Arbres de classification construits à partir de fonctions de croyance

Nicolas Sutton-Charani Sébastien Destercke Thierry Denœux

UMR CNRS 6599 Heudiasyc Université Technologique de Compiègne

BP 20529 - F-60205 Compiègne cedex - France,
nicolas.sutton-charani@hds.utc.fr
sebastien.destercke@hds.utc.fr
thierry.denoeux@hds.utc.fr

Résumé :

Les arbres de décision sont des classifieurs très populaires. Dans cet article, nous étendons une méthode de construction d'arbres de décision à deux classes, basés sur les fonctions de croyance, au cas multi-classe. Nous proposons pour cela trois extensions possibles : combiner des arbres à deux classes ou directement étendre l'estimation des fonctions de croyance au sein de l'arbre au cadre multi-classe. Des expériences sont effectuées de manière à comparer ces arbres aux arbres de décision classiques

Mots-clés :

arbres de décision, fonctions de croyance, classification

Abstract:

Decision trees are popular classification methods. In this paper, we extend to multi-class problems a decision tree method based on belief functions previously described for two-class problems only. We propose three possible extensions : combining multiple two-class trees together and directly extending the estimation of belief functions within the tree to the multi-class setting. We provide experiments, results and compare them to usual decision trees.

Keywords:

decision trees, belief function, classification

1 Introduction

La popularité des arbres de décision [2] (de classification pour les sorties catégoriques ou de régression pour les sorties continues) tient à leur interprétabilité, la simplicité de leur mise en oeuvre et leur pouvoir prédictif. La construction d'arbres de décision classiques repose sur la théorie des probabilités. Cependant, cette théorie ne permet pas de prendre en compte certains aspects comme le manque de données (petit échantillon) ou les données à valeurs incertaines. Dans ce travail, nous nous intéressons principalement au premier problème (manque de données).

La théorie des fonctions de croyance [13] offre un cadre de travail adéquate pour gérer ces types de problème. En effet, Elouedi *et al.* [8] proposent différents moyens d'adapter les arbres de décision au TBM (*Transferable Belief Model*) de manière à gérer les sorties incertaines durant la construction des arbres. Denoeux et Skarstein-Bjanger [7] proposent une autre extension ayant des liens plus étroits avec les statistiques. Elle se rapproche en cela d'approches probabilistes imprécises, notamment celle proposée par Abellan [1].

La méthode de Denoeux-Skarstein-Bjanger (DSB) ne concernant que les problèmes à deux classes, nous l'étendons à un nombre quelconque de classes par trois moyens :

- combiner les fonctions de croyance issues d'arbres à deux classes [12]
- construire des fonctions de croyance multinomiales en utilisant le modèle de Dirichlet Imprécis (IDM) [15]
- construire des fonctions de croyance multinomiales prédictives en utilisant l'approche de Denoeux [5]

La section 2 présente les notions d'arbres de décision nécessaires ainsi que la méthode de Denoeux-Skarstein-Bjanger (DSB). La section 3 étend cette méthodologie au cas multiclasse. Finalement, dans la section 4 nous com-

parons ces nouveaux classifieurs à ceux utilisant l'algorithme CART classique et discutons des effets des paramètres sur les résultats expérimentaux.

2 Pré-requis

Nous résumons les éléments de base de la théorie des fonctions de croyance nécessaire, avant de rappeler rapidement le principe des arbres de décision, d'abord dans le cadre classique probabiliste puis dans le cadre des fonctions de croyance avec la méthodologie de Denoeux-Skarstein-Bjanger (DSB).

2.1 Fonctions de croyance

Soit Ω un ensemble fini, appelé le *cadre de discernement*. Une fonction de masse de croyance est une fonction $m : 2^\Omega \to [0;1]$ telle que $m(\emptyset) = 0$ et $\sum_{A\subseteq\Omega} m(A) = 1$.

Tout sous-ensemble A de 2^Ω tel que $m(A) > 0$ est appelé ensemble focal.

