

Rapport relatif au mémoire présenté par FABRICE ROSSI pour
l'Habilitation à Diriger des Recherches
intitulé :
Contributions à l'analyse des données complexes.

Monsieur FABRICE ROSSI présente un mémoire d'habilitation à diriger des recherches sur une sélection (6 articles, 2 actes) de ses travaux totalisant, entre autres, 12 articles publiés dans des revues à comité de lecture dont Neural Networks et Neurocomputing, un soumis et 25 actes de conférences internationales faisant apparaître une participation régulière au symposium européen sur les réseaux de neurones artificiels (ESANN). Ces travaux couvrent un large spectre de compétences en techniques neuronales mais aussi en Statistique mathématique autour de 2 thèmes correspondant à deux problématiques définies par le type des données étudiées :

1. analyse de données fonctionnelles,
2. analyse de tableaux de dissimilarités.

Dans les deux cas, l'objectif est de prendre en compte dans l'analyse, par des outils mathématiques appropriés, la complexité des données en évitant la "perte d'information" de l'approche réductrice classique qui consiste à se contenter d'une représentation vectorielle élémentaire de ces données.

Il est notable que si les actes de conférences s'échelonnent de 1994 à 2006, tous les articles et 13 des conférences ont été produits ces trois dernières années, traduisant ainsi une étonnante capacité d'innovation. Notons aussi la participation à l'encadrement de 4 thèses dont deux au sein du projet AxIS de l'INRIA.

1 Données fonctionnelles

Ce travail s'inscrit dans un courant de fond, initié dans la deuxième moitié des années 70, qui vise à adapter les techniques statistiques usuelles à des données fonctionnelles c'est-à-dire des données observées sous la forme de courbes ou fonctions d'un paramètre temporel ou non. Sur le plan théorique, les variables aléatoires à valeurs dans des espaces euclidiens de dimension finie deviennent des variables aléatoires prenant leurs valeurs dans des espaces fonctionnels de dimension infinie, le plus souvent des espaces hilbertiens de façon à préserver des propriétés de projection. Pour une méthode

de modélisation donnée, il s'agit donc de définir celle-ci dans des espaces abstraits, mais possédant des propriétés analogues à celles des espaces euclidiens (définition d'un produit scalaire), avant d'en construire une estimation à la fois par échantillonnage sur les observations et discrétisation des courbes ou fonctions observées, ou encore par représentation de celles-ci dans une base fonctionnelle adaptée (Fourier, spline, ondelettes). Échantillonnage et discrétisation nécessitent alors une étude asymptotique détaillée pour s'assurer de la bonne convergence de la démarche proposée. D'autres résultats classiques en modélisation statistique : lois limites, tests vitesses de convergence ou moins classiques : inégalités de type oracle, sont souvent difficiles à obtenir et donc moins développés.

La bibliographie sur ce sujet, devenu très concurrentiel ces dernières années, est très importante et la contribution originale de Fabrice Rossi fut de compléter significativement cette démarche en traitant le cas de deux grandes familles de méthodes : les modèles neuronaux (perceptron, RBF, SOM) et les machines à vecteurs supports, de façon à ce qu'elles puissent prendre en compte efficacement des problèmes de régression ou discrimination sur données fonctionnelles.

1.1 Méthodes neuronales

Dans le cadre du perceptron multicouche, l'adaptation de ce modèle proposée par Rossi et Conan-Guez consiste à introduire une première couche de neurones pour lesquels la fonction de transfert fait intervenir l'intégrale d'un produit entre la fonction en entrée et une fonction poids ; c'est-à-dire aussi un produit scalaire dans un espace hilbertien adapté. Ayant défini ce modèle, les auteurs montrent qu'il vérifie bien les propriétés d'approximation universelle classique en méthode neuronale sous des conditions très générales. Ils posent ensuite les problèmes pratiques de discrétisation (des fonctions pour approcher les calculs d'intégrales) et estimation des paramètres (sur un compact d'un ensemble de dimension finie) en vue d'une implémentation pour aboutir à un théorème de convergence presque sûre lorsque, à la fois, le nombre d'observations tend vers l'infini et le pas de discrétisation devient suffisamment fin. Ces résultats, qui demandent des développements mathématiques très techniques, valident en profondeur la démarche proposée.

