Réduction des dimensions des données en apprentissage artificiel

Y. Bennani, S. Guérif, E. Viennet

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

Résumé. Depuis plusieurs décénies, le volume des données disponibles ne cesse de croître ; alors qu'au début des années 80 le volume des bases de données se mesurait en mega-octets, il s'exprime aujourd'hui en tera-octets et parfois même en peta-octets. Le nombre de variables et le nombre d'exemples peuvent prendre des valeurs très élevés, et cela peut poser un problème lors de l'exploration et l'analyse des données. Ainsi, le développement d'outils de traitement adaptés aux données volumineuses est un enjeu majeur de la fouille de données. La réduction des dimensions permet notament de faciliter la visualisation et la compréhension des données, de réduire l'espace de stockage nécessaire et le temps d'exploitation, et enfin d'identifier les facteurs pertinents. Dans cet article, nous présentons un panarama des techniques de réduction des dimensions essentiellement basées sur la sélection de variables supervisée et non supervisée, et sur les méthodes géométriques de réduction de dimensions.

1 Introduction

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'exemples. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure compréhension de la valeur des connaissances disponibles dans ces données. La réduction des dimensions est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ce problème. Son objectif est de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critère fixé auparavant. La sélection de ce sous-ensemble de caractéristiques permet d'éliminer les informations non-pertinentes et redondantes selon le critère utilisé. Cette sélection/extraction permet donc de réduire la dimension de l'espace des exemples et rendre l'ensemble des données plus représentatif du problème. En effet, les principaux objectifs de la réduction de la dimension sont :

- faciliter la visualisation et la compréhension des données,
- réduire l'espace de stockage nécessaire,
- réduire le temps d'apprentissage et d'utilisation,
- identifier les facteurs pertinents.

Les algorithmes d'apprentissage artificiel requièrent typiquement peu de traits (*features*) ou de variables (attributs) très significatives caractérisant le processus étudié. Dans le domaine

de la reconnaissance des formes et de la fouille de données, il pourrait encore être bénéfique d'incorporer un module de réduction de la dimension dans le système global avec comme objectif d'enlever toute information inconséquente et redondante. Cela a un effet important sur la performance du système. En effet le nombre de caractéristiques utilisées est directement lié à l'erreur finale. L'importance de chaque caractéristique dépend de la taille de la base d'apprentissage (pour un échantillon de petite taille, l'élimination d'une caractéristique importante peut diminuer l'erreur). Il faut aussi noter que des caractéristiques individuellement peu pertinentes peuvent être très informatives si on les utilise conjointement.

La réduction de la dimension est un problème complexe qui permet de réduire le volume d'informations à traiter et faciliter le processus de l'apprentissage. Nous pouvons classer toutes les techniques mathématiques de réduction des dimensions en deux grandes catégories :

- la sélection de variables : qui consiste à choisir des caractéristiques dans l'espace de mesure (figure 1),

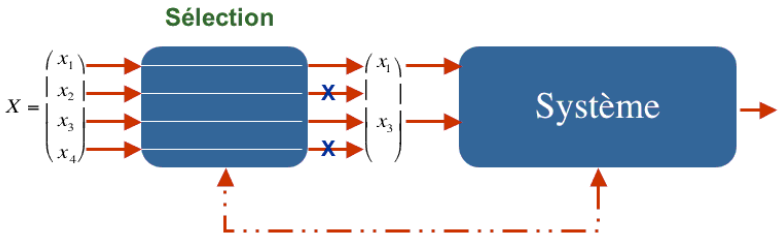

FIG. 1 – *Principe de la sélection de variables.*

- et l'extraction de traits : qui vise à sélectionner des caractéristiques dans un espace transformé (dans un espace de projection) (figure 2)

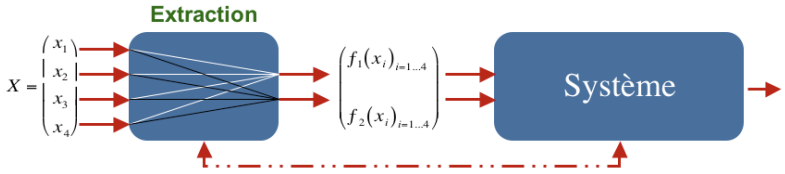

FIG. 2 – *Principe de l'extraction de caractéristiques.*

Définition 1.1 (Bennani (2001)) *Nous appelons "variables" les descripteurs d'entrée et "traits" des caractéristiques construites à partir des variables d'entrée.*

La distinction est nécessaire dans le cas des méthodes à noyaux pour lesquelles les traits ne sont pas explicitement calculés.

La première catégorie est appropriée quand l'acquisition de mesures des formes est coûteuse. Ainsi l'objectif principal de la sélection de caractéristiques dans ce cas est de réduire le nombre de mesures requises. Par contre, les techniques d'extraction de traits (deuxième catégorie) utilisent toute l'information contenue dans les formes pour la compresser et produire un

vecteur de plus petite dimension. Ces techniques projettent un vecteur forme de l'espace de représentation dans un espace de dimension plus petite. Les systèmes d'apprentissage connexionniste sont un bon exemple de cette catégorie. En effet, les modèles connexionnistes conçus pour une tache de discrimination fournissent un système avec des aptitudes intéressantes pour l'analyse du processus. Les cellules cachées d'un Perceptron multi-couches apprennent comment extraire les caractéristiques significatives du signal d'entrée.

2 Réduction des dimensions par sélection de variables

La sélection de variable est un problème difficile qui a été étudié depuis les années 70. Il revient dans l'actualité scientifique avec l'apparition des grandes bases de données et les systèmes de fouille de données «Data Mining» (Liu et Motoda, 1998; Cakmakov et Bennani, 2002; Guyon et al., 2006).

La sélection de variables a fait l'objet de plusieurs recherches en statistique, et plus particulièrement dans des domaines comme la reconnaissance des formes, la modélisation de séries chronologiques et l'identification de processus. Dans le domaine de l'apprentissage, l'étude de la problématique de la sélection de variables est assez récente. En apprentissage symbolique, de nombreuses méthodes ont été proposées pour des tâches de classement (discrimination). Dans le domaine de l'apprentissage connexionniste (Bennani, 2001, 2006), la sélection de variables a été abordée à partir d'un problème d'optimisation et de choix d'architectures des modèles, ainsi des approches très intéressantes ont émergé.

La sélection de variables est une problématique complexe et d'une importance cruciale pour les systèmes d'apprentissage. Afin de mettre en évidence les deux aspects du processus de la sélection de variables : difficulté et importance, nous allons présenter les éléments essentiels que nécessite généralement ce processus. Une définition de la sélection de variables peut s'énoncer de la façon suivante :

Définition 2.1 (Bennani (2001)) *La sélection de variables est un procédé permettant de choisir un sous-ensemble optimal de variables pertinentes, à partir d'un ensemble de variables original, selon un certain critère de performance.*

A partir de cette définition, on peut se poser trois questions essentielles :

- Comment mesurer la pertinence des variables ?
- Comment former le sous-ensemble optimal ?
- Quel critère d'optimalité utiliser ?

Ces trois questions définissent les éléments essentiels d'une procédure de sélection de variables. En effet, le problème de la sélection de variables consiste à identifier les variables permettant une meilleure séparation entre les différentes classes dans le cas d'un classement et une meilleure qualité de prédiction dans le cas d'une régression. On parle alors de "pouvoir discriminant" dans le premier cas et de "pouvoir prédictif" dans le deuxième cas, pour désigner la pertinence d'une variable. La réponse à la première question consiste à trouver une mesure de pertinence ou un *critère d'évaluation* $J(X)$ permettant de quantifier l'importance d'une variable ou d'un ensemble de variables X. La deuxième question évoque le problème du choix de la *procédure de recherche* ou de constitution du sous-ensemble optimal des variables pertinentes. La dernière question demande la définition d'un critère d'arrêt de la recherche.

Le *critère d'arrêt* est généralement déterminé à travers une combinaison particulière entre la mesure de pertinence et la procédure de recherche.

2.1 Critères d'évaluation

L'amélioration des performances d'un système d'apprentissage par une procédure de sélection de variables nécessite dans un premier temps la définition d'une mesure de pertinence. Dans le cas d'un problème de classement, on teste, par exemple, la qualité de discrimination du système en présence ou en absence d'une variable. Par contre, pour un problème de régression, on teste plutôt la qualité de prédiction par rapport aux autres variables. Commençons d'abord par définir ce qui est la pertinence d'une variable (ou d'un ensemble de variables).

Définition 2.2 (Bennani (2001)) *Une variable pertinente est une variable telle que sa suppression entraîne une détérioration des performances (pouvoir de discrimination en classement ou la qualité de prédiction en régression) du système d'apprentissage.*

Plusieurs critères d'évaluation ont été proposés, basés sur des hypothèses statistiques ou sur des heuristiques. Pour un problème de classement (discrimination), les critères d'évaluation sont souvent basés sur les matrices de dispersion intra et inter classes. En effet, ces matrices sont directement liées à la géométrie des classes et donnent une information significative sur la répartition des classes dans l'espace des formes.

On trouve aussi des critères d'évaluation qui utilisent des distances probabilistes ou des mesures d'entropie. Le critère dans ce cas est basé sur l'information mutuelle entre le classement et l'ensemble de variables. Dans le cas des systèmes d'apprentissage connexionnistes, l'évaluation des variables se fait en fonction de l'importance des poids qui est définie comme le changement de l'erreur (de classement ou de régression) dû à la suppression de ces poids.

