Classification hiérarchique de variables discrètes fondée sur l'information mutuelle en pré-traitement d'un algorithme de sélection de variables pertinentes

Hélène Daviet*,**, Ivan Kojadinovic* et Pascale Kuntz*

*LINA CNRS FRE 2729, Site Polytech Nantes, rue Christian Pauc, 44306 Nantes, France
{prenom.nom}@polytech.univ-nantes.fr
**PerformanSe SAS, Atlanpole La Fleuriaye, 44470 Carquefou, France
{prenom.nom}@performanse.fr

Résumé. Le travail présenté a pour contexte la sélection de variables pertinentes dans les problèmes de discrimination caractérisés par un grand nombre de variables potentiellement discriminantes toutes discrètes ou nominales. Dans ce cadre, nous proposons une procédure de sélection fondée sur une troncature k-additive de l'information mutuelle et utilisant une classification ascendante hiérarchique des variables potentiellement discriminantes afin de réduire le nombre de sous-ensembles dont la pertinence est estimée.

1 Introduction

Le problème de la sélection de variables en discrimination se rencontre généralement lorsque le nombre de variables, pouvant être utilisées pour expliquer la classe d'un individu, est très élevé. Le rôle de la procédure de sélection de variables consiste alors à sélectionner un sous-ensemble de variables *potentiellement discriminantes* permettant d'expliquer la classe de façon optimale ou quasi-optimale. La nécessité de ce traitement préalable est essentiellement due au fait que, généralement, l'utilisation d'un nombre de variables discriminantes trop élevé dans un modèle de discrimination détériore grandement sa capacité de *généralisation* et la compréhension de la relation modélisée.

D'un point de vue structurel, une procédure de sélection de variables peut être vue comme composée de deux éléments fondamentaux (Liu et Motoda, 1998) : une *mesure de pertinence*, utilisée pour mesurer l'*influence* d'un sous-ensemble de variables potentiellement discriminantes sur la variable qualitative à expliquer, et un *algorithme de recherche*, dont le rôle est de *parcourir* l'ensemble des sous-ensembles de variables à la recherche d'un sous-ensemble optimal ou quasi-optimal au sens de la mesure de pertinence. Du point de vue de la définition de la mesure de pertinence, les procédures de sélection de variables peuvent être essentiellement regroupées en deux classes (Liu et Motoda, 1998) : les procédures *filtres* et les procédures *modèle-dépendantes*. Dans le cas des procédures filtres, la sélection de variables est totalement indépendante du modèle de discrimination choisi et s'effectue en tant que traitement préalable à la phase d'estimation. Parmi les procédures filtres, citons le travail de Fleuret (2004), proche du notre. En revanche, dans le cas des procédures modèle-dépendantes, la mesure de pertinence

est définie à l'aide du modèle de discrimination choisi, généralement en fonction de l'erreur du modèle sur un ensemble test.

Dans le cadre de ce travail, nous nous intéressons au cas où les variables potentiellement discriminantes sont toutes discrètes ou nominales et nous proposons une procédure *filtre*, CLASSSEL, utilisant une *troncature k-additive de l'information mutuelle* (Kojadinovic, 2005) comme mesure de pertinence. L'utilisation de l'information mutuelle en tant que mesure de pertinence a déjà été considérée à de nombreuses reprises dans la littérature (voir p. ex. Hutter et Zaffalon, 2005). L'approximation que nous utilisons permet d'approcher la pertinence d'un ensemble de variables à partir des pertinences de ses sous-ensembles de faible cardinal. Afin d'éviter d'avoir à parcourir la totalité des sous-ensembles non vides de l'ensemble des variables potentiellement discriminantes ou d'avoir à recourir à des heuristiques souvent trop sous-optimales du type *sélection pas à pas*, nous proposons d'effectuer, en pré-traitement de la sélection de variables, une classification ascendante hiérarchique de l'ensemble des variables potentiellement discriminantes afin d'en identifier la *structure*.

2 L'information mutuelle en tant que mesure de pertinence

Nous considérons, dans la suite, qu'un problème de sélection de variables se présente sous la forme d'un ensemble $\aleph = \{X_1, \ldots, X_m\}$ de variables aléatoires potentiellement discriminantes et d'un vecteur aléatoire $\mathbb{Y}$ à expliquer. Comme indiqué précédemment, nous nous limitons au cas où toutes ces variables sont discrètes et prennent un nombre fini de valeurs. Dans le reste de ce document, les sous-ensembles de $\aleph$ seront notés par des majuscules doubles, p. ex. $\mathbb{X}$. De plus, lorsque nécessaire, un sous-ensemble $\mathbb{X} \subseteq \aleph$ pourra également être vu comme un vecteur aléatoire dont les composantes sont des éléments distincts de $\aleph$.

