Correction bayésienne de prédictions issues d'arbres de décision et évaluation crédibiliste

Nicolas Sutton-Charani

Correction bayésienne de prédictions issues d'arbres de décision et évaluation crédibiliste

Bayesian correction for decision tree predictions and evidential evaluation

N. Sutton-Charani

IMT Mines-Alès

Centre de recherche LGI2P
École des mines d'Alès
6 avenue de Clavières F-30 319 Alès Cedex
nicolas.sutton-charani@mines-ales.fr

Résumé :

Comme pour de nombreux classifieurs, les prédictions issues d'arbres de décision sont naturellement probabilistes. A chaque feuille de l'arbre est associée une distribution de probabilité sur les labels estimée de façon fréquentiste. Ces probabilités présentent ainsi l'inconvénient majeur d'être potentiellement non-fiables dans le cas où elles sont estimées à partir d'un faible nombre d'exemples. Les approches bayésiennes empiriques permettent la mise-à-jour de distributions de probabilité en fonction des effectifs observés. Cet article présente une approche de correction des probabilités prédictives binaires issues d'arbres de décision au travers l'utilisation d'une méthode bayésienne empirique. L'ajustement des probabilités prédictives des arbres est ainsi concentré sur les feuilles de petites tailles, ce qui entraîne une nette amélioration des performances prédictives. L'amplitude de ces corrections est utilisée pour générer des fonctions de croyance prédictives qui sont finalement évaluées par l'extension incertaine de trois indices d'évaluation de probabilités prédictives.

Mots-clés :

Correction, probabilités prédictives, arbres de décision, méthode bayésienne empiriques, fonctions de croyance prédictives.

Abstract:

As for many classifiers, decision trees predictions are naturally probabilistic, with a frequentist probability distribution on labels associated to each leaf of the tree. Those probabilities have the major drawback of being potentially unreliable in the case where they have been estimated from a limited number of examples. Empirical Bayes methods enable the updating of observed probability distributions for which the parameters of the *prior* distribution are estimated from the data. This paper presents an approach of correcting decision trees predictive binary probabilities with an empirical Bayes method. The update of probability distributions associated with tree leaves creates a correction concentrated on small-sized leaves, which improves the quality of probabilistic tree predictions. The amplitude of these corrections is used here to generate predictive belief functions which are finally evaluated through the extension of three evaluation indexes of predictive probabilities.

Keywords:

Correction, predictive probabilities, decision trees, Bayesian empirical methods, predictive belief functions.

1 Introduction

Même si les prédictions issues de l'apprentissage de modèles de classification sont généralement fournies sous une forme *précise*, elles sont souvent initialement calculées sous forme de distributions de probabilité, les labels les plus probables servant de prédictions précises au stade prédictif final ou décisionnel. Les arbres de décision constituent un modèle de classifieurs et de régresseurs basiques de l'apprentissage automatique. Une fois un arbre construit, les effectifs des labels des exemples terminant dans chaque feuille sont utilisés pour calculer ces probabilités prédictives. Les probabilités associées aux feuilles de petite taille, i.e. ne contenant qu'un faible nombre d'exemples, ne sont alors que peu fiables étant calculées à partir de peu de données.

Il existe dans la littérature classique de l'apprentissage automatique un ensemble de travaux traitant les problématiques de calibration des probabilités prédictives associés aux classi-

fieurs qui permettent de *lisser* ou corriger certain biais intrinsèques aux différents modèles prédictifs considérés [10]. Ces approches impliquent souvent l'application systématisées de fonctions mathématiques nécessitant souvent un ensemble de données dédié à l'étape de calibration et impliquant parfois des calculs lourds en termes de complexité [12] ou ne considérant qu'une sous partie des données d'apprentissage [15]. D'autres travaux basés sur des modèles crédibilistes permettent d'ajuster les modèles aux situations où des estimations sont effectuées sur de petits sous-ensembles de données [4]. Aucune de ces approches ne permet à notre connaissance un ajustement basé sur la distribution globale des données d'apprentissage et n'impliquant pas de complexité supplémentaire importante.

Cet article présente une approche permettant l'ajustement des probabilités prédictives d'arbres de classification dans le cas de classes binaires. Pour y parvenir, une méthode bayésienne empirique tenant compte de l'ensemble de l'échantillon d'apprentissage est utilisée et a pour conséquence des ajustements des probabilités associées aux feuilles relativement à leur taille.

