Apprentissage de données fonctionnelles par modèles multi-tâches : application à la prédiction de performances sportives

Arthur Leroy

Résumé:

Ce manuscrit de thèse est consacré à l'analyse de données fonctionnelles et la définition de modèles multi-tâches pour la régression et la classification non supervisée. L'objectif de ce travail est double et trouve sa motivation dans la problématique d'identification de jeunes sportifs prometteurs pour le sport de haut niveau. Ce contexte, qui offre un fil rouge illustratif des méthodes et algorithmes développés par la suite, soulève la question de l'étude de multiples séries temporelles supposées partager de l'information commune, et généralement observées à pas de temps irréguliers. La méthode centrale développée durant cette thèse, ainsi que l'algorithme d'apprentissage qui lui est associé, se concentrent sur les aspects de régression fonctionnelle à l'aide d'un modèle de processus Gaussiens (GPs) multi-tâche. Ce cadre probabiliste non-paramétrique permet de définir une loi a priori sur des fonctions, supposées avoir généré les données de plusieurs individus. Le partage d'informations communes entre les différents individus, au travers d'un processus moyen, offre une modélisation plus complète que celle d'un simple GP, ainsi qu'une pleine prise en compte de l'incertitude. Un prolongement de ce modèle est par la suite proposé via la définition d'un mélange de GPs multi-tâche. Cette approche permet d'étendre l'hypothèse d'un unique processus moyen sous-jacent à plusieurs, chacun associé à un groupe d'individus. Ces deux méthodes, nommées respectivement Magma et MagmaClust, offrent de nouvelles perspectives de modélisation ainsi que des performances remarquables vis-à-vis de l'état de l'art, tant sur les aspects de prédiction que de clustering. D'un point de vue applicatif, l'analyse se concentre sur l'étude des courbes de performances de jeunes nageurs, et une première exploration des données réelles met en évidence l'existence de différents patterns de progression au cours de la carrière. Par la suite, l'utilisation de l'algorithme Magma, entrainé sur la base de données, attribue à chaque sportif une prédiction probabiliste de ses performances futures, offrant ainsi un précieux outil d'aide à la détection. Enfin, l'extension via l'algorithme MagmaClust permet de constituer automatiquement des groupes de nageurs de part les ressemblances de leurs patterns de progression, affinant de ce fait encore les prédictions. Les méthodes détaillées dans ce manuscrit ont également été entièrement implémentées et sont partagées librement.

Mots-Clefs : Processus Gaussiens, apprentissage multi-tâche, données fonctionnelles, clustering de courbes, algorithmes EM, méthodes variationnelles