Dans ce contexte probabiliste, il est naturel de mesurer la pertinence des sous-ensembles de $\aleph$ à l'aide d'une *mesure de dépendance* (voir p. ex. Drouet-Mari et Kotz, 2001). En d'autres termes, nous considérerons qu'un sous-ensemble $\mathbb{X}$ non vide de $\aleph$ est d'autant plus pertinent que le degré de dépendance entre les vecteurs aléatoires $\mathbb{X}$ et $\mathbb{Y}$ est élevé. La mesure de dépendance à utiliser peut p. ex. être choisie parmi les mesures d'*écart à l'indépendance*. Dans le cadre de ce travail, nous avons opté pour l'*information mutuelle* en raison de ses interprétations multiples. En effet, cette quantité peut être vue comme l'écart à l'indépendance obtenu à partir de la divergence de Kullback et Leibler (1951). De plus, elle se décompose naturellement de manière additive en fonction de l'entropie de Shannon (1948).

Il est bien entendu, qu'en pratique, nous ne disposons pas de la distribution de probabilité de $(X_1, \ldots, X_m, \mathbb{Y})$ mais uniquement de n réalisations indépendantes de ce vecteur aléatoire à partir desquelles la mesure de pertinence peut être estimée.

2.1 Définition et propriétés

Considérons un couple $(\mathbb{X}, \mathbb{Y})$ de vecteurs aléatoires discrets prenant un nombre fini de valeurs. L'*information mutuelle* entre $\mathbb{X}$ et $\mathbb{Y}$ est définie comme l'*écart à l'indépendance* entre $\mathbb{X}$ et $\mathbb{Y}$ mesurée par la divergence de Kullback et Leibler. Pour deux distributions de probabilité

$p = (p_1, \ldots, p_l)$ et $q = (q_1, \ldots, q_l)$, $l \geq 2$, la divergence de Kullback et Leibler est définie par

$$KL(p, q) = \sum_{i=1}^{l} p_i \log\left(\frac{p_i}{q_i}\right), \tag{1}$$

en adoptant la convention $0 \log \frac{0}{0} = 0$. Remarquons que cette divergence n'est pas symétrique. Notons $p_{(\mathbb{X},\mathbb{Y})}$, $p_{\mathbb{X}}$ et $p_{\mathbb{Y}}$ la distribution jointe et les distributions marginales respectivement des vecteurs aléatoires $\mathbb{X}$ et $\mathbb{Y}$. L'information mutuelle entre $\mathbb{X}$ et $\mathbb{Y}$ est alors définie par

$$I(\mathbb{X}; \mathbb{Y}) = KL(p_{(\mathbb{X},\mathbb{Y})}, p_{\mathbb{X}} \otimes p_{\mathbb{Y}}), \tag{2}$$

où $\otimes$ désigne le produit tensoriel. À partir de la définition précédente, nous voyons que l'information mutuelle est symétrique et, en appliquant l'inégalité de Jensen à la divergence de Kullback et Leibler, nous obtenons que l'information mutuelle est toujours positive, et nulle si et seulement si $\mathbb{X}$ et $\mathbb{Y}$ sont indépendants (voir p. ex. Cover et Thomas, 1991).

L'information mutuelle peut être également interprétée comme la *H-information* obtenue à partir de l'entropie de Shannon (voir p. ex. Morales et al., 1996). L'entropie de Shannon d'une distribution de probabilité $p = (p_1, ..., p_l)$ est définie par

$$H(p) = -\sum_{i=1}^{l} p_i \log(p_i),$$

en adoptant la convention $0 \log 0 = 0$. La quantité $H(p)$ est toujours positive et peut être interprétée comme une mesure d'*incertitude* ou d'*information* (Rényi, 1965). Relativement à l'entropie de Shannon, l'information mutuelle entre $\mathbb{X}$ et $\mathbb{Y}$ peut se réécrire comme

$$I(\mathbb{X}; \mathbb{Y}) = H(p_{\mathbb{X}}) - E_{p_{\mathbb{Y}}}[H(p_{\mathbb{X}|\mathbb{Y}=y})] = H(p_{\mathbb{Y}}) - E_{p_{\mathbb{X}}}[H(p_{\mathbb{Y}|\mathbb{X}=x})]. \tag{3}$$

où $p_{\mathbb{X}|\mathbb{Y}=y}(x) = \frac{p_{(\mathbb{X},\mathbb{Y})}(x,y)}{p_{\mathbb{Y}}(y)}$. Ainsi, l'information mutuelle peut être interprétée comme une mesure de *réduction d'incertitude* et peut être comparée au *coefficient de détermination* en régression linéaire multiple, lequel mesure une réduction de variabilité. En réécrivant les espérances dans l'Eq. (3), nous obtenons

$$I(\mathbb{X}; \mathbb{Y}) = H(p_{\mathbb{X}}) + H(p_{\mathbb{Y}}) - H(p_{(\mathbb{X},\mathbb{Y})}). \tag{4}$$