2.2 Procédures de recherche

En général, on ne connaît pas le nombre optimal m de variables à sélectionner. Ce nombre dépendra de la taille et de la qualité de la base d'apprentissage (la quantité et la qualité d'information disponible) et de la règle de décision utilisée (le modèle). Pour un ensemble de n variables il existe $2^n - 1$ combinaisons de variables possibles où 2 représente deux choix : sélectionner ou ne pas sélectionner une variable. La recherche d'un sous-ensemble de m variables parmi n engendre un nombre de combinaison égal à :

$$\binom{n}{m} = \frac{n!}{(n-m)!\, m!} \tag{1}$$

En grande dimension (n très grand), le nombre de combinaison à examiner devient très élevé et une recherche exhaustive n'est pas envisageable. La recherche d'un sous-ensemble optimal de variables est un problème NP-difficile. Une alternative consiste à utiliser une méthode de recherche de type *Branch & Bound*, (Liu et Motoda, 1998). Cette méthode de recherche permet de restreindre la recherche et donne le sous-ensemble optimal de variables, sous l'hypothèse de monotocité du critère de sélection $J(X)$.

Le critère $J(X)$ est dit monotone si :

$$X_1 \subset X_2 \subset \ldots \subset X_m \implies J(X_1) \subset J(X_2) \subset \ldots \subset J(X_m) \tag{2}$$

où X_k est l'ensemble contenant k variables sélectionnées.

Cependant, la plupart des critères d'évaluation utilisés pour la sélection ne sont pas monotones et dans ce cas on a recours à la seule alternative basée sur des méthodes sous-optimales comme les procédures séquentielles :

- Stratégie ascendante : *Forward Selection (FS)*,
- Stratégie descendante : *Backward Selection (BS)*,
- Stratégie bidirectionnelle : *Bidirectional Selection (BiS)*.

La méthode *FS* procède par agrégations successives (par ajouts successifs de variables). Au départ l'ensemble des variables sélectionnées est initialisé à l'ensemble vide. À chaque étape k, on sélectionne la variable qui optimise le critère d'évaluation $J(X_k)$ et on la rajoute à l'ensemble des variables sélectionnées X_k. Soit X l'ensemble des variables, on sélectionne la variable x_i telle que :

$$J(X_k) = \underset{x_i \in X \setminus X_{k-1}}{max} J(X_{k-1} \cup \{x_i\}) \tag{3}$$

L'ordre d'adjonction des variables à l'ensemble des variables sélectionnées produit une liste ordonnée des variables selon leur importance. Les variables les plus importantes sont les premières variables ajoutées à la liste. Néanmoins, il faut aussi se rappeler que des variables individuellement peu pertinentes peuvent être très informatives si on les utilise conjointement.

La méthode *BS* est une procédure inverse de la précédente (par retraits successifs de variables). On part de l'ensemble complet X des variables et on procède par élimination. À chaque étape la variable la moins importante selon le critère d'évaluation est éliminée. Le procédé continu jusqu'à ce qu'il reste qu'une seule variable dans l'ensemble des variables de départ. À l'étape k, on supprime la variable x_i telle que :

$$J(X_k) = \underset{x_i \in X_{k+1}}{max} J(X_{k+1} \setminus \{x_i\}) \tag{4}$$

Une liste ordonnée selon l'ordre d'élimination des variables est ainsi obtenue. Les variables les plus pertinentes sont alors les variables qui se trouvent dans les dernières positions de la liste.

La procédure *BiS* effectue sa recherche dans les deux directions (*Forward* et *Backward*) d'une manière concurrentielle. La procédure s'arrête dans deux cas : (1) quand une des deux directions a trouvé le meilleur sous-ensemble de variables avant d'atteindre le milieu de l'espace de recherche ; ou (2) quand les deux directions arrivent au milieu. Il est clair que les ensembles de variables sélectionnées trouvés respectivement par *SFS* et par *SBS* ne sont pas égaux à cause de leurs différents principes de sélection. Néanmoins, cette méthode réduit le temps de recherche puisque la recherche s'effectue dans les deux directions et s'arrête dès qu'il y a une solution quelle que soit la direction.

2.3 Critères d'arrêt

Le nombre optimal de variables n'est pas connu a priori, l'utilisation d'une règle pour contrôler la sélection-élimination de variables permet d'arrêter la recherche lorsque aucune variable n'est plus suffisamment informative. Le critère d'arrêt est souvent défini comme une combinaison de la procédure de recherche et du critère d'évaluation. Une heuristique, souvent utilisée, consiste à calculer pour les différents sous-ensembles de variables sélectionnées une estimation de l'erreur de généralisation par validation croisée. Le sous-ensemble de variables sélectionnées est celui qui minimise cette erreur de généralisation.

2.4 Les différentes approches de sélection

Il existe trois grandes familles d'approches :

Approches « Filtres » (*Filters*) : ces méthodes sélectionnent les variables indépendamment de la méthode qui va les utiliser, elles se basent sur les caractéristiques de l'ensemble des données afin de sélectionner certaines variables et d'éliminer d'autres sous forme de pré-traitement des données.

Approches « Symbioses » (*Wrappers*) : contrairement aux approches filtre qui ignorent totalement l'influence des variables sélectionnées sur la performance de l'algorithme d'apprentissage, les approches "enveloppantes" utilisent l'algorithme d'apprentissage comme une fonction d'évaluation.

Approches « Intégrées » (*Embedded*) : ces méthodes exécutent la sélection variable pendant le processus de l'apprentissage. Le processus de la sélection de variables est effectué parallèlement au processus de classement (ou de la régression). Le sous-ensemble de variables ainsi sélectionnées sera choisi de façon à optimiser le critère d'apprentissage utilisé.

2.5 Sélection de variables et apprentissage symbolique

La sélection de variables dans le domaine de l'apprentissage symbolique (Machine Learning) est souvent limitée aux tâches de discrimination de données discrètes. De nombreuses techniques ont été proposées dans ce domaine. La méthode FOCUS (Almuallim, 1994) est basée sur une exploration exhaustive de tous les sous-ensembles de variables et choisir le plus petit sous-ensemble qui couvre le mieux la sortie cible. L'approche ABB (Liu et Motoda, 1998) estime aussi la pertinence d'une variable par une mesure de recouvrement. L'avantage de cette dernière méthode est qu'elle est monotone. La méthode RELIEF (Kira et Rendell, 1992) estime l'importance d'une variable par comparaison de cette variable et la classe correspondante sur plusieurs sous-ensembles de données. Plusieurs autres méthodes ont été basées sur l'utilisation de l'entropie croisée ou la courbe ROC. Ces méthodes sont très intéressantes mais elles sont difficilement utilisables pour les problèmes qui nous concernent dans le cadre de ce projet, i.e. en grande dimension avec des données généralement continues.

2.6 Sélection de variables et apprentissage connexionniste

La sélection de variables dans le domaine connexionniste est très attrayante et soulève de nombreux enjeux à la fois théoriques et applicatifs fondamentaux (Bennani, 2001, 2006). En effet, dans le cas des réseaux connexionnistes, le processus de la sélection de variables peut être effectué parallèlement au processus de classement - ou de la régression. Le sous-ensemble de variables ainsi sélectionnées sera choisi de façon à optimiser le critère d'apprentissage. En plus, le nombre de variables est directement lié à l'architecture et à la complexité de la fonction réalisable par le système connexionniste.

Dans le cas des systèmes d'apprentissage connexionniste, le nombre de variables est directement lié à l'architecture et à la complexité de la fonction réalisable par le modèle connexionniste. Plusieurs approches ont été proposées dans la littérature. La plupart de ces techniques

emploient la première ou la deuxième dérivée de la fonction de coût par rapport aux poids pour estimer l'importance des connexions.

Les méthodes les plus largement employées sont : *Optimal Brain Damage (OBD)* proposée par Le Cun et al. (1990), et *Optimal Brain Surgeon (OBS)* par Hassibi et Stork (1993) qui est une amélioration de la précédente. Pedersen et al. (1996) ont proposé γOBD et γOBS , où l'estimation de l'importance d'un poids est basée sur le changement associé dans l'erreur de généralisation si le poids est élagué. D'autres variantes d'*OBD* et d'*OBS* ont été proposées : *Early Brain Damage (EBD)* et *Early Brain Surgeon (EBS)* (Tresp et al., 1996). On peut citer aussi *Optimal Cell Damage (OCD)* développée par Cibas et al. (1994) qui est une extension de *OBD* pour l'élagage des variables d'entrée. Ces méthodes se basent sur l'estimation systématique de l'importance d'une connexion qui est définie comme le changement de l'erreur causé par la suppression de ce poids. L'emploi des dérivées premières pour la sélection de variables peut être trouvé par exemple dans Dorizzi et al. (1996); Moody (1994); Ruck et al. (1990). D'autres méthodes de sélection de variables utilisent les paramètres du système d'apprentissage. Certaines de ces méthodes emploient : des tests statistiques pour évaluer un intervalle de confiance pour chaque poids (M. et al., 1995), l'information mutuelle pour évaluer un ensemble de caractéristiques et sélectionner un sous-ensemble pertinent (Battiti, 1994), des mesures heuristiques basées sur l'estimation de la contribution des variables dans la prise de décision du système (Bennani et Bossaert, 1995; Yacoub et Bennani, 1997). Dans le cadre de l'apprentissage bayésien MacKay (1994); Neal (1994) proposent une méthode de sélection de variables *Automatic Relevance Determination (ARD)*. Cette méthode utilise des hypothèses de normalité sur la répartition des poids du réseau.

Dans les paragraphes qui suivent, nous allons détailler quelques méthodes en les regroupant par type.

