Pertinence des Mesures de Confiance en Classification

Relevance of Confidence Measurement in Classification

Ph. Leray[+], H. Zaragoza[*], F. d'Alché-Buc[*]

(+) ASI/PSI - INSA Rouen - BP 08 - Av. de l'Université - 76801 St-Etienne du Rouvray Cedex
(*) LIP6 - Pôle IA - 4, place Jussieu, boîte 169 - 75252 Paris Cedex 5

Résumé

Nous posons le problème de la validation locale des réseaux de neurones artificiels et plus généralement des systèmes automatiques de reconnaissance des formes dans un contexte opérationnel (industriel ou médical par exemple). Est-il possible d'associer un degré de confiance à la réponse locale d'un classifieur semi-paramétrique ? Comment le calculer ? Comment juger de sa pertinence ? Nous explorons cette question importante peu traitée dans la littérature du domaine en la formulant dans le cadre des perceptrons multicouches (PMC). Deux types de mesures heuristiques de confiance sont proposés : l'un reflétant la qualité de la modélisation opérée par le système, l'autre traduisant l'importance relative des estimations des probabilités des classes perdantes. Nous proposons aussi des critères d'évaluation de ces mesures et appliquons ensuite les méthodes proposées sur un exemple simulé et sur un problème réel de diagnostic. Les mesures de confiance proposées permettent de rejeter les exemples pour lesquels la réponse du système s'avère ambiguë.

Abstract

The use of Artificial Neural Networks and automatic pattern recognition systems in an operational environment (e.g. in industry or medicine) rises the problem of local validation : is it possible to associate a degree of confidence to the local response of a classifier ? How to calculate it ? How to judge of its relevance ? In diagnosis tasks, for example, it is crucial to reject decisions produced by the system when their associated confidence is low. This subject, despite its practical importance, has not been sufficiently discussed by the domain literature. We explore it in the framework of multi-layer perceptrons. Different heuristic measures of confidence are reviewed. We then present three methods of comparison of their usefulness based on discriminant power, performances of the resulting reject rules and correlation with the classifier's local performances. Finally, we illustrate the proposed methods on a simulated example and on a real diagnosis problem.

Mots Clés

Discrimination, Estimation de probabilités, réseaux de neurones, confiance, validation.

Keywords

Classification, probability estimation, confidence, neural networks, validation.

1. Introduction

L'utilisation des réseaux de neurones artificiels et plus généralement des systèmes automatiques de reconnaissance des formes dans un contexte opérationnel (industriel ou médical par exemple) pose le problème de leur validation locale : pour une forme présentée au système, est-il possible d'associer à la réponse du classifieur un degré de confiance ? Quelles méthodes employer pour le calculer ? Comment juger de la pertinence de celles-ci ? Ces questions couramment abordées dans le domaine de la régression ont été peu traitées dans le domaine de la discrimination. Il est pourtant crucial, par exemple dans le cas de l'aide au diagnostic, de rejeter à bon escient une décision produite par le système si la confiance associée à cette décision est faible (Dubuisson 93, 96). D'autres utilisations des mesures de confiance peuvent également être envisagées. Les mesures de confiance peuvent, par exemple, permettre de pondérer l'importance des classifieurs "experts" lorsqu'ils sont combinés ou enchaînés pour produire une décision (Tresp and Taniguchi 1995, Shimshoni and Intrator 1996).

Pour étudier cette problématique, nous nous plaçons dans le cadre de la théorie de la décision bayésienne où les décisions proviennent de l'estimation des probabilités *a posteriori* des classes. Sans rien ôter de la généralité du problème, nous l'abordons en nous appuyant sur le modèle des perceptrons multicouches (PMC). Nous considérons ces systèmes après apprentissage, c'est-à-dire dans les conditions mêmes de leur utilisation dans un contexte opérationnel. Il s'agit alors d'estimer ou de mesurer la confiance associée à chaque sortie sans connaître ni les paramètres du réseau et ni a fortiori leur distribution a posteriori. Par conséquent, l'approche bayesienne[1] fondée sur

[1] au sens de l'apprentissage bayesien

la connaissance des distributions *a priori et a posteriori* des paramètres du système (MacKay 92) ne s'applique pas.
Nous commençons par présenter deux types de mesures de confiance : l'un est fondé sur une estimation robuste de la qualité de l'approximation opérée par le réseau, l'autre traduit l'importance relative des estimations des probabilités des classes fournies par le réseau. Une fois ces mesures discutées, il s'agit de quantifier leur pertinence dans le cadre spécifique de la discrimination. Nous proposons donc différents critères traduisant la qualité des règles de décision avec rejet qu'elles engendrent et mesurant leur pouvoir discriminant. Enfin, nous calculons les différentes mesures proposées sur différents problèmes (un cas simulé et un cas réel) en appliquant les principes d'évaluation proposés et mesurons ainsi leur pertinence respective.

2. Mesures de confiance d'un classifieur

Dans le cadre de théorie de la décision bayesienne, la reconnaissance de formes peut être opérée en estimant chacune des probabilités *a posteriori* des classes puis en appliquant la règle de Bayes qui fournit la classe de plus grande probabilité *a posteriori*. Nous nous plaçons dans le cas d'un système tel qu'un perceptron multicouches qui sous des conditions appropriées fournit une approximation des probabilités *a posteriori*. Mesurer la confiance associée à la réponse d'un tel classifieur nécessite de prendre en compte la qualité de l'approximation et l'importance relative des sorties.

