Étude comparative par simulation du comportement de méthodes d'analyse discriminante, de classification et de réseaux de neurones*

Nadia Ghazzali† , Marc Parizeau†† et Josianne DeBlois‡

10 mars 1998

[1]

Résumé
Cette étude s'intéresse à l'analyse du comportement de méthodes classiques d'analyse discriminante et de classification et de méthodes neuronales lorsqu'il existe dans les données, une structure de classes bien spécifique obéissant à une certaine distribution connue. Cette analyse a été réalisée sur des données simulées selon un plan d'expérience que nous avons mis en œuvre.

[1]* Cet article est paru dans la revue *Modulad* en janvier 1998, No. 20, pages 13-34. Il a été réalisé grâce aux octrois du CRSNG OGPO155639 accordé à N. Ghazzali et OGP0155389 accordé à M. Parizeau.

†N. Ghazzali est au Département de mathématiques et de statistique, Université Laval, Québec (PQ), Canada, G1K 7P4. ghazzali@mat.ulaval.ca

††M. Parizeau est au Département de génie électrique et de génie informatique, Université Laval, Québec (PQ), Canada, G1K 7P4.parizeau@gel.ulaval.ca

‡J. DeBlois était au Département de mathématiques et de statistique, Université Laval. Elle travaille actuellement au ministère des Ressources naturelles du Québec. Québec. Canada.

1 Introduction

Pour répondre au problème épineux que pose la détermination du nombre de classes en classification automatique, Milligan & Cooper (Milligan & Cooper 1985) ont examiné trente règles d'arrêt sur des données simulées où le nombre de classes est connu d'avance. Les critères considérés d'agrégation des classes s'apparentent à la classification hiérarchique et sont le saut minimal, le saut maximal, le saut moyen et le critère de Ward (voir par exemple, Seber 1984).

Nous nous inspirons de cette étude pour analyser le comportement de différentes méthodes d'analyse discriminante et de classification hiérarchique et non hiérarchique lorsqu'il existe dans les données, une structure de classes bien spécifique possèdant une distribution connue.

Pour réaliser cette étude, nous proposons le schéma expérimental suivant qui consiste à considérer les six facteurs suivants : 1) le nombre de classes; 2) la dimension de l'espace de représentation des individus; 3) la taille des classes; 4) la dispersion spatiale à l'intérieur des classes; 5) le recouvrement entre classes et 6) l'orientation des classes. Nous supposons que les classes suivent des distributions normales; la variable réponse étant le taux de mauvaise classification.

La motivation de F. Rosenblatt (Rosenblatt 1958 et 1962), lorsqu'il a proposé son tout premier réseau de neurones artificiel, était essentiellement la compréhension et l'organisation du cerveau humain. Plusieurs années plus tard, les réseaux de neurones artificiels représentent toujours une simplification bien grossière du fonctionnement du cerveau humain. Elles connaissent cependant un grand essor dans des domaines diversifiés à cause entre autres de leur capacité d'apprentissage, comme par exemple le perceptron multicouche (Rumelhart & McClelland 1986) et leurs capacités de modélisation et de réduction, comme par exemple le réseau de Kohonen (Kohonen, 1982). Nous nous intéressons plus particulièrement, d'une part, au perceptron multicouche qui est vu comme une méthode de discrimination entre classes et, d'autre part, au réseau de Kohonen qui est considéré comme une méthode de classification. Le but est d'évaluer, lorsque c'est pertinent, la performance quant au taux de mauvaise classification des méthodes classiques et des deux méthodes neuronales sur les mêmes données.

2 Protocole expérimental

Comme nous l'avons mentionné en introduction, le schéma expérimental proposé consiste à considérer les facteurs suivants : 1) le nombre de classes; 2) la dimension de l'espace de représentation des individus; 3) la taille des classes; 4) la dispersion spatiale à l'intérieur des classes; 5) le recouvrement entre classes et 6) l'orientation des classes.

Nous fixons d'emblée le premier facteur à deux classes et nous supposons de plus que ces dernières sont distribuées selon une loi normale, situation très populaire et à la base de plusieurs travaux. Les autres facteurs sont considérés comme suit : Le deuxième facteur qui décrit le nombre de variables possède trois traitements : 2, 4 ou 6. Nos individus sont donc représentés dans un espace à 2, 4 ou 6 dimensions. Pour le troisième facteur décrivant le nombre d'individus par classe, trois traitements sont

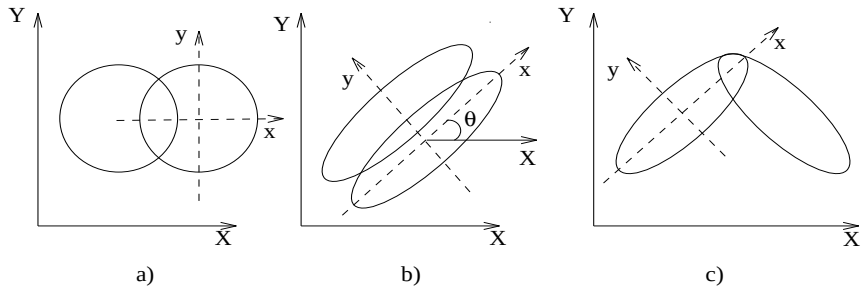


Figure 1: *Orientation des classes : a)* $\theta_1 = \theta_2 = 0°$*; b)* $\theta_1 = \theta_2 = 45°$*; c)* $\theta_1 = 45°$, $\theta_2 = -45°$

à l'étude. Le premier requiert un nombre égal d'observations, noté 50%-50%, dans chaque classe. Le deuxième impose qu'une classe contienne 20% des observations et l'autre 80%. Ce traitement sera noté 20%-80%. Le troisième traitement exige qu'une classe contienne 40% des observations et l'autre 60%. Nous le noterons 40%-60%. Le nombre total d'individus considéré est de 1000. Le quatrième facteur relatif à la dispersion spatiale à l'intérieur des classes possède deux traitements dont le premier exprime que les dispersions sont égales et le deuxième traite le cas où elles sont différentes. Quant au recouvrement entre classes, nous considérons le cas où il existe un léger recouvrement entre les classes soit 5% , et celui où le recouvrement est plus important qui a été fixé à 15%. Finalement, le dernier facteur relatif à l'orientation des nuages de points des classes dans le plan possède trois traitements.

La figure 1 illustre ces traitements dans le cas où les données sont décrites par deux variables et s'interprète comme suit. Les axes X et Y en trait plein représentent les deux variables. Les axes x et y en trait pointillé correspondent aux axes principaux des nuages de points des deux classes. Ces nuages de points sont des cercles lorsque les variances à l'intérieur des classes sont égales et des ellipses lorsque ces variances sont différentes. L'angle formé entre les axes (X, Y) et (x, y) est noté θ_1 pour la première classe et θ_2 pour la deuxième classe. La situation a) illustre le premier traitement du facteur orientation des classes où $\theta_1 = \theta_2 = 0°$. Les nuages de points n'ont donc pas subi de rotation. Le deuxième traitement est schématisé par la situation b) où les deux classes sont orientées de la même façon à un angle de 45° des axes (X, Y). De sorte que $\theta_1 = \theta_2 = 45°$ tandis que c) représente la situation où les deux classes sont orientées différemment. Auquel cas $\theta_1 = 45°$ et $\theta_2 = -45°$.

De façon plus spécifique, relativement au quatrième facteur, lorsque la dispersion spatiale est égale à l'intérieur des classes, toutes les variables ont la même variance que nous avons fixée à 1. Lorsque cette dispersion est différente, la première variable a une variance fixée à 4 et les variances de toutes les autres variables sont les mêmes égales à 1. De plus, le recouvrement a été fait seulement sur la première variable. De sorte que, par rapport au facteur orientation des classes, nous avons $3 \times 3 \times 2 \times 2 = 36$ combinaisons pour le premier traitement et $3 \times 3 \times 2 = 18$ pour chacun des deux autres traitements de ce même facteur. Nous avons donc 72 combinaisons possibles en considérant les six facteurs à l'étude. De plus, nous considérons dix répliques pour chacune de ces 72 combinaisons. Nous générons 10 fichiers de 1000 individus provenant d'une loi multinormale $N_6(\mu, \Sigma)$ où μ est le vecteur nul et Σ est la matrice identité.

Les six variables sont donc centrées en zéro, leur variance est fixée à 1 et leurs covariances sont nulles. Pour diminuer l'effet du générateur de nombres aléatoires, tous les fichiers considérés dans cette étude sont crées à partir de ces dix répliques où il suffit d'appliquer les transformations appropriées pour obtenir la combinaison désirée des facteurs.

Ainsi, pour les deux premiers traitements du facteur relatif au nombre de variables, nous considérons respectivement seulement les deux et les quatre premières colonnes des fichiers.

En ce qui concerne la taille des classes, nous divisons les fichiers en deux parties où la première correspond à la première classe et le reste des observations formant la deuxième classe. Par conséquent selon les traitements considérés, la première classe est constituée respectivement des 500, 200 et 400 premières observations.

Par rapport à la dispersion à l'intérieur des classes, lorsque la variance de la première variable doit être différente des autres, les données concernées sont multipliées par l'écart-type désiré, soit 2.

