L'arbre à nœuds probabilistes:

Une nouvelle approche à la construction d'arbres de prédictions.

Antonio Ciampi
Université McGill, Montréal, Québec, Canada
antonio.ciampi@mcgill.ca

Résumé : Nous montrons la connexion entre, d'une part, l'analyse de données symboliques, et, d'autre part, certains algorithmes d'apprentissage supervisé pour la prédiction d'une variable réponse qualitative ou quantitative. Dans le contexte des données symboliques, nous avions développé un algorithme de prédiction à arbre; cela nous avait permis de traiter des données imprécises et de construire des arbres de prédiction classiques à partir de ce type de données. Par la suite, nous avons repris le problème de la construction d'arbres pour des données précises (numérique), mais en permettant des nœuds probabilistes ou 'tendres', c'est-à-dire des nœuds correspondant à des décisions probabiliste du type : 'aller à gauche avec probabilité *p* et à droite avec probabilité 1-*p*'. Un tel arbre décrit la distribution prédictive conditionnelle de la variable réponse comme un mélange de distributions, tel que les coefficients des composantes du mélange dépendent des variables : ces coefficients sont en effet des produits de fonctions sigmoïdes de certaines variables de prédiction choisies par l'algorithme guidé par les données. Nous décrivons une approche EM pour l'estimation des paramètres du modèle correspondant. La méthode a été évaluée par des simulations et des analyses de données réelles. Nous discutons, pour conclure, les avantages et les limites de ce type d'arbre en comparaison avec les arbres conventionnels.

Mots clés : apprentissage supervisé, prédictions, arbre à nœuds probabilistes, données symboliques, données imprécises.

Abstract: We show the connection between symbolic data analysis (SDA) and certain algorithms of supervised learning for the prediction of a continuous or categorical outcome. In the context of SDA, we had previously developed a tree-growing algorithm which allowed us to handle imprecise data and to construct classical prediction trees from such data. Later we went back to tree-growing for classical (numerical) data, and proposed the notion of probabilistic or soft node, that is a node representing a decision of the type: 'go left with probability *p* and go right with probability 1-*p*'. Such a tree-shaped predictor describes the conditional predictive distribution of the outcome as a mixture of distributions, with mixing coefficients which are functions of certain predictor variables chosen by the algorithms guided by the data. We describe an EM approach to the estimation of the predictive model parameters. The method is evaluated by simulation and real data analyses. In conclusion, we discuss the advantages and the limitations of the tree with soft nodes in comparison with conventional prediction trees.

Keywords: supervised learning, prediction, trees with soft nodes, symbolic data analysis, imprecise data.

1. Introduction

Un arbre de prédiction est une structure telle qu'illustrée dans la Figure 1. L'arbre est représenté avec la racine en haut, ses branches pointant vers le bas, jusqu'aux feuilles. Le schéma illustre comment arriver a une prédiction concernant une variable dépendante (le nom d'une couleur, dans le cas de figure), quand on connait la valeur de certaines variables dites *de prédiction* (les *Y*'s) pour une observation donnée. La prédiction peut prendre la forme de *classement* (si $Y_1 < 4.8$, alors l'observation est verte) ou d'une *distribution de probabilité* (si $Y_1 < 4.8$, alors l'observation est verte avec une probabilité 3/5, noire avec une probabilité 2/5 et rouge avec une probabilité de 0). La figure montre aussi que chaque nœud de l'arbre induit une partition de l'espace de *Y*'s, d'où le nom de 'partition récursive' souvent utilisé pour les algorithmes de construction d'arbre.

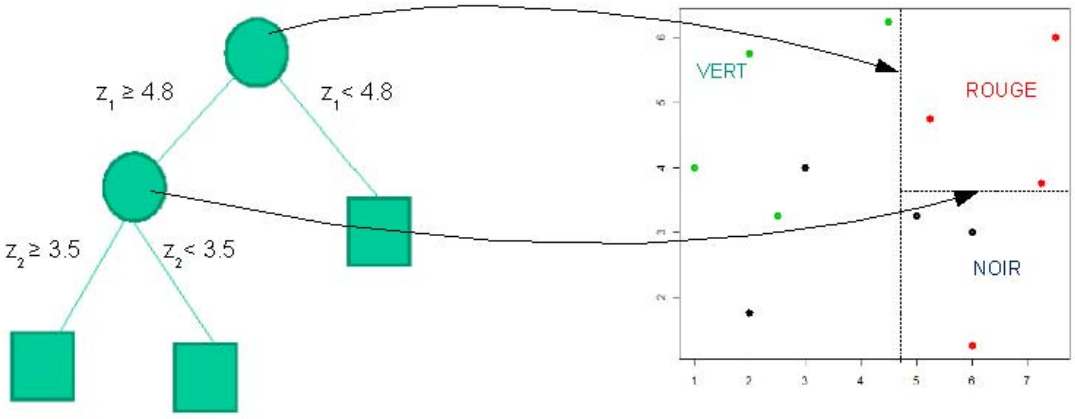

Figure 1 : Arbre de prédiction

En effet, il existe une grande variété d'algorithmes pour construire des arbres de prédiction à partir d'une base de données, dont le plus connu est sans doute CART (Classification And Regression Trees) [1]. Tous ces algorithmes partagent une idée centrale: à chaque pas d'une procédure récursive, un sous-ensemble des données est partitionné selon la valeur d'une variable de prédiction, la partition choisie étant celle qui optimise une mesure de qualité. Les variantes de cette approche de base sont très nombreuses. Cependant, nous ne mentionnerons ici que deux des choix fondamentaux qui génèrent des différentiations importantes entre algorithmes: 1) choix de la mesure de la qualité d'une partition; et 2) choix de la règle qui gouverne la taille de l'arbre obtenu par la procédure récursive. Dans nos travaux précédents, nous avons utilisé le terme 'information généralisée' [6] pour parler de la mesure de qualité d'une partition. En effet, le terme 'information' est bien approprié pour désigner la mesure de qualité d'une partition employée dans un cas particulier, celui où les données sont représentées de façon adéquate par un modèle probabiliste paramétrique. Dans ce cas, il a été démontré qu'un rapport de vraisemblance mesure l'information que nous pouvons acquérir à propos de la variable dépendante si nous l'étudions séparément sur chaque ensemble de la partition [3]. L'idée d'information est aussi

centrale dans les travaux de Quinlan, auteur d'un des premiers algorithmes de construction d'arbres dans le domaine de l'apprentissage machine [4].