2.2 Estimation

Considérons deux vecteurs aléatoires discrets $\mathbb{X}$ et $\mathbb{Y}$ prenant respectivement leurs valeurs dans $\{x_1, \ldots, x_r\}$ et $\{y_1, \ldots, y_s\}$. En considérant l'Eq. (2), il apparaît clairement que leur information mutuelle est fonction de leur distribution jointe $p_{(\mathbb{X},\mathbb{Y})}$, estimée classiquement par maximum de vraisemblance (proportions). Un estimateur naturel de l'information mutuelle est alors $\hat{I}(\mathbb{X}; \mathbb{Y}) = KL(\hat{p}_{(\mathbb{X},\mathbb{Y})}, \hat{p}_{\mathbb{X}} \otimes \hat{p}_{\mathbb{Y}})$. En utilisant la méthode *delta*, il est possible de montrer (voir p. ex. Menéndez et al., 1995) que $\sqrt{n}[\hat{I}(\mathbb{X}; \mathbb{Y}) - I(\mathbb{X}; \mathbb{Y})]$ est asymptotiquement normalement distribué, d'espérance nulle et de variance $\sigma^2_{KL}(p_{(\mathbb{X},\mathbb{Y})})$, où

$$\sigma^2_{KL}(p_{(\mathbb{X},\mathbb{Y})}) = \sum_{i=1}^{r}\sum_{j=1}^{s} p_{(\mathbb{X},\mathbb{Y})}(x_i, y_j)\left(\log\frac{p_{(\mathbb{X},\mathbb{Y})}(x_i, y_j)}{p_{\mathbb{X}}(x_i)p_{\mathbb{Y}}(y_j)}\right)^2 - KL(p_{(\mathbb{X},\mathbb{Y})}, p_{\mathbb{X}} \otimes p_{\mathbb{Y}})^2.$$

Lorsque $\mathbb{X}$ et $\mathbb{Y}$ sont indépendants, il est possible de montrer de façon similaire que l'information mutuelle suit asymptotiquement une loi du χ^2 à $(r-1)(s-1)$ degrés de liberté. Notons qu'une approche Bayésienne de l'estimation de l'information mutuelle a été récemment proposée par Hutter et Zaffalon (2005).

2.3 Généralisation de l'information mutuelle

En partant de l'Eq. (4), Abramson (1963) a proposé une généralisation naturelle de l'information mutuelle entre plus de deux vecteurs aléatoires. L'information mutuelle entre trois vecteurs aléatoires $\mathbb{X}$, $\mathbb{Y}$ et $\mathbb{Z}$ est définie par

$$I_3(\mathbb{X};\mathbb{Y};\mathbb{Z}) = H(p_{\mathbb{X}}) + H(p_{\mathbb{Y}}) + H(p_{\mathbb{Z}}) \\ - H(p_{(\mathbb{X},\mathbb{Y})}) - H(p_{(\mathbb{X},\mathbb{Z})}) - H(p_{(\mathbb{Y},\mathbb{Z})}) + H(p_{(\mathbb{X},\mathbb{Y},\mathbb{Z})}).$$

Plus généralement, pour $r \geq 2$ vecteurs aléatoires $\mathbb{X}_1,\ldots,\mathbb{X}_r$, la définition suivante a été adoptée :

$$I_r(\mathbb{X}_1;\ldots;\mathbb{X}_r) = \sum_{k=1}^{r} \sum_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,r\}} (-1)^{k+1} H(p_{(\mathbb{X}_{i_1},\ldots,\mathbb{X}_{i_k})}). \tag{5}$$

L'information mutuelle entre $r \geq 2$ vecteurs aléatoires $\mathbb{X}_1,\ldots,\mathbb{X}_r$ peut être interprétée comme une mesure de leur *interaction simultanée* (Wienholt et Sendhoff, 1996; Kojadinovic, 2005). Elle peut également être vue comme une sorte de mesure de similarité *multivoie* entre variables. Si elle est nulle, les r vecteurs aléatoires n'interagissent pas simultanément. Il est important de noter que l'information mutuelle entre plus de deux vecteurs aléatoires n'est pas nécessairement positive (Cover et Thomas, 1991).

2.4 Mesure de pertinence

Nous définissons la pertinence d'un sous-ensemble $\mathbb{X}$ de $\aleph$ par

$$\omega(\mathbb{X}) = \begin{cases} 0, & \text{si } \mathbb{X} = \emptyset, \\ I_2(\mathbb{X};\mathbb{Y}), & \text{sinon.} \end{cases} \tag{6}$$

Il peut être vérifié que cette mesure de pertinence est monotone par rapport à l'inclusion, ce qui n'est pas sans poser des problèmes pratiques (Kojadinovic, 2005). Sa version estimée à partir des données sera notée $\hat{\omega}$.

3 Approximations k-additives des mesures de pertinence

Soit $i : 2^{\aleph} \to \mathbb{R}$ la fonction d'ensemble définie par

$$i(\mathbb{X}) = \begin{cases} 0, & \text{si } \mathbb{X} = \emptyset, \\ I_{r+1}(X_{i_1};\ldots;X_{i_r};\mathbb{Y}), & \text{si } \mathbb{X} = \{X_{i_1},\ldots,X_{i_r}\}. \end{cases}$$

