

REVUE DE STATISTIQUE APPLIQUÉE

J.-P. NAKACHE

J. VILAIN

B. FERTIL

**Discrimination et régression par une méthode
neuromimétique et par la méthode de segmentation
CART : application à différentes données et
comparaison des résultats**

Revue de statistique appliquée, tome 44, n° 4 (1996), p. 19-40

http://www.numdam.org/item?id=RSA_1996__44_4_19_0

© Société française de statistique, 1996, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

DISCRIMINATION ET RÉGRESSION PAR UNE MÉTHODE NEUROMIMÉTIQUE ET PAR LA MÉTHODE DE SEGMENTATION CART : APPLICATION À DIFFÉRENTES DONNÉES ET COMPARAISON DES RÉSULTATS

J.-P. Nakache*, J. Vilain**, B. Fertil**

* INSERM U88 – Hôpital National de Saint-Maurice, 14 rue du Val d'Osne,
94415 Saint-Maurice Cedex

** INSERM U66 – CHU Pitié Salpêtrières, 91 boulevard de l'Hôpital, 75634 Paris Cedex 13

RÉSUMÉ

Les méthodes – neuromimétique et segmentation selon CART – sont exposées sans entrer dans les détails théoriques qui figurent dans plusieurs articles ou livres fournis en référence. Ces deux types de méthode sont appliqués dans un contexte de discrimination ou de régression à trois ensembles de données ayant déjà fait l'objet de plusieurs études statistiques. Les résultats sont obtenus à partir d'un échantillon d'apprentissage et leur qualité est évaluée à l'aide d'un échantillon indépendant.

Pour ces jeux de données, la méthode neuronale fournit de meilleurs résultats que ceux obtenus par la méthode CART. La méthode CART requiert sur ces exemples, quelques minutes de temps calcul pour établir la règle de décision ou de prédiction tandis que la méthode neuronale nécessite un temps calcul nettement plus long et une connaissance de l'outil très approfondie. Par contre les deux méthodes ont des temps de réponse courts quand il s'agit d'appliquer la règle de décision ou de prédiction à de nouveaux cas.

Mots-clés : Réseau de neurones, rétropropagation du gradient, apprentissage supervisé, segmentation, élagage, discrimination, régression, règle de décision ou de prédiction, risque, erreur quadratique moyenne.

ABSTRACT

Two methods – neural networks and CART tree-structured method – are presented without giving theoretical details that can be found in several articles or books provided in the references. Both methods are applied, in a discrimination or a regression context, to three datasets already used in several published statistical papers. The results are obtained from a learning sample and their accuracy is evaluated using an independent test set.

For these data, the results obtained by using neural networks are better, in terms of accuracy, than those provided by CART method. For these examples, a few minutes are required to build the decision or the prediction rule when using CART methodology, which is not the case with neural networks. In this later case an important and prohibitive computer time and

an advanced knowledge of the tool are required. But with both methods, the response time is very short when applying the rules to new cases.

Keywords : *Neural network, back-propagation, supervised learning, tree-structured classification or regression, pruning, prediction or decision rule, risk, mean square error.*

Introduction

On note depuis quelques années l'utilisation de méthodes neuromimétiques en Statistique pour effectuer des tâches de classification, de discrimination, de régression.... (Ripley, 1994b). Ces méthodes ont d'ailleurs été comparées, dans des problèmes bien définis, aux méthodes statistiques classiques (Burke *et al.*, 1994; Leon, 1994; Shadmehr *et al.*, 1990; Gallinari *et al.*, 1991). Il faut, particulièrement, mentionner le récent ouvrage de Michie *et al.* (1994) qui rapporte les résultats d'une importante étude sur la comparaison et l'évaluation d'un ensemble de techniques de classification et de discrimination, étude effectuée dans le cadre d'un projet européen ESPRIT (STATLOG).

D'autre part dans le domaine de la segmentation, la méthode CART (Breiman *et al.*, 1984) est considérée comme une méthode de référence (Nakache *et al.*, 1993; Ripley, 1993). Il nous paraît donc intéressant de comparer ces deux méthodes sur des données classiques (DIGITS, DAA, HOUSING), ayant déjà fait l'objet de différentes études statistiques.

Les données DIGITS, très connues dans la littérature, concernent un problème de reconnaissance des 10 chiffres (Breiman *et al.*, 1984). Il s'agit plus précisément d'une analyse discriminante entre 10 groupes d'observations.

Les données DAA ont été recueillies sur un ensemble de patients atteints de douleurs abdominales aiguës qui ont fait l'objet de plusieurs articles (Seroussi; ARC et AURC, 1985). Un sous-ensemble des données complètes relatif à 6 groupes diagnostiques a été soumis à l'analyse discriminante.

En ce qui concerne la régression, elle a été appliquée aux données HOUSING (Harrison et Rubinfeld, 1978) soumises, par leurs auteurs, à un modèle de régression linéaire multiple après transformation des variables indépendantes.

Les méthodes – neuromimétiques et CART – sont décrites sans trop de détails théoriques que le lecteur pourra trouver dans les nombreux articles ou livres cités en référence. Elles sont appliquées à partir d'un échantillon d'apprentissage conduisant à la construction soit d'une règle de décision (discrimination), soit d'une règle de prédiction (régression). Ces règles sont ensuite évaluées sur un échantillon indépendant. Cette évaluation est fondée, dans un contexte de discrimination sur le taux d'erreur de classement et l'adéquation des probabilités a posteriori et, dans un contexte de régression sur l'erreur relative de prédiction.

La méthode neuronale

Les réseaux de neurones font l'objet depuis quelques années d'une intense activité. Il s'agit, sur la base d'une analogie neurobiologique de la structure et du

fonctionnement du cerveau, de mettre en place des procédures nouvelles de traitement de l'information (Davallo et Naïm, 1993; Weisbuch, 1989). La capacité croissante des calculateurs est mise à profit pour simuler des neurones et construire des architectures neuronales complexes avec lesquelles il est possible de jouer tout à loisir (Masters, 1993; McClelland et Rumelhart, 1988).

Les applications de ces techniques nouvelles sont multiples. En particulier, on peut remarquer que de nombreux travaux ont été consacrés à l'analyse de données (Ripley, 1994a). Dans l'état actuel des connaissances, les réseaux de neurones sont des outils de choix pour effectuer des tâches d'identification et de classification (Gernoth et Clark, 1995, Ripley, 1994b), pour détecter des traits discriminants des signaux présentés à leur entrée (Cichocki et Unbehauen, 1993; Coustère et Fertil, 1992; Masters, 1994). Ils sont particulièrement adaptés lorsque les données sont bruitées ou incomplètes. Les données aberrantes n'ont en général que peu d'influence. Les méthodes neuronales sont notamment très utiles lorsqu'il est difficile d'identifier les paramètres d'intérêt et de formaliser les phénomènes étudiés.