Quant à l'orientation des classes dans le plan, les données subissent soit une rotation de 45° soit de $-45°$, selon le cas, à l'aide de la formule suivante :

$$\begin{array}{rcl} x & = & X\ \cos(\theta) - Y\ \sin(\theta) \\ y & = & Y\ \cos(\theta) + X\ \sin(\theta) \end{array} \quad \text{avec} \quad \theta = 45° \text{ ou } -45°$$

Signalons enfin que la première classe est toujours située à l'origine. La deuxième classe est placée en fonction du taux de recouvrement e_{th} désiré telle que :

$$e_{\text{th}} = \pi_1 \Phi \left[\frac{\log(\pi_2/\pi_1) - \frac{1}{2}\Delta^2}{\Delta} \right] + \pi_2 \Phi \left[\frac{-\log(\pi_2/\pi_1) - \frac{1}{2}\Delta^2}{\Delta} \right]$$

Cette formule est très classique et représente le taux de mauvaise classification dans le cadre d'une analyse discriminante linéaire de Fisher où les deux classes suivent des distributions normales de même variance (pour plus de détails, voir par exemple Seber 1984). La quantité e_{th} est fixée à 5% ou 15% selon le cas (niveaux du facteur recouvrement entre classes), π_1 et π_2 sont les proportions connues dans les deux classes et Δ^2 est la distance de Mahalanobis entre les deux classes.

3 Méthodes supervisées

Nous entendons ici par méthodes supervisées des méthodes qui utilisent de l'information relative à la structure des classes. En ce sens, une partie des données est dédiée à l'apprentissage (on parle de données d'entraînement) tandis que l'autre partie constitue les données à tester. Nous étudions plus particulièrement une classique et une neuronale. Il s'agit, respectivement, de l'analyse discriminante (AD) et du perceptron multicouche (PMC).

Dans cette section, nous ne décrivons pas l'analyse discriminante puisque la bibliographie est riche et abondante dans ce domaine (comme références de base, citons par exemple Lachenbruch 1975 et Seber 1984). Spécifions toutefois que nous nous plaçons dans le cadre de deux classes distibuées selon une loi normale $N_p(\mu_i, \Sigma_i)$, $(i = 1, 2)$, p est la dimension de l'espace de représentation des données $(p = 2, 4, 6)$, μ_1 et μ_2 sont les moyennes des deux classes et Σ_1 et Σ_2 sont les matrices de variances

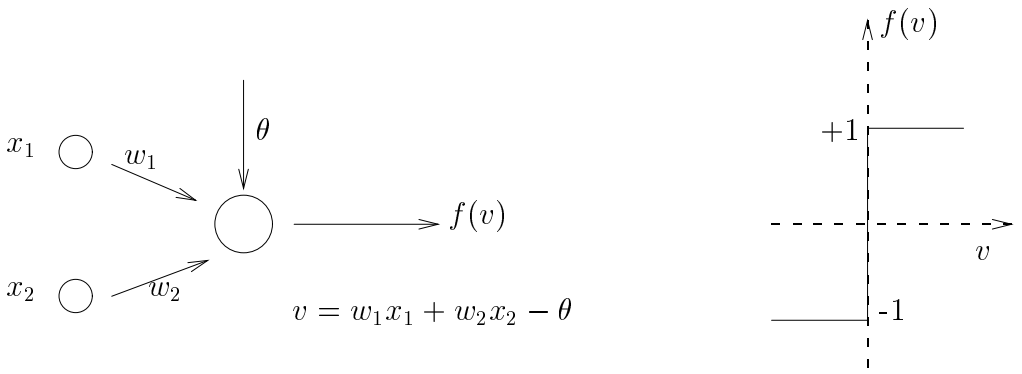


Figure 2: *Réseau perceptron simple (PS)*

correspondantes avec, dans la plupart des cas, $\Sigma_1 = \Sigma_2 = \Sigma$. La règle d'affectation d'une observation x à une classe C_i $(i = 1, 2)$ consiste à minimiser la probabilité de commettre une erreur de classification c'est-à-dire qu'on assigne x à la classe C_1 si

$$\lambda^t \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right] > \log(\pi_2/\pi_1)$$

où $\lambda = \Sigma^{-1}(\mu_1 - \mu_2)$ et $\pi_1, \pi_2, \mu_1, \mu_2$ sont définis précédemment. À noter que π_1 et π_2 peuvent être estimées par les proportions correspondantes au niveau de l'échantillon. La fonction discriminante linéaire lorsque les distributions sont parfaitement connues est donnée par :

$$D(x) = \lambda^t \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right] \tag{1}$$

Comme les résultats de l'analyse discriminante seront comparés à ceux du perceptron multicouche, nous consacrons la suite de cette section à une description succinte du réseau perceptron multicouche. Pour une description détaillée, le lecteur pourra consulter par exemple (Haykin, 1994).

3.1 Réseau perceptron multicouche (PMC)

Les réseaux de neurones sont caractérisés entre autres par leur type d'apprentissage. Le perceptron multicouche se distingue par un apprentissage supervisé qui consite à faire apprendre au réseau une tâche sous la supervision d'un professeur. Le réseau tente alors de s'orienter en suivant les instructions du professeur. Ces dernières consistent à fournir une nouvelle donnée (i.e. un vecteur d'entrées) au réseau tout en lui spécifiant le résultat attendu (i.e. vecteur de sorties désiré) pour cette donnée particulière.

En 1958, F. Rosenblatt (Rosenblatt 1958) a proposé le tout premier réseau appelé perceptron simple (PS) selon un modèle neuronal à un seul neurone tel qu' illustré à la figure 2. Cette dernière illustre le cas particulier d'un vecteur d'entrée $x = (x_i), i = 1, 2$, à deux dimensions. Les valeurs w_1 et w_2 sont respectivement les coefficients synaptiques des connexions reliant le neurone aux entrées, alors que le paramètre θ est le seuil d'activation (ou biais) associé au neurone. La fonction de seuil illustrée du côté droit de la figure 2 définit la fonction d'activation proposée à l'origine par

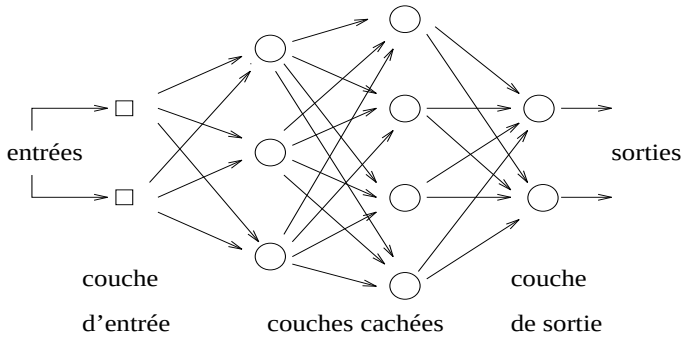


Figure 3: *Réseau perceptron multicouche (PMC)*

F. Rosenblatt. Cette dernière permet d'activer le neurone et d'émettre la valeur +1 comme sortie lorsque la somme pondérée v de ses entrées est supérieure au biais θ, et d'émettre la valeur -1 comme sortie dans le cas contraire. On peut également noter :

$$v = \sum_{i=1}^{p} w_i x_i - \theta = w^t x - \theta \tag{2}$$

où w_i est le poids associé à x_i.

L'algorithme d'apprentissage (Rosenblatt 1962) du perceptron simple consiste à choisir la frontière de décision qui minimise le nombre d'individus mal classés. Il s'agit, concrètement, d'ajuster les poids w_i et le biais θ de sorte que la diférence entre la sortie du réseau et celle désirée par le professeur soit minimum.

Ainsi, la règle de décision du perceptron simple est d'assigner l'individu présenté à la classe C_1 si la sortie du perceptron est +1 et à la classe C_2 si elle égale -1.

Nous pouvons tout de suite faire remarquer que le perceptron ainsi défini est en étroite relation avec l'analyse discriminante linéaire dans le cas où les populations sont normales puisque l'équation 1 correspond à l'équation 2. En effet, l'équation 1 peut être écrite comme suit : $D(x) = w^t x - \theta$ où $w = \lambda$ et $\theta = \frac{1}{2}\lambda^t(\mu_1 + \mu_2)$.

Ce lien avec l'analyse discriminante linéaire étant fait, le perceptron, en raison même de sa simplicité, est limité puisqu'il ne peut émettre que les valeurs +1 ou -1. Ce qui restreint le nombre de choix de la sortie désirée de la part du professeur. Il ne peut donc discriminer qu'entre deux classes dont la frontière de décison est une droite. La présence d'un seul neurone à l'intérieur du réseau semble par conséquent insuffisante d'où la mise au point de réseaux perceptron multicouche (voir figure 3).

Ce type de réseaux contient, entre ses couches d'entrées et de sorties, une ou plusieurs couches de neurones, appelées couches cachées, contenant un nombre arbitraire de neurones. Le nombre de neurones dans la couche d'entrées correspond à la dimension de l'espace dans lequel les individus sont représentés tandis que le nombre de neurones dans la couche de sortie correspond au nombre de classes présentes dans les données. À l'exception de la couche d'entrées, chaque neurone dans une couche prend la sortie des neurones de la couche précédente auxquels il est connecté comme vecteur d'entrées.

