Approche Régularisée du Traitement de Données Fonctionnelles par un Perceptron Multi-Couches[0]

Brieuc CONAN-GUEZ$*$ et Fabrice ROSSI[1] $**$

$*$ *INRIA-Rocquencourt, Domaine De Voluceau,*
BP 105 Bâtiment 18
78153 Le Chesnay Cedex, France
Brieuc.Conan-Guez@inria.fr
$**$ *LISE/CEREMADE, UMR CNRS 7534, Université Paris-IX Dauphine,*
Place du Maréchal de Lattre de Tassigny,
75016 Paris, France
Fabrice.Rossi@dauphine.fr

RÉSUMÉ. *Nous présentons dans cet article une extension naturelle du Perceptron Multi-Couches (PMC) à des données fonctionnelles : le Perceptron Multi-Couches Fonctionnel (PMCF). Contrairement à l'approche développée lors de précédents travaux, basée sur la manipulation directe des fonctions d'entrée, la méthode présentée ici s'appuie sur la projection préalable des individus sur une base de fonctions. Nous montrons que le modèle proposé est un approximateur universel et que l'estimation de ses paramètres optimaux est consistante.*

MOTS-CLÉS : *analyse de données fonctionnelles, perceptron multi-couches, approximateur universel, consistance*

1. Introduction

L'Analyse de Données Fonctionnelles (ADF, voir [RAM 97]) est une extension de l'analyse de données traditionnelles à des données fonctionnelles. Dans cette approche, chaque individu est décrit par une ou plusieurs fonctions réelles, plutôt que par un vecteur de $\mathbb{R}^n$. L'avantage majeur de l'ADF sur l'approche multivariée est de prendre en compte de manière naturelle la régularité des fonctions étudiées. Si l'on considère par exemple la croissance d'un groupe d'enfants au cours du temps, l'approche multivariée ne permet pas de modéliser correctement les liens existants entre les valeurs successives observées pour un enfant donné. En ADF en revanche, cette régularité est prise en compte de manière explicite en représentant la croissance de chaque enfant par une fonction continue. Un second avantage de l'ADF sur l'approche multivariée est qu'elle n'impose pas de contraintes sur la connaissance que l'on a de chaque fonction. En effet, dans la pratique chaque fonction est décrite par un nombre fini de couples d'entrée/sortie $(x_i, f(x_i))$. L'analyse multivariée impose que les points d'évaluation x_i soient identiques d'une fonction à une autre afin que les variables restent comparables. Dans l'approche fonctionnelle en revanche, cette contrainte n'existe plus : l'échantillonnage de chaque fonction peut différer d'une fonction à une autre (nombre de points d'évaluation et position de chacun de ces points).

Dans [ROS 01, ROS 02b, ROS 02a], nous avons présenté le Perceptron Multi-Couches Fonctionnel (PMCF) qui étend les Perceptrons Multi-Couches (PMC) classiques aux données fonctionnelles. Dans la pratique, le modèle proposé doit nécessairement être remplacé par un modèle approché car les calculs ne sont pas possibles en général. Dans nos précédents travaux, nous avons utilisé un traitement direct des données par le PMCF. Pour accélérer les traitements, nous proposons dans

0. Publié dans les actes des Neuvièmes journées de la Société Francophone de Classification.
Disponible à l'URL `http://apiacoa.org/publications/2002/sfc-proj02.pdf`
1. Les coordonnées actuelles de Fabrice Rossi sont disponibles à l'URL `http://apiacoa.org/`

le présent article, une approche basée sur une étape préliminaire de régularisation des fonctions manipulées. Cette méthode, plus traditionnelle en ADF, repose sur l'utilisation d'un opérateur de projection sur une base de fonctions choisie au préalable (B-spline, séries de Fourier).

Dans la section 2, nous présentons le PMCF. Nous introduisons l'approche régularisée dans la section 3 et nous montrons que les PMCF ainsi utilisés sont des approximateurs universels. La section 4 est consacrée aux propriétés statistiques du modèle, en particulier la consistance. La section 5 relate des résultats d'expériences sur données synthétiques.

