Affectation pondérée sur des données de type intervalle

Chérif Mballo [*,**] & Edwin Diday [**]

[*] ESIEA, Pôle E. C. D, Laval France
[**] LISE-CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16 France
mballo@esiea-ouest.fr; diday@ceremade.dauphine.fr

Résumé. On s'intéresse à la construction d'arbres de décision sur des données symboliques de type intervalle en utilisant le critère de découpage binaire de Kolmogorov-Smirnov. Nous proposons une approche permettant d'affecter un individu à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. Le but de cette méthode est de prendre en compte le positionnement de la donnée à classer par rapport à la donnée seuil de coupure.

1 Introduction

Désignons par $\Im$ l'ensemble des intervalles fermés bornés de $\Re$ (ensemble des nombres réels) : $x \in \Im$, on note $x = [i(x), s(x)]$. Indiquons par Ω un ensemble d'individus ou objets. Une variable X est de type intervalle (Bock et al., 2000) si $X(w) = [\alpha, \beta]$ $\forall w \in \Omega$, où $\alpha, \beta \in \Re$ et $\alpha \leq \beta$. Différentes méthodes de construction d'arbres de décision sur des données symboliques de type intervalle ont été proposées ((Limam, 2005) ; (Périnel, 1996)) en utilisant le critère de Gini et le likelihood. Nous privilégions ici le critère de découpage binaire de Kolmogorov-Smirnov noté KS dans la suite. Ce critère a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des données continues. Il a été étendu aux données qualitatives classiques par (Asseraf, 1998). Nous avons examiné son adaptation aux données symboliques de type intervalle ((Mballo et al., 2004), (Mballo et Diday, 2005 c)) mais dans cette approche, un individu est affecté entièrement à un nœud de l'arbre de décision. Nous proposons dans ce papier une approche permettant d'affecter un individu à la fois aux deux nœuds fils générés par le découpage d'un nœud non terminal. Nous terminerons par un exemple pour illustrer cette approche.

2 Présentation du critère de Kolmogorov-Smirnov

Supposons que l'ensemble Ω défini précédemment est une population théorique de n objets destinés à être classés par un arbre de décision. Ces objets sont décrits par $(p+1)$ variables : p variables de type intervalle $X_1, X_2, ..., X_p$ (variables explicatives) et une

variable classe Y (variable à expliquer). Soit D_{X_j} l'espace d'observations d'une variable explicative X_j. D_{X_j} est un ensemble fini d'intervalles fermés bornés de $\Re$.

Le critère KS est basé sur la fonction de répartition. Il nécessite un ordre des valeurs prises par les variables explicatives. Nous pouvons ordonner des intervalles fermés bornés de différentes façons (Diday et al., 2003) : par la borne inférieure, la borne supérieure, le centre ou la longueur. Soit $w, w' \in \Omega$ tels que $X_j(w) = x_j$ et $X_j(w') = x_j^{'}$. Désignons par « $x_j \prec x_j^{'}$ » pour indiquer que l'intervalle x_j est « *avant* » (ou « *inférieur à* ») l'intervalle $x_j^{'}$ au sens d'une des méthodes citées précédemment pour ordonner des intervalles. Notons par « $\prec_I$ » (respectivement « $\prec_S$ », « $\prec_C$ » et « $\prec_L$ ») l'ordre des intervalles par la borne inférieure (respectivement la borne supérieure, le centre et la longueur). Comme le critère KS est utilisé pour une partition binaire à expliquer, dans le cas où le nombre de classes a priori k est strictement supérieur à 2, la méthode « *twoing splitting process* » proposée par (Breiman et al., 1984) permet de les regrouper en deux groupes C_1 et C_2 appelés super classes. Pendant le processus de construction de l'arbre de décision, à chaque nœud non terminal, soit F_t^j la fonction de répartition théorique d'une variable explicative X_j et associée à C_t, $j = 1,2,...,p$ et $t = 1,2$. Ces fonctions de répartition ne sont pas connues en pratique, il faut recourir à des estimations. S'il n'y a pas d'ordre sur les données, on ne peut pas estimer la fonction de répartition théorique par la fonction de répartition empirique comme dans le cas continu (Araya et Gigon, 1992). Dans notre cas, on peut faire cette estimation car on a un ordre et l'ensemble $\{y \in D_{X_j} / y \prec x\} \bigcap \{y \in D_{X_j} / y \in C_t\}$ est toujours fini en pratique. A chaque nœud non terminal, la fonction de répartition empirique notée $\hat{F}_t^j$ qui estime la fonction de répartition théorique F_t^j en $x \in D_{X_j}$ est donnée par :

$$\hat{F}_t^j(x) = \frac{Cardinal\left(\{y \in D_{X_j} / y \prec x\} \bigcap \{y \in D_{X_j} / y \in C_t\}\right)}{Cardinal\left(\{y \in D_{X_j} / y \in C_t\}\right)}$$