Les types de réseaux de neurones sont très variés et dépendent de l'application désirée. Cependant, ils fonctionnent en général sur le même principe : *analyser des données par une méthode neuronale revient à modéliser celles-ci à l'aide d'un modèle fortement non linéaire et surparamétré*. Un réseau de neurones est souvent considéré comme une boîte noire munie d'une entrée et d'une sortie. Les données sont présentées au réseau sous la forme d'une configuration d'activation des neurones d'entrée. Cette activation est propagée à travers le réseau. Ce dernier fournit finalement une réponse sous la forme d'une configuration d'activation des neurones de sortie. L'ajustement du modèle (phase d'apprentissage dans la terminologie neuronale) s'effectue en général de manière progressive, par présentation répétée des données au système et correction partielle des paramètres (poids dans la terminologie neuronale) jusqu'à la satisfaction d'un critère d'arrêt fondé par exemple sur l'adéquation des prédictions données par le système et des valeurs souhaitées. Dans la mesure où le système est très fortement surparamétré, il est souvent possible d'obtenir un très bon apprentissage de l'ensemble des données qui lui sont présentées. Cependant, aucune garantie n'est offerte concernant les performances pour un autre ensemble de données similaires. Le critère d'arrêt est donc appliqué sur un ensemble de données indépendant de celui utilisé pour l'apprentissage. Il est courant d'observer, dans une première phase (phase de généralisation), que la qualité de l'ajustement sur cet ensemble de données augmente au cours de l'apprentissage puis qu'elle se dégrade dans une deuxième phase (apprentissage par cœur). La décision de l'arrêt est alors facile à prendre.

L'apprentissage dit *supervisé* est mis en place lorsque la sortie souhaitée est connue. C'est le cas en discrimination ou en régression. L'erreur de prédiction est alors utilisée pour améliorer l'ajustement. Le *Perceptron* est l'exemple type de réseau fonctionnant de la sorte. Il est composé de plusieurs couches de neurones (couches d'entrée, cachées, de sortie). L'apprentissage, qui s'effectue par l'algorithme dit de rétropropagation du gradient (Le Cun, 1985; Rumelhart, *et al.*, 1986), est en général plus rapide qu'en mode *non supervisé*. Dès lors qu'il existe une information pertinente dans les données, les performances sont excellentes. Ces réseaux sont couramment utilisés pour des tâches d'identification. En ce qui concerne les problèmes que nous avons abordés dans ce travail, nous avons choisi de mettre en place un réseau de ce type.

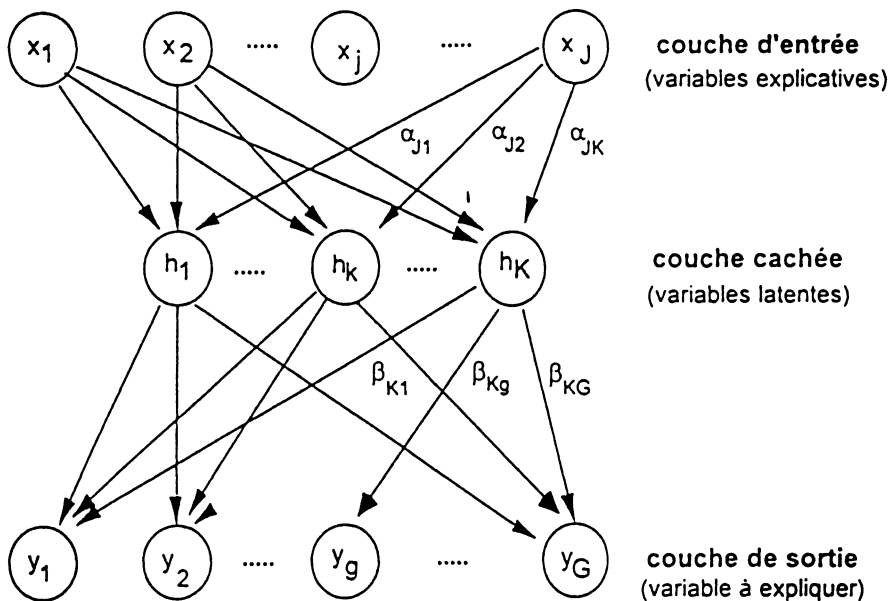
Représentation et fonctionnement d'un neurone

Les unités de base d'un réseau sont des neurones. A l'image des neurones biologiques, ils ont plusieurs entrées où viennent se connecter d'autres neurones et une ou plusieurs sorties qui leur permettent d'activer d'autres neurones. Ainsi, ils sont interconnectés pour former un réseau. La liaison entre deux neurones est caractérisée par un poids synaptique qui quantifie le degré d'interaction entre eux. Si le poids est positif, la liaison est dite activatrice; s'il est négatif, elle est inhibitrice. L'état du neurone est défini par son activation, calculée à partir de la somme pondérée (par les poids synaptiques) des activations des neurones incidents, par l'intermédiaire d'une fonction de transfert en général non linéaire. Cette activation est par la suite propagée de la même manière aux neurones suivants.

Structure du réseau

Le réseau est du type multicouches : Une couche d'entrée, une ou plusieurs couches cachées, une couche de sortie. Les neurones d'une couche donnée sont exclusivement connectés à ceux de la couche suivante. Le nombre de neurones par couche est variable. Il dépend du problème à traiter. Le nombre de neurones des couches d'entrée et de sortie dépend du format et du type des données. En général une structure en diabololo est retenue, l'importance du rétrécissement au niveau des couches cachées déterminant le degré de généralisation souhaité lors de l'apprentissage.

Architecture du réseau de neurones



Le réseau de neurones utilisé pour traiter les données en question a la configuration ci-dessus. Dans ce réseau, la valeur prise par la variable latente h_k de la couche cachée est obtenue à partir de la combinaison linéaire

$$f_k(x) = \sum_{j=1}^j \alpha_{jk} x_j$$

des observations d'un individu x en utilisant la fonction logistique

$$h_k = 1/[1 + \exp - f_k(x)]$$

dont les valeurs sont comprises entre 0 et 1 ; elles varient fortement au voisinage de $h_k(x) = 1/2$ et tendent asymptotiquement vers 1 ou 0 lorsque $f_k(x)$ tend vers plus ou moins l'infini. Il en est de même des valeurs de la variable à expliquer en sortie qui sont des fonctions des variables de la couche cachée

$$y_g = 1/[1 + \exp - \sum_{k=1}^K \beta_{kg} h_k]$$

Il s'agit d'estimer les différents coefficients α_{jk} et β_{kg} connus sous le nom de *poids synaptiques*.

Pour un individu de l'échantillon d'apprentissage le réseau fournit une valeur estimée de chaque neurone de sortie qui est comparée à la valeur observée correspondante. On en déduit un écart, fonction des poids synaptiques qui sont ensuite ajustés. La méthode habituellement utilisée pour ajuster ces poids est la rétropropagation de l'erreur par la méthode du gradient. L'algorithme de rétropropagation des réseaux multicouches vise à minimiser le carré des écarts constatés entre les valeurs attendues et les valeurs estimées à la sortie du réseau. Les écarts sont répercutés de neurone en neurone, depuis la sortie du réseau jusqu'au neurone de la couche cachée.

Remarque : dans un contexte de régression, on considère un seul neurone en sortie dont la valeur estimée pour un individu est comparée à sa valeur observée après transformation de la variable à expliquer en variable dont les valeurs sont comprises dans l'intervalle $[0,1]$.

Codage des données

La représentation des données conditionne la réussite de l'étude. A chaque neurone de la couche d'entrée, il convient d'associer une caractéristique des données à analyser (pixel pour une image, valeur d'un paramètre pour des données multidimensionnelles). Par construction, l'activation des neurones de sortie est bornée (en général $[0,1]$). Il est souhaitable d'utiliser au mieux toute la dynamique offerte par ces neurones pour représenter la réponse attendue.

