedings of OSDA95 (Ordinal Symbolic Data Analysis), 20-23 june 1995, Springer-Verlag.

Marcotorchino, F. (1984a): *Utilisation des Comparaisons par Paires en Statistique des Contingences. Partie I.* Etude du centre Scientifique IBM-France No F 069, Février 1984.

Marcotorchino, F. (1984b): *Utilisation des Comparaisons par Paires en Statistique des Contingences. Partie II.* Etude du centre Scientifique IBM-France No F 071, Octobre 1984.

Matusita, K. (1955): *Decision rules based on distance for problems of fit, two samples and estimation* . Ann. Math. Stat., vol. 16.

Matusita, K. (1977): Cluster analysis and affinity of distributions. *Recent developments in Statistics* , Barra editor, North Holland, pp. 537-544.

Messatfa, H. (1990): *Unification relationnelle des critères et structures optimales des tables de contingences.* Thèse de doctorat de l'Université Pierre et Marie Curie, 5 mars 1990.

Mingers, J. (1989): An Empirical Comparison of Selection Measures for Decision-Tree Induction. *Machine Learning 3: 319-342, 1989.* Kluwer Academic Publishers - Manufactered in the Netherlands

Müller, W., Wysotzki, F. (1994): A Splitting Algorithm, Based on a Statistical Approach in the Decision Tree Algorithm CAL5. *Proc. of the ECML 94*, 7th European Conference on ML, Catania, Sicily, April 1994.

Nakhaeizadeh, G. (1994): Interaction Between Machine Learning and Statistics, An Overview. *Proc. of the ECML 94*, 7th European Conference on ML, Catania, Sicily, April 1994.

Ouali-Allah, M. (1991): *Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques.* Thèse, Université de Rennes I, 1991.

Quinlan, J.R. (1986): Induction of Decision Trees. *Machine Learning*, pp.81-106.

Rost, B., Schneider, R., Sander, C. (1993): Progress in protein structure prediction? *Trends in Biochemical Sciences*, April, pp 120-123.

Rost, B., Sander, C. (1993): Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, pp. 584-599.

Rost, B.,Sander, C., Schneider, R. (1994): Redefining the Goals of Protein Secondary Structure Prediction *J. Mol. Biol.* 235, pp. 13-26

Solovyev, V.V., Salamov, A.A. (1994): Predicting α-helix and β-strand segments of globular proteins. *CABIOS.* Vol. 10 n. 6, pp. 661-669.

Taylor, C. C., (1994): Distance-based Decision Trees. *Proc. of the ECML 94*, 7th European Conference on ML, Catania, Sicily,

Tiago de Oliveira, J. (1982): *The δ-method for obtention of asymptotic distributions; applications.* Publ. Ins. Stat. univ. Paris, vol.XXVII (1982), pp. 49-70.

Wald, A. et Wolfovitz, J. (1944): *Statistical tests based on permutations of the observations.* Ann. Math. Stat., vol. 15.

Van de Merckt, T. (1993): Decision Trees in Numerical Attribute Spaces. *Proc. of the IJCAI 93*, 13th International Joint C. On A.I., Chambéry, France.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399