

Méthode d'apprentissage de réseaux de neurones à base de fonctions radiales avec des observations bruitées

Abd-Krim SEGHOUANE, Gilles FLEURY

École Supérieure d'Électricité - Service des Mesures,
3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette, France
Abd-krim.Seghouane@supelec.fr, Gilles.Fleury@supelec.fr

Résumé – Cet article traite du problème de l'apprentissage des réseaux de neurones à base de fonctions radiales pour l'approximation de fonctions non linéaires L_2 de \mathcal{R}^n vers \mathcal{R} . Pour ce faire, les algorithmes hybrides sont les plus utilisés. Ils font appel à des techniques d'apprentissage non supervisé pour l'estimation des centres et des paramètres d'échelle des fonctions radiales, et à des techniques d'apprentissage supervisé pour l'estimation des paramètres linéaires. La méthode d'apprentissage supervisé repose généralement sur l'estimateur des moindres carrés (MC). Cette méthode est optimale dans le cas où le jeu de données d'apprentissage $(\mathbf{x}_i, y_i)_{i=1,2,\dots,N}$ est constitué de sorties y_i , $i = 1, \dots, N$ bruitées et d'entrées \mathbf{x}_i , $i = 1, \dots, N$ exactes. Cependant lors de la collecte des données expérimentales il est rarement possible de mesurer l'entrée \mathbf{x}_i sans bruit. L'utilisation de l'estimateur des MC produit une estimation biaisée des paramètres linéaires dans le cas où le jeu de données d'apprentissage est à entrées et sorties bruitées, ce qui engendre l'estimation d'une sortie biaisée. Cet article propose l'application d'un critère basé sur le modèle d'erreur en les variables pour l'estimation des paramètres linéaires (apprentissage supervisé) dans le cas où le jeu de données d'apprentissage est à entrée et sortie bruitées. L'amélioration des performances d'approximation de fonctions non linéaires est illustrée par un exemple.

Abstract – This paper deals with the problem of learning radial basis function neural networks to approximate non linear L_2 function from \mathcal{R}^n to \mathcal{R} . Hybrid algorithms are mostly used for this task. Unsupervised learning techniques are used to estimate the center and width parameters of the radial functions and supervised learning techniques to estimate the linear parameters. Supervised learning techniques are based on the least squares estimates. They are optimal when the training set $(\mathbf{x}_i, y_i)_{i=1,2,\dots,N}$ is composed of noisy outputs y_i , $i = 1, \dots, N$ and exactly known inputs \mathbf{x}_i , $i = 1, \dots, N$. However, when collecting data under a measurement experiment, it may not be possible to avoid noisy inputs \mathbf{x}_i . Then, the use of least squares estimates produce a biased estimation of the linear parameters in the case of noisy input output training set, which leads to a biased output estimation. This paper proposes the use of a criterion based on the error in variables model to estimate the linear parameters when the training set is made up of input and output data corrupted by noise. The improved performances of non linear function approximation is illustrated with a simulation example.

1 Introduction

Les réseaux de neurones dits "Feedforward" (RNF) et les réseaux de neurones à base de fonctions radiales (RNBFR) constituent une classe de modèles paramétriques largement utilisée en identification de systèmes non linéaires [1][2]. La justification de ceci est que ces réseaux avec une seule couche cachée peuvent approximer n'importe quelle fonction continue ayant un nombre fini de discontinuités sur tout compact [3][4]. Un net regain d'intérêt pour les RNBFR à été constaté ces dernières années dans divers domaines d'application, car offrant deux avantages majeurs par rapports aux RNF habituellement utilisés. Pour un problème donné, l'utilisation d'un RNBFR conduit généralement à une structure de modèle (nombre d'unités de la couche cachée) moins complexe que celle produite par un RNF [5]. La complexité de calcul induite par leur apprentissage est moindre que celle induite par l'apprentissage des RNF grâce à l'existence d'algorithmes hybrides [6]. Les performances d'un tel réseau dépendent de la forme des fonctions radiales utilisées (gaussi-