Ainsi, à chaque nœud non terminal, le seuil de coupure est l'intervalle c^* défini par :

$$KS(c^*) = \max_{j \in \{1,2,...,p\}} \max_{x \in D_{X_j}} \left| \hat{F}_1^j(x) - \hat{F}_2^j(x) \right|$$

Comme des intervalles fermés bornés peuvent être ordonnés de différentes façons, nous proposons deux approches pour construire un arbre de décision avec le critère KS :

- L' approche « exploratoire » : elle consiste à construire un arbre pour chaque ordre et l'arbre à retenir est celui qui a la meilleure précision.

- L'approche « décisionnelle » : elle consiste à construire un seul arbre selon le principe suivant : *à chaque nœud non terminal, tous les ordres sont examinés simultanément et celui qui donne la meilleure coupe en terme d'homogénéité des nœuds fils générés est retenu*.

3 Affectation pondérée par le critère KS sur des intervalles

Considérons un nœud non terminal admis à être découpé. Soient X_{j^*} la variable explicative la plus discriminante à ce nœud et $c^* = [i(c^*), s(c^*)]$ le seuil de coupure correspondant. Par exemple avec l'ordre par la borne inférieure, les intervalles de ce nœud peuvent se représenter comme l'indique la figure (FIG. 1) où tous les intervalles notés $x = [i(x), s(x)]$ sont les intervalles à classer.

