Segmentation de données de type intervalle, diagramme et taxonomique

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. L'objectif de cet article consiste à étendre le critère de découpage binaire de Kolmogorov-Smirnov aux données de type intervalle, diagramme et taxonomique. Ce critère nécessite un ordre des valeurs prises par les variables explicatives. Nous utilisons différentes méthodes pour ordonner ce type de données. Nous présentons le format des questions binaires et la description des nœuds terminaux pour chacun de ces types de données. Nous étudions également la précision de ce critère et nous le comparons aux critères de l'entropie et de Gini.

1 Introduction

Dans le domaine de la discrimination par arbre de décision binaire, les variables explicatives sont souvent quantitatives ou qualitatives. Le critère de découpage binaire de Kolmogorov-Smirnov (noté KS dans la suite) a été introduit par (Friedman, 1977) pour une partition binaire à expliquer sur des variables continues. Ce critère a été exploré quelques années plutard par (Utgoff et Clouse, 1996) pour construire des arbres de décision binaires sur ce même type de données. Il a été étendu aux données qualitatives classiques par (Asseraf, 1998).

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 de découpage. Dans cet article, nous nous intéressons au critère de découpage binaire KS. C'est un critère présentant un bon pouvoir discriminant sur des données classiques. Nous étudions son extension dans le cas où les objets destinés à être classés par un arbre de décision sont décrits par des variables de type intervalle, diagramme et taxonomique.

Nous présentons tout d'abord différentes méthodes pour ordonner des données de type intervalle, diagramme et taxonomique. Ensuite, nous étudions l'extension du critère KS à ce type de données pour la construction d'arbres binaires de décision. Nous présentons également le format des questions binaires et la description des nœuds terminaux de l'arbre de décision pour chaque type de données. Enfin, nous comparons les critères KS, Gini et entropie sur des données de type intervalle et diagramme.

2 Ordonner des données de type intervalle, diagramme et taxonomique

L'objectif de cette section est de proposer des méthodes permettant d'ordonner des données de type intervalle, diagramme et taxonomique 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

Désignons par $\Im$ l'ensemble des intervalles fermés bornés de $\Re$ (ensemble des nombres réels) : $\Im = \{[a,b] / a,b \in \Re; a \le b\}$. Pour $x \in \Im$, on note $s(x)$ sa borne supérieure et $i(x)$ sa borne inférieure : $x = [i(x), s(x)]$. Une variable est dite de type intervalle si la valeur prise par chaque objet est un élément de $\Im$. Le domaine d'observations d'une telle variable est un ensemble fini d'intervalles fermés bornés de nombres réels.

Des auteurs comme (Fishburn, 1985) et (Pirlot et al., 1997) se sont intéressés à ordonner des intervalles. Soient x et y deux intervalles fermés bornés : $x = [i(x), s(x)]$ et $y = [i(y), s(y)]$. Désignons par « $x \prec y$ » pour indiquer que l'intervalle x est « *avant* » l'intervalle y. Ces auteurs considèrent que « $x \prec y$ » si et seulement si $s(x) < i(y)$. Nous voyons que cet ordre s'applique plus aisément dans le cas où les intervalles sont disjoints. Cependant, le cas où les intervalles sont non disjoints a été abordé par (Tsoukias et The, 2001) dans le domaine de la modélisation des préférences et par (Gioia, 2001). Dans (Diday et al., 2003), nous avions fait une première exploration des approches que nous exposons ici pour ordonner des intervalles fermés bornés de $\Re$.

Le terme « *ordre d'intervalles* » (traduction de « *interval order* » selon la terminologie Anglo-saxonne), est un terme de la théorie des ordres partiels. 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. Un ordre partiel, au sens strict, est appelé « *ordre d'intervalles* » s'il vérifie la propriété suivante :

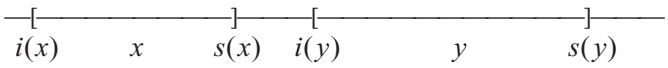
Propriété 1

Il existe une bijection (un isomorphisme) entre un ensemble E ordonné par cet ordre partiel et un ensemble d'intervalles fermés bornés Γ de la droite réelle ordonné par la relation de précédence « gauche-droite ». En d'autres termes, x et y étant deux intervalles, « $x \prec y$ » au sens de l'ordre d'intervalles dans E si et seulement si les images $[i(x), s(x)]$ et $[i(y), s(y)]$ de x et y sont telles que $s(x) < i(y)$ au sens de l'ordre des réels.

Étudier un ordre d'intervalles revient à examiner principalement deux situations : intervalles disjoints et intervalles non disjoints.

2.1.1 Intervalles disjoints

Soient x et y deux intervalles fermés bornés disjoints comme l'indique la configuration graphique ci-dessous sur une droite. Désignons par $x \prec_D y$ pour indiquer que l'intervalle x est « *strictement avant* » l'intervalle y. Nous utilisons le qualificatif « *strictement* » pour faire ressortir le fait que les intervalles sont disjoints.



La relation « $\prec_D$ » sur $\Im$ se définit par : $x \prec_D y \Leftrightarrow s(x) < i(y)$ (1)

La relation « $\prec_D$ » est antiréflexive, antisymétrique et transitive. Elle définit un ordre strict d'intervalles sur l'ensemble des intervalles fermés bornés disjoints. Elle vérifie la propriété 1.

2.1.2 Intervalles non disjoints

Soient x et y deux intervalles fermés bornés non disjoints. Suivant leur positionnement, nous pouvons distinguer les quatre cas suivants (FIG. 1) :

cas 1 : $i(x) = i(y)$ et $s(x) < s(y)$

$i(x)$ [———] $s(x)$

$i(y)$ [—————] $s(y)$

cas 2 : $i(x) < i(y)$ et $s(x) > s(y)$

$i(x)$ [—————] $s(x)$

$i(y)$ [———] $s(y)$

cas 3 : $i(x) < i(y)$ et $s(x) = s(y)$

$i(x)$ [—————] $s(x)$

$i(y)$ [————] $s(y)$

cas 4 : $i(x) < i(y)$ et $s(x) < s(y)$

$i(x)$ [————] $s(x)$

$i(y)$ [————] $s(y)$

FIG. 1 – *Différents positionnements de deux intervalles fermés bornés non disjoints*

La relation « $x \prec y \Leftrightarrow s(x) < i(y)$ » définie par (Fishburn, 1985) et (Pirlot et al., 1997) ne s'applique pas dans ce cas d'intervalles non disjoints. Pour ordonner ce type d'intervalles, nous proposons la définition suivante sur $\Im$:

Définition

Une relation binaire définie sur $\Im$ est un $\Im$-ordre strict si elle est antiréflexive, antisymétrique et transitive.

Un $\Im$-ordre strict est une relation fortement antisymétrique et transitive. Vus les différents cas de positionnement de deux intervalles non disjoints (FIG. 1), nous pouvons les ordonner soit par la borne inférieure, soit par la borne supérieure. Par analogie au cas

disjoint, nous utilisons le qualificatif « *presque* » pour faire ressortir le fait que les intervalles sont non disjoints.