Si m n'a qu'un ensemble focal, elle est dite *catégorique*, et si ses seuls ensembles focaux sont des singletons elle est dite *Bayésienne* (et est équivalente à une distribution de probabilité classique). Il est intéressant de noter que $m(\Omega)$ représente le degré d'*ignorance* de m. A partir de m, on définit

$$Bel(A) = \sum_{B\subseteq A} m(B); \quad Pl(A) = \sum_{B\cap A\neq\emptyset} m(B)$$

appellées respectivement fonction de croyance et fonction de plausibilité. Les représentations m, Bel et Pl sont équivalentes, càd il existe des transformations bijectives pour passer de l'une à l'autre. La transformée pignistique permet de transformer une fonction de masse m en une mesure de probabilité $BetP$ appelée *probabilité pignistique* [14] et telle que :

$$BetP(\{w\}) = \sum_{A\subseteq\Omega : w\in A} \frac{m(A)}{|A|}$$

2.2 Arbres de décision

Soit (X,Y) un vecteur aléatoire où $X = (X_1,...,X_J) \in \mathscr{X} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_J$ représente les attributs et $Y \in \mathscr{Y} = \{Y_1,\ldots,Y_K\}$ la classe à prédire. A partir d'un échantillon $E = \{(X^1,Y^1),\ldots,(X^n,Y^n)\}$, les arbres de décision sont des méthodes itératives de construction d'un modèle sur (X,Y) correspondant à un partitionnement de $\mathscr{X}$. Nous considérons ici des arbres binaires (i.e., modèles type CART), où chaque scission d'un noeud parent donne naissance à deux noeuds enfants.

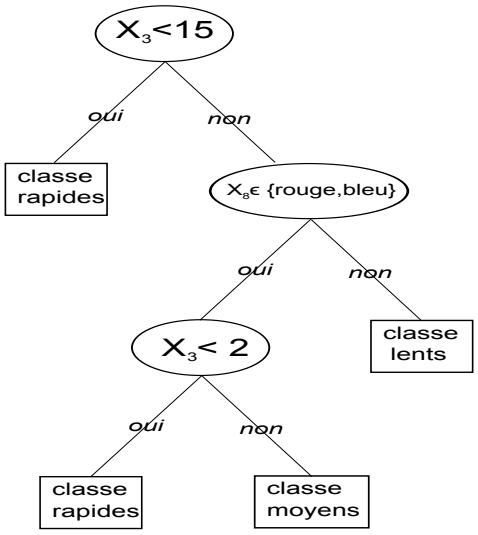

Figure 1 – Arbre de décision type CART

La méthode est la suivante : à partir d'un noeud initial contenant tout l'échantillon d'apprentissage, la scission optimale (au sein de toutes les variables et de leur valeurs) en terme de gain d'information est recherchée. Le gain d'information IG correspondant à une scission sur la variable X_k (ici continue) avec un seuil α (dans le cas discret, les seuils sont remplacés par des partitions sur X_k) est calculé comme suit :

$$IG(k,\alpha) = i(t_0) - p_L i(t_1) - p_R i(t_2) \quad (1)$$

où $i(t)$ est une mesure d'impureté d'un noeud t, t_0 le noeud initial, t_1 et t_2 ses noeuds enfant,

p_L est la proportion des individus de t_0 vérifiant $X_k < \alpha$ (i.e., $p_L = {}^{n_L}/_{n}$ où n correspond à la taille de l'échantillon dans t_0 et n_L le nombre d'individus vérifiant $X_k < \alpha$). $p_R = 1 - p_L$ est la proportion de l'échantillon ne vérifiant pas cette condition. La scission sélectionnée (k, α) est alors celle qui maximise IG (calculé comme un gain de pureté).

La méthode est ensuite appliquée récursivement à chaque noeud enfant jusqu'à ce qu'aucun gain d'information supérieur à un seuil pré-établi ne puisse être atteint. Dans ce dernier cas, le noeud est alors considéré comme une feuille prédisant la classe la plus fréquente dans la partie de l'échantillon y étant assignée. La figure 1 donne un exemple d'arbre, avec $X_3 < 15$ la première scission trouvée. Différentes mesures d'impureté existent. Dans les arbres types CART ou C4.5 [11], on utilise le plus souvent l'indice de Gini et l'entropie de Shannon, cette dernière s'écrivant

$$i_{C4.5}(t) = \sum_{k=1}^{K} p_k(t) \log(p_k(t))$$

où $p_k(t)$ est la fréquence de la classe k dans le noeud t.

