

Sélection de Variables en Apprentissage Numérique Non Supervisé ^{*}

Sébastien Guérif, Younès Bennani

Université Paris 13 - LIPN / CNRS UMR 7030
99, avenue J-B. Clément, 93430 Villetaneuse

Sebastien.Guerif@lipn.univ-paris13.fr
Younes.Bennani@lipn.univ-paris13.fr

Résumé : Cet article présente une nouvelle approche de sélection de variables pour l'apprentissage numérique non supervisé applicable à des données en grande dimensions. La méthodologie proposée repose sur le calcul d'une pondération obtenue par apprentissage à l'aide d'une extension de l'algorithme de ω -k-moyennes (Huang *et al.*, 2005) au cas des cartes auto-organisées. La pertinence d'une variable est ainsi évaluée progressivement au fur et à mesure de la découverte de la structuration en groupes des données. Les résultats expérimentaux obtenus sur deux bases de données de complexité et de dimensions différentes montrent que notre approche non supervisée sélectionne des sous-ensembles de variables pertinents comparables à ceux retenus par la plupart des méthodes supervisées.

Mots-clés : Sélection de variables, Classification, Cartes auto-organisées, Pondération, Extraction de connaissances.

1 Introduction

D'après une étude récente (KDnuggets, 2006), les techniques de classification automatique ou *clustering* font partie des méthodes d'analyse et de fouille de données les plus utilisées en 2005 ; citées par 40% des personnes interrogées, elles se placent au second rang juste derrière les arbres de décision. Malgré leur succès incontesté en analyse de données exploratoire, les techniques de *clustering* doivent s'adapter à des volumes de données toujours plus importants. En effet, au fur et à mesure de l'évolution des technologies de stockage, le volume de données disponibles a progressivement explosé en nombre d'individus mais aussi en nombre de descripteurs. Nous sommes donc très souvent confrontés au problème de la malédiction de la dimensionalité (*the curse of dimensionality*) car le nombre d'exemples nécessaires à l'apprentissage croît exponentiellement avec le nombre de variables et des techniques de réduction de dimensions

*Ce travail a été en partie financé par le Pôle de Compétitivité Cap Digital (Image, Multimedia and Vie numérique) dans le cadre du projet Infom@gic.

de l'espace de description doivent être utilisées. On distingue généralement deux types d'approche qui peuvent être utilisée conjointement : l'extraction de caractéristiques et la sélection de variables.

L'extraction de caractéristiques consiste à construire de nouveaux attributs à partir de l'ensemble des variables originales, alors que la sélection de variables permet de ne conserver qu'un sous-ensemble pertinent de variable. Dans le cadre de l'apprentissage supervisé, l'extraction de caractéristiques permet généralement d'obtenir des classificateurs plus précis, mais la sélection de variables conduit à des règles de décision plus facile à interpréter. Nous nous intéressons ici au problème de la classification automatique qui permet de construire des groupes d'individus similaires en vue de comprendre la structure d'un ensemble de données. Dans ce contexte, il nous semble pertinent d'adopter une approche de sélection de variables pour améliorer la compréhension du problème sous jacent à un ensemble d'observations par un utilisateur. Nous proposons d'abord d'étendre l'algorithme ω -k-moyennes proposé par (Huang *et al.*, 2005) aux cartes auto-organisées (Kohonen, 2001) qui permettent de visualiser efficacement des données en grande dimension. Ensuite, la pondération obtenue nous permet de sélectionner les variables les plus représentatives du problème. Cette approche permet à l'utilisateur d'accéder visuellement aux connaissances implicites représentées par un ensemble d'observations et de juger de la pertinence d'entamer une analyse plus fine de ses données en se focalisant sur un sous-ensemble pertinent d'attributs.

La suite de cet article est organisée comme suit : nous présentons différentes approches de la sélection de variables dans le contexte de l'apprentissage non supervisé dans la partie 2, avant d'introduire les algorithmes d'apprentissage des k-moyennes et des cartes auto-organisées en partie 3. La partie 4 rappelle les principes de l'algorithme ω -k-moyennes et présente l'extension au cas des cartes auto-organisées que nous proposons. La partie 5 est consacrée à la définition de la méthode de sélection de variables proposée à partir de la pondération obtenue. Nous présentons enfin différents résultats expérimentaux en partie 6 avant de conclure et d'envisager les améliorations possibles de la méthode.

2 Sélection de variables

En apprentissage supervisé, les méthodes d'extraction de caractéristiques conduisent souvent à des classificateurs plus précis que les méthodes de sélection de variables ; néanmoins, en apprentissage non supervisé, il nous semble plus pertinent d'adopter la sélection de variables car elle permet une compréhension plus aisée des résultats car les dimensions sélectionnées sont directement interprétables par l'utilisateur. Une procédure de sélection de variables comporte trois éléments essentiels : une mesure de pertinence, une procédure de recherche et un critère d'arrêt. On en distingue généralement trois types :

- les approches filtres dont la mesure de pertinence est indépendante de l'algorithme qui utilise ensuite les données,
- les approches symbioses qui évaluent la pertinence des sous-ensemble de variables à l'aide des performances du système que l'on construit,

- et les approches intégrées pour lesquelles la mesure de pertinence est directement incluse dans la fonction de coût optimisée par le système.