2.1 Interprétation des sorties d'un PMC

On considère un problème de discrimination à k classes où les sorties désirées associées aux exemples sont codées par un vecteur binaire (1 pour l'indice correspondant à la classe gagnante, 0 pour les autres). (White 1989, Richard and Lippmann 1991, Lee, Srihari and Gaborski 1991) ont montré qu'un PMC, de complexité suffisante et minimisant l'erreur quadratique moyenne[2] entre sorties calculées et sorties désirées, approche les probabilités $P(C_j|x)$ quand le nombre d'exemples d'apprentissage tend vers l'infini.
En pratique, l'apprentissage se fait avec un nombre restreint d'exemples et les paramètres $w°$ estimés nous permettent seulement d'obtenir une approximation des probabilités a posteriori :

$$\hat{P}(C_j|\boldsymbol{x}) = F_j(\boldsymbol{x},w°) \qquad (1)$$

[2] Ceci est également valable pour d'autres fonctions de coût (par exemple, la fonction de coût entropique)

où $F_j(\boldsymbol{x},w°)$ représente la j-ème sortie du PMC avec $w°$ poids obtenus par apprentissage et $P(C_j|\boldsymbol{x})$ probabilité *a posteriori* de la classe j.
Une propriété dérivable pour les sorties du PMC est donc :

$$\sum F_j(\boldsymbol{x},w°) \approx 1 \text{ , pour tout } \boldsymbol{x} \qquad (2)$$

Si l'on désire avoir une égalité stricte dans (2), il est possible d'utiliser lors de l'apprentissage la normalisation "softmax" donnée par l'équation (3):

$$F'_j(\boldsymbol{x},w) = \frac{e^{F_j(\boldsymbol{x},w)}}{\sum_l e^{F_l(\boldsymbol{x},w)}} \qquad (3)$$

Dans la pratique, cette normalisation n'est pas nécessaire. Si l'on cherche à obtenir une égalité stricte dans (2), il est plus simple, une fois que le classifieur est entraîné, de normaliser les sorties de la façon suivante :

$$\hat{P}(C_j|\boldsymbol{x}) = \frac{F_j(\boldsymbol{x},w°)}{\sum_l F_l(\boldsymbol{x},w°)} \qquad (4)$$

Pour vérifier que les sorties (normalisées par (4) ou non) sont de bonnes approximations des probabilités a posteriori, (Wan 1990) propose de calculer $\hat{P}(C_j)$:

$$\hat{P}(C_j) = \frac{1}{N}\sum_{x_i} \hat{P}(C_j|x_i) \qquad (5)$$

Il est alors possible de comparer $\hat{P}(C_j)$ avec la probabilité *a priori* de la j-ème classe (estimée par sa fréquence) $\tilde{P}(C_j) = \frac{N_j}{N}$ où N_j est le nombre d'exemples de la classe j.

2.2 Intérêt d'une mesure de confiance

De manière générale, les sorties d'un PMC utilisé en classification sont uniquement utilisées pour déterminer la classe de l'exemple présenté. Pourtant, les sorties du classifieur permettent aussi d'obtenir une mesure de la confiance que l'on peut lui donner, autorisant ainsi des décisions plus élaborées telles que le rejet d'un exemple ou la combinaison de classifieurs en fonction de leur confiance respective.

2.2.1 Règle de décision avec mesure de confiance

Dans le cas d'un problème de classification à g classes, la règle de Bayes s'écrit de la façon suivante:

$$C^*(\boldsymbol{x}) = argmax_j\, P(Cj|\boldsymbol{x}) \qquad (6)$$

où $C^*(\boldsymbol{x})$ est la classe choisie par le classifieur de Bayes.

(Chow 1970) a montré qu'il était possible d'inclure une nouvelle décision C_a dans cette règle, correspondant au rejet du résultat du classifieur lorsque la probabilité a posteriori du résultat est trop faible:

$$C^*(\boldsymbol{x}) = \begin{cases} argmax_j P(Cj \mid \boldsymbol{x}) & \text{si } \max_j P(Cj \mid \boldsymbol{x}) \geq \Theta_a \\ C_a & \text{sinon} \end{cases} \quad (7)$$

où $\Theta_a \in [0..1]$ est le seuil de rejet, paramètre libre fixé par l'utilisateur.

Ce type de rejet a été qualifié par (Dubuisson and Masson 1993) de rejet d'ambiguïté. Il permet de déterminer si un exemple est à la frontière de deux classes. (Dubuisson and Masson 1993) proposent aussi un autre type de rejet, le rejet en distance, que nous n'utiliserons pas par la suite, mais qui permet de déterminer si l'exemple proposé est aberrant (dans ce cas, il peut appartenir à une classe différente de celles considérées par le classifieur).

La valeur $max_j\ P(Cj|\boldsymbol{x})$ utilisée pour la décision de rejet (7) peut être interprétée comme une mesure de confiance dans l'exemple $\boldsymbol{x}$. Si cette valeur est nulle, l'exemple est rejeté quelque soit la valeur du seuil. De même, plus cette valeur est élevée plus on peut avoir confiance dans la décision.

Il est possible d'utiliser la règle de décision avec rejet d'ambiguïté avec un classifieur de type PMC puisque ce dernier permet d'estimer les probabilités a posteriori des classes. Par contre, nous ne connaissons pas exactement $P(Cj|\boldsymbol{x})$ mais juste une approximation de ces probabilités. Ainsi, il n'est pas possible d'estimer exactement la mesure de confiance théorique $max_j\ P(Cj|\boldsymbol{x})$. Il est, par contre, possible d'utiliser des mesures de confiance heuristiques permettant de remplacer cette mesure théorique, avec la règle de décision suivante :

$$C^\circ(\boldsymbol{x}) = \begin{cases} argmax_j \hat{P}(Cj \mid \boldsymbol{x}) & \text{si } m(\boldsymbol{x}) \geq \Theta_a \\ C_a & \text{sinon} \end{cases} \quad (8)$$

où les $\hat{P}(C_j|\mathbf{x})$ sont obtenus grâce au PMC, $C^\circ(\boldsymbol{x})$ représente la décision prise et $m(\boldsymbol{x})$ est une mesure de confiance heuristique vérifiant les hypothèses suivantes :

- $m(\boldsymbol{x}) \in [0..1]$
- $m(\boldsymbol{x}) = 0$ lorsqu'il faut toujours rejeter $\boldsymbol{x}$ (aucune confiance dans cet exemple)
- $m(\boldsymbol{x}) = 1$ s'il faut toujours accepter la décision concernant $\boldsymbol{x}$ (confiance absolue dans l'exemple).

2.2.2 Mesure de confiance et combinaison de classifieurs

Nous venons de voir qu'il est possible d'associer à un classifieur une mesure qui quantifie la confiance que nous lui donnons pour un exemple donné.

(Tresp and Taniguchi 1995) proposent d'utiliser une mesure de confiance sans faire de rejet mais pour combiner linéairement les résultats des différents classifieurs.