Mode d'apprentissage

L'algorithme de rétropropagation du gradient est la méthode de choix dans ce contexte. Au cours de la phase d'initialisation, les poids des liaisons entre les neurones

sont choisis au hasard. Puis les données sont présentées au réseau. L'apprentissage est effectué successivement pour chaque donnée, une faible correction étant apportée à chaque poids pour améliorer la prédiction. Toutes les données étant passées en revue, une nouvelle session d'apprentissage est engagée, et ainsi de suite, jusqu'à satisfaction du critère d'arrêt. On obtient ainsi un ensemble de poids qui sont une représentation de l'information acquise au cours de l'apprentissage.

Mode d'exploitation

En phase d'exploitation, les données sont codées pour être présentées à l'entrée du réseau; l'activation des neurones est propagée de l'entrée vers la sortie en utilisant le fichier de poids sélectionné. L'activité des neurones de sortie correspond à la réponse du réseau.

Applications

En ce qui concerne les problèmes que nous avons abordés dans ce travail, nous avons choisi de mettre en place des réseaux multicouches, l'apprentissage étant effectué par *l'algorithme de rétropropagation du gradient*, donc en mode supervisé.

Principales particularités de la méthode neuromimétique utilisée

Le modèle est ininterprétable en ce sens que les paramètres (poids reliant les neurones entre eux) n'ont pas *a priori* un rapport direct avec le phénomène. Il est difficile de savoir sur quelles informations se base le réseau pour effectuer des prédictions.

L'obtention de la règle de décision n'est pas rapide (de l'ordre d'une journée pour les problèmes considérés dans cette étude et sur un ordinateur de bureau). Des techniques d'ajustement adaptées au type de problème posé, peuvent réduire le temps de calcul d'un facteur 10 (Perantonis et Karras, 1995; SPAD.N : procédure NEURO, 1993).

L'utilisation de la méthode demande connaissance et maîtrise de l'outil. Le choix de l'architecture et du nombre de neurones cachés, de la fonction objet à minimiser, des paramètres de convergence pour la méthode itérative d'ajustement utilisée est fondé en grande partie sur des considérations empiriques.

Le nombre important de paramètres à ajuster réclame un échantillon de données conséquent : soit I le nombre d'individus, J le nombre de variables explicatives, H le nombre de neurones dans la couche cachée unique et K le nombre de modalités de la variable à expliquer. Alors le nombre de paramètres P est $(J + 1)H + (H + 1)K$. Le nombre 1 provient de l'introduction d'un terme constant dans la combinaison linéaire utilisée comme fonction d'entrée des neurones. Généralement J et H sont nettement supérieurs à 1 et K est inférieur à J . Dès lors $P \cong H(J + K) \cong HJ$. Le nombre de paramètres est proportionnel au nombre de variables explicatives en tenant compte du codage des variables catégorielles. On considère généralement qu'un apprentissage efficace nécessite un nombre d'individus 2 à 4 fois plus élevé que le nombre de paramètres à estimer (Master, 1993). Par ailleurs H est inférieur à 10 en pratique; I est donc de l'ordre de $40J$.

En raison de l'aspect stochastique de l'apprentissage, la règle de décision diffère d'un essai à l'autre, tout en conservant une qualité du même ordre (Fertil et Vilain, 1994).

Dans le cadre d'une discrimination, si on modifie les fonctions coûts ou les probabilités *a priori*, il faut refaire l'ajustement car le réseau de neurones n'indique que la valeur estimée de la variable à expliquer. Il est cependant possible de demander expressément au réseau de fournir les probabilités *a posteriori* (Gernoth et Clark, 1995; Jepson *et al.*, 1993; Martin, *et al.*, 1994; Nix et Weigend, 1994; Srivastava et Weigend, 1994).

La différence entre l'erreur apparente obtenue avec le fichier d'apprentissage par la méthode de resubstitution, et l'erreur réelle obtenue avec un fichier test peut être importante. Non seulement le modèle intègre les fluctuations d'échantillonnage, mais comme il est surparamétré, il peut, en faisant de l'apprentissage par cœur, réduire considérablement l'erreur apparente. Il convient donc de repenser les modalités d'applications des techniques classiques de validation : validation croisée, Jackknife et Bootstrap (Katz *et al.*, 1994; Liu, 1995; Paass, 1994) pour les utiliser avec les techniques neuronales.

Toute information fournie par les variables explicatives est susceptible d'être utilisée. Une variable qualitative est transformée en variables binaires [0-1], elles-mêmes traitées comme variables quantitatives par le réseau. Le recodage des variables est très peu contraignant. Aucune propriété particulière concernant la distribution des variables n'est recherchée. Seul un large étalement des variations entre 0,2 et 0,8 est choisi afin d'accélérer le processus d'ajustement. La méthode accepte des variables explicatives ayant une probabilité de distribution quelconque de variation non monotone avec la variable à expliquer. Elle peut aussi intégrer des interactions d'ordre élevé entre prédicteurs (exemple du problème des deux spirales (Mascioli *et al.*, 1994)). La sélection des variables explicatives, souvent nécessaire pour optimiser les performances de la règle de décision dans les méthodes classiques, est remplacée par celui plus simple, de la détermination de l'architecture du réseau, à partir d'un fichier de validation temporaire.

La méthode est robuste devant les données manquantes imprécises ou erronées. En présence d'une donnée aberrante, le réseau est capable d'allouer quelques ressources pour l'expliquer sans pour autant que les performances sur les autres données soient dégradées.

Dans le contexte d'un classement, en utilisant le fichier d'apprentissage, le réseau de neurones indique la densité de probabilité *a posteriori* pour chaque groupe, après recodage des résultats ($> 0,8$ donne 1, $< 0,2$ donne 0, et $[0,2-0,8]$ donne $[0-1]$), grâce au processus de minimisation de la somme des carrés des erreurs. Cependant ce minimum est atteint dans l'apprentissage si et seulement si la distribution des résultats est identique à celle des densités de probabilités *a posteriori* (Ruck *et al.*, 1990).

La méthode de segmentation CART

CART, sigle pour Classification And Regression Trees, est une procédure statistique non-paramétrique introduite par Breiman, Friedman, Olshen et Stone en

1984 (Breiman, *et al.*, 1984). Elle est utilisée pour analyser la relation entre une variable Y (variable dépendante ou variable à expliquer) et un ensemble de p variables X_j (variables indépendantes ou variables explicatives) de nature quelconque. On pourra se référer au livre de G. Celeux et J.-P. Nakache (1994) pour tous les détails théoriques de la méthode CART vue sous l'angle discrimination.

CART répond à un problème de discrimination ou de régression selon que la variable Y est qualitative ou quantitative. Les résultats fournis par CART se présentent sous la forme d'arbres obtenus par *partitionnement récursif binaire*. Le processus est binaire parce que les nœuds parents sont toujours divisés en deux nœuds descendants et récursif, parce que ce processus est répété en considérant chaque nœud comme un nœud parent.

Un individu, parcourant l'arbre depuis sa racine, tombe dans un nœud terminal et un seul suivant les valeurs qu'il présente pour les variables de division des nœuds intermédiaires.

Dans un contexte de discrimination un nœud terminal de l'arbre (*arbre de décision*) est affecté à une classe. Si la variable à expliquer est quantitative, une valeur prédite de Y est associée à chaque nœud terminal de l'arbre (*arbre de prédiction*).