Après un rappel des notions nécessaires en Section 2, l'approche proposée est décrite en détail en Section 3. La prise en compte de l'amplitude des ajustements de probabilités prédictives réalisés permet, en Section 4, la formalisation d'un modèle génératif de fonctions de croyance prédictives et l'extension de trois métriques d'évaluation de probabilités prédictives aux cas de prédictions incertaines crédibilistes. En Section 6, une première série d'expériences illustre l'apport de la méthodologie d'une part de façon pragmatique en termes de qualité prédictive et d'autre part par la souplesse décisionnelle qu'elle offre.

2 Pré-requis

2.1 Arbres de décision

L'apprentissage d'un arbre de décision correspond à un partitionnement récursif de l'espace des attributs ou variables prédictrices tendant à séparer au mieux les labels (classification) ou à diminuer leur variance (régression) [1]. Les données d'apprentissage se répartissent ainsi dans les différentes feuilles des arbres qui sont alors associées aux distributions de probabilité correspondant aux proportions de labels des exemples qu'elles contiennent. Dans cet article nous nous restreignons au cas de classes binaires notées $\{1, 0\}$ ou $\{+, -\}$.

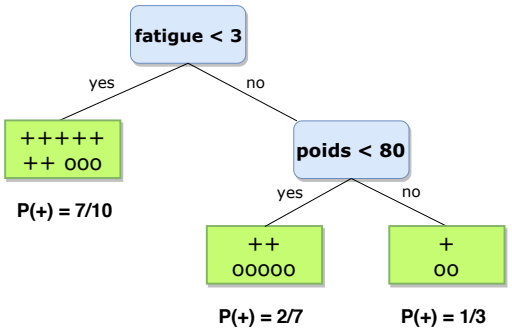

Figure 1 – Probabilités prédictives

La figure 1 est un exemple d'arbre de décision dans lequel tout exemple exprimant une fatigue inférieure à 3 aura une probabilité de classe positive estimée par $7/10$, pour tout exemple plus fatigué (plus que 3) et de poids inférieur à 80kg on aura $P(+) = 2/7$ et enfin pour un exemple exprimant une fatigue supérieure à 3 et un poids supérieure à 80kg on aura $P(+) = 1/3$. Cette dernière probabilité n'est ici estimée qu'à partir de 3 exemples, il est donc naturel de la considérer relativement peu fiable.

Différents critères d'arrêts sont envisageables pendant l'apprentissage d'un arbre de décision en fonction de la structure de l'arbre (nombre de feuille, profondeur, etc) ou en termes d'information (gain d'impureté, variance). De manière à éviter le sur-apprentissage des méthodes

d'élagage sont généralement mises en œuvre pour limiter le nombre de feuilles.

2.2 Méthode bayésienne empirique

L'inférence bayésienne est une branche importante des Statistiques qui consiste à utiliser les données disponibles pour mettre à jour la connaissance qu'on a *a priori* sur le phénomène étudié. Il résulte de cette mise à jour l'obtention de probabilités prédictives dont la qualité dépendra directement des données. Alors que l'*a priori* bayésien est généralement constitué d'une distribution de probabilité que l'utilisateur attribue subjectivement au phénomène étudié (souvent à partir de connaissance experte), pour la méthode bayésienne *empirique* [13, 3], les paramètres de cette distribution sont eux aussi estimés à partir des données. Certains auteurs considèrent ces méthodes comme des approximations de modèles bayésiens hiérarchiques.

En considérant un échantillon d'une variable binaire $x = (x_1, ..., x_N) \in \{0, 1\}^N$, pour tout sous-ensemble $x^* \subseteq x$, une estimation naturelle (et fréquentiste) de la probabilité de 1 au sein de x^*, $P(X = 1|X \in x^*)$, est sa fréquence observée $p_1^* = \frac{|x^* \cap \{x_i : x_i = 1\}|}{|x^*|}$. Cet estimateur fait implicitement l'hypothèse que l'échantillon x^* est suffisamment grand pour estimer la probabilité d'obtenir 1 par un tirage aléatoire futur en son sein. En supposant que la proportion de 1 au sein de x suit une loi $Beta(\alpha, \beta)$ (connaissance *a priori* souple), l'approche bayésienne empirique revient à estimer α et β à partir des données x puis d'ajuster p_1^* comme suit :

$$\widehat{p_1^*} = \frac{|x^* \cap \{i : \ x_i = 1\}| + \alpha}{|x^*| + \alpha + \beta} \qquad (1)$$

De cette manière p_1^* sera rapproché de l'espérance de X, notée $E[x] = \frac{\alpha}{\alpha+\beta}$ et faisant référence à tout l'échantillon x, et l'amplitude de ce rapprochement sera d'autant plus importante que la taille de x^* est petite. L'estimateur bayésien de Cette approche, illustrée pour le nombre de tirs réussis dans le baseball dans [2], est ici appliquée aux prédictions associées aux feuilles d'un arbre de classification binaire.