Deux approches, nécessitant des études de convergence différentes sont ensuite abordées pour aboutir à des solutions raisonnables en terme d'implémentation lorsqu'il s'agit de calculer des dérivées pour résoudre les problèmes d'optimisation liés à l'apprentissage des poids du réseau. Ces deux approches sont très similaires, la première propose une hypothèse simplificatrice sur la fonction des neurones d'entrée (linéarisation), la deuxième calcule une représentation préalable des données dans une base fonctionnelle (splines) pas nécessairement orthonormée. Dans ces deux cas, l'apprentissage sur données fonctionnelles s'apparente alors à un prétraitement consistant à calculer les coordonnées des fonctions en entrée du perceptron dans une base fonctionnelle appropriée de l'espace hilbertien L^2 tandis que la suite du processus est ramenée au fonctionnement classique d'apprentissage d'un réseau sur ces coefficients.

Se pose enfin les questions usuelles de contrôle de la complexité du modèle afin d'éviter les problèmes dus au sur-apprentissage. Une validation croisée permet d'optimiser le paramètre sensible, qui est le nombre de fonctions de base, donc la finesse de représentation des courbes, au même titre que le nombre de neurones, le taux d'apprentissage de l'algorithme ou encore la norme (decay) du vecteurs des poids.

Rossi et Conan-Guez montrent, sur des données classiques de la littérature (benchmarks), les bonnes propriétés prédictives de cette approche fonctionnelle des méthodes neuronales (perceptron RBF) appliquée à des courbes régulières (vagues de Breiman, spectres dans le proche infra-rouge, phonèmes...) en association à des bases de fonctions splines permettant de faire intervenir très simplement des dérivées de ces fonctions. L'avantage de la démarche proposée vient de la bonne imbrication réalisée entre base fonctionnelle et perceptron qui permet à celui-ci de profiter des très bonnes propriétés de parcimonie de l'approximation par fonctions splines et donc de bien prendre en compte les propriétés fonctionnelles des données caractérisées par leurs propriétés de régularité (dérivabilité).

A ce niveau, il aurait été intéressant de considérer d'autres types de données comme des spectres présentant de fortes singularités (chromatographies HPLC par exemple) afin de s'intéresser à d'autres types de fonctions de base (ondelettes) mieux adaptées aux caractéristiques locales des signaux.

1.2 Machines à vecteurs de support

De façon générale et comme dit en introduction de ce rapport, toute méthode de modélisation, représentation et réduction de dimension (Analyse en Composantes Principales), classification (clustering et SOM), construite à partir des outils mathématiques de base que sont les produits scalaires, normes ou distances se généralisent en la plongeant dans un espace hilbertien fonctionnel.

L'objectif est cette fois la discrimination de n courbes en deux classes et les machines à vecteurs de support (MVS) se prêtent tout aussi bien à cette démarche de généralisation. Notons d'ailleurs que cet outil, même pour des données en dimension finie, repose déjà principalement sur cette "astuce" : le problème d'estimation des vecteurs supports est basé sur l'expression d'une optimisation s'écrivant sous la forme d'une combinaison linéaire de produits scalaires. Ceci permet de définir cette optimisation dans tout espace hilbertien autoreproduisant associé à une fonction noyau et donc de ramener un problème de discrimination non linéaire à un problème linéaire par l'usage du bon noyau. C'est justement le choix de la "bonne" fonction noyau qui est important et abondamment discuté dans la littérature ; c'est celui-ci, en complément du "bon" réglage d'un coefficient de pénalisation, qui permet la bonne adéquation entre la méthode et le type des données pour aboutir à des résultats souvent très performants de discrimination.

La contribution de Rossi et Villa dans ce domaine est de deux ordres : le premier est une discussion fine et détaillée concernant le choix de certains noyaux

adaptés à certaines classes de données fonctionnelles et aux problèmes qu'elles soulèvent : noyaux gaussiens et polynomiaux et transformations par projection, centrage, dérivation... Le deuxième est l'étude et la preuve que les propriétés de consistance des MVS sont préservées par projection sur une base hilbertienne tronquée. Ils montrent également que la constante de pénalisation peut être optimisée sur un intervalle (pas seulement un espace dénombrable) en un temps de calcul raisonnable et cette propriété est intégrée à la preuve de consistance. Celle-ci nécessite l'adaptation des résultats très techniques les plus récents sur les inégalités "oracle" à la base du contrôle de l'erreur de prévision.

Les expériences réalisées montrent encore les très bonnes propriétés des outils implémentés sur différents jeux de données. D'autres travaux, encore en cours et notamment en collaboration avec Nathalie Villa, abordent des problèmes de taille industrielle.