enne généralement), de la distribution de leurs centres dans l'espace d'entrée, de leurs matrices de covariance (paramètres d'échelle des fonctions de la base) et des paramètres reliant les unités de la couche cachée aux unités de la couche de sortie (paramètres linéaires). L'ensemble de ces paramètres est ajusté lors d'une phase d'apprentissage au cours de laquelle un ensemble de paires entrée-sortie expérimentales $(\mathbf{x}_i, y_i)_{i=1,2,\dots,N}$ est utilisé pour permettre aux RNBFR d'acquérir une relation entrée-sortie non linéaire. Généralement les algorithmes d'apprentissage utilisés supposent uniquement la sortie expérimentale y_i bruitée. Ceci est clairement une supposition erronée, car toute donnée produite expérimentalement est bruitée. Il est donc nécessaire que ces algorithmes d'apprentissage tiennent compte du bruit qui affecte l'entrée expérimentale \mathbf{x}_i . Dans cet article on s'intéressera à l'ajustement des paramètres linéaires des RNBFR dans un algorithme hybride dans le cas où les entrées-sorties expérimentales sont toutes deux bruitées. Les centres et les paramètres de largeur des unités de la couche cachée, étant obtenus

par des méthodes d'apprentissage non supervisé, ne sont pas affectés par le bruit d'entrée car ne font pas intervenir la relation entrée-sortie recherchée.

2 Formulation du problème

L'approximation par un RNBFR d'une fonction inconnue $f : \mathcal{R}^n \rightarrow \mathcal{R}$ avec $f \in L_2$ à partir d'un ensemble de données expérimentales constitué de paires entrée-sortie $(\mathbf{x}_i, y_i)_{i=1, \dots, N}$, consiste à déterminer le nombre optimal de fonctions constituant la base de fonctions radiales ainsi que le réglage de leurs paramètres. Le modèle généralement utilisé pour la description des données $(\mathbf{x}_i, y_i)_{i=1, \dots, N}$ est le modèle de régression :

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \dots, N \quad (1)$$

ou ε_i est un échantillon de bruit gaussien de moyenne nulle et de variance finie σ_s^2 . Cependant ce modèle suppose que les données \mathbf{x}_i ne sont pas bruitées, or ceci est une supposition fautive car toute donnée produite expérimentalement est bruitée (bruit dû aux appareils de mesure). Il est donc nécessaire de tenir compte du bruit sur les données d'entrées et de sa propagation aux données de sortie lors de l'estimation des paramètres. L'utilisation dans ce cas du modèle (1) induira un biais sur les paramètres estimés du modèle et donc sur la sortie prédite. On supposera le bruit sur le système v , le bruit de mesure de l'entrée η et le bruit de mesure de la sortie e , stationnaires, indépendants, gaussien de moyennes nulles et de variances (matrice de covariance) finies connues, soit :

$$\varepsilon = e + v, \quad (2)$$

représentant le bruit de mesure de la sortie et le bruit sur le système comme illustré en figure 1. Les échantillons de bruit d'entrée η_i et de sortie ε_i pour différent i sont supposés indépendants. Un modèle adapté à la description des données $(\mathbf{x}_i, y_i)_{i=1, \dots, N}$ dans ce cas est le modèle d'erreur en les variables [7] qui s'exprime comme suit :

$$\begin{aligned} y_i &= f(\mathbf{x}_i^*) + \varepsilon_i, & i = 1, \dots, N \\ \mathbf{x}_i &= \mathbf{x}_i^* + \eta_i \end{aligned} \quad (3)$$

Le problème dans ce cas consiste à estimer les paramètres du RNBFR ainsi que le vecteur des entrées expérimentales exactes $(\mathbf{x}_i^*)_{i=1}^N$.

