Optimisation directe des poids de modèles dans un prédicteur Bayésien naïf moyenné

Romain Guigourès*, Marc Boullé*

*Orange Labs
2 avenue Pierre Marzin
22307 Lannion Cedex

{romain.guigoures, marc.boulle}@orange-ftgroup.com

Résumé. Le classifieur Bayésien naïf est un outil de classification efficace en pratique pour de nombreux problèmes réels, en dépit de l'hypothèse restrictive d'indépendance des variables conditionnellement à la classe. Récemment, de nouvelles méthodes permettant d'améliorer la performance de ce classifieur ont vu le jour, sur la base à la fois de sélection de variables et de moyennage de modèles. Dans cet article, nous proposons une extension de la sélection de variables pour le classifieur Bayésien naïf, en considérant un modèle de pondération des variables utilisées et des algorithmes d'optimisation directe de ces poids. Les expérimentations confirment la pertinence de notre approche, en permettant une diminution significative du nombre de variables utilisées, sans perte de performance prédictive.

1 Introduction

Le classifieur naïf Bayésien est un outil largement utilisé dans les problèmes de classification supervisée. Il a pour avantage de se montrer efficace pour de nombreux jeux de données réels (Hand et Yu, 2001). Cependant, l'hypothèse naïve d'indépendance des variables peut, dans certains cas, dégrader les performances du classifieur.

Ainsi, des méthodes proposant de réaliser de la sélection de variables ont vu le jour (Langley et Sage, 1994). Elles consistent en la mise en place d'heuristiques d'ajout et de suppression de variables afin de sélectionner le meilleur sous-ensemble de variables maximisant un critère et donc les performances du classifieur, selon une approche wrapper (Guyon et Elisseeff, 2003). Il a été montré par Boullé (2007) que moyenner un grand nombre de classifieurs Bayésiens naïfs sélectifs, réalisés avec différents sous-ensembles de variables, revenait à ne considérer qu'un seul modèle avec une pondération sur les variables.

Dans cet article, on se propose de trouver un modèle optimal par optimisation directe des poids des variables. On introduit un critère, basé sur la vraisemblance d'un modèle, fonction continue d'un vecteur de poids. Une descente de gradient est ensuite utilisée pour l'optimisation, le critère étant continu et dérivable sur l'ensemble de définition du vecteur de poids des variables. Une méthode de régularisation est introduite pour minimiser le nombre de variables