MLFP: Un algorithme d'apprentissage de métrique pour la classification de données déséquilibrées

Rémi Viola[1,2], Rémi Emonet [1], Amaury Habrard[1], Guillaume Metzler[1], and Marc Sebban[1]

[1]Laboratoire Hubert Curien UMR 5516, Université de Lyon, Université Jean-Monnet, F-42023, Saint-Etienne, France
[2]Direction Générale des Finances Publiques, Ministère de l'Économie et des Finances, France

Résumé

Apprendre à partir de données déséquilibrées reste une tâche complexe en apprentissage, tant sur le plan théorique que pratique. Dans cet article, nous abordons cette problématique en utilisant une stratégie basée sur l'apprentissage de métrique. Contrairement aux méthodes se basant sur la même approche, notre algorithme apprend une nouvelle métrique qui ne sera utilisée que lorsqu'une (nouvelle) donnée est comparée à un exemple d'apprentissage de la classe minoritaire (ou classe positive). D'un point de vue géométrique, cela revient à rapprocher artificiellement des exemples positifs de cette (nouvelle) donnée sans modifier les caractéristiques de la classe majoritaire. La stratégie mise en œuvre permet d'étendre les frontières de décisions autour des données positives. En terme de performance, cela se traduit par une meilleure F-mesure, critère de performance très souvent employé dans ce contexte, par rapport aux algorithmes de l'état de l'art. Au-delà de cette contribution algorithmique, notre article présente une étude théorique basée sur la *stabilité uniforme*. Cette étude nous donne des garanties de généralisation sur les taux de faux positifs et de faux négatifs. Les expériences, effectuées sur plusieurs ensembles de données déséquilibrées, montrent l'efficacité de notre méthode par rapport aux algorithmes d'apprentissage de métrique existants. Avec la méthode proposée, nous sommes en mesures de rivaliser voire d'obtenir de meilleures performances qu'avec des algorithmes spécifiquement dédiés au traitement de données déséquilibrées.

Mots-clef : Imbalanced Binary Classification, Metric Learning, Nearest-Neighbors, Theoretical Guarantees