Combinaison d'étiquettes floues/possibilistes pour la sélection de variables

Fuzzy/possibilistic labels combination for feature selection

Dahbia Semani Carl Frélicot Pierre Courtellemont

Laboratoire d'Informatique – Image – Interaction (UPRES EA 2118),
Université de La Rochelle, avenue M. Crépeau, 17042 La Rochelle Cedex 1, France
{dahbia.semani,carl.frelicot,pierre.courtelllemont}@univ-lr.fr

Résumé

Cet article aborde le problème de la sélection de variables en classification supervisée. Les méthodes de sélection sont fondées sur un algorithme de sélection et un critère d'évaluation pour mesurer la pertinence des sous-ensembles de variables. Nous présentons une mesure d'ambiguïté permettant de définir un nouveau critère. Cette mesure est fondée sur une combinaison d'étiquettes représentant le degré de typicalité ou d'appartenance aux classes en présence. Le nouveau critère est comparé, sur des jeux de données réelles, avec d'autres critères de la littérature.

Mots Clef

Sélection de variables, critères d'évaluation, mesure d'ambiguïté, opérateurs flous.

Abstract

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 data sets.

Keywords

Feature selection, Evaluation criteria, Ambiguity measure, Fuzzy operators.

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* est l'une des plus importantes techniques de pre-traitement de données qui permet de réduire la taille des informations à traiter et faciliter l'étape d'apprentissage. Des traitements plus sophistiqués peuvent alors être utilisés dans des espaces de dimension réduite, les performances peuvent augmenter lorsque les variables non pertinentes ou redondantes disparaissent, etc.
Dans cet article, nous traitons 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 comprend généralement les trois étapes suivantes :

- un algorithme de recherche, permettant d'explorer l'espace des combinaisons de variables,
- un critère d'évaluation, pour comparer différents, sous-ensembles de variables,
- un critère d'arrêt.

Nous nous intéressons, dans cet article, aux critères d'évaluation. Ainsi, nous proposons une mesure d'ambiguïté permettant de définir un nouveau critère. Cette mesure est fondée sur une combinaison d'étiquettes représentant le degré de typicalité 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 état de l'art sur les algorithmes de sélection de variables et les critères d'évaluation sont dressés 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 *sélection* ou de recherche de sous-ensembles de variables ont été proposés dans la littérature et des études comparatives existent [9, 23]. Ces algorithmes diffèrent selon qu'ils sont appliqués à un domaine particulier (reconnaissance de formes statistique, structurelle, etc), selon la nature de la solution trouvée (optimale ou non, unique ou non, selon la nature de l'algorithme utilisé (déterministe, paramétrique, etc). En reconnaissance de formes statistique, les algorithmes de sélection sont généralement groupés en deux catégories: les algorithmes optimaux garantissant une solution optimale du problème, et les algorithmes sous optimaux [16].
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 [8]. La seule alternative permettant de trouver la solution optimale est l'algorithme *Branch and Bound* (B&B)[27]. Cet algorithme exige que le critère d'évaluation soit monotone avec le nombre de variables sélectionnées mais la plupart d'entre eux ne satisfont pas cette condition. Plusieurs versions de l'algorithme B&B, ont été proposés [34, 40] ; ils restent le plus souvent inapplicables en grande dimension ($p > 10^2$) car la complexité dans leur pire des cas est exponentielle.
Des heuristiques basées sur des parcours séquentiels leur sont souvent préférées. Elles consistent à rajouter ou à éliminer itérativement des variables [12]. 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, comme elles n'explorent pas tous les sous-ensembles possibles de variables et ne permettent pas de retour arrière pendant la recherche ; elles sont donc sous-optimales. 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. L'algorithme *"plus l- take away r"* [36] est une généralisation de cette méthode. L'algorithme commence par élargir le sous-ensemble de variables en répétant l fois la procédure SFS, et puis d'en éliminer des variables en répétant r fois la procédure SBS. Le choix des paramètres l et r influe beaucoup sur la qualité des résultats ainsi que le temps de calcul. Les méthodes flottantes (*Sequential Floating Search methods*, [30]) sont une extension de l'algorithme *"plus l-take away r"*. L'acronyme des versions *forward* et *backward* de ces méthodes sont respectivement SFFS et SBFS. Elles sont considérées comme les méthodes sous-optimales les plus efficaces [16] et supportent l'utilisation de critères d'évaluation non monotones [29]. L'algorithme SFFS consiste à appliquer après chaque étape *forward* autant d'étapes *backward* que le sous-ensemble de variables F_q correspondant améliore le critère d'évaluation $J(F_q)$ à ce niveau de recherche:

Algorithme 1: L'algorithme SFFS

Données: $V = \{x_j \mid j = 1, ..., p\}$
// V : ensemble de variables initial //
// J : critère (à maximiser par exemple) //
Sortie: $F_q = \{y_j \in V \mid j = 1, ..., q\},\ q \leq p$
// F_q : sous-ensemble de q variables sélectionnées //

Initialisation:
$q := 0$
$F_q := \emptyset$

Étape 1 (*forward*)
$x_+ :=$argmax$_{x_j \in V \setminus F_q} J(F_q \cup \{x_j\})$
$F_{q+1} := F_q \cup \{x_+\}$
$q := q + 1$;

Étape 2 (*backward*)
$x_- :=$argmax$_{y_j \in F_q} J(F_q \setminus \{y_j\})$
si $J(F_q \setminus \{x_-\}) > J(F_{q-1})$ **alors**
    $F_{q-1} := F_q \setminus \{x_-\}$;
    $q := q - 1$;
    Aller à **Étape 2**
sinon
    Aller à **Étape 1**
fin

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 $J(F_q)$ ou sur $J(F_q) - J(F_{q-1})$. 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. Une version adaptative des méthodes séquentielles flottantes a été proposée [35]. Dans celle-ci, les étapes *forward* et *backward* sont répétées un nombre de fois variable, ce nombre étant dynamiquement contrôlé par la valeur du critère. Les deux versions, *Adaptive SFFS* (ASFFS) et ASBFS généralisent les méthodes SFFS et SBFS. Elles permettent de se rapprocher de la solution optimale malheureuse-

ment au prix d'une complexité accrue. Tous les algorithmes mentionnés ci-dessus sont déterministes, c'est à dire, qu'ils produisent le même sous-ensemble de variables pour un problème donné à chaque exécution de l'algorithme. Des algorithmes génétiques, donc aléatoires, ont également été proposés [32] ; ils sont comparables en performance mais très coûteux en temps [23]. 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 (voir Algorithme 1).

3 Critères d'évaluation

Deux approches sont couramment utilisées pour évaluer la pertinence d'un sous-ensemble de variables sélectionnées [24] : celles fondées uniquement sur les données (*filter*) et celles qui tiennent également compte de la règle de classement utilisée après la phase de sélection (*wrapper*).
Pour ces dernières, présentées longuement dans [17, 20], les paramètres nécessaires à la règle de classement sont estimés à partir de l'ensemble de variables sélectionnées et le critère d'évaluation est simplement la probabilité d'erreur. Celle-ci est 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 [20][3]. Il est évident que la probabilité d'erreur estimée est un critère d'évaluation particulièrement bien adapté aux problèmes de classification. Cependant, elle diminue de fait la capacité à généraliser des méthodes de type *wrapper* puisqu'elles sont attachées à un couple (données, règle).
Dans cet article nous nous intéressons aux mesures d'évaluation de type *filter*, qui sont généralement divisées en plusieurs catégories :

1. Mesures de distance probabiliste :
Appelées aussi *mesures de discrimination/divergence*, elles maximisent les distances probabilistes (ex : de Mahalanobis, de Battacharyya ou de divergence [37]) entre les fonctions de densités ou entre les probabilités a posteriori des classes. L'hypothèse de distribution gaussienne est souvent nécessaire.

2. Mesures de distance inter-calsses :
Ces mesures sont fondées sur des transformations scalaires des matrices de covariance inter – et intra – classes (ex : le lambda de Wilks [18]). Elles ne requièrent pas d'hypothèses probabilistes et sont bien adaptées au cas où les classes ont des distrbutions de même matrice de covariance.

3. Mesures d'information :
Issues de la théorie de l'information [7], elles évaluent le gain apporté part les variables via les probabilités a posteriori (ex : entropie de Shannon ou dérivées [1, 21]). Si ces probabilités tendent vers une valeur égale quelle que soit la classe, le gain est minimal et l'incertitude est maximale.

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 [15], information mutuelle [39, 38]). Il a été montré que toute mesure de dépendance peut être reformulée comme une mesure de catégorie 1 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 [10], les tests statistiques [25, 37]).

4 Un nouveau critère d'évaluation

Illustrons l'approche que nous proposons par un exemple en dimension $p = 2$. Soient les deux classes non séparables ω_1 ($\triangle$) et ω_2 ($\star$) représentées à la figure 1. Les projections des classes sur les 2 variables x_1 et x_2 montrent un chevauchement des classes plus important sur x_1 que sur x_2. La variable la plus significative pour le problème de classification supervisée est donc x_2.