Les méthodes connexionnistes de sélection de variables sont en général de type "backward". L'idée générale est de faire converger un réseau jusqu'à un minimum local en utilisant toutes les variables et de faire ensuite la sélection. L'étape de sélection consiste à trier les variables par ordre croissant de pertinence, supprimer la ou les variables les moins pertinentes et ré-entraîné le réseau avec les variables restantes. Ce processus continue tant qu'un certain critère d'arrêt n'est pas satisfait. Les méthodes qui suivent cette procédure comportent donc deux phases : une phase d'apprentissage et une phase d'élagage qui peuvent être alternées. On peut dire qu'une "vraie" procédure connexionniste de sélection de variables suit l'algorithme général suivant :

1. Atteindre un minimum local
2. Calculer la pertinence de chaque entrée
3. Trier les entrées par ordre croissant de pertinence
4. Supprimer les entrées dont la pertinence cumulée est inférieur à un seuil fixé
5. Recommencer en 1. Tant que les performances estimées sur une base de validation ne chutent pas

Les méthodes de sélection de variables en apprentissage connexionniste peuvent se regrouper en trois grandes familles :

- Les méthodes d'ordre zéro
- Les méthodes du premier ordre
- Les méthodes du second ordre

2.6.1 Méthodes d'ordre zéro

Pour estimer la pertinence d'une variable, les mesures d'ordre zéro utilisent les valeurs des paramètres du système d'apprentissage (les valeurs des connexions, la structure, ...). Par exemple la mesure de pertinence HVS (Yacoub et Bennani, 1997) repose sur les paramètres et la structure du réseau connexionniste. Dans le cas d'un Perceptron multicouches à une seule couche cachée, cette mesure est définie par :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \sum\limits_{j \in Hidden} \left[\frac{|\omega_{ji}|}{\sum\limits_{i' \in Input} |\omega_{ji'}|} \times \sum\limits_{k \in Output} \frac{|\omega_{kj}|}{\sum\limits_{j' \in Hidden} |\omega_{kj'}|} \right] \\ \text{critère d'évaluation} & J(X_k) = \sum\limits_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} + \text{réapprentissage} \\ \text{critère d'arrêt} & \text{test statistique} \end{cases} \tag{5}$$

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \sum\limits_{j \in Hidden} \left[\frac{|\omega_{ji}|}{\sum\limits_{i' \in Input} |\omega_{ji'}|} \times \sum\limits_{k \in Output} \frac{|\omega_{kj}|}{\sum\limits_{j' \in Hidden} |\omega_{kj'}|} \right] \\ \text{critère d'évaluation} & J(X_k) = \sum\limits_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} + \text{réapprentissage} \\ \text{critère d'arrêt} & \text{test statistique} \end{cases} \tag{6}$$

Une autre méthode d'ordre zéro très efficace a été proposée par MacKay (1994) : *Automatic Relevance Determination (ARD)*. Dans cette méthode la pertinence d'une variable est estimée par la variance de ses poids : la variable est éliminée si la variance correspondante est faible.

2.6.2 Méthodes du premier ordre

La dérivée de la fonction ψ que représente un système d'apprentissage connexionniste - un réseau - par rapport à chacune de ses variables est très utilisée comme mesure de pertinence des variables. Si une dérivée est proche de zéro pour tous les exemples, alors la variable correspondante n'est pas utilisée par le réseau, et peut donc être supprimé.

Dans le cas des PMC - Perceptrons multicouches -, cette dérivée peut se calculer comme une extension de l'algorithme d'apprentissage. Comme ces dérivées peuvent prendre aussi bien des valeurs positives que négatives, produisant une moyenne proche de zéro, c'est la moyenne des valeurs absolues qui est généralement utilisée - ce sont les grandeurs des dérivées qui nous intéressent. On trouve beaucoup de mesures de pertinences basées sur cette approche.

La sensibilité de l'erreur à la suppression de chaque variable est utilisée par Moody dans Moody (1994). Une mesure de sensibilité est calculée pour chaque variable x_i pour évaluer la variation de l'erreur en apprentissage si cette variable est supprimée du réseau. Le remplacement d'une variable par sa moyenne supprime son influence sur la sortie du réseau. La définition de la pertinence est :

$$\zeta_i = R(\omega) - \tilde{R}(\overline{x}_i, \omega) \tag{7}$$

$$\text{avec } \tilde{R}(\overline{x}_i, \omega) = \frac{1}{N} \sum_{k=1}^{N} \left\| y^k - \psi(x_1^k, \ldots, \overline{x}_i^k, \ldots, x_n^k) \right\|^2 \tag{8}$$

N est la taille de la base d'apprentissage. Quand cette taille est très grande, Moody propose d'utiliser une approximation qui donne la méthode de sélection suivante :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i \overset{N\to\infty}{\cong} \frac{1}{N} \sum_{k=1}^{N} \left(x_i^k - \overline{x}_i\right) \left(y^k - \psi(x^k, \omega)\right) \frac{\partial \psi(x^k, \omega)}{\partial x_i} \\ \text{critère d'évaluation} & J(X_k) = \sum_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{variation des performances en test} \end{cases} \tag{9}$$

Ruck et al. (1990) proposent la méthode suivante :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \sum_{k=1}^{N} \sum_{j \in Output} \left| \frac{\partial \psi_j(x^k, \omega)}{\partial x_i} \right| \\ \text{critère d'évaluation} & J(X_k) = \sum_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{seuil : moyenne des pertinences} \end{cases} \tag{10}$$

Refenes et Zapranis (1999) utilisent l'élasticité moyenne de la sortie par rapport à chaque variable :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \frac{1}{N} \sum_{k=1}^{N} \left| \frac{\partial \psi(x^k, \omega)}{\partial x_i} \times \frac{x_i}{\psi(x^k, \omega)} \right| \\ \text{critère d'évaluation} & J(X_k) = \sum_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{seuil : moyenne des pertinences} \end{cases} \tag{11}$$

Dans le cas des réseaux à fonctions radiales RBF - *Radial Basis Functions* -, Dorizzi et al. (1996) utilisent le quantile à 95% de la distribution des valeurs absolues des dérivées de chaque variable.

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = q_{.95}\left[\left|\dfrac{\partial\psi(x,\omega)}{\partial x_i}\right|\right] \\ \text{critère d'évaluation} & J(X_k) = \sum\limits_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{seuil : moyenne des pertinences} \end{cases} \tag{12}$$

Pour un problème de discrimination, Rossi (1996) propose de ne considérer que les exemples qui sont près des frontières interclasses :

$$x^k \in frontier \;\equiv\; \left\|\nabla_{x^k}\psi(x^k,\omega)\right\| > \epsilon \tag{13}$$

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \frac{1}{|Output|} \sum\limits_{x^k \in frontier} \sum\limits_{j \in Output} \dfrac{\left|\frac{\partial\psi_j(x^k,\omega)}{\partial x_i}\right|}{\left\|\frac{\partial\psi_j(x^k,\omega)}{\partial x}\right\|} \\ \text{critère d'évaluation} & J(X_k) = \sum\limits_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{seuil : moyenne des pertinences} \end{cases} \tag{14}$$

2.6.3 Méthodes du second ordre

Pour estimer la pertinence d'une variable, les méthodes du second ordre calculent la dérivée seconde de la fonction de coût par rapport aux poids. Ces mesures sont des extensions des techniques d'élagage des poids. La technique d'élagage la plus populaire est *Optimal Brain Damage (OBD)* proposée par Le Cun et al. (1990). *OBD* est basée sur l'estimation de la variation de la fonction de coût $R(w)$ lorsqu'un poids est supprimé du réseau. Cette variation peut être approximée à l'aide d'un développement en série de Taylor :

$$\delta\tilde{R}(\omega_i) = \sum_i \frac{\partial\tilde{R}(\omega)}{\partial\omega_i}\delta\omega_i + \frac{1}{2}\sum_i\sum_j \frac{\partial^2\tilde{R}(\omega)}{\partial\omega_i\partial\omega_j}\delta\omega_i\delta\omega_j + O\left(\delta\omega^3\right) \tag{15}$$

Sous l'hypothèse que le réseau connexionniste a atteint un minimum local, le premier terme de droite de cette formule est nul. Pour simplifier les calculs, Le Cun et al. (1990) supposent en outre que la matrice Hessienne est nulle et le coût est localement quadratique. On obtient

alors la formule simplifiée suivante :

$$\delta\tilde{R}(\omega_i) \approx \frac{1}{2}\sum_i \frac{\partial^2\tilde{R}(\omega)}{\partial\omega_i^2}\delta\omega_i^2 + O\left(\delta\omega^3\right) \quad (16)$$

$$\approx \frac{1}{2}H_{ii}\delta\omega_i^2 \quad (17)$$

La pertinence d'une connexion est alors estimée par :

$$pertinence(\omega_i) \approx \frac{1}{2}H_{ii}\omega_i^2 \quad (18)$$

La méthode de sélection de variables *Optimal Cell Damage (OCD)* développée par Cibas et al. (1994) est basée sur la mesure de pertinence ci-dessus. Dans *OCD*, l'importance de chaque variable s'obtient en sommant les importances des connexions qui partent de celle-ci :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \frac{1}{2}\sum\limits_{j\in fan-Out(i)} \dfrac{\partial^2\tilde{R}(w)}{\partial\omega_{ji}^2}\omega_{ji}^2 \\ \text{critère d'évaluation} & J(X_k) = \sum\limits_{x_i\in X_k}\zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{test statistique} \end{cases} \quad (19)$$

où $fan - Out(i)$ est l'ensemble des neurones qui utilisent comme entrée la sortie du neurone i.