– Ordonner par la borne inférieure

Désignons par $x \prec_I y$ pour indiquer que x est « *presque avant* » y. Nous distinguons deux cas selon le positionnement des bornes inférieures des deux intervalles : si elles sont différentes, alors leur position détermine l'ordre des intervalles; si elles sont égales, alors l'ordre des intervalles est déterminé par la position des bornes supérieures. Nous définissons la relation « $\prec_I$ » par la formule (2) suivante :

$$x \prec_I y \Leftrightarrow \begin{cases} i(x) < i(y) & \text{si } i(x) \neq i(y) \\ s(x) < s(y) & \text{sinon} \end{cases} \qquad (2)$$

La relation « $\prec_I$ » est un $\Im$-ordre strict total. Si x et y sont deux intervalles fermés bornés non disjoints tels que $x \prec_I y$, alors : $\exists t \in x / \forall t' \in y, t \leq t'$.

– Ordonner par la borne supérieure

Par analogie au cas précédent, désignons par $x \prec_S y$ pour indiquer que y est « *presque après* » x. Nous distinguons deux cas selon le positionnement des bornes supérieures des intervalles : si elles sont différentes, alors leur position détermine l'ordre des intervalles; si elles sont égales, alors l'ordre des intervalles est déterminé par la position des bornes inférieures. Nous définissons la relation « $\prec_S$ » par la formule (3) suivante :

$$x \prec_S y \Leftrightarrow \begin{cases} s(x) < s(y) & \text{si } s(x) \neq s(y) \\ i(x) < i(y) & \text{sinon} \end{cases} \qquad (3)$$

La relation « $\prec_S$ » est un $\Im$-ordre strict total. Si x et y sont deux intervalles fermés bornés non disjoints tels que $x \prec_S y$, alors : $\exists t' \in y / \forall t \in x, t' \geq t$.

Pour les intervalles non disjoints, on peut énoncer le résultat suivant :

Théorème

Les relations binaires de la borne inférieure et de la borne supérieure forment un ordre strict sur l'ensemble des intervalles fermés et bornées de $\Re$.

Analysons maintenant les différents cas de positionnement (FIG. 1) de deux intervalles non disjoints x et y pour étudier la différence des deux $\Im$-ordres stricts « $\prec_I$ » et « $\prec_S$ ». Le cas 2 nous montre que l'intervalle y est strictement inclus dans l'intervalle x : $\forall t' \in y, \exists t_1, t_2 \in x / t_1 < t' < t_2$. Pour le cas 1 (respectivement cas 3), nous avons une inclusion avec égalité des bornes inférieures (respectivement supérieures). L'intersection des deux intervalles du cas 4 est également non vide mais ne sont pas inclus. Par abus de langage, nous parlerons d'inclusion.

Étant donnés deux intervalles x et y, nous remarquons que, dans le cas d'inclusion non stricte (FIG. 1 : cas 1; 3 ; 4), si x est « *presque avant* » y ($x \prec_I y$), alors y est « *presque après* » x ($x \prec_S y$). Les relations « $\prec_I$ » et « $\prec_S$ » déterminent donc la même relation de « *précédence* » entre deux intervalles pour les cas 1 ; 3 ; 4. En revanche, pour le cas 2 (inclusion stricte), si x est « *presque avant* » y ($x \prec_I y$), alors x est aussi « *presque après* » y ($y \prec_S x$). C'est la principale différence entre les relations « $\prec_I$ » et « $\prec_S$ ».

Nous pouvons aussi ordonner les intervalles par le centre ou la longueur. Soient $c(x)$ le centre et $l(x)$ la longueur d'un intervalle x. Désignons par « $\prec_C$ » et « $\prec_L$ » l'ordre par le centre et la longueur, on a : $x \prec_C y \Leftrightarrow c(x) \leq c(y)$ et $x \prec_L y \Leftrightarrow l(x) \leq l(y)$. Chacune de ces deux relations est réflexive et transitive et définit un $\Im$-préordre total.

2.2 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 individus ont les mêmes modalités. Par exemple, pour une variable ayant q modalités notées $(m_1, m_2, ..., m_q)$, la description d'un individu 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 individus 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 individu est alors simplement notée par $(h_1, h_2, ..., h_q)$. Soient $H = (h_1, h_2, ..., h_q)$ et $G = (g_1, g_2, ..., g_q)$ 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 :

$$H \prec_{Pa} G \Leftrightarrow Pa(H) \leq Pa(G) \qquad (4)$$

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 $\dfrac{\sum_{t=1}^{q} p_t \times h_t}{\sum_{t=1}^{q} p_t}$. 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 :

$$H \prec_{Lex} G \Leftrightarrow [\exists s \in \{1,2,...,q\} / h_s < g_s \text{ et } \forall r \in \{1,2,...,s-1\}, h_r = g_r] \qquad (5)$$

2.3 Ordonner des données de type taxonomique

Une variable taxonomique est une application de l'ensemble des individus dans un ensemble de valeurs ordonnées totalement ou partiellement (Diday, 1998). Le domaine d'observations a une structuration hiérarchique. C'est une variable organisée en arbre exprimant plusieurs niveaux de généralité. Les feuilles n'ont pas toujours le même niveau. Les modalités d'un niveau m sont regroupées au niveau supérieur $(m+1)$. Considérons l'exemple de la variable taxonomique « *âge* » (FIG. 2). La racine correspond au nom de la

variable taxonomique (ce n'est pas une description d'un individu). On parlera de taxonomie binaire si chaque nœud a au plus deux fils. Dans le cas où un nœud peut avoir n fils avec $n > 2,$ on parlera de taxonomie n-aire. Pour ordonner ce type de données, nous utilisons le principe de parcours d'arbres et d'arbre binaire de recherche (Gondran et Minoux, 1985). Un parcours d'arbre est une façon d'ordonner les nœuds d'un arbre afin de les parcourir. Dans notre cas, il n'y aura pas de traitement de la racine car elle représente le nom de la variable (ce n'est pas une valeur à traiter). Nous partons du principe que les fils d'un nœud d'une variable taxonomique sont toujours disposés de gauche à droite selon un ordre croissant de préférence. Soient x et y deux modalités d'une variable taxonomique. Désignons par « $x \prec y$ » pour indiquer que x est « *avant* » (ou « *inférieure à* ») y. Examinons d'abord les principes de numérotation en profondeur et en largeur avant de définir la relation « $\prec$ ».