2.1 Mesure de pertinence

En apprentissage supervisé, la pertinence d'un sous-ensemble de variables est souvent évaluée d'après un critère de performance du modèle que l'on cherche à construire et qui dépend de la tâche à accomplir : classement ou prédiction. Définir une mesure de pertinence pour l'apprentissage non supervisé est un problème difficile car on ne dispose ni de valeur à prédire ni de classe à attribuer. Malgré la difficulté inhérente à l'absence de valeurs cibles, les différentes approches que l'on trouve dans la littérature semblent s'accorder autour d'une définition de la pertinence d'un sous-ensemble de variables que l'on pourrait énoncer ainsi : “*en apprentissage non supervisé, un sous-ensemble de variables pertinent est minimal et permet de mettre en évidence la structure en groupes naturels d'un ensemble de données.*”

Une partie des approches de sélection de variables proposées pour l'apprentissage non supervisé se focalise sur l'élimination des attributs redondants. Ainsi, P. Mitra et al. définissent l'indice de compression maximale de l'information (*maximal information compression index*) comme la plus petite valeur propre de la matrice de corrélation des variables prises deux à deux et proposent une procédure itérative d'élimination des attributs redondants en s'appuyant sur cette mesure de dissimilarité (Mitra *et al.*, 2002). J. Vesanto et al. proposent une détection visuelle des corrélations en se basant sur la construction d'une carte auto-organisées (Vesanto & Ahola, 1999). Ces deux approches s'appuient sur une mesure de corrélation linéaire entre variables et elles ne permettent d'éliminer qu'une partie de la redondance. Ainsi, si on considère un couple de variables aléatoires (X, Y) tel que pour toute réalisation x de la variable X , la réalisation de Y est $y = x^2$, le coefficient de corrélation linéaire entre ces deux variables sera proche de zéro bien que l'ensemble $\{X, Y\}$ soit redondant car la réalisation de la variable Y peut se déduire sans peine de celle de X . De manière naturelle, on peut penser lever cette limitation majeure en remplaçant la mesure de corrélation linéaire par une mesure d'information mutuelle dont nous rappelons la définition ci-dessous :

$$I(X, Y) = - \int_{x,y} p(X = x, Y = y) \log \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} dx \quad (1)$$

$$= H(X) + H(Y) - H(X, Y) \quad (2)$$

$$\text{avec } H(X) = - \int_x p(X = x) \log p(X = x) dx \quad (3)$$

où $H(X)$ est l'entropie au sens de Shannon associée à la variable aléatoire X . Néanmoins, l'évaluation de cette mesure nécessite de connaître d'une part les densités de probabilité des variables aléatoires X et Y , et d'autre part leur densité de probabilité conjointe. Bien entendu, ces informations ne sont en pratique pas disponibles et l'estimation, à la fois rigoureuse et efficace, de l'information mutuelle demeure un problème difficile (Kraskov *et al.*, 2004).

Une autre partie des méthodes tentent d'identifier un sous-ensemble qui permet de mettre en avant les groupes naturels présents dans un ensemble de données. Ainsi, dans (Sorg-Madsen *et al.*, 2003) les auteurs conjecturent que dans un ensemble de variables, celles qui sont indépendantes ou presque autres correspondent vraisemblablement à du bruit et peuvent donc être éliminées. Ils décrivent ainsi une approche "filtre" basée sur une estimation de l'information mutuelle, avant de l'hybrider avec une approche symbiose qui utilise la précision d'un classificateur naïf de Bayes pour la prédiction des classes d'un modèle de mélange préalablement appris. D'autres méthodes basées sur les modèles de mélange ont été proposées, elles diffèrent par la mesure de pertinence utilisée : le maximum de vraisemblance et la séparabilité sont utilisés dans (Dy & Brodley, 2000), alors qu'une mesure de saillance (*saliency*) est définie dans (Law *et al.*, 2004) sous l'hypothèse qu'une variable indépendante des groupes n'est pas pertinente. Cette dernière hypothèse est également utilisée par (Guérif & Bennani, 2006) qui proposent d'utiliser une classification à deux niveaux combinée à la valeur test proposée par (Morineau, 1984) pour identifier les variables les plus significatives ; le premier niveau de la classification est formée par une carte auto-organisée qui est segmentée en utilisant l'algorithme des k-moyennes associé à l'indice de Davies-Bouldin pour fixer le nombre de groupes (Davies & Bouldin, 1979; Vesanto & Alhoniemi, 2000) .

2.2 Procédure de recherche

Trouver une solution optimale suppose une recherche exhaustive parmi les $2^n - 1$ sous-ensembles de variables possibles, et bien que des méthodes de recherche efficaces comme le *Branch and Bound* ait été proposées, elles s'appuient sur une propriété de monotonie de la mesure de pertinence qui est en pratique difficile à assurer (Bennani, 2001a; Yacoub, 2006). Une recherche exhaustive est dès lors inapplicable même pour un nombre de variables de l'ordre de quelques dizaines. En pratique, on utilise des approches sous-optimales comme les algorithmes de recherche gloutonne ou les méthodes de recherche aléatoire.