Dans *OBD* et *OBS*, la sensibilité d'un poids ne peut être évaluée correctement qu'autour d'un minimum local de la fonction de coût. Tresp et al. (1996) proposent deux extensions d'*OBD* et d'*OBS* : *Early Brain Damage (EBD)* et *Early Brain Surgeon (EBS)*. *EBD* et *EBS* peuvent être utilisées avec le "*early stopping*" comme critère d'arrêt de l'apprentissage. Dans *EBD*, par exemple, la sensibilité d'un poids est donnée par la formule suivante :

$$pertinence(\omega_i) = \frac{1}{2}\frac{\partial^2\tilde{R}(w)}{\partial\omega_{ji}^2}\,\omega_{ji}^2 - \frac{\partial\tilde{R}(w)}{\partial\omega_{ji}}\,\omega_{ji} + \frac{\left(\dfrac{\partial\tilde{R}(w)}{\partial\omega_{ji}}\right)^2}{\dfrac{\partial^2\tilde{R}(w)}{\partial\omega_{ji}^2}} \quad (20)$$

A partir de cette définition de pertinence et de la même façon que *OCD*, Leray et Gallinari

(2001) propose la méthode *ECD (Early Cell Damage)* :

$$\begin{cases} \text{pertinence d'une variable} & \zeta_i = \frac{1}{2} \sum_{j \in fan-Out(i)} \frac{\partial^2 \tilde{R}(w)}{\partial \omega_{ji}^2} \omega_{ji}^2 - \frac{\partial \tilde{R}(w)}{\partial \omega_{ji}} \omega_{ji} + \frac{\left(\frac{\partial \tilde{R}(w)}{\partial \omega_{ji}}\right)^2}{\frac{\partial^2 \tilde{R}(w)}{\partial \omega_{ji}^2}} \\ \text{critère d'évaluation} & J(X_k) = \sum_{x_i \in X_k} \zeta_i \\ \text{procédure de recherche} & \textit{Backward} \\ \text{critère d'arrêt} & \text{test statistique} \end{cases} \tag{21}$$

Pour cette méthode on supprime les variables une par une et on peut utiliser la technique de *early stopping* pour arrêter l'apprentissage.

2.7 Sélection de variables et apprentissage non supervisé

Contrairement à la sélection de variables pour les systèmes d'apprentissage supervisé, relativement peu d'approches ont été proposées pour l'apprentissage non-supervisé (classification automatique ou *clustering*). En effet, le problème de la sélection de variables en classification automatique est un problème beaucoup plus difficile que dans le cas supervisé (discrimination) où les données sont étiquetées (Guyon et al., 2006). Un autre problème important associé à la classification concerne la détermination automatique du nombre de groupes (*clusters*) qui est clairement influencé par l'issue de la sélection des variables. Enfin, la question ouverte est comment évaluer/comparer les résultats de plusieurs classifications ?

Le théorème d'impossibilité proposé par (Kleinberg, 2002) indique qu'il n'existe pas de méthode de classification qui vérifie simultanément les trois propriétés suivantes :

- invariance à l'unité de mesure des distances : la multiplication par un scalaire de distance utilisée par un algorithme ne modifie pas la partition qu'il découvre,
- exhaustivité : pour toute partition de l'ensemble des individus, il existe une distance qui permette à l'algorithme de classification de la découvrir,
- consistance : si une partition de l'ensemble des individus.

2.7.1 Approches filtres non supervisées

La majeure partie des approches filtres que l'on rencontre en apprentissage non supervisé peuvent être regroupées en deux catégories :

- celles qui s'appuient sur le calcul de corrélations ou sur l'estimation de l'information mutuelle entre variables et qui permettent d'éliminer soit les variables redondantes (Mitra et al., 2002; Vesanto et Alhoniemi, 2000), soit les variables non pertinentes (Sorg-Madsen et al., 2003) ;
- celles qui se fondent sur la notion de densité de l'espace des données (Dash et al., 2002; He et al., 2006; Pal et al., 2000).

Corrélation et information mutuelle

On considère souvent qu'un sous-ensemble optimal de variables pertinentes doit être minimal, cela a conduit certains auteurs à proposer des méthodes de sélection de variables qui se focalise sur l'élimination des dimensions redondantes. Ainsi, Mitra et al. (2002) 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. Vesanto et Ahola (1999) proposent une détection visuelle des corrélations en se basant sur la construction d'une carte auto-organisées.

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 :

$$I(X,Y) = -\int_{x,y} p(X=x, Y=y)\ log\ \frac{p(X=x, Y=y)}{p(X=x)\,p(Y=y)}\ dx \tag{22}$$

$$= H(X) + H(Y) - H(X,Y) \tag{23}$$

$$\text{avec } H(X) = -\int_{x} p(X=x)\ log\, p(X=x)\, dx \tag{24}$$

où $H(X)$ est l'entropie au sens de Shanon 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).

Lorsqu'on cherche à segmenter un ensemble d'individus en apprentissage non supervisé, on considère qu'une dimension pertinente est généralement liée à une variable latente qui indique le groupe des observations. Ainsi, en supposant qu'il existe au moins deux variables pertinentes, (Sorg-Madsen et al., 2003) proposent d'utiliser une stratégie ascendante guidée par une mesure de dépendance entre chaque couple de variables ; les dimensions indépendantes de toutes les autres sont alors associées à du bruit et sont éliminées. Ils utilisent deux mesures de liaisons différentes : l'information mutuelle et le pouvoir prédictif mutuel.

Utilisation de la densité des données

L'objectif de l'apprentissage non supervisé est de découvrir et de comprendre la structure d'un ensemble d'individus ; ainsi, une variable distribuée uniformément peut être considérée comme non pertinente car elle ne met aucune structure en exergue. En se basant sur cette observation, Dash et al. (2002) proposent une nouvelle mesure d'entropie pour guider une approche ascendante. Une formulation légèrement différente de cette observation a été proposée par He et al. (2006) : *« dans de nombreux problèmes d'apprentissage comme la classification,*

la structure locale de l'espace des données est plus importante que la structure globale. » Ils proposent alors une mesure d'évaluation, le score laplacien (*Laplacian Score*), qui est basée sur le respect de la structure d'un graphe de voisinage entre les individus. Antérieurement, Pal et al. (2000) avaient utilisé un degré d'appartenance de deux individus à un même ensemble flou qu'ils faisaient apprendre à un réseau neuromimétique de type perceptron multicouche. Les poids du réseaux indiquent alors la contribution de chaque dimension au degrés d'appartenance et peuvent utilisés commme mesure d'évaluation.

Approches symbioses et intégrées

Sorg-Madsen et al. (2003) complètent leur approche « filtre » présentée plus haut en l'hybridant avec une approche « symbiose » : après avoir éliminer les variables non pertinentes, ils estiment les paramètres d'un modèle de mélange. Ils se ramènent ainsi au cas de l'apprentissage supervisé et construisent un classificateur naïf de Bayes dont la précision est utilisée pour guider une procédure de recherche ascendante. Dans Dy et Brodley (2000, 2004), les auteurs restent dans le cadre non supervisé en proposant différentes approches ascendantes basées sur les modèles de mélanges qui sont guidées soit le maximum de vraisemblance, soit la séparabilité des classes.

Guérif et Bennani (2006) utilisent une classification à deux niveaux combinée à la valeur test (Morineau, 1984) pour identifier les variables les plus significatives ; le premier niveau de la classification est formée par une carte auto-organisée (Kohonen, 2001) qui est segmentée en utilisant l'algorithme des k-moyennes associé à l'indice de Davies-Bouldin (Davies et Bouldin, 1979) pour fixer le nombre de groupes (Vesanto et Alhoniemi, 2000). La statistique Λ de Wilks est utilisée pour stopper leur procédure en se basant sur la séparabilité des classes.

La sélection de variables peut être vue comme un problème de sélection de modèles. Ainsi, Raftery et Dean (2006) adoptent une recherche séquentielle bidirectionnelle et ils retiennent le modèle optimal au sens du critère BIC (*Bayesian Information Criterion*) (Schwarz, 1978). Leur méthode permet de considérer à chaque étape différents modèles plus ou moins contraints comportant un nombre variable de groupes. Dans Law et al. (2004), les auteurs définissent une mesure de saillance (*saliency*) et ajoutent de nouveaux paramètres aux modèles de mélange pour intégrer la sélection de variables directement à la fonction de coût optimisée par l'algorithhme EM. Ils déterminent le nombre de composantes du mélange selon le critère MML (*Minimum Message Length*).

2.7.2 Evaluation et critères de validité

Dans le contexte de la classification automatique, il est naturel de s'interroger sur la validité de la partition obtenue. Les groupes découverts correspondent-ils à nos connaissances à priori ? Correspondent-ils vraiment à l'ensemble d'objets dont on dispose ? De deux classifications, laquelle est la plus pertinente ? Ces différentes questions permettent de distinguer trois catégories de critères (Jain et Dubes, 1988) :

- les **critères externes** qui permettent de répondre à la première question et de mesurer l'adéquation entre une partition et les connaissances à priori dont on dispose ;
- les **critères internes** qui quantifient l'adéquation entre une partition et l'idée subjective que l'on se fait d'une « bonne » classification ; ainsi, les propriétés les plus communément recherchées sont la compacité et la séparabilité des groupes découverts ;

- les **critères relatifs** qui s'intéressent à la troisième et dernière question et à défaut de donner une appréciation absolue de la validité d'une partition, ils permettent d'ordonner plusieurs classifications et d'en choisir « une meilleure ».

