REVUE DE STATISTIQUE APPLIQUÉE

W. H. ZHU
C. GUICHENEY
J.-L. BERDAGUÉ
J. JOUSSET

Application des réseaux perceptron multicouches au contrôle de la qualité des aliments par analyse sensorielle. Comparaison des résultats avec différentes méthodes d'analyse discriminante

Revue de statistique appliquée, tome 45, n° 2 (1997), p. 39-57
<http://www.numdam.org/item?id=RSA_1997__45_2_39_0>

Rev. Statistique Appliquée, 1997, **XLV** (2), 39-57

APPLICATION DES RÉSEAUX PERCEPTRON MULTICOUCHES AU CONTRÔLE DE LA QUALITÉ DES ALIMENTS PAR ANALYSE SENSORIELLE – COMPARAISON DES RÉSULTATS AVEC DIFFÉRENTES MÉTHODES D'ANALYSE DISCRIMINANTE

W.H. Zhu (1), C. Guicheney (2), J.-L. Berdagué (3), J. Jousset (2)

(1) CISIA, 1 Avenue Herbillon, 94160 Saint-Mandé, France
Ceremade, Université de Paris IX Dauphine, 75775 Paris Cedex 16, France
(2) Laboratoire de Physique Corpusculaire de Clermont-Ferrand, IN2P3-CNRS,
Université Blaise Pascal, 63177 Aubière Cedex, France
(3) INRA de Theix, Station de Recherche sur la Viande,
63122 Saint-Genès-Champanelle, France

RÉSUMÉ

L'objectif de cet article est de présenter les performances des Réseaux Perceptrons Multicouches appliqués à un problème de caractérisation sensorielle de produits agro-alimentaires. Il s'agit de contrôler la qualité de différents aliments à partir de descripteurs provenant d'une analyse sensorielle.

Dans l'exposé, nous proposons plusieurs méthodes d'analyse discriminante pour essayer de construire le modèle le mieux adapté au problème. Puis nous comparons les résultats obtenus par les réseaux de neurones à ceux d'une analyse discriminant linéaire (AFD) et d'une segmentation (CART). On peut choisir une architecture ou «élaguer» à partir d'un réseau sur-dimensionné en mettant l'accent sur la technique de validation croisée qui permet de déterminer le réseau optimal. Dans l'exemple choisi, il apparaît clairement que les réseaux ont un pouvoir de classification supérieur à celui d'autres méthodes discriminantes.

Mots-clés : *analyse discriminante, réseau de neurones, segmentation, apprentissage, validation, validation croisée, classification, analyse des sensibilités.*

ABSTRACT

The main purpose of this article is to show the multilayered neural network performances applied to a sensory caracterization problem.

In order to design a tool for quality control of farm-products, a Fisher linear discriminant analysis, a non-parametric classification tree and a mulilayered neural network are studied in the same set of sensory data. The techniques and results are compared using the cross-validation strategy.

Keywords : *data analysis, neural networks, decision tree, training, validation, cross-validation, classification, sensitivity analysis.*

Dans cet article, nous allons réaliser sur des données réelles une comparaison des performances en reconnaissance de formes de différentes techniques d'analyse multidimensionnelles. Contrairement aux études théoriques qui ont été exposées récemment [Celeux 94, Lechevallier 95, Lebart 95], il s'agit ici de mettre en compétition différentes stratégies dans le domaine agroalimentaire. Notre objectif est d'illustrer les caractéristiques essentielles des modèles (linéaires ou non-linéaires, paramétriques ou non paramétriques) et particulièrement de préciser leurs champs d'application. Il s'agit également de mettre en perspective dans un environnement statistique réel (exploration de graphiques des résultats) l'utilisation de techniques complémentaires; et de donner des éléments sur les possibilités de leur utilisation à travers des logiciels statistiques répandus...; Mais il s'agit aussi de donner des compléments méthodologiques, d'insister sur leur portée pratique par des remarques s'appuyant sur une mise en œuvre réelle des techniques de discrimination.

Afin de remplir ces objectifs, nous avons sélectionné un jeu de données qui est assez typique de ce que l'on peut rencontrer en analyse sensorielle.

Avec ces données, nous allons étudier la qualité et la performance du modèle neuronal appliqué à un problème de caractérisation sensorielle de produits agro-alimentaires. Il s'agit de contrôler la qualité de différents aliments à partir de descripteurs provenant d'une analyse sensorielle. Nous nous proposons de réaliser, sur un même échantillon de données, une comparaison entre les performances de reconnaissance de formes du Perceptron simple, quadratique,ou des Réseaux de Perceptron Multicouches (RPM), avec celles de l'Analyse Factorielle Discriminante (AFD) et celles de la segmentation par arbre binaire [méthode CART].

1. Protocole des analyses sensorielles

Une série de trois fromages d'Emmenthal originaires de trois ateliers de fabrication différents a été dégustée lors d'une séance unique par 41 dégustateurs. Les évaluations sensorielles étaient de type profil utilisant une échelle de notation structurée à 5 niveaux. Trois séances de dégustation préalables avaient permis au jury de définir le vocabulaire utilisé lors de la dégustation, sachant que l'ensemble des juges appartenait déjà à la profession fromagère. Les descripteurs sensoriels sur lesquels ont porté les notations étaient les suivants :

– aspect olfactif : odeur d'Emmenthal (OEM), odeur rance (ORA), odeur piquante (OPI).