Ces fonctions mesure l'homogénéité d'un noeud en terme de classes, mais ne dépendent pas de la taille des échantillons (uniquement de leur fréquence). Par contre, la méthode et la mesure d'impureté de la méthode DSB tiennent compte de la taille des échantillons.

2.3 Méthode DSB pour des échantillons à deux classes

Cette méthode utilise les mêmes principes que CART, mais diffère au niveau du calcul du gain d'information : premièrement elle prend en argument des fonctions de masse au lieu de simples fréquences et deuxièmement elle utilise une mesure d'impureté combinant non-spécificité (imprécision) et conflit (variabilité).

Les fonctions de masses sont construites à partir de l'approche de Dempster appliquée au cas binomial ($aDeB$) :

$$\begin{cases} m_{aDeB}(\{Y_1\}) = \frac{n_1}{n+1} \\ m_{aDeB}(\{Y_2\}) = \frac{n_2}{n+1} \\ m_{aDeB}(\mathscr{Y}) = \frac{1}{n+1}, \end{cases} \quad (2)$$

où n correspond à la taille totale de l'échantillon et n_1, n_2 au nombre d'individus de classes respectives Y_1, Y_2. Denœux et Skarstein-Bjanger proposent alors d'utiliser la mesure d'impureté suivante [10], appliquée à m_{aDeB} :

$$U_\lambda(m) = (1-\lambda)N(m) + \lambda D(m) \quad (3)$$

où $N(m) = \sum_{A \subseteq \mathscr{Y}} m(A) \log_2 |A|$ mesure la non-spécificité et $D(m) = -\sum_{A \subseteq \mathscr{Y}} m(A) \log_2 BetP(A)$ la variabilité.

Les deux parties sont pondérées par un hyper-paramètre $\lambda \in [0,1]$. On peut remarquer que $m(\mathscr{Y})$ (l'imprécision) diminue quand n augmente. En utilisant U_λ comme mesure d'impureté $i(t)$, le gain d'information (1) peut être négatif. Cela constitue un critère d'arrêt naturel pendant la construction de l'arbre, aucune scission n'est ainsi effectuée quand tous les gains possibles d'information sont négatifs. Klir propose de fixer λ à 0.5, cependant en classification λ peut se déterminer par validation croisée.

Exemple : Soit t un noeud contenant 5 individus de classes 1 et 2 individus de classe 2 ($n = 7$), on a alors

$m(\{Y_1\}) = \frac{5}{8}$, $m(\{Y_2\}) = \frac{2}{8}$ et $m(\Omega) = \frac{1}{8}$.

On a donc : $BetP(\{Y_1\}) = \frac{5}{7}$,

$BetP(\{Y_2\}) = \frac{2}{7}$ et $BetP(\Omega) = 1$

On obtient alors :

– $N(m) = m(\Omega) = \frac{1}{8}$
– $D(m) = -[\frac{5}{8} * log_2(\frac{5}{7}) + \frac{2}{8} * log_2(\frac{2}{7})] \approx 0.61$

et donc $U_{0.5}(m) \approx 0.37$

Le tableau 1 présente les résultats obtenus

à l'aide d'arbres de classification CART et d'arbres de classification utilisant la méthode DSB sur des jeux de données à 2 classes. Le critère d'arrêt est le suivant : continuer à *scinder* tant que $IG > \beta$ pour les arbres CART classiques ($IG > 0$ pour ceux utilisant U_λ comme mesure d'impureté) et tant que les noeuds enfants issus de la scission contiennent au moins 10 individus.

Les arbres CART classiques sont optimisés sur le seuil β alors que ceux basés sur U_λ le sont sur le paramètre λ, par une validation croisée à 10 couches sur l'échantillon d'apprentissage. Les résultats montrent que les deux méthodes obtiennent des résultats comparables en terme de bonne prédiction.