Les stratégies gloutonnes les plus utilisées sont les méthodes séquentielles dont font partie la méthode de sélection avant (*forward selection*) (Dy & Brodley, 2000; Raftery & Dean, 2006), la méthode d'élimination arrière (*backward elimination*) (Guérif & Bennani, 2006) et les méthodes bidirectionnelles comme la méthode *stepwise* ou celle proposée par (Sorg-Madsen *et al.*, 2003) qui combine une approche filtre par sélection à une approche symbiose par élimination. La méthode de sélection avant débute avec un ensemble vide et progresse en ajoutant une à une les variables les plus intéressantes. A l'inverse, la méthode d'élimination arrière commence par l'ensemble de toutes les variables dont les moins pertinentes sont supprimées tour à tour. Les méthodes bidirectionnelles combinent ces deux modes de recherche. Les algorithmes génétiques font partie des méthodes de recherche aléatoire qui sont parfois utilisées.

On reproche généralement à la méthode de sélection avant de ne pas prendre en compte le problème de la pertinence mutuelle : deux variables peuvent être intéressantes lorsqu'elles sont présentes conjointement dans le sous-ensemble sélectionné bien que chacune prise séparément n'apporte aucune information pertinente (un exemple classique de cette situation est le cas du *XOR*.) La méthode d'élimination arrière permet de

traiter correctement ce problème mais n'est malheureusement pas efficace du point de vue calculatoire. Enfin, les algorithmes génétiques sont difficiles à mettre en œuvre et souvent coûteux en temps de calcul.

2.3 Critère d'arrêt

Le nombre de variables à sélectionner fait généralement partie des inconnues du problème et il doit être déterminé automatiquement à l'aide d'un critère d'arrêt de la procédure de sélection de variables. Ce problème se ramène à celui de la sélection de modèles et de nombreux auteurs utilisent soit des critères de maximum de vraisemblance (Dy & Brodley, 2000; Raftery & Dean, 2006) soit des critères de séparabilité des classes (Dy & Brodley, 2000; Guérif & Bennani, 2006). Il convient de noter que les critères de séparabilité utilisés dans (Dy & Brodley, 2000) et (Guérif & Bennani, 2006) ne sont plus utilisables lorsque le nombre de variables est important car leur évaluation fait intervenir soit l'inversion soit le calcul du déterminant de matrices de covariance. D'autres auteurs utilisent une combinaison de la mesure de pertinence et de la procédure de recherche. Ainsi, dans (Sorg-Madsen *et al.*, 2003) les auteurs utilisent pour leur approche un critère d'information mutuelle et stoppent la procédure de sélection avant lorsqu'aucune variable ne peut plus être ajoutée. La procédure d'élimination arrière de leur approche symbiose est stoppée à l'aide d'un seuil sur les performances du classificateur bayésien naïf qu'ils utilisent.

3 Algorithme des k-moyennes et Cartes auto-organisées

3.1 Notations

Soit $X = \{x_i \in \mathbb{R}^n : i = 1, \dots, N\}$ un ensemble d'observations, N et n correspondent respectivement au nombre d'exemples et à la dimension de l'espace de description. Une partition $P = \{P_j : j = 1, \dots, K\}$ de X peut être représentée sous forme d'une matrice de partition $U = (u_{ij})$ dont les indices $i = 1, \dots, N$ et $j = 1, \dots, K$ font respectivement référence à une observation $x_i \in X$ et à un groupe $P_j \in P$; ainsi, $u_{ij} = 1$ indique que l'observation $x_i \in X$ appartient au groupe $P_j \in P$ et pour tout $i = 1, \dots, N$ nous avons $\sum_j u_{ij} = 1$.

3.2 Algorithme des k-moyennes

Une des méthodes les plus utilisées en classification est l'algorithme des k-moyennes dans lequel chaque groupe est représenté par son référent $z_j \in \mathbb{R}^n$ qui appartient à l'espace de description. Cet algorithme optimise la fonction de coût suivante :

$$R_{KM}(U, Z) = \sum_i \sum_j u_{ij} \times d^2(x_i, z_j) \quad (4)$$

où U est une matrice de partition, Z est la matrice dont les lignes correspondent aux référents des différents groupes et où $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ est la mesure de dissimilarité utilisée pour regrouper les observations; la mesure la plus utilisée est la distance

euclidienne dont la définition est rappelée ci-dessous :

$$d^2(x_i, z_j) = \sum_k (x_{ik} - z_{jk})^2 \quad (5)$$

où x_i et z_j sont respectivement une observation de X et le référent du groupe $P_j \in P$. La fonction de coût (4) peut être optimisée en sélectionnant aléatoirement les référents initiaux et en itérant les deux étapes suivantes jusqu'à convergence :

1. Optimisation des affectations : les référents des groupes \hat{Z} sont fixés et la fonction de coût $R_{KM}(U, \hat{Z})$ est optimisée en affectant à chaque observation le groupe du référent le plus proche,

$$u_{ij} = \begin{cases} 1, & \text{if } j = \underset{P_l \in P}{\operatorname{argmin}} d^2(x_i, \hat{z}_l), \\ 0, & \text{sinon.} \end{cases} \quad (6)$$

2. Optimisation des référents : la matrice de partition \hat{U} est fixée et la fonction de coût $R_{KM}(\hat{U}, Z)$ est optimisée en remplçant chaque référent par le centroïde du groupe qu'il représente.

$$z_j = \frac{\sum_i \hat{u}_{ij} x_i}{\sum_i \hat{u}_{ij}} \quad (7)$$

Cet algorithme est très instable et manque de robustesse face à la présence de valeurs abérantes ; on lui reproche très souvent de converger vers des optima locaux qui sont fortement conditionnés par l'initialisation des référents.