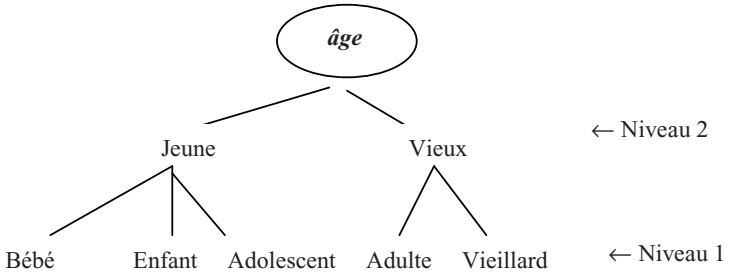


FIG. 2 – *Variable « âge »*

2.3.1 Numérotation en profondeur

Nous considérons tout d'abord le cas où la taxonomie est binaire. Le parcours de la variable taxonomique, dans le but de trouver un ordre croissant de ses modalités, consistera à parcourir récursivement les sous arbres gauche et droit à partir de la racine de la façon suivante : traiter d'abord le sous-arbre (ou nœud) gauche ; traiter la racine de ce sous-arbre (ou père de ce nœud) gauche et enfin traiter le sous-arbre (ou nœud) droit de la même façon. Ce type de parcours est connu sous le nom de « *parcours infixé* » en théorie de parcours d'arbres. Le traitement se fera en respectant la propriété suivante des arbres binaires de recherche (Gondran et Minoux, 1985) :

Propriété 2
Pour tout nœud, les numéros apparaissant dans le sous-arbre gauche sont strictement inférieurs à son numéro et ceux du sous-arbre droit strictement supérieurs.

Le principe de numérotation des nœuds est le suivant : on part de la racine principale (sans la numéroter car c'est le nom de la variable) et on suit la branche gauche jusqu'au dernier nœud n'ayant pas de fils, on le numérote (premier nœud numéroté). On numérote le nœud père correspondant. Ensuite on passe à la partie droite en appliquant le même principe. Le traitement est du type : *gauche–père–droit*. L'ordre ainsi obtenu est un ordre croissant.

Dans le cas d'une taxonomie n-aire, le principe consiste à choisir, parmi les n fils d'un nœud ($n > 2$), les p premiers à partir de la gauche comme groupe de fils gauche (les $(n - p)$ restants forment ainsi le groupe de fils droit) et d'appliquer le même déroulement que précédemment (les numéros d'un groupe sont consécutifs).

2.3.2 Numérotation par niveau

Nous définissons des niveaux de la taxonomie à partir du bas (FIG. 2). Le niveau correspondant à la racine n'est pas numéroté du fait que la racine n'intervient pas dans le traitement. Le principe consiste à numéroter tous les nœuds d'un niveau (en commençant toujours par le fils gauche) avant de passer au niveau supérieur. Ce principe de numérotation est aussi connu sous l'appellation de numérotation en largeur. Dans le cas d'une taxonomie n-aire, on applique le même principe énoncé à la section 2.3.1.

2.3.3 Relation d'ordre sur des données de type taxonomique

Le principe de numérotation exposé aux sections 2.3.1 et 2.3.2 définit une bijection assignant un unique numéro à chaque nœud (Castillo et al., 1997). Chaque nœud correspond à une modalité de la variable taxonomique et chaque description d'un objet d'une population décrite par une telle variable est une modalité de cette variable. Nous pouvons définir la relation « $\prec$ » de la façon suivante :

$$x \prec y \Leftrightarrow \textit{numéro}(x) < \textit{numéro}(y) \qquad (6)$$

où $\textit{numéro}(a)$ désigne le numéro de la modalité a (numéro obtenu à la numérotation des nœuds de la variable taxonomique selon le principe opté). La relation « $\prec$ » est antiréflexive, antisymétrique et transitive. C'est un ordre total strict sur l'ensemble des modalités d'une variable taxonomique. On note « $\prec_P$ » l'ordre en profondeur et « $\prec_N$ » celui par niveau.

3 Le critère de Kolmogorov-Smirnov pour la segmentation de données de type intervalle, diagramme et taxonomique

Considérons une population théorique Ω de n objets (ou individus) $w_1, w_2, ..., w_n$ destinés à être classés à l'aide d'un arbre de décision. Au départ, Ω est muni d'une partition en k classes. Soient $K = \{1,2,...,k\}$ et Y la variable déterminant l'appartenance d'un objet à l'une de ces classes (variable à expliquer) : $\forall w \in \Omega,\ Y(w) = y \in K.$ Chaque objet est aussi décrit par un vecteur $X = (X_1, X_2, ..., X_P)$ de p variables explicatives, chacune pouvant être de type intervalle, diagramme ou taxonomique. Soit D_{X_j} l'espace d'observations de X_j, $j = 1,..., p.$ D_{X_j} est alors un ensemble fini d'intervalles fermés bornés de nombres réels ou de diagrammes ou de modalités selon que X_j est de type

intervalle, diagramme ou taxonomique. X_j définit une application de $\Omega \to D_{X_j}$ telle que $X_j(w) = x_j \quad \forall w \in \Omega$. Le vecteur X est alors défini de $\Omega \longrightarrow D_X = \prod_{j=1}^{p} D_{X_j}$ tel que $X(w) = x = (x_1, x_2, ..., x_p) \in D_X, \forall w \in \Omega$. Avec les applications X et Y, un objet $w \in \Omega$ est simplement modélisé par le couple $(x, y) \in D_X \times K$. Ce couple « description, classe » contient les seules informations susceptibles d'être connues sur les objets à classer au départ du processus de construction d'un arbre de décision. Toute autre information résultera d'hypothèses ou de connaissances globales sur les lois qui régissent les objets de Ω. Le processus de construction d'un arbre de décision doit aboutir à la construction d'une règle de décision permettant d'affecter un nouvel objet à l'une des classes a priori. Soit $\hat{Y}$ cette fonction permettant d'affecter une classe à un objet à l'arrêt de la construction de l'arbre : $\hat{Y} : D_X \to K$. L'idéal serait d'aboutir ou de s'approcher au mieux, à l'issue d'un processus d'identification de $\hat{Y}$, à la situation $Y = \hat{Y} \circ X$ où « $\circ$ » est la composition des applications.

3.1 Présentation du critère de Kolmogorov-Smirnov

Pour utiliser le critère KS sur des données continues, (Friedman, 1977) suppose que les coûts de mauvais classement et les probabilités a priori des classes sont inversement proportionnels. Cette hypothèse permet de ne pas estimer les probabilités a priori. Le principe proposé par (Friedman, 1977) consiste à considérer ce problème à k classes comme une succession de problèmes à deux classes (un contre le reste). Cette approche n'explore pas tous les sous-groupes : par exemple, si $k = 4$ classes notées 1, 2, 3 et 4, cette méthode n'explore pas les groupes comme « $\{1,2\}$ contre $\{3,4\}$ ». (Gustafson et al., 1980) proposent de faire une agrégation des classes en amont de la construction de l'arbre. (Celeux et Lechevallier, 1982) développent une méthode consistant à faire l'agrégation des classes à chaque nœud non terminal pendant la construction de l'arbre. Deux années plus tard, (Breiman et al., 1984) proposent une stratégie similaire à celle de (Celeux et Lechevallier, 1982) appelée « *twoing splitting process* » consistant à explorer toutes les paires de parties d'un ensemble de k classes en les regroupant en deux groupes C_1 et C_2 appelés « *super classes* ». Pendant le développement de l'arbre, à chaque nœud non terminal, tous les objets appartenant à C_t sont considérés comme étant tous de la classe t avec $t \in \{1,2\}$. Ces auteurs établissent des résultats (Propriété 3 ci-dessous) permettant de réduire considérablement le nombre de super classes à examiner : X_j étant une application de $\Omega \to D_{X_j}$, on peut définir l'image réciproque $X_j^{-1}(x)$ pour tout $x \in D_{X_j}$: $X_j^{-1}(x) = \{w \in \Omega / X_j(w) = x\}$.