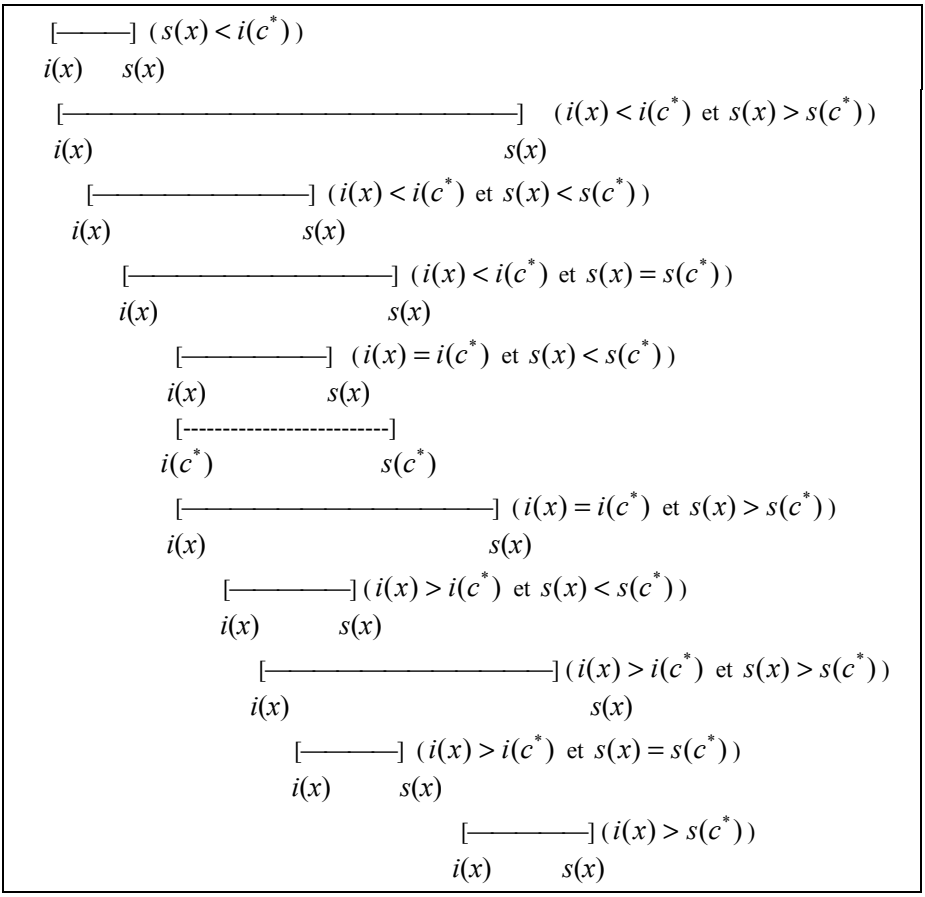

FIG. 1 – *Positionnement d'intervalles ordonnés par la borne inférieure*

Pour une affectation pure, tous les intervalles « *avant* » le seuil c^* sont affectés au nœud fils gauche (intervalles au dessus de c^*) et le reste au nœud fils droit. Cette méthode traite de la même façon un intervalle à classer disjoint avec le seuil c^* et celui non disjoint avec ce seuil. Notre objectif est de proposer une méthode d'affectation prenant en compte la position de l'intervalle à classer par rapport à l'intervalle seuil. Le but est d'affecter l'intervalle à classer à la fois à tous les deux nœuds fils générés par le découpage d'un nœud non terminal en attribuant des poids. Notons par $p_g(x)$ et $p_d(x)$ les poids à attribuer à un intervalle x à classer respectivement aux nœuds fils gauche et droit. Selon le positionnement des deux intervalles x et c^*, nous distinguons deux cas :

- **Cas 1 : les deux intervalles x et c^* sont disjoints** : si $s(x) < i(c^*)$, alors $p_g(x) = 1$ et $p_d(x) = 0$ et si $s(c^*) < i(x)$, alors $p_g(x) = 0$ et $p_d(x) = 1$.

- **Cas 2 : les deux intervalles x et c^* sont non disjoints** : on pose :

$E(x,c^*) = \max(s(x), s(c^*)) - \min(i(x), i(c^*))$ la longueur de leur étendue ;

$I(x,c^*) = \min(s(x), s(c^*)) - \max(i(x), i(c^*))$ celle de leur intersection ;

$g(x,c^*) = [\max(i(x), i(c^*)) - \min(i(x), i(c^*))]$ la longueur du débordement à gauche de l'un des deux intervalles et $d(x,c^*) = [\max(s(x), s(c^*)) - \min(s(x), s(c^*))]$ celle du débordement à droite. Avec ces notations, on montre facilement que :

$$E(x,c^*) = g(x,c^*) + I(x,c^*) + d(x,c^*)$$

Nous définissons alors les poids $p_g(x)$ et $p_d(x)$ de la façon suivante :

$$p_g(x) = \frac{g(x,c^*) + \frac{I(x,c^*)}{2}}{E(x,c^*)} \text{ et } p_d(x) = \frac{d(x,c^*) + \frac{I(x,c^*)}{2}}{E(x,c^*)}$$

Dans chacun des cas (disjoints et non disjoints), les poids $p_g(x)$ et $p_d(x)$ vérifient :

$0 \le p_g(x) \le 1$; $0 \le p_d(x) \le 1$ et $p_g(x) + p_d(x) = 1$. Dans le cas où l'intervalle à classer x est l'intervalle seuil de coupure c^*, on a : $p_g(c^*) = p_d(c^*) = 0.5$.

Les méthodes permettant d'affecter une donnée à la fois à tous les nœuds fils générés par le partitionnement d'un nœud non terminal ont été largement développées sur des données continues notamment par (Quinlan, 1990) et (Yuan et Shaw, 1995). Dans le cadre de l'analyse des données symboliques, ces méthodes ont été explorées par (Périnel, 1996) sur des données imprécises de type intervalle (le centre de l'intervalle est le seuil de coupure).

4 Exemple illustratif

Considérons le tableau (TAB. 1) où les 12 objets sont décrits par deux variables de type intervalle X_1 et X_2 et une variable classe Y ayant trois modalités notées « 1 », « 2 » et

« 3 » (Ciampi et al., 2000). Le paramètre retenu pour arrêter le développement de l'arbre est l'effectif minimum d'un nœud. Il est fixé à 3. Au niveau de l'arbre de décision, inutile de préciser l'ordre sélectionné car les poids d'un objet à classer sont calculés uniquement en fonction de sa description et du seuil de coupure. Nous obtenons la figure (FIG. 2) avec l'approche « décisionnelle » de construction d'arbres de décision par le critère KS.

Variables / *Objets*	X_1	X_2	Y
w_1	[1,3]	[1.5,2]	1
w_2	[2.5,3.5]	[3,5]	1
w_3	[3.5,6.5]	[3,3.5]	1
w_4	[5,7]	[1.5,4.5]	1
w_5	[4,8]	[0.5,2]	2
w_6	[7,7.5]	[2.5,5]	2
w_7	[7,8]	[5.5,6.5]	2
w_8	[4,6.5]	[4,5.5]	2
w_9	[3,6]	[6,6.5]	3
w_{10}	[0.5,1.5]	[3,5]	3
w_{11}	[1.5,2.5]	[5.5,6]	3
w_{12}	[1,4]	[2.5,4]	3