Tableau 1 – Efficacité des arbres par rapport à la mesure d'impureté utilisée

Jeu de données	#var	%err CART	%err U_λ
Blood transfusion	4	23.5%	24.2%
Statlog heart	13	28%	25.7%
Tic-tac	9	21.5%	11.5%
Breast-cancer	10	5.9%	4.7%
Pima	8	27.3%	25.1%
Haberman	3	26.6%	26%

L'approche de Dempster est peu pratique lors du passage à plusieurs classes (cas multinomial), pour lequel il n'existe pas d'expression simple comme (2). Nous proposons donc trois moyens plus abordables de gérer de tels cas : *éclater* un problème de classification à K classes (K $\geq$ 3) en C_k^2 problèmes à 2 classes en utilisant la méthode de combinaison de classifieurs binaires de Quost [12], utiliser l'approche IDM ou le modèle multinomial de Denœux.

3 Cas multi-classes

3.1 Combinaisons de classifieurs binaires

Dans [12], Quost présente une méthode permettant de résoudre un problème de classification multi-classes en combinant des classifieurs entraînés sur des sous-échantillons à deux classes. Il propose d'apprendre (à partir du sous-échantillon correspondant) une fonction de croyance conditionnelle pour chaque paire $\{Y_i, Y_j\}$, $1 \leq i < j \leq K$ de classes et de les combiner en une fonction de croyance globale sur $\mathscr{Y}$ en utilisant une procédure d'optimisation.

Nous proposons d'utiliser cette méthode en utilisant les arbres obtenus par méthode DSB comme classifieurs binaires pour apprendre les fonctions de croyance conditionnelles. A notre connaissance, c'est la première fois que les arbres sont utilisés comme classifieurs de bases dans cette approche.

Cette méthode est différente de celle présentée par Vannoorenberghe ([16]) où K arbres à 2 classes sont construits considérant pour chacun d'eux "une classe contre les autres" puis dont les fonctions de croyance de sortie sont combinées en faisant la moyenne des K masses étendues (de $\{Y_k; \overline{Y_k}\}$ dans $\mathscr{Y}$).

Les arbres de décision sont bien adaptés à ce genre de combinaison au vu de leur simplicité. Cependant, on peut remarquer que l'optimisation sur le λ devient alors problématique, étant donné que $K(K-1)/2$ classifieurs doivent être appris à chaque étape de l'optimisation.

3.2 IDM

L'IDM fut introduit dans le cadre des "probabilités imprécises" par Walley [17]. Le modèle obtenu par l'IDM est une fonction de croyance particulière, et nous pouvons donc l'utiliser dans notre approche. L'imprécision de l'IDM est contrôlée par un hyper-paramètre $s \in \mathbb{R}^+$. A partir d'un échantillon aléatoire $Y^1, ..., Y^n$, Walley montre que les bornes probabilistes obtenues pour la classe Y_k à partir de l'IDM sont $P_k^- = {}^{n_k}/_{n+s}$ et $P_k^+ = {}^{n_k}/_{n+s}$ où n_k est le nombre d'occurrences de Y_k. La fonction de masse correspondante est telle que :

$$\begin{cases} m_{IDM}(\{Y_j\}) = n_j/(n+s) & j = 1,...,K \\ m_{IDM}(\mathscr{Y}) = s/(n+s) \end{cases} \tag{4}$$

On peut remarquer que l'on retrouve l'équation (2) pour K=2 et s=1. En utilisant m_{IDM}, U_λ peut alors calculer l'impureté d'un noeud au sein d'un arbre multiclasse. La forme analytique de U_λ appliquée à m_{IDM} est alors :

$$U_\lambda(m_{IDM}) = \frac{(1-\lambda)s}{n+s}\log_2(K) - \frac{\lambda}{n+s}\sum_{k=1}^{K} n_k \log_2\left[\frac{Kn_k+S}{K(n+s)}\right] \tag{5}$$

Notons que l'imprécision de l'IDM ne dépend que de la taille de l'échantillon n, et pas de sa distribution de probabilité sur $\mathscr{Y}$. Aussi, cette approche n'est pas directement basée sur la théorie des fonctions de croyance. Ceci n'est pas le cas pour le modèle multinomial de Denœux qui offre une alternative intéressante.

