Un critère d'évaluation pour la sélection de variables

Dahbia Semani, Carl Frélicot, Pierre Courtellemont

Laboratoire d'Informatique – Image – Interaction
Université de La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle Cedex, France
{dahbia.semani,carl.frelicot,pierre.courtellemont}@univ-lr.fr

Résumé. Cet article aborde le problème de la sélection de variables dans le cadre de la classification supervisée. Les méthodes de sélection reposent sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables. Nous présentons un nouveau critère d'évaluation fondé sur une mesure d'ambiguïté. Cette mesure est fondée sur une combinaison d'étiquettes représentant le degré de spécificité ou d'appartenance aux classes en présence. Les tests menés sur de nombreux jeux de données réels et artificiels montrent que notre méthode est capable de sélectionner les variables pertinentes et d'augmenter dans la plupart des cas les taux de bon classement.

1 Introduction

En reconnaissance des formes, les données sont des vecteurs réalisations de variables qui correspondent à des mesures réalisées sur un système physique ou à des informations collectées lors d'une observation d'un phénomène. Ces variables ne sont pas toutes aussi informatives : elles peuvent correspondre à du bruit, être peu significatives, corrélées ou non pertinentes pour la tâche à réaliser. La sélection de variables a pour objectif de réduire le nombre de ces variables et donc réduire la taille des informations à traiter. Des traitements plus sophistiqués peuvent alors être utilisés dans des espaces de dimension réduite, l'étape d'apprentissage est facilitée, les performances peuvent augmenter lorsque les variables non pertinentes ou redondantes disparaissent, etc.

Nous traitons, dans cet article, le problème de la sélection de variables dans le cadre de la reconnaissance de formes statistique et plus particulièrement dans le cadre de la classification supervisée (ou classement). Dans ce cas, la sélection de variables a pour objectif de réduire la complexité en sélectionnant le sous-ensemble de variables de taille minimale sans que les performances de la règle de classement diminuent trop voire même augmentent.

Une méthode de sélection repose sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables. Nous nous intéressons aux critères d'évaluation. Ainsi, nous proposons un nouveau critère d'évaluation fondé sur une mesure d'ambiguïté. Cette mesure repose sur la combinaison d'étiquettes représentant le degré de spécificité ou d'appartenance aux classes en présence. Des opérateurs d'agrégation issus de la logique floue sont utilisés pour la combinaison de ces étiquettes.

Cet article est organisé comme suit. Un bref état de l'art sur les algorithmes de sélection de variables et les critères d'évaluation est dressé aux sections 2 et 3. Nous

présentons notre critère d'évaluation d'un ensemble de variables à la section 4. La section 5 est consacrée à sa validation sur des jeux de données artificiels et réels utilisés dans la littérature. Enfin, nous concluons ce travail et dressons quelques perspectives.

2 Algorithmes de recherche

Un nombre important d'algorithmes de recherche de sous-ensembles de variables ont été proposés dans la littérature et des études comparatives existent (Jain et Zongker 1997)(Liu et Motoda 1998)(Kudo et Sklansky 2000). Ces algorithmes sont généralement groupés en deux catégories : les algorithmes optimaux garantissant une solution optimale du problème, et les algorithmes sous-optimaux (Jain et Zongker 1997).

Le problème de sélection d'un sous-ensemble de q variables parmi $p >> q$ est un problème combinatoire. La recherche exhaustive du meilleur sous-ensemble parmi les $(p!)/(q!(p-q)!)$ possibles est irréaliste. Une alternative permettant de trouver la solution optimale est l'algorithme *Branch and Bound* (B&B) (Narendra et Fukunaga 1977) dont la complexité est exponentielle. Des heuristiques fondées sur des parcours séquentiels lui sont souvent préférées. Elles consistent à rajouter ou à éliminer itérativement des variables (Devijver et Kittler 1982). Dans les approches séquentielles, il est possible de partir d'un ensemble de variables vide et d'ajouter des variables à celles déjà sélectionnées (*Sequential Forward Selection*, SFS) ou de partir de l'ensemble de toutes les variables et d'éliminer des variables parmi celles déjà sélectionnées (*Sequential Backward Selection*, SBS). Ces méthodes sont connues pour leur simplicité de mise en œuvre et leur rapidité. Cependant, elles sont sous-optimales car elles n'explorent pas tous les sous-ensembles possibles de variables et ne permettent pas de retour arrière pendant la recherche. Pour réduire cet effet, des méthodes alternent les procédures SFS et SBS, permettant ainsi d'ajouter des variables et puis d'en retirer d'autres. C'est le cas, par exemple, des méthodes flottantes (*Sequential Floating Search methods*, (Pudil et al. 1994)). Les acronymes des versions *forward* et *backward* de ces méthodes sont respectivement SFFS et SBFS. Elles sont considérées comme les méthodes sous-optimales les plus efficaces (Jain et Zongker 1997). L'algorithme SFFS consiste à appliquer après chaque étape *forward* autant d'étapes *backward* que le sous-ensemble de variables S_q correspondant améliore le critère d'évaluation $J(S_q)$ à ce niveau de recherche (voir Algorithme 1). Les deux étapes de l'algorithme sont alternées jusqu'à ce qu'une condition d'arrêt soit vérifiée. Parmi celles-ci, citons : une borne sur q, un seuil sur $J(S_q)$ ou sur $J(S_q) - J(S_{q-1})$. Dans l'algorithme SBFS, le même principe est appliqué sauf que les deux étapes sont inversées. Le nombre de variables ajoutées ou éliminées à chaque étape est déterminé dynamiquement en fonction de la valeur du critère d'évaluation ; par conséquent, aucun paramètre n'est à régler au préalable.