Le succès des arbres en analyse de données est dû en grande partie à la simplicité de la règle de prédiction qu'ils décrivent : selon les réponses à un nombre, souvent petit, de questions binaires, on arrive à une prédiction directe de la variable dépendante. En revanche, si on n'est intéressé qu'à obtenir de prédictions correctes, il existe un très grand choix de techniques pour construire des règles de prédictions plus performantes. La pratique montre que les réseaux de neurones, les 'support vector machines' ou SVM et même les modèles de régression classiques, éventuellement obtenus avec des méthodes modernes de sélection et/ou de transformation de variables, peuvent produire des prédictions plus performantes que les algorithmes de construction d'arbres. Il y a aussi une famille de méthodes dites 'ensemblistes' basées sur des techniques de ré-échantillonnage, qui produisent des prédictions de qualité exceptionnelle, mais basées non pas sur un modèle de prédiction unique, mais sur un ensemble de modèles. En particulier, la méthode des Forêts Aléatoires (Random Forest) peut être vue comme une extension des arbres, puisqu'elle extrait des données un grand nombre d'arbres, lesquels, de façon coopérative, produisent une règle de prédiction [5].

On serait donc tenté de conclure que les arbres sont dépassés et qu'il faut accepter une certaine perte de facilité d'interprétation en échange d'une performance supérieure pour les techniques plus récentes. Ce n'est pas notre point de vue. Nous sommes convaincus que l'on peut garder une bonne mesure de facilité d'interprétation tout en améliorant le processus de construction d'arbre. Nous proposons notamment de transformer radicalement l'algorithme de partition récursive, *local* par sa nature, en le rendant *global*, c'est-à-dire en faisant intervenir toutes les données convenablement pondérées à chaque étape de la construction.

La section suivante relie l'idée principale de ce travail à certains développements de l'analyse de donnée symboliques [6]. Dans les sections 3 et 4, nous introduisons la notion de *nœuds probabilistes* et d'arbre à nœuds probabilistes respectivement. Deux exemples d'analyse de données réelles sont présentés dans la section 5. La section 6 contient le résultat de quelques expériences d'évaluation empirique de l'arbre probabiliste. Conclusions et directions de recherches courantes et futures sont esquissées dans la section 7.

2. Arbres pour données symboliques

Un arbre classique est construit à partir de données numériques, c'est-à-dire, d'une matrice dont les lignes sont les valeurs (numériques) que des variables prennent sur des observations. Un arbre classique, tel que celui de la Figure 1, peut servir, une fois obtenu, comme règle pour attribuer une valeur de la variable dépendante à toute (nouvelle) observation connue uniquement sur les variables explicatives. Que faire si nous voulions classer une observation dite 'symbolique' [1] ?

Ici nous nous limiterons à considérer un type simple de données symboliques: des données imprécises, avec l'imprécision représentée par une plage de valeurs possibles. Par exemple, une observation spécifique pourrait être : $z^* = (z_1, z_2) = ([4.5, 5.0], [3.4, 3.9])$, ce qui équivaut à spécifier que la valeur de z_1 se situe entre 4.5 et 5.0, et que celle de z_2 se situe entre 3.4 et 3.9. Supposons que z^* est observée sur un élément de la même population à partir de laquelle on a construit l'arbre de la Figure 1.

Comment classer une telle observation? Quinlan [10] a suggéré de l'envoyer, à chaque nœud, 'un peu à gauche' et 'un peu à droite'; plus spécifiquement, au nœud défini par une contrainte sur la variable z_i, on calcule la fraction de la plage de z_i de z^* satisfaisant la contrainte, et on envoie cette fraction de l'observation à gauche et la fraction complémentaire à droite. Nous pouvons, bien sûr, interpréter cela comme une décision probabiliste, avec probabilités estimées par les fractions observées, sous l'hypothèse implicite que la valeur (réelle mais inconnue) de z_i soit distribuée uniformément sur sa plage. La Figure 2 illustre cet exemple.

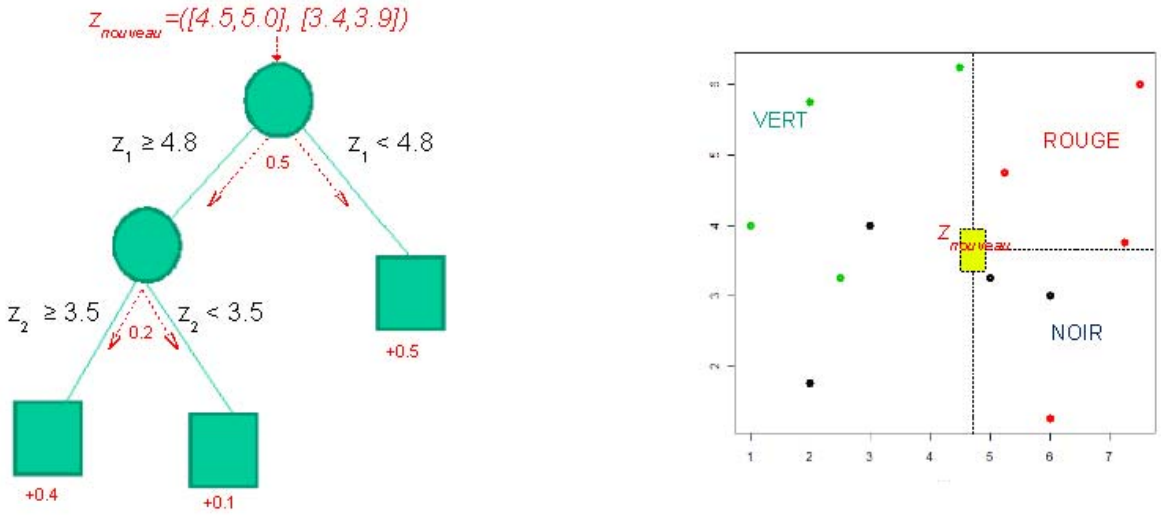

Figure 1: Classement d'une donnée imprécise (Quinlan, 1990).