3 Architecture et algorithme

3.1 Architecture des RNBFR

Une fonction continue $f : \mathcal{R}^N \rightarrow \mathcal{R}$, avec $f \in L_2$ peut être décrite par une combinaison linéaire de fonctions élémentaires appelées noyaux. Cette décomposition peut être générée par un réseau de neurones où chaque noyau est implanté par une unité de la couche cachée, comme représenté sur la figure 2. Chaque noyau de la couche cachée est associé à une région de l'espace d'entrée \mathcal{R}^N appelée région d'action. Dans le cas des RNBFR, la réponse

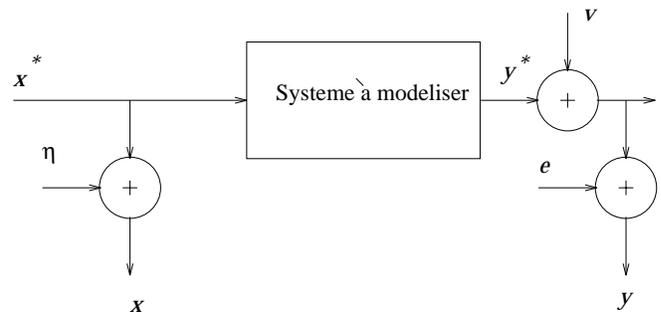


FIG. 1: Illustration du principe de génération de données expérimentales.

des noyaux est fonction de la distance entre l'entrée et un paramètre interne, appelé centre, modulée par un paramètre d'échelle :

$$\phi_l(\mathbf{x}) = \phi\left(\frac{\|\mathbf{x} - \mathbf{C}_l\|^2}{d_l^2}\right) \quad (4)$$

La distance généralement utilisée est la norme euclidienne sur \mathcal{R}^N . La réponse d'un RNBFR à une entrée $\mathbf{x} \in \mathcal{R}^N$ est donnée par :

$$\hat{f}_{RNBFR}(\mathbf{x}) = \lambda_0 + \sum_{l=1}^L \lambda_l \phi\left(\frac{\|\mathbf{x} - \mathbf{C}_l\|^2}{d_l^2}\right) \quad (5)$$

Divers types de fonctions peuvent être utilisées comme noyaux [8], la fonction gaussienne restant cependant la plus utilisée. Les performances d'un tel réseau en approximation dépendent donc du nombre de centres et par conséquent des noyaux, de leurs positions, de la valeur des paramètres d'échelle et de la méthode utilisée pour l'apprentissage de la relation entrée-sortie recherchée (ajustement des paramètres linéaires λ_l , $l = 1, \dots, L$).

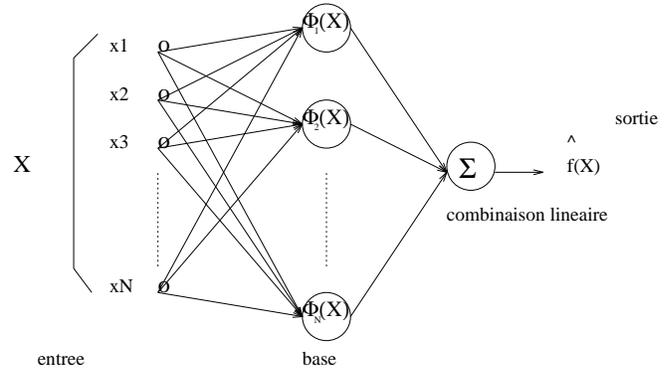


FIG. 2: Réseau de neurones à base de fonctions radiales, RNBFR.

3.2 Apprentissage d'un RNBFR

Le vecteur des paramètres d'un RNBFR, \mathbf{P} , est constitué des centres, des paramètres d'échelles et des paramètres linéaires et, est généralement ajusté de manière à min-

imiser l'EQM :

$$J(\mathbf{P}) = \frac{1}{N} \sum_{i=1}^N (y_i - f_{RNBFN}(\mathbf{x}_i, \mathbf{P}))^2, \quad (6)$$

sur un jeu de données d'apprentissage $(\mathbf{x}_i, y_i)_{i=1}^N$. Pour ceci deux principales stratégies sont proposées dans la littérature. La première est une méthode supervisée ou directe utilisant des algorithmes tel que le gradient [9] pour la recherche du vecteur optimal :

$$\mathbf{P}^* = \arg \min_{\mathbf{P}} J(\mathbf{P}). \quad (7)$$

La deuxième est une méthode hybride utilisant des techniques d'apprentissage non supervisé [10] pour l'ajustement des paramètres de la couche cachée, les centres et les paramètres d'échelle (ou en d'autres termes pour la recherche de la base de fonctions radiales optimale pour la projection), et des techniques d'apprentissage supervisé pour l'ajustement des paramètres linéaires. Les techniques d'apprentissage supervisé sont généralement basées sur l'estimateur des MC. Cette méthode est plus utilisée en pratique car plus rapide et moins chère en calcul que la méthode directe. Le problème est que ces méthodes (hybride ou directe) supposent toutes deux que le jeu de données d'apprentissage est décrit par le modèle de régression (1) et donc ne tiennent pas compte du bruit pouvant affecter l'entrée.