De leur côté, (Shimshoni and Intrator 1996) proposent une règle de décision avec rejet mettant les différents classifieurs en compétition en fonction de leur mesure de confiance.

Dans (Leray 1998), nous proposons une combinaison de classifieur utilisant un réseau bayesien dont la probabilité des noeuds (avant inférence) tient compte de la décision prise par la règle (7).

2.3 Définition de mesures de confiance

Dans (Zaragoza et d'Alché-Buc 1998, Leray 98) nous proposons de classer les mesures de confiance heuristiques en deux familles. Le premier type de mesures est basé sur la qualité du modèle. Puisque le modèle fournit une approximation des probabilités a posteriori des classes, il est légitime de s'interroger sur la qualité de cette approximation avant de prendre une décision. L'autre famille de mesures de confiance s'appuie sur la décision du classifieur. Après avoir présenté ces deux types de mesures, nous proposons ensuite de les combiner pour prendre en compte leurs qualités respectives.

2.3.1. Confiance par rapport à la qualité du modèle

Il existe plusieurs moyens de déterminer la confiance dans un estimateur. Le moyen le plus couramment utilisé est l'estimation de la variance de l'estimateur pour un exemple donné, par des méthodes robustes telles que le bootstrap ou la validation croisée. La méthode du Boostrap (Efron and Tibshirani 1993, Tibshirani 1996) mesure la "variabilité" de l'estimation due à l'algorithme d'apprentissage et au nombre limité de données.

Pour des problèmes de classification, (Shimshoni and Intrator 1996) proposent d'utiliser comme mesure de confiance la variance calculée à partir des ensembles de bootstrap :

$$\mathrm{Var}[\,\hat{P}(C_j|\mathbf{x})\,] = \frac{1}{N_b}\sum_b \left(\hat{P}_b(C_j|\boldsymbol{x}) - \overline{\hat{P}(C_j|\boldsymbol{x})}\right)^2 \quad (9)$$

avec $$\overline{\hat{P}(C_j|\mathbf{x})} = \frac{1}{N_b}\sum_b \hat{P}_b(C_j|\boldsymbol{x}) \quad (10)$$

où $\hat{P}_b(C_j|\boldsymbol{x})$ est l'estimation de $P(Cj|\boldsymbol{x})$ obtenue grâce au b-ième ensemble de bootstrap et N_b le nombre total d'ensembles de bootstrap. La qualité de l'estimation est inversement proportionnelle à la variance. Une variance faible signifie que l'estimation de la probabilité a posteriori opérée par le réseau est peu sensible aux données d'apprentissage , ce qui conforte l'idée d'un modèle fiable. A contrario, une variance forte traduit la faiblesse de la fiabilité car l'exemple proposé ne sera pas toujours bien classé. Nous avons alors proposé dans (Zaragoza et d'Alché-Buc 1998, Leray 1998) une mesure de confiance basée sur la variance de la classe la plus probable, normalisée pour répondre aux contraintes posées en §2.2 :

$$m_0(\boldsymbol{x}) = 1 - \frac{Var(\hat{P}(C_{j^\circ}|\boldsymbol{x}))}{(1/2)^2} \qquad (11)$$

où $j^\circ = argmax_j \overline{\hat{P}(C_j|\mathbf{x})}$.

Dans le meilleur des cas, la modélisation est parfaite donc $Var(\hat{P}(C_{j^\circ}|\boldsymbol{x})) = 0$. L'équation (11) donne alors $m_0(\boldsymbol{x})=1$ ce qui correspond à une confiance absolue dans le modèle.

Dans le pire des cas, la loi $\hat{P}(C_j|\mathbf{x})$ fournie par les différents modèles de bootstrap suit une loi aléatoire uniforme entre 0 et 1 (de variance égale à $(1/2)^2$). Dans ce cas, l'équation (11) donne $m_0(\boldsymbol{x})=0$ ce qui correspond bien à une confiance nulle dans le modèle.

La mise en oeuvre de ce type de méthode est malheureusement assez lourde puisqu'elle nécessite autant d'apprentissages que d'ensembles de bootstrap. Ceci est rédhibitoire lorsque le problème général est lui même décomposé en un grand nombre de tâches. Une autre méthode pour estimer la variance de l'estimateur consiste à utiliser un réseau de neurones qui devra apprendre cette variance. Cependant, cette méthode introduite par plusieurs auteurs (Lippmann, Kukolich and Shalian 1995, Nix and Weigend 1995) nécessite l'existence d'un grand nombre d'exemples de manière à procéder au second apprentissage de manière indépendante par rapport à l'apprentissage du classifieur initial. Nous ne retiendrons donc pas cette méthode.

2.3.2 Confiance par rapport à la décision

Supposons maintenant que nous possédons un classifieur fiable (i.e. un PMC, ou un ensemble de PMC obtenus par bootstrap[3], dont les sorties nous fournissent une bonne approximation des probabilités a posteriori des classes).

Il est possible de déterminer des mesures de confiance heuristiques basées sur ces probabilités et qui nous donneront la confiance par rapport à la décision du classifieur pour un exemple donné. Les contraintes définies en §2.2 pour une mesure de confiance quelconque peuvent être explicitées en conséquence :

- [1] $m(\boldsymbol{x}, \hat{P}) \in [0..1]$[4]
- [2] $m(\boldsymbol{x}, \hat{P}) = 1$
  lorsque $\{ \hat{P}(C_j|\mathbf{x}) \} = \{1,0,...,0\}$
- [3] $m(\boldsymbol{x}, \hat{P}) = 0$
  lorsque $\{ \hat{P}(C_j|\mathbf{x}) \} = \{1/g,...,1/g\}$

où $\{ \hat{P}(C_j|\mathbf{x}) \} = \{ \hat{P}(C_j|\mathbf{x}), \forall j \in \{1,..,g\} \}$ est trié dans l'ordre décroissant des valeurs des probabilités.

Notons aussi P_{max} le plus grand élément de cet ensemble et $j^\circ = argmax_j \hat{P}(C_j|\mathbf{x})$ son indice.