S'il est possible d'utiliser un arbre classique, construit sur des données numériques, afin de classer une observation symbolique, serait-il possible de construire un arbre classique à partir de données symbolique? Nous avons donné une réponse affirmative à cette question dans [8]. Sans entrer dans les détails de la solution proposée, voici, dans la Table 1, une matrice de données symboliques de taille réduite (exemple artificiel) : nous avons ici des plages de variables de prédiction pour 12 observations; en revanche, chaque observation appartient sans ambigüité à une des trois classes possibles.

Ces données sont représentées en deux dimensions dans la Figure 3a), et l'arbre obtenu en appliquant l'algorithme développé en [8], est représenté dans la Figure 3b).

Finalement, on peut se poser une autre question : est-il possible de construire des règles probabilistes en forme d'arbre à partir de données classiques?

Cet article propose une réponse à cette question.

z_1	z_2	Classe
[1.0, 3.0]	[1.5, 2.0]	1
[2.5, 3.5]	[3.0, 5.0]	1
[3.5, 6.5]	[3.0, 3.5]	1
[5.0, 7.0]	[1.5, 4.5]	1
[4.0, 4.8]	[0.5, 2.0]	2
[7.0, 7.5]	[2.5, 5.0]	2
[7.0, 8.0]	[5.5, 6.5]	2
[4.0, 6.5]	[4.0, 5.5]	2
[3.0, 6.0]	[6.0, 6.5]	3
[0.5, 1.5]	[3.0, 5.0]	3
[1.5, 2.5]	[5.5, 6.0]	3

Table 1 : Exemple de données symboliques

--

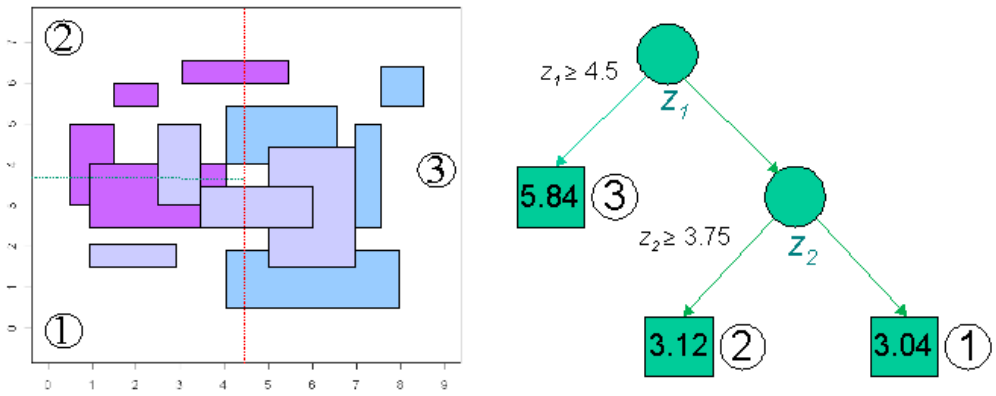

Figure 2 : Arbre dur obtenu à partir de données imprécises

3. Nœuds probabilistes

Un nœud probabiliste représente une règle décisionnelle probabiliste binaire : ‘ Au nœud *i,* prendre le chemin de gauche avec la probabilité $p(z_i)$ ou le chemin de droite avec la probabilité complémentaire $q(z_i) = 1- p(z_i)$’. La partition probabiliste correspondante est définie par la fonction $p(z_i)$, et nous limiterons la recherche de ladite fonction à une classe de fonctions sigmoïdes non-décroissantes, dont les valeurs sont comprises entre 0 et 1. Cette partition peut, en principe, être de meilleure qualité, c'est-à-dire qu’elle peut être associée à un gain d’information plus important, que celui de la meilleure partition binaire nette. En effet, on ne peut faire que mieux en cherchant dans un domaine plus vaste; or, la fonction indicatrice de l’intervalle $[a, \infty)$, $I\,[z \geq a]$, c'est-à-dire la fonction qui prend la valeur 1 sur l’intervalle $[a, \infty)$ et 0 ailleurs, est un cas limite de fonction sigmoïde non-décroissante, et elle représente un nœud ‘classique’ ou dur. Choisir la meilleure partition binaire nette associée à un nœud dur, revient à déterminer l’unique paramètre a, alors que, selon la classe de sigmoïdes retenue, on gagne en flexibilité. Pour spécifier et simplifier ultérieurement notre problème, nous nous limiterons dans cet article à chercher $p(z)$ dans la famille logistique à deux paramètres, définie par :

avec :

$$p(z;a,b) = g(\frac{z-a}{b})$$

$$g(z) = \frac{1}{1+e^{-z}} \ .$$

La Figure 4 illustre la flexibilité de la famille logistique. Nous nommerons les paramètres a et b, *coupure* et *tendresse* respectivement (le dernier terme traduit ‘softness’ en analogie avec le vocable ‘bois tendre’ qui traduit ‘soft wood’). Nous utiliserons aussi occasionnellement le terme *nœud dur* (en analogie avec ‘bois dur’ correspondant à l’anglais ‘hard wood’) au lieu de nœud classique (non-probabiliste). Évidemment, un nœud dur a une tendresse zéro.