Critères externes

Les critères externes se ramènent au problème ancien de la comparaison de partitions et une littérature abondante est disponible sur le sujet (Fowlkes et Mallows, 1983; Hubert et Arabie, 1985; Meilă, 2003, 2005, 2007; Rand, 1971; Wallace, 1983). Une manière simple de comparer deux partitions $\mathcal{C}$ et $\mathcal{C}'$ consiste à contruire une table de contingence (figure 3) qui donne une appréciation intuitive de leur adéquation. Le calcul de la plupart des critères s'appuie d'ailleurs soit directement sur cette table soit un comptage des accords et des désaccords que l'on peut déduire à l'aide des formules de linéarisation suivantes (Hubert et Arabie, 1985; Jain et Dubes, 1988) :

$$N_{00} = \frac{1}{2}\left(n^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i.}^2 + \sum_{j=1}^{K'} n_{.j}^2\right)\right) \quad (25)$$

$$N_{11} = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1) \quad (26)$$

$$N_{01} = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{.j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \quad (27)$$

$$N_{10} = \frac{1}{2}\left(\sum_{i=1}^{K'} n_{i.}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \quad (28)$$

Les nombres de paires d'objets qui sont séparés ou regroupés dans les deux partitions sont notés respectivement N_{00} et N_{11}. N_{01} indique le nombre de paires d'objets séparés dans la première partition et regroupés dans la seconde. De manière analogue, N_{10} désigne le nombre de paires d'objets regroupés dans la première partition et séparés dans la seconde.

	$\mathcal{C}'_1$	$\dots$	$\mathcal{C}'_j$	$\dots$	$\mathcal{C}'_{K'}$	
$\mathcal{C}_1$	n_{11}	$\dots$	n_{1j}	$\dots$	$n_{1K'}$	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathcal{C}_i$	n_{i1}	$\dots$	n_{ij}	$\dots$	$n_{iK'}$	$n_{i.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathcal{C}_K$	n_{K1}	$\dots$	n_{Kj}	$\dots$	$n_{KK'}$	$n_{K.}$
	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.K'}$	N

FIG. 3 – *Exemple de table de contingence entre de deux partitions $\mathcal{C} = \{\mathcal{C}_i : i = 1, \dots, K\}$ et $\mathcal{C}' = \{\mathcal{C}'_j : j = 1, \dots, K'\}$ d'un même ensemble de N objets ; les marges $n_{i.}$ et $n_{.j}$ indiquent respectivement les effectifs des classes $\mathcal{C}_i$ et $\mathcal{C}'_j$.*

Il convient de remarquer que les critères de comparaison que l'on peut construire à partir de N_{00}, N_{11}, N_{01} et N_{10} correspondent à des mesures de dissimilarité binaires dont un grand nombre peuvent s'exprimer sous la forme suivante (Li, 2006) :

$$d_{\alpha,\delta} = \frac{N_{10} + N_{01}}{\alpha N_{11} + N_{10} + N_{01} + \delta N_{00}} \tag{29}$$

où α et δ sont deux paramètres qui permettent de pondérer la prise en compte respective des regroupement ou séparation simultanées d'une paire d'objets dans deux partitions. La table 1 rappelle la définition de quelques mesures et le lecteur intéressé en trouvera une présentation plus complète de ces mesures ou de leur propriété dans Albatineh et al. (2006), Jouve et al. (2001), Lourenço et al. (2004), Roux (1985) et Li (2006).

Mesure	Similarité	Dissimilarité
Sokal & Sneath (I)	$\frac{\frac{1}{2}N_{11}}{\frac{1}{2}N_{11}+N_{10}+N_{01}}$	$\frac{N_{10}+N_{01}}{\frac{1}{2}N_{11}+N_{10}+N_{01}}$
Rogers & Tanimoto	$\frac{\frac{1}{2}(N_{11}+N_{00})}{\frac{1}{2}(N_{11}+N_{00})+N_{10}+N_{01}}$	$\frac{N_{10}+N_{01}}{\frac{1}{2}(N_{11}+N_{00})+N_{10}+N_{01}}$
Jaccard	$\frac{N_{11}}{N_{11}+N_{10}+N_{01}}$	$\frac{N_{10}+N_{01}}{N_{11}+N_{01}+N_{10}}$
Rand	$\frac{N_{11}+N_{00}}{N_{11}+N_{10}+N_{01}+N_{00}}$	$\frac{N_{10}+N_{01}}{N_{11}+N_{10}+N_{01}+N_{00}}$
Czekanowski-Dice	$\frac{2N_{11}}{2N_{11}+N_{10}+N_{01}}$	$\frac{N_{10}+N_{01}}{2N_{11}+N_{10}+N_{01}}$
Sokal & Sneath (II)	$\frac{2(N_{11}+N_{00})}{2(N_{11}+N_{00})+N_{10}+N_{01}}$	$\frac{N_{10}+N_{01}}{2(N_{11}+N_{00})+N_{10}+N_{01}}$
Kulczynski (II)	$\frac{1}{2}\left(\frac{N_{11}}{N_{11}+N_{10}}+\frac{N_{11}}{N_{11}+N_{01}}\right)$	$1-\frac{1}{2}\left(\frac{N_{11}}{N_{11}+N_{10}}+\frac{N_{11}}{N_{11}+N_{01}}\right)$
Ochiai	$\frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}}$	$1-\frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}}$
Russel & Rao	$\frac{N_{11}}{N_{11}+N_{10}+N_{01}+N_{00}}$	$1-\frac{N_{11}}{N_{11}+N_{10}+N_{01}+N_{00}}$

TAB. 1 – *Quelques mesures de similarité et de dissimilarité binaire.*

D'autres critères se calculent directement à partir de la table de contingence. Le critère de Larsen, le critère de Meilă & Heckerman (Meilă, 2005, 2006), la distance de transfert maximum (Charon et al., 2006) ou la variation d'information (Meilă, 2003, 2005, 2007) en sont des exemples. Enfin, il convient de rappeler qu'une part non négligeable de la similarité entre deux partitions doit être attribuée au hazard et que cela a conduit de nombreux auteurs à proposer des méthodes de correction d'indices ; le lecteur qui souhaite approfondir cette question est invité à consulter Albatineh et al. (2006), Fowlkes et Mallows (1983), Hubert et Arabie (1985) ou encore Meilă (2007).

Critères internes

Les critères internes visent à quantifier l'adéquation entre une partition et l'idée subjective que l'on se fait d'une « bonne » classification en se basant uniquement sur les propriétés des données. Les valeurs de ce type de critère sont généralement très dépendantes du jeu de données utilisé et déterminer une valeur de référence peut s'avérer coûteux. Nous ne détaillerons pas ce point dans le cadre de cet article mais le lecteur est invité à consulter Jain et Dubes (1988) pour obtenir davantage d'information.

Critères relatifs

Les critères relatifs sont les plus largement utilisés et permettent d'ordonner différentes partitions en fonction de l'idée subjective que l'on se fait d'une « bonne » classification ; ainsi, on recherche généralement des groupes compacts et bien séparés. Etablir une liste exhaustive de ce type de critères dépasse largement le cadre de cet article et seules les définitions de deux indices de ce type sont rappelées ci-dessous :

– **Indice de Dunn :** dans le cas d'une classification dure, l'indice de Dunn (Halkidi et al., 2001, 2002a,b) tient compte à la fois de la compacité et de la séparabilité des groupes : la valeur de cet indice est d'autant plus faible que les groupes sont compacts et bien séparés. Notons que la complexité de l'indice de Dunn devient prohibitive dès qu'on manipule de grands ensembles d'objets ; il est par conséquent rarement utilisé.

$$I_{Dunn} = \frac{\min\{D_{min}(\mathcal{C}_i, \mathcal{C}_j) \ : \ i \neq j\}}{\max\{S_{max}(\mathcal{C}_i)\}} \tag{30}$$

où $D_{min}(\mathcal{C}_i, \mathcal{C}_j)$ est la distance minimale qui sépare un objet du groupe $\mathcal{C}_i$ d'un objet du groupe $\mathcal{C}_j$ et où $S_{max}(\mathcal{C}_i)$ est la distance maximale qui sépare deux objets du groupe $\mathcal{C}_i$:

$$D_{min}(\mathcal{C}_i, \mathcal{C}_j) = \min\{\|x - y\| \ : \ x \in \mathcal{C}_i \text{ et } y \in \mathcal{C}_j\} \tag{31}$$

$$S_{max}(\mathcal{C}_i) = \max\{\|x - y\| \ : \ (x, y) \in \mathcal{C}_i \times \mathcal{C}_i\} \tag{32}$$

– **Indice de Davies-Bouldin :** dans le cas d'une classification dure, l'indice de Davies-Bouldin (Davies et Bouldin, 1979) tient compte à la fois de la compacité et de la séparabilité des groupes : la valeur de cet indice est d'autant plus faible que les groupes sont compacts et bien séparés. Cet indice dont la complexité en $\theta\left(K \times (N + K)\right)$ est raisonnable favorise les groupes hypersphériques et il est donc particulièrement bien adapté pour une utilisation avec la méthode des K-moyennes.

$$I_{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} \left\{ \frac{S_c(\mathcal{C}_k) + S_c(\mathcal{C}_l)}{D_{ce}(\mathcal{C}_k, \mathcal{C}_l} \right\} \tag{33}$$

où $S_c(\mathcal{C}_i)$ est la distance moyenne entre un objet du groupe $\mathcal{C}_i$ et son centre, et où $D_{ce}(\mathcal{C}_i, \mathcal{C}_j)$ est la distance qui sépare les centres des groupes $\mathcal{C}_i$ et $\mathcal{C}_j$:

$$S_c(\mathcal{C}_i) = \frac{1}{N_i} \sum_{i=1}^{N_i} \|x - \omega_i\| \tag{34}$$

$$D_{ce}(\mathcal{C}_i, \mathcal{C}_j) = \|\omega_i - \omega_j\| \tag{35}$$

2.7.3 Approches symbioses et intégrées non supervisées

De nombreuses approches de sélection de variables pour la classification automatique utilisent les modèles de mélanges comme cadre théorique (Dy et Brodley, 2000, 2004; Law et al., 2004; Raftery et Dean, 2006; Sorg-Madsen et al., 2003) et s'appuient sur l'algorithme EM (*Expectation Maximization*) (Dempster et al., 1977).