Rajoutons une contrainte du type :

- [4] $m(\boldsymbol{x}, \hat{P}_1) > m(\boldsymbol{x}, \hat{P}_2)$ pour
  $\{ \hat{P}_1(C_j|\mathbf{x}) \} = \{P_{1max}, \frac{1-P_{1max}}{g-1}, ..., \frac{1-P_{1max}}{g-1} \}$
  et $\{ \hat{P}_2(C_j|\mathbf{x}) \} = \{P_{2max}, 1-P_{2max}, 0...0\}$

Cette dernière contrainte indique que nous avons d'avantage confiance dans un modèle qui minimise les valeurs des probabilités des classes "perdantes". Puisque la probabilité de la bonne classe est P_{max}, et avec la propriété $\sum \hat{P}(C_j|\mathbf{x}) = 1$, on obtient la relation suivante :

$$\sum_{j \neq j^\circ} \hat{P}(C_j|\mathbf{x}) = 1 - P_{\max}$$

Les modèles $\hat{P}_1$ et $\hat{P}_2$ utilisés dans la contrainte [4] correspondent aux deux cas limites de cette équation. Dans le premier cas, les probabilités des classes perdantes sont bien minimisées, dans l'autre c'est le contraire.

Un certain nombre d'heuristiques peuvent être utilisées comme mesure de confiance. Nous allons passer en revue les plus courantes et regarder si elles vérifient nos quatre contraintes.

---

[3] Dans ce cas, il suffit de remplacer dans chaque équation le terme $\hat{P}(C_j|\mathbf{x})$ représentant l'estimation des probabilités a posteriori par le PMC par $\overline{\hat{P}(C_j|\mathbf{x})}$ (10) estimée grâce aux ensembles de bootstrap.

[4] Dans la plupart des cas, nous travaillerons avec un seul modèle et donc une seule évaluation des probabilités a posteriori. Nous omettrons alors d'indiquer $\hat{P}$ comme argument de la mesure de confiance.

La mesure de confiance la plus couramment utilisée (notons la m_1) est celle issue de la règle de décision bayesienne (7) :

$$m_1(\mathbf{x}) = P_{max} = max_j \ \hat{P}(C_j|\mathbf{x}) \qquad (12)$$

m_1 obéit bien aux contraintes [1] à [3], mais pas à la contrainte [4] (sauf si l'on transforme l'inégalité stricte en inégalité large).
(12) prend uniquement en compte la plus grande probabilité a posteriori, mais il est possible de prendre en compte la probabilité a posteriori de toutes les classes en utilisant une des deux mesures (13) et (14) qui vérifient bien les contraintes précédentes :

$$m_2(\mathbf{x}) = 1 - \frac{1}{g} \ H\{ \hat{P}(C_j|\mathbf{x}) \} \qquad (13)$$

où $H\{ \hat{P}(C_j|\mathbf{x}) \} = - \sum_j \ \hat{P}(C_j|\mathbf{x}) \ log \hat{P}(C_j|\mathbf{x})$ est l'entropie des probabilités a posteriori.

$$m_3(\mathbf{x}) = P_{max} - P_{2max} \qquad (14)$$

où P_{2max} est le deuxième plus grand élément de $\{ \hat{P}(C_j|\mathbf{x}) \}$.

Toujours dans l'idée de la contrainte [4] (i.e. il faut que P_{max} soit le plus grand possible et que les autres probabilités soient les plus petites possibles), d'autres heuristiques ont été proposées par (Lengellé et al. 1991, Dubuisson, Masson and Frélicot 1996) dans le cadre du diagnostic avec rejet. Ces mesures sont basées sur la distance entre $\{ \hat{P}(C_j|\mathbf{x}) \}$ et la sortie "idéale" du classifieur {1,0,...0}. Nous proposons de considérer respectivement la distance quadratique et l'entropie croisée entre ces deux valeurs, ce qui nous donne les équations (15) et (16) normalisées pour obéir aux contraintes [1] à [3].

$$m_4(\mathbf{x}) = 1 - \alpha \ D\_quad(\{ \hat{P}(C_j|\mathbf{x}) \},\{1,0,...0\}) \qquad (15)$$

$$m_5(\mathbf{x}) = 1 - \beta \ Ent\_Crois(\{ \hat{P}(C_j|\mathbf{x}) \},\{1,0,...0\}) \qquad (16)$$

où α et β sont les coefficients de normalisation non détaillés ici.
Il est possible d'inventer bien d'autres heuristiques de cette sorte, basées sur d'autres types de distance (distance de Kullback-Liebler, mesures de séparabilité, (cf. Devijver and Kittler 1982)). En pratique, comme nous le montrons dans la troisième partie, la plupart de ces mesure sont équivalentes et il est difficile d'en choisir une a priori.

2.3.3 Confiance par rapport à la qualité du modèle et à la décision

Nous avons présenté en §2.3.1 une mesure de confiance m_0 prenant en considération la qualité de l'approximation des probabilités a posteriori par notre classifieur. Nous avons ensuite passé en revue en §2.3.2 plusieurs mesures de confiance m_i basées sur les résultats du classifieur. Dans (Zaragoza et d'Alché-Buc 1998) nous avons alors proposé de combiner ces deux types de mesures de confiance.
Puisque le résultat de la combinaison doit rester une mesure de confiance et vérifier les contraintes générales présentées en §2.2, nous proposons dans (Leray 98) d'effectuer la combinaison linéaire convexe donnée par :

$$m_6(\mathbf{x}) = a \ m_0(\mathbf{x}) + b \ m_i(\mathbf{x}) \qquad (17)$$

avec $a \geq 0$, $b \geq 0$, $a + b = 1$. m_0 est donnée par l'équation (12) et m_i est une de mesures proposées en §2.3.2.
Les paramètres a et b peuvent être estimés à partir des données, en utilisant un nouvel ensemble d'exemples pour éviter tout biais, ou en inversant la base d'apprentissage et celle de validation (en conservant la même base de test) si le nombre d'exemples est limité.
L'apprentissage de m_6 pose le même problème que l'apprentissage d'un classifieur : les sorties désirées de la combinaison (i.e. la mesure de confiance théorique $max_j \ P(Cj|\mathbf{x})$ du classifieur bayesien) ne sont pas connues. Il faut alors prendre comme valeurs désirées la fonction $Correct(\mathbf{x})$ qui donne 1 si $\mathbf{x}$ est bien classé et 0 sinon.
La détermination de a et b est ensuite effectuée par minimisation du coût :

$$\sum \| m_6(\mathbf{x}) - Correct(\mathbf{x}) \|^2$$

sous les contraintes liées à la convexité.