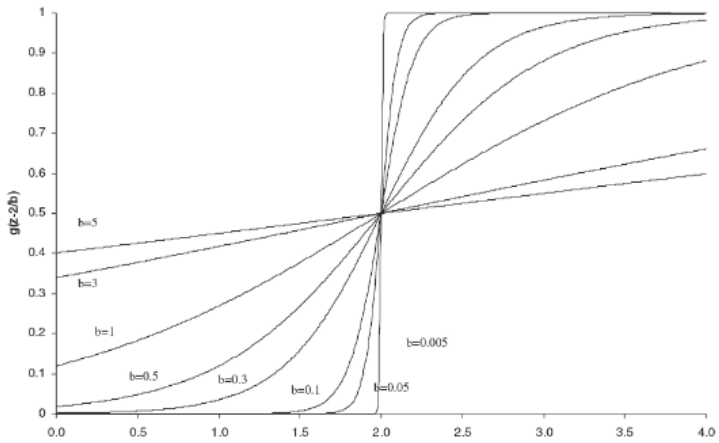

Figure 3: La fonction logistique pour valeurs variées des paramètres 'coupure' (a) et 'tendresse' (b)

Si l'objet de notre prédiction est une variable binaire y (deux classes), nous pouvons calculer le gain d'information *empirique* associé au nœud N en comparant deux modèles de prédiction : le modèle trivial, $p = P[y = 1] = 1$, et celui associé à l'équation :

$$p(z) = p_{gauche} g(\frac{z-a}{b}) + p_{droite}(1 - g(\frac{z-a}{b})) \qquad (1).$$

Alors, le gain d'information associé à N est estimé par la Statistique du Rapport de Vraisemblance (SRV) de la comparaison (gain d'information empirique). Il est intéressant de remarquer que le second modèle représente la distribution de y comme un mélange de distributions de Bernoulli.

4. Arbres à nœuds probabilistes

Un nœud probabiliste (tendre) permet d'utiliser toutes les données pour estimer les paramètres du modèle de prédiction. Par contre, l'estimation des paramètres π_{gauche} et π_{droite} associés à un nœud classique (dur) est strictement *locale*. Par conséquent, dans la construction d'un arbre dur on estime les paramètres sur des sous-échantillons de plus en plus petits, ce qui est une des causes de la variance excessive et de l'instabilité typique des arbres classiques.

Un arbre *tendre*, c'est-à-dire un arbre avec des nœuds probabilistes, est obtenu par un algorithme qui permet de faire intervenir toutes les données dans l'estimation de tous les paramètres. Ceci augmente la capacité d'emprunter de la puissance des données voisines ('borrowing strength'), un aspect très prisé dans la modélisation statistique, qui est à la base du grand succès des modèles de régression.

Rappelons d'abord que le modèle de prédiction d'une variable binaire y associé à un arbre dur est de la forme :

$$p(z) = p_1 I_1(z) + p_2 I_2(z) + ... + p_L I_L(z) \qquad (2)$$

où I_k, $k = 1,2,\ldots L$, représente la fonction indicatrice de la k-ème feuille de l'arbre, et p_k la probabilité que y soit égale à 1 pour une observation appartenant à la k-ème feuille. Les I sont, évidemment, des produits de fonctions indicatrices correspondant aux questions associées aux nœuds de l'arbre.

L'idée-clef de ce travail est de définir un arbre de prédiction tendre en substituant les I's de l'équation (2) avec des produits de sigmoïdes correspondant à des nœuds tendres, dénotés J :

$$p(z) = p_1 J_1(z) + p_2 J_2(z) + ... + p_L J_L(z) \qquad (3).$$

Une fois défini le modèle correspondant à une structure d'arbre tendre T, il est simple de lui associer un *gain d'information* IC(T), défini comme le rapport de vraisemblance qui compare ce modèle au modèle trivial ($p = P[y = 1]$ = constante). En pratique, ce gain est estimé par la SRV correspondante.

Plutôt que de développer un formalisme général, nous nous limiterons ici à donner un exemple illustrant la formule (3). Dans la Figure 5, nous représentons un arbre avec un nœud dur et un nœud tendre. SBP est une abréviation de tension artérielle systolique (Systolic Blood Pressure en Anglais). En effet un nœud défini par une variable de prédiction binaire ne peut être que dur, à cause d'un problème d'identification (voir [9]).

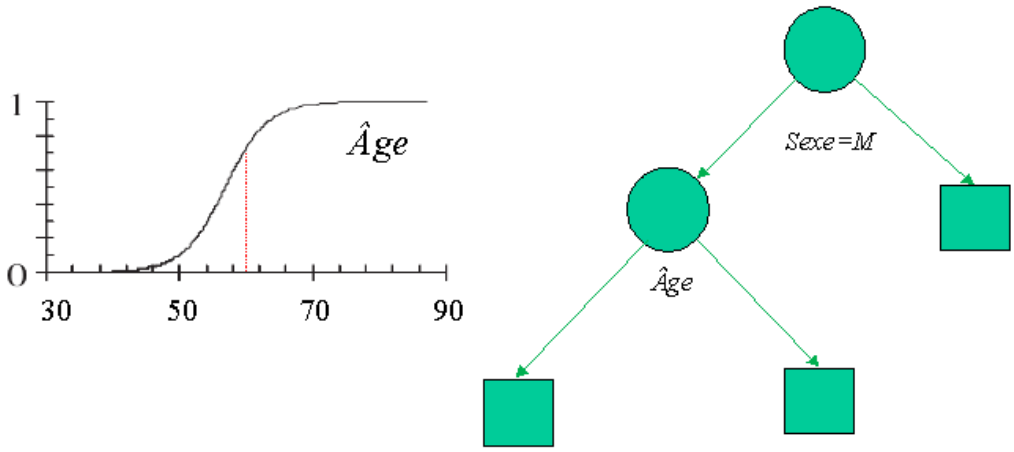

Figure 4: Un arbre tendre

Or, le modèle de prédiction associé à cet arbre est :

$$p(Sex, Age) = p_1 I[Sex = M]g(Age \geq 65] + p_2 I[Sex = M](1 - g(Age \geq 65]) + p_3 I[Sex = F] \qquad (4).$$

Un simple algorithme récursif de construction d'arbre peut être développé selon la règle suivante :

Élargir l'arbre courant en y ajoutant une branche de telle sorte que l'incrément du gain d'information empirique soit maximal.