2.3 Evaluation de probabilités prédictives

Même si l'évaluation d'un classifieur est souvent réalisée à partir de prédictions précises en les comparant au vrais labels au travers différentes métriques (e.g. justesse, précision, rappel) elle peut cependant être faite au niveau des probabilité prédictives, donc en amont. Trois métriques d'évaluation de prédictions probabilistes binaires à valeurs dans $[0, 1]$ sont ici présentées.

En notant $y = (y_1, ..., y_N) \in \{1, 0\}^N$ les *vrais* labels et $p = (p_1, ..., p_N) = [P(y_1 = 1), ..., P(y_N = 1)]$ les probabilités prédictives de classe $\{1\}$, le Tableau 1 résume les définitions de l'entropie croisée ou *log-perte*, du score de Brier et de l'aire sous la courbe ROC. Les deux premières métriques mesurent l'écart entre les observations et les probabilités prédites, en pénalisant les probabilités des labels les moins probables. La courbe ROC est un moyen classique pour quantifier le potentiel prédictif d'un classifieur binaire, elle représente le taux de vrais positifs (i.e. *sensibilité*) en fonction du taux de faux positifs ($1-$*spécificité*) en considérant une variation du seuil d'attribution des classes λ, on a $ROC(p, y, \lambda) = sensibilité_{p,y}^{\lambda}(1 - specificité_{p,y}^{\lambda})$. L'aire AUC sous la courbe ROC est un indicateur reconnu de la qualité des prédictions probabilistes.

Un bon classifieur binaire sera donc caractérisé par des valeurs d'entropie croisée et de score de Brier proches de 0 et une valeur de AUC proche de 1. Il est cependant à noter que ces trois métriques sont définies pour des prédictions incertaines classiques, *i.e.* probabilistes.

Tableau 1 – Métriques d'évaluation de prédictions probabilistes binaires $(p_1, ...p_N)$ au vu des vrais labels $(y_1, ...y_N)$

Nom de la métrique	Définition
entropie croisée	$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$
score de Brier	$\frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2$
aire sous la courbe ROC	$\int_0^1 ROC(p, y, \lambda)d\lambda$

3 Correction bayésienne empirique des probabilités prédictives d'arbres de décision

L'approche présentée dans ce papier consiste à corriger les probabilités prédictives des H feuilles d'un arbre de classification binaire par une méthode bayésienne empirique. On fait l'hypothèse que la proportion de label $\{1\}$ au sein des feuilles d'un arbre suit une loi $Beta(\alpha, \beta)$ (sans fixer *a priori* les valeurs des paramètres α et β). Cette hypothèse est classique en inférence bayésienne où la loi $Beta$ est la loi conjuguée a priori de la loi binomiale. On peut aussi remarquer que la loi Beta est à la fois un cas particulier de la loi de Dirichlet (très utilisée en Statistique Bayésienne) et une généralisation de la loi uniforme ($\mathcal{U}_{[0,1]} = Beta(1, 1)$).

Une fois l'hypothèse de la loi Beta formulée, ses paramètres α et β sont estimés à partir de l'ensemble $\{p_1^1, ..., p_1^H\}$ des proportions de label $\{1\}$ au sein des H feuilles de l'arbre considéré. De manière à pénaliser les feuilles de petite taille, on créera un échantillon E contenant chaque proportion p_1^h répétée un nombre de fois égale à la taille de la feuille considérée, i.e. au nombre d'exemples qu'elle contient. L'estimation des paramètres α et β peut être ensuite réalisée sur E par différentes approches (moments, maximum de vraisemblance, moindres carrés, etc).

Les proportions p_1 au sein des feuilles sont enfin corrigées selon l'équation (1).