Les réseaux PMC nécessitent une fonction d'activation dérivable. Nous utilisons la fonction sigmoïde de pente 1 appelée également fonction logistique de la forme suivante : $f(v) = \dfrac{1}{1+e^{-v}}, \quad v \in I\!R$.

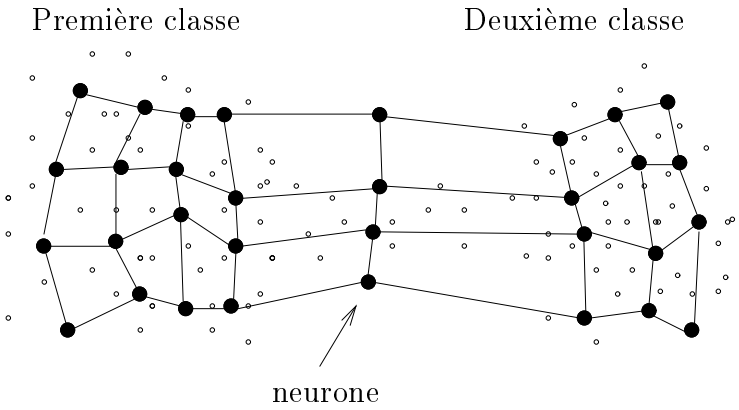


Figure 4: *Exemple de configuration du réseau de Kohonen*

Chaque neurone appartenant à une couche cachée ou à la couche de sortie du réseau crée un hyperplan et l'intersection de ces hyperplans détermine les régions de décision définissant les classes présentes dans les données.

Les réseaux PMC ont été longtemps inexploités à cause essentiellement de l'absence d'un algorithme d'apprentissage adéquat pour ajuster les nombreux poids associés au réseau. Il a fallu attendre les travaux de Rumelhart et McClelland (Rumelhart & McClelland, 1986) en 1986 qui ont mis en œuvre l'algorithme de rétropropagation des erreurs. Ce dernier peut être employé pour n'importe quelle architecture de PMC.

4 Méthodes non supervisées

Les méthodes non supervisées considérées dans ce travail sont essentiellemnt les méthodes de classification automatique et le réseau de Kohonen (Kohonen 1982 et 1990). À l'inverse des méthodes supervisées, elles n'exploitent aucune connaissance a priori sur les données, et de ce fait la structure des classes.

Plus particulièrement, nous considérons les méthodes classiques de classification hiérarchique à savoir le saut minimal, le saut maximal, le saut moyen, la méthode de Ward et celle de Lance et Williams aussi appelée méthode flexible (voir par exemple, Seber 1984). Nous étudions également la méthode des k-means (MacQueen 1967).

4.1 Réseau de Kohonen

En ce qui concerne le réseau de Kohonen, il se caractérise par un apprentissage non supervisé. En ce sens, il ne reçoit aucune aide ou indication et prend seul les décisions quant à son apprentissage.

Ce réseau possède une structure différente du PMC puisque les neurones ne sont plus disposés sur différentes couches mais plutôt sur une carte dite d'auto-organisation. Bien que la dimension de cette carte puisse être arbitraire, elle comporte généralement deux dimensions et est discrétisée en un nombre limité de positions dont chacune correspond à un neurone du réseau (voir figure 4). Dans notre application, nous considérons un réseau composé de 4×8 neurones. Un poids différent est associé

à chaque connexion, reliant les entrées aux neurones, et le vecteur des poids d'un neurone correspond à un point dans l'espace des entrées (à la figure 4, l'espace des entrées est à deux dimensions; les ∘ correspondent aux données d'entraînement et les • représentent les vecteurs de poids des neurones).

On peut donc associer à chaque neurone du réseau de Kohonen, une région de l'espace des entrées définie par un centre. En ce sens, ce type de réseau s'apparente à la méthode du k-means. L'objectif lors de l'apprentissage étant de modifier les poids des neurones (déplacer les centres) pour approximer le mieux possible les données d'entraînement.

L'algorithme d'apprentissage fonctionne dans un mode compétitif et utilise un concept de voisinage (représenté à la figure 4 par des segments de droite reliant les neurones). Une nouvelle donnée est d'abord choisie aléatoirement parmi l'ensemble des données d'entraînement. Celle-ci est ensuite comparée avec chaque neurone en calculant sa distance (par exemple euclidienne) avec le vecteur de poids correspondant. Le neurone gagnant est celui dont la distance est minimum. Ses poids de même que ceux de tous les autres neurones qui sont dans son voisinage sont ajustés de manière à réduire cette distance. L'apprentissage est ainsi répété en faisant décroître de façon monotone la taille du voisinage jusqu'à un nombre prédéterminé d'itérations. Pour plus de détails sur le fonctionnement exact de l'algorithme, le lecteur pourra se référer à (Kohonen, 1990).

5 Discussion des résultats

Le protocole expérimental présenté à la section §2 nous a permis de générer 72 fichiers dont chacun contient 1000 observations suivant une loi multinormale dont la moyenne est fonction du taux de recouvrement choisi et dont la variance est déterminée par le facteur dispersion spatiale à l'intérieur des classes. Le tableau 1 donne pour chacun de ces 72 fichiers ses caratéristiques en fonction des facteurs à l'étude, et ce pour 2, 4 ou 6 variables. Rappelons ici que nous analysons 10 répliques de chacun de ces fichiers et, par conséquent, nous disposons au total de 720 fichiers.

Pour une meilleure présentation des résultats obtenus, nous avons schématisé chacune des combinaisons des facteurs à l'étude comme suit, en rappelant toutefois que seul le cas de deux classes nous intéresse. Les nombres se trouvant à l'intérieur de chaque classe correspondent aux proportions associées à chacune d'elles. Lorsque la variance de la première variable est égale à celle des autres variables, les classes possèdent une forme circulaire. Une classe de forme allongée sur le premier axe indique par contre que la variance de la première variable est différente des autres. L'orientation des classes par rapport aux axes permet de distinguer les traitements utilisés pour ce facteur. Pour ce qui est du nombre de variables et du taux de recouvrement, les traitements utilisés sont toujours indiqués explicitement.

Prenons par exemple les situations illustrées à la figure 5. La situation (a) correspond à des classes bien balancées où les variances des variables sont toutes égales et les classes sont orientées avec un angle nul. La situation (b) concerne un cas très débalancé où la variance de la première variable est différente de celle des autres variables. La position des classes montre que les deux classes sont orientées à 45° par rapport aux axes.

Table 1: *Description des fichiers pour 2, 4 ou 6 variables*

Fichier	**Orientation**	**Taille**	**Var*(X)**	e_{th} (%)	e_{emp} (%)
1	$\theta_1 = 0°, \theta_2 = 0°$	50% - 50%	$\sigma_{\text{X}}^2 = 1$	5	4.98
2	$\theta_1 = 0°, \theta_2 = 0°$	50% - 50%	$\sigma_{\text{X}}^2 = 1$	15	14.92
3	$\theta_1 = 0°, \theta_2 = 0°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	5	4.98
4	$\theta_1 = 0°, \theta_2 = 0°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	15	14.92
5	$\theta_1 = 0°, \theta_2 = 0°$	20% - 80%	$\sigma_{\text{X}}^2 = 1$	5	4.85
6	$\theta_1 = 0°, \theta_2 = 0°$	20% - 80%	$\sigma_{\text{X}}^2 = 1$	15	14.7
7	$\theta_1 = 0°, \theta_2 = 0°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	5	4.86
8	$\theta_1 = 0°, \theta_2 = 0°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	15	14.7
9	$\theta_1 = 0°, \theta_2 = 0°$	40% - 60%	$\sigma_{\text{X}}^2 = 1$	5	4.97
10	$\theta_1 = 0°, \theta_2 = 0°$	40% - 60%	$\sigma_{\text{X}}^2 = 1$	15	14.9
11	$\theta_1 = 0°, \theta_2 = 0°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	5	4.97
12	$\theta_1 = 0°, \theta_2 = 0°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	15	14.9
13	$\theta_1 = 45°, \theta_2 = 45°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	5	5.15
14	$\theta_1 = 45°, \theta_2 = 45°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	15	15.22
15	$\theta_1 = 45°, \theta_2 = 45°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	5	5.05
16	$\theta_1 = 45°, \theta_2 = 45°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	15	14.97
17	$\theta_1 = 45°, \theta_2 = 45°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	5	5.18
18	$\theta_1 = 45°, \theta_2 = 45°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	15	15.28
19	$\theta_1 = 45°, \theta_2 = -45°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	5	4.99
20	$\theta_1 = 45°, \theta_2 = -45°$	50% - 50%	$\sigma_{\text{X}}^2 = 4$	15	14.94
21	$\theta_1 = 45°, \theta_2 = -45°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	5	4.84
22	$\theta_1 = 45°, \theta_2 = -45°$	20% - 80%	$\sigma_{\text{X}}^2 = 4$	15	14.68
23	$\theta_1 = 45°, \theta_2 = -45°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	5	4.99
24	$\theta_1 = 45°, \theta_2 = -45°$	40% - 60%	$\sigma_{\text{X}}^2 = 4$	15	14.94

$*$ indique que les variances des autres variables sont fixées à 1.
e_{th} est le recouvrement théorique.
e_{emp} est la moyenne des recouvrements empiriques pour deux variables.