CART n'est pas la seule technique basée sur une structure d'arbre. Les principales méthodes développées avant CART, comme AID (Morgan et Sonquist, 1963), THAID (Morgan et Messenger, 1972) et CHAID (Kass, 1980), sont basées sur la recherche d'une règle d'arrêt de division la plus adéquate, ce qui produit généralement des résultats insatisfaisants. Une innovation majeure de CART a été d'éliminer toute règle d'arrêt de division et de poursuivre par conséquent le plus longtemps possible la division des nœuds.

En présence d'un problème de discrimination, nous avons à différencier plusieurs groupes ou classes. L'idée est de sélectionner chaque division de telle sorte que les données de chacun des deux nœuds descendants soient «plus purs» que les données dans le nœud parent. Cette sélection est basée sur la notion d'*impureté d'un nœud*. Aucun critère n'est utilisé pour arrêter la division des nœuds, ce qui conduit à l'obtention d'un grand arbre (nœuds terminaux en nombre important). Ce grand arbre est ensuite successivement élagué en coupant les branches non informatives. Une *séquence de sous-arbres emboîtés*, dont le premier est le grand arbre et le dernier la racine, est ainsi construite.

CART utilise finalement, soit la méthode de l'échantillon test, soit la méthode de validation croisée (suivant la taille de l'échantillon) pour, d'une part estimer le taux d'erreur associé à chaque sous-arbre de la séquence, et d'autre part sélectionner le «meilleur» sous-arbre qui est basé sur *l'erreur* et sur une *pénalité* appliquée aux grands arbres.

Ayant obtenu le «meilleur» sous-arbre, le problème est alors de déterminer le classement de toute observation qui, parcourant l'arbre depuis sa racine, tombe dans un des nœuds terminaux. La règle de décision la plus simple consiste à affecter cette observation au groupe le plus représenté dans ce nœud. Cette règle est bien entendu valable si les probabilités *a priori* sont estimées par les fréquences des groupes dans l'échantillon et si de plus, les coûts de mauvaise classification sont pris égaux à 1. Il n'y a, d'autre part, aucune difficulté à construire cette règle dans le cas général (Breiman *et al.*, 1984; Gueguen et Nakache, 1988).

Si la variable à expliquer est quantitative, il s'agit de prédire la valeur numérique de cette variable pour un individu donné. La technique est rigoureusement identique à celle de CART utilisée dans un contexte de discrimination. La seule différence réside dans le choix du critère de division d'un nœud qui est, ici, basé sur la *variance-intra nœuds descendants* de la variable Y à expliquer (Breiman *et al.*, 1984).

Ayant obtenu le «meilleur» sous-arbre de régression, la règle de prédiction de la valeur Y pour un individu est simple : l'individu, qui parcourt l'arbre depuis sa racine, tombe dans un des nœuds terminaux. Sa valeur prédite de Y est la moyenne des valeurs de Y des individus de l'échantillon d'apprentissage inclus dans ce nœud terminal.

Principales particularités de la méthode CART

La procédure CART présente les avantages suivants : les variables explicatives, qui peuvent être de n'importe quelle nature (qualitatives, binaires ou quantitatives), sont utilisées sans codage préalable. Les individus présentant des données manquantes ne sont pas exclus de l'analyse. Les *variables suppléantes* (Breiman, *et al.*, 1984, Gueguen et Nakache, 1988) permettent d'autre part, une affectation à un groupe ou une prédiction de la variable à expliquer d'un nouvel individu dont les données sont manquantes pour certaines des variables de division des nœuds de l'arbre. La procédure est robuste vis-à-vis de quelques données aberrantes ou erronées. Les résultats obtenus sont invariants par rapport à des transformations monotones des variables explicatives. Enfin dans un contexte de discrimination il est possible, sans aucune difficulté, de fixer les *probabilités a priori* d'appartenance aux différents groupes ainsi que les *coûts de mauvaise classification*, et ce dans le but d'attribuer, si nécessaire, des poids différents aux groupes.

Evaluation de la qualité de la règle de décision ou de la règle de prédiction

Qualité de la règle de décision

En discrimination, une fois construite la règle de décision, il est important d'en évaluer la performance (ou qualité), c'est-à-dire le risque qui lui est associé, ou encore l'erreur qu'elle produira lorsqu'elle sera appliquée à de nouvelles observations.

Dans les applications présentées dans cet article, les probabilités *a priori* d'appartenance aux différents groupes ont été estimées par les fréquences des groupes dans l'échantillon servant à cette évaluation, et de plus, les coûts de mauvaise classification ont été pris égaux à 1. Sous cette hypothèse, l'estimation du risque est égale au pourcentage d'observations de l'échantillon d'évaluation mal classées par la règle de décision.

Qualité d'une discrimination fondée sur l'adéquation des probabilités a posteriori

Il existe en pratique des exemples où le point de vue purement décisionnel n'est pas adéquat. Par exemple dans un problème de diagnostic médical, au lieu

de déterminer la classe diagnostique à laquelle est affectée un patient, il peut être préférable de fournir uniquement les estimations des probabilités *a posteriori* d'appartenance aux différentes classes diagnostiques. Dans ce cas, il n'y a pas de règle de décision proprement dite, la discrimination consiste à fournir, pour chaque observation de l'échantillon E servant à l'évaluation, le vecteur

$$\{\hat{P}(G_1/x_i), \dots, \hat{P}(G_k/x_i)\}$$

décrivant les k probabilités *a posteriori* d'appartenance aux groupes.

La mesure de la qualité de la discrimination peut être estimée par *l'écart quadratique moyen* (Breiman *et al.*, 1984; Gueguen *et al.*, 1996). Si les probabilités *a priori* sont estimées par les fréquences des groupes de l'échantillon E , l'expression de ce score est :

$$SQM = \frac{1}{N} \sum_{i \in E} \sum_{s=1}^k [Z_{is} - \hat{P}(G_s/x_i)]^2$$

où N est la taille de E et $z_{is} = 1$ si le groupe d'origine de x_i est G_s et $z_{is} = 0$ sinon.

Remarque : On pourra trouver des détails théoriques sur l'estimation du risque d'une règle de décision et du score quadratique moyen ainsi que les estimations de leurs variances respectives, dans l'article de Gueguen *et al.* (1996).

Qualité d'une règle de prédiction

En régression, comme en discrimination, il est tout aussi important d'évaluer la performance de la règle de prédiction de la valeur de la variable à expliquer quand elle est appliquée à de nouvelles observations constituant un échantillon test E . Une estimation de la qualité de la régression est généralement fournie par :

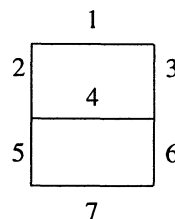
$$ERP = \frac{\sum_{i \in E} (y_i - \hat{y}_i)^2}{\sum_{i \in E} (y_i - \bar{y})^2}$$

qui représente la part de la variance de Y dans l'échantillon non expliquée par la régression. Cette estimation de la qualité de la régression est aussi connue sous le nom d'estimation de *l'erreur relative de prédiction* (CART, 1984).

Application aux données DIGITS

Ces données concernent les nombres (*digits*) de 0 à 9 apparaissant sur les montres ou les écrans d'ordinateurs. Chaque nombre est caractérisé par un vecteur $(x_1, \dots, x_j, \dots, x_7)$ à 7 dimensions composé de 0 et de 1. Les différents profils sont les suivants :

Digit	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1
0	1	1	1	0	1	1	1



Il s'agit de données bruitées générées par simulation avec une introduction d'erreurs dans le profil de chaque digit : chaque observation x_j a une probabilité égale à 0,1 d'être changée en son contraire. 200 profils (20 par digit), ainsi obtenus, constituent l'échantillon d'apprentissage divisés en 10 groupes (digits) qu'il est question de reconnaître. D'autre part, pour évaluer la règle discriminante construite à partir de cet échantillon d'apprentissage, un échantillon de 2000 profils (200 par digit) a été généré de la même façon.