4 Construction du critère

Le critère développé ici peut être utilisé pour la phase d'apprentissage supervisé dans une méthode hybride. Le but étant la recherche de la valeur des paramètres linéaires (poids de sortie $\boldsymbol{\lambda}$) qui maximiseraient les performances du RNBFN pour un jeu de données à entrée et sortie bruitées. L'application des méthodes d'apprentissage non supervisé utilisées pour les entrées non bruitées (l'algorithme du "k-means clustering" [11] pour les centres et l'approche des M-plus proches voisins pour les paramètres d'échelle [6]) avec des entrées bruitées n'influent pas sur les performances du RNBFN. La justification est que la relation entrée-sortie recherchée n'intervient pas dans l'implémentation de ces algorithmes. Partant du modèle d'erreur en les variables :

$$\begin{aligned} y_i &= f(\mathbf{x}_i^*) + \varepsilon_i, & i &= 1, \dots, N \\ \mathbf{x}_i &= \mathbf{x}_i^* + \boldsymbol{\eta}_i \end{aligned} \quad (8)$$

ou $\mathbf{X}^* = (\mathbf{x}_1^*, \dots, \mathbf{x}_N^*)$ représente le vecteur des entrées exactes, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ le vecteur des entrées observées, $\boldsymbol{\eta}_i$ et ε_i sont deux échantillons indépendants de bruits normaux de moyennes nulles et de matrices de covariances diagonales à éléments identiques respectivement de valeur σ_e pour $\boldsymbol{\eta}_i$ et de variance σ_s pour ε_i . Dans ce cas les quantités inconnues sont le vecteur $\boldsymbol{\lambda}$ des paramètres linéaires ainsi que le vecteur des entrées exactes \mathbf{X}^* . La densité de probabilité conjointe de $\boldsymbol{\lambda}$ et \mathbf{X}^* *a posteriori* (compte tenu des résultats de mesures) est donnée par la

règle de Bayes

$$\varphi(\boldsymbol{\lambda}, \mathbf{X}^*/Y, \mathbf{X}) = \frac{\varphi(Y/\mathbf{X}^*, \boldsymbol{\lambda}, \mathbf{X})\varphi(\mathbf{X}/\boldsymbol{\lambda}, \mathbf{X}^*)\varphi(\boldsymbol{\lambda}, \mathbf{X}^*)}{\varphi(Y, \mathbf{X})}. \quad (9)$$

Comme Y et \mathbf{X} sont des résultats de mesures, donc connus, $\varphi(Y, \mathbf{X})$ est uniquement une constante de normalisation destinée à assurer que $\varphi(\boldsymbol{\lambda}, \mathbf{X}^*/Y, \mathbf{X})$ est bien une densité de probabilité,

$$\varphi(Y, \mathbf{X}) = \int \int \varphi(Y, \mathbf{X}/\boldsymbol{\lambda}, \mathbf{X}^*)\varphi(\boldsymbol{\lambda}, \mathbf{X}^*)d\boldsymbol{\lambda}d\mathbf{X}^*. \quad (10)$$

L'application du critère du maximum *a posteriori* avec l'hypothèse d'uniformité de la distribution conjointe $\varphi(\boldsymbol{\lambda}, \mathbf{X}^*)$ sur (6) est équivalent à son application sur

$$\varphi(\boldsymbol{\lambda}, \mathbf{X}^*/Y, \mathbf{X}) \propto \varphi(Y/\mathbf{X}^*, \boldsymbol{\lambda}, \mathbf{X})\varphi(\mathbf{X}/\boldsymbol{\lambda}, \mathbf{X}^*), \quad (11)$$

de plus on a,

$$\varphi(Y/\mathbf{X}^*, \boldsymbol{\lambda}, \mathbf{X}) \propto \varphi(Y/\mathbf{X}^*, \boldsymbol{\lambda}), \quad (12)$$

