Affectation pondérée par le critère de Kolmogorov-Smirnov sur des données de type intervalle et diagramme

Chérif Mballo
Laboratoire de bioinformatique, Département d'informatique
Université du Québec à Montréal, Case Postale 8888
Succursale Centre Ville, Montréal (QC) H3C 3P8 Canada
Courriel : mballo.cherif@courrier.uqam.ca

Résumé. Le critère de découpage binaire de Kolmogorov-Smirnov a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des variables continues. Nos travaux antérieurs nous ont permis de l'étendre dans le cas où les objets destinés à être classés par un arbre de décision sont décrits par des variables de type intervalle et diagramme ((Mballo et Diday, 2004), (Mballo et al., 2004)) en adoptant une affectation pure. Dans cet article, nous proposons une méthode permettant d'affecter une donnée à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. Cette approche d'affectation est basée sur des poids et tient compte de la position de la donnée à classer par rapport à celle seuil de coupure.

1 Introduction

Avec l'avènement de l'analyse des données symboliques (Bock et Diday, 2000), on assiste à la mise au point de méthodes de construction d'arbres de décision sur des données de type intervalle et diagramme ((Périnel, 1996), (Aboa, 2002), (Vrac, 2002), (Limam, 2005)). Pour construire l'arbre de décision, ces auteurs utilisent l'entropie, le critère de Gini, le gain ratio et le likelihood comme critère d'évaluation de la qualité d'une coupure.

Dans cet article, nous nous intéressons au 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 variables continues. Il a été également exploré quelques années plutard par (Utgoff et Clouse, 1996) sur ce même type de données. (Asseraf, 1998) s'est intéressé à son extension aux données qualitatives. Il présente un bon pouvoir discriminant sur des données classiques. Dans ((Mballo et Diday, 2004), (Mballo et al., 2004)), nous l'avons étendu aux données de type intervalle et diagramme mais dans cette approche, une donnée est entièrement affectée à un nœud (affectation pure). Comme ce critère nécessite un ordre des données, nous présentons tout d'abord quelques méthodes pour ordonner des intervalles (Diday et al., 2003) et des diagramme. La possibilité d'estimer la fonction de répartition théorique par celle empirique nous permet d'adapter ce critère aux données de type intervalle et diagramme. Nous présentons à la section 4 une méthode permettant d'affecter une donnée à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. La motivation de cette approche d'affectation est de prendre en compte le positionnement de la donnée à classer par rapport à la donnée seuil de coupure en définissant des poids. Des exemples illustrant cette approche d'affectation sont également présentés.

2 Ordonner des données de type intervalle et diagramme

Nous présentons dans cette section des méthodes permettant d'ordonner des données de type intervalle et diagramme dans le but d'examiner l'adaptation du critère KS de construction d'arbres binaires de décision à ce type de données.

2.1 Ordonner des intervalles

Soit $\Im$ l'ensemble des intervalles fermés bornés de $\Re$ (ensemble des nombres réels) : $x \in \Im$, on note $x = [i(x), s(x)]$. Différentes méthodes permettent d'ordonner des intervalles selon leur positionnement (Diday et al., 2003). Un ordre d'intervalles est une relation d'ordre partiel, c'est-à-dire une relation réflexive, antisymétrique et transitive. On peut considérer la version stricte de l'ordre partiel en imposant l'antiréflexivité, mais la relation obtenue par ajout de la diagonale est bien réflexive, antisymétrique et transitive.

2.1.1 Intervalles disjoints

Désignons par $x \prec_D y$ pour indiquer que *x* est « *strictement avant* » *y* où *x* et *y* sont deux intervalles fermés bornés disjoints. Nous utilisons le qualificatif « *strictement* » pour faire ressortir le fait que les intervalles sont disjoints. La relation « $\prec_D$ » se définit par : $x \prec_D y \Leftrightarrow s(x) < i(y)$. Cette relation « $\prec_D$ » est antiréflexive, antisymétrique et transitive et définit ainsi un ordre strict d'intervalles sur l'ensemble des intervalles fermés bornés disjoints.

2.1.2 Intervalles non disjoints

Par analogie au cas disjoint, nous utilisons le qualificatif « *presque* » pour faire ressortir le fait que les intervalles sont non disjoints. Soient *x* et *y* deux intervalles non disjoints. Suivant leur positionnement, nous distinguons deux cas :

– **Ordonner par la borne inférieure** : désignons par $x \prec_I y$ pour indiquer que *x* est « *presque avant* » *y*. Nous distinguons deux cas: si les deux bornes inférieures sont égales, alors l'ordre est déterminé par la position des bornes supérieures et si les deux bornes inférieures sont différentes, alors l'ordre est déterminé par la position de ces bornes inférieures. Nous définissons la relation « $\prec_I$ » par la formule suivante :

$$x \prec_I y \Leftrightarrow \begin{cases} i(x) < i(y) & \text{si } i(x) \neq i(y) \\ s(x) < s(y) & \text{sinon} \end{cases}$$

– **Ordonner par la borne supérieure** : désignons par $x \prec_S y$ pour indiquer que *y* est « *presque après* » *x*. Nous distinguons deux cas comme précédemment : si les deux bornes supérieures sont égales, alors l'ordre est déterminé par la position des bornes inférieures et si

les deux bornes supérieures sont différentes, alors l'ordre est déterminé par la position de ces bornes supérieures. Nous définissons la relation « $\prec_S$ » par la formule suivante :

$$x \prec_S y \Leftrightarrow \begin{cases} s(x) < s(y) & \text{si } s(x) \neq s(y) \\ i(x) < i(y) & \text{sinon} \end{cases}$$

Chacune des deux relations « $\prec_I$ » et « $\prec_S$ » est antiréflexive, antisymétrique et transitive et définit un ordre strict d'intervalles sur $\Im$.