– aspect gustatif : intensité des caractères fruités (FRU), goût de noisette (NOI) et mauvais goût (MGO), (vocables très utilisés dans la profession fromagère et qui présentent une conotation hédonique certaine), piquant (PIQ) (sensation rapidement perçue sur la langue pendant la mastication), salé (SAL), sucré (SUC), amer (AME).

– aspect texture : apprécié à la main et en bouche pour la fermeté (FER), l'élasticité (ELA), le caractère collant (COL) ou granuleux (GRA) de la pâte.

– aspect visuel : coloration de la pâte (COU) et intensité des exsudations (EXS) de graisse sur la tranche des échantillons.

Une description détaillée des conditions de la dégustation et du mode de présentation des échantillons est décrite dans [Berdagué *et al.*, 90].

2. Présentation et utilisation des données

Les résultats bruts des analyses sensorielles ont fourni un tableau de 16 variables (ou descripteurs sensoriels) × 41 juges × 3 fromages (ou classes), soit 16 variables × 123 observations ou individus.

A partir de ces données, trois essais de classification des fromages par les juges ont été réalisés : l'un par AFD; un autre par Perceptron simple ou quadratique, et le dernier par RPM. Pour ces trois méthodes, une phase d'apprentissage a été effectuée à partir de 25 juges tirés au hasard parmi les 41 donc à partir d'un tableau constitué des 1200 données (16 variables × 25 juges × 3 fromages). Cette étape a été suivie d'une phase de validation des modèles obtenus par AFD et RPM à partir des données restantes, soit 16 variables × 16 juges × 3 fromages = 768 données.

On dispose donc d'un échantillon d'apprentissage à 75 individus et d'un échantillon test à 48 individus.

Une analyse factorielle a été préalablement effectuée, à partir des 25 juges (75 individus) actifs qui ont participé au calcul des axes discriminants. Les 16 juges (48 individus) illustratifs ont été ensuite et projetés sur les axes factoriels lors de l'étape de validation. Tous les axes issus de l'analyse factorielle précédente ont été retenus pour l'AFD et les réseaux neuronaux.

3. Présentation des réseaux de neurones multicouches à rétropopagation du gradient de l'erreur

L'architecture des réseaux multicouches est organisée en niveaux ou couches de neurones : une couche d'entrée qui a autant de neurones que de variables explicatives, une de sortie qui a autant de neurones que de classes à discriminer et un ou plusieurs niveaux intermédiaires appelés couches cachées. Les connexions interneuronales s'effectuent d'une couche à l'autre et jamais sur une même couche. Chaque neurone d'un niveau (niv) est ainsi directement connecté à tous les neurones de la couche suivante (niv + 1).

Tous les neurones j de niveau niv de ce réseau produisent une réponse O_i sur les neurones i de la couche de niveau niv + 1. Cette réponse, ou entrée des neurones i, s'obtient en calculant une somme pondérée des sorties O_j des $K(\text{niv})$ neurones de la couche de niveau niv auxquels ils sont connectés. Cette somme est ensuite transformée par une fonction sigmoïde non linéaire dérivable g :

$$O_i = g\left(\sum_{j=1}^{K(\text{niv})} W_{ij}O_j - \theta_j\right) \qquad i = 1 \ldots K(\text{niv}+1)$$

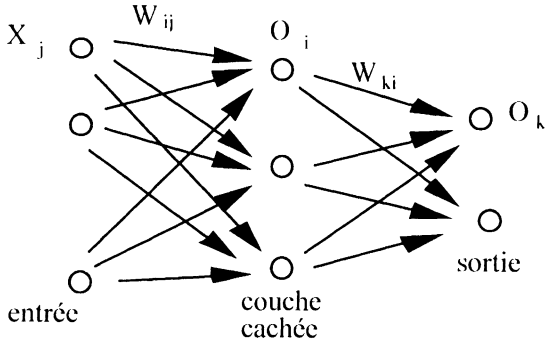

FIGURE 1
Exemple d'un réseau de neurone

avec

θ_i :	seuil du neurone i;
W_{ij} :	poids de la connexion entre les neurones i et j;
$g(x)$:	$(1 - e^{-kx})/(1 + e^{-kx})$: fonction de transfert;
O_i et O_j :	sorties des neurones i et j;
$K(\text{niv})$ et $K(\text{niv} + 1)$:	nombre de neurones des couches respectivement de niveau niv et niv + 1.

L'apprentissage d'un tel réseau est supervisé et utilise un algorithme de rétropropagation du gradient de l'erreur [Rummelhard 86] qui se décompose de la manière suivante :

– **Propagation du signal d'un événement à travers le réseau** : un individu à classer par le réseau est représenté en entrée par un vecteur X à k dimensions. Ensuite, tous les neurones calculent leur réponse respective.

– **Rétropropagation du gradient d'une erreur** : pour chaque individu présenté t, le réseau cherche à minimiser, sur la couche de sortie, une erreur quadratique (ou coût) E commise entre la réponse effective O_k des K neurones de la couche de sortie et la réponse désirée d_k :

$$E = \sum_{k=1}^{K} (O_k - d_k)^2 \tag{1}$$

Pour ce faire, le gradient de E est rétropropagé en modifiant la valeur des poids :

$$W_{ij}(t) = W_{ij}(t-1) + \Delta W_{ij}(t)$$

avec $\Delta W_{ij}(t) = -\varepsilon \frac{\partial E}{\partial W_{ij}} + \alpha \Delta W_{ij}(t-1)$ où ε et α sont deux paramètres d'apprentissage du réseau.