Principes des modèles de mélange

On suppose que l'ensemble d'individus dont on dispose a été obtenu en fusionnant plusieurs sous-populations qui suivent chacune une loi de probabilité propre. La probabilité qu'un individu x soit issu de ce mélange de paramètres $\theta = (\alpha_1, \theta_1, \ldots, \alpha_i, \theta_i, \ldots)$ est alors donnée par :

$$p(x|\theta) = \sum_i \alpha_i \times p_i(x|\theta_i) \tag{36}$$

où les coefficients de mélange α_i satisfont $\sum_i \alpha_i = 1$, et où les densités de probabilité de chaque sous-population $\mathcal{C}_i$ sont données par les lois $p_i(x|\theta_i)$ de paramètres θ_i. Rappelons que toute distribution continue peut être approximée à l'aide d'un modèle de mélange dès lors que ses composantes sont assez nombreuses et que leurs paramètres sont bien choisis.

L'estimation du nombre et des paramètres de composantes est un problème difficile et dans la plupart des applications seuls les mélanges de lois normales sont considérés. Lorsqu'on impose de plus que toutes les lois normales du mélange aient la matrice identité comme matrice de covariance, on retrouve le cas des k-moyennes.

Algorithme EM

L'algorithme le plus répandu pour estimer les paramètres d'un mélange est l'algorithme EM (*Expectation Maximization*) introduit par Dempster et al. (1977). Il consiste à itérer les deux phases suivantes jusqu'à ce que l'amélioration de la log vraisemblance du modèle soit inférieure à un seuil $\epsilon > 0$ fixé :

1. **Estimation :** on suppose fixés les paramètres $\hat{\theta} = \left(\hat{\alpha}_1, \hat{\theta}_1, \hat{\alpha}_2, \hat{\theta}_2, \ldots\right)$ du modèle et on calcule la probabilité $p(x|\hat{\theta}_i)$ qu'un objet $x \in \Omega$ ait été généré par la composante correspondant à la sous-population $\mathcal{C}_i$:

$$p(x|\hat{\theta}_i) = \frac{\alpha_i \times p(x|\hat{\theta}_i)}{\sum_k \alpha_k \times p(x|\hat{\theta}_k)} \tag{37}$$

2. **Maximisation :** on suppose cette fois fixée la partition floue de l'ensemble des objets $x \in \Omega$ dont les degrés d'appartenance sont donnés par les probabilités $p(x|\hat{\theta}_i)$. On cherche alors les paramètres $\widetilde{\theta}$ du modèle qui maximisent sa log vraisemblance

$$\log L(\theta|\Omega) = \sum_{x \in \Omega} p(x|\theta) \tag{38}$$

$$\widetilde{\theta} = \arg\max_{\theta} \{\log L(\theta|\Omega)\} \tag{39}$$

Les coefficients optimaux du mélange sont définis par :

$$\widetilde{\alpha_i} = \frac{1}{N} \sum_{x \in \Omega} x \times p(x|\hat{\theta}_i) \tag{40}$$

où N est le nombre d'individus présents dans Ω.

2.7.4 Stabilité

Une étude de la stabilité d'une classification est un moyen d'évaluer la validité des groupes formés ; en d'autres termes, cela permet de vérifier que les groupes formés ne sont pas le fruit du hasard mais correspondent effectivement à une structure cachée mais présente dans les données. Ainsi, Ben-Hur et al. (2002) proposent de combiner les techniques de ré-échantillonnage à l'indice de Jaccard pour déterminer le nombre de groupes naturels d'un ensemble d'individus. Fred et Jain (2002, 2005) définissent une nouvelle mesure de similarité entre individus en s'appuyant sur la stabilité des groupes formés par différents algorithmes ou en utilisant différentes valeurs de paramètres ; deux individus sont d'autant plus similaires qu'ils sont souvent regroupés.

2.8 Réduction des dimensions par extraction de traits

Les méthodes utilisées pour l'extraction de traits sont très variées, et nous ne prétendons pas de ce court document en faire le tour. Nous rappellerons brièvement les principes des méthodes linéaires (ACP, MDS), puis décrirons quelques méthodes non linéaires qui ont fait l'objet de nombreuses études depuis cinq ans. Nous nous intéressons en particulier aux méthodes utilisant des graphes, comme Isomap, LLE et leurs variantes. Dans le cadre du projet InfoMagic, nous nous pencherons plus particulièrement sur les applications de ces méthodes au traitement des données textuelles.

On considère un espace d'observations χ, qui n'est pas nécessairement R^n, ce qui permet de généraliser les méthodes proposées aux cas où l'on ne dispose pas d'une représentation vectorielle des données à traiter, par exemple les données structurées (arbres ou graphes). L'espace de caractéristiques H est relié à l'espace d'observation par une application :

$$\begin{aligned} \Phi \quad : \quad & \chi \rightarrow H \\ & x \mapsto \phi(x) \end{aligned}$$

Les données d'apprentissage sont un ensemble fini de points $\{x_i\}$, ou bien, dans le cas de l'apprentissage supervisé, un ensemble fini de couples (point, étiquette) $\{(x_i, y_i)\}$.

2.8.1 Méthodes linéaires

Nous rappelons brièvement les principes de deux méthodes classiques d'analyse de données, qui sont le fondement de plusieurs méthodes non linéaires plus récentes.

Analyse en Composantes Principales (ACP)

L'analyse en composantes principales (ACP) est une ancienne approche (aussi connue sous le nom de transformation de Karhunen Loeve dans la communauté du traitement de signal), qui effectue une réduction de dimension par projection des points orignaux dans un sous-espace vectoriel de dimension plus réduite. L'ACP détermine des axes de projections orthogonaux, qui maximisent la variance expliquée. Dans la base formée par ces axes, les coordonnées ne sont pas corrélées.

L'ACP maximise la variance de la projection dans l'espace de caractéristiques, ce qui est équivalent à minimiser l'erreur quadratique moyenne de reconstruction.

L'ACP se calcule en diagonalisant la matrice de corrélations, le plus souvent en utilisant une décomposition en valeurs singulières (SVD). L'analyse en composantes principales est très utilisée car elle est simple à mettre en oeuvre. Elle est limitée par son caractère linéaire : il est facile d'imaginer des situations dans lesquelles l'ACP n'apporte aucune information utilisable (par exemple, des données réparties sur un tore en dimension n).

Multi-Dimensional Scaling (MDS)

Dans de nombreux cas, on connaît les distances entre les points d'un ensemble d'apprentissage (on peut utiliser une mesure de similarité plus sophistiquée que la distance euclidienne, comme indiquée dans la section suivante), et on cherche à obtenir une représentation en faible dimension de ces points. La méthode de positionnement multidimensionnel (MDS) permet de construire cette représentation. L'exemple classique est d'obtenir la carte d'un pays en partant de la connaissance des distances entre chaque paire de villes. L'algorithme MDS est basé sur une recherche de valeurs propres.

MDS permet de construire une configuration de m points dans R^d à partir des distances entre m objets. On observe donc $m(m-1)/2$ distances. Il est toujours possible de générer un positionnement de m points en m dimensions qui respecte exactement les distances fournies. MDS calcule une approximation en dimension $d < m$.

L'algorithme est le suivant :

- Moyennes des distances carrées par rangées : $\mu_i = \frac{1}{n} \sum_j D_{ij}$
- Double centrage (distance carrée vers produit scalaire) :

$$P_{ij} = -\frac{1}{2} \left(D_{ij} - \mu_i - \mu_j + \sum_i \mu_i \right)$$

- Calcul des vecteurs propres v_j et valeurs propres λ_j principales de la matrice P (avec les λ_j^2 les plus grands).
- La i-ème coordonnée réduite de l'exemple j est $\sqrt{\lambda_j}\, v_{ij}$

Notons que la matrice de distance D doit être semi définie positive. Les méthodes linéaires comme l'ACP et le MDS ne donnent des résultats intéressants que si les données sont situées sur un sous-espace linéaire. Elles ne peuvent traiter le cas où les données sont sur une variété très non linéaire.

2.8.2 Méthodes non-linéaires

Les méthodes linéaires reposent (au moins implicitement) sur l'utilisation d'une distance euclidienne (liée au produit scalaire ordinaire). Dans de nombreuses applications, la distance euclidienne n'a pas grand sens ; elle suppose en particulier que toutes les variables sont comparables entre elles (elles doivent donc avoir été convenablement normalisées). La théorie des espaces de Hilbert permet de définir d'autres produits scalaires, basés sur des fonctions noyaux $k(x, y)$. k est alors une mesure de similarité entre les points de l'ensemble à traiter. Le noyau k défini implicitement une application de l'espace d'origine vers un "espace de caractéristiques" H. La dimension de l'espace H est éventuellement infinie. De nombreuses méthodes statistiques peuvent s'exprimer en ne recourant qu'à des produits scalaires entre les points à traiter et les exemples d'apprentissage. Si l'on remplace le produit scalaire habituel par un noyau k, on rend la méthode non-linéaire ; c'est le "truc du noyau" (*kernel trick*), qui a fait l'objet de nombreuses recherches depuis son introduction par Vapnik Boser et al. (1992) dans le cadre des machines à vecteurs de support (SVM).