L'objectif de cet article n'étant pas le problème de la recherche de sous-ensembles de variables, nous avons opté pour un algorithme séquentiel flottant de type SFFS.

3 Critères d'évaluation

Deux approches sont couramment utilisées pour évaluer la pertinence d'un sous-ensemble de variables sélectionnées (Langley 1994) : l'approche de type *filter* et celle de

Algorithme 1: L'algorithme SFFS.

Données : $\mathcal{S} = \{x_j \mid j = 1, p\}$// $\mathcal{S}$: ensemble de variables initial //
// J : critère (à maximiser par exemple) //
Sortie : $S_q = \{y_j \in \mathcal{S} |\ j = 1, q\},\ q \leq p$
// S_q : sous-ensemble de q variables sélectionnées //
Initialisation : $q := 0; \quad S_q := \emptyset;$
Étape 1 (*forward*)
$x_+ := \arg\max_{x_j \in \mathcal{S} \setminus S_q} J(S_q \cup \{x_j\});$
$S_{q+1} := S_q \cup \{x_+\}; \quad q := q + 1;$
Étape 2 (*backward*)
$x_- := \arg\max_{y_j \in S_q} J(S_q \setminus \{y_j\});$
si $J(S_q \setminus \{x_-\}) > J(S_{q-1})$ **alors**
  $S_{q-1} := S_q \setminus \{x_-\}; \quad q := q - 1;$
  Aller à **Étape 2** ;
sinon
  Aller à **Étape 1** ;
fin

type *wrapper*. Dans la première approche, les critères sont fondés uniquement sur les données et sont donc totalement indépendants du discriminateur utilisé. Les variables sont alors filtrées avant le processus d'apprentissage et de classification. Contrairement à l'approche *filter*, l'approche *wrapper*, présentée longuement dans (Langley 1994) (Kohavi et John 1997), tient également compte de la règle de classement dans le calcul du critère d'évaluation. Celui-ci est simplement la probabilité d'erreur estimée sur l'ensemble des données, éventuellement à l'aide d'une procédure de bootstrap ou de validation croisée si le nombre de variables est conséquent et le nombre d'exemples est insuffisant (Kohavi et John 1997). Il est évident que l'approche *wrapper* est particulièrement bien adaptée aux problèmes de classification. Cependant, sa capacité en généralisation est faible puisqu'elle est attachée à un couple (données, règle de classement). De plus, elle est coûteuse en temps puisque la sélection boucle sur le processus d'apprentissage.

Dans cet article nous nous intéressons aux mesures d'évaluation de type *filter*, qui sont généralement divisées en plusieurs catégories :

1. **Mesures de distance inter-classes :** Elles essayent de conserver la meilleure séparabilité des classes pour réaliser le moins d'erreurs possible. On s'inspire alors du principe de l'analyse factorielle discriminante (AFD) pour la recherche du sous-espace de représentation pour lequel l'inertie intra-classes est minimale (classes compactes) et l'inertie inter-classes est maximale (classes séparées) (ex : le lambda de Wilks (Kittler 1986)). Ces mesures ne requièrent pas d'hypothèse probabiliste et sont bien adaptées au cas où les classes ont des distributions de même matrice de covariances.
2. **Mesures de distance probabiliste :** Appelées aussi *mesures de divergence*, elles maximisent les distances probabilistes (ex : de Mahalanobis, de Battacharyya ou de divergence (Kittler 1986)) entre les densités conditionnelles ou entre

les probabilités a posteriori des classes. Elles sont largement utilisées dans la littérature (Jain et Zongker 1997)(Kudo et Sklansky 2000)(Pudil et al. 1994). Notons que l'hypothèse de distribution gaussienne est souvent nécessaire pour ce type de mesures.