Une autre approche consisterait à ordonner les intervalles par le centre ou la longueur.

2.1 Ordonner des diagrammes

On appelle diagramme un ensemble fini de modalités pondérées (ordonnées ou pas) ou un ensemble fini d'intervalles disjoints pondérés. Une variable est dite de type diagramme si la valeur prise par chaque individu de la population est un diagramme (« diagrammes à bandes ou histogrammes » ou « diagramme à bâtons »). Le domaine d'observations d'une telle variable est un ensemble fini de diagrammes. Ce sont les variables modales dans (Bock et Diday, 2000), mais nous avons pris l'appellation « diagrammes » dans le cadre de cet article à cause du traitement que nous envisageons faire sur ce type de données. Par exemple, au niveau de l'étude de l'ordre, nous utiliserons des alternatives consistant à introduire des poids au niveau des modalités, ce qui n'est pas possible avec les variables modales dans (Bock et Diday, 2000). Mais du point de vue syntaxique et sémantique, c'est le même type de données. C'est la nature des modalités qui indique les appellations « diagrammes à bandes ou histogrammes » ou « diagramme à bâtons ». Pour une variable de ce type de données, tous les objets ont les mêmes modalités. Par exemple, pour une variable ayant q modalités notées $(m_1, m_2, ..., m_q)$, la description d'un objet est $\left(m_1(h_1), m_2(h_2), ..., m_q(h_q)\right)$ où $h_1, h_2, ..., h_q$ sont respectivement les valeurs prises aux modalités $m_1, m_2, ..., m_q$. Dans le cas où les diagrammes sont normalisés, les hauteurs vérifient : $0 \leq h_t \leq 1 \ \forall t = 1,2,...,q$ et $\sum_{t=1}^{q} h_t = 1.$ Les descriptions des objets diffèrent selon les valeurs prises au niveau des modalités. Nous nous intéressons uniquement à ces valeurs. Nous confondrons dans la suite ces deux appellations (« histogramme ou diagramme à bandes » et « diagramme à bâtons ») en une seule appellation « diagramme » car nous les traiterons de la même manière concernant l'ordre. La description d'un objet est alors simplement notée par $\left(h_1, h_2, ..., h_q\right)$. Soient $H = \left(h_1, h_2, ..., h_q\right)$ et $G = \left(g_1, g_2, ..., g_q\right)$ deux diagrammes. Désignons par « $H \prec G$ » pour indiquer que H est « *avant* » (ou « *inférieur à* ») G. Pour ordonner ce type de données, nous utilisons les caractéristiques de position et de dispersion d'une distribution et l'ordre lexicographique.

– Ordonner par un paramètre

Le principe consiste à calculer le paramètre (moyenne, médiane, écart-type, mode, étendue) pour chaque diagramme et d'ordonner les diagrammes en fonction de l'ordre des valeurs obtenues du paramètre. Désignons par « $H \prec_{Pa} G$ » pour indiquer que H est « *avant* » (ou « *inférieur à* ») G en ordonnant par le paramètre « Pa », alors on a :

$$H \prec_{Pa} G \Leftrightarrow Pa(H) \leq Pa(G)$$

La relation « $\prec_{Pa}$ » est un préordre total.

Pour calculer la moyenne, nous utilisons la méthode suivante : soient $p_1, p_2, ..., p_q$ les poids attribués respectivement aux modalités $m_1, m_2, ..., m_q$ de la variable diagramme considérée (ces poids varient selon l'importance accordée à une modalité), la moyenne d'un diagramme H est $\frac{\sum_{t=1}^{q} p_t \times h_t}{\sum_{t=1}^{q} p_t}$. Cette moyenne est aussi appelée moyenne pondérée ou « ordered weighted average » selon la terminologie Anglo-saxonne.