En utilisant des notions de combinatoire telles que la transformée de Möbius (Rota, 1964), il peut être montré que i est une *représentation équivalente à* ω (Kojadinovic, 2005). D'un point

de vue pratique, cela signifie que les nombres $(\omega(\mathbb{X}))_{\mathbb{X}\subseteq\aleph}$ peuvent être obtenus à partir des coefficients $(i(\mathbb{X}))_{\mathbb{X}\subseteq\aleph}$, et *vice versa*. Plus précisément, à partir de l'Eq. (5) et en utilisant la *transformation zeta* (Rota, 1964), il a été montré que

$$i(\mathbb{X}) = \sum_{\mathbb{T}\subseteq\mathbb{X}} (-1)^{|\mathbb{T}|+1}\omega(\mathbb{T}) \qquad \text{et} \qquad \omega(\mathbb{X}) = \sum_{\mathbb{T}\subseteq\mathbb{X}} (-1)^{|\mathbb{T}|+1} i(\mathbb{T}), \qquad \forall\, \mathbb{X}\subseteq\aleph.$$

Il s'ensuit que la pertinence d'un sous-ensemble $\mathbb{X} = \{X_{i_1}, \ldots, X_{i_r}\}$ de $\aleph$ peut être réécrite comme

$$\begin{aligned}\omega(\mathbb{X}) = &\sum_{X_j\in\mathbb{X}} I_2(X_j;\mathbb{Y}) - \sum_{\{X_j,X_k\}\subseteq\mathbb{X}} I_3(X_j;X_k;\mathbb{Y}) \\ &+ \sum_{\{X_j,X_k,X_l\}\subseteq\mathbb{X}} I_4(X_j;X_k;X_l;\mathbb{Y}) - \cdots + (-1)^{r+1} I_{r+1}(X_{i_1};\ldots;X_{i_r};\mathbb{Y}). \quad (7)\end{aligned}$$

La pertinence de $\mathbb{X}$ est ainsi calculée d'abord en sommant les pertinences des singletons contenus dans $\mathbb{X}$, puis en soustrayant les informations mutuelles entre paires de variables de $\mathbb{X}$ et $\mathbb{Y}$, ensuite en ajoutant les informations mutuelles entre variables des sous-ensembles de 3 éléments de $\mathbb{X}$ et $\mathbb{Y}$, etc. Les informations mutuelles qui sont rajoutées ou enlevées peuvent être vues comme des *termes correcteurs* ou des termes d'*ordre supérieur* et s'apparentent aux termes d'interaction utilisés dans le contexte de l'*analyse de variance* ou des *modèles log-linéaires* (Agresti, 2002).

Afin d'obtenir une approximation de l'information mutuelle moins coûteuse en terme de temps de calcul, nous proposons de procéder à une *troncature k-additive* de ω pour un $k \in \{1,\ldots,m\}$ fixé, c.-à-d. de négliger les *termes correcteurs* d'ordre supérieur à k dans l'Eq. (7). La troncature k-additive de ω est simplement définie par

$$\omega^{(k)}(\mathbb{X}) = \sum_{\substack{\mathbb{T}\subseteq\mathbb{X}\\ |\mathbb{T}|\leq k}} (-1)^{|\mathbb{T}|+1} i(\mathbb{T}), \qquad \mathbb{X}\subseteq\aleph.$$

À partir de l' Eq. (7), nous voyons ainsi qu'approcher ω par sa troncature k-additive $\omega^{(k)}$ est équivalent à considérer que l'information mutuelle entre plus de k variables potentiellement discriminantes et $\mathbb{Y}$, est négligeable.

Prendre la troncature 1-additive de ω en tant que mesure de pertinence est équivalent à considérer que la pertinence d'un sous-ensemble est égale à la somme des pertinences des singletons qu'il contient, c.-à-d. que ω est additive. Dans la plupart des situations réelles, une telle simplification est trop extrême car, généralement, l'ensemble des variables potentiellement discriminantes contient des variables redondantes.

La troncature 2-additive apparaît plus appropriée car elle prend partiellement en compte les interactions entre variables potentiellement discriminantes sans être trop complexe en terme de nombre de coefficients. En effet, $\omega^{(2)}$ est complètement définie à partir de ses valeurs sur les singletons et les paires de variables potentiellement discriminantes, c.-à-d., pour tout $\mathbb{X}\subseteq\aleph$ non vide, il peut être montré (voir p. ex. Kojadinovic, 2005) que

$$\omega^{(2)}(\mathbb{X}) = \sum_{\{X_i,X_j\}\subseteq\mathbb{X}} \omega(\{X_i,X_j\}) - (|\mathbb{X}|-2)\sum_{X_i\in\mathbb{X}} \omega(\{X_i\}).$$

Utiliser $\omega^{(2)}$ est très avantageux du point de vue du temps de calcul : une fois les pertinences des singletons et des paires de $\aleph$ estimées, la pertinence approchée de n'importe quel sous-ensemble de $\aleph$ peut être immédiatement calculée à l'aide de l'équation précédente. Du point de vue de la qualité de l'approximation, nous pouvons voir, en considérant l'Eq. (7), que plus la dépendance entre variables de $\aleph$ est faible, meilleure sera l'approximation de ω par sa troncature 2-additive.