Propriété 3

Pour tout $x \in D_{X_j}$, *l'étude de l'évènement* $X_j^{-1}(x)$ *est analogue à l'étude de son complémentaire* $\Omega - X_j^{-1}(x)$ *pour tout* $j \in \{1,2,...,p\}$.

Preuve

Soit $x, x' \in D_{X_j}$ tel que $X_j^{-1}(x') = \Omega - X_j^{-1}(x)$. Soit $l \in \{1,2,...,k\}$ une classe a priori représentée par le groupe noté G_l. Soit $\pi_l = P(Y = l)$ la probabilité a priori de la classe l. Désignons par F_l^j la fonction de répartition théorique de la classe l pour la variable explicative X_j. Pour tout $j \in \{1,2,...,p\}$ et pour tout $l \in \{1,2,...,k\}$, on a :

$$F_l^j(x) + F_l^j(x') = \frac{P(\{X_j = x\} \cap G_l)}{\pi_l} + \frac{P(\{X_j = x'\} \cap G_l)}{\pi_l}$$

$$= \frac{P(\{\{X_j = x\} \cap G_l\} \cup \{\{X_j = x'\} \cap G_l\})}{\pi_l}$$

$$= \frac{P(\{\{X_j = x\} \cap G_l\} \cup \{\{\Omega - \{X_j = x\}\} \cap G_l\})}{\pi_l}$$

$$= \frac{P(G_l)}{\pi_l} = \frac{\pi_l}{\pi_l} = 1, \text{ donc } F_l^j(x) + F_l^j(x') = 1 \ \forall \ l \in \{1,2,...,k\} \text{ et}$$

$\forall \ j \in \{1,2,...,p\}$. D'où $F_l^j(x) = 1 - F_l^j(x')$.

Avec l'approche « *twoing splitting process* », on sait que le cardinal de l'ensemble des parties de $K = \{1,2,...,k\}$ est 2^k. La propriété 3 permet de diviser ce cardinal par 2, on obtient 2^{k-1}. Or la paire formée de l'ensemble vide et de K n'a aucune importance pour notre étude. Finalement, le nombre de paires de super classes à examiner pendant le processus de construction de l'arbre de décision est de $2^{k-1} - 1$. Cette complexité exponentielle a été réduite à une complexité polynomiale par (Asseraf, 1998) sur des données qualitatives.

3.2 Extension du critère de Kolmogorov-Smirnov aux données de type intervalle, diagramme et taxonomique

Le critère KS est basé sur la fonction de répartition. Étudier l'extension de ce critère au cas où chaque objet est décrit par des variables explicatives de type intervalle, diagramme ou taxonomique revient à étudier l'extension de la notion de fonction de répartition à ce type de données. Nous pouvons ordonner ce type de données de différentes façons (section 2). Pour deux valeurs x_j et $x_j^{'}$ d'une variable explicative X_j correspondant aux descriptions de

deux objets w et w' de la population Ω, désignons par « $x_j \prec x_j^{'}$ » pour indiquer que x_j est « *avant* » (ou « *inférieure à* ») $x_j^{'}$. La fonction de répartition a pour rôle de compter toutes les valeurs « *avant* » (ou « *inférieures à* ») un certain seuil. C'est la probabilité de l'ensemble des réalisations d'une variable aléatoire X_j qui sont « *avant* » une réalisation x : $F_{X_j}(x) = P\left(\left\{w \in \Omega / X_j(w) \prec x\right\}\right)\ \forall x \in D_{X_j}$. Soient F_1^j et F_2^j les fonctions de répartition théoriques d'une variable explicative X_j associées respectivement aux super classes C_1 et C_2. Comme le plus souvent nous disposons d'un échantillon, ces fonctions de répartition ne sont pas connues. Il faut alors recourir à des estimations. S'il n'y a pas d'ordre sur les observations des variables explicatives, on ne peut pas estimer la fonction de répartition théorique F_t^j par la fonction de répartition empirique notée $\hat{F}_t^j$ comme dans le cas d'une variable continue. Dans notre cas, on peut faire cette estimation car nous pouvons ordonner les observations de chaque variable explicative X_j et l'ensemble $\left\{y \in D_{X_j} / y \prec x\right\} \bigcap \left\{y \in D_{X_j} / y \in C_t\right\}$ est toujours fini en pratique (la taille de l'ensemble d'apprentissage est toujours finie). A chaque nœud non terminal, la fonction de répartition empirique $\hat{F}_t^j$ qui estime F_t^j en $x \in D_{X_j}$, pour tout $j \in \{1,2,...,p\}$ et pour tout $t \in \{1,2\}$ est donnée par :

$$\hat{F}_t^j(x) = \frac{Cardinal\left(\left\{y \in D_{X_j} / y \prec x\right\} \bigcap \left\{y \in D_{X_j} / y \in C_t\right\}\right)}{Cardinal\left(\left\{y \in D_{X_j} / y \in C_t\right\}\right)} \qquad (7)$$

Ce sont les proportions réelles des observations qui sont « *avant* » l'observation x pour chaque variable explicative X_j relative à une super classe C_t à chaque nœud non terminal. En d'autres termes, c'est la probabilité conditionnelle de l'ensemble des observations qui sont « *avant* » l'observation x sachant la super classe C_t à chaque nœud non terminal. Ainsi, le critère KS est défini par :

$$KS = \sup_{x \in D_{X_j}} \left| \hat{F}_1^j(x) - \hat{F}_2^j(x) \right| \ \forall j = 1,2,...,p \qquad (8)$$

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. Selon la nature de la variable explicative sélectionnée X_{j^*}, le seuil de coupure c^* est soit un intervalle fermé borné de nombres réels, soit un diagramme, soit une modalité d'une variable taxonomique. 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, diagramme ou taxonomique pouvant être ordonnées de différentes façons, la question à se poser est de savoir quel ordre utiliser. Dans (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. Le résultat (Proposition ci-dessous) a été établi dans (Mballo, 2005) avec diverses applications sur des données de type intervalles (Mballo et Diday, 2006) :

Proposition

Les arbres de décision obtenus en combinant seulement deux des différents ordres d'un type de données (intervalle et diagramme) ne sont pas équivalents en général.