– Ordre lexicographique

Nous pouvons aussi ordonner des diagrammes par l'ordre lexicographique :

$$H \prec_{Lex} G \Leftrightarrow [\exists s \in \{1,2,...,q\} / h_s < g_s \text{ et } \forall r \in \{1,2,...,s-1\}, h_r = g_r]$$

3 Le critère de Kolmogorov-Smirnov sur des données de type de type intervalle et diagramme

Considérons une population théorique de n objets destinés à être classés par un arbre de décision. Ces objets sont décrits par p variables de type intervalle ou diagramme $X_1, X_2, ..., X_P$ (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 ou un ensemble fini de diagrammes. Le critère KS permet de séparer une population en deux sous-populations plus homogènes en se basant sur les deux fonctions de répartition des classes a priori (cas de deux classes) pour chaque variable explicative. Dans le cas où le nombre de classes a priori k est supérieur à 2, les fonctions de répartition sont induites par le regroupement de ces k classes en deux groupes appelés super classes par la méthode « twoing splitting process » (Breiman et al. 1984). Cette méthode est utilisée pour générer deux super classes G_1 et G_2 auxquelles sont associées deux fonctions de répartitions F_1 et F_2 d'une variable explicative. La fonction de répartition a pour rôle de compter toutes les valeurs inférieures à un certain seuil. Les deux fonctions de répartition

théoriques F_1 et F_2 ne sont pas connues en pratique. S'il n'y a pas d'ordre sur les observations, on ne peut pas estimer la fonction de répartition théorique F_i par la fonction de répartition empirique $\hat{F}_i$ (i = 1,2) comme dans le cas continu. Dans notre cas, nous pouvons faire cette estimation car on peut ordonner les données (intervalles ou diagrammes) et l'ensemble $\{y \in D_{X_j} / y \leq x\} \cap \{y \in D_{X_j} / y \in G_i\}$ est toujours fini en pratique. Selon l'ordre choisi pour ordonner les observations d'une variable explicative de type intervalle ou diagramme X_j, la fonction de répartition empirique $\hat{F}_i^j$ qui estime F_i^j en $x \in D_{X_j}$ est donnée par :

$$\hat{F}_i^j(x) = \frac{Cardinal\left(\{y \in D_{X_j} / y \leq x\} \cap \{y \in D_{X_j} / y \in G_i\}\right)}{Cardinal\left(\{y \in D_{X_j} / y \in G_i\}\right)}$$

Ce sont les proportions réelles des observations pour chaque variable explicative X_j relative à une classe a priori (ou super classe). Ainsi, le critère KS est défini par :

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

C'est une extension naturelle du critère KS, seulement l'argument sélectionné pour le seuil de coupure est ici une donnée qui n'est pas un réel comme dans le cas classique, mais un intervalle ou un diagramme selon la nature de la variable explicative sélectionnée X_{j^*}.
Comme dans le cas de données continues, on peut utiliser toutes les autres étapes communes à tout type de variable pour construire l'arbre de décision.

Les données de type intervalle ou diagramme pouvant être ordonnées de différentes façons, la question à se poser est de savoir quel ordre utiliser pour construire l'arbre de décision. Dans (Mballo et Diday, 2006), (Mballo, 2005), (Mballo et Diday, 2005), (Mballo et Diday, 2004) et (Mballo et al., 2004), nous avons exploré séparément les différents ordres de chaque type de données en construisant un arbre pour chaque ordre. Dans cet article, nous utilisons l'approche suivante : *à chaque nœud non terminal pendant le processus de développement de l'arbre, toutes les méthodes pour ordonner les données sont examinées et celle qui donne la meilleure coupe en termes d'homogénéité des nœuds fils générés est retenue*.

4 Affectation pondérée

L'approche de construction d'arbres de décision par le critère KS utilisée dans ((Mballo et Diday, 2004), (Mballo et al., 2004)) affecte entièrement un individu à un nœud de l'arbre de décision quelque soit le positionnement de sa description par rapport à la donnée seuil de coupure. Nous proposons dans cette section une approche permettant d'affecter une donnée à la fois aux deux nœuds fils générés par le partitionnement d'un nœud non terminal. Cette approche a pour but de prendre en compte le positionnement de la donnée à classer par rapport à celle seuil de coupure en définissant des poids à gauche et à droite.

4.1 Cas des données de type intervalle

Considérons un nœud non terminal contenant des intervalles destinés à être classés en utilisant le critère KS comme critère d'évaluation de la qualité d'une coupure. Supposons que l'ordre sélectionné à ce nœud est celui par la borne inférieure par exemple. Indiquons par $c^* = [i(c^*), s(c^*)]$ l'intervalle seuil de coupure. Les intervalles de ce nœud peuvent alors se représenter comme l'indique la figure (FIG. 1) où tous les intervalles indiqués par $x = [i(x), s(x)]$ sont les intervalles à classer.