étant donné que la dépendance de Y à \mathbf{X} est la même que la dépendance de Y à \mathbf{X}^* , on obtient,

$$\varphi(\boldsymbol{\lambda}, \mathbf{X}^*/Y, \mathbf{X}) \propto \varphi(Y/\mathbf{X}^*, \boldsymbol{\lambda})\varphi(\mathbf{X}/\mathbf{X}^*), \quad (13)$$

et

$$\begin{aligned} \{\hat{\boldsymbol{\lambda}}, \hat{\mathbf{X}}^*\} &= \arg \min_{\boldsymbol{\lambda}, \mathbf{X}^*} \sum_{k=1}^N \left[\left(\frac{y_k - f_{RNBFN}(\mathbf{x}_k^*, \boldsymbol{\lambda})}{\sigma_s} \right)^2 \right. \\ &\quad \left. + \left(\frac{\|\mathbf{x}_k - \mathbf{x}_k^*\|}{\sigma_e} \right)^2 \right]. \end{aligned} \quad (14)$$

L'amélioration des performances d'approximation de fonctions non linéaires dans le cas où toutes les observations sont bruitées, lors de l'utilisation de ce critère est illustré par l'exemple suivant.

5 Exemple d'application

Afin d'illustrer la méthode développée dans cet article, nous allons utiliser un RNBFN pour l'approximation de la fonction $f : \mathcal{R} \rightarrow \mathcal{R}$ définie par

$$y = f(x) = (1 + 2x - x^2)e^{-x^2}, \quad (15)$$

avec $x \in [-4, 4]$. Le jeu de données expérimentales utilisé pour l'apprentissage du RNBFN est à entrée et sortie bruitées, et est illustré sur la figure 3. Le bruit ajouté aux données est obtenu par échantillonnage de la loi normale centrée de variance $\sigma_e = 0.02$ pour les entrées et de variance $\sigma_s = 0.15$ pour les sorties. Un RNBFN avec trois fonctions gaussiennes de projection est utilisé pour l'approximation. Afin d'effectuer une comparaison des résultats, deux méthodes d'apprentissage hybrides sont utilisées. La même technique d'apprentissage non supervisé est utilisée pour les deux méthodes [6], à savoir l'algorithme de "k-means clustering" pour le calcul de la valeur des centres des fonctions gaussiennes et un paramètre d'échelle identique pour les trois fonctions radiales de la base. Par

contre l'estimateur des MC est utilisé comme technique d'apprentissage supervisé pour la première méthode et l'estimation au sens du critère établi en (14) sera utilisée comme technique d'apprentissage supervisé pour la deuxième méthode. Soit f_1 la fonction obtenue par l'ajustement des paramètres linéaires du RNBFR par la première méthode et f_2 la fonction obtenue par l'ajustement des paramètres linéaires du RNBFR par la deuxième méthode. La valeur de l'EQM entre la réponse du réseau et la valeur de la fonction donnée en (15) est donnée dans le tableau 1. On peut remarquer une nette amélioration des performances pour f_2 .

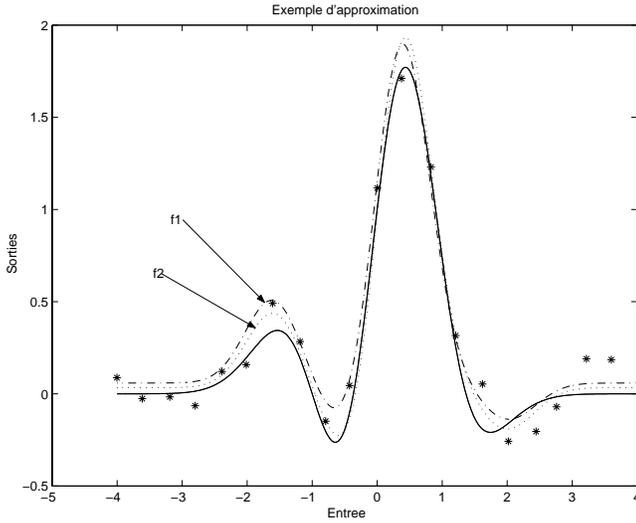


FIG. 3: Exemple de résultat d'approximation, '*' données d'apprentissage, '-' estimation obtenue par RNBFR avec MC et '...' estimation obtenue par RNBFR avec le critère (14).