4 Classification hiérarchique ascendante de variables pour identifier la *structure* de $\aleph$

Le deuxième élément fondamental d'une procédure de sélection de variables est un algorithme de recherche. Afin d'éviter d'avoir à parcourir la totalité des sous-ensembles non vides de $\aleph$ ou d'avoir à recourir à des heuristiques souvent trop sous-optimales du type *sélection pas à pas*, nous proposons d'effectuer une classification ascendante hiérarchique de $\aleph$ afin d'en identifier la *structure*.

4.1 Classification ascendante hiérarchique de variables fondée sur l'information mutuelle

Un algorithme de classification ascendante hiérarchique est classiquement défini par deux éléments : une mesure de similarité (ou dissimilarité) et un critère d'agrégation entre classes. Les partitions compatibles avec la hiérarchie de classes obtenue sont généralement évaluées (en vue p. ex. du choix d'une partition) en fonction de leur *homogénéité* et de leur *séparation*. Une première façon simple de mesurer l'homogénéité et la séparation d'une partition consiste à calculer le *diamètre* moyen et l'*écart* moyen respectivement de ses classes (voir p. ex. Hansen et Jaumard, 1997).

Pour la mesure de similarité, nous avons opté une fois de plus pour l'information mutuelle, cette fois-ci normalisée. La similarité entre deux variables X_i et X_j de $\aleph$ est ainsi définie par

$$I^*(X_i; X_j) = \frac{I_2(X_i; X_j)}{\min[H(p_{X_i}), H(p_{X_j})]}.$$

Il peut être vérifié que la quantité $I^*(X_i; X_j)$ est comprise entre 0 et 1 (Joe, 1989). De plus, $I^*(X_i; X_j) = 1$ si et seulement si X_i et X_j sont fonctionnellement dépendantes (Joe, 1989, Th. 2.3). Comme critère d'agrégation, nous avons choisi le lien moyen, souvent considéré comme une alternative "robuste" au lien simple ou au lien complet.

4.2 L'algorithme CLASSSEL

Idéalement, l'objectif serait de retenir, parmi les partitions les plus homogènes compatibles avec la hiérarchie obtenue, la moins fine. D'un point de vue pratique, il faut tempérer l'objectif précédent en trouvant un compromis entre une forte homogénéité et un faible nombre de classes. Nous nous contentons ici d'identifier un "coude" sur le graphique donnant le diamètre moyen des partitions compatibles en fonction de leur taille (Hardy, 1996). L'heuristique que nous proposons alors est de n'estimer que la pertinence des sous-ensembles composés d'au

plus une variable de chaque classe, les variables d'une même classe pouvant être considérées comme "suffisamment dépendantes". Cette approche nous pousse ainsi à privilégier l'homogénéité de la partition retenue. En effet, notre heuristique sera d'autant plus efficace que les classes regrouperont des variables très proches. Ainsi, n'en choisir qu'au plus une par classe ne devrait pas empêcher l'algorithme d'évaluer des sous-ensembles de variables quasi-optimaux. De plus, la pertinence des sous-ensembles étant mesurée par le biais de la troncature 2-additive de l'information mutuelle pénalisant les sous-ensembles contenant des variables liées, un certain degré de dépendance inter-classes est envisageable en pratique.

Une fois une partition compatible sélectionnée, il est demandé à l'utilisateur de donner le nombre maximal p de variables discriminantes à retenir. Les grandes lignes de l'algorithme CLASSSEL, dont une première version a été implémentée sur la plateforme R (R Development Core Team, 2005) sont données ci-après :

Algorithme 1 L'algorithme CLASSSEL.

Nécessite:
$\mathcal{P} = \{\mathbb{X}_1, \mathbb{X}_2, ..., \mathbb{X}_k\}$: une partition de $\aleph$ en k classes
p : le cardinal maximal des sous-ensembles de variables pertinentes à renvoyer

Renvoie:
Un ensemble de sous-ensembles de variables pertinentes

$q \leftarrow \min(p, k)$
pour $i = 1, \ldots, q$ **faire**
  pour chaque sous-ensemble $\mathbb{X} \subseteq \aleph$ de cardinal i composé d'au plus une variable de chaque classe de $\mathcal{P}$ **faire**
    calculer sa pertinence $\omega^{(2)}(\mathbb{X})$
    stocker le couple $\left(\mathbb{X}, \omega^{(2)}(\mathbb{X})\right)$
  fin pour
  Afficher parmi les sous-ensembles de cardinal i considérés dans la boucle précédente, celui qui a la plus forte pertinence
fin pour

À ce stade de notre travail, nous évaluons tous les sous-ensembles de cardinal inférieur à q contenant au plus une variable de chaque classe de la partition retenue. Le nombre de sous-ensembles de variables parcourus est ainsi de l'ordre de $O(|\mathbb{X}_1| \times \cdots \times |\mathbb{X}_k|)$, ce qui n'est envisageable que pour des problèmes de faible taille. Cet algorithme de parcours clairement non satisfaisant sera remplacé par une heuristique pour réduire le nombre de sous-ensembles de variables à évaluer. Cette approche exhaustive été implémentée dans le seul but de tester l'intérêt de la classification en tant que pré-traitement d'une procédure de sélection de variables.