L'avantage de la mesure de confiance m_6 est de prendre en compte à la fois la qualité du modèle et les résultats obtenus par le classifieur. De plus les pondérations accordées à ces deux critères sont estimées à partir des données, et peuvent varier en fonction du problème considéré.
Dans le cas où le nombre d'exemples n'est pas suffisant pour estimer les coefficients de cette combinaison, des résultats acceptables ont été obtenus en donnant la même pondération aux deux types de mesures de confiance ($a = b = 0.5$), ce qui nous donne l'heuristique suivante :

$$m_7(\mathbf{x}) = \frac{m_0(\mathbf{x}) + m_i(\mathbf{x})}{2} \qquad (18)$$

3. Qualité d'une mesure de confiance

Toutes les mesures de confiance que nous venons de présenter reposent sur des heuristiques. Nous allons maintenant proposer plusieurs outils permettant de mesurer, pour un problème donné, l'efficacité de ces mesures. Ces outils seront utilisés dans la partie suivante pour comparer les mesures de confiance sur différents problèmes.
Il existe autant de méthodes heuristiques pour déterminer la qualité d'une mesure de confiance que de mesures de confiance.

Nous avons retenu deux familles représentatives. D'une part, la courbe Performance/Rejet nous donne les performances d'un classifieur utilisant une règle de décision avec rejet. D'autre part, nous essayerons de déterminer directement le pouvoir discriminant de la mesure de confiance.

3.1 Courbe Performance/Rejet

La courbe Performance/Rejet reproduit l'évolution des performances d'un classifieur utilisant la règle de décision avec rejet (8) lorsque le seuil de rejet varie. La figure 1 nous en donne un exemple. L'axe horizontal représente le pourcentage d'exemples rejetés par la règle de décision tandis que l'axe vertical représente le pourcentage d'exemples bien classés parmi les exemples non rejetés. Notons que pour des pourcentages de rejets importants, l'évaluation des performances du classifieur s'effectue sur un nombre peu élevé d'exemples et n'est donc pas forcément 'robuste'. Nous avons représenté ce phénomène sur la figure 1 par des fluctuations intervenant sur la partie droite de la courbe.

Dans certains cas, il peut être utile de rajouter sur cette courbe le taux de bonne classification de chacune des classes, ou de celles les moins représentées. Imaginons un problème où l'une des classes (dite 'faible') possède peu d'exemples et supposons que le classifieur la reconnaît moins bien que les autres classes. Certaines mesures de confiance peuvent avoir tendance à rejeter uniquement les points de cette classe, améliorant les performances globales du classifieur au détriment de la reconnaissance de la classe 'faible'. Ce phénomène est alors facilement détectable grâce aux différentes courbes.

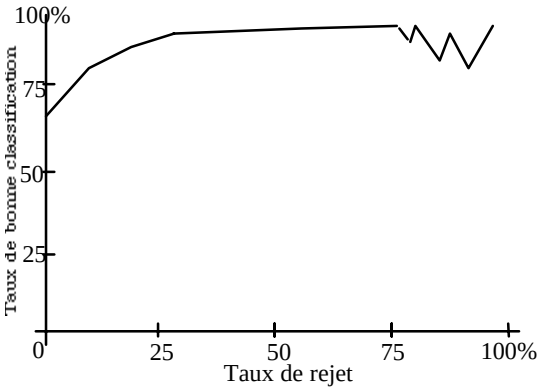

Figure 1 : Exemple de courbe Performance/Rejet. L'axe horizontal représente le pourcentage d'exemples rejetés par la règle de décision tandis que l'axe vertical représente le pourcentage d'exemples bien classés parmi ceux non rejetés. Plus le taux de rejet est important, moins il y a de points pour évaluer les performances (fluctuations).

3.2 Pouvoir discriminant de *m(x)*

Considérons le problème de discrimination à deux classes : la classe des exemples bien classés (B^+) et celle des exemples mal classés (B^-).

Pour une mesure de confiance 'idéale', la règle de décision avec rejet (9) se transforme de la façon suivante :

$$\begin{cases} \boldsymbol{x} \in B^+ & \text{si } m(\boldsymbol{x}) \geq \theta_a \\ \boldsymbol{x} \in B^- & \text{sinon} \end{cases} \qquad (19)$$

La mesure de confiance doit permettre de déterminer si l'exemple $\boldsymbol{x}$ est bien classé ou non par le classifieur.

Nous proposons dans (Leray 1998) de mesurer le pouvoir de discrimination des mesures de confiance par rapport aux classes B^+ et B^- afin d'estimer leur importance. Nous présentons ci dessous deux mesures simples, le coefficient de corrélation et la surface de confusion, ainsi que des mesures plus complexes.

3.2.1 Coefficient de corrélation

Reprenons la fonction *Correct(**x**)*, fonction nous indiquant si le classifieur utilisé classe bien ou non l'exemple $\boldsymbol{x}$. Nous avons :

$$\begin{cases} Correct(x) = 0 & si\ x \in B^- \text{ (ensemble des points mal classés)} \\ Correct(x) = 1 & si\ x \in B^+ \text{ (ensemble des points bien classés)} \end{cases}$$

Cette fonction représente la mesure de confiance optimale puisque elle permet de reconnaître les exemples à rejeter (les mal classés). Nous proposons donc de calculer, à partir des exemples de la base de test, le coefficient de corrélation entre cette fonction et une mesure de confiance donnée :

$$Q_1(m) = Corrélation(m, Correct) \qquad (20)$$

La figure 2 donne une interprétation de ce coefficient. D'après la définition du coefficient de corrélation, Q_1 mesure la répartition des points autour de la droite d'ajustement entre *Correct(**x**)* et *m(**x**)*.