Il est à noter que d'autres travaux [12, 15] permettent une *calibration* des probabilités prédictives par différentes approches utilisant soit uniquement les distributions des exemples contenues dans les feuilles de façon indépendante, soit la distribution de tout l'échantillon d'apprentissage mais en appliquant des transformations systématiques basées sur des estimations nécessitant de nombreux calculs (souvent obtenus par validation croisée). La méthode proposée dans cet article utilise à la fois l'ensemble de la distribution des données d'apprentissage et reste très simple en termes de complexité, les paramètre α et β de l'Equation (1) n'étant estimés qu'une seule fois pour tout l'arbre de classification considéré.

4 Génération de fonctions de croyance prédictives

L'incertitude exprimée dans les probabilités prédictives d'un classifieur est principalement aléatoire. En effet elle repose sur le modèle mathématique sous-jacent au classifieur et sur des estimations fréquentistes. La connaissance de l'ajustement bayésien empirique et de son amplitude peut permettre d'incorporer de l'incertitude épistémique aux probabilités prédictives. Il est en effet naturel de considérer peu fiables des probabilités prédictives estimées sur un faible nombre d'exemples (et donc fortement ajustées). Nous proposons en première approche de générer une fonction de croyance à partir d'une probabilité prédictive en utilisant son ajustement bayésien empirique et en attribuant un poids à l'ignorance égale à l'amplitude de cet ajustement. Cette modélisation sous-entend que, plus une probabilité prédictive est corrigée (*i.e.* plus le sous-échantillon considéré est petit), moins elle est fiable. Si on note p_1 et $\widehat{p}_1$ une probabilités prédictive de la classe $\{1\}$ et son ajustement, on obtient :

$$\begin{cases} m(\{1,0\}) &= |p_1 - \widehat{p}_1| \\ m(\{1\}) &= \widehat{p}_1 - \frac{|p_1 - \widehat{p}_1|}{2} \\ m(\{0\}) &= 1 - \widehat{p}_1 - \frac{|p_1 - \widehat{p}_1|}{2} \end{cases} \quad (2)$$

De manière à faire le moins d'hypothèses possible, une fois la masse de l'ignorance définie, la correction s'effectue symétriquement entre les 2 classes. On peut alors remarquer qu'on a $Bel(\{0\}) = 1 - Pl((\{1\})$ et $Bel(\{1\}) = 1 - Pl((\{0\})$.

Remarque **:** Ce modèle de fonction de croyance peut aussi s'écrire sous forme de probabilité imprécise : $p_1 \in [p_1^-, p_1^+]$ avec

$$\begin{cases} p_1^- &= p_1 - \frac{|p_1 - \widehat{p}_1|}{2} \\ p_1^+ &= p_1 + \frac{|p_1 - \widehat{p}_1|}{2} \end{cases} \quad (3)$$

5 Evaluation incertaine

De manière à conserver l'incertitude prédictive contenue dans le modèle (3) jusqu'au stade de l'évaluation des classifieurs, il est possible de considérer les métriques définies dans le Tableau 1 de façon ensembliste ou *intervalliste*. En effet à un ensemble de probabilités prédictives correspond naturellement un ensemble de valeurs prises par ces métriques d'évaluation. Une probabilité imprécise $[p^-, p^+]$ calculée selon le modèle (3) sera donc évaluée de façon imprécise par un intervalle défini comme suit :

$$err([p^-, p^+], y) = \left[\min_{p \in [p^-, p^+]} err(p, y), \max_{p \in [p^-, p^+]} err(p, y) \right]$$

où err sera une des métriques définies dans le Tableau 1.

Il va de soi qu'en procédant de cette manière, plus les feuilles de l'arbre de décision évalué seront petites, plus les probabilités associées à ces feuilles seront ajustées et plus l'évaluation des dits arbres sera imprécise (i.e. plus les intervalles obtenus seront larges). Cette façon de procéder représente donc bien un moyen de propager l'incertitude épistémique relative à la structure de l'arbre jusqu'à son évaluation. Il est cependant à noter que cette solution requiert un parcours effectif de tout l'intervalle $[p^-, p^+]$ ce qui implique une complexité calculatoire importante.