L'IDM étant le modèle utilisé par Abellan [1] (mais avec une measure d'impureté basée sur une approche probabiliste imprécises), il sera intéressant de comparer cette extension avec la méthode proposée par Abellan [1].

3.3 Modèle multinomiale de Denœux

Denœux [5] propose d'utiliser les intervalles de confiance de Goodman [9] pour construire une fonction de croyance prédictive. La première étape est la construction d'intervalles de probabilité [4] (probabilités inférieures et supérieures sur les singletons) puis de les transformer en fonctions de croyance.

A partir d'un échantillon $Y^1,...,Y^n$ avec $Y^k \in \mathscr{Y} = \{Y_1,...,Y_K\}$, ces intervalles de probabilité $[P_k^-, P_k^+]$ sont donnés, pour Y_k (k=1,...,n), par :

$$P_k^- = \frac{q+2n_k-\sqrt{\Delta_k}}{2(n+q)} \quad et \quad P_k^+ = \frac{q+2n_k+\sqrt{\Delta_k}}{2(n+q)} \tag{6}$$

où q est le quantile d'ordre $1-\alpha$ de la loi du chi-deux à un degré de liberté, et où $\Delta_k = q(q+\frac{4n_k(n-n_k)}{n})$. Comme démontré dans [5], la mesure de confiance inférieure (i.e., $P^-(A) = max(\sum_{Y_k \in A} P_k^-, 1 - \sum_{Y_k \notin A} P_k^-)$) obtenue par ces intervalles dans les cas $K = 2$ ou 3 est une fonction de croyance.

On remarque que les fonctions de croyance obtenues suivent le principe de Hacking (voir [5] pour plus de détails), mais que la solution pour $K = 2$ n'est pas équivalente à celle de Eq. (2).

Dans le cas $K > 3$, l'inverse de Möbius de P^- peut être négative et n'est donc pas une fonction de croyance en général. Différentes méthodes incluant de la programmation linéaire sont proposées dans [5] pour l'approximer par une fonction de croyance. En outre, dans le cas particulier où les classes sont ordinales, Denœux propose un algorithme restreint à un certain ensemble d'éléments focaux. Une fonction de croyance valide est ainsi obtenue et peut être utilisée par U_λ pour créer des arbres multiclasses.

4 Experiences

Nous commençons par comparer les résultats des différents classifieurs puis discutons de l'effet de l'hyper-paramètre λ sur U_λ.

4.1 Comparaisons entre les classifieurs

Nous comparons les trois extensions proposées avec l'algorithme CART classique. Le tableau 2 présente les caractéristiques de six jeux de données UCI multi-classes. Le tableau 3 présente les résultats expérimentaux sur les jeux de données précédents comparant l'efficacité de quatre classifieurs :

- des arbres CART classiques utilisant l'indice de Gini (CART) ;
- des arbres basés sur U_λ avec m_{IDM} (IDM) ;
- des combinaisons d'arbres à deux classes basés sur U_λ (combi), pour lequel le nombre moyen de noeuds n'est pas spécifié (ce pa-

Jeux de données	# d'attributs	# classes	taille jeu apprentissage	taille jeu de test
Iris	4	3	113	37
Balance scale	4	3	469	156
Wine	13	3	134	44
Car	6	4	1152	576
Page blocks	10	5	3649	1824
Forest-fires	12	6	345	172