Il est important de remarquer que ce sont les poids W_{ij} qui portent la connaissance du problème à traiter. Concrètement, les valeurs de ces poids contribuent à l'activation des neurones.

Ces deux étapes sont itérées en présentant plusieurs fois au réseau l'ensemble des individus afin de faire converger l'erreur E vers un minimum.

4. Construction des réseaux neuronaux

Nous utilisons d'abord un Réseau Perceptron Multicouches dont les neurones sont organisés en plusieurs couches : une couche d'entrée, une couche de sortie et une ou plusieurs couches intermédiaires.

Architecture du Perceptron Simple et Quadratique

Le nombre de neurones dans la couche d'entrée ne dépend que du nombre de variables explicatives, alors que le nombre de neurones dans la couche de sortie ne dépend que du nombre de classes à discriminer. Dans notre application, le réseau comporte donc 16 neurones dans la couche d'entrée qui reçoivent 16 variables explicatives (coordonnées factorielles); et 3 neurones dans la couche de sortie, chacun affecté à une classe d'Emmenthal. Le Perceptron Simple est un réseau sans couche intermédiaire (ou cachée) entre les neurones de la couche d'entrée et les neurones de la couche de sortie [Zhu 95]. Les poids qui lient le couple (i, j) des neurones d'entrée au neurone k de sortie ne sont pas pris en compte. Le Perceptron Quadratique, lui, contient non seulement les connexions simples précédentes, mais aussi les connexions entre les combinaisons des entrées $x(i)$ et $x(j)$ avec le neurone dans la couche de sortie.

Architecture du Réseau Perceptron Multicouches

Un réseau Perceptron Multicouches est un réseau qui comprend une couche d'entrée et une couche de sortie identiques respectivement à celles du Perceptron Simple et Quadratique, c'est-à-dire qu'il contient pour notre application une couche d'entrée de 16 neurones et une couche de sortie de 3 neurones. Il contient également des couches intermédiaires que l'on appelle «couches cachées».

Le nombre de couches cachées

A l'heure actuelle, il n'existe pas de méthode générale réellement convaincante pour fixer de façon idéale le nombre de couches cachées. Un réseau utilise une couche cachée de neurones pour créer sa propre représentation interne en fonction du problème à résoudre. Cette couche est alors considérée comme un niveau de prétraitement de l'information avant la décision finale. Ainsi, pour s'approcher empiriquement du nombre optimal de couches cachées, il est conseillé de respecter la complexité du problème et la nature des données. La détermination de l'architecture du réseau dépend aussi de la capacité de l'ordinateur. Plus un réseau a de couches cachées (donc plus de connexions), plus le traitement nécessite de mémoire et de

temps de calcul. En effet, le temps d'apprentissage est proportionnel à $N * P_m * L$ où N, P_m et L sont respectivement la taille de l'échantillon d'apprentissage, la taille du réseau (nombre total de neurones) et le nombre de présentations de l'échantillon d'apprentissage. [Wang 92] a montré que l'emploi de réseaux à deux couches cachées suffit. Nous avons constaté expérimentalement que si l'on effectue rigoureusement une sélection[1] de variables explicatives, l'emploi d'un RPM à une ou deux couches cachées suffit.

Le nombre de neurones dans la couche cachée

La détermination du nombre de neurones dans la couche cachée est délicate. Comme indiqué dans [Lechevallier 95], pour un problème donné et des échantillons de taille fixe, un réseau «sous-estimé» (le nombre de neurones dans la couche cachée ou le nombre de couches cachées est insuffisant), aura un nombre de degrés de liberté trop faible; la fonction d'erreur aura donc une composante de biais importante et un terme de variance faible; le modèle est stable, mais a une performance relativement faible. Par contre, un réseau «sur-estimé» (un grand nombre des couches cachées ou un grande nombre de neurones dans la couche cachée), possède un grand nombre de degrés de liberté et l'optimisation à partir d'échantillons d'apprentissage différents conduira à des solutions pouvant être différentes, ce qui correspond à une composante de variance importante. Le modèle neuronal devient instable. Sous l'hypothèse que le RPM ne dispose que d'une seule couche cachée, nous utilisons la technique de rééchantillonnage pour la sélection du réseau optimal. Dans le cas où le réseau possède deux couches cachées, nous utilisons la procédure de détermination de la topologie optimale du réseau, décrite dans [Zhu 95]. Au delà de trois couches cachées[2], le nombre optimal de neurones contenus par les couches cachées ne peut être fixé qu'expérimentalement. Par conséquent, un nombre trop élevé de neurones entraîne d'inévitables redondances et donc pénalise le temps de calcul inutilement employé. Différentes études sont effectuées selon le cas.

5. Comparaisons des résultats

La qualité de l'apprentissage dépend de plusieurs éléments qui peuvent varier : le choix du coefficient du gradient, les valeurs initiales des poids, la durée de l'apprentissage représenté par le nombre de présentation des échantillons d'apprentissage. Pour mettre en œuvre les phénomènes que nous intéressent, nous avons essayé de limiter certaines de ces variations. Nous précisons donc chaque fois les conditions des expériences. Pour les modèles d'architectures différentes, nous fixons nos paramètres à des valeurs «raisonnables» déterminées après quelques essais. Les valeurs initiales des poids sont tirées aléatoirement suivant une distribution uniforme.