3. **Mesures d'information :** Issues de la théorie de l'information, elles évaluent le gain apporté par les variables via les probabilités *a posteriori*. Si ces probabilités tendent vers une valeur égale quelle que soit la classe, le gain est minimal et l'incertitude (c'est-à-dire l'entropie) est maximale. Ces mesures sont généralement fondées sur le calcul d'entropie (ex : entropie de Shannon, entropie croisée (Koller et Sahami 1996) ou dérivées (Mitra et al. 2002)) ou bien sur le principe MDL (*Minimum Description Length*)(Pfahringer 1995). Citons également l'indice de Gini (Breiman et al. 1984) et la fonction gain (Quinlan 1986) utilisés par les méthodes de sélection par arbre de décision.
4. **Mesures de dépendance :** Ces mesures quantifient le pouvoir de prédiction d'une variable à partir d'une autre variable, donc le degré de redondance (ex : coefficient de corrélation (Hall 2000), information mutuelle (Torkkola 2003)). Il a été montré que toute mesure de dépendance peut être reformulée comme une mesure de catégorie 2 ou 3.

Il existe d'autres critères d'évaluation difficilement classables dans une des catégories de mesures présentées ci-dessus (ex : la mesure de consistance (Dash et al. 2000), les tests statistiques (Liu et Setiono 1995)). De manière générale, il existe peu d'études comparatives des critères d'évaluation, et aucune d'entre elles ne permet de conclure qu'un critère particulier est meilleur qu'un autre.

4 Un nouveau critère d'évaluation

En classification, une bonne solution au problème de la sélection serait de trouver l'espace de représentation $\mathbb{R}^q$ restreint à q variables sur lequel les projections des classes se chevauchent le moins possible. Plus les classes se chevauchent, plus l'ambiguïté de la classification est importante. Nous proposons d'utiliser une mesure d'ambiguïté entre les projections des classes sur les sous-ensembles de variables pour définir un critère d'évaluation. Cette mesure est fondée sur la combinaison d'étiquettes floues/possibilistes représentant le degré de spécificité ou d'appartenance des données $\mathbf{x}$ aux classes en présence.

4.1 Étiquetage

Considérons un point $\mathbf{x} = (x_1, x_2, ..., x_p)^t$ dans un espace de représentation de dimension p et un ensemble de c classes $\omega = \{\omega_1, \omega_2, ..., \omega_c\}$. On peut lui associer un vecteur d'étiquettes à l'aide d'une fonction : $\Re^p \rightarrow [0,1]^c$, $\mathbf{x} \mapsto \mu(\mathbf{x}) = (\mu_1(\mathbf{x}), ..., \mu_c(\mathbf{x}))^t$ où $\mu_i(\mathbf{x})$ représente le degré d'appartenance de $\mathbf{x}$ à la classe ω_i. D'un point de vue mathématique et non sémantique, les étiquettes sont *floues* si $\sum_{j=1}^{c} \mu_j(\mathbf{x}) = 1$, comme par exemple la fonction utilisée dans les méthodes de classification non supervisée de

type FCM (*Fuzzy C-Means*, (Bezdek 1987)) :

$$\mu_i(\mathbf{x}) = 1 \Big/ \sum_{j=1}^{c} \left(\frac{d(\mathbf{x}, \mathbf{p}_i)}{d(\mathbf{x}, \mathbf{p}_j)} \right)^{2/(m-1)} \quad (1)$$

où m est un paramètre à fixer dans l'intervalle $]1, +\infty[$ et $d(\mathbf{x}, \mathbf{p}_i)$ est une distance entre $\mathbf{x}$ et un prototype de la classe ω_i ; la moyenne estimée est le prototype le plus usuel. Si les étiquettes ne sont pas normalisées, elles sont dites *possibilistes* au sens où elle représentent un degré de spécificité (Krishnapuram et al. 1993), comme par exemple :

$$\mu_i(\mathbf{x}) = \frac{\lambda_i}{\lambda_i + d(\mathbf{x}, \mathbf{p}_i)} \quad (2)$$

où les λ_i sont des paramètres à fixer dans l'intervalle $]1, +\infty[$. C'est cette fonction d'étiquetage que nous utiliserons à la section 5 avec $\lambda_i = 1$ $(\forall i = 1, c)$ et la distance de Mahalanobis.

4.2 La mesure d'ambiguïté

Après l'étape d'étiquetage, nous disposons, pour chaque vecteur $\mathbf{x}$, d'un vecteur d'étiquettes μ_i $(i = 1, c)$. Nous voulons quantifier l'ambiguïté entre les classes en combinant les étiquettes μ_i. Pour cela, nous utilisons des opérateurs de combinaison issus de la logique floue et tout à fait adaptés à notre problème. Le lecteur intéressé trouvera des présentations des opérateurs d'agrégation par exemple dans (Dubois et Prade 1985).