3.3 Cartes auto-organisées

Les cartes auto-organisées - *Self-Organizing Maps (SOM)* - ont été introduites au début des années 80 comme une méthode de visualisation de données multidimensionnelles par le professeur Teuvo Kohonen (Kohonen, 2001). Cette approche peut être vue comme une extension de l'algorithme des k-moyennes dans laquelle les référents sont soumis à une contrainte de voisinage qui permet de préserver l'ordre topologique de l'espace de description. Une carte auto-organisée peut-être vue comme un ensemble de référents en bijection avec l'ensemble des sommets d'un graphe non orienté qui définit la topologie de l'espace de projection. La longueur du plus court chemin entre deux sommets du graphe définit une distance δ dans l'espace de projection. Les termes référent, unité ou neurone seront utilisés sans distinction dans la suite du documents. La préservation de l'ordre topologique signifie que des individus voisins dans l'espace d'entrée seront affectés au même référent ou à des référents voisins sur la carte. Cette propriété peut être obtenue en introduisant une contrainte de voisinage dans la fonction de coût optimisée par l'algorithme des k-moyennes (4) de la manière suivante :

$$R_{SOM}(U, Z) = \sum_i \sum_j u_{ij} \times \left[\sum_l h_{jl} \times d^2(x_i, z_l) \right] \quad (8)$$

où h_{jl} est la valeur de la fonction de voisinage entre les unités j et l . On utilise le plus souvent un noyau gaussien de paramètre λ comme fonction de voisinage :

$$h_{jl} = \exp\left(-\frac{\delta^2(j,l)}{2\lambda^2}\right) \quad (9)$$

La fonction de coût des cartes auto-organisée ou SOM (8) peut être optimisée soit par une approche de descente du gradient, soit par un procédure similaire à celle décrite ci-dessus pour la fonction de coût des k-moyennes. La deuxième étape doit être modifiée pour prendre en compte la fonction de voisinage et les référents sont remplacés par la moyenne pondérée des individus qui sont affectés à une unité de son voisinage. Il est important de rappeler ici que le paramètre λ du noyau gaussien doit décroître progressivement et qu'une valeur trop faible en début d'apprentissage risque d'empêcher le processus d'auto-organisation.

4 Apprendre une pondération pendant la classification

4.1 L'algorithme ω -k-moyennes

Dans (Huang *et al.*, 2005), les auteurs proposent d'introduire des coefficients de pondération dans la fonction de coût des k-moyennes en remplaçant la distance euclidienne classique par une distance euclidienne pondérée définie par :

$$d_\omega^2(x_i, z_j) = \sum_k \omega_k^\beta \times (x_{ik} - z_{jk})^2 \quad (10)$$

avec $\omega_k \geq 0$ et $\sum_k \omega_k = 1$. La fonction de coût est modifiée ainsi :

$$R_{\omega KM}(U, Z, W) = \sum_i \sum_j u_{ij} \times d_\omega^2(x_i, z_j) \quad (11)$$

où W est un vecteur colonne formé par les coefficients de pondération utilisés pour le calcul de d_ω . La procédure d'optimisation des k-moyennes présentée dans la partie précédente peut être utilisée pour la fonction de coût définie ci-dessus (11) ; les coefficients de pondération \hat{W} sont fixés pendant les deux premières étapes et d'après le théorème donné dans (Huang *et al.*, 2005), la fonction de coût (11) est optimisée par rapport à W en ajoutant la troisième étape suivante :

3. Optimisation de la pondération : la matrice de partition \hat{U} et les référents des groupes \hat{Z} sont fixés et il est montré dans (Huang *et al.*, 2005) que la fonction de coût $L(\hat{U}, \hat{Z}, W)$ atteint son minimum pour les valeurs suivantes des coefficients de pondération :

$$\omega_k = \begin{cases} 0, & \text{if } D_k = 0, \\ \left(\sum_t \left[\frac{D_k}{D_t}\right]^{\frac{1}{\beta-1}}\right)^{-1}, & \text{sinon.} \end{cases} \quad (12)$$

$$\text{avec } D_k = \sum_i \sum_j \hat{u}_{ij} \times (x_{ik}, \hat{z}_{jk})^2 \quad (13)$$

4.2 L'algorithme ω -SOM

L'algorithme ω -k-moyennes présenté ci-dessus peut être étendu aux cartes auto-organisatrices en introduisant une contrainte de voisinage entre les référents pour préserver l'ordre topologique de l'espace d'entrée. Ainsi, les fonctions de coût de l'algorithme ω -k-moyennes et des cartes auto-organisées peuvent être combinées comme suit :

$$R_{\omega SOM}(U, Z, W) = \sum_i \sum_j u_{ij} \times \left[\sum_l h_{jl} d_{\omega}^2(x_i, z_l) \right] \quad (14)$$

L'optimisation de la fonction de coût $R_{\omega SOM}$ (14) est réalisée par le même algorithme que pour $R_{\omega KM}$ (11). Bien que la définition (13) des termes D_k utilisés dans (12) soit modifiée comme suit :

$$D_k^{(SOM)} = \sum_i \sum_j \hat{u}_{ij} \times \left[\sum_l h_{jl} (x_{ik} - z_{lk})^2 \right] \quad (15)$$

la preuve du théorème proposé dans (Huang *et al.*, 2005) reste valide. Dans l'algorithme de haut niveau (algorithme 1) qui résume la procédure d'optimisation ω -SOM, t et T_{max} correspondent respectivement à l'itération courante et nombre de période d'apprentissage.

