La régression Partial Least-Squares boostée

Jean-François DURAND[1]

Université Montpelier II, France
E-Mail : jf.durand@club-internet.fr
Site web : www.jf-durand-pls.com

Résumé Ce papier présente la régression Partial Least-Squares (PLS) comme appartenant à la famille de méthodes de boosting à fonction coût L_2. D'une part, la régression PLS linéaire classique appartient à cette catégorie en considérant une variable latente ou composante principale comme base d'apprentissage (*base learner*) rendant robuste le modèle face au problème de la rareté des données et de la multi-corrélation des variables. D'autre part, l'usage des B-splines et de leurs produits tensoriels dans la construction de la base d'apprentissage, exploite de façon naturelle le potentiel du boosting L_2 de PLS pour produire des modèles non-linéaires additifs qui capturent les effets principaux ainsi que les interactions significatives. La performance du boosting PLS en régression comme en classification supervisée est montrée sur trois exemples.
Keywords :Boosting, Partial Least-Squares, B-splines, Produits Tensoriels
This paper presents the Partial Least-Squares regression (PLS) in the framework of the boosting methods with L_2 loss. First, the ordinary PLS regression already belongs to that family by considering the latent variables or principal components as base learners producing robust linear models that overcome the problems of the scarcity of the observations as well as the multi-collinearity of the predictor variables. Most of all, the use B-splines and their tensor products to construct the base learner, typically provides PLS with L_2-boosts leading to non-linear additive models that capture main effects as well as relevant interactions. The performances of the different PLS boosts in both regression and classification are shown on three exemples.
Keywords : Boosting, Partial Least-Squares, B-splines, Tensor Products

1 Introduction

La régression linéaire Partial Least-Squares (Wold et al., 1983), (Tenenhaus, 1998), en bref PLS, est une méthode statistique de prévision très populaire en chimiométrie à sa création et maintenant dans tous les domaines industriels et économiques. Les débuts de PLS dans le monde académique ont été plus laborieux peut-être à cause de sa formulation algorithmique. Depuis quelques années cependant, la recherche statistique a pris à son compte l'introduction d'algorithmes dans la définition des méthodes de prévision. Un exemple est donné par le boosting L_2 dont la version séminale est le "twicing" (Tukey, 1977). L'idée consiste à améliorer une méthode peu performante comme celle des moindres carrés, en l'appliquant de façon récurrente sur les données mal prédites que sont les résidus, pour construire, étape par étape, un modèle additif. Les méthodes modernes de boosting transforment à chaque étape, les variables explicatives par une fonction de classe paramétrique, appelée la base d'apprentissage (*base learner*). Cela peut être un arbre de décision (Hastie et al., 2001), une spline de lissage (Bühlmann et Bin Yu, 2000)...

Le premier objectif de cet article est de montrer comment une variable latente PLS, combi-