5 Expérimentations

Afin d'étudier la qualité des sous-ensembles de variables renvoyés par l'algorithme CLASSSEL, nous avons considéré deux problèmes de sélection de variables : un problème artificiel dont nous connaissons la structure et le problème classique Soybean (Newman et al., 1998).

Dans le cadre du problème artificiel, nous considérons un ensemble de 35 variables discrètes potentiellement discriminantes d'une 36 ème. Ce problème a la structure suivante :

- $X_1, \ldots, X_5$ et $X_{21}, X_{22}, X_{28}, X_{29}$, sont mutuellement indépendantes, à valeurs dans $\{1, 2, 3, 4\}$, et distribuées selon une loi uniforme ;
- $X_6, \ldots, X_{10}$ sont définies par $X_i = 4 - X_{i-5}$;
- $X_{11}, \ldots, X_{15}$ sont définies par $X_i = X_{i-10}^2$;
- $X_{16}, \ldots, X_{20}$ sont définies par $X_i = \min(X_1, X_2)$;
- $X_{23} = 3X_1 + 1$ et $X_{24} = 2X_2 - 1$;
- $X_{25} = X_1^3$;
- $X_{26} = X_6 + X_{25}$ et $X_{27} = X_7 + X_{26}$;
- $X_{30} = X_1 - 1$ si $X_1 < 3$ et $X_{30} = X_1 + 1$ sinon ;
- $X_{31} = X_1$, $X_{32} = 2 - X_{31}$ et $X_{33} = X_6 + X_7$;
- $X_{34} = X_4 - X_5 + X_3$;
- $X_{35} = X_2 + X_3$ si $X_2 < 3$, $X_{35} = X_1$ si $X_2 < 3$, $X_3 < 3$ et $X_{35} = X_4$ sinon ;
- la variable aléatoire Y à expliquer est définie par $Y = \max(X1, X2, X3) + \min(X_4, X_5)$.

Nous voyons ainsi que les variables $X_6, \ldots, X_{20}$, $X_{23}, \ldots, X_{27}$ et $X_{30}, \ldots, X_{35}$ sont redondantes par rapport aux variables $X_1, \ldots, X_5$. Les variables $X_{21}, X_{22}, X_{28}, X_{29}$ sont quant à elles non pertinentes. Enfin, $n = 800$ réalisations du vecteur aléatoire $(X_1, \ldots, X_{35}, Y)$ ont été générées.

Le deuxième problème que nous avons considéré est fondé sur le jeu de données classique Soybean (Newman et al., 1998). Il est composé de 35 variables discrètes potentiellement discriminantes et de 307 individus. Les individus sont répartis en 19 classes, les quatre dernières étant très peu représentées.

5.1 Protocole expérimental

Pour chacun des deux problèmes, l'algorithme CLASSSEL a renvoyé q sous-ensembles de variables potentiellement explicatives de cardinal 1 à q. Pour chaque sous-ensemble de cardinal i renvoyé, nous avons construit un arbre de décision à l'aide de l'algorithme CART (Breiman et al., 1984) en utilisant un ensemble d'apprentissage contenant 70 % des individus pris aléatoirement (distribution uniforme sur l'ensemble des individus). Nous avons ensuite défini l'erreur d'apprentissage par le nombre d'individus mal classés et l'erreur de test par le nombre d'individus (parmi les 30 % restants) dont la classe a été mal prédite. Ces deux indicateurs permettent d'évaluer la qualité de l'arbre construit avec un nombre i restreint de variables explicatives et par conséquent la pertinence du sous-ensemble de variables en question. Afin d'obtenir des résultats plus "robustes", pour chaque sous-ensemble de taille i, nous avons généré 500 échantillons d'apprentissage comme indiqué ci-dessus et appliqué l'algorithme CART. Le critère de qualité retenu est le nombre moyen d'individus mal classés et le nombre moyen d'individus mal prédits sur ces 500 répétitions.

Afin de comparer l'algorithme CLASSSEL avec des approches filtres existantes, nous avons effectué ces mêmes tests sur des sous-ensembles obtenus avec l'approche additive de Lewis (1992) implantée dans le logiciel *Weka* (Witten et Frank, 2005), qui calcule l'information mutuelle entre chaque variable potentiellement discriminante et la classe à prédire, et ne retient que les variables les plus porteuses d'information.

Enfin, nous avons aussi comparé les résultats de CLASSSEL avec ceux obtenus pour des sous-ensembles de variables explicatives générés aléatoirement (distribution uniforme sur l'en-

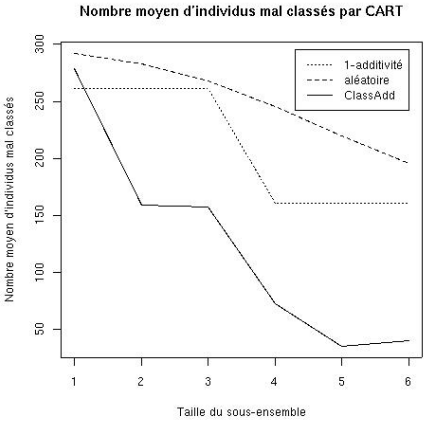

FIG. 1 – *Nombre moyen d'individus mal classés (ensemble d'apprentissage) avec les données artificielles.*