---

[1] La sélection des variables peut être effectuée en utilisant des critères statistiques. L'étape de sélection systématique de variables n'est pas le principal objectif de cet exposé. Pour cela, nous utilisons la procédure Demod ou Fuwil de SPAD.N, Stepdisc de SAS ainsi que la segmentation par arbre binaire du logiciel SPAD.S. L'élagage des variables explicatives en utilisant une analyse des sensiblités [Moody 1992] est présenté dans le paragraphe suivant.

[2] A notre connaissance, l'emploi des réseaux ayant plus de trois couches est rare dans la littérature et dans la pratique car un tel réseau est très complexe à analyser.

Le coefficient d'apprentissage ε a été pris égal à une série de valeurs decroîssantes pendant l'apprentissage. Nous pouvons également permettre au réseau durant l'apprentissage d'ajuster au fur et à mesure la valeur du coefficient ε. Ceci a pour objectif d'obtenir une convergence rapide et bonne pour la famille des modèles employés. La fonction à rétropropager durant l'apprentissage est la fonction quadratique.

5.1 L'utilisation des RPM à une seule couche cachée

5.1.1 Apprentissage

Comme nous disposons d'un échantillon d'apprentissage, nous pouvons comparer plusieurs techniques d'analyse discriminante en les appliquant au même ensemble de données. Nous utilisons d'abord l'analyse factorielle discriminante (AFD), la segmentation par arbre binaire (CART) et le réseau Perceptron Simple et Quadratique. Nous obtenons les résultats dans les tableaux 1 à 4.

TABLEAU 1
Classement des observations ayant participé à l'apprentissage de l'AFD et de la segmentation par arbre binaire

Individus actifs	AFD		CART		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	24 96%	1 4%	24 96%	1 4%	25 100%
classe 2	24 96%	1 4%	25 100%	0 0%	25 100%
classe 3	24 96%	1 4%	18 72%	7 28%	25 100%
Ensemble	72 96%	3 4%	67 89.33%	8 10.67%	

On connaît l'importance qu'il y a à bien choisir le réseau sur lequel on va apprendre. On trouve également certaines approches permettant de bien comprendre les phénomènes en jeu et leur influence sur les choix qui sont à faire concernant notamment la taille du réseau [Gallinari 95, Guicheney 92, Zhu 95]. Dans cette application, nous montrons à la fois comment on procède pour choisir un réseau effectif et comment on peut intégrer dans l'architecture de ce réseau des connaissances qui vont influer sur la solution finale.

La démarche empirique exhaustive pour la sélection de modèle consiste à tester sur le problème traité un grand nombre de solutions différentes, décrivant un sous-ensemble M de modèles jugés appropriés par l'utilisateur et à choisir la meilleure

TABLEAU 2
Classement des observations ayant participé à l'apprentissage du Perceptron Simple et du Perceptron Quadratique

Individus actifs	Perceptron Simple		Perceptron Quadratique		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	24 96%	1 4%	24 96%	1 4%	25 100%
classe 2	24 96%	1 4%	25 100%	0 0%	25 100%
classe 3	25 100%	0 0%	25 100%	0 0%	25 100%
Ensemble	73 97.33%	2 2.67%	74 98.67%	1 1.33%	

solution. Nous explorons la famille M pour un critère donné en faisant varier le nombre de neurones dans la couche cachée. Nous utilisons ici un échantillon test qui sert à sélectionner la bonne architecture. Dans la pratique, nous employons les RPM qui ne possèdent qu'une seule couche cachée. Le nombre de neurones dans la couche cachée varie entre 2 et 13. Nous obtenons les résultats dans le tableau 3.

Le réseau ayant deux neurones dans la couche cachée est sélectionné comme un réseau optimal. Les apprentissages des réseaux Perceptron simple, Perceptron Quadratique et Perceptron Multicouches sont obtenus en prenant le paramètre $\varepsilon = 0.01$ après 30 présentations des individus actifs.

En comparant les résultats des cinq méthodes utilisées, nous constatons que l'AFD ne permet pas une classification des individus aussi performante que celle obtenue par le Réseau Perceptron Multicouches et que celle obtenue par le Perceptron Simple ou Quadratique. Cette différence est due au fait que l'AFD utilise des fonctions discriminantes linéaires qui ne prennent pas en compte la non linéarité des données sensorielles. En effet, après normalisation des données par la technique de centrage (élimination des décalages) et réduction (élimination des différences de grandeur), il persiste toujours une non linéarité de comportement des juges au niveau des trois produits.

5.1.2 Validation

Comme nous disposons d'un échantillon test, nous pouvons tester la qualité des différentes techniques de discrimination : AFD, méthode CART et réseaux de neurones. La validation de ces techniques est donc effectuée sur cet échantillon (ensemble des individus illustratifs), c'est-à-dire sur l'ensemble des notations des

TABLEAU 3
Fonctions d'erreur (cf. formule (1)) et pourcentages de bien classés en fonction du nombre de neurones dans la couche cachée

Nombre de neurones dans la couche cachée	% de bien classés sur l'échantillon		Coût E calculé sur l'échantillon	
	App.	Test	App.	Test
2	100%	85.42%	0.008667	0.182960
3	100%	83.33%	0.009754	0.201759
4	100%	81.25%	0.011284	0.207427
5	100%	79.17%	0.010320	0.210513
6	100%	81.25%	0.011634	0.210644
7	100%	81.25%	0.010876	0.202289
8	100%	77.08%	0.010416	0.214959
9	100%	81.25%	0.010921	0.198971
10	100%	79.17%	0.012546	0.208303
11	100%	81.25%	0.009223	0.198718
12	100%	81.25%	0.017488	0.213949
13	100%	81.25%	0.008726	0.194522

TABLEAU 4
Classement des observations ayant participé à l'apprentissage du Perceptron Multicouches

Individus actifs	Perceptron Multicouches		Total
	Nombre de Bien classés	Nombre de Mal classés	
classe 1	25 100%	0 0%	25 100%
classe 2	25 100%	0 0%	25 100%
classe 3	25 100%	0 0%	25 100%
Ensemble	75 100%	0 0%	

16 juges non utilisées à l'apprentissage de l'AFD, du Perceptron Simple et du RPM. Nous constatons que le RPM offre une généralisation (taux de classement) meilleure que celle de l'AFD. Le classement de l'échantillon test des 48 individus est donné par les tableaux 6 à 8.