Méthode neuronale

Dans le schéma du réseau de neurones construit à partir de l'échantillon d'apprentissage, la couche d'entrée comprend 7 neurones correspondant aux 7 variables binaires de chacun des profils de l'échantillon d'apprentissage. La sortie est codée sur 10 neurones, chacun d'eux représente une classe ou digit. Pour l'apprentissage, l'appartenance à la classe correspond à une activation de 0,8 pour le neurone concerné, 0,2 pour les autres. En phase d'exploitation, à chaque individu testé correspond une distribution d'activation parmi les neurones de sortie. La classe est déterminée par le neurone d'activité la plus grande. Plus la valeur de l'activation est proche de 0,8 et/ou plus l'écart avec les autres activations est grand, plus la prédiction est sûre.

Le nombre de neurones de la couche cachée doit être déterminé de manière empirique. En général, l'erreur sur le fichier d'apprentissage diminue lorsque le nombre de neurones cachés augmente. Cependant, dans le cas présent, l'indétermination structurelle des données (il existe des individus de classes différentes qui ne sont pas séparables) conduit à une limite minimale de cette erreur (17%). Celle-ci est pratiquement atteinte dès que le nombre de neurones cachés est supérieur ou égal à 9. Il n'est donc pas nécessaire d'utiliser un fichier de validation.

Le tableau 1 fournit le résumé du classement des 2 000 observations de l'échantillon d'évaluation. L'erreur de classement est de 27,05% avec un écart-type de 0,99% et la valeur du score quadratique moyen est de 0,4218.

Remarque : Ripley (1993) présente dans son article plusieurs exemples dont une comparaison sur les digits entre CART et un type de réseau de neurones identique au nôtre (MLP), mais avec des résultats différents : 35% d'erreurs avec un fichier de 200 individus bruités à 10%. Il semble *a priori* que le nombre de neurones cachés

TABLEAU 1
Méthode neuronale appliquée aux données DIGITS.
Classement des 2 000 observations de l'échantillon d'évaluation
(27,05% MC $\pm$ 0,99%; SQM = 0,4218)

		Affectation										
		1	2	3	4	5	6	7	8	9	0	Total
Origine	1	151	0	1	15	0	1	20	0	0	12	200
	2	0	124	12	2	5	22	0	17	0	18	200
	3	2	19	127	2	10	2	19	2	12	5	200
	4	17	1	0	158	2	2	4	13	1	2	200
	5	0	0	2	10	154	21	2	1	9	1	200
	6	0	0	0	0	14	146	0	10	1	29	200
	7	15	0	3	7	1	1	168	0	1	4	200
	8	0	17	1	2	3	16	2	138	9	14	200
	9	0	2	4	26	19	1	3	12	119	14	200
	0	0	1	1	0	6	4	2	11	1	174	200
												2000

choisi soit trop élevé, puisque 20 neurones cachés ont été utilisés au lieu de 9, ce qui conduit à un rapport du nombre d'individus au nombre de paramètres trop faible ($200/370 \cong 0,55$).

Méthode CART

La figure 1 représente l'arbre de décision obtenu à partir de l'échantillon d'apprentissage. Ce sous-arbre retenu, après élagage et utilisation de la méthode de validation croisée, contient 10 nœuds terminaux. Le digit le plus représenté et sa fréquence sont mentionnés à l'intérieur de chaque nœud terminal. Pour chaque nœud intermédiaire, la figure 1 fournit son nombre d'observations ainsi que la variable ayant servi à le diviser. Le nombre d'observations de chaque nœud terminal est indiqué sous le nœud.

Le tableau 2 présente le résumé du classement, par la règle de décision associée à l'arbre, des 2 000 observations de l'échantillon d'évaluation. La qualité de cette règle de décision est appréciée par une estimation de l'erreur de classement égale à 30,15% avec un écart-type de 1,02% et une valeur du score quadratique moyen égale à 0,5167.

Comparaison des résultats obtenus par les deux méthodes

Ces résultats montrent que la plus petite valeur du score quadratique est obtenue par la méthode neuronale. D'autre part, la comparaison des estimations des erreurs de

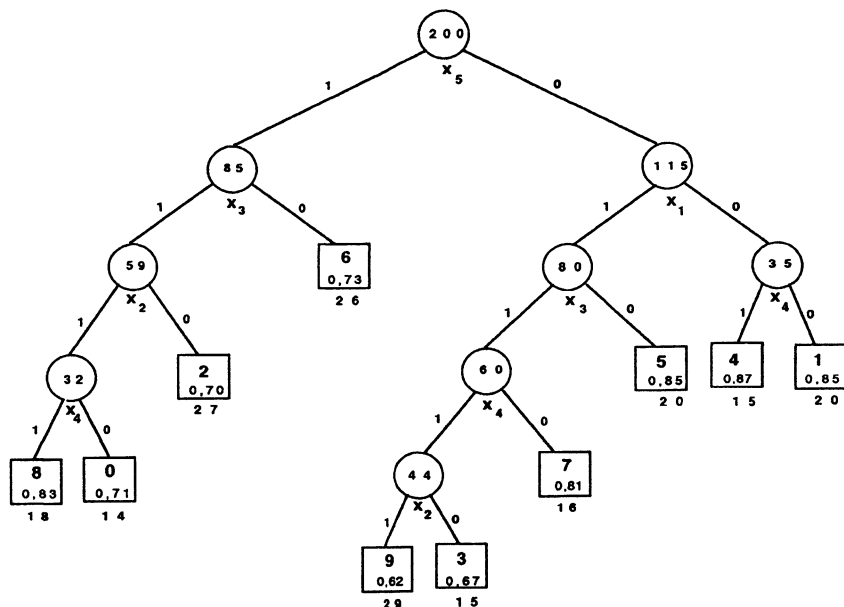


FIGURE 1
*Arbre de décision obtenu à partir de l'échantillon
d'apprentissage des données DIGITS*

TABEAU 2
Méthode CART appliquée aux données DIGITS
Classement des 2 000 observations de l'échantillon d'évaluation
(30,15% MC $\pm$ 1,02%; SQM = 0,5167)

		Affectation										Total
		1	2	3	4	5	6	7	8	9	0	
Origine	1	144	18	2	15	0	4	17	0	0	0	200
	2	0	134	13	2	5	23	3	19	0	1	200
	3	1	22	124	11	11	1	16	2	12	0	200
	4	16	1	2	148	1	3	2	14	12	1	200
	5	0	0	1	25	142	9	1	1	11	0	200
	6	0	0	0	0	14	171	0	10	1	4	200
	7	14	18	14	2	16	2	128	0	3	3	200
	8	0	17	1	0	3	17	2	141	11	8	200
	9	0	2	4	15	19	1	15	12	130	2	200
	0	2	10	1	0	6	20	14	11	1	135	200
												2000

classement au moyen du test de MacNemar (Siegel, 1956) de comparaison de deux proportions appariées (tableau 3) conduit à une différence hautement significative entre elles, avec une plus petite erreur pour la méthode neuronale.