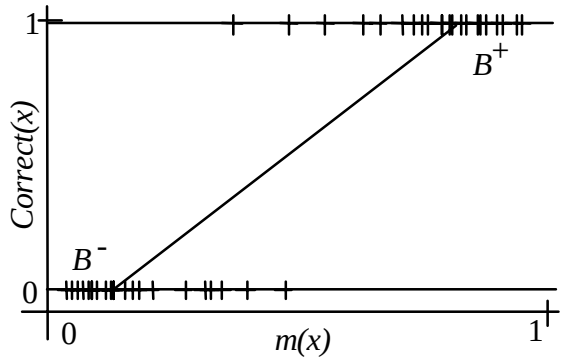

Figure 2 : Corrélation entre la mesure de confiance m et la fonction Correct.

Si $Q_1 = 0$ alors la mesure de confiance ne permet pas de rejeter les exemples. De même, plus Q_1 tend vers *1*, plus la mesure de confiance est utile.
Il peut arriver que Q_1 soit négatif. Cela signifie que la mesure de confiance et la fonction *Correct* sont anti-corrélées. Cela peut arriver lorsque la mesure de confiance ne vérifie pas la contrainte de croissance (plus *m* augmente, plus on a confiance dans l'exemple).

3.2.2 Surface de confusion

Une autre méthode simple pour mesurer le pouvoir de discrimination d'une mesure de confiance est de calculer de manière numérique (à l'aide d'un simple histogramme par exemple) la surface de confusion entre les densités de probabilité $P(m|B^+)$ et $P(m|B^-)$ (cf. figure 3) pour un nombre fini d'exemples. Notons $Q_2(m)$ cette surface.
Quand $Q_2=0$ la mesure de confiance sépare complètement les classes B^+ et B^-, ce qui correspond à une mesure de confiance optimale. Ensuite, plus Q_2 tend vers *1*, moins la mesure de confiance est utile.

3.2.3 Autres mesures de pouvoir discriminant

Il existe bien d'autres mesures de pouvoir discriminant moins heuristiques que les précédentes comme par exemple le lambda de Wilks . Sous l'hypothèse que les distributions de B^+ et B^- sont unimodales, le lambda de Wilks appliqué à la mesure de confiance nous donne une mesure de qualité que nous noterons $Q_3(m)$.
Plus Q_3 est faible, plus la mesure de confiance est utile (i.e. mieux elle sépare les exemples bien classés et mal classés).

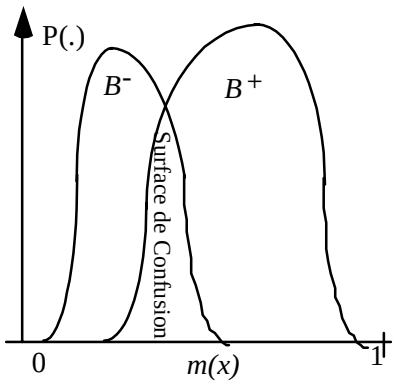

Figure 3 : Surface de confusion entre les distributions de la mesure de confiance pour les classes "exemples bien classés" (B^+) et "exemples mal classés" (B^-) .

4. Quelques exemples

Pour étudier empiriquement l'intérêt des mesures de confiance proposées, nous avons appliquées ces mesures à différents types de jeux de données :

- un jeu de données simulées constitué de deux classes gaussiennes en dimension 20, sans bruit.
- le problème des vagues de Breiman, problème bien connu du domaine, où les données sont non linéairement séparables et bruitées.
- un problème réel de détection de pannes de tracteurs
- un problème de classification d'alarmes dans le cadre de la gestion en temps réel du trafic téléphonique.

Dans tous les cas, nous avons évalué l'intérêt des mesures en calculant les différents critères présentés précédemment. Il ressort d'abord de ces expériences que les mesures $m_1,\ldots, m_5$ sont très proches et pertinentes sauf dans le cas où l'échantillon d'apprentissage représente mal les classes en présence. Les résultats sur les problèmes 1 et 4 étant présentés en détail dans (Leray 98), nous invitons le lecteur à s'y reporter et présenterons en détail les simulations numériques concernant les problèmes de Breiman et de la détection de pannes.
Pour tous ces exemples, les classifieurs utilisés sont des perceptrons multicouches (avec une seule couche cachée de 10 neurones). Une comparaison très poussée des différentes mesures de qualité ne peut se faire que si l'on possède un intervalle de confiance précis sur chaque mesure. Ce n'est malheureusement pas le cas. En posant l'hypothèse forte que la mesure de qualité suit une loi normale de variance $\sigma=1/2$, nous utilisons un intervalle de confiance plus simple de la forme :

$$IC = Q_i(m) \pm \frac{\sigma}{\sqrt{N}} \qquad (21)$$

4.1 Vagues de Breiman

Nous présentons dans le tableau 1 et la figure 4 les résultats obtenus sur le problème des vagues de Breiman, problème de classification à trois classes symétriques.
Les différentes heuristiques proposées pour mesurer la confiance dans le classifieur donnent à peu près toutes les mêmes résultats. De même, les mesures de qualité présentées dans la table 2 donnent des résultats proches. Les mesures de qualité Q_1 et Q_3 classent les mesures de confiance dans le même ordre, différent de l'ordre proposé par Q_2. La figure 5 nous propose un classement des mesures de confiance en tenant compte des deux facteurs de qualité : les mesures de confiance les plus intéressantes sont alors m_3, m_4 et m_5.

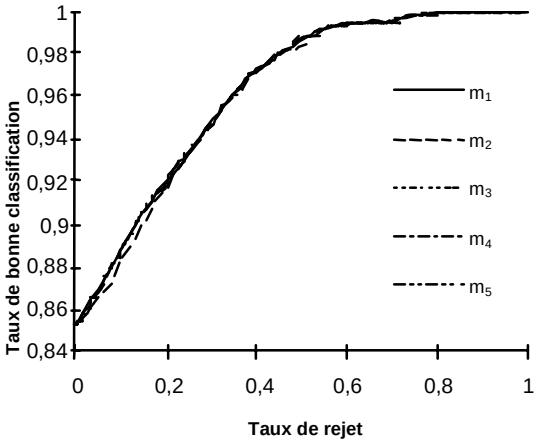

Figure 4 : Problème des vagues de Breiman. Courbes Performance/Rejet pour les mesures de confiance m_1 à m_5.