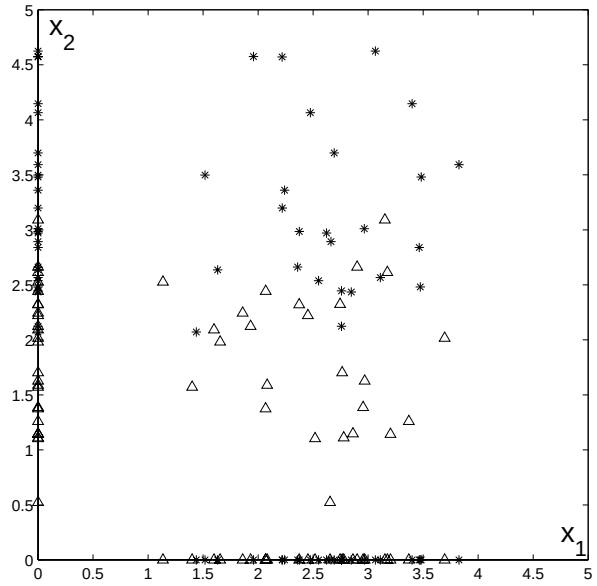

Figure 1: Deux classes ω_1 ($\triangle$) et ω_2 ($\star$) et leurs projections sur x_1 et x_2

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 des données $\mathbf{x}$ aux classes.

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 $\mu_k(\mathbf{x})$ représente le degré d'appartenance de $\mathbf{x}$ à la classe ω_k. D'un point de vue mathématique et non sémantique, les étiquettes sont *floues* si $\sum_{l=1}^{c} \mu_l(\mathbf{x}) = 1$, comme par exemple la fonction utilisée dans les méthodes de classification non supervisée de type FCM (*Fuzzy C-Means*, [2]) :

$$\mu_k(\mathbf{x}) = 1 \Big/ \sum_{l=1}^{c} \left(\frac{d(\mathbf{x}, \mathbf{p}_k)}{d(\mathbf{x}, \mathbf{p}_l)} \right)^{2/(m-1)} \quad (1)$$

où m est un paramètre à fixer dans l'intervalle $]1, +\infty[$ et $d(\mathbf{x}, \mathbf{p}_k)$ est une distance entre $\mathbf{x}$ et un prototype de la classe ω_k ; 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 typicalité [22], comme par exemple :

$$\mu_k(\mathbf{x}) = \frac{\lambda_k}{\lambda_k + d(\mathbf{x}, \mathbf{p}_k)} \quad (2)$$

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

$$d^2(\mathbf{x}, \mathbf{p}_k) = (\mathbf{x} - \mathbf{p}_k)^t \, \Sigma_k^{-1} \, (\mathbf{x} - \mathbf{p}_k) \quad (3)$$

où Σ_k est la matrice de covariance de la classe ω_k.

4.2 La mesure d'ambiguïté

Après l'étape d'étiquetage, nous disposons, pour chaque vecteur $\mathbf{x}$, d'un vecteur d'étiquettes $\mu_k, (k = 1, c)$. Nous voulons quantifier l'ambiguïté entre les classes en combinant les étiquettes μ_k. 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 [11, 13]. Nous avons choisi d'utiliser les normes (et conormes) triangulaires ou *t-normes* (et *t-conormes*).

Définition 1.
Une t-norme est une fonction $\top : [0,1] \times [0,1] \rightarrow [0,1]$ *qui satisfait les axiomes suivants :* $\forall \mu, \nu, \eta \in [0,1]$

$$\mu \top \nu = \nu \top \mu \quad (4)$$
$$\nu \leq \eta \Rightarrow \mu \top \nu \leq \mu \top \eta \quad (5)$$
$$\mu \top (\nu \top \eta) = (\mu \top \nu) \top \eta \quad (6)$$
$$\mu \top 1 = \mu \quad (7)$$

Définition 2.
La t-conorme duale est la fonction $\perp : [0,1] \times [0,1] \rightarrow [0,1]$ *définie par :*

$$\mu \perp \nu = 1 - (1 - \mu) \top (1 - \nu) \quad (8)$$

et satisfait donc les axiomes (4), (5), (6) et :

$$\mu \perp 0 = \mu \quad (9)$$

Les axiomes (5), (7) et (9) impliquent :

$$\mu \top \nu \leq \mu \quad (10)$$
$$\mu \leq \mu \perp \nu \quad (11)$$

ce qui a pour conséquences que :

- 0 est absorbant pour une t-norme, 1 pour une t-conorme,
- *min* est la plus grande des t-normes, *max* est la plus petite des t-conormes :

$$\mu \top \nu \leq min(\mu, \nu) \leq max(\mu, \nu) \leq \mu \perp \nu \quad (12)$$