Nous avons choisi d'utiliser les normes (et conormes) triangulaires ou *t-normes* (et *t-conormes*). Ces normes peuvent être vues comme l'extension au cas multi-valué des opérateurs d'intersection $\cap$ et d'union $\cup$ de la théorie classique des ensembles, ou des connecteurs logiques ET, OU de la logique booléenne. Quelques-unes de ces normes $(\top,\perp)$ sont données au tableau 1 ; le lecteur intéressé peut en trouver une synthèse dans (Klir et Yuan 1995). Dans toute la suite, $\top$ désigne une t-norme arbitraire et $\perp$ sa t-conorme duale. Remarquons que les normes de Yager généralisent celles de Łukasiewcz $(m = 1)$, qu'on retrouve les normes *probabilistes* (*produit* et *somme*) en posant $\gamma = 1$ dans les normes de Hamacher.

Standard	$\top$	$\min(\mu_1, \mu_2)$
	$\perp$	$\max(\mu_1, \mu_2)$
Hamacher	$\top$	$\frac{\mu_1\mu_2}{\gamma+(1-\gamma)(\mu_1+\mu_2-\mu_1\mu_2)}$
$(\gamma \geq 0)$	$\perp$	$\frac{\mu_1+\mu_2-\mu_1\mu_2-(1-\gamma)\mu_1\mu_2}{1-(1-\gamma)\mu_1\mu_2}$
Yager	$\top$	$\max\left(1-((1-\mu_1)^m+(1-\mu_2)^m)^{1/m}, 0\right)$
$(m = 1, 2...)$	$\perp$	$\min\left((\mu_1^m+\mu_2^m)^{1/m}, 1\right)$

TAB. 1 – t-normes $(\top)$ et t-conormes $(\perp)$

Dans (Frélicot et al. 2003), nous avons introduit un nouvel opérateur d'agrégation dans le cadre de la classification supervisée avec double option de rejet et en particulier

option de rejet d'ambiguïté. Nous l'avons baptisé *OU-2 flou*, noté $\overset{2}{\perp}$, défini par :

$$\overset{2}{\bigperp}_{i=1,c}\mu_i = \bigtop_{i=1,c}\left(\bigperp_{j=1,c;j\neq i}\mu_j\right) \tag{3}$$

Nous avons également montré que dans le cas où la t-norme utilisée est le *min*, $\overset{2}{\perp}_{i=1,c}\mu_i$ est égal au deuxième plus grand des μ_i. Cet opérateur possède plusieurs propriétés mathématiques dont certaines découlent naturellement de celles des t-normes et t-conormes (ex : bornes, monotonie, continuité, symétrie, etc.) (Frélicot et al. 2003). L'une des propriétés importante du $\overset{2}{\perp}$, dite de *compensation faible*, est la suivante :

Proposition : *pour tout entier* $c \geq 2$ *et tout* $(\mu_1, ..., \mu_c) \in [0,1]^c$*, on a :*

$$\bigtop_{i=1,c}\mu_i \leq \overset{2}{\bigperp}_{i=1,c}\mu_i \leq \bigperp_{i=1,c}\mu_i \tag{4}$$

Un moyen assez naturel de mesurer l'ambiguïté entre les classes est d'effectuer le rapport entre le *deuxième plus grand* et le *plus grand* des μ_i. Utiliser ce rapport n'est pas une idée neuve ; il a été défini pour la première fois dans (Frélicot 1992). La mesure d'ambiguïté que nous proposons ici la généralise :

$$A(\mathbf{x}) = \overset{2}{\bigperp}_{i=1,c}\mu_i(\mathbf{x}) \Big/ \bigperp_{i=1,c}\mu_i(\mathbf{x}) \tag{5}$$

La propriété de compensation faible (4) implique que, pour chaque t-norme $\top$ et sa t-conorme duale $\perp$:

$$A(\mathbf{x}) \leq 1 \tag{6}$$

4.3 Le critère d'évaluation proposé

Nous proposons d'utiliser la mesure d'ambiguïté (5) pour définir le nouveau critère d'évaluation d'un sous-ensemble S_q de q variables suivant :

$$J_A(S_q) = \sum_{\mathbf{x}} A^{[q]}(\mathbf{x}) \tag{7}$$

où l'exposant $^{[q]}$ indique que la mesure d'ambiguïté est définie à partir d'étiquettes $\mu_i^{[q]}$ représentant le degré d'appartenance, donné par exemple par (2), du point $\mathbf{x}$ à la classe ω_i dans l'espace $\mathbb{R}^q$. L'équation (6) permet de borner supérieurement $J_A(S_q)$:

$$J_A(S_q) \leq N \tag{8}$$

où N est le nombre d'exemples dans la base d'apprentissage.

Il s'agit alors d'utiliser un algorithme de recherche, comme par exemple SFFS, pour sélectionner l'ensemble des q variables parmi les p d'origine qui rendent le critère $J_A(S_q)$ minimum.