6 Expériences

Dans cette section, une série d'expériences est mise en oeuvre de manière à illustrer l'intérêt pratique du modèle de correction bayésienne empirique présenté dans cet article. A partir de six jeux de données benchmark issus des sites UCI [1] et Kaggle [2], des validations croisées à 10 couches sont effectuées avec pour chaque couche de chaque jeu de données, l'apprentissage d'arbres de décision correspondant à différentes complexités sur les neuf autres couches et l'évaluation sur la couche en question à l'aide des trois métriques explicitées dans le Tableau 1. Le paramètre 'cp' (de la fonction rpart sous R) permet de contrôler cette complexité, il représente le gain d'information minimal de chaque coupure considérée pendant l'apprentissage des arbres. Les arbres nommés 'pruned' sont appris avec une complexité maximale ($cp = 0$) et sont ensuite élagués selon l'approche classique de l'algorithme $CART$ [1]. Les boites à moustache de type 'classique' et 'EB' font respectivement référence aux approches classiques et ajustées par méthode bayésienne empirique. Les évaluations sont constituées des calculs de métriques précises présentées dans le Tableau 1 ainsi que de leurs extensions incertaines définies en Section 5. Ces étapes sont répétées 150 fois de manière à rendre robuste au bruit les résultats et seules les évaluations moyennes des probabilités prédictives des arbres sont ici représentées. Les codes utilisés pour l'implémentation de l'ensemble des expériences présentées ci-après sont disponibles sur https://github.com/lgi2p/empiricalBayesDecisionTrees .

Le tableau 2 représente les caractéristiques des différents jeux de données utilisés en termes de nombre d'exemples (N), de nombre d'attributs ou variables prédictrices (J) et de nombre de classes ou labels (K). Les Tableaux 3, 4, 5, 6, 7 et 8 contiennent les évaluations moyennes

1. https ://archive.ics.uci.edu/ml/datasets.html
2. https ://www.kaggle.com/datasets

Tableau 2 – Dimensions des jeux de données

	N	J	K
banana	5300	2	2
bankLoan	5000	12	2
banknote	1372	4	2
mammo	830	5	2
pima	768	8	2
ticTacToe	958	9	2

calculées pour chaque jeu de données et pour chaque type d'arbre, sur l'ensemble des 150 validations croisées effectuées, avant et après corrections. Les Figures 2 et 3 illustrent les distributions de ces résultats sous forme de boites à moustache pour l'entropie croisée *précise* sur le jeu de données 'banana' et pour le score de Brier *précis* sur le jeu de données 'bankLoan' au regard des supports des évaluations incertaines correspondantes (seuls les extremums des métriques incertaines sont représentés).

Tableau 3 – Entropies croisées avant correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.410	0.351	0.355	0.419	0.425	0.359
bankLoan	1	0.665	0.568	0.571	0.571	0.566
banknote	0.330	0.329	0.322	0.331	0.347	0.329
mammo	0.444	0.437	0.414	0.409	0.450	0.416
pima	0.886	0.884	0.742	0.655	0.565	0.630
ticTacToe	0.221	0.221	0.213	0.210	0.549	0.214

Tableau 4 – Entropies croisées après correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.287	0.300	0.334	0.418	0.424	0.358
bankLoan	0.640	0.582	0.568	0.571	0.571	0.566
banknote	0.173	0.174	0.180	0.225	0.346	0.328
mammo	0.419	0.419	0.410	0.406	0.450	0.416
pima	0.557	0.557	0.545	0.541	0.564	0.630
ticTacToe	0.188	0.188	0.188	0.191	0.550	0.214

On observe des entropies croisées des arbres corrigées quasiment systématiquement inférieures à celles des arbres non-corrigés. Cette augmentation de performance est nette pour les grands arbres (appris avec une faible

Tableau 5 – Scores de Brier avant correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.084	0.086	0.097	0.129	0.131	0.086
bankLoan	0.229	0.204	0.190	0.192	0.192	0.189
banknote	0.040	0.040	0.043	0.057	0.098	0.041
mammo	0.130	0.130	0.125	0.123	0.140	0.126
pima	0.188	0.188	0.184	0.183	0.189	0.186
ticTacToe	0.061	0.061	0.061	0.062	0.187	0.062

Tableau 6 – Scores de Brier après correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.083	0.086	0.097	0.129	0.131	0.086
bankLoan	0.219	0.198	0.190	0.192	0.192	0.189
banknote	0.040	0.040	0.042	0.057	0.098	0.041
mammo	0.129	0.129	0.125	0.123	0.140	0.126
pima	0.183	0.183	0.179	0.179	0.188	0.186
ticTacToe	0.060	0.060	0.060	0.062	0.187	0.061