TABLEAU 3
Comparaison des % MC obtenus par les méthodes CART et NEURONES
appliquées aux données DIGITS - ($\chi^2 = 18,45$) à 1 ddl)

		Méthode CART		
		MC	BC	
Méthode neuronale	MC	468	73	541
	BC	135	1324	1459
		603	1397	2000

Application aux données DAA

Ces données concernent des patients présentant des douleurs abdominales aiguës. Elles ont été recueillies par les chirurgiens de l'association ARC-AURC (Flamant, *et al.*, 1981). Ces patients sont divisés en 6 groupes correspondant à 6 diagnostics. Les données ont été divisées aléatoirement en deux échantillons : un échantillon d'apprentissage (EA) de 3 114 patients utilisé pour construire la règle de décision, et un échantillon d'évaluation (EE) de 1 427 patients (30% de patients tirés au hasard de l'échantillon total) utilisé pour évaluer la règle. Les six groupes et leurs tailles respectives sont :

	EA	EE
Appendicites	1 060	483
Cholecystites	387	231
Colique néphrétiques	141	55
Douleurs non spécifiques	964	432
Péritonites	224	81
Occlusions	338	145

Ces données concernent les mesures de 54 variables qualitatives (âge en classes, sexe, 44 variables cliniques, 6 variables radiologiques et 2 variables biologiques) relevées sur chacun des patients. Tous les détails concernant ces données peuvent être trouvés dans l'article de Flamant *et al.* (1981) et celui de Séroussi *et al.* (Seroussi; ARC et AURC, 1985).

Méthode neuronale

En ce qui concerne l'application des réseaux de neurones à ces données, un neurone a été associé à chacune des variables binaires, et un neurone à chacune des

modalités des variables à plus de 2 modalités, soit finalement 65 neurones d'entrée. Six neurones de sortie (un neurone par groupe diagnostic) ont été utilisés pour prédire le diagnostic. L'échantillon de 3 114 patients servant à construire la règle de décision a été ici divisé en un fichier d'apprentissage comprenant 2 075 sujets et un fichier de validation de 1 039 sujets. La règle de décision construite à l'aide de ces deux fichiers, a été testée sur les 1 427 sujets de l'échantillon d'évaluation en vue de mesurer la performance de cette règle.

Le traitement de l'activation des neurones, en phase d'apprentissage comme en phase d'exploitation est identique à l'étude précédente. De même il convient de choisir le nombre de neurones de la couche cachée. Nous avons fait varier ce nombre de 1 à 16. Le meilleur résultat correspondant à chaque cas a été déterminé à l'aide du fichier de validation. Quatre neurones cachés suffisent pour cette étude.

Méthode CART

Pour ce qui est de la méthode CART, elle a été appliquée, comme dans l'exemple précédent, (i) en estimant les probabilités *a priori* d'appartenance aux différents groupes par les fréquences des groupes dans l'échantillon de base, et (ii) en prenant des coûts de mauvaise classification égaux à 1. La règle ainsi construite a été évaluée en l'appliquant aux sujets de l'échantillon d'évaluation des 1 427 sujets.

Comparaison des résultats obtenus par les deux méthodes

Les tableaux 4 et 5 fournissent le résumé du classement de ces sujets par les deux méthodes. La méthode neuronale conduit à une estimation de l'erreur de classement égale à 28,45% avec un écart-type de 1,19% et à un score quadratique moyen égal à 0,4144. Pour la méthode CART on obtient une erreur de 35,18% avec un écart-type de 1,26% et un score quadratique moyen égal à 0,5097.

TABLEAU 4
Méthode CART appliquée aux données DAA.
Classement des 1 427 sujets de l'échantillon d'évaluation
(35,18% MC $\pm$ 1,26%; SQM = 0,5097)

	Affectation						Total
	Append	Chol.	Neph.	DNS	Périt.	Occl.	
Append	340	5	0	125	1	12	483
Chol.	32	152	0	26	1	20	231
Neph.	7	3	27	16	0	2	55
Origine DNS	116	36	1	261	0	18	432
Périt.	19	5	0	3	48	6	81
Occl.	21	8	0	19	0	97	145
							1427

TABLEAU 5
Méthode neuronale appliquée aux données DAA.
Classement des 1 427 sujets de l'échantillon d'évaluation
(28,45% MC $\pm$ 1,19%; SQM = 0,4144)

		Affectation						
		Append	Chol.	Neph.	DNS	Périt.	Occl.	Total
Origine	Append	367	6	1	97	2	10	483
	Chol.	11	169	0	37	1	13	231
	Neph.	4	0	37	13	0	1	55
	DNS	119	32	2	263	0	16	432
	Périt.	10	2	0	2	61	6	81
	Occl.	8	3	0	10	0	124	145
								1427

TABLEAU 6
Comparaison des % MC obtenus par les deux méthodes CART et NEURONES
appliquées aux données DAA - ($\chi^2 = 31,15$ à 1 ddl)

		Méthode CART		
		MC	BC	
Méthode neuronale	MC	290	116	406
	BC	212	809	1021
		502	925	1427

Comme pour les données DIGITS, les meilleurs résultats sont fournis par la méthode neuronale tant du point de vue du score quadratique moyen que de la comparaison des erreurs de classement par le test de Mac Nemar (Tableau 6).

Application aux données HOUSING

Il s'agit de données recueillies par Harrison et Rubinfeld (Harrison et Rubinfeld, 1978). Elles concernent le prix médian des logements (en milliers de dollars) situés dans différents quartiers de Boston (506 au total). Ce prix représente la variable Y à expliquer par 13 variables indépendantes caractérisant le quartier : taux de criminalité (CRIM); % de la superficie du quartier occupé par les logements (ZN); % de grandes industries (IND); traversée du quartier par la rivière Charles (CHAS); concentration en oxide nitrogénique (NOX); nombre moyen de chambres des logements (RM); % de logements construits avant 1940 (AGE); distance au centre administratif (DIS); accessibilité aux autoroutes (RAD); taux d'imposition (TAX);

rapport élève/enseignant (P/T); % de population noire (B); % de population de faible niveau social (LSTAT).

L'échantillon total de 506 observations a été divisé aléatoirement en un échantillon d'apprentissage (337 observations) utilisé pour construire la règle de prédiction et un échantillon d'évaluation (169 observations) servant à mesurer la performance de cette règle. Il s'agit, dans le cas de ces données, d'effectuer une méthode de régression, et par conséquent, de construire une règle de prédiction de la valeur de la variable à expliquer Y pour une nouvelle observation.

Méthode neuronale

L'activation des neurones étant limitée entre 0 et 1, chaque valeur a été recodée afin d'obtenir approximativement une variation entre 0,2 et 0,8, laissant ainsi au neurone une marge de débordement, soit $0,2 + 0,6(x - x_{\min})/(x_{\max} - x_{\min})$. L'ajustement des paramètres du modèle est facilité si la distribution de chaque variable est étalée. D'où le codage de deux variables «taux de criminalité» = $1n(\text{CRIM})/10$ et «% de population noire» = $1n(65 - B)$.

Chacune des 13 variables indépendantes explicatives est représentée par un neurone, de même que la variable à expliquer. Le nombre de couches cachées est de un, comme c'est généralement le cas. Un choix supérieur ne s'impose que dans les situations plus complexes, par exemple en présence de plusieurs variables à expliquer.