Algorithm 1 Algorithme ω -SOM

Initialiser W et Z aléatoirement.

for $t = 1, \dots, T_{max}$ **do**

Optimiser $R_{\omega SOM}(U, \hat{W}, \hat{W})$: chaque individu est affecté à son référent le plus proche au sens de la distance d_{ω} ,

Optimiser $R_{\omega SOM}(\hat{U}, Z, \hat{W})$: les référents sont remplacés par la moyenne pondérée des individus affecté à une unité de leur voisinage sur la carte,

Optimiser $R_{\omega SOM}(\hat{U}, \hat{Z}, W)$: les coefficients de pondération sont mis à jour d'après le théorème proposé dans (Huang *et al.*, 2005) étendu aux cas des cartes auto-organisées.

end for

5 De la pondération à la sélection de variables

5.1 Mesure de pertinence

En présence de données fortement bruitées, l'évaluation de la pertinence relative des variables est sujette à des erreurs d'estimation qui peuvent conduire à un échec du processus de sélection. Ainsi, la suppression d'une variable pertinente dans le cas d'une recherche par élimination arrière est irrémédiable (Guérif & Bennani, 2006), et l'ajout d'une variable par une méthode de sélection avant n'est jamais remis en cause. Notons qu'une erreur d'estimation peut également conduire à un arrêt prématuré du processus.

Les coefficients de pondération calculés par l'algorithme ω -SOM indique l'importance relative des différentes variables : plus le poids ω_k de la k -ième variable est important, plus celle-ci contribue à la classification obtenue. Leur optimisation progressive en présence de l'ensemble des variables permet, d'une part, une remise des erreurs d'estimation antérieures, et d'autre part, une prise en considération de la pertinence mutuelle des différentes dimensions. On dispose ainsi d'une mesure de pertinence robuste et stable qui permet d'ordonner les dimensions.

5.2 Procédure de recherche et critère d'arrêt

Une des limitations majeures de la méthodes d'élimination arrière utilisée dans (Guérif, 2006) est son manque d'efficacité du point de vue calculatoire ; chaque suppression de variable est suivie d'une nouvelle exécution de l'algorithme ω -SOM. La procédure pourrait être améliorée en utilisant des techniques de réoptimisation adaptées mais nous avons remarqué un phénomène à son point d'arrêt qu'il nous semble intéressant d'exploiter. En effet, ce dernier correspond systématiquement à une augmentation significative du poids de la variable candidate à l'élimination ; il est ainsi possible de le détecter efficacement.

Considérons les variables selon l'ordre croissant de leur pertinence et notons ρ_i le rapport des coefficients de pondération des variables i et $i + 1$:

$$\rho_i = \frac{\omega_{i+1}}{\omega_i} \quad (16)$$

Lorsque les contributions de deux variables consécutives sont proches, la valeur de ce rapport est proche de 1 et une valeur supérieure caractérise une augmentation sensible du poids des variables. Le théorème de Bienaymé-Tchebitchev permet de fixer le seuil indiquant une valeur significativement supérieure à 1 des ρ_i sans faire d'hypothèse sur leur distribution. Il dit que pour une variable aléatoire continue X de distribution quelconque, d'espérance \bar{x} et d'écart-type σ , on a :

$$p(\bar{x} - k \sigma \leq X \leq \bar{x} + k \sigma) \geq 1 - \frac{1}{k^2} \quad (17)$$

où k un réel positif. Nous l'utilisons pour détecter la première augmentation significative du poids de la variable candidate à l'élimination ; toutes les dimensions dont le poids est supérieur ou égal sont alors conservées.

6 Résultats expérimentaux

6.1 Jeux de données

Nous avons utilisé différents jeux de données de taille et de complexité variables pour évaluer notre approche et nous présentons ici les résultats obtenus sur deux d'entre eux : le premier a été mis à disposition de la communauté par l'Université d'Irvine (D.J. Newman & Merz, 1998) et le second a été proposé pendant la compétition sur la sélection de variables organisée à l'occasion de la conférence NIPS'2003 (Guyon *et al.*, 2006) :

- La base *waveform* est composée de 5000 individus issus de 3 classes. Chaque classe a été générée à partir d'une combinaison de 2 sur 3 vagues de *base* et un bruit gaussien a été ajouté à chaque dimension. La base originale comportait 21 dimensions mais 19 dimensions additionnelles distribuées selon une loi normale ont été ajoutées.
- La base *madelon* est un problème 2 classes proposé à l'origine pendant la compétition sur la sélection de variables organisée lors de la conférence NIPS'2003 (Guyon *et al.*, 2006). Les exemples sont situés sur les sommets d'un hypercube en dimension 5, mais 15 attributs redondants et 480 dimensions bruitées ont été ajoutés. Les dimensions bruitées suivent la même loi de probabilité que les variables pertinentes mais elles sont indépendantes des étiquettes qui ont été affectées aléatoirement aux sommets. Le jeu de données original était séparé en trois parties (apprentissage, validation et test) mais nous n'avons utilisé que les 2600 exemples de l'ensemble d'apprentissage et de validation pour lesquels les classes étaient connues.