Tableau 7 – AUC avant correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.946	0.936	0.920	0.872	0.867	0.937
bankLoan	0.778	0.772	0.747	0.722	0.720	0.755
banknote	0.843	0.838	0.821	0.799	0.777	0.827
mammo	0.852	0.848	0.835	0.819	0.789	0.839
pima	0.839	0.836	0.826	0.812	0.771	0.821
ticTacToe	0.861	0.858	0.850	0.838	0.758	0.846

Tableau 8 – AUC après correction

dataset \ cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.948	0.937	0.920	0.872	0.867	0.937
bankLoan	0.781	0.773	0.747	0.722	0.720	0.755
banknote	0.845	0.840	0.822	0.799	0.777	0.827
mammo	0.854	0.850	0.836	0.819	0.789	0.838
pima	0.842	0.839	0.828	0.813	0.771	0.821
ticTacToe	0.864	0.860	0.851	0.839	0.758	0.845

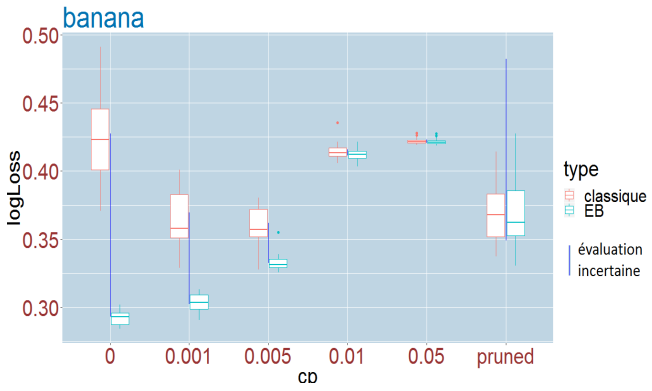

Figure 2 – Entropie croisée précise et incertaine en fonction de la complexité

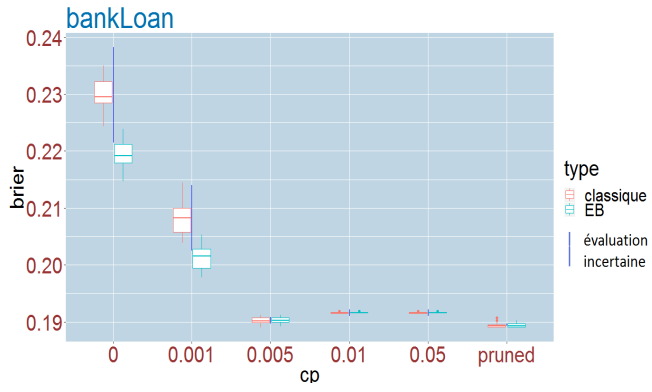

Figure 3 – Score de Brier précis et incertain en fonction de la complexité

valeur de l'hyper-paramètre 'cp' et contenant ainsi de nombreuses feuilles), un peu moins visible pour les petits arbres et très limitée pour les arbres élagués ('pruned'). Le même phénomène de gain en performance proportionnel à la taille des arbres est globalement observable pour le score de Brier et l'indice AUC mais dans de plus petites proportions.

Les intervalles formés par les évaluations incertaines correspondent à peu près aux intervalles formés par les évaluations précises sans et avec corrections bayésiennes. On peut toutefois remarquer sur les Figure 2 et 3 qu'il arrive que les évaluations incertaines *sortent* quelque fois de ces bornes naturelles (arbres élagués de la Figure 2 et *grands* arbres de la Figure 3), mettant ainsi en avant la non-convexité des métriques d'évaluation des probabilités prédictives incertaines.

7 Conclusion

Le modèle de correction bayésienne empirique présenté dans cet article pour les probabilités prédictives issues d'arbres de décision présente un intérêt certain en termes de performances prédictives et cet intérêt concerne surtout les arbres de grande taille. Le fait que les corrections bayésiennes n'améliorent quasiment pas les performances des arbres *petits* (i.e. de faible complexité) ou élagués laisse à penser que la correction bayésienne représente une sorte d'équivalent ou d'alternative à l'élagage, qu'en ramenant les probabilités prédictives des petites feuilles vers leurs moyennes globales (i.e. calculées au sein de l'échantillon d'apprentissage total) elle diminue le phénomène de surapprentissage.