Figure 5: *Exemples de schéma des deux classes.*

Finalement, le taux de mauvaise classification peut être exprimé de deux façons différentes, soit à l'aide du taux moyen e_{moy} et du taux pondéré e_{pd} définis comme suit :

$$e_{\text{moy}} = \frac{1}{2} \left[\frac{e_1}{n_1} + \frac{e_2}{n_2} \right] \tag{3}$$

où e_i représente le nombre d'individus mal classés dans la classe C_i et n_i correspond à la taille de la classe C_i, avec $n_1 + n_2 = n = 1000$ pour chacun de nos fichiers.

$$e_{\text{pd}} = \pi_1 \frac{e_1}{n_1} + \pi_2 \frac{e_2}{n_2} = \frac{e_1 + e_2}{n} \qquad \text{car} \quad \pi_i = \frac{n_i}{n} \quad i = 1, 2 \tag{4}$$

Ce dernier taux correspond donc au pourcentage du nombre d'individus total mal classés dans chaque fichier. On remarque facilement que les taux moyen et pondéré de mauvaise classification sont équivalents lorsque les classes sont équiprobables.

5.1 Analyse discriminante et PMC

Précisons tout d'abord que dans le cadre de l'analyse discriminante, nous avons quatre résultats différents et ceci pour chacun des 720 fichiers. En effet, nous avons considéré deux cas concernant les probabilités a priori, soit le cas où ces probabilités sont proportionnelles aux tailles des classes et le cas où elles sont égales, c'est-à-dire des probabilités a priori de 0.5 pour chacune des deux classes. De plus, pour chacun de ces cas, nous avons calculé les taux moyen et pondéré de mauvaise classification.

Mentionnons également que nous avons pris tour à tour chacune des dix répliques de chaque fichier comme fichier d'entraînement. Nous avons ainsi 100 taux de mauvaise classification pour chacune des 72 combinaisons différentes des facteurs étudiés et ce, pour chacun des quatre cas considérés, soit le taux pondéré ou moyen obtenu avec les probabilités a priori proportionnelles ou égales.

Le tableau 2 présente les résultats obtenus pour quelques fichiers dont les classes sont débalancées, ce qui nous permet de comparer les différents taux de mauvaise classification considérés. Les paramètres $\bar{x}$ et s représentent respectivement la moyenne et l'écart-type des 100 taux de mauvaise classification correspondant à la situation étudiée. Comme les taux pondéré et moyen sont équivalents dans le cas où les probabilités a priori sont égales (voir équations (3) et (4)), nous ne présentons que le taux pondéré associé à ce cas.

On peut premièrement remarquer que les taux pondérés sont moins importants dans le cas où les probabilités a priori sont proportionnelles que lorsque celles-ci sont considérées égales. Ceci est essentiellement dû au fait que, lorsque les probabilités

Table 2: *Quelques résultats de l'analyse discriminante.*

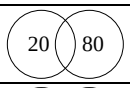
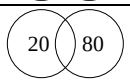
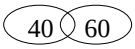
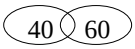
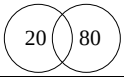
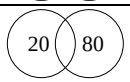
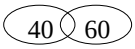
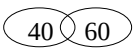
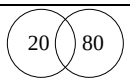
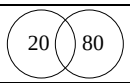
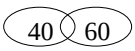
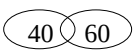
Situation étudiée	# de var.	Recouvrement (%)	Proportionnelles		Égales
			Pondéré	Moyen	Pondéré
	2	5	$\bar{x} = 4.9$ $s = 0.6$	$\bar{x} = 8.4$ $s = 1.0$	$\bar{x} = 6.7$ $s = 0.7$
	2	15	$\bar{x} = 14.8$ $s = 1.1$	$\bar{x} = 29.6$ $s = 2.3$	$\bar{x} = 21.7$ $s = 1.0$
	2	5	$\bar{x} = 5.0$ $s = 0.6$	$\bar{x} = 5.2$ $s = 0.7$	$\bar{x} = 5.1$ $s = 0.6$
	2	15	$\bar{x} = 15.4$ $s = 0.7$	$\bar{x} = 16.3$ $s = 0.9$	$\bar{x} = 15.8$ $s = 0.6$
	4	5	$\bar{x} = 4.9$ $s = 0.7$	$\bar{x} = 8.4$ $s = 1.1$	$\bar{x} = 6.7$ $s = 0.7$
	4	15	$\bar{x} = 14.8$ $s = 1.1$	$\bar{x} = 29.6$ $s = 2.3$	$\bar{x} = 21.8$ $s = 1.1$
	4	5	$\bar{x} = 5.0$ $s = 0.6$	$\bar{x} = 5.2$ $s = 0.7$	$\bar{x} = 5.1$ $s = 0.6$
	4	15	$\bar{x} = 15.4$ $s = 0.6$	$\bar{x} = 16.3$ $s = 0.8$	$\bar{x} = 15.9$ $s = 0.7$
	6	5	$\bar{x} = 4.9$ $s = 0.6$	$\bar{x} = 8.4$ $s = 1.1$	$\bar{x} = 6.7$ $s = 0.7$
	6	15	$\bar{x} = 14.9$ $s = 1.0$	$\bar{x} = 29.7$ $s = 2.1$	$\bar{x} = 21.8$ $s = 1.1$
	6	5	$\bar{x} = 5.0$ $s = 0.6$	$\bar{x} = 5.2$ $s = 0.7$	$\bar{x} = 5.1$ $s = 0.5$
	6	15	$\bar{x} = 15.3$ $s = 0.6$	$\bar{x} = 16.2$ $s = 0.9$	$\bar{x} = 15.8$ $s = 0.6$

Table 3: *Répartition des individus dans les classes en fonction des probabilités a priori (cas débalancé (20%-80%), variances égales, recouvrement 5%).*

	Prob. a priori égales		Prob. a priori proportionnelles	
	Classés dans la classe 1	Classés dans la classe 2	Classés dans la classe 1	Classés dans la classe 2
Provenant de la classe 1 (200)	189	11	176	24
Provenant de la classe 2 (800)	60	740	23	777
Taux pondéré	7.1%		4.7%	
Taux moyen	6.5%		7.4%	

a priori sont proportionnelles, la frontière de séparation est plus proche de la classe ayant le moins d'individus. Alors que cette frontière est pratiquement au milieu de l'intersection entre les deux classes lorsque les probabilités a priori sont égales.

On peut également remarquer que lorsque les probabilités a priori sont proportionnelles dans la situation 20%-80% (classes très débalancées), le taux moyen est presque le double du taux pondéré. Alors que cette différence devient moins importante lorsque les classes sont peu débalancées (situation 40%-60%). Ceci peut s'expliquer par la définition même des deux taux moyen et pondéré puisque le premier revient à prendre la moyenne des taux relatifs à chacune des deux classes alors que le deuxième prend en considération le poids associé à chacune des classes, et ceci est d'autant important lorsque les classes sont très débalancées.

Le taux pondéré avec les probabilités a priori proportionnelles aux tailles des classes représente plus adéquatement la structure de nos classes. C'est ce dernier qui servira dans le reste de la section à faire des comparaisons et à discuter des résultats obtenus.

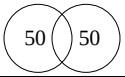
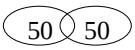
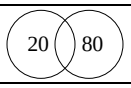
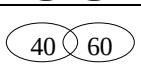
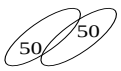
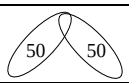
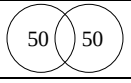
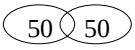
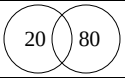
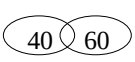
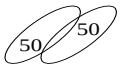
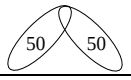
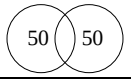
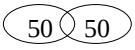
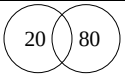
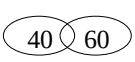
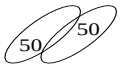
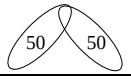
Le tableau 3 donne la répartition des individus lorsque les probabilités a priori sont considérées égales d'une part et proportionnelles à la taille des classes d'autre part. La relation entre les taux pondérés et moyens obtenus pour chacune des situations est comparable à celle constatée au tableau 2.

Pour ce qui est du perceptron multicouche, nous considérons également les taux pondérés de mauvaise classification, afin de pouvoir comparer les résultats obtenus avec ceux de l'analyse discriminante.

La couche d'entrée de ce réseau contient deux, quatre ou six neurones, dépendant du nombre de variables considéré, tandis que le nombre de neurones artificiels dans la couche de sortie est fixé à deux, soit le nombre de classes présentes dans les données. Après différents essais, nous avons opté pour une seule couche cachée contenant quatre neurones, soit deux fois le nombre de classes considéré. Comme pour l'analyse discriminante, chacune des dix répliques a été considérée comme fichier d'entraînement.