TABLEAU 5
Temps de calcul et nombre de paramètres des modèles discriminants

	AFD	Perceptron Simple	Perceptron Quadratique	Perceptron Multicouches	CART
Temps*	0.3	0.3	2	0.5	0.7
Nombre de paramètres	17	17	153**	43**	4***

* Le temps (en minute) est mesuré sur PC 80486 DX2-66.

** Le nombre de paramètres neuronaux est le nombre de connexions neuronales. Pour le Perceptron quadratique, le nombre est égal à 17 + 17 * (17 – 1)/2 = 153, pour le Perceptron multicouches, il est égal à 16 * 2 + 2 * 3 + 2 + 3 = 43.

*** Le nombre de paramètres mentionné ci-dessus est le nombre de segments terminaux après l'élagage de l'arbre maximun. La segmentation est une modèlisation statistique non paramétrique.

TABLEAU 6
Classement des individus dans l'échantillon test après validation de l'AFD et de la segmentation

Individus illustratifs	AFD		CART		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	14 87.75%	2 12.5%	12 75%	4 25%	16 100%
classe 2	14 87.75%	2 12.5%	15 93.75%	1 6.25%	16 100%
classe 3	12 75%	4 25%	7 43.75%	9 56.25%	16 100%
Ensemble	40 83.33%	8 16.67%	34 70.83%	14 19.17%	

5.2 L'arrêt de l'apprentissage

L'algorithme d'apprentissage itératif dans les réseaux de neurones explore l'espace des paramètres du réseau selon une trajectoire qui sera fonction de l'algorithme utilisé et des conditions initiales. Les praticiens des réseaux de neurones ont rapidement constaté que trop apprendre à partir d'un échantillon de taille limitée nuit aux performances en validation. Il s'agit du phénomène de *sur-apprentissage* : après une première phase où les poids appris sont des statistiques représentatives du problème, ils commencent à capturer les idiosyncrasies de l'ensemble d'apprentissage. L'erreur sur l'ensemble d'apprentissage décroît continûment au cours du temps puis se stabilise, alors que l'erreur en test passe par un minimum avant de croître. Si l'ensemble de

TABLEAU 7
Classement des individus dans l'échantillon test après validation du Perceptron Simple et du Perceptron Quadratique

Individus illustratifs	Perceptron Simple		Perceptron Quadratique		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	14 87.75%	2 12.5%	14 87.75%	2 12.5%	16 100%
classe 2	14 87.75%	2 12.5%	14 87.75%	2 12.5%	16 100%
classe 3	12 75%	4 25%	12 75%	4 25%	16 100%
Ensemble	40 83.33%	8 16.67%	40 83.83%	8 16.67%	

TABLEAU 8
Classement des individus dans l'échantillon test après validation du Perceptron Multicouches

Individus illustratifs	Perceptron Multicouches		Total
	Nombre de Bien classés	Nombre de Mal classés	
classe 1	15 93.75%	1 6.25%	16 100%
classe 2	13 81.25%	3 18.75%	16 100%
classe 3	13 81.25%	3 18.75%	16 100%
Ensemble	41 85.42%	7 14.58.75%	

test est représentatif des données, c'est en ce minimum qu'il faut arrêter l'apprentissage (figure 2). La pratique courante est d'utiliser, pour fixer l'arrêt de l'apprentissage, un échantillon test afin de ne pas biaiser favorablement l'estimation du modèle. La popularisation de cette procédure d'arrêt [Roméder 69] est un point important dans la

communauté neuromimétique. L'analyse théorique en est encore peu avancée. [Finnoff *et al.* 93] ont réalisé une étude empirique comparative de différentes méthodes visant à biaiser l'apprentissage, d'où il ressort qu'arrêter avant la convergence est une méthode plus efficace.

5.3 Evolution des performances au cours de l'apprentissage

Le phénomène de sur-apprentissage est visible même sur des architectures simples. Les figures 2 (erreur quadratique) et 3 (taux de bien classés) montrent l'évolution des performances dans le temps d'une architecture à 2 neurones dans la couche cachée. Alors que l'erreur sur l'échantillon d'apprentissage décroît continûment au cours du temps, sur l'échantillon test les performances atteignent rapidement un optimum dont la position exacte dépend de l'expérience, puis les performances se dégradent jusqu'à se stabiliser.