Mesure de confiance	Mesure de qualité		
	Q_1	Q_2	Q_3
m_1	0.448 [0.433 - 0.463]	0.440 [0.425 - 0.455]	0.799 [0.784 - 0.814]
m_2	0.389 [0.374 - 0.404]	0.438 [0.423 - 0.453]	0.849 [0.833 - 0.864]
m_3	0.448 [0.433 - 0.463]	0.430 [0.415 - 0.445]	0.799 [0.784 - 0.814]
m_4	0.461 [0.446 - 0.476]	0.421 [0.406 - 0.436]	0.788 [0.773 - 0.803]
m_5	0.455 [0.440 - 0.470]	0.433 [0.418 - 0.448]	0.793 [0.778 - 0.808]

Table 1 : Problème des vagues de Breiman. Mesures de qualité Q_1 à Q_3 des mesures de confiance m_1 à m_5. L'intervalle entre crochets indique l'intervalle de confiance correspondant.

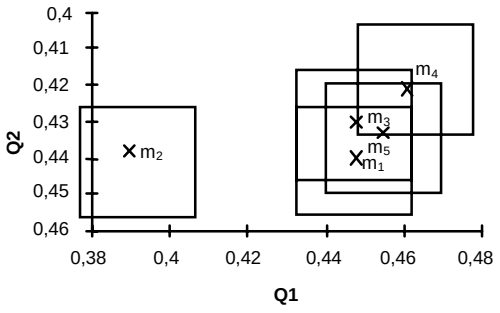

Figure 5 : Problème des vagues de Breiman. Ordre des différentes mesures de confiance en fonction de Q_1 et Q_2. L'axe correspondant à Q_2 est inversé de manière à trouver les meilleures mesures de confiance en haut à droite de la figure.

4.2 Détection de pannes

Nous présentons ici l'application des mesures de confiance à un problème réel, application proposée dans (Zaragoza et d'Alché-Buc 1998). Il s'agit d'une tâche de détection de pannes de tracteurs à partir de courbes de régime moteur, avec 50 variables. Les résultats proposés ci-dessous portent sur un sous-problème difficile, la détection d'une panne spécifique (la surconsommation du moteur) peu représentée dans la base d'exemples (moins d'un tiers des exemples).

Nous avons vu dans l'exemple précédent que les différentes mesures de confiance par rapport aux sorties du classifieur donnaient des résultats similaires. Nous allons ici nous intéresser à m_1. La figure 6 montre que la courbe Performance/Rejet de m_1 (courbe "Perf. Globales" en trait épais) est globalement satisfaisante.

Par contre, la même courbe tracée pour la classe "panne détectée" montre que m_1 n'est pas robuste pour les classes faibles, elle élimine les exemples de panne bien détectés, entraînant une chute des performances pour cette classe (courbe "Perf. Classe-" en trait épais). Cela se traduit au niveau des mesures de qualité de la table 2 par un coefficient de corrélation Q_1 négatif pour m_1 par rapport à la classe "panne".

Puisque la sortie du classifieur n'est pas une mesure de confiance robuste, regardons ce que donne l'autre type de mesure de confiance dont le but est justement de quantifier la confiance par rapport au modèle.

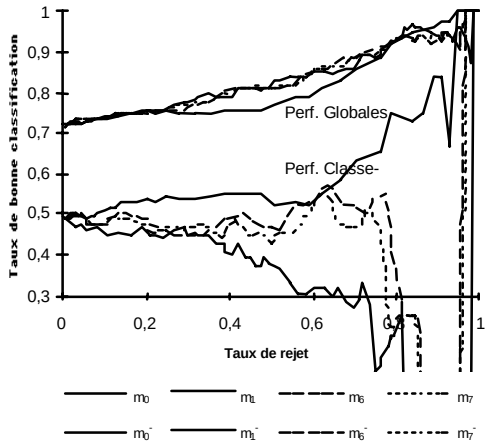

Figure 6 : Détection de Pannes. Courbes Performance/Rejet pour les mesures de confiance par rapport au modèle (m_0), par rapport à la sortie du classifieur (m_1) et pour les combinaisons des deux (m_6 et m_7). Les courbes du haut (Perf. Globales) représentent les performances globales du classifieur, celles du bas (Perf. Classe-) représentent les performances pour la classe la moins représentée.

La figure 6 propose les courbes Performance/Rejet de m_0 pour le classifieur (courbe "Perf. Globales" en trait plein) et pour la classe "panne" (courbe "Perf. Classe-" en trait plein). De manière globale, m_0 est légèrement moins bonne que m_1, par contre elle donne de très bons résultats pour la classe "panne".
Il semble judicieux de combiner m_0 et m_1, obtenant ainsi m_6 (combinaison linéaire dont les coefficients sont appris grâce aux données de test) et m_7 (moyenne de m_0 et m_1). La qualité de ces deux mesures est donnée, table 2, en utilisant uniquement la mesure de qualité Q_1.
Il est visible que m_6 et m_7 sont plus intéressantes que m_0 et m_1 pour la mesure globale. Par contre, phénomène étonnant, il semble que la moyenne simple (m_6) donne des résultats aussi bons que la combinaison apprise sur les données (m_7). Cela met bien en évidence la difficulté d'apprendre correctement la combinaison: il faut avoir à peu près autant d'exemples bien classés que d'exemples mal classés pour l'apprentissage, ce qui est en contradiction avec le fait que l'on recherche un classifieur le plus performant possible, et donc le moins possible d'exemples mal classés.
Avec m_1 les performances pour la classe la moins représentée chutaient dès que l'on rejetait des points. m_6 et m_7 permettent d'atténuer ce phénomène : les performances stagnent autour de 50% jusqu'à un taux de rejet important (80%). Ces deux mesures de performances semblent donc plus robustes.

Mesure de confiance	**Mesure de qualité Q_1**	
	Globale	*Classe-*
m_0	0.236 [0.221 - 0.251]	0.176 [0.161 - 0.191]
m_1	0.270 [0.255 - 0.285]	-0.196 [-0.211 - -0.181]
m_6	0.296 [0.281 - 0.311]	0.024 [0.009 - 0.039]
m_7	0.295 [0.280 - 0.310]	0.059 [0.044 - 0.074]

Table 2 : Détection de Pannes. Mesure de qualité Q_1 pour les mesures de confiance par rapport au modèle (m_0), par rapport à la sortie du classifieur (m_1) et pour les combinaisons des deux (m_6 et m_7). La colonne (Globale) représente la qualité globale de la mesure de confiance tandis que celle de droite (Classe-) représente la qualité de la mesure de confiance par rapport à la classe la moins bien représentée. L'intervalle entre crochets indique l'intervalle de confiance correspondant.