Le tableau 4 présente les résultats obtenus avec l'analyse discriminante de même qu'avec le perceptron multicouche. Les paramètres $\bar{x}$ et s représentent respective-

Table 4: *Taux pondérés de l'analyse discriminante et du PMC.*

Situation étudiée	# de var.	Recouvrement 5%		Recouvrement 15%	
		AD	Perceptron	AD	Perceptron
	2	$\bar{x} = 5.0$ $s = 0.5$ *	$\bar{x} = 5.2$ $s = 0.5$	$\bar{x} = 15.3$ $s = 0.7$	$\bar{x} = 15.6$ $s = 0.8$
	2	$\bar{x} = 5.0$ $s = 0.5$ *	$\bar{x} = 5.4$ $s = 0.6$	$\bar{x} = 15.3$ $s = 0.7$	$\bar{x} = 15.7$ $s = 0.9$
	2	$\bar{x} = 4.9$ $s = 0.6$	$\bar{x} = 5.1$ $s = 0.6$	$\bar{x} = 14.8$ $s = 1.1$ *	$\bar{x} = 14.9$ $s = 1.0$
	2	$\bar{x} = 5.0$ $s = 0.6$ *	$\bar{x} = 5.3$ $s = 0.7$	$\bar{x} = 15.4$ $s = 0.7$	$\bar{x} = 15.6$ $s = 0.7$
	2	$\bar{x} = 5.1$ $s = 0.6$ *	$\bar{x} = 5.5$ $s = 0.8$	$\bar{x} = 15.3$ $s = 0.7$	$\bar{x} = 15.7$ $s = 0.9$
	2	$\bar{x} = 5.3$ $s = 0.6$	$\bar{x} = 5.7$ $s = 0.9$	$\bar{x} = 15.0$ $s = 0.9$ *	$\bar{x} = 15.4$ $s = 0.9$
	4	$\bar{x} = 5.1$ $s = 0.5$	$\bar{x} = 5.5$ $s = 0.7$	$\bar{x} = 15.4$ $s = 0.7$	$\bar{x} = 15.8$ $s = 0.9$
	4	$\bar{x} = 5.1$ $s = 0.5$	$\bar{x} = 5.7$ $s = 0.6$	$\bar{x} = 15.4$ $s = 0.7$	$\bar{x} = 15.6$ $s = 0.7$
	4	$\bar{x} = 4.9$ $s = 0.7$	$\bar{x} = 5.1$ $s = 0.6$	$\bar{x} = 14.8$ $s = 1.1$ *	$\bar{x} = 15.1$ $s = 1.1$
	4	$\bar{x} = 5.0$ $s = 0.6$ *	$\bar{x} = 5.6$ $s = 0.8$	$\bar{x} = 15.4$ $s = 0.6$	$\bar{x} = 15.7$ $s = 0.8$
	4	$\bar{x} = 5.1$ $s = 0.6$ *	$\bar{x} = 5.4$ $s = 0.7$	$\bar{x} = 15.3$ $s = 0.8$	$\bar{x} = 15.7$ $s = 1.0$
	4	$\bar{x} = 5.3$ $s = 0.7$	$\bar{x} = 5.7$ $s = 0.8$	$\bar{x} = 15.1$ $s = 0.9$	$\bar{x} = 15.4$ $s = 1.0$
	6	$\bar{x} = 5.1$ $s = 0.5$	$\bar{x} = 5.5$ $s = 0.6$	$\bar{x} = 15.3$ $s = 0.7$	$\bar{x} = 15.9$ $s = 0.8$
	6	$\bar{x} = 5.1$ $s = 0.5$	$\bar{x} = 5.7$ $s = 0.8$	$\bar{x} = 15.3$ $s = 0.7$	$\bar{x} = 15.9$ $s = 0.9$
	6	$\bar{x} = 4.9$ $s = 0.6$	$\bar{x} = 5.2$ $s = 0.6$	$\bar{x} = 14.9$ $s = 1.0$	$\bar{x} = 15.0$ $s = 1.0$
	6	$\bar{x} = 5.0$ $s = 0.6$ *	$\bar{x} = 5.6$ $s = 0.9$	$\bar{x} = 15.3$ $s = 0.6$	$\bar{x} = 15.6$ $s = 0.7$
	6	$\bar{x} = 5.1$ $s = 0.6$ *	$\bar{x} = 5.5$ $s = 0.6$	$\bar{x} = 15.4$ $s = 0.8$	$\bar{x} = 15.8$ $s = 0.9$
	6	$\bar{x} = 5.3$ $s = 0.7$	$\bar{x} = 5.8$ $s = 0.8$	$\bar{x} = 15.0$ $s = 0.9$	$\bar{x} = 15.7$ $s = 1.1$

ment la moyenne et l'écart-type des 100 taux de mauvaise classification obtenus pour chacune des combinaisons de facteurs présentées. Afin de pouvoir mieux évaluer nos méthodes, nous avons comparé ces taux pondérés avec les taux empiriques correspondants à l'aide de 72 tests de Student couplés. Une étoile ($*$) à côté d'un résultat signifie que le taux obtenu n'est pas significativement différent, au seuil théorique de 5%, du taux empirique correspondant.

On peut remarquer que le nombre de situations où le taux obtenu avec l'analyse discriminante diffère significativement du taux empirique correspondant augmente avec le nombre de variables. À part quelques exceptions, les taux pondérés obtenus avec l'analyse discriminante sont supérieurs aux taux empiriques correspondants.

Quant au perceptron multicouche, ce dernier constat a été observé pour toutes les situations présentées.

À l'aide de tests de Student couplés, nous avons ensuite comparé les taux pondérés de l'analyse discriminante et du perceptron multicouche . Pour toutes les situations illustrées dans le tableau 4, il existe une différence significative, au seuil de 5%, entre les résultats de ces deux méthodes, le perceptron donnant un taux supérieur à celui de l'analyse discriminante.

Finalement pour déterminer quels sont les facteurs qui, parmi ceux considérés, influencent le taux de mauvaise classification obtenu avec l'analyse discriminante et le perceptron multicouche, une analyse de variance à cinq facteurs à effets fixes a été effectuée pour chacune des deux méthodes. Nous avons considéré les effets principaux de ces facteurs de même que toutes les interactions doubles et triples existantes entre ces facteurs. À noter que les interactions impliquant le facteur qui concerne la variance de la première variable et celui de l'orientation des nuages définissant les classes n'ont pas été considérées, du fait qu'elles étaient inexistantes. En effet, il était inutile de faire une rotation des classes dont les variances de chacune des variables étaient égales. Ces analyses de variance ont été faites sur un échantillon de 7200 observations (100 taux pondérés pour chacune des 72 combinaisons des facteurs).

Puisque les hypothèses de normalité et d'homogénéité des variances des résidus n'étaient pas vérifiées pour les deux méthodes, nous avons tout d'abord essayé une transformation qui peut parfois être utile en présence de proportions. Nous avons donc refait les analyses en prenant comme variable réponse la cosécante du radical du taux pondéré de mauvaise classification ($\tau' = \arcsin(\sqrt{\tau})$). Malheureusement, cette transformation n'a pas permis de valider les hypothèses. Nous avons donc utilisé un test non paramétrique, le test de Kruskal-Wallis, c'est-à-dire que nous avons refait des analyses de variance, mais cette fois sur les rangs des taux pondérés de mauvaise classification.

Nous présentons seulement les résultats significatifs obtenus avec les tests paramétrique (Fisher) et non paramétrique (Kruskal-Wallis), qui sont compilés au tableau 5. Par exemple, si l'interaction entre les facteurs A et B est significative, les résultats des effets principaux de ces deux facteurs ne sont pas présentés.

Les deux analyses, paramétrique et non paramétrique, montrent que le taux pondéré de mauvaise classification obtenu avec l'analyse discriminante est influencé par le recouvrement existant entre les deux classes, la taille des deux classes de même que l'orientation des nuages définissant les classes, l'interaction triple entre ces trois facteurs étant significative avec un seuil observé de 0.0001.

D'après les deux types d'analyses, le taux pondéré de mauvaise classification

Table 5: *Résultats des analyses de variance des méthodes supervisées.*

Méthode	**Source**	**Fisher**	**Kruskal-Wallis**
A. discriminante	orientation $*$ taille $*$ recouvrement	0.0001	0.0001
Perceptron	variable	0.0001	0.0001
	variance $*$ recouvrement	0.0448	0.0551
	taille $*$ variance	0.0583	0.0375
	orientation $*$ taille $*$ recouvrement	0.0001	0.0001

obtenu avec le perceptron multicouche est également influencé par les trois mêmes facteurs que l'analyse discriminante, leur interaction triple étant également significative avec le même seuil observé de 0.0001, de même que par le nombre de variables, l'effet principal de ce facteur étant significatif au seuil observé de 0.0001. Plus le nombre de variables augmente, plus le taux pondéré de mauvaise classification du perceptron est élevé.

Par contre, les deux interactions doubles présentées sont significatives seulement pour une des deux analyses. Il faut toutefois remarquer que la différence entre les seuils observés des deux analyses est très faible pour chacune des interactions doubles et que les seuils observés ne sont que légèrement inférieurs ou supérieurs au seuil théorique de 5%.

5.2 Classification et réseau de Kohonen

Pour les différents critères de classification considérés, nous avons calculé également deux taux de mauvaise classification : le taux moyen et le taux pondéré, tels que définis précédemment. Contrairement aux méthodes supervisées, nous avons retenu ici le taux moyen de mauvaise classification. Ce choix a été principalement fait en raison du comportement du saut minimal puisque plus les classes sont débalancées, plus ce taux est faible. En effet, ce critère a tendance à regrouper tous les individus dans une classe à l'exception d'un individu qui forme la deuxième classe.