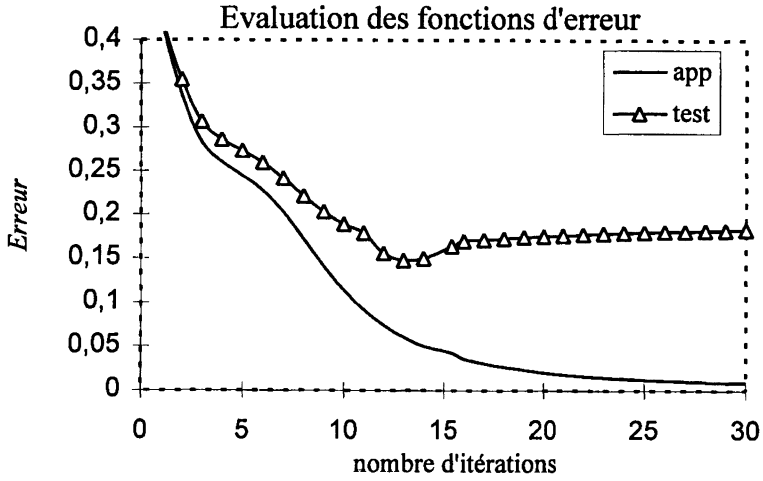

FIGURE 2
Evolution des performances en apprentissage (app) et en test (test) en fonction du nombre d'itérations

Les deux figures 2 et 3 montrent donc très clairement le phénomène de sur-apprentissage au cours du temps. Celui-ci est observable sur une large gamme d'architecture. Toutefois, il disparaît dans deux cas : le premiers cas est celui où l'échantillon d'apprentissage apporte au système choisi toutes les informations qu'il peut capturer sur le problème, c'est-à-dire quand les données sont très nombreuses. Le second, qui est plus compliqué à analyser, correspondant aux réseaux trop complexes [Gallinari 95].

5.4 Evaluation des performances vis-à-vis des échantillon utilisés

Lorsqu'on procède comme précédemment pour la sélection d'un modèle neuronal, on explore une famille M de réseaux en faisant varier le nombre de neurones

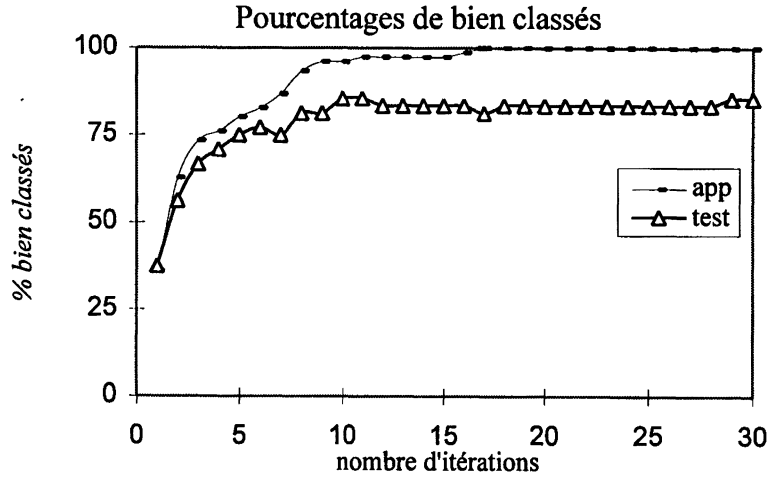

FIGURE 3
Evolution des pourcentages de bien classés en fonction du nombre d'itérations pour l'échantillon d'apprentissage (app) et l'échantillon test (test)

dans la couche cachée. L'estimation des performances dépend alors de l'échantillon utilisé; elle devient d'autant plus sensible que plusieurs facteurs interviennent pour perturber cette évaluation. De nombreuses procédures ont été proposées pour cela dans la littérature statistique (par exemple la *validation croisée* [Efron 83]). Comme nous disposons de données peu nombreuses, nous adoptons la technique de la *validation croisée*.

Les données sont constituées, par un tirage aléatoire (sur les 41 * 3 = 123 juges * fromages), de 5 ensembles disjoints, chacun de 25 individus environ. Chacun de ces ensembles est à tour de rôle échantillon-test, le reste (à savoir les 4 autres ensembles) constituant l'ensemble d'apprentissage. On peut alors pour chaque méthode calculer la moyenne des cinq pourcentages de bien- classés et écart-type associé en apprentissage et en test. les résultats sont données dans le tableau 9.

TABLEAU 9
Tableau de statistiques du taux de bien classés

Modèle	Apprentissage		test	
	Moyenne	Ecart-type	Moyenne	Ecart-type
A.D. Linéaire	95.0%	1.7%	90.6%	8.8%
Réseaux Multicouches	99.0%	0.6%	93.76%	5.28%
Perceptron Quadratique	96.3%	0.8%	92.4%	7.5%

Nous trouvons que les RPM et les Perceptron Quadratique donnent des résultats meilleurs que l'AFD. Les pourcentages de bien classés obtenus par RPM et par le

Perceptron Quadratique sont non seulement plus grands, mais aussi moins variables que ceux fournis par l'AFD. Les modèles neuronaux sont donc plus robustes que le modèle discriminant linéaire.

6. Technique d'élagage des variables d'entrée

Cette technique permet la réduction du nombre de paramètres du RPM en éliminant certaines variables d'entrée. Afin de tester quelles sont les variables les plus significatives pour le calcul des sorties finales neuronales, nous utilisons *l'analyse des sensibilités*. Nous définissons la «sensibilité» de la variable d'entrée X_j de la façon suivante :

$$s(X_j) = \widetilde{E}_q(\overline{X}_j, w) - \widetilde{E}_q(X_j, w)$$

avec $\overline{X}_j = \dfrac{1}{N}\sum_{n=1}^{N} X_j(n)$, $\widetilde{E}_q(w) = \dfrac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(O_k - d_k)^2$.