6.2 Méthode d'évaluation

Nous avons utilisé une validation croisée pour évaluer les performances de notre approche : 10 % des exemples ont été utilisés pour l'apprentissage et les 90 % restants pour l'évaluation. La classe des exemples étant disponible, nous avons utilisé les performances d'un classificateur basé sur les k plus proches voisins (k -ppv) pour évaluer la pertinence du sous-ensemble de variables sélectionné. Le caractère significatif de l'amélioration de la précision a été évalué en comparant les résultats aux performances moyennes obtenues par permutations des coefficients de pondération. Chaque expérimentation a été répétée pour les valeurs allant de 2 à 10 du paramètre β .

6.2.1 Pertinence du sous-ensemble sélectionné

Les figures 2, 3 et 4 montrent respectivement le taux d'erreurs des classificateurs (k -ppv) sur la base *waveform* pour les valeurs 2, 5, et 10 du paramètre β . Les pires résultats sont obtenus avec une sélection aléatoire des attributs et les meilleurs en appliquant la pondération sur le sous-ensemble de dimensions sélectionnées. La courbe de référence correspond à l'utilisation de la distance euclidienne non pondérée calculée sur l'ensemble des dimensions. On peut remarquer que la courbe de référence et celles qui correspondent à l'utilisation d'une distance pondérée sur l'ensemble des dimensions (avec ou sans permutation des coefficients de pondération) se ressemblent lorsque la valeur du paramètre β augmente ; ce phénomène est dû à l'effet lissant de β qui est illustré par la figure 1. Il convient de souligner que le sous-ensemble de variables sélectionné par notre approche conduit à une amélioration significative de la précision de près de 10 points indépendamment de la valeur de β .

Le comportement des courbes correspondant à des distances pondérées l'ensemble des dimensions (avec ou sans permutation des poids) se confirme sur le deuxième jeu de données comme le montre les figures 5, 6 et 7. Bien que la précision des classificateurs obtenus restent médiocres, on notera une nette amélioration (plus de 10 points) en utilisant seulement les variables du sous-ensemble sélectionné. Plusieurs raisons expliquent ce taux d'erreurs élevé. D'une part la complexité importante de ce jeu de données et

Sélection de variables en apprentissage numérique non supervisé

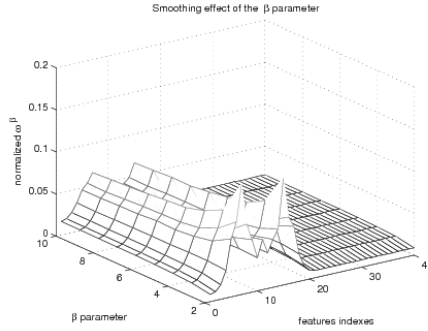


FIG. 1 – Effet de lissage du paramètre β (poids médians obtenus pour la base *waveform*) : les coefficients de pondération ω_j^β ont été normalisés pour avoir $\sum \omega_j^\beta = 1$. On peut observer qu'au fur et à mesure que la valeur du paramètre β augmente, la courbe des poids se lisse.

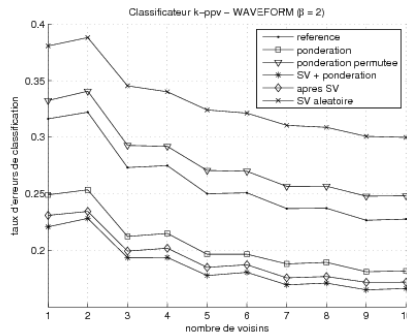


FIG. 2 – Performances d'un classificateur k-ppv en utilisant 10 % de la base *waveform* pour l'apprentissage avec $\beta = 2$.

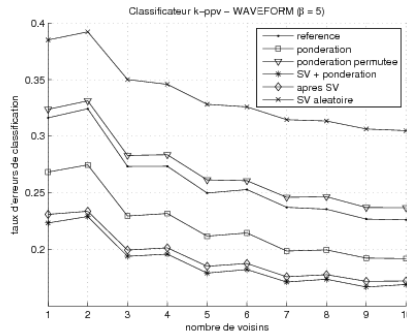


FIG. 3 – Performances d'un classificateur k-ppv en utilisant 10 % de la base *waveform* pour l'apprentissage avec $\beta = 5$.

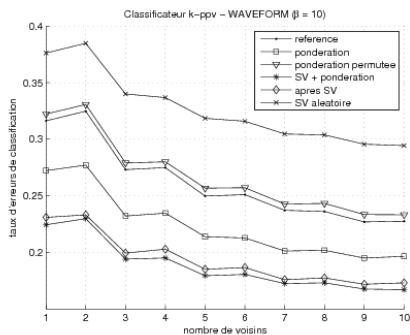


FIG. 4 – Performances d’un classificateur k-ppv en utilisant 10 % de la base *waveform* pour l’apprentissage avec $\beta = 10$.

la faible proportion utilisée pour l’apprentissage ; en effet, seuls 260 exemples ont été utilisés pour l’apprentissage, ce qui représente environ la moitié du nombre de dimensions. D’autre part, notre approche se focalise sur l’élimination du bruit et ne traite pas le problème de la redondance et il est indiqué dans (Guyon *et al.*, 2006) qu’un taux d’erreurs d’environ 10 % par un classificateur k-ppv peut être atteint en utilisant seulement les 5 variables pertinentes et non redondantes.