Si on devait classer notre critère dans une des catégories énoncées à la section 3, ce serait celle des mesures d'information. Cependant, et contrairement à notre critère la plupart des critères d'information (ex : indice de Gini, la fonction gain) définissent des méthodes de sélection univariées : les variables sont supposées indépendantes et sont donc évaluées séparément sans tenir compte d'une éventuelle interaction entre les variables.

5 Résultats et discussion

5.1 Protocole

Nous avons testé l'approche proposée sur onze jeux de données artificiels et réels issus de la base de données UCI (Blake et Merz 1998) ; un résumé en est fait au tableau 2 où N, c et p représentent le nombre d'exemples, le nombre de classes et le nombre de variables respectivement. Nous les avons divisés en trois groupes :

1. **Données artificielles en dimension faible (Gr 1) :** Il s'agit de trois jeux de données, en dimension faible, composés de deux classes, spécialement dédiés à la sélection : *Monk-1* qui contient trois variables pertinentes $\{x_1, x_2, x_5\}$ et trois autres non pertinentes $\{x_3, x_4, x_6\}$, *Monk-2* dont toutes les variables sont pertinentes, *Monk-3* contenant lui aussi trois variables pertinentes $\{x_2, x_4, x_5\}$ et trois qui ne le sont pas $\{x_1, x_3, x_6\}$; dans ce jeu, 5% des instances ont été changées de classe afin de tester la robustesse.
2. **Données réelles en dimension faible (Gr 2) :** Il s'agit de trois jeux de données classiques en classification : *Iris*, *Breast Cancer Wisconsin* et *Pima Indian Diabetes* dont la description complète est donnée à l'adresse : *http ://www.ics.uci.edu/∼mlearn/MLRepository.html*.
3. **Données réelles en dimension moyenne (Gr 3) :** Décrits à la même url, il s'agit de cinq jeux de données moins classiques, mais utilisés en sélection à cause de leur dimension moyenne ($10 < p < 100$) : *Cleveland Heart Disease*, *Dna*, *Ionosphere*, *Sonar*, et *Segment*.

Précisons que les instances avec des valeurs manquantes ont été éliminées. Pour certains jeux de données (Gr 1 et *Sonar*), un ensemble d'apprentissage et un ensemble test sont disponibles ; nous avons fusionné les deux ensembles.

Deux types de résultats sont présentés dans cette section. Nous comparons dans un premier temps les taux de bon classement obtenus, sur les onze jeux, avec les q variables sélectionnées par le critère proposé avec ceux obtenus en utilisant les p variables. Dans un second temps, nous comparons les performances du critère proposé avec d'autres issus de la littérature sur le groupe de données numéro 2. Les dimensions des jeux du groupe de données numéro 3 étant importantes, l'estimation des taux a été réalisée selon une procédure 10-VC (*Validation Croisée*). Pour réduire le biais dû au caractère aléatoire de la construction des ensembles test et apprentissage par la procédure 10-VC, cette dernière est répétée en réalisant dix essais indépendants. Le critère J_A étant de type *filter*, les performances ont été établies pour trois discriminateurs classiques : la

Jeu	**Nom**	N	c	p	q^*
#1	*Monk-1*	556	2	6	3
#2	*Monk-2*	601	2	6	6
#3	*Monk-3*	554	2	6	3
#4	*Iris*	150	3	4	2
#5	*Pima*	768	2	8	-
#6	*Breast*	683	2	9	-
#7	*Cleve*	297	2	13	-
#8	*Segment*	2310	7	19	-
#9	*Ionosphere*	351	2	34	-
#10	*Dna*	318	2	57	-
#11	*Sonar*	208	2	60	-

TAB. 2 – Résumé des jeux de données. q^* indique le nombre de variables pertinentes.

règle des *k-Plus Proches Voisins* (k-PPV), la règle de *Bayes Quadratique* sous hypothèse gaussienne (BQ) et l'algorithme C4.5. Rappelons que la règle BQ correspond à la règle du *Maximum A Posteriori* (MAP) en considérant une matrice de covariance Σ_i propre à chaque classe ω_i. Pour les k-PPV, le nombre k de voisins retenu est celui donnant le meilleur taux de classement dans la plage $[1, 30]$. Enfin, précisons que pour le critère J_A, nous avons utilisé les normes standard et d'Hamacher avec $\gamma = 1$ et $\gamma = 0$. Bien que les résultats soient comparables, ceux reportés correspondent à chaque fois aux normes ayant donné les meilleures performances.

5.2 Validation des sélections de variables

La validation des sous-ensembles de variables sélectionnées par le critère proposé est montrée en comparant, pour chaque jeu de données, les taux de bon classement obtenus avec les p variables d'origine et q variables sélectionnées. Comme nous avons réalisé dix essais de la procédure 10-VC, nous disposons donc de dix taux estimés, nous avons reporté les intervalles de confiance à 95% sur la moyenne de ces taux dans le tableau 3.