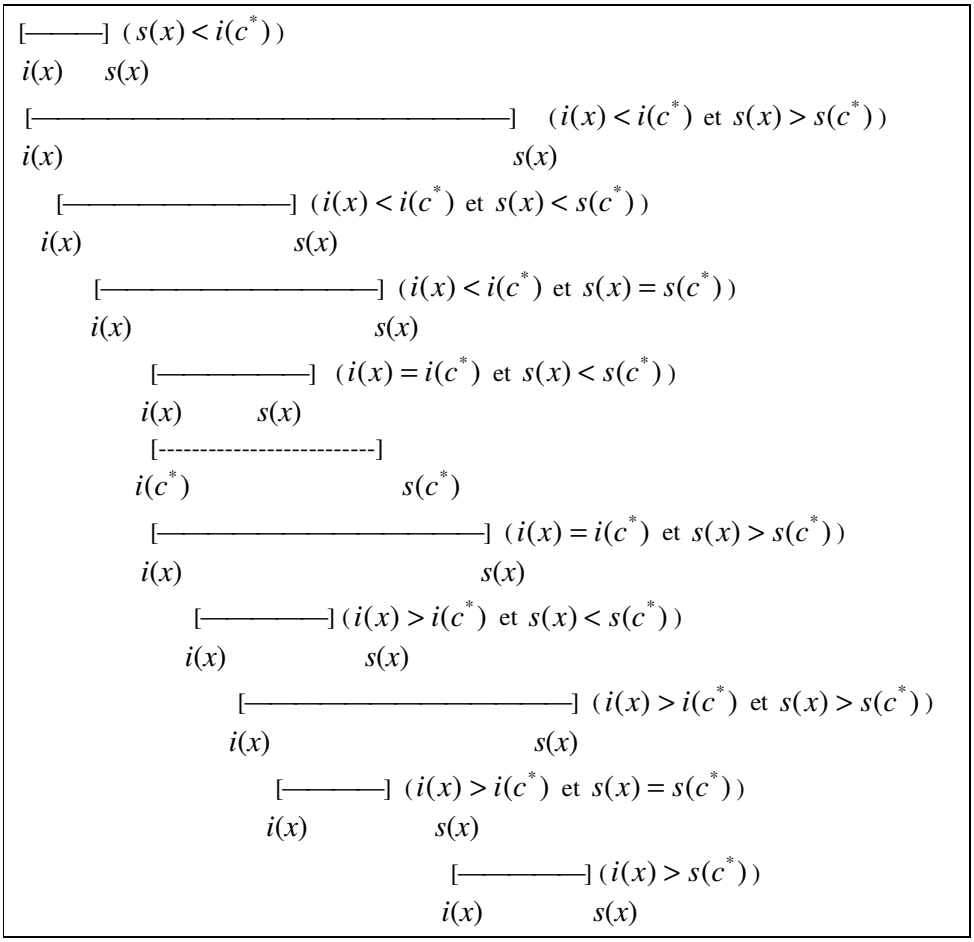

FIG. 1 – *Positionnement d'intervalles ordonnés par la borne inférieure*

Pour une affectation pure, tous les intervalles « *avant* » l'intervalle seuil c^* sont affectés au nœud fils gauche et le reste au nœud fils droit. L'affectation pure traite de la même façon

un intervalle disjoint au seuil de coupure c^* et celui non disjoint avec ce seuil. Notre objectif est de proposer une méthode d'affectation prenant en compte le positionnement de chaque intervalle à classer par rapport à l'intervalle seuil de coupure. Le but est d'affecter l'intervalle à classer à la fois à tous les deux nœuds fils gauche et droit générés par le découpage d'un nœud non terminal suivant des proportions ou poids. Notons par $p_g(x)$ et $p_d(x)$ les poids à attribuer à un intervalle à classer x respectivement aux nœuds fils gauche et droit générés par la coupure de seuil c^*. Selon le positionnement de x et c^*, nous distinguons le cas où ces deux intervalles sont disjoints du cas où ils ne sont pas disjoints pour définir les poids $p_g(x)$ et $p_d(x)$ de l'intervalle à classer x.

4.1.1 Cas où les deux intervalles sont disjoints

Dans le cas où les deux intervalles x et c^* sont disjoints, nous avons deux cas : soit l'intervalle à classer est « *complètement avant* » l'intervalle seuil, soit il est « *complètement après* » l'intervalle seuil (au sens de l'ordre sélectionné au nœud courant pour ordonner les intervalles).

- si l'intervalle à classer x est « *complètement avant* » (ou complètement à gauche) l'intervalle seuil c^* ($s(x) < i(c^*)$), alors $p_g(x) = 1$ et $p_d(x) = 0$. Dans ce cas, l'intervalle x est entièrement affecté au nœud fils gauche.
- si l'intervalle x est « *complètement après* » (ou complètement à droite) l'intervalle seuil c^* ($s(c^*) < i(x)$), alors $p_g(x) = 0$ et $p_d(x) = 1$. Dans ce cas, l'intervalle x est entièrement affecté au nœud fils droit.

4.1.2 Cas où les deux intervalles sont non disjoints

La figure (FIG. 2) donne les différents positionnements de deux intervalles non disjoints x et c^*.