TAB. 1: Erreur d'approximation

Approximation	EQM
f_1	0.0105
f_2	0.0039

6 Conclusion

L'identification des paramètres d'un RNBFR à partir d'un jeu de données expérimentales constitué d'entrées et de sorties bruitées a été étudiée. L'utilisation des algorithmes hybrides existant ne permet pas l'obtention d'une bonne estimation de la réponse du RNBFR pour une entrée donnée. En effet l'utilisation d'une méthode basée sur l'estimateur des moindres carrés ne tient pas compte de l'erreur présente sur les entrées d'apprentissage et de leur propagation aux sorties correspondantes. Ceci va donc biaiser l'estimation des paramètres linéaires du RNBFR, l'estimation de la valeur des centres et des paramètres d'échelles ne faisant pas intervenir la relation entrée-sortie recherchée. Pour la réduction de ce biais nous avons indro-

duit une méthode d'apprentissage supervisé reposant sur la minimisation d'un critère basé sur le modèle d'erreur en les variables. L'estimation nécessite dans ce cas la connaissance des variances (matrice de covariance dans le cas multidimensionnel) du bruit à l'entrée σ_e et à la sortie σ_s . Une estimation de ses grandeurs peut être obtenue par la répétition des mesures. Cependant, la complexité de calcul augmente car en plus des paramètres linéaires, les entrées expérimentales non bruitées doivent être estimées. Le gain obtenu par l'utilisation de cette méthode est la réduction du biais sur les paramètres linéaires et donc l'amélioration des performances d'approximation. Un exemple de simulation montrant cette amélioration lors de l'utilisation de ce critère dans la phase supervisée de l'apprentissage a été présenté.

Références

- [1] K. S. Narendra, and K. Parthasarathy. *Identification and control of dynamical systems using neural networks*. IEEE Transactions on Neural Networks, Vol. 1, pp. 4-27, 1990.
- [2] S. Chen, and S. A. Billings. *Neural networks for non-linear system modelling and identification*. International Journal of Control, Vol. 56, pp. 319-346, 1992.
- [3] K. Hornik, M. Stinchcombe, and H. White. *Multi-layer feedforward networks are universal approximators*. Neural Networks, Vol. 2, pp. 359-366, 1989.
- [4] J. Parks, and I. W. Sandberg. *Universal approximation using radial-basis function networks*. Neural Computation, Vol. 3, pp. 246-257, 1991.
- [5] S. Lee, and R. M. Kil. *A Gaussian potential function network with hierarchically self-organizing learning*. Neural Networks, Vol. 4, pp. 207-224, 1991.
- [6] J. E. Moody, and C. J. Darken. *Fast learning in networks of locally-tuned processing units*. Neural Computation, Vol. 1, pp. 281-294, 1989.
- [7] P. M. Reilly, and H. P. Leal. *A Bayesian study of the error-in-variables model*. Technometrics, Vol. 23, pp. 221-231, 1981.
- [8] B. A. Whitehead, and T. D. Choate. *Cooperative-competitive genetic evolution of radial basis function centers and widths for time series prediction*. IEEE Transactions on Neural Networks, Vol. 7, pp. 869-880, 1996.
- [9] N. B. Karayiannis. *Reformulated radial basis neural networks trained by gradient descent*. IEEE Transactions on Neural Networks, Vol. 10, pp. 657-671, 1999.
- [10] N. B. Karayiannis, and G. W. Mi. *Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques*. IEEE Transactions on Neural Networks, Vol. 8, pp. 1492-1506, 1997.
- [11] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum, New York, 1981.