De manière générale, les taux de succès obtenus après sélection sont, dans la plupart des cas, meilleurs que ceux obtenus par les p variables et ce pour les trois discriminateurs. Notre critère a été capable de sélectionner toutes les variables pertinentes pour les jeux *Monk-1*, *Monk-3* et *Iris*. En revanche, pour le jeu *Monk-2* il n'a retenu que cinq variables parmi les six variables pertinentes, les performances sont donc plus faibles dans le cas des règles BQ et k-PPV. Une baisse des performances est également enregistrée pour le jeu de données réel *Dna* dans le cas des discriminateurs BQ et k-PPV, ainsi que pour le jeu *Monk-3* dans le cas de la règle des k-PPV, probablement en raison du bruit.

Nous remarquons que la sélection influe beaucoup sur le résultat de la classification pour les deux discriminateurs BQ et k-PPV. En revanche, C4.5 semble peu sensible aux résultats de la sélection dans le cas où les données sont de dimension faible. Pour

Jeu	Nom	Critère	BQ (%)	k-PPV (%)	C4.5 (%)
#1	*Monk-1*	$p = 6$	81.96±0.36	96.67±0.31	97.84±2.56
		$J_A(q = 3)$	83.33±0.52 +	100.0±0.00 +	100.0±0.00 +
#2	*Monk-2*	$p = 6$	75.74±0.44	85.29±0.37	65.74±0.76
		$J_A(q = 5)$	70.28±0.27−	78.48±0.90−	65.74±0.76
#3	*Monk-3*	$p = 6$	91.62±0.12	98.19±0.23	98.92±0.90
		$J_A(q = 3)$	91.73±0.05	93.54± 2.89 −	98.90±0.91
#4	*Iris*	$p = 4$	97.20±0.20	96.26±0.63	94.66±6.28
		$J_A(q = 2)$	97.13±0.32	96.27±0.56	94.64±6.18
#5	*Pima*	$p = 8$	74.06±0.51	74.60±0.57	74.60±3.90
		$J_A(q = 3)$	75.57±0.30 +	75.75±0.51 +	75.63±3.93
#6	*Breast*	$p = 9$	95.17±0.16	97.11±0.24	93.71±1.90
		$J_A(q = 3)$	96.27±0.11 +	96.82±0.20	93.56±1.51
#7	*Cleve*	$p = 13$	82.52±0.35	66.70±1.13	77.15±2.01
		$J_A(q = 8)$	82.19±0.59	76.36±0.77 +	79.86±2.46 +
#8	*Segment*	$p = 19$	81.15±1.06	81.78±1.17	96.08±0.51
		$J_A(q = 4)$	89.83±0.08 +	93.57±0.16 +	97.09±0.90 +
#9	*Ionosphere*	$p = 34$	89.23±0.37	86.50±0.35	89.18±1.51
		$J_A(q = 17)$	91.20±0.28 +	88.43±0.40 +	90.89±0.73 +
#10	*Dna*	$p = 57$	97.96±0.60	88.46±0.60	81.45±7.48
		$J_A(q = 37)$	97.04±0.60−	86.67±0.88 −	83.01±7.02
#11	*Sonar*	$p = 60$	76.59±1.58	81.30±0.78	73.49±4.26
		$J_A(q = 30)$	85.48±0.58 +	84.33±0.63 +	78.93±3.86 +

Amélioration (+) ou dégradation (−) statistiquement significative (95%)

TAB. 3 – Moyennes et intervalles de confiance des taux de bon classement obtenus avec la procédure *10-VC*.

un nombre de variables plus élevé ($10 < p < 100$) nous avons noté une amélioration des performances pour quatre jeux de données parmi cinq. Rappelons que C4.5 opère une sélection des variables lors de la construction des arbres de décision. C'est sans doute la raison pour laquelle des jeux ont donné des performances similaires avec et sans sélection.

5.3 Comparaison avec d'autres critères

Nous avons comparé notre critère avec d'autres appartenant à différentes catégories de critères citées dans la section 3 :

1. **La distance probabiliste de Mahalanobis** ($Maha$).
2. **Distance floue** (Campos et al. 2001) :

$$DF(S_q) = -\sum_i \sum_j \left(\sum_x d_{\mathbf{x}}^{[q]} (\omega_i, \omega_j)^2 \right)^{1/2} \tag{9}$$

où $d_{\mathbf{x}}^{[q]}(\omega_i, \omega_j)$ est la distance entre les classes ω_i et ω_j en dimension q dans une boule $\mathcal{B}$ de diamètre τ centrée en $\mathbf{x}$, définie par :

$$d_{\mathbf{x}}^{[q]}(\omega_i, \omega_j) = inf_{\mathbf{y},\mathbf{z} \in \mathcal{B}(\mathbf{x},\tau)} \left| \mu_i^{[q]}(\mathbf{y}) - \mu_j^{[q]}(\mathbf{z}) \right| \tag{10}$$

où $\mu_i(\mathbf{y})$ est défini par (2) avec une distance Euclidienne si $\mathbf{y} \in \omega_j$, dans le cas contraire il est égal à 0. Le paramètre τ a été fixé à 0.5 dans tout les tests.