Dans cet article, nous utilisons l'approche suivante pour construire l'arbre de décision : *à chaque nœud non terminal pendant le processus de développement de l'arbre de décision, toutes les méthodes pour ordonner les valeurs des variables explicatives sont examinées et celle qui donne la meilleure coupe en termes d'homogénéité des nœuds fils générés est retenue*. Avec cette approche, toutes les méthodes pour ordonner les données des variables explicatives sont examinées simultanément à chaque nœud non terminal de l'arbre de décision.

Comme nous construisons l'arbre de décision sur des données non habituelles, étudions maintenant le format des questions binaires sur ce type de données.

3.3 Format des questions binaires

Soit X_{j^*} la variable explicative la plus discriminante à une étape du développement de l'arbre de décision et soit $c^* \in D_{X_{j^*}}$ le seuil de coupure. Au niveau de l'arbre, nous nous intéressons à la notation explicite de la question binaire $X_{j^*} \prec c^*$ permettant de partitionner la population d'un nœud non terminal en deux sous-populations plus homogènes. L'objectif est de tenir compte de l'ordre sélectionné pour la coupure du nœud.

3.3.1 Cas d'une variable intervalle

Dans le cas où X_{j^*} est de type intervalle, le seuil de coupure est alors un intervalle $c^* = [i(c^*), s(c^*)] \in D_{X_{j^*}}$. Si l'ordre sélectionné à l'étape courante est celui par le centre (respectivement, la longueur), nous notons la question binaire par $X_{j^*} \leq_C v$ (respectivement, $X_{j^*} \leq_L v$) où v désigne le centre (respectivement, la longueur) de c^*. Dans le cas où l'ordre sélectionné est celui par la borne inférieure (respectivement, supérieure), nous notons la question binaire au niveau de l'arbre de décision par $X_{j^*} \prec_I [i(c^*), s(c^*)]$ (respectivement, $X_{j^*} \prec_S [i(c^*), s(c^*)]$) où « $\prec_I$ » (respectivement, « $\prec_S$ ») désigne l'ordre des intervalles par la borne inférieure (respectivement supérieure). Au lieu de prendre la borne inférieure ou la borne supérieure de

l'intervalle par analogie aux ordres par le centre et la longueur, nous avons adopté cette notation mentionnant l'intervalle seuil de coupure pour les raisons suivantes :

- Supposons par exemple que c'est l'ordre par la borne inférieure qui est sélectionné. En cas d'égalité des bornes inférieures, nous savons que l'ordre est déterminé par la position des bornes supérieures (section 2.1.2). Considérons par exemple un intervalle à classer $x = [i(x), s(x)]$. Si la question binaire avec l'ordre par la borne inférieure était notée $X_{j^*} \leq_I i(c^*)$, dans le cas où $i(x) = i(c^*)$, l'algorithme affecterait x au nœud gauche de l'arbre sans comparer les bornes supérieures de x et c^*. Avec l'approche adoptée, en cas d'égalité des bornes inférieures, l'algorithme compare les bornes supérieures.

- Le même raisonnement s'applique dans le cas où l'ordre sélectionné est celui par la borne supérieure : en cas d'égalité des bornes supérieures, c'est la position des bornes inférieures qui déterminera le chemin de l'intervalle à classer.

3.3.2 Cas d'une variable diagramme

Soit q_{j^*} le nombre de modalités de la variable discriminante de type diagramme X_{j^*}, le seuil de coupure $c^* = (c_1, c_2, ..., c_q);\ 0 \leq c_t \leq 1 \quad \forall t \in \{1,2,...,q\}$ et $\sum_{t=1}^{q_{j^*}} c_t = 1$. Si l'ordre sélectionné au niveau du seuil de coupure est un ordre établi à partir d'un paramètre (moyenne, médiane, écart-type, mode ou étendue), alors nous notons la question binaire par $X_{j^*} \leq_{Pa} v^*$ où « Pa » indique le paramètre utilisé pour ordonner les diagrammes et v^* la valeur de ce paramètre au seuil de coupure c^*. Tous les diagrammes G tels que $Pa(G) \leq v^*$ sont affectés au nœud fils gauche, le reste au nœud fils droit. Avec l'ordre lexicographique, soit $G = (g_1, g_2, ..., g_q) \in D_{X_{j^*}}$ tel que $\exists s \in \{1,2,...,q_{j^*}\} / g_s < c_s$ et $\forall r \in \{1,2,...,s-1\}, g_r = c_r$. La question binaire au niveau de l'arbre de décision est $X_{j^*} \leq_{Lex} c_s$. Tous les diagrammes G tels que $g_s < c_s$ sont affectés au nœud fils gauche, le reste au nœud fils droit.

3.3.3 Cas d'une variable taxonomique

Notons « $\prec_P$ » l'ordre en profondeur et « $\prec_N$ » celui par niveau (section 2.3). Dans le cas où la question binaire porte sur une variable de type taxonomique X_{j^*}, le seuil de coupure c^* est une modalité de cette variable. Au niveau de l'arbre de décision, on note la question binaire par $X_{j^*} \prec_P c^*$ ou $X_{j^*} \prec_N c^*$ selon l'ordre sélectionné.

3.4 Descriptions des nœuds terminaux

A l'exception de la racine, chaque nœud de l'arbre est obtenu par une succession de nœuds reliés entre eux par des arêtes (nous appelons « arête » la ligne reliant deux nœuds). La description d'un nœud est la conjonction des descriptions fournies sur les arêtes appartenant au chemin menant à ce nœud. Si un nœud N est obtenu après r coupures, sa description d_N est $d_N = \bigwedge_{j=1}^{r} [X_j R_j d_j] \wedge c$ où c est la classe attribuée à ce nœud (classe majoritaire). Il devient alors possible de déterminer la description de chaque nœud terminal selon la nature des données. Désignons par X_j la variable explicative la plus discriminante sélectionnée à une étape donnée de la construction de l'arbre de décision. Si X_j est de type :

- intervalle, soit m (respectivement M) le minimum (respectivement maximum) des bornes inférieures (respectivement supérieures) de tous les intervalles contenus dans le nœud généré par la coupure, alors $[X_j R_j d_j] = [X_j \subseteq [m, M]]$;

- diagramme, $[X_j R_j d_j] = [X_j = d_j]$ où d_j est le « meilleur » diagramme du nœud généré par la coupure (« meilleur » au sens de l'ordre sélectionné à cette coupure) ;

- taxonomique, $[X_j R_j d_j] = [X_j \subseteq M_j]$ où M_j est l'ensemble des modalités des objets du nœud généré par la coupure.

Pour classer un nouvel objet à partir des règles de décision obtenues de l'arbre, il suffit de comparer la description de l'objet avec le seuil de coupure à chaque étape comme dans le cas de données continues (comparaison au sens de l'ordre sélectionné à l'étape courante).