Ces opérateurs peuvent être vus 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 [11, 19]. 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 *probabilisites* (*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	$\top$	$\max\left(1 - ((1-\mu_1)^m + (1-\mu_2)^m)^{1/m}, 0\right)$
$(m = 1, 2...)$	$\perp$	$\min\left((\mu_1^m + \mu_2^m)^{1/m}, 1\right)$

Table 1: t-normes ($\top$) et t-conormes ($\perp$)

Dans [33], nous avons introduit un nouvel opérateur d'agrégation, que nous avons baptisé *OU-2 flou*, noté $\underset{2}{\perp}$, défini par :

$$\overset{2}{\bigperp}_{k=1,c} \mu_k(x) = \bigperp_{k=1,c} \left(\mu_k(x) \top \bigperp_{l=1,c,l\neq k} \mu_l(x) \right) \quad (13)$$

dans le cadre de la classification supervisée avec double option de rejet et en particulier option de rejet d'ambiguïté.

Nous avons depuis montré que :

- dans le cas où la t-norme est le min, $\overset{2}{\perp}_{k=1,c}\mu_k$ est égal au deuxième plus grand des μ_k,
- il existe d'autres opérateurs $\overset{2}{\perp}$ partageant cette propriété,
- $\overset{2}{\perp}$ possède les propriétés de bornes suivantes:

$$\overset{2}{\bigperp}_{k=1,c} 0 = 0 \text{ et } \overset{2}{\bigperp}_{k=1,c} 1 = 1 \quad (14)$$

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

$$\forall(\top, \perp), \quad \bigtop_{k=1,c} \mu_k \leq \overset{2}{\bigperp}_{k=1,c} \mu_k \leq \bigperp_{k=1,c} \mu_k \quad (15)$$

que nous appellerons *compensation faible*, propriété que l'on peut rapprocher de (12).

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_k. Utiliser ce rapport n'est pas une idée neuve ; nous l'avons définie pour la première fois dans [14]. La mesure d'ambiguïté que nous proposons ici la généralise :

$$A(\mathbf{x}) = \overset{2}{\bigperp}_{k=1,c} \mu_k(\mathbf{x}) \Big/ \bigperp_{k=1,c} \mu_k(\mathbf{x}) \quad (16)$$

Notons que la compensation faible (15) implique que :

$$\forall(\top, \perp), \quad A(\mathbf{x}) \leq 1 \quad (17)$$

La figure 2 montre les isosurfaces de $A(\mathbf{x})$ défini pour $c = 2$ étiquettes dans le cas des normes de Hamacher avec $\gamma = 1$, c'est-à-dire le *produit* comme t-norme et la *somme probabiliste* comme t-conorme. Les zones claires correspondant aux valeurs élevées de $A(\mathbf{x})$ traduisent bien une situation d'ambiguïté au sens du classement.

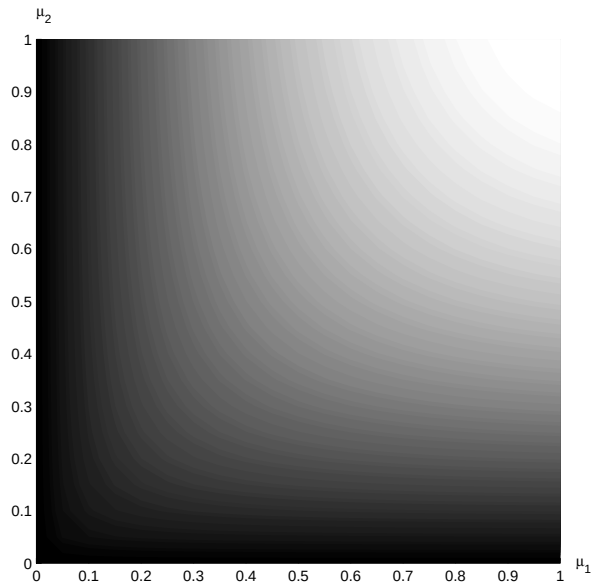

Figure 2: Isosurfaces de la mesure d'ambiguïté – cas de 2 étiquettes

4.3 Le critère d'évaluation proposé

Nous proposons d'utiliser la mesure d'ambiguïté (16) pour définir le nouveau critère d'évaluation d'un sous-ensemble de q variables suivant :

$$FA(q) = \sum_{\mathbf{x}} A^{[q]}(\mathbf{x}) \quad (18)$$

où l'exposant $^{[q]}$ indique que la mesure d'ambiguïté est définie à partir d'étiquettes $\mu_k^{[q]}$ représentant le degré d'appartenance, donné par exemple par (2), du point $\mathbf{x}$ à la classe ω_k dans l'espace de représentation restreint à q variables. L'équation (17) permet de borner supérieurement $FA(q)$:

$$FA(q) \leq N \quad (19)$$

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 $FA(q)$ minimum. Sur les données présentées à la figure 1, ce critère suggère de sélectionner la variable x_2 puisqu'on obtient les résultats suivants : $FA(\{x_2\}) < FA(\{x_1, x_2\}) < FA(\{x_1\})$.