3. **Entropie** (Mitra et al. 2002) :

$$E(S_q) = -\sum_{\mathbf{x}} \sum_{\mathbf{y}} \left[S^{[q]}(\mathbf{x},\mathbf{y}) \ \log S^{[q]}(\mathbf{x},\mathbf{y}) + (1 - S^{[q]}(\mathbf{x},\mathbf{y})) \ \log(1 - S^{[q]}(\mathbf{x},\mathbf{y})) \right] \quad (11)$$

où $S^{[q]}(\mathbf{x},\mathbf{y}) = e^{-\alpha D^{[q]}(\mathbf{x},\mathbf{y})}$ est une mesure de similarité entre $\mathbf{x}$ et $\mathbf{y}$ avec q variables, α est une constante positive et D la distance définie par :

$$D^{[q]}(\mathbf{x},\mathbf{y}) = \left[\sum_{l=1}^{q} \left(\frac{x_l - y_l}{max_l - min_l} \right)^2 \right]^{\frac{1}{2}} \quad (12)$$

où max_l (resp. min_l) est la valeur maximum (resp. minimum) de la l-ème variable. L'entropie est une mesure d'information qui est maximale lorsque les données sont uniformément distribuées.

Nous avons mené des tests pour comparer les taux de bon classement obtenus avec les variables sélectionnées pour les différents critères et avec les trois discriminateurs BQ, k-PPV et C4.5. Nous avons reporté, à titre d'exemple, dans le tableau 4 les taux de bon classement obtenus en utilisant la règle BQ. L'analyse des résultats nous a permis de dégager les remarques suivantes :

- Hormis les jeux artificiels pour lesquels le critère $Maha$ n'a pas été capable de sélectionner toutes les variables pertinentes, les performances obtenues par J_A et $Maha$ sont de manière générale comparables, quel que soit le discriminateur.
- Excepté pour le jeu *Monk-3*, les performances obtenues par le critère J_A sont meilleures que celles obtenues par DF en utilisant les règles BQ et k-PPV. Dans le cas de la règle C4.5, les résultats obtenus avec les deux critères sur les jeux artificiels et réels de dimension faible sont proches. Pour les jeux réels de dimension moyenne, les performances du critère J_A sont meilleures que celles de DF.
- Contrairement au critère J_A, la mesure E n'a pas été capable de sélectionner les variables pertinentes pour les trois jeux artificiels ; les performances obtenues par J_A sont donc largement meilleures. Les deux variables pertinentes ont été retenues par les deux critères pour le jeu $Iris$, les taux de bon classement sont alors identiques. Pour les autres jeux les performances de J_A sont en général meilleures que celles de E, quel que soit le discriminateur.

6 Conclusion

Dans le cadre de la sélection de variables en classification supervisée, nous avons présenté une mesure d'ambiguïté permettant de définir un nouveau critère d'évaluation d'un (sous-)ensemble de variables. Cette mesure est fondée sur une combinaison d'opérateurs d'agrégation d'étiquettes floues/possibilistes représentant le degré d'appartenance aux classes en présence. Le critère proposé peut être associé à n'importe

Jeux	p **variables**	J_A	$Maha$	DF	E
Monk-1	81.96±0.36	83.33±0.52+	66.55±0.00−	82.55±0.42	47.09±1.48−
Monk-2	75.74±0.44	70.28±0.27−	62.68±0.53−	62.59±0.56−	63.84±0.18−
Monk-3	91.62±0.12	91.73±0.05	90.70±0.50−	92.33±0.22+	46.57±1.00−
Iris	97.20±0.20	97.13±0.32	97.40±0.15	94.67±0.42−	97.07±0.33
Pima	74.06±0.51	75.57±0.30+	75.40±0.17+	74.39±0.25	74.73±0.22+
Breast	95.17±0.16	96.27±0.11+	95.99±0.18+	95.23±0.07	94.73±0.17−
Cleve	82.52±0.35	82.19±0.59	80.57±0.80−	74.04±0.62−	81.72±0.47−
Segment	81.15±1.06	89.83±0.08+	85.45±3.32+	80.78±2.86	70.29±2.70−
Ionosphere	89.23±0.37	91.20±0.28+	90.79±0.30+	89.57±0.22	89.49±0.31
Dna	97.96±0.60	97.04±0.60−	97.33±1.08	96.63±0.64−	94.59±1.27−
Sonar	76.59±1.58	85.48±0.58+	83.27±1.55+	79.61±1.15+	68.12±0.76−

Amélioration (+) ou dégradation (−) statistiquement significative (95%)

TAB. 4 – Taux de bon classement obtenus dans le cas de la règle BQ.

quel discriminateur. Ses performances ont été comparées avec d'autres critères issus de la littérature. Les tests menés sur de nombreux jeux de données réels et artificiels ont montré que le critère proposé est capable de sélectionner les variables pertinentes et d'augmenter dans la plupart des cas les taux de bon classement. Les perspectives de ce travail concernent l'étude des propriétés mathématiques de la mesure d'ambiguïté et la définition de nouvelles mesures à partir de normes de base autres que des normes triangulaires.

