TELECOM ParisTech

Institut Mines-Télécom

CHAIRE
MACHINE LEARNING
FOR BIG DATA

Paris, le 27/11/2017

Offre de stage

Sujet : Détection d'anomalies et observations fonctionnelles

Possibilité de poursuivre sur une thèse

Encadrement

Stephan Clémençon , Anne Sabourin, Florence D'Alché

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : printemps 2018

Équipe(s) d'accueil de la thèse

département IDS, équipe S2A (Signa, Statistique et Apprentissage)

Mots clés

Données fonctionnelles, détection d'anomalies, espaces fonctionnels

Sujet détaillé

Whereas the vast majority of anomaly detection/scoring techniques applied to temporal observations documented in the statistical literature strongly rely on probabilistic modelling and/or rigid assumptions on the type of anomalies to be detected (\textit{e.g.} change points corresponding to a shifted distribution), the information collected by complex infrastructures such as aircrafts or energy networks, equipped with more and more sensors in the era of IoT and Big Data, offer a wealth of information that cannot be processed by means of these traditional techniques. Gathered with the purpose of monitoring such sophisticated systems, these observations generally take the form of high dimensional functional data, sampled at various frequencies and possibly incomplete, and slip from statistical modelling due to their complexity. This internship is a part of a research project aiming at developing machine-learning methods tailored to this setting, as well as a (theoretical/experimental) validity framework for the latter. It shall build on recent preliminary works carried out by the academic team in the field of anomaly scoring/detection, combining machine-learning approaches with extreme value theory and on the strong expertise of the industrial partner in the data available and their preprocessing. Beyond statistical performance and computational feasability, special attention will be paid to the problem of producing interpretable prediction rules, a crucial issue

regarding practical application. In particular, the design of visualization techniques dedicated to represent in a summary fashion the degree of similarity of detected anomalies will be considered throughout the project.

La Chaire Machine Learning for Big Data

Le traitement statistique des masses de données convoque à la fois mathématiques appliquées et informatique, à travers une discipline en plein essor : le Machine Learning ou apprentissage statistique.

Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée à hauteur de près de 2 M€ par quatre entreprises partenaires : Criteo, PSA Peugeot Citroën, Safran et BNP Paribas, la Chaire Machine Learning for Big Data est portée par le mathématicien Stéphan Clémençon, Enseignant-Chercheur, Professeur au sein du Département du Traitement du Signal et des Images à Télécom ParisTech.

Proposant cinq axes de recherche méthodologiques, enrichis par des applications industrielles concrètes, cette Chaire a pour objectif d'animer, en interaction avec ses partenaires, une activité de recherche de pointe en Machine Learning, ainsi que de proposer des programmes de formation.

La variété des données aujourd'hui disponibles (nombres, images, textes, signaux), leur grande dimension et leur volumétrie rendent souvent inopérantes les méthodes statistiques traditionnelles reposant sur le prétraitement humain et un long travail de modélisation. Le Machine Learning vise donc à élaborer et étudier des algorithmes, à vocation prédictive le plus souvent, permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon performante.

Les avancées technologiques, l'omniprésence des capteurs (systèmes embarqués, objets connectés, Internet…) et l'explosion des réseaux sociaux s'accompagnent d'un véritable déluge de données, propulsant les sciences de l'information au centre du processus de valorisation des masses de donnés. En plus de la collecte et du stockage, l'enjeu est de pouvoir analyser ces données afin d'optimiser les décisions et mettre au point de nouvelles applications.

Au-delà du buzz médiatique dont il fait l'objet, le Big Data est donc un sujet stratégique majeur, au cœur d'enjeux économiques et sociétaux considérables. Son impact est désormais perçu dans presque tous les secteurs de l'activité humaine : de la recherche scientifique à la médecine en passant, entre autres, par la finance, le bâtiment, l'e-commerce, la défense ou les transports.

En savoir plus sur la Chaire, ses axes de recherche, ses activités, ses publications :

http://machinelearningforbigdata.telecom-paristech.fr

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique / reconnaissance des formes

- Traitement de la parole, traitement du langage naturel

- Bon niveau en programmation (Java, C/C++, Python)

- Bon niveau d'anglais

Candidatures

à envoyer à stephan.clemencon@telecom-paristech.fr:

- Curriculum Vitae

- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)

- Relevés de notes des années précédentes

- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[1] Denis Bosq. Linear processes in function spaces: theory and applications , volume 149. Springer Science & Business Media, 2012.
[2] Ferraty Frederic and Vieu Philippe. Nonparametric Functional Data Analysis . Springer Series in Statistics, 2006.
[3] R. Brault, M. Heinonen, F. D'Alche-Buc. Random Fourier Features For Operator-Valued Kernels . ACML, 2016.
[4] Ramsay James and Silverman B. W. Functional Data Analysis . Springer Series in Statistics, 2005.
[5] Albert Thomas, Stephan Clemencon, Vincent Feuillard, and Alexandre Gramfort. Learning hyperparameters for unsupervised anomaly detection. In Anomaly Detection Workshop , 2016.
[6] Albert Thomas, Stephan Clemencon, Alexandre Gramfort, and Anne Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In Aarti Singh and Jerry Zhu, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics , volume 54 of Proceedings
of Machine Learning Research , pages 1011{1019, Fort Lauderdale, FL, USA, 20{22 Apr 2017. PMLR.
[7] S. Clemencon, A. Thomas. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere
Submitted, available at https://hal.archives-ouvertes.fr/hal-01519728
[8] W. Polonik. Minimum volume sets and generalized quantile processes. Stochastic Processes and their
Applications , 69(1):1{24, 1997.
[9] N. Goix, A. Sabourin, and S. Clemencon. Learning the dependence structure of rare events: a nonasymp-
totic study. In Proceedings of the International Conference on Learning Theory, COLT'15 , 2015.
[10] N. Goix, A. Sabourin, and S. Clemencon. Sparse representation of multivariate extremes with applica-
tions to anomaly ranking. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS'16 , 2016.
[11] N. Goix, A. Sabourin, and S. Clemencon. Sparsity in multivariate extremes with applications to anomaly
detection. http://arxiv.org/abs/1507.05899 , to appear in JMVA 2017.
[12] R. Vert and J.P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. J.
Machine Learning Research , 17:817{854, 2006.
[13] C. Scott and R. Nowak. Learning Minimum Volume Sets. Journal of Machine Learning Research , 7:665-704, 2006.