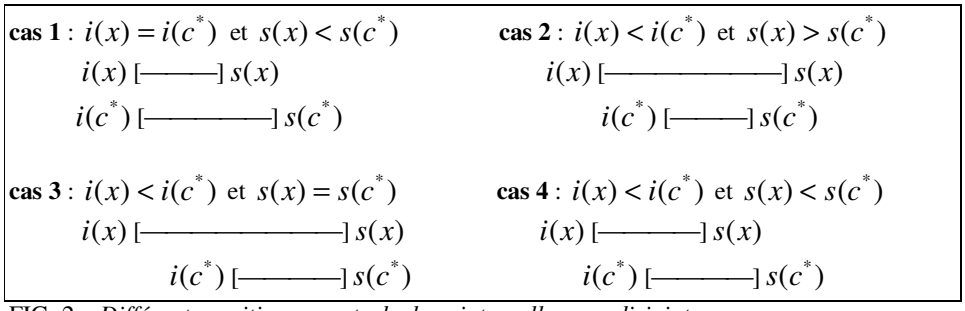

FIG. 2 – *Différents positionnements de deux intervalles non disjoints*

Posons : $E(x,c^*) = \max(s(x),s(c^*)) - \min(i(x),i(c^*))$ la longueur de l'étendue des deux intervalles x et c^* et $I(x,c^*) = \min(s(x),s(c^*)) - \max(i(x),i(c^*))$ celle de leur intersection. Nous appelons partie « propre » le débordement de l'un des intervalles. Pour deux intervalles non disjoints, nous avons donc deux parties « propres » car on peut avoir un débordement à gauche ou à droite (FIG. 3).

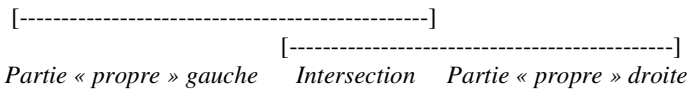

FIG. 3 – *Différentes parties de deux intervalles non disjoints*

Dans le cas où les deux bornes inférieures (respectivement supérieures) des deux intervalles x et c^* sont égales, alors la partie « propre » à gauche (respectivement à droite) est un point. Comme nous travaillons avec des intervalles fermés bornés, un point peut s'écrire sous la forme d'un intervalle fermé borné où les deux bornes inférieure et supérieure sont égales. La longueur d'une partie réduite à un point est donc égale à zéro. La partie « propre » à gauche a pour longueur $g(x,c^*) = [\max(i(x),i(c^*)) - \min(i(x),i(c^*))]$ et celle à droite a pour longueur $d(x,c^*) = [\max(s(x),s(c^*)) - \min(s(x),s(c^*))]$. Pour définir les deux poids $p_g(x)$ et $p_d(x)$ à attribuer à l'intervalle à classer x à gauche et à droite respectivement, nous partons du principe que la longueur de l'intersection des deux intervalles x et c^* doit être partagée « équitablement » à droite et à gauche. La longueur de la partie « propre » gauche (respectivement droite) revient au poids attribué à gauche (respectivement droite). En utilisant les notations précédentes, on montre facilement la relation suivante :

$$E(x,c^*) = [g(x,c^*) + \frac{I(x,c^*)}{2}] + [d(x,c^*) + \frac{I(x,c^*)}{2}]$$

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

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

Remarque 1:

- Dans chacun des cas disjoints et non disjoints, on a :

$0 \le p_g(x) \le 1$; $0 \le p_d(x) \le 1$ et $p_g(x) + p_d(x) = 1$

- L'intervalle seuil de coupure est affecté à gauche et à droite avec le même poids égal à 0.5 $(p_g(c^*) = p_d(c^*) = 0.5)$.

4.2 Cas des données de type diagramme

Supposons maintenant que la variable explicative la plus discriminante à un nœud non terminal est une variable de type diagramme X_{j^*}. Soit H^* le diagramme seuil sélectionné pour la coupure et soit H un diagramme à classer. Soit q le nombre de modalités de la variable explicative X_{j^*} et $m_1, m_2, ..., m_q$ ses modalités. Tout diagramme H est alors de la forme $H = (h_1, h_2, ..., h_q)$ où h_t est la valeur prise à la modalité m_t pour tout $t \in \{1,2,...,q\}$. Les diagrammes étant normalisés, nous avons : $0 \leq h_t \leq 1$ pour tout $t \in \{1,2,...,q\}$ et $\sum_{t=1}^{q} h_t = 1$. Le diagramme seuil de coupure $H^* = (h_1^*, h_2^*, ..., h_q^*)$. Désignons par $p_g(H)$ et $p_d(H)$ les poids à attribuer au diagramme à classer H aux nœuds fils gauche et droit respectivement. Le but consiste à définir ces poids en fonction du diagramme seuil de coupure H^* et de la méthode d'ordre sélectionnée à cette coupure pour ordonner les diagrammes. Nous avons vu à la section 2 que nous pouvons ordonner des diagrammes de plusieurs façons : moyenne, médiane, écart-type, mode, étendue et l'ordre lexicographique. La moyenne (respectivement l'écart-type et l'étendue) d'un diagramme ne correspond pas à une valeur prise à une modalité de ce diagramme contrairement au mode. Dans le cas où le nombre de modalités est impair, la médiane correspond aussi à une valeur d'une modalité (la valeur centrale des valeurs ordonnées des modalités). L'ordre lexicographique est à son tour déterminé par les valeurs prises au niveau des modalités. Comme notre objectif est de définir les poids $p_g(H)$ et $p_d(H)$ en fonction de la méthode d'ordre et des valeurs prises au niveau des modalités, nous nous intéressons aux méthodes d'ordre par la médiane, le mode et l'ordre lexicographique.

