Journal de la Société Française de Statistique
Vol. 152 No. 3 (2011)

Nouveaux défis en apprentissage statistique

Title: New and challenging problems in statistical learning

Charles Bouveyron[1], Florence Forbes[2] et Stéphane Girard[2]

L'apprentissage statistique joue de nos jours un rôle croissant dans de nombreux domaines scientifiques aussi variés que l'imagerie, la biologie, l'astronomie, l'économie ou la sociologie. Les progrès scientifiques réalisés ces dernières années ont permis d'augmenter sensiblement les capacités de mesure et de calcul, et il est à présent difficile pour un opérateur humain de traiter de façon exhaustive ces données dans un temps raisonnable. L'apprentissage statistique se propose alors de prendre le relais sur l'humain en analysant de façon automatique ces données dans le but d'aider les opérateurs à la prise de décision. De plus, de nombreux domaines d'application génèrent de nouveaux problèmes théoriques en statistique. Par exemple, la classification de données de très grande dimension ou la classification de données corrélées sont des problèmes particulièrement présents en analyse d'images et biologie et pour lesquels les outils théoriques sont encore à développer.

Il est par conséquent important de proposer des méthodes statistiques adaptées aux problèmes modernes posés par les différents champs d'application. Outre l'importance de la performance des méthodes proposées, elles devront également apporter une meilleure compréhension des phénomènes observés. Afin de conforter les contacts entre les différentes communautés et de favoriser l'émergence de nouvelles idées, une série de colloques d'audience internationale sur le thème « Challenging problems in statistical learning » a été organisée sous le nom de Statlearn depuis 2009 : Paris en 2009 et 2010 puis Grenoble en 2011. Nous rapellons à cette occasion que la prochaine édition du colloque Statlearn aura lieu à Lille les 5 & 6 avril 2012.

L'objectif de ce numéro spécial est de donner un aperçu des présentations faites lors des colloques Statlearn. Le lecteur pourra en particulier remarquer combien les sujets des articles sont variés, preuve de la diversité des problèmes d'apprentissage statistique.

L'article de Flora Jay, Michaël Blum, Eric Frichot et Olivier François présente des familles de modèles hiérarchiques bayésiens dédiés à l'analyse de la structure génétique de populations. Ils appliquent avec succès leurs modèles à des données génétiques de populations humaines et végétales. Christophe Biernacki et Alexandre Lourme abordent quant à eux un problème nouveau dans le domaine de la classification non supervisée : la classification simultanée de populations différentes. La méthode proposée, basée sur des modèles de mélange, permet de classer simultanément plusieurs échantillons provenant de populations différentes. La méthode est notamment

1. Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne
E-mail : charles.bouveyron@univ-paris1.fr
2. Equipe Mistis, INRIA Grenoble Rhône-Alpes & LJK
E-mail : florence.forbes@inria.fr and E-mail : stephane.girard@inria.fr

 ISSN: 2102-6238

illustrée sur des oiseaux d'une même espèce mais d'origines différentes. La contribution de Nathalie Villa-Vialanex et Fabrice Rossi traite de la classification et de la visualisation de grands réseaux. L'analyse des données de type réseau (également appelées données relationnelles) est actuellement un sujet très actif en apprentissage statistique du fait de l'intérêt croissant pour les réseaux sociaux par exemple. Leur travail se base sur une classification hiérarchique des sommets du réseau considéré et propose une méthode innovante permettant de visualiser le réseau de manière simplifiée à différents niveaux de détails. Julie Carreau et Stéphane Girard s'intéressent dans leur article à l'estimation des quantiles extrêmes spatiaux par une méthode de vraisemblance pondérée. La pondération est établie à partir d'une notion de distance entre le point d'intérêt pour l'estimation et les observations dans un espace latent. Le travail d'Alain Celisse et Tristan Mary-Huard présente une étude théorique de l'estimation du taux d'erreur par validation croisée de l'algorithme de classification k-NN (k plus proches voisins). Cette étude propose en particulier une expression explicite de ce taux d'erreur. Le problème est étudié à la fois du point de vue de l'apprentissage passif et de celui de l'apprentissage actif. Charles Bouveyron et Camille Brunet s'intéressent également à un problème de classification mais non-supervisée et en grande dimension. Ils proposent deux nouvelles techniques d'estimation de l'espace latent pour l'algorithme Fisher-EM qui modélise et classe les données de grande dimension dans un espace latent de faible dimension. Enfin, l'article de Florence Forbes, Benoît Scherrer et Michel Dojat traite de la classification non-supervisée coopérative. Le but d'une telle classification est de produire une double partition des données. La méthode proposée se base sur un modèle à variables cachées et sur des champs de Markov. Une application concrète au clustering coopératif de tissus et structures d'images IRM est proposée pour illustrer la méthode.

Nous espérons que ce numéro spécial permettra aux lecteurs du Journal de la SFdS de mesurer la diversité des recherches menées dans le domaine de l'apprentissage statistique. Sa lecture donnera peut-être à certains l'envie de contribuer à ce domaine passionnant et très actif de la Statistique. Pour conclure, nous tenons à remercier l'Université Paris 1 Panthéon-Sorbonne, l'INRIA, le LJK ainsi que le groupe "Statistique et Image" de la Société Française de Statistique qui ont soutenu financièrement les colloques Statlearn. Nos remerciements vont également à Philippe Besse et Gilles Celeux qui sont à l'initiative de ce numéro spécial. Nous remercions enfin les auteurs pour leurs contributions ainsi que les relecteurs anonymes pour la qualité de leur travail.

 ISSN: 2102-6238