2 Tableaux de dissimilarités

Cette thématique plus récente a été développée dans le cadre d'un projet INRIA. L'objectif en est l'analyse de données non plus fonctionnelles ou vectorielles mais présentant une structure arborescente comme l'exemple cité d'un texte codé en XML. La contribution de Fabrice Rossi, en collaboration avec Aïcha El Golli (doctorante), Yves Lechevallier et Brioux Coan-Guez est double. La première a consisté à proposer un algorithme efficace pour adapter les cartes auto-organisatrices à des données de type dissimilarités et en faire ainsi une méthode "générique" sur ce type de données. La deuxième fut de définir des dissimilarités adaptées à l'analyse de l'usage d'un site web c'est-à-dire aux données contenues dans les fichiers log retraçant le cheminement des utilisateurs.

Un premier algorithme proposé (DSOM pour Dissimilarity SOM) minimise une "énergie" en alternant itérativement et indépendamment deux phases (affectation et représentation). La preuve de convergence de cet algorithme est une adaptation directe d'une preuve existante et ses propriétés sont comparées avec celles des algorithmes concurrents notamment par des mises en œuvre sur des données synthétiques. Néanmoins, la complexité de cet algorithme supérieure à $\mathcal{O}(MN^2)$ (M : nombre de neurones, N : nombre d'observations) le rend inadapté au traitement de données réelles souvent volumineuses. Une adaptation de la phase de représentation de l'algorithme, ainsi que deux astuces d'implémentation réduisant le volume des calculs permet de diminuer considérablement les temps de calcul ($\mathcal{O}(N^2 + NM^2)$) et rend donc l'outil d'autant plus opérationnel sur des grands corpus de données réelles comme par exemple ceux associés à la gestion des sites web. Utilisé sur celui de l'INRIA, il apporte des informations précieuses sur la façon dont ce site est utilisé et perçu par les visiteurs extérieurs.

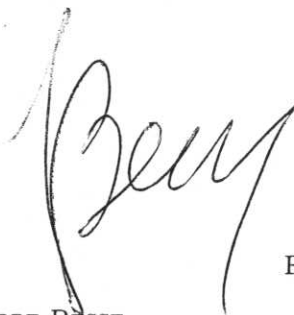
Commentaires

Tous les travaux présentés ont été réalisés en collaboration. Loin de dévaloriser le candidat, cette remarque souligne son dynamisme et ses capacités d'animation de la recherche.

L'analyse de données fonctionnelles comme celle de tableaux de dissimilarités, offrent des perspectives de développement encore considérables. Les pistes proposées par Fabrice Rossi concernent à la fois l'obtention de résultats théoriques afin d'affiner la connaissance de ces méthodes mais aussi plus d'expérimentation et de "savoir faire" pour les adapter précisément à certaines classes de problèmes ou types de données.

La diversité du travail n'empêche pas Fabrice Rossi de présenter un mémoire synthétique et bien structuré, mettant clairement en évidence ses apports dont ceux des outils mathématiques qu'il a judicieusement adaptés à ses problématiques, son travail d'implémentation informatique et les nombreux tests d'efficacité qu'il a mis en route afin de montrer expérimentalement la pertinence de la démarche.

En conclusion, Fabrice Rossi occupe une place originale à l'interface entre Mathématiques et Informatique. Soucieux d'apporter des réponses concrètes aux problèmes abordés par le développement et le codage d'algorithmes efficaces, il ne néglige par pour autant les preuves mathématiques, parfois très techniques, de convergence, validant en profondeur les outils mis en œuvre en allant toujours chercher les outils théoriques adaptés à cet objectif. Cette place, souvent ingrate et difficile à tenir entre deux disciplines, est pourtant une contribution essentielle à une fructueuse interaction et au dialogue entre différentes équipes. Nous pouvons espérer que Fabrice Rossi s'attachera également, dans l'avenir, à publier dans des revues de Mathématiques pour un dialogue qui alimentera aussi cette communauté en problématiques très motivantes. Le travail présenté montre que, confronté à des problèmes difficiles, il a rapidement acquis une grande maturité pour la direction, l'animation et l'encadrement de la recherche. Je suis très favorable à ce qu'il soutienne en l'état son mémoire pour obtenir l'Habilitation à Diriger des Recherches.



PHILIPPE BESSE
Professeur

Institut National des Sciences Appliquées de Toulouse.
Laboratoire de Statistique et Probabilités — UMR CNRS 5583

UNIVERSITE PAUL SABATIER
LABORATOIRE
DE STATISTIQUE ET PROBABILITES
U.M.R. C 5583
31062 TOULOUSE CEDEX 9

Fait à Toulouse, le 30 octobre 2006