La notion de noyau peut être utilisée pour la réduction de dimension, comme nous allons le voir dans la section suivante.

Kernel PCA

La première approche permettant d'appliquer l'ACP au cas de données situées sur une variété non linéaire est d'effectuer des approximations locales : on calcule une ACP pour un groupe de points proches les uns des autres. Cette approche pose le problème de la définition des voisinages et du traitement des points nouveaux rencontrés loin des exemples connus.

Une autre approche, formalisée par B. Schölkopf en 1998 Schölkopf et al. (1999), utilise le *kernel trick* pour rendre non linéaire l'ACP traditionnelle. En effet, le calcul de l'ACP ne fait intervenir que des produits scalaires entre les points (pour le calcul de la matrice de covariance) et ne considère jamais les coordonnées d'un point isolé. Si l'on remplace le produit scalaire par un noyau, on calcule donc les composantes principales dans l'espace de caractéristiques H, et on peut ainsi accéder à des corrélations d'ordre supérieur entre les variables observées. Remarquons que l'on peut calculer la projection d'un point ne faisant pas partie de l'ensemble d'apprentissage, ce qui n'est pas le cas de toutes les méthodes de réduction de dimension non linéaires.

Isometric feature mapping (Isomap)

Isomap (Tenenbaum et al., 2000) est une méthode de réduction de dimension qui, comme MDS, part de la connaissance de la matrice des distances entre les paires de points. Le but est cette fois de trouver une variété (non linéaire) contenant les données. On exploite le fait que pour des points proches, la distance euclidienne est une bonne approximation de la distance géodésique sur la variété. On construit un graphe reliant chaque point à ses k plus proches voisins. Les longueurs des géodésiques sont alors estimées en cherchant la longueur du plus court chemin entre deux points dans le graphe. On peut alors appliquer MDS aux distances obtenues afin d'obtenir un positionnement des points dans un espace de dimension réduite.

Locally Linear Embedding (LLE)

LLE (*locally linear embedding*, ou plongement localement linéaire) (Roweis et Saul, 2000)

a été présenté en même temps qu'Isomap et aborde le même problème par une voie différente. Chaque point est ici caractérisé par sa reconstruction à partir de ses plus proches voisins. LLE construit une projection vers un espace linéaire de faible dimension préservant le voisinage.

Segmentation spectrale (spectral clustering)
La segmentation spectrale (*spectral clustering*) (Weiss, 1999; Ng et al., 2002) est une technique de réduction de dimension couplée à une segmentation. Le but est de regrouper les données de chaque segment (*cluster*) sur une sous-variété linéaire séparée de faible dimension.

Méthodes supervisées (S-Isomap)
Lorsque des informations sur les classes présentes dans les données sont disponibles (classification supervisée), il est possible d'en tenir compte lors de la construction de la matrice de distances utilisée par les méthodes de réduction de dimension (Vlachos et al., 2002; Geng et al., 2005). Cette technique permet à peu de frais d'améliorer la précision du résultat. Les auteurs proposent de l'utiliser pour construire un classificateur, qui semble obtenir des résultats corrects sur les données testées (benchmarks académiques).

3 Liste des problèmes résiduels importants

3.1 Sélection de variables

Variabilité du sous-ensemble de variable sélectionnées
Beaucoup de méthodes de sélection de variables sont sensibles à des petites perturbations des conditions expérimentales. Si les données ont des variables redondantes, différents sous-ensembles de variables avec le même pouvoir prédictif peuvent être obtenus en fonction des conditions initiales de l'algorithme d'apprentissage : la suppression ou l'ajout de quelques variables ou d'exemples d'apprentissage, ou l'addition de bruit. Cette variabilité est indésirable parce que (i) la variance est souvent le symptôme "d'un mauvais" modèle qui ne généralise pas bien ; (ii) les résultats ne sont pas reproductibles ; et (iii) un sous-ensemble de variables ne sera pas représentatif du problème. Une méthode possible pour stabiliser la sélection de variables consiste à utiliser des techniques de “bootstraps”. Le processus de sélection de variables est répété avec les sous-échantillons des données d'apprentissage. L'union des sous-ensembles de variables choisies dans les divers “bootstraps” est prise comme le sous-ensemble "stable" final. Ce sous-ensemble commun peut être au moins aussi pertinent que le meilleur sous-ensemble des “bootstraps”. L'analyse du comportement des variables à travers les divers “bootstraps” pourra aussi fournir une nouvelle compréhension du problème.

Procédure de recherche Forward vs. Backward
Le choix de la procédure de recherche est un sujet encore ouvert. Il est souvent dit que la procédure de recherche Forward est, de point de vue calculatoire, plus efficace que la procédure Backward pour produire les sous-ensembles de variables pertinentes. Cependant, la recherche par la procédure Forward ne tient pas compte du contexte d'autres variables non incluses encore. Le problème de l'importance mutuelle est absent dans cette procédure de recherche.

Sélection des exemples
Les problèmes duels de sélection/construction de variables sont ceux de sélection/construction de forme (exemple/observation). La symétrie des deux problèmes montre que certains algorithmes de sélection de variables peuvent s'appliquer aussi au choix d'exemples pour les méthodes à noyaux par exemple. La similitude et la complémentarité des deux problèmes sont évidentes. Particulièrement, des exemples mal étiquetés peuvent inciter le choix de fausses variables pertinentes. Au contraire, si l'étiquetage est fortement fiable, le choix de fausses variables pertinentes peut être évité en se concentrant sur les exemples du voisinage de la frontière de décision.

Problème inverse (causalité)
Dans certains domaines d'application, particulièrement dans la bio-informatique, la sélection d'un sous-ensemble de variables pertinentes En diagnostic, par exemple, il est important d'identifier les facteurs qui ont déclenché une maladie particulière ou démêler la chaîne d'événements des causes aux symptômes. En effet, le problème de la causalité représente une tâche plus stimulante que la juste sélection de variables pertinentes. Au coeur de ce problème est la distinction entre la corrélation et la causalité. Les données disponibles pour les chercheurs en apprentissage artificiel et en statistique nous permettent seulement d'observer des corrélations. Par contre le problème de causalité n'est que rarement exploré.

3.2 Extraction de traits

Les méthodes de réduction de dimension LLE et Isomap mentionnées ci-dessus ont de fort liens avec l'analyse en composantes principales non-linéaire (kernel PCA) (Bengio et al., 2004a; Burges, 2005). Isomap et LLE ne permettent pas le calcul efficace de la projection d'un nouveau point, ne faisant pas partie de l'ensemble d'apprentissage. On peut contourner le problème en estimant une application de l'espace d'origine vers l'espace réduit, par exemple à l'aide d'un réseau connexionniste. Cette approche est toutefois coûteuse et peu précise. L'interprétation de la réduction de dimension comme un apprentissage des fonctions propres d'un opérateur donné par un noyau dépendant des données permet une extension naturelle aux points hors échantillon (Bengio et al., 2004b; Paiement, 2003). Cette interprétation des algorithmes en termes de noyaux est aussi approfondie dans Ham et al. (2004).

Notons aussi que les méthodes présentées sont susceptibles de mal se comporter en présence de bruit sur les données. Elles sont aussi sensibles à l'accroissement du nombre de dimension de l'espace initial (elles n'évitent donc pas le curse of dimensionality) car elles reposent en dernier ressort sur une recherche des plus proches voisins dans cet espace (Bengio et al., 2005).

Dans le cadre de la première phase du projet, nous proposons d'étudier les applications des algorithmes de réduction de dimension présentés au traitement de données textuelles, et de comparer leurs propriétés à celle des méthodes connexionnistes.

Références

Albatineh, A. N., M. Niewiadomska-Bugaj, et D. Mihalko (2006). On similarity indices and correction for chance agreement. *Journal of Classification 23*(2), 301–313.

Almuallim, H. (1994). Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence 69*, 279–306.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks 5*(4), 537–550.

Ben-Hur, A., A. Elisseff, et I. Guyon (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing 7*, pp. 6–17.

Bengio, Y., O. Delalleau, et N. Le Roux (2005). The curse of dimensionality for local kernel machines. Technical Report 1258.

Bengio, Y., O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, et M. Ouimet (2004a). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Comp. 16*(10), 2197–2219.

Bengio, Y., J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, et M. Ouimet (2004b). Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In S. Thrun, L. Saul, et B. Scholkopf (Eds.), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Bennani, Y. (2001). *Systèmes d'apprentissage connexionnistes : sélection de variables*, Volume 15(3-4) of *Revue d'Intelligence Artiticielle*. Paris, France : Hermes Science Publications.

Bennani, Y. (2006). *Apprentissage Connexionniste*. Editions Hermès Science.

Bennani, Y. et F. Bossaert (1995). A neural network based variable selector. In C. H. Dagli, M. Akay, C. L. Chen, B. R. Fernandez, et J. Ghosh (Eds.), *ANNIE'95*, Volume 5, St. Louis, Missouri, USA, pp. 425–430. ASME Press.

Boser, B. E., I. M. Guyon, et V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *COLT '92 : Proceedings of the fifth annual workshop on Computational learning theory*, New York, NY, USA, pp. 144–152. ACM Press.

Burges, C. J. C. (2005). *Geometric methods for feature extraction and dimensional reduction - a guided tour*, pp. 59–92. Springer.

Cakmakov, D. et Y. Bennani (2002). *Feature Selection for Pattern Recognition*. Informa Press, Ed.