4.2.1 Les diagrammes sont ordonnés par la médiane ou le mode

Supposons que l'ordre sélectionné au niveau du nœud courant non terminal donnant le seuil de coupure H^* est celui par la médiane ou le mode. Soit m^* la modalité du diagramme seuil H^* correspondant à ce paramètre (médiane ou mode). Pour chaque diagramme à classer H, nous définissons les poids $p_g(H)$ et $p_d(H)$ de la façon suivante : $p_g(H)$ *est la somme de toutes les valeurs des modalités jusqu'à la modalité* m^* *et* $p_d(H)$ *est la somme des valeurs des modalités restantes*.

4.2.2 Les diagrammes sont ordonnés par l'ordre lexicographique

Supposons maintenant que c'est l'ordre lexicographique qui a été sélectionné à cette étape courante du processus de partitionnement récursif. Pour les deux diagrammes H et H^*, soit H est « *avant* » H^*, soit H^* est « *avant* » H. Si H est « *avant* » H^*,

$\exists s \in \{1,2,...,q\} / h_s < h_s^*$ et $\forall r \in \{1,2,...,s-1\}, h_r = h_r^*$. Si H^* est « *avant* » H, $\exists s \in \{1,2,...,q\} / h_s^* < h_s$ et $\forall r \in \{1,2,...,s-1\}, h_r^* = h_r$. Soit m^* la modalité du diagramme seuil H^* correspondant à la valeur h_r^*. Nous définissons les poids $p_g(H)$ et $p_d(H)$ de la façon suivante : $p_g(H)$ *est la somme de toutes les valeurs des modalités jusqu'à la modalité* m^* *et* $p_d(H)$ *est la somme des valeurs des modalités restantes*.

Remarque 2 : Comme les diagrammes sont normalisés, nous avons : $0 \leq p_g(H) \leq 1$; $0 \leq p_d(H) \leq 1$ et $p_g(H) + p_d(H) = 1$.

Après avoir exposé cette méthode d'affectation pondérée sur des données de type intervalle et diagramme, nous présentons une méthode permettant d'obtenir les effectifs des nœuds terminaux à l'arrêt du processus de développement de l'arbre.

4.3 Calcul des effectifs des nœuds terminaux

Comme en segmentation standard, il faut trouver un moyen de calculer les effectifs de chaque nœud et pour chaque classe. Au niveau de chaque nœud, le poids d'un objet est calculé comme étant le produit de ces poids du chemin qu'il a suivi de la racine à ce nœud. A un nœud donné, l'effectif d'un objet est alors un nombre réel supérieur ou égal à zéro car produit de poids compris entre 0 et 1. L'effectif d'une classe à un nœud est calculé comme étant la somme des effectifs de tous les objets appartenant à cette classe. L'effectif du nœud est alors la somme des effectifs des classes. Avec cette approche, on obtient les effectifs des nœuds terminaux. Pour arrêter le développement de l'arbre, les critères d'arrêt utilisés dans le cas d'une affectation pure sont aussi valables. Par exemple, on peut fixer un effectif minimum d'un nœud terminal. La classe d'un nœud terminal est celle ayant le plus grand effectif. Une fois que nous avons l'effectif de chaque nœud terminal, la précision peut alors être calculée comme dans le cas d'une affectation pure. Les règles de décision sont aussi notées comme en affectation pure car une classe est attribuée à chaque nœud terminal de l'arbre de décision.

Nous présentons maintenant des exemples pour illustrer le mécanisme de construction d'un arbre de décision en utilisant cette méthode d'affectation pondérée par le critère KS.

4.4 Exemples

4.4.1 Exemple sur des données de type intervalle

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 (Ciampi et al., 2000). La figure (FIG. 4) présente l'arbre de décision obtenu par l'approche d'affectation pondérée exposée ci-dessus. Le paramètre retenu pour arrêter le développement de l'arbre est celui d'un effectif minimum d'un nœud, il est fixé à 3 (donc tout nœud ayant un effectif strictement inférieur à 6 ne sera pas partitionné).

	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}	[1,3]	[1.5,2]	3

TAB. 1 – *Données de type intervalle*

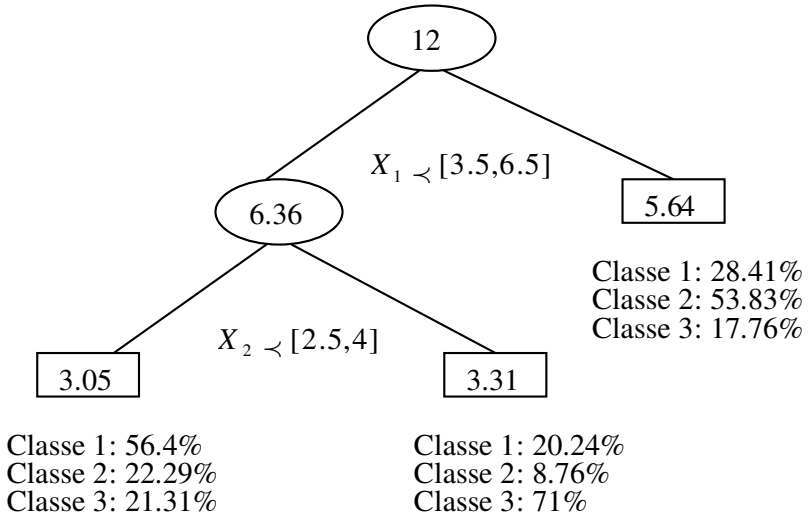

FIG. 4 – *Arbre de décision obtenu par l'affectation pondérée sur le tableau (TAB. 1)*