6.2.2 Stabilité du sous-ensemble sélectionné

Les variables 3 à 19 de la base *waveform* ont été sélectionnées à chaque exécution et pour toutes les valeurs de β ; les variables 2 et 20 ont été sélectionnées pour un des 10 sous-ensembles de données. Le sous-ensemble sélectionné est également relativement stable sur la base *madelon* ; ainsi, les variables 29, 65, 106, 129, 154, 242, 282, 319, 337, 339, 434, 443, 452, 454, 473, 476 et 494 de la base *madelon* ont été sélectionnées pour plus de la moitié des exécutions pour chaque des valeurs de β . Néanmoins, une légère diminution de la stabilité est observée lorsque la valeur de β augmente ; nous l’attribuons à son effet lissant.

6.2.3 Découverte des “vraies” classes

Une fois la sélection de variables réalisée, on cherche généralement à avoir un aperçu de la structuration de nos données. L’approche de classification à deux niveaux proposée par (Vesanto & Alhoniemi, 2000) peut être utilisée à cette fin : une carte auto-organisée est entraînée en utilisant seulement les attributs sélectionnés avant d’être segmentée par la méthodes des k-moyennes. Le nombre de groupes est alors déterminés à l’aide de l’indice de Davies-Bouldin (Davies & Bouldin, 1979) qui est défini ainsi :

$$I_{DB} = \frac{1}{K} \sum_{j=1}^K \max_{j \neq l} \left\{ \frac{\sum_i u_{ij} \|x_i - z_j\|^2 + u_{il} \|x_i - z_l\|^2}{\|z_k - z_l\|^2} \right\} \quad (18)$$

Sélection de variables en apprentissage numérique non supervisé

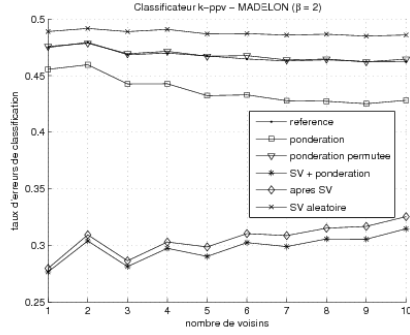


FIG. 5 – Performances d’un classificateur k-ppv en utilisant 10 % de la base *madelon* pour l’apprentissage avec $\beta = 2$.

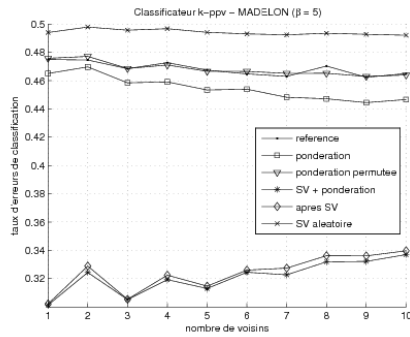


FIG. 6 – Performances d’un classificateur k-ppv en utilisant 10 % de la base *madelon* pour l’apprentissage avec $\beta = 5$.

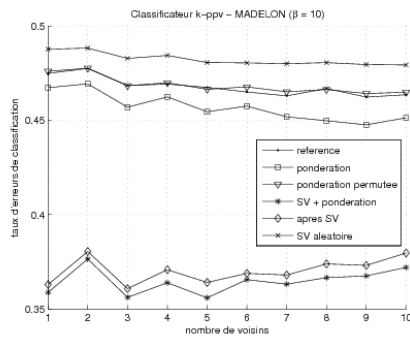


FIG. 7 – Performances d’un classificateur k-ppv en utilisant 10 % de la base *madelon* pour l’apprentissage avec $\beta = 10$.

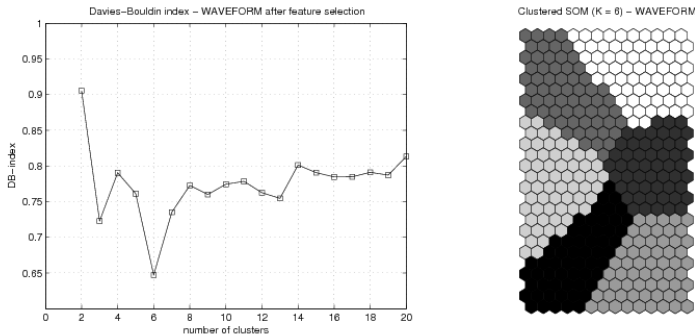


FIG. 8 – A gauche : Indice de Davies-Bouldin pour différentes segmentation de la carte auto-organisée entraînée sur la base *waveform* réduite. Le minimum 0.65 est atteint pour $K = 6$. A droite : Segmentation en 6 groupes de la carte

TAB. 1 – Répartition des classes originales dans les 6 groupes identifiés : 1, 2 et 3 correspondent aux classes réelles de la base *waveform* ; en affectant la classe majoritaire à chaque groupe, le taux d'erreur de classification s'élève à 0,30.

	1	2	3	Pureté
Groupe 1	5	0	545	99,09 %
Groupe 2	602	2	3	99,18 %
Groupe 3	6	622	1	98,89 %
Groupe 4	577	615	0	51,59 %
Groupe 5	0	414	444	51,75 %
Groupe 6	502	0	662	56,87 %

où K est le nombre de groupes. Cette procédure de classification tend à identifier des groupes hyper-sphériques compacts bien séparés.