Ici X_j est la j-ème variable d'entrée du réseau, $X_j(n)$ est la valeur prise par la variable X_j sur l'observation n; K est le nombre de neurones de la couche de sortie, $\widetilde{E}_q(w)$ est la fonction d'erreur quadratique calculée sur l'échantillon d'apprentissage de taille N. La sensibilité $s(X_j)$ mesure la contribution à la fonction d'erreur en remplaçant la variable X_j par sa moyenne $\overline{X}_j$. Comme les données sont normalisées (centrage et réduction) en entrée du réseau, ceci revient statistiquement à fournir une hypothèse de nullité ($w_{ij} = 0$).

Le remplacement de X_j par sa moyenne $\overline{X}_j$ a pour avantage de ne pas changer la configuration initiale du réseau.

Nous introduisons, pour la sélection du meilleur modèle, le critère suivant :

$$\widehat{C}(w) = \left(1 + 2\frac{P_m(w)}{N}\right)\widetilde{E}_q(w)$$

Ici, $P_m(w)$ est le nombre de connexions (et donc le nombre de paramètres ou de pondérations à estimer) dans le modèle neuronal, et N est le nombre d'individus dans l'échantillon d'apprentissage.

Quand N est très grand, ce critère est équivalent au critère *VCG* (critère de la Validation Croisée Généralisé) et *EPF* d'Akaike (Erreur de Prévision Finale d'Akaike) ([Eubank 88]).

$$VCG(w) \approx \frac{1}{\left(1 - \dfrac{P_m(w)}{N}\right)^2}\widetilde{E}_q(w) \quad EPF(w) = \widetilde{E}_q(w)\left(\frac{1 + \dfrac{P_m(w)}{N}}{1 - \dfrac{P_m(w)}{N}}\right)$$

L'analyse des sensiblilités vis-à-vis des 16 variables donne les erreurs calculées en retirant l'une de 16 variables. Les résultats sont obtenus sur la figure 4.

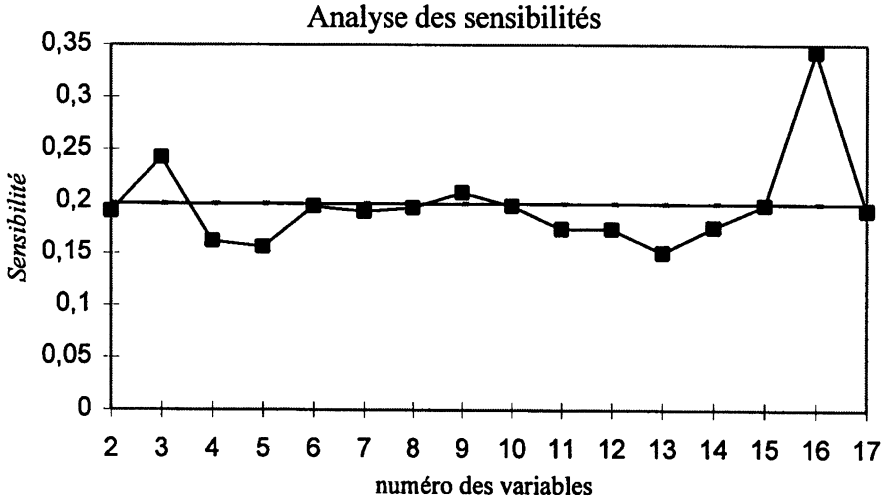

FIGURE 4

L'analyse des sensibilités vis-à-vis des variables d'entrée. La ligne en pointillés représente la moyenne des erreurs calculées sur 16 réseaux multicouches en retirant sucessivement l'une des 16 variables. Les réseaux sont des RPM avec deux neurones dans la couche cachée. L'apprentissage est effectué après 30 présentations de l'échantillon d'apprentissage.

On voit clairement sur la figure 4 que les variables 4, 5, 11, 12, 13 et 14 sont moins sensibles que les autres. Cependant les variables 3 et 16 sont relativement sensibles.

Si on effectue un nouvel apprentissage en retirant les variables 4, 5, 11, 12, 13 et 14, on obtient un réseau multicouches assez simple ayant 31 paramètres (élimination de 12 connexions). L'apprentissage sur l'ensemble des 10 variables significatives donne les résultats contenus dans les tableaux 10 à 13.

Par conséquent, nous constatons que le réseau (10,2,3) ayant 10 neurones dans la couche d'entrée, 2 neurones dans la couche cachée et 3 neurones dans la couche de sortie donne une qualité d'apprentissage presque aussi bonne que le réseau complet (16,2,3) (où tous les poids de connexions sont retenus). La qualité en généralisation est meilleure que celle fournie par le réseau complet, à savoir 87.5% contre 85.4% des invidus correctement classés.

La comparaison des tableaux 1, 2 et 4 d'une part, des tableaux 6 à 8 d'autre part montre que la prise en compte de la non linéarité des données par le réseau donne des classes mieux séparées.

Une étude approfondie est réalisée sur le comportement des juges par rapport aux produits à classer. Le réseau fait ressortir de façon beaucoup plus nette que la majorité des juges utilisés pour la validation a un comportement consensuel pour la caractérisation des trois produits. Certains juges ont un comportement atypique

TABLEAU 10

Classement des observations ayant participé à l'apprentissage de l'AFD et de la segmentation par arbre binaire sans les variables {4,5,11,12,13,14}

Individus actifs	AFD		CART		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	24 96%	1 4%	24 96%	1 4%	25 100%
classe 2	24 96%	1 4%	25 100%	0 0%	25 100%
classe 3	24 96%	1 4%	22 88%	3 12%	25 100%
Ensemble	72 96%	3 4%	71 94.67%	4 5.33%	

TABLEAU 11

Classement des observations ayant participé à l'apprentissage du Perceptron Multicouches sans les variables {4,5,11,12,13,14}.