3.5 Exemple illustratif sur des données de type intervalle

Nous présentons dans cette section un exemple de construction d'un arbre de décision à l'aide du critère KS sur des données de type intervalle. Le tableau (TAB. 1) 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. Les variables explicatives sont : X_1 : *Taux de croissance de la population* (par an) ; X_2 : *Superficie totale du pays* (en milliers de kilomètres carrés) ; X_3 : *Espérance de vie* (en années) et X_4 : *Taux d'analphabétisme des femmes*. La variable à

[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

prédire Y : *Nom_catégorie* a deux modalités (catégorie 1 et catégorie 2). En appliquant l'approche de construction d'arbre de décision présentée à la fin de la section 3.2, on obtient la figure (FIG. 3). Toutes les quatre méthodes d'ordre d'intervalles ont été examinées à chaque nœud non terminal pendant le développement de l'arbre mais seulement deux ont été sélectionnées. La lecture des règles de décision se fait comme dans le cas classique. La règle de décision correspondant au premier nœud terminal à partir de la gauche est (w désigne un individu) : Si $(X_3(w) \leq_C 72.15 \wedge X_2(w) \prec_I [2,1221]),$ alors $Y(w) = 1.$ Pour classer un nouveau individu w' tel que $X_2(w') = [a,b]$ et $X_3(w') = [c,d],$ on procède comme suit :

- pour la coupe $X_3 \leq_C 72.15,$ on compare 72.15 et $\frac{c+d}{2}$. Si $\frac{c+d}{2} \leq 72.15,$ alors w' prend la branche gauche, sinon, ce sera la branche droite.

- pour la coupe $X_2 \prec_I [2,1221],$ on compare les bornes inférieures des deux intervalles $[a,b]$ et $[2,1221].$ Si $a < 2,$ alors w' prend la branche gauche et si $a > 2,$ alors w' prend la branche droite (dans ces deux cas, la position des bornes supérieures n'est pas examinée). Si $a = 2,$ on compare les bornes supérieures de ces deux intervalles. Dans ce cas, si $b \leq 1221,$ w' prend la branche gauche et si $b > 1221,$ w' prend la branche droite.

La description du deuxième nœud terminal à partir de la gauche (section 3.4) est : $d_2 = [X_3 \subseteq [40.6,78.2]] \wedge [X_2 \subseteq [3,17075]] \wedge 2.$

Le principe est le même pour des données de type diagramme et taxonomique.

	X_1	X_2	X_3	X_4	Y
DAMQ	[0.9,1.8]	[757,9976]	[67.2,78.5]	[0,14.6]	1
SDAMQ	[0.8,2.6]	[11,1285]	[64,76]	[4.3,38.7]	2
DEUR	[0,0.5]	[41,547]	[77.2,79.3]	[0,0]	1
SDEUR	[-0.4,0.2]	[20,17075]	[66.1,78.2]	[0,0]	2
DOCE	[0.9,1.2]	[18,7682]	[68.4,78.7]	[0,9.2]	1
SDOCE	[0,2.7]	[3,462]	[55.6,70.5]	[21,52]	2
DAFR	[0.8,1.9]	[2,1221]	[56.7,70.7]	[15.4,63.9]	1
SDAFR	[0.9,2.9]	[196,945]	[40.6,52.3]	[15.3,72.3]	2
DASI	[0.3,2.9]	[20,9597]	[65.1,80.5]	[1,23.7]	1
SDASI	[1.4,2.6]	[185,3287]	[62.3,71.9]	[4.8,54.6]	2

TAB. 1 – *Données*

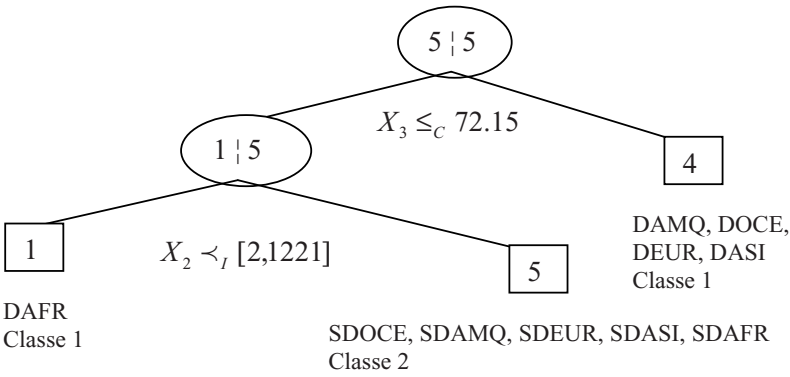


FIG. 3 – *Arbre de décision obtenu des données du tableau (TAB. 1)*

4 Précision et comparaison des critères KS, Gini et entropie

Dans cette section, nous examinons la précision du critère KS et nous le comparons aux critères de Gini et de l'entropie sur des données de type intervalle et diagramme. Nous axons notre attention sur le risque réel $\hat{R}_r$ (pourcentage d'objets mal classés de l'échantillon de test). Pour estimer ce paramètre, nous utilisons la technique « *hold-out method* ». Elle consiste à diviser aléatoirement la base de données en deux parties : l'apprentissage (les deux tiers de la base de données) sert à construire l'arbre et la partie restante (échantillon de test) est destinée à valider le résultat de l'apprentissage. Ces deux échantillons sont indépendants : ce qui est donné à l'un est retiré à l'autre. Les tableaux (TAB. 2 et TAB. 3) présentent les fichiers[2] utilisés pour estimer le risque réel. Ce sont des fichiers obtenus à partir de bases de données classiques à l'aide du module DB2SO (**D**ata **B**ase **to** **S**ymbolic **O**bjects) du logiciel SODAS[3] (**S**tatistical **O**fficial **D**ata **A**nalysis **S**ystem). Nos individus ne sont pas des individus au sens classique du terme (individus de premier ordre), mais des concepts (individus de second ordre). C'est pourquoi les bases de données ne sont pas assez volumineuses. Pour chaque tableau, la colonne « *Nb_cl* » indique le nombre de classes a priori de la variable classe, « *Répartition* » donne la répartition des objets par classe et « *Nb_var* » indique le nombre de variables explicatives. Par exemple, la notation (20 ;10) indique que 20 objets sont de la classe « 1 » et 10 de la classe « 2 » pour une variable classe ayant deux modalités.