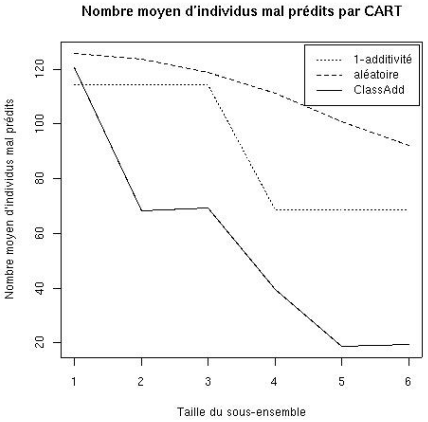

FIG. 2 – *Nombre moyen d'individus mal prédits (ensemble de test) avec les données artificielles.*

semble des variables). Pour chaque i, $i \in \{1, ..., q\}$, nous avons généré 500 sous-ensembles de variables de cardinal i et appliqué l'algorithme CART comme indiqué précédemment.

5.2 Résultats expérimentaux

5.2.1 Problème artificiel

La partition retenue est celle en 12 classes :

$$\{X_1, X_6, X_{11}, X_{23}, X_{25}, X_{26}, X_{27}, X_{30}, X_{31}, X_{32}\},$$

$$\{X_2, X_7, X_{12}, X_{24}\}, \{X_3, X_8, X_{13}\},$$

$$\{X_4, X_9, X_{14}\}, \{X_5, X_{10}, X_{15}\},$$

$$\{X_{16}, X_{17}, X_{18}, X_{19}, X_{20}, X_{33}\},$$

$$\{X_{21}\}, \{X_{22}\}, \{X_{28}\}, \{X_{29}\}, \{X_{34}\}, \{X_{35}\}$$

Les Figures 1 et 2 présentent les résultats de l'algorithme CART appliqué aux sous-ensembles renvoyés par CLASSSEL, aux sous-ensembles générés aléatoirement et à ceux obtenus avec un filtre additif pour le jeu de données artificielles. Le fait que les données contiennent un grand nombre de variables redondantes explique que l'algorithme CLASSSEL retourne des ensembles ayant un meilleur pouvoir explicatif de la classe que l'algorithme additif classique. Enfin, les sous-ensembles retournés par CLASSSEL sont nettement meilleurs que les sous-ensembles générés aléatoirement. Ce résultat conforte l'heuristique de la classification consistant à tenir compte de la structure de l'ensemble des variables potentiellement discriminantes.

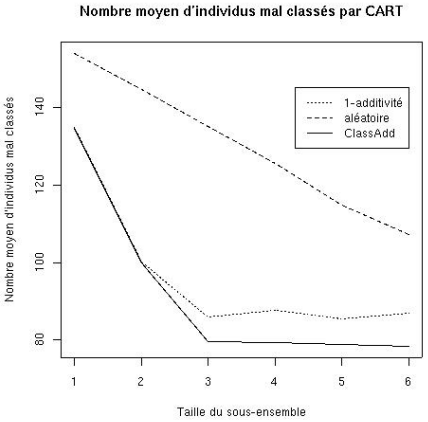

FIG. 3 – *Nombre moyen d'individus mal classés (ensemble d'apprentissage) avec les données Soybean.*

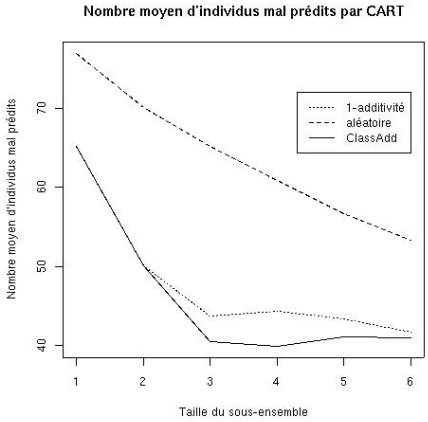

FIG. 4 – *Nombre moyen d'individus mal prédits (ensemble test) avec les données Soybean.*

5.2.2 Les données Soybean

La partition retenue est celle en 20 classes :

$$\{X_1\}, \{X_2\}, \{X_3, X_{26}, X_{27}\}, \{X_4\}, \{X_5\},$$

$$\{X_6\}, \{X_7\}, \{X_8, X_{25}\}, \{X_9\}, \{X_{10}\},$$

$$\{X_{11}, X_{19}, X_{21}, X_{22}, X_{28}, X_{29}\},$$

$$\{X_{12}, X_{13}, X_{14}, X_{15}\}, \{X_{16}\}, \{X_{17}\}, \{X_{18}\},$$

$$\{X_{20}\}, \{X_{23}\}, \{X_{24}\}, \{X_{30}, X_{31}, X_{32}, X_{33}, X_{34}\}, \{X_{35}\}$$

Les Figures 3 et 4 présentent les résultats de l'algorithme CART appliqué sur les sous-ensembles obtenus par CLASSSEL, les sous-ensembles générés aléatoirement et ceux obtenus avec un filtre additif pour le jeu de données Soybean. Nous retrouvons le même type de résultats que pour les données artificielles.