L'échantillon d'apprentissage a d'abord été divisé aléatoirement en deux (2/3, 1/3) pour réserver un échantillon de validation temporaire (112 individus), utilisé dans le calcul du critère d'arrêt d'ajustement. Le nombre de neurones cachés a varié de 6 à 30 en 5 points : 6, 10, 14, 20 et 30. La valeur de 20 correspond à l'erreur quadratique moyenne minimum pour le fichier de validation temporaire. Un excès de neurones cachés, ce qui semble être le cas, n'est pas critique si la durée d'apprentissage a été limité par un fichier de validation. Le même rôle peut être joué par deux ou plusieurs neurones. L'ajustement a été refait avec le fichier complet d'apprentissage avec un nombre équivalent d'itérations.

Méthode CART

L'arbre binaire de régression, qui fait l'objet de la figure 2, a été construit à l'aide de 337 observations de l'échantillon d'apprentissage. Cet arbre, retenu après élagage et validation croisée, contient 13 nœuds terminaux.

Comparaison des résultats obtenus par les deux méthodes

Le réseau de neurones, appliqué à l'échantillon d'évaluation, fournit une estimation de la mesure de la qualité de la règle de prédiction ERP égale à $7,14/85,124 = 0,084$. La même mesure pour le réseau composé de 6 neurones cachés, ajusté avec un échantillon d'apprentissage réduit, est de $11,2/85,124 = 0,132$.

La mesure de la qualité de la règle de prédiction fournie par l'arbre de régression et appliquée aux 169 observations de l'échantillon d'évaluation est égale $16,09/85,124 = 0,189$.

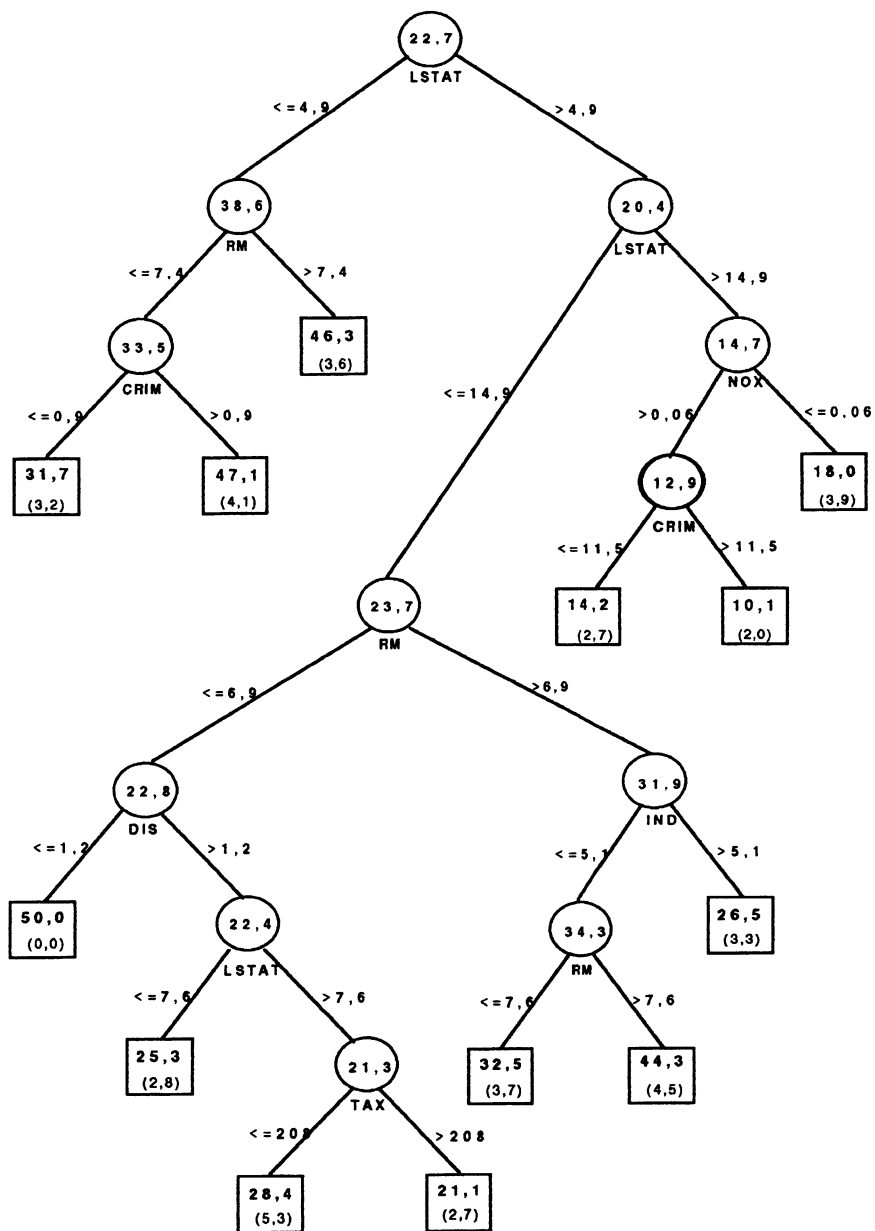


FIGURE 2
Arbre de décision obtenu à partir de l'échantillon
d'apprentissage des données HOUSING

La méthode bootstrap a été utilisée pour estimer la mesure de la qualité de la règle de prédiction. Dans ce but, 1 000 échantillons «bootstraps» ont servi à estimer cette mesure, chacun d'eux étant constitué de 169 observations tirées au hasard avec remise dans l'échantillon d'évaluation. Cette méthode fournit les estimations de la qualité de la règle de prédiction suivantes : pour la méthode neuronale $0,084 \pm 0,011$ et pour la méthode CART $0,188 \pm 0,022$.

Comme pour les deux premiers exemples, les résultats obtenus par la méthode neuronale sont les meilleurs en terme de qualité de la règle de prédiction.

Temps de calcul

Il faut distinguer les temps de calcul nécessaires pour l'obtention des règles des temps de réponse dans l'application de ces règles à de nouveaux cas. Alors que quelques minutes sont nécessaires pour obtenir les règles d'une analyse suivant la méthode CART, le temps de calcul requis pour l'application de la méthode neuronale se compte en jours. Plus précisément, les temps de mise en œuvre sur un macintosh Quadra 650 ont été les suivants :

- pour les données DIGITS en utilisant une carte accélératrice DSP, le cycle typique d'apprentissage a duré 2 heures et la détermination du nombre de neurones cachés a pris 24 heures,
- pour les données DAA, 3 heures ont été nécessaires pour le cycle typique d'apprentissage et 8 jours pour déterminer le nombre de neurones cachés,
- pour les données HOUSING, un cycle typique prend 24 heures. L'ensemble a demandé 6 jours de temps de calcul.

Par contre dans les deux méthodes, la prédiction ou la décision est obtenue instantanément pour un nouveau cas.

Logiciels

Les résultats de la méthode neuronale proviennent du programme BP, une implémentation de l'algorithme de rétropropagation effectué par McClelland et col. (McClelland et Rumelhart, 1988).

Les résultats de la méthode CART ont été obtenus par le programme SPAD.S (SPAD.S, 1993) qui comprend deux modules : DISAR et REGAR qui effectuent respectivement une analyse discriminante ou une régression par arbre.

Références bibliographiques

- BREIMAN L., FRIEDMAN J.H., OHLSEN R.A., STONE C. J., Classification and Regression Trees. Belmont, 1984.
- BURKE H.B., ROSEN D.B., GOODMAN P.H., Comparing Artificial neural networks to other statistical methods for medical outcome prediction. In proceeding : IEEE Int. Conference on Neural Networks, Orlando, Fl, p. 2213, 1994.

