Processus gaussiens parcimonieux pour la classification générative de données hétérogènes

Charles Bouveyron, Mathieu Fauvel, Stéphane Girard

Processus gaussiens parcimonieux pour la classification générative de données hétérogènes

Charles Bouveyron [1] & Mathieu Fauvel [2] & Stéphane Girard [3]

[1] *Laboratoire SAMM (EA 4543), Université Paris 1 Panthéon-Sorbonne*
charles.bouveyron@univ-paris1.fr
[2] *UMR 1201 DYNAFOR INRA & Université de Toulouse*
mathieu.fauvel@ensat.fr
[3] *Equipe MISTIS, INRIA Rhône-Alpes & LJK*
stephane.girard@inria.fr

Résumé. Nous proposons dans ce travail une famille de processus gaussiens parcimonieux permettant de construire, à partir d'un échantillon de taille finie, un classifieur génératif dans un espace de dimension (potentiellement) infinie. Ces modèles parcimonieux permettent en particulier d'utiliser des transformations non-linéaires des données projetant les observations dans un espace de dimension infinie. Nous montrons qu'il est possible de construire directement le classifieur depuis l'espace des observations au travers d'une fonction noyau. La méthode de classification proposée permet ainsi de classer des données de types variés (données qualitatives, données fonctionnelles, réseaux, ...). En particulier, il est possible de classer des données hétérogènes en combinant plusieurs fonctions noyaux. La méthodologie est également étendue au cas de la classification non supervisée (clustering).

Mots-clés. Classification, processus gaussiens, modèles parcimonieux, noyau

Abstract. This work presents a family of parsimonious Gaussian process models which allow to build, from a finite sample, a model-based classifier in a infinite dimensional space. These parsimonious models allow in particular to use non-linear mapping functions which project the observations in an infinite dimensional space. It is also demonstrated that the building of the classifier can be directly done from the observation space through a kernel. The proposed classification method is thus able to classify data of various types (categorical data, functional data, networks, ...). In particular, it is possible to classify mixed data by combining different kernels. The methodology is as well extended to the unsupervised classification case (clustering).

Keywords. Classification, Gaussian process, parsimonious models, kernel

1 Introduction

Depuis les travaux pionniers de Fisher en analyse discriminante, un très grand nombre de méthodes ont été proposés pour permettre la classification de données de types variés.

En effet, il existe une grande variété de données telles que les données quantitatives, catégorielles et binaires mais aussi des textes, des fonctions, des séquences, des images et plus récemment des réseaux. Par exemple, les biologistes sont souvent intéressés à classer les séquences biologiques (séquences ADN, les séquences de protéines), des réseaux (interactions, co-expression entre gènes), des images (imagerie cellulaire, classification des tissus) ou des données structurées (structure de gènes, information sur les patients). L'espace d'observation des données peut être donc $\mathbb{R}^p$ si des données quantitatives sont considérées, $L^2([0,1])$ si des données fonctionnelles sont considérées (séries temporelles par exemple) ou $\{A, C, G, T\}^p$ si les données sont qualitatives avec quatre modalités possibles sur les p variables (séquences d'ADN par exemple). Par ailleurs, les données à classer peuvent être un mélange de différents types de données : données quantitatives et qualitatives ou données qualitatives et réseaux.

Les méthodes de classification peuvent être divisées en deux familles principales : méthodes génératives et méthodes discriminatives. Les méthodes génératives, qui modélisent chaque classe à l'aide d'une distribution de probabilité, sont généralement très appréciées pour leurs bons résultats et leur facilité d'interprétation. En revanche, elles ne peuvent pas traiter tous les types de données. A l'inverse, les méthodes discriminatives et en particulier les méthodes à noyaux ont la capacité de classer des données de types très variés. Pour cela, les méthodes à noyaux projettent les observations dans un espace de dimension potentiellement infinie (le *feature space*) grâce à une fonction linéaire ou non (le *feature map*). Ces méthodes sont en général très performantes, cependant l'interprétation des résultats est difficile. De plus, ces méthodes souffrent d'une compléxité calculatoire importante.

2 La famille de processus gaussiens parcimonieux

Nous proposons donc de combiner l'approche de projection dans un *feature space* des méthodes à noyaux avec la modélisation et la règle de décision des méthodes génératives. Nous supposons que les données d'apprentissage $\{(x_1, z_1), ..., (x_n, z_n)\} \in E \times \{1, ..., k\}$ sont des réalisations indépendantes d'un couple de vecteur aléatoire (X, Z) où z_j indique la classe de l'observation x_j. Ces données sont projetées dans un espace F à l'aide d'une fonction φ et le vecteur aléatoire $Y = \varphi(X)$ est de plus supposé être, conditionnellement à $Z = i$, un processus gaussien de moyenne μ_i et d'opérateur de covariance Σ_i, $i = 1, ..., k$. La règle du maximum a posteriori (MAP) peut être ensuite utilisée pour classer une nouvelle observation $\varphi(x)$ dont la classe est inconnue. Malheureusement, si la fonction φ projette les données dans un espace de dimension infinie, il ne sera pas possible de construire avec une erreur raisonnable la règle de classification à partir d'un échantillon de taille finie.

Pour pallier ce problème, nous proposons de contraindre la décomposition spectrale de l'opérateur de covariance de chaque classe C_i de sorte que celle-ci ait au plus $d_i + 1$ valeurs

propres différentes, $i = 1, ..., k$. Les plus petites valeurs propres pour chaque classe sont donc supposées égales et de plus communes entre les classes. Cette modélisation peut également être contrainte pour donner naissance à 7 autre processus gaussiens parcimonieux. Si $d_i < n_i$ pour chaque classe où n_i est le nombre d'observations de la $i$ème classe, la règle du MAP peut être construite efficacement à partir d'un échantillon de taille finie. De plus, en exploitant le "*kernel trick*", on montre qu'il est possible de calculer la règle de classification du MAP au travers d'une fonction noyau sans avoir une connaissance explicite de la fonction φ.

Il est également possible d'étendre cette modélisation au contexte de la classification non supervisée (clustering). Dans ce cas, l'algorithme EM devra être utilisé car l'appartenance des données aux classes est inconnue (Z est donc considérée comme une variable manquante).

3 Application à des données non-quantitatives

Les méthodes supervisées et non supervisées basées sur la famille de processus gaussiens parcimonieux peuvent donc être appliquées à tous les types de données pour lesquels il est possible de définir une fonction noyau. En particulier, nous avons appliqué nos méthodes à des données fonctionnelles (pour lesquelles la fonction φ est connue), des données qualitatives (grâce à un noyau basé sur la distance de Hamming), des données de type réseau (avec le noyau du laplacien du graphe) et des données hétérogènes (quantitatives et qualitatives). La classification des données hétérogènes a été obtenue en combinant un noyau RBF calculé sur les données quantitatives et un noyau basé sur la distance de Hamming pour les données qualitatives.