6 Conclusion et perspectives

Dans cet article, nous avons proposé un algorithme de sélection de variables, CLASSSEL. La pertinence d'un sous-ensemble de variables est estimée grâce à une troncature 2-additive de l'information mutuelle. Cette approximation permet d'estimer la pertinence de n'importe quel sous-ensemble, quelle que soit sa taille, en n'utilisant que la pertinence de ses paires et de ses singletons. Afin de réduire la taille de l'espace de recherche, nous avons proposé de faire une classification des variables potentiellement discriminantes en pré-traitement de l'algorithme.

Les premiers résultats obtenus semblent satisfaisants en ce qui concerne la qualité des sous-ensembles retournés. Une des prochaines étapes de notre travail portera sur la recherche d'une heuristique pour le parcours des sous-ensembles afin d'avoir un algorithme de recherche d'un coût raisonnable en terme de nombre de sous-ensembles examinés. De plus, nous avons constaté, lors de nos expérimentations, l'influence du nombre de classes retenues sur les résultats. En effet, la partition obtenue contraint fortement l'ensemble des sous-ensembles de variables potentiellement discriminantes possibles et un mauvais choix du nombre de classes peut entraîner des résultats peu satisfaisants. Dans l'algorithme actuel, le choix du nombre de classes est graphique et donc assez subjectif, ce qui peut poser problème compte-tenu de l'importance de ce choix pour le bon fonctionnement de CLASSSEL. Par la suite, nous allons donc travailler sur le critère de sélection du nombre de classes.

Enfin, pour plus de robustesse, nous envisageons le passage à une version probabiliste de la mesure de pertinence s'inspirant de l'approche de l'*analyse de la vraisemblance du lien* (Lerman, 1981). La mesure de pertinence serait alors définie par

$$\hat{\omega}(\mathbb{X}) = F\left(\hat{I}(\mathbb{X}, \mathbb{Y})\right), \qquad \forall\, \mathbb{X} \subseteq \aleph,\ \mathbb{X} \neq \emptyset.$$

où F est la fonction de répartition de l'estimateur $\hat{I}(\mathbb{X}; \mathbb{Y})$ de $I(\mathbb{X}; \mathbb{Y})$ sous l'hypothèse d'indépendance entre $\mathbb{X}$ et $\mathbb{Y}$ (cf. Section 2.2).

Références

Abramson, N. (1963). *Information Theory and Coding*. New-York : McGraw Hill.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley. Second edition.

Breiman, L., J. Freidman, R. Olshen, et C. Stone (1984). *Classification and Regression Tree*. Wadsworth.

Cover, T. et J. Thomas (1991). *Elements of Information Theory*. John Wiley and Sons.

Drouet-Mari, D. et S. Kotz (2001). *Correlation and dependence*. London : Imperial College Press.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research 5*, 1531–1555.

Hansen, P. et B. Jaumard (1997). Cluster analysis and mathematical programming. *Mathematical programming 79*, 191–215.

Hardy, A. (1996). On the number of clusters. *Computational Statistics and Data Analysis 23*, 83–96.

Hutter, M. et M. Zaffalon (2005). Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis 48*, 633–657.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Am. Statist. Assoc. 84*, 157–164.

Kojadinovic, I. (2005). Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics and Data Analysis 49*(4), 1205–1227.

Kullback, S. et R. A. Leibler (1951). On information and sufficiency. *Ann. Math. Stat. 22*, 79–86.

Lerman, I. (1981). *Classification et analyse ordinale de données*. Paris : Dunod.

Lewis, D. D. (1992). Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of Speech and Natural Language Workshop*, San Mateo, California, pp. 212–217. Morgan Kaufmann.

Liu, H. et H. Motoda (1998). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers.

Menéndez, M., D. Morales, L. Pardo, et M. Salicrú (1995). Asymptotic behaviour and statistical applications of divergence measures in multinomial populations : a unified study. *Statistical papers 36*, 1–29.

Morales, D., L. Pardo, et I. Vajda (1996). Uncertainty of discrete stochastic systems : general theory and statistical theory. *IEEE Trans. on System, Man and Cybernetics 26*(11), 1–17.

Newman, D., S. Hettich, C. Blake, et C. Merz (1998). UCI repository of machine learning databases.

R Development Core Team (2005). *R : A language and environment for statistical computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-00-3.

Rényi, A. (1965). On the foundations of information theory. *Review of the International Statistical Institute 33*(1), 1–14.

Rota, G.-C. (1964). On the foundations of combinatorial theory. I. Theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 2*, 340–368 (1964).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal 27*, 379–623.

Wienholt, W. et B. Sendhoff (1996). How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos 6*(1), 101–117.

Witten, I. H. et E. Frank (2005). *Data Mining : Practical Machine Learning Tools and Techniques* (Second ed.). Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Summary

In the framework of subset variable selection for supervised classification involving only discret variables, we propose a selection algorithm using a computationally efficient relevance measure based on a k-additive truncation of the mutual information and involving an agglomerative hierarchical clustering of the set of potentially discriminatory variables in order to reduce the number of subsets whose relevance is estimated.