4.4.2 Exemple sur des données de type diagramme

Le tableau (TAB. 2) est un extrait de la base de données « *développement des pays du monde* »[1]. C'est une base de données établie en 2000 et constituée à partir d'indicateurs de développement acceptés par la banque mondiale, le fond monétaire international et les Nations Unies. Au début, il y avait cinquante pays répertoriés sur les cinq continents. Ces pays sont divisés en deux catégories économiques : pays développés (catégorie 1) et pays en développement (catégorie 2). Le croisement de ces deux catégories avec les cinq continents donne dix concepts. Ces concepts ou individus de second ordre sont les individus de cet exemple. Par exemple, les concepts DEUR et SDEUR signifient respectivement « pays développés » et « pays en développement » en Europe. Ces concepts sont décrits par deux variables explicatives de type diagramme X_1 : *Nature_Régime* et X_2 : *Code_Religion*. La variable classe est Y : *Nom_catégorie*. La variable explicative X_1 a six modalités : *Présidentiel*, *Parlementaire*, *Présidentiel-Parlementaire*, *Monarchie*, *Constitutionnel* et *Populaire*. La variable explicative X_2 a dix modalités : *Catholicisme*, *Protestantisme*, *Athéisme*, *Anglicanisme*, *Animisme*, *Hindouisme*, *Islamisme*, *Judaïsme*, *Bouddhisme* et *Shintoïsme*. Au niveau du tableau de données, nous présentons seulement les valeurs prises au niveau des modalités. Par exemple, la description (0.75,0.25,0,0,0,0) de « SDAMQ » à la variable explicative X_1 signifie que cet individu prend la valeur 0.75 à la modalité *Présidentiel*, 0.25 à la modalité *Parlementaire* et 0 pour toutes les autres modalités. Cette description pouvait être notée par : (*Présidentiel* (0.75),*Parlementaire* (0.25),*Présidentiel-Parlementaire* (0),*Monarchie* (0),*Constitutionnel* (0),*Populaire* (0)), mais nous avons adopté la notation (0.75,0.25,0,0,0,0) par soucis de simplification. Pour cette description, 75% des pays du concept « SDAMQ » sont de régime « *Présidentiel* » et 25% de régime « *Parlementaire* ». La figure (FIG. 5) présente l'arbre de décision obtenu par le critère KS sur ce tableau de données.

	X_1	X_2	Y
DAMQ	(0.85,0.14,0,0,0,0)	(0.71,0.28,0,0,0,0,0,0,0,0)	1
SDAMQ	(0.75,0.25,0,0,0,0)	(1,0,0,0,0,0,0,0,0,0)	2
DEUR	(0,0.83,0.16,0,0,0)	(0.16,0.66,0.16,0,0,0,0,0,0,0)	1
SDEUR	(0.16,0.83,0,0,0,0)	(1,0,0,0,0,0,0,0,0,0)	2
DOCE	(0,1,0,0,0,0)	(0,0.33,0,0.33,0.33,0,0,0,0,0)	1
SDOCE	(0,1,0,0,0,0)	(0,0,0,0,1,0,0,0,0,0)	2
DAFR	(0.4,0.2,0.2,0.2,0,0)	(0.2,0,0,0,0,0.2,0.6,0,0,0)	1
SDAFR	(0.83,0.16,0,0,0,0)	(0,0.16,0.16,0,0.16,0,0.5,0,0,0)	2
DASI	(0.2,0.4,0,0,0.2,0.2)	(0,0,0,0,0.2,0,0.2,0.2,0.2,0.2)	1
SDASI	(0.4,0.2,0,0,0.4,0)	(0.2,0,0,0,0,0.2,0.4,0,0.2,0)	2

TAB. 2 – *Données de type diagramme*

[1] Rapport de Stage (DESS Informatique Décisionnelle) de Ravelomanantsoa H., Université Paris Dauphine (2002), disponible à l'URL http://www.ceremade.dauphine.fr/%7Etouati/pays1.htm