Dans cet article les corrections bayésiennes ne sont réalisées qu'au stade prédictif, i.e. au niveau des feuilles. Adopter la même démarche pendant tout le processus d'apprentissage est envisageable en procédant de la même manière au niveau des calculs d'impureté (donc pour toutes les coupures considérées). Il sera alors intéressant de comparer les résultats obtenus

avec ceux de [4] qui poursuit le même but (pénaliser les petites feuilles) en se basant sur un modèle fréquentiste *ajusté* de façon crédibiliste où un poids de $\frac{1}{N+1}$ est attribué à l'ignorance lors de l'évaluation des gains d'information des différentes coupures à partir des probabilités de chaque classe. Les approches de génération de fonctions de croyance prédictives basées sur l'utilisation de la vraisemblance crédibiliste [6, 9] représentent aussi une alternative intéressante à laquelle il sera important de se comparer tant en termes de performances prédictives que relativement à la sémantique sous-jacente. Il est important de remarquer que les deux approches précédemment citées se basent sur la distribution des exemples au sein des feuilles de façon individuelle, en cas d'échantillon d'apprentissage déséquilibré elles ne permettent pas de correction en direction de la distribution générale comme c'est le cas avec le modèle bayésien empirique. Dans le même esprit, l'approche présentée dans ce papier pour le contexte de classification binaire pourrait être étendue au cas multiclasse en réutilisant les méthodes employées dans [14]. De façon plus générale, tous les classifieurs dont l'apprentissage ou l'utilisation au stade prédictif implique des calculs de probabilités fréquentistes pourraient potentiellement bénéficier de ce type de correction bayésienne.

Le modèle de génération de fonctions de croyance prédictives et surtout l'extension des métriques d'évaluation au contexte crédibiliste proposés dans ce travail pourraient être largement enrichis par une modélisation plus fine de l'incertitude issue des probabilités prédictives initiales et de leurs corrections. Il serait par exemple possible d'utiliser les distances proposées dans [8] de manière à complexifier la représentation des métriques d'évaluation incertaines au-delà des simples intervalles. Il serait aussi souhaitable d'estimer directement les bornes des intervalles d'évaluation incertaine sans avoir à les parcourir de façon effective à partir d'approches à base de simulations comme dans [5] ou de résultats d'optimisation tels que dans [7].

Références

[1] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. Classification And Regression Trees. *Chapman & Hall*, 1984.

[2] L. P. Brown. Empirical Bayes in-season prediction of baseball batting averages. *The annals of applied statistics*, 2(1) : 113-152, 2008.

[3] G. Casella. An Introduction to Empirical Bayes Data Analysis. *American Statistician - AMER STATIST*, 39(5) : 83-87, 1985.

[4] T. Denœux and M. Bjanger. Induction of decision trees from partially classified data using belief functions. *international conference on systems, man and cybernetics (SMC 2000)*, 4 : 2923-2928, 2000.

[5] T. Denœux, M-H. Masson, P-A Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets and Systems*, 153(1) : 1-28, 2005.

[6] T. Denœux. Likelihood-based belief function : Justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7) : 1535-1547, 2014.

[7] S. Destercke, O. Strauss. Kolmogorov-Smirnov Test for Interval Data. *Information Processing and Management of Uncertainty (IPMU), 2014*, pp.416-425.

[8] A-L. Jousselme, P. Maupin. Distances in evidence theory : Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 53(2) : 118-145, 2012.

[9] O. Kanjanatarakul, S. Sriboonchitta, T. Denœux. Forecasting using belief functions : An application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5) : 1113-1128, 2014.

[10] M. Kuhn, K. Johnson. Applied Predictive Modeling. *Springer*, 2013.

[11] V-L. Nguyen, S. Destercke, M-H. Masson and E. Hüllermeier. Reliable Multi-class Classification based on Pairwise Epistemic and Aleatoric Uncertainty. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 2018*, pp. 5089-5095.

[12] A. Niculescu-Mizil, R. Caruana. Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22Nd International Conference on Machine Learning (ICML '05), 2005*, pp. 625-632.

[13] H. Robbins. *An Empirical Bayes Approach to Statistics*. University of California Press, 1 : 157-163, 1956.

[14] N. Sutton-Charani, S. Destercke, T. Denœux. Arbres de classification construits à partir de fonctions de croyance. *21ème acte des Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2012)*, 2012.

[15] B. Zadrozny, C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, 2001.