Si on devait classer ce 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 [5], la fonction gain [31]) 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 11 jeux de données artificiels et réels issus de la base de données UCI [4] ; un résumé en est fait au tableau 2 où N, c et p représentent le nombre d'instances, le nombre de classes et le nombre de variables respectivement. Nous les avons divisés en trois groupes :

Gr 1. Données artificielles en dimension faible
Il s'agit de 3 jeux de données, en dimension faible, composés de 2 classes, spécialement dédiés à la sélection : *Monk-1* qui contient 3 variables pertinentes $\{x_1, x_2, x_5\}$ et 3 autres non pertinentes $\{x_3, x_4, x_6\}$, *Monk-2* dont toutes les variables sont pertinentes, *Monk3* contenant lui aussi 3 variables pertinentes $\{x_2, x_4, x_5\}$ et 3 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.

Gr 2. Données réelles en dimension faible
Il s'agit de 3 jeux de données classiques en classification : *Iris*, *Breast Cancer Wisconsin* et *Pima Indian Diabetes* dont la description complète est donnée à *http://www.ics.uci.edu/∼mlearn/MLRepository.html*

Gr 3. Données réelles en dimension élevée
Décrits à la même url, il s'agit de 5 jeux de données moins classiques, mais utilisés en sélection à cause de leur dimension élévée : *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 performances du critère proposé avec d'autres issus de la littérature sur le groupe de données numéro 2. Les nombres de classes et de variables étant raisonnables pour ces jeux, les taux de bon classement ont été estimés selon une procédure *Leave-One-Out* (LOO). Dans un second temps, nous comparons les taux de bon classement obtenus, sur les 11 jeux, avec les q variables sélectionnées par le critère proposé avec ceux obtenus en utilisant les p variables. Les dimensions des jeux du goupe de données numéro 3 étant importantes, l'estimation des taux a été réalisée selon une procédure 10-CV (*Validation Croisée*). Pour réduire le biais dû au caractère aléatoire de la construction des ensembles test et apprentissage par la procédure 10-CV, cette dernière est répétée en réalisant 10 essais indépendants. Le critère FA étant de type *filter*, les performances ont été établies pour deux discriminateurs classiques : la règle des *k-Plus Proches Voisins* (k-PPV) et la règle de *Bayes Quadratique* sous hypothèse gaussienne (BQ). 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 FA, 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.

Jeu	Nom	N	c	p
#1	$Monk-1$	556	2	6
#2	$Monk-2$	601	2	6
#3	$Monk-3$	554	2	6
#4	*Iris*	150	3	4
#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

Table 2: Résumé des jeux de données

5.2 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. **Entropie** [26]:

$$E(q) = - \sum_{\mathbf{x}} \sum_{\mathbf{y}} \Big[S^{[q]}(\mathbf{x},\mathbf{y}) \ \log S^{[q]}(\mathbf{x},\mathbf{y}) + (1 - S^{[q]}(\mathbf{x},\mathbf{y})) \ \log(1 - S^{[q]}(\mathbf{x},\mathbf{y})) \Big] \quad (20)$$

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 :

$$D^{[q]}(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{q} \left(\frac{x_j - y_j}{max_j - min_j} \right)^2 \right]^{\frac{1}{2}}, \quad (21)$$

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

2. **Index d'évaluation flou** [28]:

$$IF(q) = \frac{1}{N(N-1)} \sum_{\mathbf{x}} \sum_{\mathbf{y}\neq\mathbf{x}} \Big[\mu_{\bullet}^{[q]}(\mathbf{x},\mathbf{y})\ (1-\mu_{\bullet}^{[p]}(\mathbf{x},\mathbf{y})) + \mu_{\bullet}^{[p]}(\mathbf{x},\mathbf{y})\ (1-\mu_{\bullet}^{[q]}(\mathbf{x},\mathbf{y}))\Big] \quad (22)$$