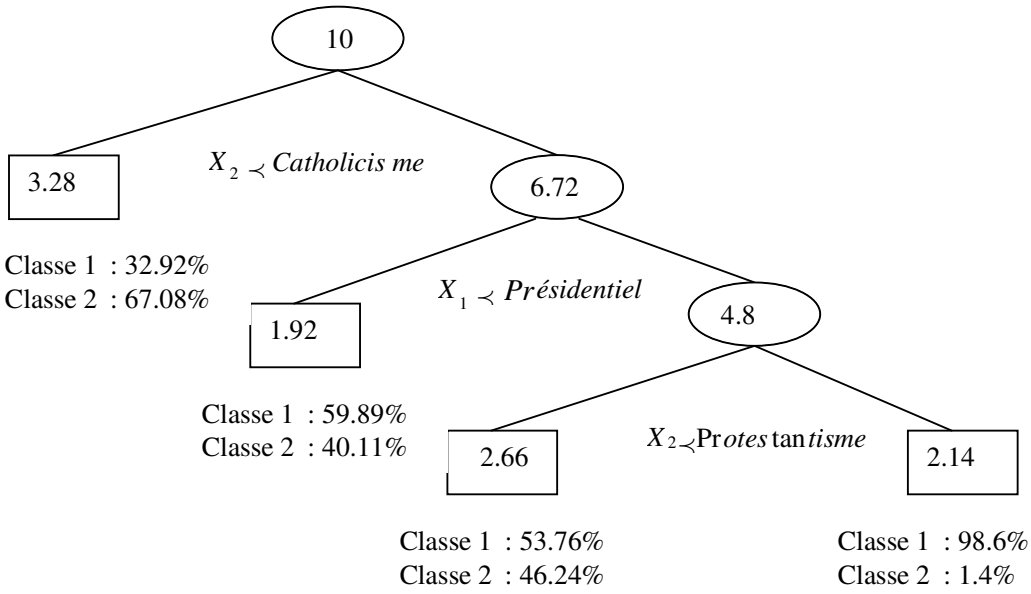

FIG. 5 – *Arbre de décision obtenu par l'affectation pondérée sur le tableau (TAB. 2)*

5 Conclusion et perspectives

Nous avons proposé une méthode d'affectation permettant de prendre en compte le positionnement de la description d'un objet à classer par rapport au seuil de coupure. Avec cette approche, chaque objet pourra se retrouver à plusieurs nœuds terminaux de l'arbre de décision suivant des proportions variées. Cette approche présente un avantage car elle permet d'affecter différemment une donnée selon sa position avec le seuil de coupure. La différence se fait surtout remarquer dans le cas des intervalles où les cas disjoints et non disjoints sont traités séparément, contrairement à une approche d'affectation pure qui les traite de la même façon.

Dans le cas de l'affectation pure ((Mballo et Diday, 2004), (Mballo et al., 2004)), le critère KS a été comparé aux critères de Gini et de l'entropie. Dans la suite, nous envisageons suivre la même voie pour cette méthode d'affectation pondérée. Nous projetons également d'examiner l'extension d'autres méthodes d'apprentissage à ce type de données, comme par exemple les réseaux Bayésiens et les SVM (Support Vector Machine). Pour les réseaux Bayésiens, le problème consiste à examiner comment définir des probabilités conditionnelles a posteriori sur des variables de type intervalle et diagramme. Dans le cadre des SVM, il s'agira de chercher à expliquer une variable classe à l'aide d'un vecteur de modèles.

Références

Aboa, J. P. Y. (2002) *Méthodes de segmentation sur un tableau de variables aléatoires*. Thèse de Doctorat, Spécialité Mathématiques Appliquées, Université Paris Dauphine.

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

Bock, H. H. et 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, et 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*, Lyon, France, 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, Spécialité Informatique ; Université Paris Dauphine.

Mballo, C. et E. Diday (2006) The criterion of Kolmogorov-Smirnov for binary decision tree: Application to interval valued variables; In « *Analysis of symbolic and spatial data : mining complex data structures* », Paula Brito & Monique Noirhomme-Fraiture (Guest Editors), *Intelligent Data Analysis*, Volume 10, Number 4/2006, pp 325-341.

Mballo, C. (2005) *Ordre, codage et extension du critère de Kolmogorov-Smirnov pour la segmentation de données symboliques*. Thèse de Doctorat en Informatique, Université Paris Dauphine, France.

Mballo, C. et E. Diday (2005) Arbres de décision sur des données de type intervalle : évaluation et comparaison ; *Revue des Nouvelles Technologies de l'Information (RNTI)*; Numéro E-3, pp 67-78.

Mballo, C. et E. Diday (2004) Kolmogorov-Smirnov for decision trees on interval and histogram variables; in "*Studies in classification, Data Analysis and Knowledge organization*: *Classification, Clustering and Data Mining Applications*", Part IV: *Symbolic Data Analysis*; editors: D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul; Springer; pp 341-350.

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

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

Utgoff, P. E. et J. A. Clouse (1996) A Kolmogorov-Smirnov metric for decision tree induction; *Technical report*, Volume 3, University of Massachusetts.

Vrac, M. (2002) *Analyse et modélisation de données probabilistes par décomposition de mélanges de copules et application à la climatologie.* Thèse de Doctorat, Université Paris Dauphine, Spécialité Mathématiques Appliquées.

Summary

The Kolmogorov-Smirnov binary splitting criterion was introduced by (Friedman, 1977) for binary partition on continuous variables. Our previous work allowed us to extend it in the case of interval and diagram variables ((Mballo and Diday, 2004), (Mballo et al. 2004)) by adopting a pure assignment of the object to be classified. In this paper, we propose an approach that consists in assigning an object at once to the two children nodes generated by the splitting of a non terminal node. This approach is based on weights and its main aim is to take account the position of the data to be classified with regard to the selected data for the cutting.