5. Conclusion et Perspectives

Nous avons montré que les sorties d'un système estimant les probabilités *a posteriori* des classes permettent de définir une mesure locale de la confiance dans le classifieur. Cette mesure de confiance peut être employée pour fixer une règle de décision avec rejet des exemples ambigus (exemples bruités, exemples de mélange). Un autre exemple d'application fréquent en cas de diagnostic d'un système complexe est la combinaison de plusieurs classifieurs en fonction de leur confiance respective.
Deux familles de mesures de confiance ont été présentées :

- les mesures permettant de quantifier la confiance dans le PMC utilisé comme modèle,
- les mesures permettant de quantifier la confiance dans la décision du PMC pour un exemple donné.

Ces deux familles de mesures peuvent être combinées et leur évaluation peut être effectuée grâce aux critères fondés sur la courbe Performance/Rejet et sur le pouvoir discriminant de ces mesures.
Les résultats numériques montrent d'une part, l'intérêt d'utiliser les mesures de confiance proposées pour élaborer des règles pertinentes et pour combiner efficacement différents classifieurs "experts".

Par la suite, nous privilégierons deux directions de recherche :

- la poursuite de l'analyse empirique des mesures de confiance,
- l'incorporation de ces mesures dans la procédure d'apprentissage.

Plus précisément, nous compléterons cette étude par l'analyse du comportement des mesures de confiance en fonction de la taille de l'ensemble d'exemples appris (comportement asymptotique), puis en fonction de la complexité des réseaux employés. Cela devrait nous permettre de relier la notion de confiance avec l'erreur en généralisation.
Ensuite, nous nous intéresserons à la procédure d'apprentissage des classifieurs (en particulier, des réseaux de neurones artificiels).
Vu l'intérêt des mesures de confiance pour la prise de décision, il apparaît opportun d'incorporer dans la fonction de coût à minimiser une mesure de la confiance. Cette approche est à comparer avec les travaux de (Shapire et al. 98, Masson et al. 98, Lemaire 99) qui emploient pour l'apprentissage de leurs systèmes le principe de la maximisation de la marge, notion qui correspond à l'idée de mesure de confiance restreinte aux exemples d'apprentissage.

Références

[1] Chow, C.K. 1970. On Optimum Recognition Error and Reject Tradeoff. IEEE Transactions in Information Theory 16(1):41-46.

[2] Devijver, P.A. and Kittler, J. 1982. Pattern Recognition : a Statistical Approach. Prentice Hall.

[3] Dubuisson, B. and Masson, M.H. 1993. A Statistical Decision Rule With Incomplete Knowledge About Classes. Pattern Recognition 26(1):155-165.

[4] Dubuisson, B.; Masson, M.H. and Frélicot, C. 1996. Some Topics in Using Pattern Recognition for System Diagnosis. Engineering Simulation 13:863-888.

[5] Efron, B. and Tibshirani, R.J. 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57. Chapman & Hall.

[6] Lee, D.-S.; Srihari, S.N. and Gaborski, R. 1991. Bayesian and neural network pattern recognition : a theoritical connection and empirical results with handwritten characters. in Artificial Neural Networks and Statisticial Pattern Recognition : Old and New Connections. Sethi, I.K. and Jain A.K. eds. Elsevier Science Publishers B.V.

[7] Lengellé, R.; Hao, Y.; Schaltenbrand, N. and Denoeux, T. 1991. Ambiguity and Distance Rejection Using Multilayer Neural Networks. in Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks (ANNIE'91). 299-304.

[8] Lemaire, V.; Bernier, O.; Collobert, D. and Clérot, F. 1999. Une nouvelle fonction de coût régularisante dans les réseaux de neurones artificiels : Application à la classification. in Proceedings of CAP'99.

[9] Leray, P. 1998. Apprentissage et Diagnostic de Systèmes Complexes: réseaux de neurones et réseaux bayesiens. Application à la gestion en temps réel du trafic téléphonique français. Thèse de Doctorat de l'Université Paris 6.

[10] Lippmann, R.P.; Kukolich, L. and Shahian, D. 1995. Predicting the Risk of Complications in Coronary Artery Bypass Operations using Neural Networks. Neural Information Processing System 7.

[11] Mason, L.; Bartlett, P. and Baxter, J. 1998. Direct optimization of margins improves generalization in combined classifiers, Technical Report, Deparment of Systems Engineering, Australian National University.

[12] MacKay D., The Evidence Framework Applied to Classification, Neural Computation, 1992, 4(2):720-736.

[12] Nix, D.A. and Weigend, A.S. 1995. Learning Local Error Bars for Nonlinear Regression. Neural Information Processing System 7.

[13] Richard, M.D. and Lippmann, R.P. 1991. Neural Network Classifiers Estimate Bayesian a posteriori Probabilities. Neural Computation 3:461-483.

[14] Schapire, R.E.; Freund, Y.; Bartlett, P. and Lee, W.S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics, 26(5):1651-1686.

[15] Shimshoni, Y. and Intrator, N. 1996. Classifying Seismic Signals by Intregrating Ensemble of Neural Networks. in Proceedings of ICONIP'96.

[16] Tibshirani, R.J. 1996. A Comparison of Some Error Estimates for Neural Network Models. Neural Computation 8:152-163.

[17] Tresp, V. and Taniguchi, M. 1995. Combining estimators using non-constant weighting functions. Neural Information Processing Systems 7:419-426.

[18] Wan, E.A. 1990. Neural Network Classification : a Bayesian Interpretation. IEEE Transactions on Neural Networks 1(4):303-305.

[19] White, H. 1989. Learning in Artificial Neural Networks : A Statistical Perspective. Neural Computation 1:425-464.

[20] Zaragoza, H. and d'Alché-Buc, F. 1998. Confidence Measures for Neural Network Classifiers. in proceedings of IPMU'98.