où $\mu_{\bullet}^{[p]}(\mathbf{x},\mathbf{y})$ et $\mu_{\bullet}^{[q]}(\mathbf{x},\mathbf{y})$ sont les degrés d'appartenance de $\mathbf{x}$ et $\mathbf{y}$ à la même classe dans les espaces respectifs de dimension p et q. Le degré $\mu_{\bullet}(\mathbf{x},\mathbf{y})$ peut être défini par $1-\frac{d(\mathbf{x},\mathbf{y})}{d_m}$ si $d(\mathbf{x},\mathbf{y}) \leq d_m$, 0 sinon; $d(\mathbf{x},\mathbf{y})$ est la distance euclidienne et $d_m = \max_{\mathbf{x},\mathbf{y}} d(\mathbf{x},\mathbf{y})$. IF appartient à la catégorie des mesures de distance inter-classes ; il diminue avec cette distance inter-classes.

3. **Distance floue** [6]:

$$DF(q) = -\sum_k \sum_l \left(\sum_x d_{\mathbf{x}}^{[q]}(\omega_k,\omega_l)^2\right)^{1/2} \quad (23)$$

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

$$d_{\mathbf{x}}^{[q]}(\omega_k,\omega_l) = inf_{\mathbf{y},\mathbf{z}\in\mathcal{B}(\mathbf{x},\tau)} \left|\mu_k^{[q]}(\mathbf{y}) - \mu_l^{[q]}(\mathbf{z})\right| \quad (24)$$

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

Nous avons reporté dans le tableau 3 les taux de bon classement obtenus avec le même nombre q de variables sélectionnées pour les différents critères FA, E, IF et DF ; les résultats sur le jeu *Iris* ne sont pas donnés car égaux pour tous les critères. Les meilleurs résultats pour le critère proposé ont été obtenus pour les normes d'Hamacher avec $\gamma = 1$ pour *Pima* et $\gamma = 0$ pour *Breast*. Sur ces deux jeux, le critère que nous proposons a conduit à un meilleur taux de bon classsement que pour les autres critères, quel que soit le discriminateur. On constate même un taux plus élevé sur les données *Pima* avec les $q = 3$ variables qu'avec l'ensemble des $p = 8$ variables.

5.3 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 10 essais de la procédure 10-CV, nous disposons donc de 10 taux estimés, nous avons reporté les intervalles de confiance à 95% sur la moyenne de ces taux dans le tableau 4.

Jeux	Critères	BQ (%)	k-PPV (%)
Pima $p=8$ $q=3$	p variables	73.96	76.30
	FA	**75.91**	**77.34**
	E	67.58	67.45
	IF	68.36	76.76
	DF	67.84	66.93
Breast $p=9$ $q=3$	p variables	95.02	97.36
	FA	**96.05**	97.07
	E	94.73	95.75
	IF	94.58	96.93
	DF	93.56	94

Table 3: Taux de bon classement – *LOO*.

Jeux	Critères	BQ (%)	k-PPV (%)
$Monk-1$	$p=6$	81.96 ±0.36	96.67 ±0.31
	$FA(q=3)$	**83.36 ±0.36**	**100 ±0.00**
$Monk-2$	$p=6$	75.74 ±0.44	85.29 ±0.37
	$FA(q=6)$	**75.74 ±0.44**	**85.29 ±0.37**
$Monk-3$	$p=6$	91.62 ±0.12	98.19 ±0.23
	$FA(q=3)$	90.58 ±0.50	95.40 ±0.41
Iris	$p=4$	97.20 ±0.20	96.26 ±0.63
	$FA(q=2)$	**97.13 ±0.32**	**96.27 ±0.56**
Pima	$p=8$	74.06 ±0.51	74.60 ±0.57
	$FA(q=3)$	**75.57 ±0.30**	**75.75 ±0.51**
Breast	$p=9$	95.17 ±0.16	97.11 ±0.24
	$FA(q=3)$	**96.27 ±0.11**	96.82 ±0.20
Cleveland	$p=13$	82.52 ±0.35	66.70 ±1.13
	$FA(q=5)$	**82.95 ±0.46**	**82.22 ±0.82**
Segment	$p=19$	81.15 ±1.06	81.78 ±1.17
	$FA(q=4)$	**89.83 ±0.08**	**93.57 ±0.16**
Ionosphere	$p=34$	89.23 ±0.37	86.50 ±0.35
	$FA(q=8)$	**92.36 ±0.23**	**90.17 ±0.49**
Dna	$p=57$	97.96 ± 0.60	88.46 ±0.60
	$FA(q=25)$	96.48 ±0.59	86.67 ±0.88
Sonar	$p=60$	76.59 ±1.58	81.30 ±0.78
	$FA(q=23)$	**85.05 ±1.23**	**83.12 ±1.04**

Table 4: Moyennes et intervalles de confiance des taux de bon classement – *10-CV*

Gr 1. Données artificielles en dimension faible - *Monk-1* : les 3 variables sélectionnées sont les 3 variables pertinentes $\{x_5, x_2, x_1\}$. Les taux de bon classement sont logiquement meilleurs qu'avec les 6 variables, quel que soit le discriminateur ; il atteint même 100% dans le cas du discriminateur k-PPV.