La partie gauche de la figure 8 montre la valeur de l'indice de Davies-Bouldin pour différentes segmentation de la carte construite à partir des variables sélectionnées pour la base *waveform*. D'après ce critère, la meilleure segmentation comporte 6 groupes et elle est présentée sur la partie droite de la figure 8. La matrice de confusion (table 1) met en évidence que les 6 groupes identifiés correspondent soit à une classes réelles, soit à une des zones de recouvrement important comme le confirme la positions relative des groupes sur la carte. En procédant de manière analogue pour la base *madelon*, nous obtenons la matrice de confusion présentée à la tables 2.

7 Conclusion

Nous avons proposé une approche de sélection de variables performante pour l'apprentissage non supervisé. Elle s'appuie sur le calcul d'une pondération pendant une classification basée sur les carte auto-organisées qui sont plus robustes et moins sen-

TAB. 2 – Répartition des classes originales dans les 14 groupes identifiés : -1 and +1 correspondent aux classes réelles de la base *madelon* ; en affectant la classe majoritaire à chaque groupe, le taux d’erreur de classification s’élève à 0,31.

	-1	+1	Pureté
Groupe 1	160	47	77,29 %
Groupe 2	112	113	50,22 %
Groupe 3	105	65	61,76 %
Groupe 4	72	48	60,00 %
Groupe 5	51	155	75,24 %
Groupe 6	49	107	68,59 %
Groupe 7	96	55	63,58 %
Groupe 8	87	169	66,02 %
Groupe 9	117	21	84,78 %
Groupe 10	66	138	67,65 %
Groupe 11	10	139	93,29 %
Groupe 12	163	22	88,11 %
Groupe 13	64	104	61,90 %
Groupe 14	148	117	55,85 %

sibles à l’initialisation que l’algorithme de k-moyennes. L’apprentissage d’une pondération permet en outre une évaluation progressive de la pertinence des variables au fur et à mesure que la structuration du jeu de données est découverte. On dispose ainsi d’un critère d’évaluation non supervisée robuste et fiable des différentes dimensions. On est ainsi en mesure d’ordonner les variables par ordre de pertinence et nous proposons une méthode pour déterminer automatiquement une valeur de coupure. Notre méthode est en mesure de traiter des base de données de dimension importante et est applicable lorsque le nombre de variables dépasse celui des individus. Nous envisageons de poursuivre ce travail en le combinant avec une approche de sélection avant qui permette d’éliminer la redondance en se basant sur une estimation de l’information mutuelle.

Références

- BENNANI Y. (2001a). *La sélection de variables*, p. 351–371. In Bennani (2001b).
- Y. BENNANI, Ed. (2001b). *Systèmes d’apprentissage connexionnistes : sélection de variables, Numéro spécial de la Revue d’Intelligence Artificielle*. Hermès, Paris.
- Y. BENNANI, Ed. (2006). *Apprentissage Connexionniste*. Hermès Sciences, Paris.
- DAVIES D. L. & BOULDIN D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, **1**(2), 224–227.
- D.J. NEWMAN, S. HETTICH C. B. & MERZ C. (1998). UCI repository of machine learning databases.
- DY J. G. & BRODLEY C. E. (2000). Feature Subset Selection and Order Identification for Unsupervised Learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML’2000)*, Stanford University, CA.

- GUÉRIF S. (2006). *Réduction de dimensions en Apprentissage numérique non supervisée*. PhD thesis, Université Paris 13.
- GUÉRIF S. & BENNANI Y. (2006). Selection of clusters number and features subset during a two-levels clustering task. In *Proceedings of the 10th IASTED International Conference Artificial intelligence and Soft Computing 2006*, p. 28–33.
- GUYON I., GUNN S., NIKRAVESH M. & ZADEH L. (2006). *Feature Extraction, Foundations and Applications*, Editors. Series Studies in Fuzziness and Soft Computing, Physica-Verlag. Springer.
- HUANG J. Z., NG M. K., RONG H. & LI. Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(5), 657–668.
- KDNUGETTS (2006). 2006 kdnuggets poll on data mining/analytic techniques, http://www.kdnuggets.com/polls/2006/data_mining_methods.htm.
- KOHONEN T. (2001). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Berlin, Heidelberg, New York : Springer, third extended edition edition.
- KRASKOV A., STÖGBAUER H. & GRASSBERGER P. (2004). Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys*, **69**(6 Pt 2).
- LAW M. H. C., FIGUEIREDO M. A. T. & JAIN A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(9), 1154–1166.
- MITRA P., MURTHY C. & PAL S. (2002). Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(4).
- MORINEAU A. (1984). *Note sur la caractérisation statistique d'une classe et les valeurs-tests*. Bulletin technique 2, Centre international de statistique et d'informatique appliquées, Saint-Mandé, France.
- RAFTERY A. & DEAN N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.
- SORG-MADSEN N., THOMSEN C. & PEÑA J. (2003). Unsupervised feature subset selection. In *Proceedings of the Workshop on Probabilistic Graphical Models for Classification, ECML/PKDD*, p. 71–82.
- VESANTO J. & AHOLA J. (1999). Hunting for Correlations in Data Using the Self-Organizing Map. In H. BOTHE, E. OJA, E. MASSAD & C. HAEFKE, Eds., *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, p. 279–285 : ICSC Academic Press.
- VESANTO J. & ALHONIEMI E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, **11**(3), 586–600.
- YACCOUB M. (2006). *Techniques d'élagage et sélection de variables*, chapter 9, p. 249–277. In Bennani (2006).