2. PMC Fonctionnel

Soit μ une mesure finie sur $\mathbb{R}^r$, $L^2(\mu)$ l'espace des fonctions mesurables de carré intégrable, et w un élément de cet espace. On définit alors un neurone fonctionnel comme étant la fonction qui a tout élément g de $L^2(\mu)$ associe le scalaire $T(b + \int wg\, d\mu)$ (où b est un réel et où T est une fonction d'activation de $\mathbb{R}$ dans $\mathbb{R}$). Dans la suite de cet article, la fonction w portera le nom de "fonction de poids".

Comme le neurone fonctionnel est à valeurs réelles, on peut construire un PMCF en combinant des neurones fonctionnels et des neurones numériques selon une organisation par couches à la manière d'un PMC classique. La première couche du réseau est composée exclusivement de neurones fonctionnels alors que les couches suivantes ne sont constituées que de neurones numériques. Par exemple, un PMCF avec une couche cachée et une sortie réelle est défini par l'expression suivante :

$$H(g) = \sum_{k=1}^{K} a_k T\left(b_k + \int g w_k \, d\mu\right), \qquad [1]$$

où g et les w_k sont des élément de $L^2(\mu)$, et où les a_k et les b_k sont des scalaires.

3. L'approche par projection

3.1. *Projection*

Comme on peut le voir dans l'équation 1, l'évaluation du PMCF implique le calcul des intégrales de la première couche $\int g w_k \, d\mu$, ce qui n'est pas possible directement en pratique. Dans [ROS 01, ROS 02b, ROS 02a], nous résolvons ce problème en remplaçant chaque intégrale par une moyenne empirique. Dans le présent article, nous effectuons au contraire un calcul exact grâce à une représentation régularisée des fonctions g et w_k (stratégie usuelle en ADF).

Pour cela, on considère $(\phi_p)_{p\in\mathbb{N}^*}$ une base topologique de $L^2(\mu)$, et on note Π_P l'opérateur de projection sur l'espace vectoriel engendré par les P premiers éléments de la base (noté $span(\phi_1, ..., \phi_P)$), i.e. $\Pi_P(g) = \sum_{p=1}^{P}(\int \phi_p g \, d\mu)\phi_p$. Cette étape de projection permet de réaliser une première simplification du modèle : on ne cherche plus à évaluer le PMCF sur une fonction (i.e. $H(g)$), mais sur son projeté (i.e. $H(\Pi_P(g))$).

De la même façon, on représente les fonctions de poids par les éléments de $span(\psi_1, ..., \psi_Q)$, l'espace vectoriel engendré par les Q premiers éléments d'une seconde base topologique $(\psi_q)_{q\in\mathbb{N}^*}$ de $L^2(\mu)$. Si l'on considère la fonction de poids définie par $w = \sum_{q=1}^{Q} \alpha_q \psi_q$, et la fonction d'entrée g, l'expression des intégrales de l'équation 1 se simplifie en :

$$\int w \Pi_P(g) \, d\mu = \sum_{p=1}^{P} \sum_{q=1}^{Q} (\int \phi_p \, g \, d\mu) \alpha_q \int \phi_p \psi_q \, d\mu \qquad [2]$$

Il nous reste donc à calculer les $\int \phi_p \psi_q \, d\mu$ (expressions indépendantes des fonctions d'entrée et des α_q). En choisissant des bases appropriées (par exemple B-splines ou séries de Fourier), l'évaluation exacte de ces intégrales est aisée.

Le point important à noter est que le PMCF est à présent paramétré par un nombre fini de poids réels (chaque fonction de poids est en effet totalement décrite par ses coordonnées $(\alpha_q)_{1\leq q\leq Q}$ sur

$span(\psi_1, ..., \psi_Q)$). L'apprentissage du PMCF peut donc être réalisé par les algorithmes d'optimisation usuels (algorithme de type gradient). De plus, un algorithme de rétro-propagation peut être appliqué à ce modèle pour un calcul efficace du gradient.

Notons que pour obtenir ce calcul simplifié, le lien entre les paramètres numériques des fonctions de poids et ces fonctions elles-mêmes doit nécessairement être linéaire dans l'approche régularisée, alors qu'on peut utiliser des modèles non linéaires (par exemple un PMC) pour l'approche directe de [ROS 02b].