- *Monk-2* : les 6 variables, toutes pertinentes, ont été retenues.
- *Monk-3* : là encore, les 3 variables sélectionnées sont les 3 variables pertinentes $\{x_5, x_2, x_4\}$. Les taux obtenus sont cependant plus faibles qu'avec les 6 variables, probablement en raison du bruit.

Gr 2. Données réelles en dimension faible

- *Iris* : les deux variables $\{x_3, x_4\}$ connues comme étant les plus discriminantes pour ce jeu de données, ont été sélectionnées. Les taux de bon classement sont légèrement meilleurs qu'avec les 4 variables, pour les deux discriminateurs.
- *Pima*, *Breast* : les commentaires faits à la sous-section précédente restent valides.

Notons que pour les 3 jeux, au plus la moitié des variables initiales ont été sélectionnées.

Gr 3. Données réelles en dimension élevée

- *Cleveland* : 5 variables ont été sélectionnées. Elles donnent des performances comparables à celles obtenues pour les 13 variables d'origine dans le cas de la règle BQ mais sont améliorées de manière très significative dans le cas de la règle des k-PPV.
- *Segment, Ionosphere et Sonar* : sur ces 3 jeux de données, les performances obtenues avec les variables sélectionnées sont améliorées de manière très significative, quel que soit le discriminateur considéré. La réduction du nombre de variables est là encore très importante puisque plus de la moitié des variables d'origine n'ont pas été retenues.
- *Dna* : ce jeu de données réelles est le seul cas d'échec que nous ayons rencontré puisque, dans le cas du discriminateur BQ, les intervalles de confiance sur les taux moyens avec q et p variables ne se chevauchent pas. Cependant, on constate que la réduction du nombre de variables est très significative.

Les figures 3 et 4 résument graphiquement les résultats du tableau 4 pour les deux règles utilisées (BQ et k-PPV respectivement) pour les 11 jeux de données dans l'ordre du tableau 2. Les taux de bon classement obtenus avec les p variables d'origine sont représentés en trait pointillé, ceux obtenus avec les q variables sélectionnées le sont en trait plein. On peut évidement réitérer les remarques précédentes, en particulier que les seules baisses de performances sont observées pour les jeux *Monk-3* et *Dna*.

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'aggrégation d'étiquettes floues/possibilistes représentant le degré d'appartenance aux classes en présence. Le critère proposé peut être associé à n'importe quel discriminateur. Ses performances ont été comparées avec d'autres critères récents 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.

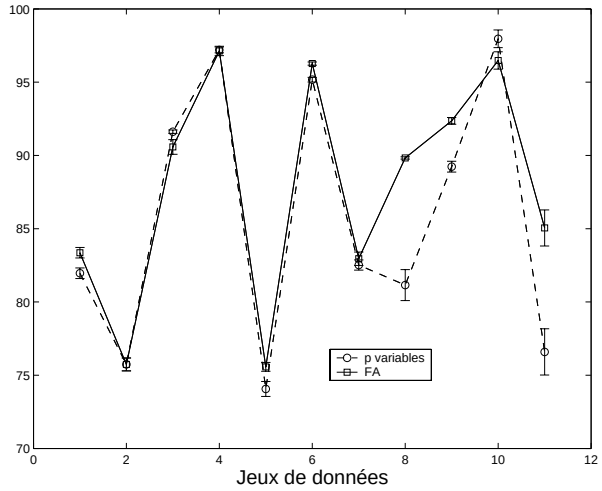

Figure 3: Taux de bon classement (%) pour les q variables sélectionnées et les p variables d'origine – 11 jeux de données, BQ