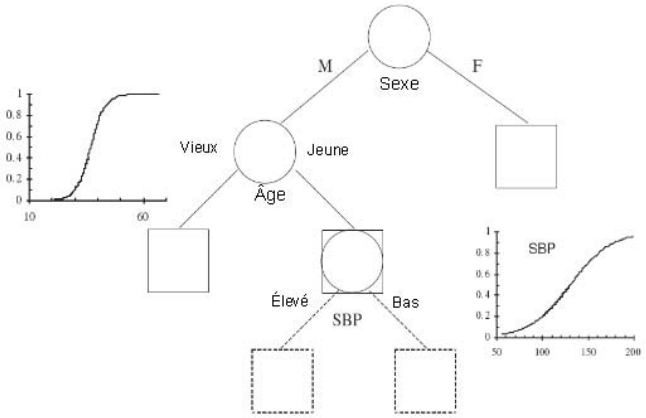

Figure 5 : Construction d'un arbre tendre à partir des données.

L'algorithme décrit dans [8] est bâti, à quelques détails près, autour de la règle ci-dessus, et en arrêtant la construction là où un critère approprié, à savoir le AIC ou le BIC [10], est optimal.

Afin de calculer le gain d'information et le critère à chaque pas de l'algorithme, il faudra maximiser le logarithme de la vraisemblance de l'arbre courant et de l'arbre élargi.

Puisque le modèle de base est celui d'un mélange de distributions, nous avons travaillé avec un algorithme de type EM. Cet algorithme ne maximise pas directement la vraisemblance du modèle de mélange, mais maximise itérativement la vraisemblance de données dites *complètes*, c'est-à-dire de données hypothétiques semblable aux données observées, mais augmentées par l'indicateur de classe (distribution) *non-observé* de chaque observation. Puisque cet indicateur de classe n'est pas connu, il faudra, à chaque itération, le remplacer avec par une estimation, basée, elle, sur ce qu'on a effectivement observé. En effet, l'algorithme alterne, jusqu'à convergence, entre une étape E (Espérance), et une étape M (Maximisation). Dans l'étape E, on remplace les indicateurs de classe par leurs espérances, et dans l'étape M on maximise la vraisemblance des données complètes telle que mise à jour dans l'étape E précédente. Sans donner les détails (voir [9]), nous écrivons ici, à titre d'exemple, le logarithme de la vraisemblance du modèle de prédiction représenté par l'arbre de la Figure 6 (avant l'ajout de la branche en pointillé). Puisque cet arbre a un seul nœud tendre, il y a un seul indicateur de classe, dénoté par ζ_i. On a donc :

$$
\begin{aligned}
l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{ & y_i \zeta_i I[\mathrm{sex}_i = \mathrm{M}] \log(p_1) + (1 - y_i)\zeta_i I[\mathrm{sex}_i = \mathrm{M}] \log(1 - p_1) \\
& + y_i(1 - \zeta_i) I[\mathrm{sex}_i = \mathrm{M}] \log(p_2) + (1 - y_i)(1 - \zeta_i) I[\mathrm{sex}_i = \mathrm{M}] \log(1 - p_2) \\
& + y_i I^{\mathrm{c}}[\mathrm{sex}_i = \mathrm{M}] \log(p_3) + (1 - y_i) I^{\mathrm{c}}[\mathrm{sex}_i = \mathrm{M}] \log(1 - p_3) \} \\
& + \sum_{i=1}^{n} \{ \zeta_i \log(g((\mathrm{age}_i - a)/b)) + (1 - \zeta_i) \log(g^{\mathrm{c}}((\mathrm{age}_i - a)/b)) \}
\end{aligned}
$$

.

5. Exemples d'analyses

Le premier exemple concerne le développement d'une règle de prédiction pour le diabète de type II, basée sur des facteurs de risque connus. Nous présentons ici un arbre dur et un arbre probabiliste obtenus à partir d'une base de données publique. Il s'agit de données recueillies sur 532 femmes appartenant à la culture amérindienne Pima et vivant dans la région de Phoenix, Arizona. Le but de la récolte de ces données était de comprendre les liens entre plusieurs facteurs de risque et la présence de diabète de type II. Ici nous nous intéressons, plus simplement, au développement d'une règle efficace de prédiction par arbre. Les variables de prédiction, avec leur abréviation anglaise correspondante, sont les suivantes :

- *Glucose (Glu)*
- *Tension artérielle diastolique (bp)*
- *Triceps skinfold thickness (skin)*
- *Body Mass Index (BMI)*

– *Diabetes Pedigree Function (Ped)*

– *Age*.

La figure suivante montre l'arbre dur (7 a)) et l'arbre tendre (7 b)) construit à partir de ces données.