Tableau 2 – Jeux de donnée UCI utilisés lors des expériences

	CART			IDM			Combi		Multi		
datasets	*R*	*t*	*nb*	*R*	*t*	*nb*	*R*	*t*	*R*	*t*	*nb*
iris	2.0%	0	5	2.0%	0	5	2.0%	1	2.0%	6	5
balance-scale	20.2%	2	27	25.0%	0	27	17.8%	2	15.9%	29	27
wine	11.9%	0	7	8.5%	0	9	13.6%	1	13.6%	19	22
car	17.7%	1	17	17.7%	0	17	15.6%	9	32.3%	9	1
pageblocks	4.8%	53	23	4.7%	38	27	5.0%	140	5.2%	1801	25
forests-fire	43.6%	0	13	43.0%	0	21	43.0%	15	43.0%	78	1

Tableau 3 – Efficacité (*R*=taux d'erreur *t*=temps de calcul en secondes *nb*=nombre de noeuds moyen par arbre) des arbres en fonction du modèle d'assignement des masses

ramètre étant moins pertinent dans ce cas) ;
– des arbres basés sur U_λ avec $m_{Multinomiale}$ (multi).

La stratégie de construction des arbres est la suivante : continuer à scinder tant que $IG > 0$ pour CART et les arbres basés sur U_λ et que la taille des sous-échantillons de chaque noeud enfant est supérieure à 10. La profondeur des arbres est limitée à 5. La valeur de λ a été fixée à 0.5.

Partant du fait que nous nous intéressons surtout à l'efficacité des modèles et pas à leur simplicité, aucun de ces arbres n'a été post-élagué (définir une stratégie d'élagage adéquate pour des arbres de décision basé sur U_λ constitue l'objet de futures recherches).

Pour les jeux de données à 3 classes, le modèle multinomial utilise simplement la fonction de croyance induite par P^- alors que pour *Page blocks* les fonctions de croyance sont approximées par programmation linéaire et l'algorithme des classes ordinales est utilisé pour $Forest-fires$.

Comme nous pouvons le constater, les différentes efficacités des classifieurs sont compétitives bien que les temps de calcul soient allongés pour le modèle multinomial, du fait de l'utilisation de nombreux éléments focaux et de la programmation linéaire.

4.2 Discussion à propos de λ

Les figures 2 et 3 montrent l'impact de λ sur la complexité (en nombre de noeuds) de l'arbre obtenu et sur son efficacité sur le jeu de donnée "Pima". Nous constatons que cette complexité augmente avec λ, ce qui confirme l'intuition que $1-\lambda$ peut être interprété comme un indicateur de l'importance donnée au manque d'individus dans un noeud (i.e. la non-spécificité $N(m)$) et à la tendance de IG à être négative. Ceci suggère que l'optimisation de ce paramètre devrait intégrer la complexité des arbres comme critère. Le paramètre λ semble n'avoir qu'une faible influence sur l'efficacité des arbres.

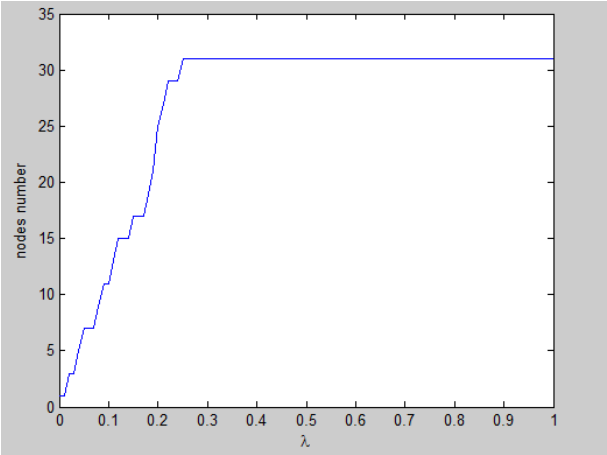

Figure 2 – Nombre de noeuds en fonction de λ (gauche) et taux d'erreur en fonction de λ (droite) pour le jeu de données Pima