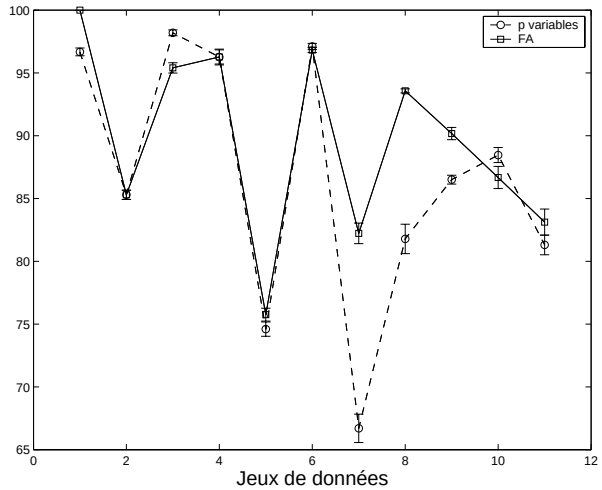

Figure 4: Taux de bon classement (%) pour les q variables sélectionnées et les p variables d'origine – 11 jeux de données, k-PPV

References

[1] S. Basu, C.A. Micchelli, and P. Olsen. Maximum entropy and maximum likelihood criteria for feature selection. *Proc. IEEE Int'l. Symp. Circuits and Systems*, III:267–270, 2000.

[2] J. Bazdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, second edition, 1987.

[3] J. Bi, K. Bennett, Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[4] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. *Wadsworth, Belmont, CA*, 1984.

[6] T.E. Campos, I. Bloch, and R.M. Cesar Jr. Feature selection based on fuzzy distances between clusters: First results on simulated data. *Lecture Notes in Computer Science*, 2013, 2001.

[7] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[8] T.M. Cover and J.M. Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Transactions on Systems Man and Cybernetics*, SMC-7:657–661, 1977.

[9] M. Dash and H Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

[10] M. Dash, H. Liu, and H. Motoda. Consistency based feature selection. pages 98–109, 2000.

[11] M. Detyniecki. *Mathematical Aggregation Operators and their Application to Video Querying*. PhD thesis, Université de Paris 6, December 2000.

[12] P.A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, 1982.

[13] D. Dubois and H. Prade. A review of fuzzy set aggregation connectives. *Information Sciences*, 36:85–121, 1985.

[14] C. Frélicot and B. Dubuisson. A multi-step predictor of membership function as an ambiguity reject solver in pattern recognition. In *: 4th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems(IPMU)*, pages 709–715, Palma de Mallorca, Spain, July 1992.

[15] M.A. Hall. Correlation based feature selection for discrete and numeric class machine learning. *Proc. 17th Int'l. Conf Machine Learning*, 2000.

[16] A. Jain and D. Zongker. Feature selection: Evaluation, application and small sample performance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[17] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994.

[18] J. Kittler. *Feature selection and extraction*. in: Handbook of Pattern Recognition and Image Processing, Y. Fu, ed.,, New York, Academic Press, 1986.

[19] G.J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, NJ, USA, 1995.

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

[21] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[22] R. Krishnapuram and J. M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy systems*, 1(2):98–110, May 1993.

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

[24] P. Langley. Selection of relevant features in machine learning. *in AAAI Fall Symposium on Relevance*, pages 140–144, 1994.

[25] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. *In Proceedings of 7th IEEE Int'l Conference on Tools with Artificial Intelligence*, pages 388–391, 1995.

[26] P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, March 2002.

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

[28] S.K. Pal, R.K. De, and J. Basak. Unsupervised feature evaluation: A neuro-fuzzy approach. *IEEE Trans. Neural Network*, 11:366–376, 2000.

[29] P. Pudil, F. J. Ferri, J. Novovicov'a, and J. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. *12th International Conference on Pattern Recognition*, pages 279–283, 1994.

[30] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

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

[32] P.K. Sankar and P.P. Wang. *Genetic Algorithms and Pattern Recognition*. Boca Raton, FL: CRC Press, 1996.

[33] D. Semani, C. Frélicot, and L. Mascarilla. Discrimination possibiliste avec options de rejet : une nouvelle approche. *In: 12èmes Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, Novembre 2002.

[34] P. Somol, P. Pudil, and J. Grim. Branch & bound algorithm with partial prediction for use with recursive and non-recursive criterion forms. *Proceedings Int. Conf. on Advances in Pattern Recognition, Lecture Notes in Computer Science LNCS 2013, Springer, Rio de Janeiro*, pages 230–239, 2001.

[35] P. Somol, P. Pudil, J. Novovicova, and P. Paclik. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163, 1999.

[36] S. D. Stearns. On selecting features for pattern classifiers. *In Third Int. Conf. on Pattern Recognition, Coronado, CA*, pages 71–75, 1976.

[37] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press Inc. - ISBN 0-12-686140-4, 1999.

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

[39] G. D. Tourassia, E. D. Frederick, and M. K. Markey. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28:2395–2402, December 2001.

[40] B. Yu and B. Yuan. A more efficient branch and bound algorithm for feature selection. *Pattern Recognition*, 26:883–889, 1993.