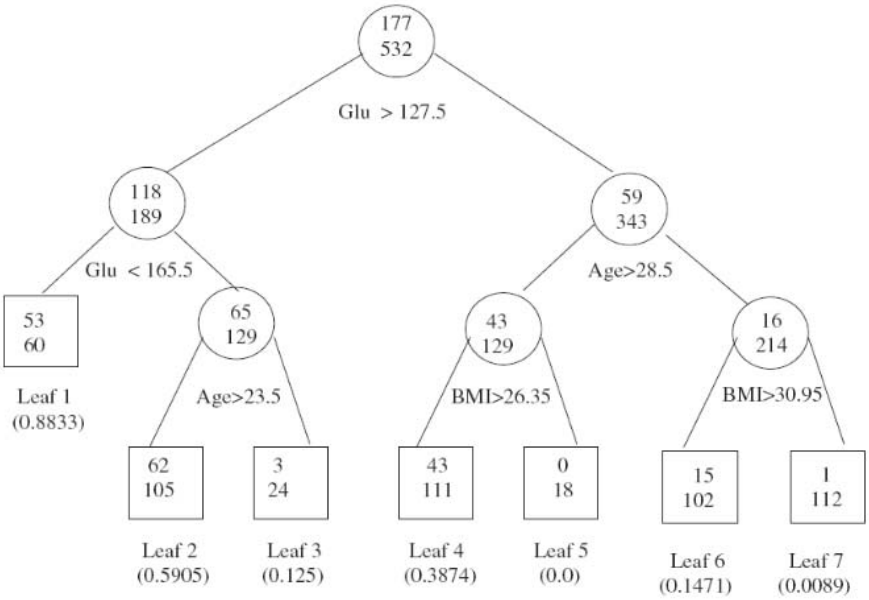

Figure 7a): Données PIMA: arbre dur

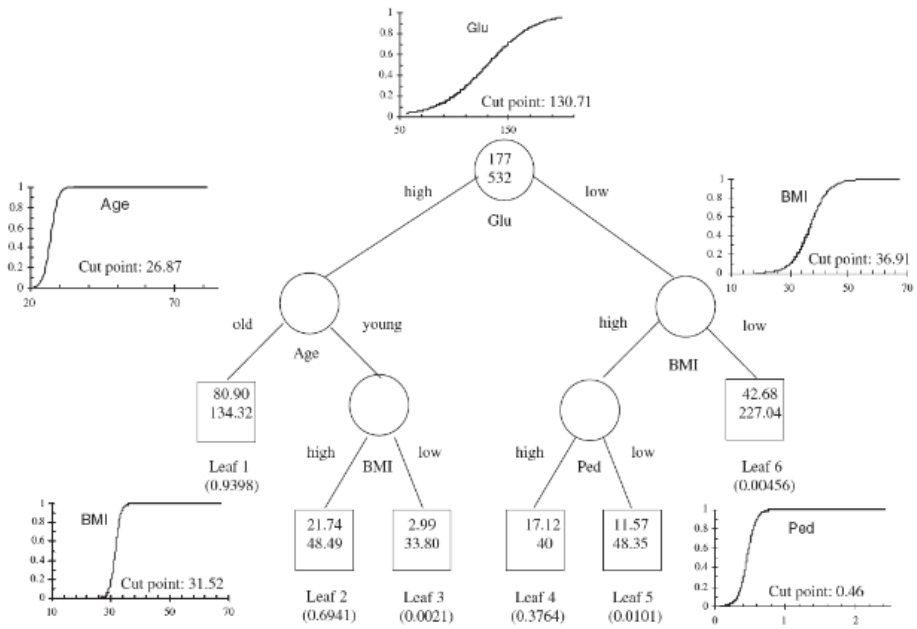

Figure 6b) : Données PIMA: arbre tendre

Les prédictions des deux modèles sont comparables (erreur de prédiction : 19.1% pour l'arbre tendre et 21.0% pour l'arbre dur) mais légèrement favorables à l'arbre tendre. La règle de prédiction associée à l'arbre tendre est, dans ce cas, plus simple que celle associée à l'arbre dur : si, d'une part, l'arbre dur crée des nœuds basés sur moins de variables, ces variables apparaissent plusieurs fois, suggérant que des coupures dures ne sont peut-être pas très appropriées. Par contre, l'arbre tendre contient plus de variables mais moins de répétitions.

Comme deuxième exemple, nous avons choisi un problème de classification avec quatre classes. La variable dépendante est donc une variable discrète à quatre niveaux, et non pas une variable binaire. Des modifications relativement simples nous permettent de généraliser la méthode de construction d'arbre à cette situation. Le premier pas est d'écrire les modèles de base sous la forme de l'équation (2), mais avec des paramètres vectoriels remplaçant les paramètres scalaires dans cette équation. Les données CRABES dont l'analyse est présentée ici sont des mesures de caractéristiques morphologiques de 200 crabes appartenant à quatre classes : Bleu Mâle (BM), Bleu Femelle (BF), Orange Male (OM) et Orange Femelle (OF). Les caractéristiques morphologiques sont (les abréviations en parenthèse sont dérivées du terme anglais correspondant) : taille du lobe frontal (FL), largeur postérieure (RW), longueur de la carapace (CL), largeur de la carapace (CW), et épaisseur du corps (BD). L'arbre tendre obtenu avec notre algorithme est représenté dans la Figure 8.

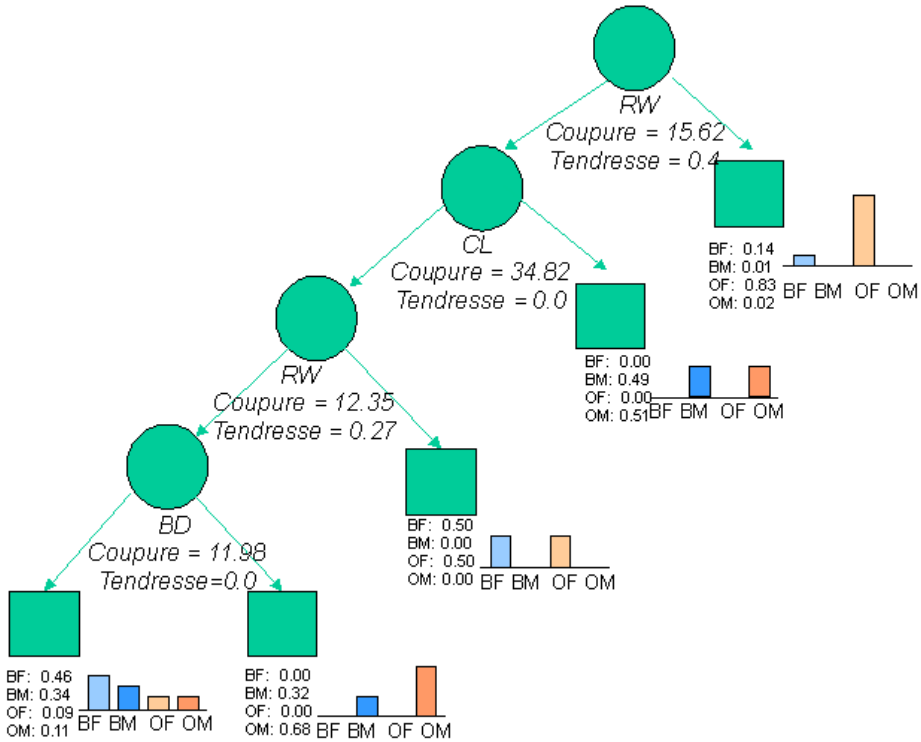

Figure 8 : Données CRABES : arbre tendre.