Individus actifs	Perceptron Multicouches		Total
	Nombre de Bien classés	Nombre de Mal classés	
classe 1	25 100%	0 0%	25 100%
classe 2	24 96%	1 4%	25 100%
classe 3	25 100%	0 0%	25 100%
Ensemble	75 98.67%	0 1.33%	

TABLEAU 12

Classement des individus dans l'échantillon test après validation de Perceptron Multicouches sans les variables {4,5,11,12,13,14}

Individus illustratifs	Perceptron Multicouches		Total
	Nombre de Bien classés	Nombre de Mal classés	
classe 1	15 93.75%	1 6.25%	16 100%
classe 2	13 81.25%	3 18.75%	16 100%
classe 3	14 87.5%	2 12.5%	16 100%
Ensemble	72 87.5%	6 12.5%	

TABLEAU 13

Classement des individus dans l'échantillon test après validation de l'AFD et de la segmentation par arbre binaire sans les variables {4,5,11,12,13,14}.

Individus actifs	AFD		CART		Total
	Nombre de Bien classés	Nombre de Mal classés	Nombre de Bien classés	Nombre de Mal classés	
classe 1	14 87.5%	2 12.5%	12 75%	4 25%	16 100%
classe 2	14 87.5%	2 12.5%	15 93.75%	1 6.25%	16 100%
classe 3	13 81.25%	3 18.75%	8 50%	8 50%	16 100%
Ensemble	41 85.42%	7 14.58%	35 72.92%	13 27.08%	

(en particulier les juges numérotés 2, 13 et 41 ont fourni par deux fois des réponses conduisant à un mauvais classement). Cela révèle les différences importantes au niveau de leur analyse sensorielle des produits.

7. Conclusion

Cette étude a permis de démontrer la capacité des RPM à apporter des solutions nouvelles et plus performantes pour classifier ou caractériser des aliments par analyse sensorielle. Nous pouvons également utiliser les RPM pour sélectionner des variables.

Ainsi, il apparaît clairement que des applications importantes des réseaux de neurones peuvent être envisagées à différents niveaux du contrôle de la qualité des denrées alimentaires. En effet, les RPM prouvent leur aptitude à affiner la sélection des juges ayant un comportement atypique lors de la formation de jury de dégustation. De plus, ils sont capables de mieux exploiter l'information issus des analyses sensorielles.

8. Références

[Berdagué 90] Berdagué J.-L., Grappin R., «Caractérisation sensorielle de l'Emmental français "Grand-Cru"». II : Analyses sensorielles. Le Lait, 70, p 133-145 (1990).

[Berdagué 91] Berdagué J.-L., Grappin R., Caractérisation sensorielle des aliments par analyse factorielle discriminante : l'apport du centrage et de la réduction des données, Lebensm.-Wiss. u.-Technol., 24, p 298-302 (1991).

[Celeux 94] Celeux G., Nakache J.-P., «Analyse discriminante sur variables qualitatives», *Technica*, 1994.

[Efron 83] Efron B., Gong G., «A leisurely look at the bootstrap, the jacknife and cross-validation», *American Stat.* 37 p 36-48, 1983.

[Eubank 88] Eubank Rondall L., «Spline smoothing and non parametric regression», Marcel Dekkel Inc., 1988.

[Finnoff 93] Finnoff W., Hergert, Zimmermann H.G., «Improving model selection by non convergent methods», *Neural Networks* 6, p 589-599, 1993.

[Gallinari 95] Gallinari P., Gascuel O., «Statistique, apprentissage et généralisation; applications aux réseaux de neurones», rapport de recherche LIRMM, Montpellier, 1995.

[Guicheney 92] Guicheney C., «Développement de la technique des réseaux neuromimétiques en physique des particules. Etude de la réaction $e^+e^- -> Z^0 -> \gamma H$». Thèse de l'Université Blaise Pascal, 1992.

[Lebart 95] Lebart L., Morineau A., Piron M., «Statistique exploratoire multidimensionnelle». Dunod, 1995.

[Lechevallier 95] Lechevallier Y., Ciampi A., «Réseaux de neurones et modèles statistiques», *Revue Modulad* 15, 1995.

[Moody 92] Moody J.-E., «Principaled Architecture Selection for neural networks» in *advance in neural information processing system* 4, p 683-690, 1992.

[Roméder 69] Roméder J.-M., «Méthodes et programmes d'analyse discriminante», Dunod, 1969.

[Rummelhart 86] : Rummelhart D., Hinton G.E., Williams R.J., «Learning representation by error backpropagation» dans *Paralled distributed processing/exploration in micro-struture of cognition*. MIT presse, 1986.

[SPAD.N 94] «Système Portable d'Analyse des Données Numériques», CISIA, Saint-Mandé, France, 1994.

[SPAD.S 94] «Segmentation par arbre de décision binaire – Discrimination et régression», CISIA, Saint-Mandé, France, 1994.

[Thomassone 88] Thomassone R., «Comment interpréter les résultats d'une analyse factorielle discriminante». Institut Technique des Céréales et des Fourrages, 8 Avenue du Président wilson, 75116 Paris, p 1-56 (1988).

[ZHU 95] Zhu W.H., «Modèles statistiques et Approche neuronale», Thèse de l'Université de Paris Dauphine, 1995.