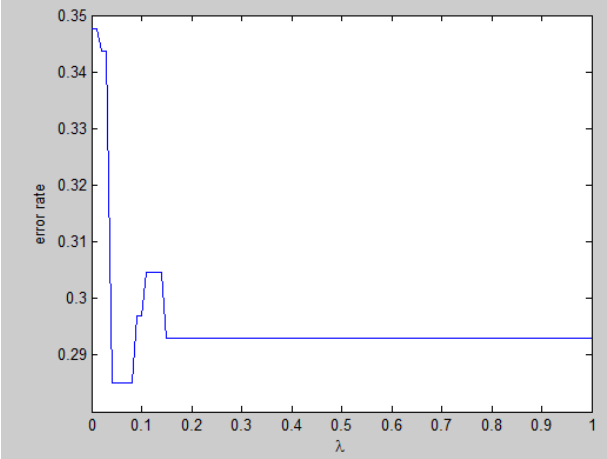

Figure 3 – Nombre de noeuds en fonction de λ (gauche) et taux d'erreur en fonction de λ (droite) pour le jeu de données Pima

5 Conclusion

Dans ce papier, nous avons étendu la méthode de Skarstein Bjanger de construction d'arbres de décision au cas multi-classe, proposant trois moyens d'y parvenir. L'IDM ne s'inscrit pas vraiment dans le cadre de la théorie des fonctions de croyance mais génère des fonctions de croyance simples ; le modèle multinomiale de Denœux est plus élaboré, correspond mieux à l'approche des fonctions de croyance, mais nécessite de lourds efforts calculatoires ; la combinaison de classifieurs à deux classes est efficace mais rend l'interprétation des résultats difficile (tout du moins plus longue), générant un nombre quadratique d'arbres de décision.

Nous avons montré que les méthodes présentées ont des capacités prédictives comparables aux méthodes classiques. Cependant, ce travail n'est en réalité qu'un point de départ avec divers perspectives : un des principaux intérêts de l'utilisation de fonctions de croyance est la capacité de gérer des données incertaines en entrée ou en sortie, une option que nous devrions intégrer à la méthode présente dans de futurs travaux (en utilisant, par exemple, une extension de l'algorithme EM pendant l'apprentissage des arbres [3] [6]). Une autre extension intéressante serait d'adapter ce modèle aux sorties continues et aux problèmes de régression. Des expériences supplémentaires étudiant le comportement de la méthode devraient aussi être effectuées.

Références

[1] J. Abellan and S. Moral. Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning*, 39(2-3) :235–255, June 2005.

[2] Breiman, Friedman, Olshen, and Stone. *Classification And Regression Trees*. Chapman and Hall/CRC, 1984.

[3] A. Ciampi. Growing a tree classifier with imprecise data. *Pattern Recognition Let-*

ters, 21(9) :787–803, Aug. 2000.

[4] L. de Campos, J. Huete, and S. Moral. Probability intervals : a tool for uncertain reasoning. *Int. J. Uncertainty Fuzziness Knowledge-Based Syst.*, 1 :167–196, 1994.

[5] T. Denoeux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3) :228–252, Aug. 2006.

[6] T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. on Know. and Data Eng. (in press)*, 2011.

[7] T. Denoeux and M. Bjanger. Induction of decision trees from partially classified data using belief functions. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 4, pages 2923–2928. IEEE, 2000.

[8] Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees : theoretical foundations. *International Journal of Approximate Reasoning*, 28(2-3) :91–124, 2001.

[9] L. A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2) :247–254, 1965.

[10] G. J. Klir. *Uncertainty and information : foundations of generalized information theory*. Wiley-IEEE Press, 2006.

[11] J. Quinlan. Induction of decision trees. *Machine Learning*, 1 :81–106, Oct. 1986.

[12] B. Quost and T. Denoeux. Pairwise Classifier Combination using Belief Functions. *Pattern Recognition Letters*, 28 :644–653, 2006.

[13] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[14] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66 :191–234, 1994.

[15] L. V. Utkin. Extensions of belief functions and possibility distributions by using the imprecise dirichlet model. *Fuzzy Sets and Systems*, 154(3) :413–431, 2005.

[16] P. Vannoorenberghe and T. Denoeux. Handling uncertain labels in multiclass problems using belief decision trees. *Intelligence*, 2002.

[17] P. Walley. Inferences from multinomial data : Learning about a bag of marbles. *Journal of the Royal Statistical Society*, Series B, :3–57, 1996.