Les figures 6 et 7 illustrent les résultats obtenus par les différents critères de classification considérés pour le cas de deux variables. Elles montrent le minimum, la moyenne et le maximum obtenus pour chacun des critères utilisés. La discussion qui suit prend également en considération les résultats obtenus pour 4 et 6 variables.

Encore une fois dans le but de pouvoir mieux comparer nos méthodes, les taux obtenus ont été comparés aux taux empiriques correspondants en utilisant des tests de Student couplés. Les résultats qui ne sont pas significativement différents des taux empiriques correspondants, au seuil théorique de 5%, sont indiqués par une étoile ($*$) à droite des figures.

Les points saillants qui ressortent de cette analyse se résument comme suit. Premièrement, le critère du k-means est celui qui donne le plus souvent les taux de mauvaise classification les plus faibles. Ce critère est suivi de près par le critère de Ward. Cette bonne performance s'explique par le fait que ces critères, en plus de tenir compte du poids des classes, forment celles-ci de façon à ce que les individus

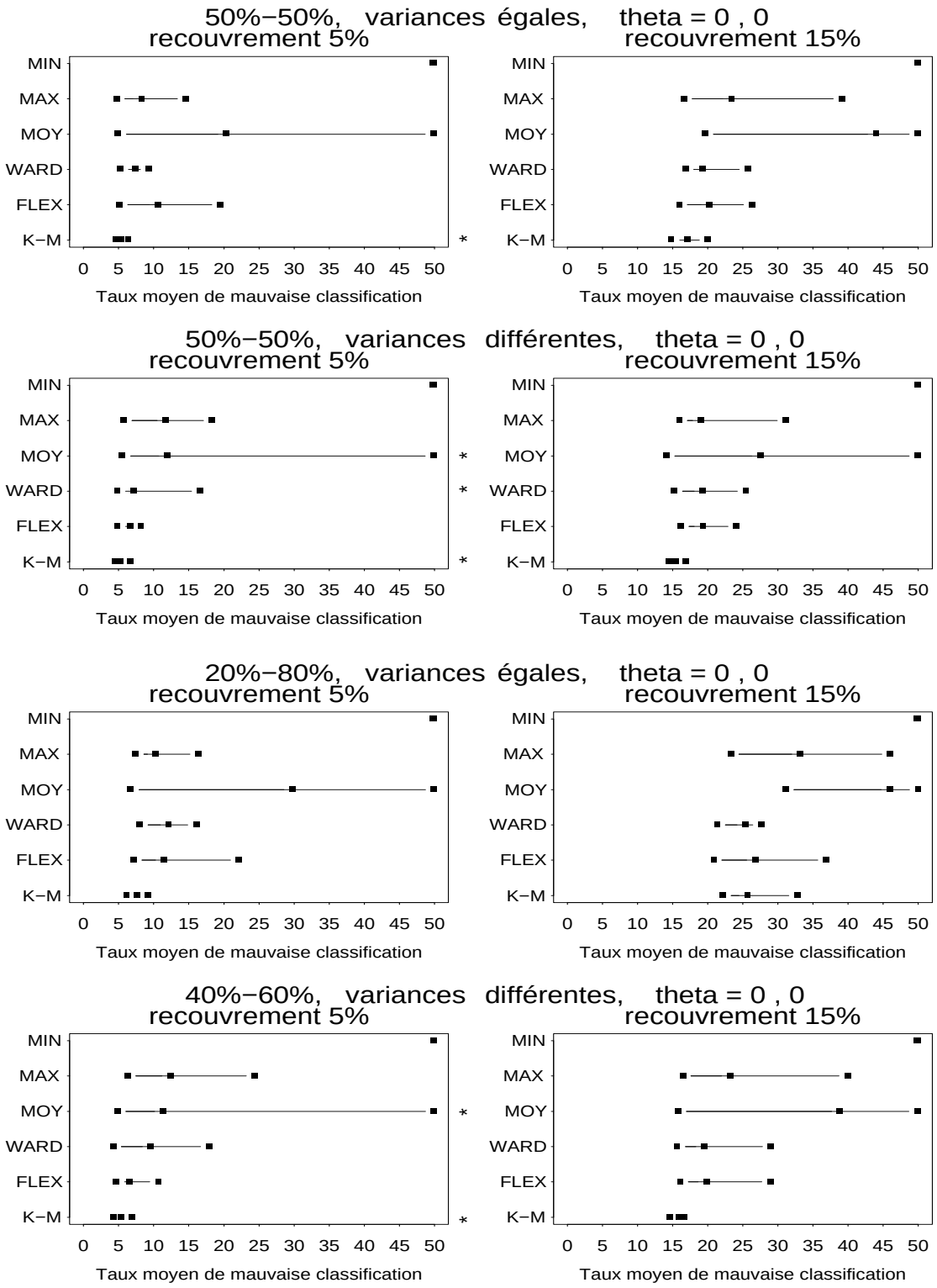


Figure 6: *Résultats des critères de classification pour 2 variables.*

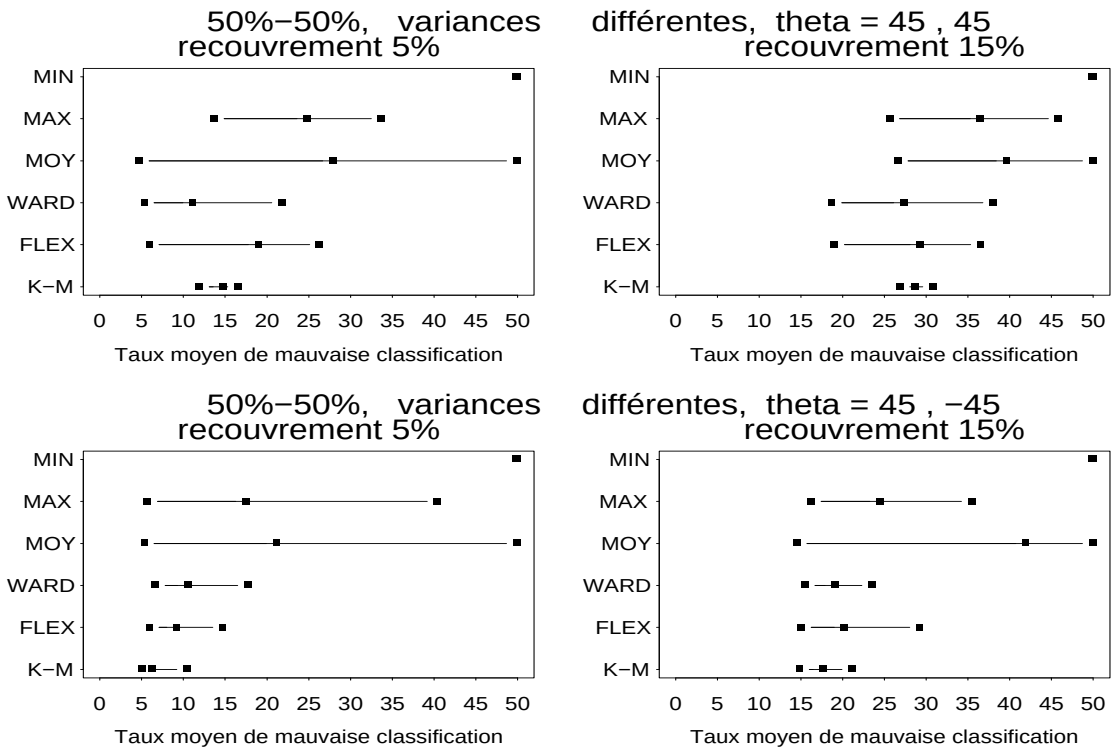


Figure 7: *Résultats des critères de classification pour 2 variables (suite).*

soient le plus près possible de leur centre de gravité respectif, centre qui est recalculé après chaque affectation. Par conséquent, ces critères ont la capacité de fournir des classes plus ou moins balancées.

Deuxièment, le critère de la moyenne donne, dans certains cas, des taux qui ne diffèrent pas de façon significative (au seuil de 5%) des taux empiriques, bien que la moyenne de ses taux se situe bien au-dessus des taux empiriques. Ce résultat, bien que surprenant à première vue, est dû à la grande variabilité qui existe dans les résultats obtenus par ce critère causée par quelques valeurs extrêmes.

La nature de nos données ne favorise pas le critère du saut minimal. Sa moins bonne performance s'explique par la propriété de chaînage qui est d'autant plus présente en raison du recouvrement existant entre les classes.

Lorsque les classes sont très débalancées (20%-80%) et qu'il existe un recouvrement assez important (15%) entre elles, aucun des critères étudiés ne performe bien, les meilleurs donnant un taux de mauvaise classification autour de 25%. On s'y attendait un peu, car c'est une situation où les classes sont assez difficiles à définir.

Un autre cas difficile mentionné précédemment est celui où les classes sont orientées à 45° par rapport aux axes puisque tous les critères de classification donnent des taux nettement supérieurs aux taux désirés.