3.2. *Approximateur universel*

On démontre dans [CON 02] que le PMCF avec une étape préliminaire de régularisation est un approximateur universel. Plus précisément, on a :

Corollaire *Soit F une fonction continue d'un sous-ensemble compact K de $L^2(\mu)$ dans $\mathbb{R}$, soit ε un réel strictement positif, et soit T une fonction d'activation continue non polynomiale. Alors il existe un entier P positif, un entier Q positif et un PMC Fonctionnel H dont les fonctions de poids sont restreintes au sous-ensemble $span(\psi_1, ...\psi_Q)$ tel que $\parallel Ho\Pi_P - F \parallel_\infty < \varepsilon$.*

4. Implantation

4.1. *Connaissance limitée de chaque fonction*

Dans la pratique, chaque fonction d'entrée g^i n'est connue que grâce à un nombre fini de couples d'entrée/sortie $(x_j^i, g^i(x_j^i) + \varepsilon_j^i)$. Une manière naturelle de modéliser cette connaissance est d'interpréter chaque x_j^i comme étant la réalisation d'une variable aléatoire X_j^i définie sur l'espace $(\Omega, \mathcal{A}, \mathcal{P})$ et à valeurs dans $\mathbb{R}^r$ (les X_j^i sont indépendantes identiquement distribuées de loi P_X). De même, on suppose que ε_j^i est une réalisation de la variable aléatoire $\mathcal{E}_j^i$ avec $E(\mathcal{E}_j^i) = 0$ et $E(\mathcal{E}_j^i)^2 = \sigma^2$ (les $\mathcal{E}_j^i$ sont indépendantes identiquement distribuées). Il est alors naturel de considérer des fonctions g^i éléments de $L^2(P_X)$.

On définit $\Pi_P(g^i)^m$ comme la fonction $\sum_{p=1}^P \beta_p \phi_p$ qui minimise l'expression $\sum_{j=1}^m (g^i(x_j^i) + \varepsilon_j^i - \sum_{p=1}^P \beta_p \phi_p(x_j^i))^2$. Dans la pratique, on ne cherche donc plus à évaluer l'expression $H(\Pi_P(g^i))$ mais la valeur approchée $H(\Pi_P(g^i)^m)$.

4.2. *Consistance*

Dans la pratique, à la connaissance limitée de chaque fonction s'ajoute le fait que nous ne connaissons qu'un nombre fini de couples d'entrée/sortie (g^i, y^i) (y^i est la valeur à prédire connaissant la fonction g^i). Malgré cette double limitation, on démontre dans [CON 02] un résultat de consistance qui est résumé ici.

On modélise chaque g^i comme étant une réalisation de la variable aléatoire fonctionnelle G^i (les G^i sont indépendantes et identiquement distribuées définies sur $(\Omega, \mathcal{A}, \mathcal{P})$ et à valeurs dans un sous-ensemble compact de $C(\mathcal{K}, \mathbb{R})$, où $\mathcal{K}$ est compact). On modélise de la même manière chaque y^i comme étant une réalisation de la variable aléatoire réelle Y^i (les Y^i sont i.i.d. définies sur $(\Omega, \mathcal{A}, \mathcal{P})$). On définit enfin la fonction de coût $l : l(g, y, w)$ mesure l'adéquation entre la valeur prédite par le PMCF (i.e. la valeur $H_w(\Pi_P(g))$ où w est le vecteur poids du PMCF) et la valeur à prédire y. On cherche donc à minimiser l'espérance de l, i.e., $\lambda(w) = E(l(G, Y, w))$. On définit de manière analogue l'erreur empirique par $\lambda_m^n(w)(\omega) = \frac{1}{n}\sum_{i=1}^n l(G^i(\omega), Y^i(\omega), w)^m$, où $l(G^i(\omega), Y^i(\omega), w)^m$ mesure l'adéquation entre la valeur prédite par le PMCF à partir du projeté empirique et la valeur à prédire.