Comme les échantillons de test et d'apprentissage sont indépendants, la précision de l'estimation ne dépend que du nombre d'objets n_t de l'ensemble de test et de $\hat{R}_r$. Pour n_t

[2] Disponibles librement à : http://www.ceremade.dauphine.fr/%7Etouati/exemples.htm
[3] Disponible librement à : http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm

assez grand (au-delà de 100 objets), l'intervalle de confiance de $\hat{R}_r$ à $x\%$ est donné par (Cornuéjols et Miclet, 2002) :

$$\left[\hat{R}_r - \varphi(x)\sqrt{\frac{\hat{R}_r}{n_t}(1-\hat{R}_r)}; \hat{R}_r + \varphi(x)\sqrt{\frac{\hat{R}_r}{n_t}(1-\hat{R}_r)} \right] \qquad (9)$$

où $\varphi(x)$ prend en particulier les valeurs du tableau (TAB. 4). La relation (9) signifie que la probabilité que le risque réel soit à l'intérieur de cet intervalle est supérieur à $x\%$. Les intervalles de confiance ne dépendent que de la taille n_t de l'échantillon de test. Nous présentons les intervalles de confiance des estimateurs des risques réels des bases de données ayant un ensemble de test de plus de 100 objets. Si on désigne par m_t le nombre d'objets mal classés de l'échantillon de test, alors $\hat{R}_r = \frac{m_t}{n_t}$. L'arrêt du processus de développement de l'arbre est basé sur un effectif minimum des nœuds terminaux. Cet effectif varie en fonction de la taille de l'ensemble d'apprentissage. Les figures (FIG. 4 et FIG. 5) présentent les résultats obtenus pour chaque critère sur les deux types de données intervalle et diagramme. Pour les intervalles de confiance des estimateurs du risque réel, seuls les fichiers F_13, F_14, F_15, F_16 de type intervalle et F_14, F_15 de type diagramme satisfont un ensemble de test dont le cardinal est supérieur à 100 individus. Les tableaux (TAB. 5, TAB. 6, TAB. 7) présentent les intervalles de confiance des estimateurs des risques réels de ces fichiers.

Fichier	Nom du fichier	Taille	Nb_cl	Répartition	Nb_var
F_1	Voyage	21	5	(5;4;3;6;3)	2
F_2	Wine	23	9	(2;2;2;2;6;5;2;1;1)	21
F_3	Joueur	29	2	(11;18)	7
F_4	Iris de Fisher	30	3	(10;10;10)	4
F_5	Wave	30	3	(10;10;10)	21
F_6	Auto	33	4	(10;8;8;7)	8
F_7	Football	45	4	(16;21;7;1)	7
F_8	Accident	48	3	(36;9;3)	5
F_9	Temperature_1988	60	8	(7;13;5;6;12;7;3;7)	12
F_10	Shuttle	102	7	(78;1;1;15;5;1;1)	9
F_11	Cholesterol	193	2	(99;94)	2
F_12	Age_color	231	9	(23;27;24;27;28;27;27;22;26)	10
F_13	Glucose	690	2	(349;341)	2

F_14	Profession_size	720	5	(36;441;32;35;176)	1
F_15	Temperatur_74-88	900	2	(475;425)	12
F_16	Regions_NU	10000	4	(1900 ;2300 ;3620 ;2180)	4

TAB. 2 – *Inventaire des fichiers de données de type intervalle*

Fichier	Nom du fichier	Taille	Nb_cl	Répartition	Nb_var
F_1	Joueur	29	2	(13 ;16)	4
F_2	Clinique	43	2	(18 ;25)	3
F_3	Musique	33	9	(5 ;10 ;2 ;1 ;4 ;4 ;5 ;1 ;1)	11
F_4	Médecine	36	3	(12 ;12 ;12)	10
F_5	Football	45	2	(27 ;18)	4
F_6	Accident	48	3	(36 ;9 ;3)	11
F_7	Merovigian	59	3	(15 ;25 ;19)	6
F_8	Statut_matrimonial	119	2	(62 ;57)	8
F_9	Régions_UK_1	135	2	(77 ;58)	6
F_10	Cholesterol	193	2	(107 ;86)	2
F_11	Statut_profession	213	4	(47 ;57 ;42 ;67)	7
F_12	Age_color	231	4	(50 ;51 ;82 ;48)	10
F_13	Régions_UK_2	406	2	(337 ;69)	6
F_14	Glucose	690	2	(367 ;323)	2
F_15	Profession_size	720	5	(36 ;441 ;32 ;35 ;176)	14

TAB. 3 – *Inventaire des fichiers de données de type diagramme*

$x\%$	50%	68%	80%	90%	95%	98%	99%
$\varphi(x)$	0.67	1	1.28	1.64	1.96	2.33	2.58

TAB. 4 – *Valeurs de* $\varphi(x)$

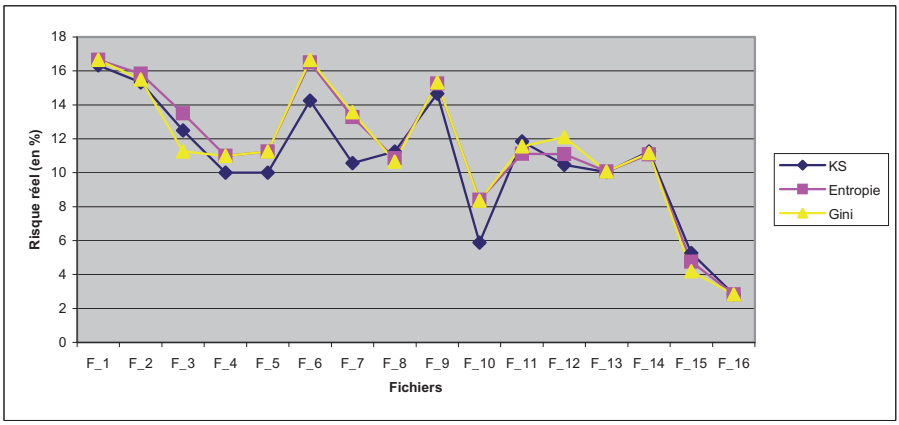


FIG. 4 – *Estimation du risque réel sur les données de type intervalle par critère*

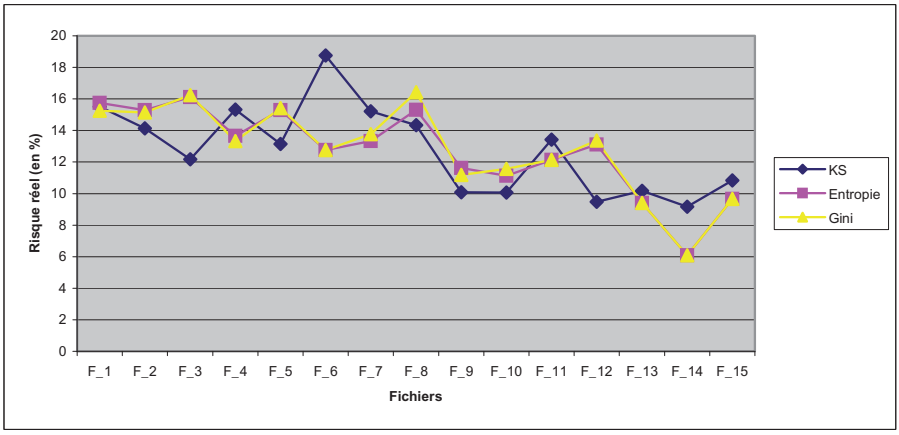


FIG. 5 – *Estimation du risque réel sur les données de type diagramme par critère*