En ce qui concerne le réseau de Kohonen, une carte rectangulaire de 4 par 8 neurones a été choisie après plusieurs expérimentations. Comme ce réseau n'est pas conçu pour édifier directement une classiffication, nous devions décider d'une partition à deux classes des 32 neurones constituant cette carte et caractérisés par leurs vecteurs poids. Cette partition peut être choisie de façon aléatoire ou encore en affectant le même nombre de neurones à chacune des classes. Cependant, pour mieux représenter

la nature de nos données, nous optons pour une partition obtenue par le critère du k-means. On attribue, par la suite, chaque individu au neurone le plus proche et on calcule le taux de mauvaise classification correspondant.

Mentionnons également que, comme pour les méthodes supervisées, chacune des dix répliques d'une combinaison de facteurs a servi de fichier d'entraînement. La matrice de poids résultante obtenue par le Kohonen a été par la suite utilisée pour tester les dix fichiers. De sorte qu'à chacune des combinaisons des facteurs à l'étude correspond 100 taux moyens de mauvaise classification.

Le tableau 6 compile les résultats obtenus avec les deux critères de classification, le moins bon et le meilleur, et le réseau de Kohonen. Les paramètres $\bar{x}$ et s représentent respectivement la moyenne et l'écart-type des taux de mauvaise classification obtenus pour chacune des situations présentées. Les meilleurs résultats de la classification proviennent soit du k-means (k), soit du critère de Ward (w) ou encore du critère flexible (f), tandis que les moins bons proviennent du saut minimal (m).

Les taux moyens qui ne sont pas significativement différents, au seuil théorique de 5%, des taux empiriques correspondants sont indiqués par une étoile ($*$). On peut alors remarquer que les taux moyens obtenus par le Kohonen sont tous significativement plus élevés que les taux empiriques correspondants.

À l'aide de tests de Student, nous avons comparé les taux moyens des critères de classification et du Kohonen. Pour toutes les situations envisagées dans cette étude, le Kohonen performe mieux que les critères du saut minimal, du saut maximal et du saut moyen, au seuil théorique de 5%.

Plus spécifiquement, des 36 situations illustrées au tableau 6, on observe que le Kohonen est meilleur que le k-means, le critère de Ward et le critère flexible dans 17, 32 et 33 cas, respectivement. Il donne des résultats équivalents à ces derniers trois critères dans 6, 2 et 2 cas, respectivement.

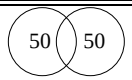
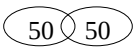
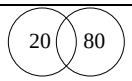
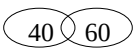
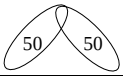
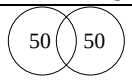
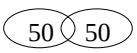
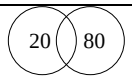
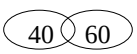
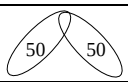
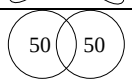
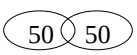
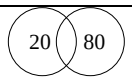
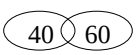
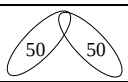
Quant à la comparaison spécifique du Kohonen versus le k-means, rappelons tout d'abord que le Kohonen utilise le k-means pour partitionner ces neunones d'entrée. Les 17 cas sur 36 où le Kohonen est meilleur que le k-means se résument dans les trois situations suivantes : 1) les classes sont balancées ou très débalancées avec un recouvrement important; 2) les classes sont orientées différemment à 45° et $-45°$ quelque soit le recouvrement considéré et 3) les classes sont orientées toutes les deux à 45° avec un faible recouvrement. Cette dernière situation donne exceptionnellement un résultat moins bon pour le Kohonen comparativement à Ward ou au critère flexible, et ceci dépend du nombre de variables à l'étude. Mais, comme nous l'avons déjà fait remarquer auparavant, cette situation semble présenter des difficultés à tous les critères de classification.

À souligner, toujours par rapport à cette situation, que lorsque le recouvrement est plus important, le Kohonen et le k-means donnent des résultats équivalents à une exception près.

Cependant, le k-means performe mieux que le Kohonen principalement lorsque les variances à l'intérieur des classes, qui sont balancées ou peu débalancées, sont différentes avec une orientation de 0°, c'est-à-dire des classes allongées n'ayant pas subi de rotation, à l'exception toutefois du cas de deux variables ou les performances sont les mêmes. Et ceci est vrai quelque soit le recouvrement considéré.

Signalons finalement que le cas relatif aux classes très débalancées est partagé par le critère de Ward, du k-means et du Kohonen dépendant du recouvrement et du

Table 6: *Taux moyens du Kohonen et de certains critères de classification.*

Situation étudiée	# de var.	Para-mètre	Recouvrement 5%			Recouvrement 15%		
			Koh.	Classification		Koh.	Classification	
				pire	meilleur		pire	meilleur
	2	$\bar{x}$	5.1	49.9 (m)	5.3 (k)∗	15.6	49.9 (m)	17.1 (k)
		s	0.5	0.04	0.6	0.8	0.04	1.8
	2	$\bar{x}$	5.3	49.9 (m)	5.2 (k)∗	15.4	49.9 (m)	15.4 (k)
		s	0.6	0.04	0.7	0.7	0.04	0.7
	2	$\bar{x}$	9.1	49.8 (m)	7.7 (k)	24.5	49.8 (m)	25.4 (w)
		s	1.1	0.09	1.0	1.5	0.09	2.3
	2	$\bar{x}$	6.5	49.9 (m)	5.4 (k)∗	16.0	49.9 (m)	15.8 (k)
		s	0.8	0.03	0.8	0.7	0.03	0.7
	2	$\bar{x}$	12.7	49.9 (m)	11.1 (w)	27.5	49.9 (m)	27.3 (w)
		s	1.5	0.03	5.5	1.6	0.04	6.8
	2	$\bar{x}$	5.3	49.9 (m)	6.2 (k)	15.1	49.9 (m)	17.6 (k)
		s	0.7	0.03	1.6	0.9	0.03	2.1
	4	$\bar{x}$	7.2	49.9 (m)	5.9 (k)	18.1	49.9 (m)	20.8 (w)
		s	1.1	0.00	0.7	1.2	0.00	1.8
	4	$\bar{x}$	5.5	49.9 (m)	5.2 (k)∗	15.7	49.9 (m)	15.4 (k)
		s	0.7	0.00	0.7	0.7	0.03	0.7
	4	$\bar{x}$	9.8	49.8 (m)	10.1 (k)	29.2	49.8 (m)	28.0 (w)
		s	1.3	0.10	3.2	3.7	0.10	3.7
	4	$\bar{x}$	6.3	49.9 (m)	5.4 (k)∗	16.2	49.9 (m)	15.8 (k)
		s	0.9	0.02	0.8	0.7	0.02	0.6
	4	$\bar{x}$	12.3	49.9 (m)	14.2 (w)	28.4	49.9 (m)	28.3 (f)
		s	2.3	0.03	7.7	2.4	0.03	3.0
	4	$\bar{x}$	5.4	49.9 (m)	6.2 (k)	15.8	49.9 (m)	18.4 (k)
		s	0.7	0.03	1.7	0.9	0.03	2.5
	6	$\bar{x}$	7.3	49.9 (m)	6.5 (k)	18.7	49.9 (m)	21.9 (w)
		s	0.9	0.00	0.9	1.4	0.00	3.6
	6	$\bar{x}$	5.6	49.9 (m)	5.2 (k)∗	16.0	49.9 (m)	15.4 (k)
		s	0.7	0.00	0.7	0.9	0.00	0.8
	6	$\bar{x}$	9.9	49.8 (m)	10.6 (k)	30.1	49.8 (m)	30.6 (w)
		s	1.4	0.09	3.8	3.6	0.09	4.3
	6	$\bar{x}$	5.6	49.9 (m)	5.4 (k)∗	16.3	49.9 (m)	15.8 (k)
		s	0.6	0.02	0.9	0.8	0.02	0.9
	6	$\bar{x}$	13.2	49.9 (m)	12.1 (f)	28.6	49.9 (m)	28.7 (k)
		s	2.9	0.00	3.5	2.6	0.00	1.3
	6	$\bar{x}$	5.6	49.9 (m)	6.3 (k)	16.1	49.9 (m)	18.0 (k)
		s	0.8	0.00	1.8	0.9	0.00	2.1

nombre de variables à l'étude.

Comme pour les méthodes supervisées, nous avons effectué des analyses de variance pour exhiber les facteurs qui influencent le taux moyen de mauvaise classification des différents critères de classification et du réseau de Kohonen. Tout comme celles effectuées précédemment, seules les interactions existantes d'ordre inférieur ou égal à 3 ont été considérées. Ces analyses ont été effectué sur un échantillon de 720 fichiers pour tous les critères de classification et sur 7200 fichiers pour le Kohonen. Les conclusions tirées sont basées sur un seuil théorique de 5%.

D'après les premières analyses effectuées, les hypothèses de normalité et d'homogénéité des variances des résidus n'étaient pas vérifiées pour chacun des critères. Nous avons donc de nouveau essayé une transformation de la variable réponse, soit la cosécante du radical du taux moyen de mauvaise classification ($\tau' = \arcsin(\sqrt{\tau})$). Cette transformation n'a été utile en partie que pour le critère de Ward car elle a permis d'obtenir la normalité des résidus. Pour tous les autres critères, nous avons utilisé le test de Kruskal-Wallis. Les conclusions tirées à partir de ce test allant essentiellement dans le même sens que celles du test paramétrique, nous nous limitons à présenter uniquement les résultats significatifs du test de Fisher (voir tableau 7).