- CART, A software classification and regression trees. Yorshire Ct. Lafayette, California : California Statistical Software, inc., 1984.
- CELEUX J.P., NAKACHE J.-P., Analyse discriminante sur variables qualitatives. Polytechnica Ed, 1994.
- CICHOCKI A., UNBEHAUEN R., Neural Networks for Optimization and Signal Processing, p. 526. Stuttgart : Wiley, J. & Sons Ltd & Teubner, B.G., 1993.
- COUSTERE C., FERTIL B., Un nouvel outil pour l'analyse de données : les réseaux de neurones. Applications en microbiologie, *Bulletin de la société Française de Microbiologie* 7, 10, 1992.
- DAVALO E., NAIM P., Des réseaux de neurones, 2 édition, p. 232. Paris : Editions Eyrolles, 1993.
- FERTIL B., VILAIN J., Multiple Learning Sessions to Improve Predictions and Evaluate Reliability of Neural Networks. In proceeding : 4th International conference on Artificial Neural Networks, Sorrento, Italy, 1994, p. 1323.
- FLAMANT Y., LACAINE F., HAY J.M., MAILLARD J.N., Syndromes douloureux aigus de l'abdomen. Aide au diagnostic par ordinateur, *Nouvelle Presse Médicale* 10, 3367, 1981.
- GALLINARI P., THIRIA S., BADRAN F., FOGELMAN-SOULIE F., On the relations between discriminant analysis and multilayer perceptrons, in Neural Networks (USA), vol. 4, n°3, p. 349-69, 1991.
- GERNOTH K.A., CLARK J.W., Neural networks that learn to predict probabilities : Global models of nuclear stability and decay, *Neural Networks* 8, 291, 1995.
- GUEGUEN A., NAKACHE J.-P., Méthode de discrimination basée sur la construction d'un arbre de décision binaire, *Rev. Stat. Appl.*, 36, 19, 1988.
- GUEGUEN A., NICOLAU J., NAKACHE J.-P., Utilisation des réseaux probabilistes en analyse discriminante sur variables qualitatives, *Rev. Stat. Appl.*, 1996.
- HARRISON D., RUBINFELD D.L., Hedonic prices and the demand for clean air, *J. Envir. Econ. and Management* 5, 81, 1978.
- JEPSON B., COLLINS A., EVANS A., Post-neural network procedure to determine expected prediction values and their confidence limits, *Neural Computing & applications* 1, 224, 1993.
- KASS G.V., An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 29, 119, 1980.
- KATZ A.S., KATZ S., LOWE N., Fundamentals of the bootstrap based analysis of neural network's accuracy. In proceeding : WCNN, San Diego, USA, 1994, p. 673.
- LE CUN Y., Learning Scheme for asymmetric threshold networks. In proceeding : Cognitiva 85, Paris, France, p. 599, 1985.
- LEON M.A., Binary response forecasting : comparaison between neural networks and logistic regression analysis, in proc. of WCN 2, 244, 1994.
- LIU Y., Unbiased Estimate of Generalization Error and Model Selection in Neural Network, *Neural Networks* 8, 215, 1995.

- MARTIN C.E., ROGERS S.K., RUCK D.W., Neural network Bayes error estimation, *in proc. of IEEE ICNN* 305, 1994.
- MASCIOLI F.M.F., MARTINELLI G., LAZZARO D., comparison of Constructive Algorithms for Neural Networks, *in proc. of ICANN* 1, 731, 1994.
- MASTERS T., Practical neural network recipes in C++, p. 493. San Diego, CA, *Academic Press, Inc*, 1993.
- MASTERS T., Signal and image processing with neural networks, a C++ sourcebook. New York, Wiley & sons, inc., 1994.
- MCCLELLAND J.L., RUMELHART D.E., Explorations in parallel distributed processing. A handbook of models, programs and exercises, p. 344. Cambridge, MA, MIT Press, 1988.
- MICHIE D., SPIEGELMALTER D.J., TAYLOR C.C., Machine learning Neural and Statistical Classification, Ed. Ellis Horwood N.Y., 1994.
- MORGAN J. A., MESSENGER R.C., A modal search technique for predictive nominal scale multivariate analysis, *J. Amer. Statis. Ass.* 67, 768, 1972.
- MORGAN J.A., SONQUIST J.N., Problems in the analysis of survey data and a proposal, *J. Amer. Statist. Ass.* 58, 415, 1963.
- NAKACHE J.P., GOLMARD J.L., GUEGUEN A., Comparison of the performance of the conditional independence based model and the CART tree-structured discrimination model applied to a large medical data set. In proceeding : MIE 93, Jérusalem (Israël), 1993.
- NIX D.A., WEIGEND A.S., Estimating the Mean and Variance of the Target Probability Distribution, *in proc. of IEEE ICNN* 55, 1994.
- PAASS G., Assessing predictive accuracy by the bootstrap algorithm. In proceeding, 4th International conference on Artificial Neural Networks, Sorrento, Italy, p. 823, 1994.
- PERANTONIS S.J., KARRAS D.A., An efficient constrained learning algorithm with momentum acceleration, *Neural Networks* 8, 237, 1995.
- RIPLEY B.D., Statistical aspects of neural networks. In : O.E. Barndorff-Nielsen, J.L. Jensen, and W.S. Kendall (ed.), *Networks and Chaos – Statistical and Probabilistic Aspects*, p. 40, Chapman & Hall, 1993.
- RIPLEY B.D., Network methods in statistics. In : F.P. Kelly (ed.), *Probability, Statistics, Optimisation, a Tribute to Peter Whittle*, p. 241, Wiley, 1994a.
- RIPLEY B.D., Neural networks and flexible regression and discrimination. In : K.V. Mardia (ed.), *Advances in Statistics* 2, p. 39. Carfax : Abingdon, 1994b.
- RUCK D.W., ROGERS S.K., KABRISKY M., OXLEY M.E., SUTER B.W., The multilayer perceptron as an approximation to a Bayes optimal discriminant function, *IEEE Transactions on Neural Networks* 1, 296, 1990.
- RUMELHART D.E., HINTON G.E., WILLIAMS R.J., Learning internal representations by error propagation. In : D. E. Rumelhart and J.L. McClelland (ed.), *Parallel distributed processing : explorations in the microstructure of cognition*, Vol. 1, Foundations, p. 318. Cambridge, MA, MIT Press, 1986.

- SEROUSSI B., ARC et AURC, Comparison of discrimination methods; application to the acute abdominal pain diagnosis. In : D. Tfistsis (ed.), Lecture notes in médical informatics, Objective medical decision making, Vol. 28, p. 12 : Springer-Verlag, 1985.
- SIEGEL S., Nonparametric statistics for the behavioral sciences, McGraw-Hill Intern. Book Company, 1956.
- SHADMEHR R., D'ARGENIO D.Z., A comparison of a neural network based estimator and two statistical estimators in a sparse and noisy data environment, in proc. of IJCNN 1, 289, 1990.
- SPAD.N., Système portable d'Analyse des Données – procédure NEURO. Saint-Mandé, CISIA, 1993.
- SPAD.S., Système Portable d'Analyse des Données – Segmentation. Saint-Mandé : CISIA, 1993.
- SRIVASTAVA A.N., WEIGEND A.S., Computing the probability density in connectionist regression, in proc. of WCNN 2, 311, 1994.
- WEISBUCH G., Dynamique des systèmes complexes, p. 212. Paris, InterEditions & Editions du CNRS, 1989.