Dans les grandes lignes, l'interprétation fournie par cet arbre de classification à nœuds probabilistes est assez claire et semblable à celle qu'on pourrait donner à un arbre hypothétique de la même topologie mais ayant des coupures dures : a) la taille postérieure sépare les femelles orange du reste; b) parmi les crabes de taille postérieure petite, les mâles ont une carapace plus longue que celle des femelles; c) il y a une tendance pour les crabes bleus à être plus petits que les crabes oranges; d) en général, il est plus difficile de discriminer entre les sexes qu'entre les couleurs.

D'autre part, la règle de classification associée à cet arbre est plus complexe que celle associée à l'arbre dur correspondant. Par exemple, considérons le crabe #134; ses mesures sont : FL = 18.4, RW = 13.4, CL = 37.9, CW = 42.2, BD = 17.7. Au nœud racine, le crabe #134 (un mâle orange) est envoyé à gauche avec une probabilité 0.996; le deuxième nœud étant dur, ce crabe est envoyé à droite avec une probabilité de 1, c'est-à-dire avec une certitude pratiquement parfaite; il est donc classé, si l'on adopte la règle de la majorité, comme mâle orange.

6. Évaluation

Afin d'évaluer la méthode de construction d'arbres à nœuds probabilistes, nous avons adopté l'approche de Breiman [11]. Nous résumons ici les résultats pour une variable dépendante binaire; les détails sont discutés dans [9]. Nous avons sélectionné six bases de données de domaine public (Table 2), et pour chaque base de données nous avons répété les étapes suivantes 100 fois :

1. Sélectionner 10% des données au hasard et les garder à part comme échantillon test (ET); utiliser les 90% restant comme échantillon d'apprentissage (EA).
2. Utiliser l'EA pour construire un arbre dur ainsi qu'un arbre probabiliste.
3. Sur les ET, évaluer la performance de chaque arbre en calculant: l'erreur de classification, l'aire sous la courbe ROC, le score de Brier, la déviance, et le nombre de feuilles.

Un sommaire des résultats est contenu dans les Tables 3-6.

	Taille de l'échantillon originale	**Taille de l'échantillon utilisable**	**Taille de l'Échantillon Test**	**Nombre de variables**
Cancer du sein	683	683	68	9 (C = 9; D = 0)*
PIMA	532	532	50	7 (C = 7; D = 0)
Maladie cardiovasculaire	303	296	30	8 (C = 5; D = 3)
Maladie du foie	345	345	35	5 (C = 5; D = 0)
Diabète 2	403	367	37	8 (C = 6; D = 2)
Cancer de la prostate	502	482	48	13 (C = 10; D = 3)

Table 2: Caractéristiques de 6 bases de données utilisées pour l'évaluation

**C = variable continue, D =variable discrète.*

	Arbre tendre				Arbre dur				p-value*	$E_H > E_S$	$E_H >= E_S$
	Mean	Std	Min	Max	Mean	Std	Min	Max			
Breast cancer	3.998	2.499	0	10.29	5.31	2.715	0	11.76	<0.0001	58%	84%
Pima Indian	22.86	5.58	12	34	26.12	5.797	10	38	<0.0001	69%	77%
Heart disease	25.37	7.429	3.33	43.33	33.73	7.803	10	56.67	<0.0001	82%	86%
Liver disease	37.74	7.556	20	54.29	50.63	8.634	22.86	71.43	<0.0001	86%	93%
Diabetes 2	15.68	5.34	2.7	27.03	14.62	5.47	2.7	27.03	0.0007	3%	78%
Prostate cancer	36.92	6.92	18.75	56.25	39.19	6.65	18.75	60.42	0.0013	54%	66%

Table 3: Résultats de l'évaluation: Erreur de classification moyenne sur les Échantillons Test

** t-test apparié pour la différence en erreur de classification.*

	Arbre tendre				Arbre dur			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Breast cancer	99.13	0.82	94.74	100	97.23	1.96	91.76	100
Pima Indian	83.75	5.42	69.71	95.93	79.48	5.85	61.9	93.21
Heart disease	81.34	7.45	61.38	99.04	69.39	9.32	47.69	91.63
Liver disease	66.29	9.33	44.77	86.84	51.59	5.85	30.72	75.95
Diabetes 2	74.54	10.03	50	100	64.71	11.38	33.33	81.82
Prostate cancer	62.65	7.72	40.91	78.15	54.55	6.42	40.3	71.3

Table 4 : Résultats de l'évaluation: moyenne de l'indice c (aire sous la courbe ROC en %) sur les Échantillons Test

	Arbre tendre				Arbre dur				# infinie*
	Mean	Std	Min	Max	Mean	Std	Min	Max	
Breast cancer	15.87	7.94	4.12	48.8	20.16	11.02	2.57	58.62	41
Pima Indian	48.03	9.11	32.63	71.08	50.22	8.05	29.66	68.66	12
Heart disease	33.08	7.93	17.39	64.24	38.25	7.01	21.63	55.13	7
Liver disease	45.52	4.09	36.33	58.84	49.33	2.98	40.61	63.85	1
Diabetes 2	28.51	8.08	12.17	57.25	29.56	7.83	14.42	56.16	5
Prostate cancer	66.01	7.78	49.72	85.73	64.16	4.38	55.63	80.36	1

Table 5: Résultats de l'évaluation: Déviance moyenne sur les Échantillons Test

	Arbre tendre				Arbre dur			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Breast cancer	5.01	0.86	3	8	8.76	3.66	4	17
PIMA Indian	6.92	1.19	5	10	5.12	2.05	2	14
Heart disease	9.13	2.11	6	15	3.98	1.74	2	12
Liver disease	3.15	0.41	3	5	1.64	0.95	1	6
Diabetes 2	6	1.29	3	9	2.16	0.84	1	7
Prostate cancer	10.88	3.53	4	18	0.8	0.9	1	6