On note $w_m^n(\omega)$ un minimiseur de $\lambda_m^n(w)(\omega)$ et W^* l'ensemble des minimiseurs de $\lambda(w)$. On démontre dans [CON 02] que pour presque tout $\omega \in \Omega$, $\lim_{n\to\infty} \lim_{m\to\infty} d(w_m^n(\omega), W^*) = 0$.

5. Simulations

Nous avons comparé l'approche proposée ici à celle développée dans [ROS 02a] sur des données synthétiques pour des problèmes de discrimination de fonctions. Par exemple, nous cherchons à séparer les fonctions de la forme $f_d(x) = \sin(2\pi(x-d))$ des fonctions de la forme $g_d = \sin(4\pi(x-d))$. Nous engendrons des fonctions d'exemple pour une classe de la façon suivante :

1. d est choisi aléatoirement dans $[0, 1]$ (distribution uniforme) ;
2. on choisit aléatoirement et uniformément dans $[0, 1]$ 25 points d'évaluation ;
3. on ajoute aux fonctions un bruit gaussien de moyenne nulle et d'écart-type 0.7.

Nous avons entraîné les PMCF avec un algorithme de gradient conjugué. L'ensemble d'apprentissage consiste en 50 fonctions de chaque classe. On choisit les meilleurs paramètres pour chaque PMCF grâce à un ensemble de validation contenant 50 fonctions de chaque classe. Enfin, les performances sont évaluées sur un ensemble de test contenant 150 fonctions pour chaque classe.

Les fonctions de poids sont représentées pour les deux approches par des B-splines ce qui donne deux architectures strictement identiques et donc directement comparables. L'approche par régularisation utilise de la même façon des B-splines pour représenter les fonctions d'entrées. Quand on utilise 3 neurones fonctionnels, l'approche directe de [ROS 02a] donne un taux de reconnaissance de 94.4%, alors que l'approche régularisée permet d'atteindre 97%. De plus cette dernière est environ 20 fois plus rapide (dans sa phase d'apprentissage) que l'approche directe.

D'autres expériences (par exemple l'étude en dimension 2 proposée dans [ROS 02a]) montrent que l'approche régularisée est efficace quand la dimension d'entrée des fonctions manipulées reste faible. La possibilité d'utiliser des fonctions de poids dépendant non linéairement de leurs paramètres avec l'approche directe, donne à celle-ci une plus grande parcimonie : les PMCF "directs" utilisent moins de paramètres que les PMCF "régularisés".

6. Conclusion

Nous avons montré dans cet article que l'approche régularisée du PMCF possède deux propriétés théoriques importantes : le PMCF est un approximateur universel et l'estimation des paramètres optimaux est consistante. Les temps de calcul liés à l'approche par projection sont nettement inférieurs à ceux de l'approche directe. Ce gain de performance a cependant un prix : le choix des fonctions de poids est réduit à un sous-espace vectoriel de $L^2(\mu)$. Dans l'approche directe n'importe quelle famille de régresseurs peut être employée (par exemple des PMC). Le but de nos travaux actuels et futurs est la comparaison de ces deux approches sur des données réelles.

7. Bibliographie

[CON 02] Conan-Guez B., Rossi F., Projection Based Functional Multi Layer Perceptrons, rapport, april 2002, LISE/CEREMADE & INRIA, `http://www.ceremade.dauphine.fr/`.

[RAM 97] Ramsay J., Silverman B., *Functional Data Analysis*, Springer Series in Statistics, Springer Verlag, June 1997.

[ROS 01] Rossi F., Conan-Guez B., Fleuret F., Functional Multi Layer Perceptrons, rapport n°0134, december 2001, LISE/CEREMADE & INRIA, `http://www.ceremade.dauphine.fr/`.

[ROS 02a] Rossi F., Conan-Guez B., Fleuret F., Functional Data Analysis With Multilayer Perceptrons, *Proceedings of IJCNN 2002 (WCCI 2002)*, vol. 3, Honolulu, Hawai, USA, May 2002, IEEE/NNS/INNS, p. 2843–2848.

[ROS 02b] Rossi F., Conan-Guez B., Fleuret F., Theoretical Properties of Functional Multilayer Perceptrons, *Proceedings of ESANN 2002*, Bruges, Belgium, April 2002, p. 7–12.