	90%	95%	99%
F_13 (intervalle)	[0.054,0.146]	[0.045,0.155]	[0.028,0.173]
F_14 (intervalle)	[0.065,0.16]	[0.056,0.169]	[0.038,0.187]

F_15 (intervalle)	[0.023,0.082]	[0.017,0.088]	[0.005,0.099]
F_16 (intervalle)	[0.022,0.035]	[0.02,0.036]	[0.018,0.038]
F_14 (diagramme)	[0.048,0.136]	[0.039,0.144]	[0.022,0.161]
F_15 (diagramme)	[0.062,0.155]	[0.053,0.164]	[0.035,0.181]

TAB. 5 – *Intervalles de confiance des estimateurs par le critère KS*

	90%	95%	99%
F_13 (intervalle)	[0.055,0.147]	[0.046,0.156]	[0.028,0.173]
F_14 (intervalle)	[0.064,0.158]	[0.055,0.167]	[0.037,0.185]
F_15 (intervalle)	[0.019,0.076]	[0.014,0.081]	[0.002,0.092]
F_16 (intervalle)	[0.022,0.035]	[0.02,0.036]	[0.018,0.038]
F_14 (diagramme)	[0.024,0.097]	[0.017,0.104]	[0.003,0.118]
F_15 (diagramme)	[0.052,0.141]	[0.044,0.149]	[0.027,0.166]

TAB. 6 – *Intervalles de confiance des estimateurs par le critère de l'entropie*

	90%	95%	99%
F_13 (intervalle)	[0.055,0.147]	[0.046,0.156]	[0.028,0.173]
F_14 (intervalle)	[0.064,0.159]	[0.055,0.168]	[0.037,0.186]
F_15 (intervalle)	[0.015,0.068]	[0.001,0.074]	[-0.001,0.084]
F_16 (intervalle)	[0.022,0.0	[0.02,0.0	[0.018,0.03

	35]	36]	8]
F_14 (diagramme)	[0.024,0.097]	0.017,0.104]	[0.003,0.118]
F_15 (diagramme)	[0.052,0.141]	[0.044,0.149]	[0.027,0.166]

TAB. 7 – *Intervalles de confiance des estimateurs par le critère de Gini*

Les risques réels varient fortement d'un fichier à un autre suivant la taille. Ils sont très élevés pour les fichiers de petite taille en général. Ils diminuent en fonction de l'augmentation du nombre d'objets de l'ensemble de test. Plus la taille de l'échantillon de test est grande, plus l'intervalle de confiance est réduit : on remarque par exemple que le fichier F_16 a des intervalles de confiance très réduits par rapport à ceux des autres fichiers. En général, ces résultats montrent un « léger » avantage pour le critère KS par rapport aux critères de Gini et de l'entropie car les risques réels associés à ce critère sont légèrement inférieurs à ceux des deux autres critères, surtout sur les données de type intervalle.

Il serait intéressant de voir le comportement de ces critères sur des bases de données plus volumineuses. Cependant, comme la modélisation porte sur des concepts (individus de second ordre), il n'est pas facile de trouver de volumineuses bases de données.

5 Conclusion et perspectives

Nous avons proposé différentes méthodes pour ordonner des données de type intervalle, diagramme et taxonomique. Dans le cas de non inclusion stricte d'intervalles, les méthodes que nous avons exposées pour ordonner des intervalles sont compatibles. Les ordres sur les données de type diagramme sont basés sur les paramètres de position et de dispersion d'une distribution. Une variable taxonomique étant une variable organisée en arbre, nous avons utilisé les principes de parcours d'arbres pour ordonner les modalités d'une telle variable.

La possibilité d'estimer la fonction de répartition théorique par la fonction de répartition empirique nous a permis d'adapter le critère KS aux variables de type intervalle, diagramme et taxonomique munies d'un ordre. En utilisant ce critère comme critère d'évaluation de la qualité d'une coupure, il est alors possible de classer un tableau de données où les objets sont décrits par des variables de type intervalle, diagramme ou taxonomique à l'aide d'un arbre de décision. Au lieu de construire un arbre pour chaque ordre possible d'une variable explicative donnée, nous utilisons des relations d'ordre et de préordre sur une même variable explicative pour sélectionner le seuil de coupure. Nous avons également présenté le format des questions binaires et la description des nœuds terminaux pour ce type de données.

Dans cette voie de recherche consistant à étendre les méthodes et algorithmes de l'analyse des données classiques à d'autres types de données, deux aspects semblent être ignorés jusqu'à présent : les réseaux Bayésiens et les SVM (Support Vector Machine). Pour les réseaux Bayésiens, le problème consiste à examiner comment définir des probabilités conditionnelles sur des objets décrits par des variables non classiques, par exemple des probabilités conditionnelles a posteriori étendues à des variables de type intervalle, diagramme, etc. 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.

Castillo E., J. M. Gutiérrez et A. S. Hadi (1997) *Expert Systems and Probabilistic Network Models*; Monographs in Computer Science, Springer.

Celeux, G. et Y. Lechevallier (1982) Méthodes de segmentation non paramétriques. *Revue de Statistique Appliquée*, volume XXX, Numéro 4, pp 39-53.

Cornuéjols, A. et L. Miclet (2002) *Apprentissage artificiel : concepts et algorithmes* ; eyrolles.

Diday, E. (1998) L'analyse des données symboliques : un cadre théorique et des outils. *Cahier de recherche du CEREMADE*, UMR 7534, Numéro 9821, Université Paris Dauphine ; 1998.

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.

Fisburn, P. C. (1985) *Interval orders and interval graphs: A study of partially ordered sets*; A Wiley-Interscience Publication.

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

Gioia, F. (2001) *Metodi Statistici per Variabili di Intervallo*, Ph.D.Thesis, Università di Napoli Federico II.

Gondran, M. et M. Minoux (1985) *Graphes et algorithmes* ; 2^{e} édition, Eyrolles, Paris.

Gustafson, D. E; S. Gelfand et S. K. Mitter (1980) A non parametric multiclass partitioning method for classification; *Proceedings of the 5th International Conference on Pattern Recognition*.

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.

Pirlot, M. et P. H. Vincke (1997) *Semi-orders : properties, representations, applications.* Kluwer Academic Publisher.

Tsoukias, A. et N. A. The (2001) Numerical representation of PQI interval orders. *Cahier du LAMSADE, Numéro 184*, 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 goal of this paper consists to adapt Kolmogorov-Smirnov's binary splitting criterion to interval-valued, diagram and taxonomical data for decision tree induction. This criterion requires an order on the values of the explanatory variables. We use different methods to order this type of data. We present the format of the binary questions and the description of the terminal nodes for each of this type of data. We study also the precision of this criterion and we compare it to the criteria of entropy and Gini.