SAFRAN AEROSPACE · DEFENCE · SECURITY
criteo
PSA PEUGEOT CITROËN
BNP PARIBAS La banque d'un monde qui change

TELECOM ParisTech
CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 26/11/2016

Offre de stage

Sujet : Detection of emerging events in spatial time series

Poursuite en thèse possible

Encadrement

Florence d'Alché-Buc, florence.dalche@telecom-paristech.fr

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : à partir de février 2017

Équipe(s) d'accueil de la thèse

Equipe Statistique, Signal & Apprentissage du département TSI, département SES.

Mots clés

Time series, social network, spatial network, event/outbreak forecasting, structured time series, operator-valued kernels, vector autoregressive models, heterogeneous data.

Sujet détaillé

In many fields such as health monitoring, epidemiology, transport or energy, massive and complex real-world datasets are accumulated through time and space, giving rise to spatial time series. A recurrent critical task consists in forecasting outbreaks or rare events from spatial time series. Although many works have been developed in this area, this problem still raises many issues such as dealing with incomplete measures, data acquired at different time scale, heterogeneity of time series (sensors measurements, sells, text message, search engine queries, …) and inherent group-structure of the data. Eventually the very large scale of data requires to adapt forecasting tools to be efficient in inference and learning.

The main goal of internship is to build upon multi-task autoregressive models and graph-vased approaches to provide a general framework for outbreak/event detection in spatial time series. Approaches based on operator-valued kernel autoregressive models will be especially studied and extended while paying attention to the general context of Markov Random Fields. Implementation will be based on a home-made open-source library Operalib.

Preferably the intern will have a strong background in advanced machine learning and ideally in time-seires modeling, although the later is not mandatory. A knowledge of graph -based approaches in Machine Learning is also welcome.

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 données structurées
- Séries temporelles, traitement du langage naturel

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

Candidatures

à envoyer à florence.dalche@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

Aramaki E, Maskawa S, Morita M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP'11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 1568–1576. Available from: http://dl.acm.org/citation.cfm?id=2145432.2145600.

Romain Brault, Markus Heinonen, Florence d'Alché-Buc, Random Fourier Features for Operator-valued kernels, ACML 2016.

Romain Brault, Néhémy Lim, Florence d'Alché-Buc, Scaling up Vector Autoregressive models with operator-valued Random Fourier features, AALTD'16, joint workshop to ECML/PKDD 2016, to appear in LNCS.

Céline Brouard, Marie Szafranski, Florence d'Alché-Buc, Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels

17(176):1−48, (2016).

Céline Brouard, Huibin Shen, Kai Dührkop, Florence d'Alché-Buc, Sebastian Böcker, Juho Rousu: Fast metabolite identification with Input Output Kernel Regression. Bioinformatics 32(12): 28-36 (2016).

Néhémy Lim, Florence d'Alché-Buc, Cédric Auliac, George Michailidis:Operator-valued kernel-based vector autoregressive models for network inference. Machine Learning 99(3): 489-513 (2015)

Maxime Sangnier, Olivier Fercoq, Florence d'Alché-Buc, Joint Quantile Regression in vector-valued RKHSs, NIPS 2016.

Néhémy Lim, Florence d'Alché-Buc, Cédric Auliac, George Michailidis:Operator-valued kernel-based-vector autoregressive models for network inference. Machine Learning 99(3): 489-513 (2015)

Daniel Oliveira, Daniel B. Neill, James H. Garrett Jr., and Lucio Soibelman. Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network. *Journal of Computing in Civil Engineering* 25(1): 21-30, 2011.

Daniel B. Neill and Gregory F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning,*79, (2010).

Thapen N, Simmie D, Hankin C, Gillard J. DEFENDER: Detecting and Forecasting Epidemics Using Novel Data-Analytics for Enhanced Response. PLoS ONE 11(5): e0155417.