TAB. 1 – *Données initiales*

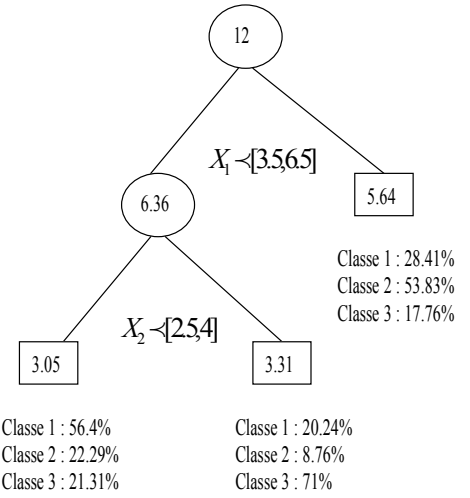

FIG. 2 – *Arbre de décision obtenu avec la méthode d'affectation pondérée*

5 Conclusion et perspectives

Dans ce papier, nous avons proposé une méthode d'affectation permettant de prendre en compte le positionnement de l'intervalle à classer par rapport à l'intervalle seuil de coupure. Avec cette approche, chaque objet pourra se retrouver à plusieurs nœuds terminaux de l'arbre de décision suivant des proportions variées. Dans le cas de l'affectation pure, toutes les deux approches « exploratoire » (Mballo et Diday, 2005 c) et « décisionnelle » (Mballo et Diday, 2005 a) ont été explorées et le critère KS a été comparé aux critères de Gini et de l'entropie (Mballo et Diday, 2005 b). Dans la suite, nous envisageons de suivre la même voie pour cette méthode d'affectation pondérée.

Références

Araya, R. and P. Gigon, (1992). Segmentation trees : a new help building expert systems and neural networks. *Comstat*, volume 1, pp 119-124.

Asseraf, M. (1998). Extension et Optimisation pour la Segmentation de la distance de Kolmogorov-Smirnov. *Thèse de Doctorat,* Université Paris Dauphine.

Bock, H. H. and E. Diday (2000). *Analysis of symbolic data : Exploratory methods for extracting statistical information from complex data*. Springer-Verlag, Berlin- Heidelberg

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Ciampi, A., E. Diday, J. Lebbe, and E. Perinel (2000). Growing a tree classifier with imprecise data. *Pattern Recognition Letters*, Number 21, pp 787-803.

Diday, E., F. Gioia et C. Mballo (2003). Codage qualitatif d'une variable intervalle, *Comptes rendus des XXXV^{ième} Journées de Statistique*, pp 415-418.

Friedman, J. H. (1977). A recursive partitioning decision rule for non parametric classification. *IEEE Transactions on Computers*, C-26, Number 4, pp 404-408.

Limam M. M. (2005). Méthodes de description de classes combinant classification et discrimination en analyse des données symboliques ; *Thèse de Doctorat*, Université Paris Dauphine.

Mballo, C. and E. Diday (2005 a). The criterion of Kolmogorov-Smirnov for binary decision tree: Application to interval valued variables; *Intelligent Data Analysis*; (A Paraître).

Mballo, C. et E. Diday (2005 b). Comparaison des critères de Kolmogorov-Smirnov, de Gini et de l'entropie sur des données de type intervalle. *Comptes rendus des XII^{ième} Rencontres de la Société Francophone de Classification*, pp 207-210.

Mballo, C. et E. Diday (2005 c). Arbres de décision sur des données de type intervalle : évaluation et comparaison ; EGC 2005 ; pp 67-78.

Mballo, C., M. Asseraf and E. Diday (2004). Binary decision trees for interval and taxonomic variables. *A Statistical Journal for Graduates Students (incorporating Data & Statistics)*, Volume 5, Number 1, April 2004, pp 13-28.

Périnel, E. (1996). Segmentation et Analyse de données symboliques : Application à des données probabilistes imprécises. *Thèse de Doctorat,* Université Paris Dauphine.

Quinlan, J. R. (1990). Probabilistic decision trees. In *Kodratoff, Y. and Michalski, R. S. (Eds.), Machine learning* III, pp 140-152.

Yuan, Y. and M. J. Shaw (1995). Induction of fuzzy decision trees. *Fuzzy sets and systems*, 69, pp 125-139.

Summary

One interests to the construction of decision trees on interval data using the Kolmogorov-Smirnov's binary splitting criterion. We propose an approach that consists in assigning an individual at once to the two children nodes generated by the splitting of a non terminal node. The aim of this method is to take account the position of the data to be classified with regard to the selected data for the cutting.