HAL open science

Classification

Charles Bouveyron, Francois Caron, Marie Chavent

To cite this version:

Charles Bouveyron, Francois Caron, Marie Chavent. Classification. Journées MAS et Journée en l'honneur de Jacques Neveu, Aug 2010, Talence, France. inria-00496744

HAL Id: inria-00496744

https://inria.hal.science/inria-00496744v1

Submitted on 1 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Classification

Session organisée par Charles Bouveyron et Francois Caron et Marie Chavent

La classification a pour objet de regrouper des données en classes possédant des caractéristiques similaires. La classification peut être supervisée lorsque l'on dispose d'un ensemble d'apprentissage labellisé, semi-supervisée ou non supervisée. Elle apparait dans de nombreuses applications telles que la fouille de texte, la reconnaissance vocale ou l'analyse de données génomiques. L'objectif de cette session est d'offrir un panorama des approches statistiques pour la classification de données (modèles de mélange, SVM, processus de Dirichlet, etc.) et d'en présenter diverses applications.

Adresses des organisateurs :

Charles Bouveyron
Laboratoire SAMM
Université Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris France
E-mail : `charles.bouveyron@univ-paris1.fr`
<`http://samm.univ-paris1.fr/~charles-bouveyron`>

Francois Caron
Equipe ALEA INRIA Bordeaux Sud-Ouest, IMB
Université Bordeaux1, 351 Cours de la libération, 33400 Talence France
E-mail : `Francois.Caron@inria.fr`
<`http://www.math.u-bordeaux1.fr/~fcaron`>

Marie Chavent
Equipe CQFD INRIA Bordeaux Sud-Ouest, IMB
Université Bordeaux2, 3 ter place de la Vctoire, 33000 Bordeaux France
E-mail : `Marie.Chavent@u-bordeaux2.fr`
<`http://www.math.u-bordeaux1.fr/~chavent`>

Session : Classification

Classification générative des données de grande dimension : état de l'art et avancées récentes

par **Charles Bouveyron**

La classification générative a du faire face ces dernières années à l'augmentation de la dimension des données et au fléau de la dimension qui lui est associée. Aprés une brève introduction à la classification générative, l'exposé passera tout d'abord en revue les méthodes récentes de classification dédiées aux données de grande dimension. Quelques avancées récentes seront ensuite présentées, concernant notamment la sélection de dimensions intrinséques et le clustering dans un sous-espace discriminant.

Adresse :
Charles Bouveyron
SAMOS-MATISSE
Université Paris1, 90 rue de Tolbiac, 75013 Paris France
E-mail : `charles.bouveyron@univ-paris1.fr`
<`http://samos.univ-paris1.fr/~charles-bouveyron`>

Session : Classification

Modèèles hybrides génératifs-discriminatifs : théorie et applications

par **Guillaume Bouchard**

Les paradigmes d'apprentissage génératif et discriminatif pour résoudre les problèmes de prédiction en l'apprentissage automatique sont souvent mis en opposition, l'un permettant de bien modéliser la structure des données mais dont la prédiction est fortement biaisée, l'autre permettant de créer une règle de décision asymptotiquement optimale, mais souvent difficile à interpréter. Ils ont souvent été étudiés dans différentes sous-communautés, mais au cours des dix dernières années, il y a un intérêt croissant pour essayer de comprendre et tirer parti des avantages des deux approches. Nous présenterons notre compréhension actuelle des approches génératives et discriminatives ainsi que leur combinaison à travers des résultats théoriques et empiriques. En particulier, nous verrons que les méthodes hybrides génératives-discriminatives permettent de résoudre des taches de classification supervisée pour lesquelles une représentation vectorielle des données est difficile à construire, comme les problèmes de détection de panne ou de reconnaissance de paraphrase/intrication textuelle.

Adresse :
Guillaume Bouchard
Xerox Research
E-mail : `guillaume.bouchard@xerox.com`
<`http://www.xrce.xerox.com/`>

Session : Classification

Progress and open challenges in extremely high-dimensional medical outcome prediction

par **Kevin Bleakley**

Using biological data for medical decisions requires "extremely high" prediction accuracy ; mistakes can lead to death. Very few current statistical methods are good enough to be used in life-threatening clinical decisions, e.g. choice of low vs high chemotherapy dose for breast cancer patients. Difficulties include (1) the above moral reason, (2) high-dimensionality of data (p >> n) and (3) the possibility that data does not contain enough information to construct a near-perfect classification rule. I will review the current state-of-the-art in high-dimensional biological decision-making, showing what statistical methods are being used, their success (or lack of), and suggest possible future research directions. In particular, I will describe Next Generation Sequencing approaches, their faster-than-exponential drop in cost, and implications for the next five years at the statistics/biology interface.

Adresse :

Kevin Bleakley

PostDoc, Mines ParisTech/Institut Curie/INSERM U900

Institut Curie - Centre de recherche Biologie du developpement - U900

26 rue d'Ulm, 75248 Paris cedex 05, France

E-mail : `kevbleakley@gmail.com`

<`http://cbio.ensmp.fr/~kbleakley`>

Session : Classification

Nonparametric bayesian modelling of co-exposures to various pesticides to determine cocktails

par **Amélie Crepet**

This work introduces a specific application of Bayesian nonparametric models in food risk analysis framework. The goal is to determine mixture of pesticides residues which are simultaneously present in the diet, to give directions for future toxicological experiments for studying possible combined effects of those mixtures. Namely, the distribution of the exposures to a large number P of pesticides is assessed from the available consumption data and contamination analyses. We propose to model the co-exposures to the P pesticides by a Dirichlet process mixture based on a multivariate Gaussian kernel so as to determine clusters of individuals with similar co-exposure patterns. The posterior distributions and the optimal partition are computed through a Gibbs sampler based on stick-breaking priors. To reduce computational time due to the high dimensional data, a random block sampling is used. Other nonparametric Bayesian models such as models based on Indian Buffet process will be developed to propose a simultaneously classification of the individuals and the pesticides in groups.

Adresse :

Amélie Crepet
AFFSA-DERNS-AQR-PC
27-31 Av. Général Leclerc, 94701 Maisons-Alfort, France
E-mail : a.crepet@affsa.fr