Table 6: Résultats de l'évaluation: Numéro moyen de feuilles

Nous notons que :

a) Pour toutes les bases de données, à l'exception de Diabète 2, l'arbre à nœuds probabilistes a une plus petite erreur de classification que l'arbre dur. L'amélioration moyenne de cette erreur varie entre 4% et 13%.

b) La performance de l'arbre à nœuds probabilistes est partout supérieure du point de vue de la déviance (il a une plus petite valeur de la déviance), et de l'index c (il a une plus grande valeur de c).

c) Les résultats pour le score de Brier, qui ne sont pas présentés ici, montrent aussi que la performance de l'arbre à nœuds probabilistes est meilleure que celle de l'arbre dur.

7. Discussion

Nous avons présenté ici les concepts fondamentaux de l'arbre à nœuds probabilistes, ou arbre tendre, ainsi qu'un algorithme qui permet de construire un tel arbre à partir de données. L'idée clef est celle du nœud probabiliste. Un nœud probabiliste peut, d'une part, être conçu comme un outil artificiel pour obtenir une meilleure décision. D'autre part, dans des circonstances particulières, un nœud probabiliste peut être la représentation d'une variable binaire latente, par exemple la présence d'une caractéristique au niveau génétique qui n'est pas observée, voire pas observable. Dans ce dernier cas, la probabilité 'd'aller à gauche' sachant une variable continue, n'est que la probabilité conditionnelle qu'une observation appartienne à une classe latente d'intérêt particulier (par exemple une variante pathogénique d'un gène).

Les résultats empiriques présentés ici indiquent que l'arbre à nœuds probabilistes peut fournir un outil de prédiction plus performant que l'arbre dur, toute en offrant une clef interprétative intéressante. À la différence d'un arbre dur, l'arbre probabiliste définit des classes de façon plus qualitative que quantitative; par exemple, une feuille d'un arbre à nœuds probabilistes peut caractériser un groupe à risque élevé de développer un cancer, par des 'valeurs très élevées' de certains variables et par des

'valeurs très basses' de certaines autres. En général, chaque feuille est un 'cas limite', et chaque observation est décrite comme un mélange de ces cas limites. En même temps, l'arbre à nœuds probabilistes offre une règle pour faire des prédictions individuelles et, comme nos résultats l'indiquent, ces prédictions semblent être de bonne qualité. Cette dualité entre, d'une part, une interprétation globale moins nette et, de l'autre, une prédiction individuelle plus complexe, peut, de prime abord, paraitre moins intéressante que la simplicité d'un arbre dur. Ce dernier offre un modèle qui est à la fois général et individuel; ce modèle spécifie très exactement le sens de l'expression 'valeur élevée' en termes d'un seuil (point de coupure), et attribue chaque observation à une feuille unique. Nous notons aussi que les arbres tendres ont tendance à avoir plus de feuilles et moins de répétitions d'une même variable que les arbres durs correspondant. Ceci est probablement une conséquence directe du gain de flexibilité représenté par les nœuds tendres, et du fait que chaque observation contribue, bien qu'en proportions différentes, à toutes les étapes de la construction de l'arbre tendre.

Pour que l'arbre à nœuds probabilistes puisse être adopté dans la pratique courante de l'analyse de données, il faudra montrer que la perte de simplicité d'interprétation est justifiée par le gain en réalisme et en qualité de la prédiction. Il faudra aussi s'assurer que l'algorithme, moyennant de modifications dont l'ampleur n'est pas encore connue, puisse opérer efficacement sur des bases de données plus grandes que celles sur lequel il a été testé. Nous essayons dans nos recherches courantes d'étudier ces problématiques. Plus spécifiquement, nous sommes en train d'évaluer l'algorithme par simulations et par analyses de données réelles, traitant des variables dépendantes catégorielles, et quantitatives. Nous étudions aussi des alternatives plus efficaces à la cette forme actuel de l'algorithme EM.

Pour des développements futurs, nous considérons la généralisation de l'algorithme au traitement de variables dépendantes plus complexes, tels que temps de survie censurés et variables multidimensionnelles qualitatives et quantitatives.

Pour conclure, nous remarquons que notre itinéraire de recherche résumé ici nous a amené successivement de la construction d'arbres classiques à partir de données classiques, à la construction d'arbres classiques à partir de données symboliques; et de là, à la notion et à la construction d'arbres à nœuds probabilistes à partir de données classiques. Il serait intéressant d'avancer vers le développement d'arbres à nœuds probabilistes pour données symboliques.

Références

[1] L. Billard, E. Diday (2006). Symbolic Data Analysis: conceptual statistics and data mining. New York, NY: Wiley

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

[3] L. Breiman. (1997): 'Bagging predictors'. *Machine Learning*, 26:123–140.

[4] L. Breiman (2001): 'Random Forests'. *Machine Learning,* 1:5-32

[5] K. P. Burnham, and D. R. Anderson. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach (2nd Edition)*. New York, NY: Springer-Verlag.

[6] A. Ciampi (1991): 'Generalized Regression Trees'. *Computational Statistics & Data Analysis*, 12:57-78.

[7] A. Ciampi, E. Diday, J. Lebbe, E. Perinel, R. Vignes. (2000): 'Growing a tree classifier with imprecise data'. *Pattern Recognition Letters*, 21:787-803.

[8] A. Ciampi, A. Couturier, Shaolin Li. (2002): 'Prediction trees with soft noeuds for binary outcomes'. *Statistics in Medicine*, 21:1145–1165.

[9] S. Kullback (1968). Information Theory and Statistics. New York, NY: Dover

[10] J. R. Quinlan (1990). Probabilistic decision trees. In: Kodratoff, Y., Michalski, R. (Eds.), *Machine Learning III*, pp. 140-152.

[11] J. R. Quinlan. (1993): *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.