Charon, I., L. Denoeud, A. Guenoche, et O. Hudry (2006). Maximum transfer distance between partitions. *Journal of Classification 23*(1), 103–121.

Cibas, T., F. Fogelman, P. Gallinari, et S. Raudys (1994). Variable selection with optimal cell damage. In *ICANN'94*, Volume 1, pp. 727–730.

Dash, M., K. Choi, P. Scheuermann, et H. Liu (2002). Feature selection for clustering - a filter solution. In *ICDM*, pp. 115–122. IEEE Computer Society.

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 1*(2), 224–227.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*(39), 1–38.

Dorizzi, B., G. Pellieux, F. Jacquet, T. Czernikov, et A. Munoz (1996). Variable selection using generalized rbf networks : Application to forecast french t-bonds. In *IEEE-IMACS'96, Lille*.

Dy, J. G. et C. E. Brodley (2000). Feature Subset Selection and Order Identification for Unsu-

pervised Learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML'2000)*, Stanford University, CA.

Dy, J. G. et C. E. Brodley (2004). Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research 5*, 845–889.

Fowlkes, E. B. et C. L. Mallows (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association 78*(383), 553–569.

Fred, A. et A. Jain (2002). Evidence accumulation clustering based on the k-means algorithm. In *Proceedings of the International Workshops on Structural and Syntactic Pattern Recognition (SSPR)*.

Fred, A. et A. Jain (2005). Combining Multiple Clustering Using Evidence Accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*(6), 835–850.

Geng, X., D.-C. Zhan, et Z.-H. Zhou (2005). Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B 35*(6), 1098–1107.

Guérif, S. et Y. Bennani (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*, pp. 28–33.

Guyon, I., S. Gunn, M. Nikravesh, et L. Zadeh (2006). *Feature Extraction, Foundations and Applications, Editors*. Series Studies in Fuzziness and Soft Computing, Physica-Verlag. Springer.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2001). On clustering validation techniques. *Intelligent Information Systems Journal 17*(2-3), 107–145.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2002a). Cluster validity methods : Part i. *SIGMOD Record*.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2002b). Cluster validity methods : Part ii. *SIGMOD Record*.

Ham, J., D. D. Lee, S. Mika, et B. Scholkopf (2004). A kernel view of the dimensionality reduction of manifolds. In C. E. Brodley (Ed.), *ICML*. ACM.

Hassibi, B. et D. Stork (1993). Second order derivatives for networks pruning : Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*, pp. 164–171. Morgan Kaufmann Publishers.

He, X., D. Cai, et P. Niyogi (2006). Laplacian score for feature selection. In Y. Weiss, B. Schölkopf, et J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, pp. 507–514. Cambridge, MA : MIT Press.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Jain, A. K. et R. C. Dubes (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

Jouve, B., P. Kuntz, et F. Velin (2001). Extraction de structures macroscopiques dans des grands graphes par une approche spectrale. *Extraction des Connaissances et Apprentissage 1*(4).

Kira, K. et L. Rendell (1992). A practical approach to feature selection in machine learning. In *Proceedings of International Conference on Machine Learning*, pp. 249–256.

Kleinberg, J. (2002). An impossibility theorem for clustering. In *Proceedings of the 16th conference on Neural Information Processing Systems*.

Kohonen, T. (1995,1997,2001). *Self-Organizing Maps* (Third Extended Edition ed.), Volume 30 of *Springer Series in Information Sciences*. Berlin, Heidelberg, New York : Springer.

Kraskov, A., H. Stögbauer, et P. Grassberger (2004). Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys 69*(6 Pt 2).

Law, M. H. C., M. A. T. Figueiredo, et A. K. Jain (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*(9), 1154–1166.

Le Cun, Y., J. Denker, et S. Solla (1990). Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, pp. 598–605. Morgan Kaufmann Publishers.

Leray, P. et P. Gallinari (2001). *De l'utilisation d'OBD pour la sélection de variables dans les perceptrons multicouches*, pp. 373–391.

Li, T. (2006). A Unified View on Clustering Binary Data. *Machine Learning 62*(3), 199–215.

Liu, H. et H. Motoda (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

Lourenço, F., V. Lobo, et F. Bação (2004). Binary-based similarity measures for categorical data and their application in self-organizing maps.

M., C., G. B., G. Y., M. M., et M. C. (1995). Neural modeling for time sseries : A statistical stepwise method for weight elimination. *IEEE Trans. on Neural Networks 6*(6).

MacKay, D. (1994). *Bayesian methods for backpropagation networks*, Chapter 6. New York, USA : Springer-Verlag.

Meilă, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf et M. K. Warmuth (Eds.), *COLT*, Volume 2777 of *Lecture Notes in Computer Science*, pp. 173–187. Springer.

Meilă, M. (2005). Comparing clusterings : an axiomatic view. In L. D. Raedt et S. Wrobel (Eds.), *ICML*, pp. 577–584. ACM.

Meilă, M. (2006). Comparing clusterings - an information based distance. *in print*.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis 98*(5), 873–895.

Mitra, P., C. Murthy, et S. Pal (2002). Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(4).

Moody, J. (1994). Prediction risk and architecture selection for neural networks. In V. Cherkassky, J. Friedmann, et H. Wechsler (Eds.), *From Statistics to Neural Networks - Theory and Pattern Recognition Application*.

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.

Neal, R. (1994). *Bayesian learning for neural networks*. Ph. D. thesis, University of Toronto, Canada.

Ng, A., M. Jordan, et Y. Weiss (2002). On spectral clustering : Analysis and an algorithm. In *NIPS 14 - Advances in Neural Information Processing Systems*.

Paiement, J.-F. (2003). Généralisation d'algorithmes de réduction de dimension. Master's thesis, Université de Montréal.

Pal, S. K., R. K. De, et J. Basak (2000). Unsupervised Feature Evaluation : A Neuro-Fuzzy Approach. *IEEE Transactions on Neural Networks 11*(2), 366–376.

Pedersen, M., L. Hansen, et J. Larsen (1996). Pruning with generalization based weight saliencies : γobd, γobs. In *Advances in Neural Information Processing Systems 8*. Morgan Kaufmann Publishers.

Raftery, A. et N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Assocation 101*(473), 168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Refenes, A.-P. N. et A. Zapranis (1999). Neural model identification, variable selection and model adequacy. *Journal of Forecasting 18*(5), 299–332.

Rossi, F. (1996). Attribute suppression with multi-layer perceptron. In *Proceedings of IEEEI-MACS'96, Lille, France.*

Roux, M. (1985). *Algorithmes de classification*. Paris : Masson.

Roweis, S. T. et L. K. Saul (2000). Nonlinear Dimensionality Reduction by Local Linear Embedding. *Science 290*, 2323–2326.

Ruck, D. W., S. K. Rogers, et M. Kabrisky (1990). Feature selection using a multilayer perceptron. *International Journal on Neural Network Computing 2*(2), 40–48.

Schölkopf, B., A. J. Smola, et K.-R. Müller (1999). Kernel principal component analysis. *Advances in kernel methods : support vector learning*, 327–352.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Sorg-Madsen, N., C. Thomsen, et J. Peña (2003). Unsupervised feature subset selection. In *Proceedings of the Workshop on Probabilistic Graphical Models for Classification, ECML/PKDD*, pp. 71–82.

Tenenbaum, J., V. de Silva, et J. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 2319–2323.

Tresp, V., R. Neuneier, et H. G. Zimmermann (1996). Early brain damage. In M. Mozer, M. Jordan, et T. Petsche (Eds.), *Advances in Neural Information Processing Systems (NIPS 1996)*, pp. 669–675. MIT Press.

Vesanto, J. et J. Ahola (1999). Hunting for Correlations in Data Using the Self-Organizing Map. In H. Bothe, E. Oja, E. Massad, et C. Haefke (Eds.), *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, pp. 279–285. ICSC Academic Press.

Vesanto, J. et E. Alhoniemi (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks 11*(3), 586–600.

Vlachos, M., C. Domeniconi, D. Gunopulos, G. Kollios, et N. Koudas (2002). Non-linear

dimensionality reduction techniques for classification and visualization. In *Proceeding of the International Conference on Knowledge Discovery and Data mining (KDD)*, pp. 645–651. ACM.

Wallace, D. L. (1983). A Method for Comparing Two Hierarchical Clusterings : Comment. *Journal of the American Statistical Association 78*(383), 569–576.

Weiss, Y. (1999). Segmentation using eigenvectors : A unifying view. In *ICCV'99 : Proceedings of the International Conference on Computer Vision*, Volume 2, Washington, DC, USA, pp. 975. IEEE Computer Society.

Yacoub, M. et Y. Bennani (1997). HVS : A heuristics for variables selection in multilayer neural network classifiers. In C. H. Dagli, M. Akay, C. L. Chen, B. R. Fernandez, et J. Ghosh (Eds.), *ANNIE'97*, Volume 7, St. Louis, Missouri, USA, pp. 527–532. ASME Press.

Summary

Since several years, the volume of available data does not stop growing; whereas at the beginning of the eighties the amount of databases was measured in mega-bytes, it is expressed today in tera-bytes and sometimes even in peta-bytes. The number of variables and the number of examples can take very high values, and that can causes some problems for data exploration and analysis process. Thus, the development of processing tools adapted to these massive databases is a major stake for data mining. The dimensionality reduction makes it possible and facilitates visualization and understanding of the data, reduce the storage space and the running time, and finally identify the relevant features. In this article, we present a review about dimensionality reduction techniques primarily based on variables selection in supervised and unsupervised learning, and some geometrical methods for non linear dimensionality reduction.