Références

Bezdek J. (1987), Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 2nd edition, 1987.

Blake C.L. et Merz C.J. (1998), UCI repository of machine learning databases, 1998.

Breiman L., Friedman J.H., Olshen R.A. et Stone C.J. (1984), Classification and Regression Trees, Kluwer Academic Publishers, 1984.

Campos T.E., Bloch I. et Cesar Jr. R.M. (2001), Feature Selection Based on Fuzzy Distances between Clusters : First Results on Simulated Data, In : Lecture Notes in Comp. Sci., 2001.

Dash M., Liu H. et Motoda H. (2000), Consistency based feature selection, Proc. of Pacific Asia Conf. on Knowledge Discovery and Data Mining, pp. 98–109, 2000.

Devijver P.A. et Kittler J. (1982), Pattern Recognition : A Statistical Approach, Prentice-Hall, London, 1982.

Dubois D. et Prade H. (1985), A review of fuzzy set aggregation connectives, Information Sciences, 36 :85–121, 1985.

Frélicot C. (1992), Un système adaptatif de diagnostic prédictif par reconnaissance des formes floue, Thèse de doctorat, Université de Technologie de Compiègne, 1992.

Frélicot C., Fruchard A. et Mascarilla L. (2003), Une classe d'opérateurs pour la mesure d'ambiguïté, Rencontres Francophones sur la Logique Floue et ses Applications (LFA), pp. 123–130, 2003.

Hall M.A. (2000), Correlation based feature selection for discrete and numeric class machine learning, In : 17th Int. Conf. on Machine Learning, 2000.

Jain A. et Zongker D. (1997), Feature selection : Evaluation, application and small sample performance, IEEE Trans. on PAMI, 19(2) :153–158, 1997.

Kittler J. (1986), Feature selection and extraction, In : Handbook of Pattern Recognition and Image Processing (Y. Fu, Edition), Academic Press, New York, 1986.

Klir G.J. et Yuan B. (1995), Fuzzy Sets and Fuzzy Logic : Theory and Applications, Prentice Hall, 1995.

Kohavi R. et John G.H. (1997), Wrappers for feature subset selection, Artificial Intelligence, 97(1-2) :273–324, 1997.

Koller D. et Sahami M. (1996), Toward optimal feature selection, In : 13th Int. Conf. on Machine Learning, pp. 284–292, 1996.

Krishnapuram, R. et Keller J.M. (1993), A possibilistic approach to clustering, IEEE Trans. on Fuzzy systems, 1(2) :98–110, May 1993.

Kudo M. et Sklansky J. (2000), Comparison of algorithms that select features for pattern classifiers, Pattern Recognition, 33(1) :25–41, January 2000.

Langley P. (1994), Selection of relevant features in machine learning, In AAAI Fall Symposium on Relevance, pp. 140–144, 1994.

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

Liu H. et Setiono R. (1995), Chi2 : Feature selection and discretization of numeric attributes, IEEE Int. Conf. on Tools with Artificial Intelligence, pp. 388–391, 1995.

Mitra P., Murthy C.A. et Pal S.K. (2002), Unsupervised feature selection using feature similarity, IEEE Trans. on PAMI, 24(3) :301–312, March 2002.

Narendra P.M. et Fukunaga K. (1977), A branch and bound algorithm for feature subset selection, IEEE Trans. on Computers, 26(9) :917–922, 1977.

Pfahringer, B. (1995), Compression-based feature subset selection, Knowledge Discovery and Data Mining, pp. 234–239, 1995.

Pudil P., Novovicová J. et Kittler J. (1994), Floating search methods in feature selection, Pattern Recognition Letters, 15 :1119–1125, 1994.

Quinlan J.R. (1986), Induction of decision trees, Machine Learning, 1(1) :81–106, 1986.

Torkkola K. (2003), Feature extraction by non-parametric mutual information maximization, Journal of Machine Learning Research, 3 :1415–1438, 2003.

Summary

This paper addresses the feature selection problem for supervised classification. Feature selection methods are based on a selection algorithm and a criterion function assessing how effective feature subsets are. We propose an ambiguity measure that allows to define a new evaluation criterion. It is based on a combination of labels representing the degree of typicality to the classes. The new criterion is compared to others found in the literature on various real and artificial data sets.