Prédictions contrôlées en apprentissage automatique

Alexander Gammerman & Vladimir Vovk

Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
alex@cs.rhul.ac.uk, vovk@cs.rhul.ac.uk

Résumé: Les récentes avancées obtenues en apprentissage automatique rendent possible la conception d'algorithmes efficients de prédiction pour des ensembles de données à grand nombre de paramètres. Cet article décrit une nouvelle méthode pour contrôler les prédictions élaborées par de nombreux algorithmes, incluant les machines à vecteurs support, la régression ridge à noyau, les plus proches voisins par noyau et bien d'autres méthodes correspondant à l'actuel état de l'art. Les prédictions contrôlées pour les étiquettes de nouveaux objets comportent des mesures quantitatives de leur précision et de leur fiabilité. Nous prouvons que ces mesures sont valides sous hypothèse de randomisation, traditionnelle en apprentissage automatique : les objets et leurs étiquettes sont supposés indépendamment générés par la même distribution de probabilité. En particulier, il devient possible de contrôler (aux fluctuations statistiques près) le nombre de prédictions erronées en choisissant un niveau de confiance approprié. La validité étant assurée, l'objectif restant pour les prédictions contrôlées est l'efficience : prendre au mieux les caractéristiques des nouveaux objets ainsi que l'information disponible pour produire des prédictions aussi précises que possible. Ceci peut être obtenu avec succès en utilisant toute la puissance des méthodes modernes d'apprentissage automatique.

Mots-clés: prédicteurs conformes, apprentissage en ligne, étrangeté d'une prédiction, induction, transduction

Abstract: Recent advances in machine learning make it possible to design efficient prediction algorithms for data sets with huge numbers of parameters. This article describes a new technique for `hedging' the predictions output by many such algorithms, including support vector machines, kernel ridge regression, kernel nearest neighbours, and by many other state-of-the-art methods. The hedged predictions for the labels of new objects include quantitative measures of their own accuracy and reliability. These measures are provably valid under the assumption of randomness, traditional in machine learning: the objects and their labels are assumed to be generated independently from the same probability distribution. In particular, it becomes possible to control (up to statistical fluctuations) the number of erroneous predictions by selecting a suitable confidence level. Validity being achieved automatically, the remaining goal of hedged prediction is efficiency: taking full account of the new objects' features and other available information to produce as accurate predictions as possible. This can be done successfully using the powerful machinery of modern machine learning.

Keywords: conformal predictors, on line procedure, strangeness, induction, transduction

1 Introduction

Les deux principaux aspects du problème de prédiction, la classification supervisée et la régression, sont des thèmes usuels de la statistique et de l'apprentissage automatique. Les techniques classiques de classification supervisée et de régression sont capables de traiter des ensembles de données conventionnels de faible dimension et petite taille ; cependant, les tentatives faites pour appliquer ces techniques à des ensembles actuels de données de grande dimension se heurtent à de sérieuses difficultés conceptuelles et calculatoires. Plusieurs nouvelles méthodes, et tout d'abord les machines à vecteur support [42,43] et autres méthodes à noyau, ont été récemment développées en apprentissage automatique, avec l'objectif explicite de traiter des ensembles de données de grande taille et de grande dimension.
Un inconvénient caractéristique de ces nouvelles méthodes est l'absence de mesure utilisable de confiance pour les prédictions. Par exemple, certaines limites supérieures strictes obtenues par la théorie d'apprentissage PAC (probablement approximativement correcte) pour la probabilité d'erreur dépasse souvent 1, même pour des ensembles de données relativement simples ([51], p. 249).