Pour le critère du saut minimal, seule la taille des classes influence de façon significative le taux moyen de mauvaise classification avec un seuil observé de 0.0001. Lorsque les classes sont très débalancées (20%-80%), le taux est significativement plus faible que lorsqu'elles sont de même taille (50%-50%) ou légèrement débalancées (40%-60%), ces deux traitements n'étant pas différents de façon significative.

En ce qui concerne le saut maximal, tous les facteurs considérés ont un effet significatif sur le taux moyen de mauvaise classification, à l'exception du nombre de variables. Deux interactions doubles, celle concernant le recouvrement entre les deux classes et d'une part, la taille des classes et d'autre part, la variance de la première variable, sont significatives, de même que l'effet simple de l'orientation des nuages définissant les classes. En effet, tous les traitements de ce facteur diffèrent de façon significative et c'est lorsque l'angle des classes par rapport aux axes est nul qu'on retrouve le taux le plus faible, suivi respectivement des orientations de $45°$ et $-45°$ et de celles de $45°$ pour les deux classes.

Tous les facteurs ont également un effet significatif sur le taux donné par le critère de la moyenne. En effet, l'interaction triple entre le nombre de variables, la variance et le recouvrement est significative, ainsi que l'interaction double entre la taille et la variance et celle concernant l'orientation des classes et le recouvrement.

Le critère de Ward est lui aussi influencé par tous les facteurs étudiés. En plus de l'interaction triple entre l'orientation, la taille et le recouvrement, il existe une interaction significative entre le nombre de variables et la variance de la première variable d'une classe.

L'analyse de variance effectuée sur le taux moyen obtenu par le critère flexible montre qu'en plus de deux interactions triples, soit celle entre la taille, le recouvrement et d'une part, le nombre de variables, et d'autre part l'orientation des classes, l'effet simple de la variance est significatif. Un taux plus faible est obtenu lorsque la variance de la première variable est fixée à 1, comme les autres variables, plutôt qu'à 4.

Pour la méthode du k-means, trois interactions entre les facteurs s'avèrent si-

Table 7: *Résultats des analyses de variance des méthodes non suprevisées.*

Critère	**Source**	**Fisher**
Saut minimal	taille	0.0001
Saut maximal	orientation	0.0001
	taille ∗ recouvrement	0.0321
	variance ∗ recouvrement	0.0001
Saut moyen	orientation ∗ recouvrement	0.0001
	taille ∗ variance	0.0177
	variable ∗ variance ∗ recouvrement	0.0002
Ward	variable ∗ variance	0.0237
	orientation ∗ taille ∗ recouvrement	0.0119
Flexible	variance	0.0001
	taille ∗ recouvrement ∗ variable	0.0250
	taille ∗ recouvrement ∗ orientation	0.0060
k-means	orientation ∗ recouvrement	0.0011
	taille ∗ recouvrement	0.0001
	variable ∗ variance ∗ recouvrement	0.0007
Kohonen	variable ∗ taille ∗ orientation	0.0001
	variable ∗ taille ∗ variance	0.0265
	variable ∗ taille ∗ recouvrement	0.0004
	variable ∗ orientation ∗ recouvrement	0.0001
	variable ∗ variance ∗ recouvrement	0.0001
	taille ∗ orientation ∗ recouvrement	0.0001
	taille ∗ variance ∗ recouvrement	0.0001

gnificatives. Il y a tout d'abord celle entre le nombre de variables, la variance et le recouvrement. On retrouve également deux interactions doubles, une concernant l'orientation des nuages définissant les classes et le recouvrement et l'autre interaction se situant entre ce dernier facteur et la taille des classes.

Finalement, toutes les interactions triples pour le réseau Kohonen sont significatives.

5.3 Conclusion

Cet article présente une étude basée sur un procédé de simulation original des données afin d'évaluer la performance, d'une part, de méthodes supervisées et, d'autre part, de méthodes non supervisées.

Les résultats obtenus pour les méthodes supervisées, confrontant une méthode classique à une méthode neuronale, montrent que l'analyse discriminante donne un taux pondéré de mauvaise classification plus faible que le perceptron multicouche.

Ceci étant dit, même si l'analyse discriminante est significativement meilleure que le perceptron, les taux obtenus sont tout de même très proches, soit à 2.7% près. Cette relative bonne performance de l'analyse discriminante est obtenue alors que 1) nos données présentent une situation qui lui est *idéale* puisqu'elles proviennent de

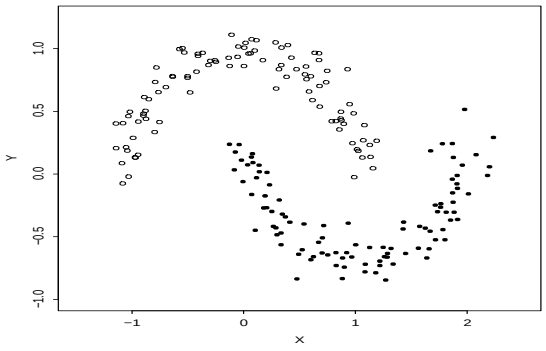


Figure 8: *Deux classes concaves imbriquées.*

distributions normales et 2) nous avons retenu les taux pondérés avec probabilités a priori proportionnelles aux tailles des classes, pour mieux respecter la structure de nos classes. Les taux moyens sont, dans ce cas, presque deux fois plus grands lorsque les classes sont très débalancées. Alors que l'information relative à la répartition des individus dans les classes n'a pas été fournie explicitement au perceptron. Il est donc difficile pour ce dernier de faire mieux que l'analyse discriminante, à moins de l'entraîner plus longtemps, ce qui n'est pas souhaitable.

Cependant, le perceptron multicouche peut être très performant dans certains cas comme, par exemple, le cas de deux classes concaves imbriquées et distinctes, tel qu'illustré à la figure 8. Ce réseau donne un taux de mauvaise classification de 0% alors que celui obtenu par l'analyse discriminante est de 9%. Ceci est principalement dû au fait que ce réseau peut tracer une frontière de séparation non linéaire. Pour ce qui est des méthodes non supervisées, nous avons confronté six critères classiques de classification entre eux et avec le réseau de Kohonen, même si ce dernier n'est pas conçu pour édifier directement une classification.

Les six critères de classification se séparent en deux groupes. L'un est formé de critères réalisant une bonne performance quant à la nature de nos données. Il s'agit du k-means, du critère de Ward et du critère flexible. L'autre groupe contient le saut minimal, le saut maximal et le saut moyen qui ont obtenu une faible performance. Plus particulièrement, le saut minimal est de loin celui qui performe le moins bien. Et pourtant sur les classes concaves imbriquées illustrées à la figure 8, c'est celui, parmi les six critères de classification, qui donne le meilleur résultat soit 0% de taux de mauvaise classification. Les cinq autres critères donnant des taux allant de 13.5% à 18%, le moins bon résultat est obtenu par le k-means.

À la lumière des 36 situations présentées dans ce travail, le réseau de Kohonen semble globalement donner de meilleurs résultats que le groupe des trois critères de classification les plus performants. Il est cependant suivi de très près du k-means qui donne de meilleurs résultats principalement lorsque les variances à l'intérieur des classes, qui sont balancées ou peu débalancées, sont différentes avec une orientation de 0°, c'est à dire des classes allongées n'ayant pas subi de rotation, à l'exception toutefois du cas de deux variables ou les performances sont les mêmes. Et ceci est vrai quelque soit le recouvrement considéré.

Dans le cas des deux classes concaves imbriquées, le Kohonen est aussi performant que le saut minimal, moyennant un choix judicieux de la carte d'auto-organisation.

Toute cette étude a porté exclusivement sur deux classes alors que le nombre de classes est l'un des six facteurs sur lesquels portent la planification. Les perspectives de ce travail se situent, entre autres, à ce niveau où nous aimerions analyser l'impact du nombre de classes sur la performance des méthodes lorsque ce nombre augmente. Nous aimerions également étudier le cas de classes dont les distributions ne sont pas nécessairement normales.

6 Références

Haykin, S. (1994) *Neural Networks: A comprehensive Foundation.* IEEE Press.

Kohonen, T. (1982) "Self-Organized Formation of Topologically Correct Feature Maps". *Biological Cybernetics*, 43, 59–69.

Kohonen, T. (1990) *Self-Organized and Associative Memory Analysis and Machine Intelligence.* Springer-Verlag.

Lachenbruch, P.A. (1975) *Discriminant Analysis.* Hafner Press.

MacQueen, J.B. (1967) "Some methods for classification and analysis of multivariate observations". *Proc. Fifth Berkely Symp. Math. Stat. Prob.*, 1, 281–297.

Milligan, W.G. & Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a data set", *Psychometrika*, 50, No.2, 159–179.

Rosenblatt, F. (1958) "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". *Psychological Review*, 65, No. 6, 386–408.

Rosenblatt, F. (1962) *Principles of Neurodynamics.* Spartan Books.

Rumelhart, D.E. & McClellan, J.L. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press.

Seber, G.A.F. (1984) *Multivariate Observations.* Wiley Series in probability and mathematical statistics.