CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 24/11/16

Offre de stage

Sujet : Supervised Reconstruction of Large Graphs

Possibilité de poursuivre sur une thèse

Encadrement

Stephan Clémençon (Telecom ParisTech), Guillaume Papa (Telecom ParisTech), Aurélien Bellet (INRIA Magnet)

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : mars/avril 2017

Équipe(s) d'accueil de la thèse

Département TSI, équipe Statistiques et Applications (STA)

Mots clés

Random graphs, link prediction, graph reconstruction, large-scale networks

Sujet détaillé

The problem of predicting connections between a set of data points described by features X finds many applications. This statistical learning problem is motivated by a variety of applications such as systems biology and social network analysis. It has recently been the subject of a good deal of attention in the machine learning literature, and is also known as supervised link prediction. The learning task can be formulated as the minimization of a reconstruction risk, whose natural empirical version is the average prediction error over the n (n - 1)/2 pairs of nodes in a training graph of size n. Under restrictive structural assumptions, stipulating that the marginal graph (ignoring the features X attached to each node/vertex) is of Bernoulli type and that the probability of occurrence of an edge between two vertices described by feature vectors X and X' given the labeled graph only depends on the pair (X,X') (combined with standard complexity assumptions on the set of candidate prediction rules), excess risk bounds of the order O (1/n) for the empirical risk minimizers have been established by [19] (improving upon the results obtained in the seminal contribution of [18]) based on a representation of the objective functional very similar to the second Hoeffding decomposition for U-statistics of degree two. The computational complexity of finding an empirical risk minimizer scaling as O (n^2), since the empirical graph reconstruction risk involves summing up over n (n - 1)/2 terms, performance of minimizers of computationally cheaper Monte-Carlo estimates of the empirical reconstruction risk, built by averaging over B <<n^2 pairs of vertices drawn with replacement have also been investigated in order to scale up the empirical graph reconstruction procedure.

While the results obtained in this domain rely on very restrictive structural assumptions, the goal of the present research internship is to investigate to which extent they can be extended to much more realistic situations (i.e. more complex graph structures), where the degree distribution of the graph follows a power law, as in a scale-free networks, in particular.

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, Probabilité, Optimisation
- Bon niveau en programmation (e.g. Java, C/C++, Python)
- Bon niveau d'anglais

Candidatures

à envoyer à stephan.clemencon@telecom-paristech.fr, guillaume.papa@telecom-paristech. :

- 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.

Référence

[1] Spielman, D. (2005). Fast Randomized Algorithms for Partitioning, Sparsification, and Solving Linear Systems. 338 Lecture notes from IPCO Summer School 2005.

[5] Cukierski, W., Hamner, B., and Yang, B. (2011). Graph-based features for supervised link prediction. In IJCNN.

[7] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A., Snyder, M., Greenblatt, J., andGerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302(5644):449–453.

[8] Janson, S. and Nowicki, K. (1991). The asymptotic distributions of generalized U-statistics with applications to random graphs. Probability Theory and Related Fields, 90:341–375.

[9] Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data. Pharmacogenomics,
2(4):373–385.

[10] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In CIKM.

[11] Lichtenwalter, R., Lussier, J., and Chawla, N. (2010). New perspectives and methods in link prediction. In KDD.

[14] Mattick, J. and Gagen, M. (2005). Accelerating networks. Science, 307(5711):856–858.

[16] Vert, J.-P., Qiu, J., and Noble, W. S. (2007). A new pairwise kernel for biological network inference with support vector machines. BMC Bioinformatics, 8(10).

[17] Vert, J.-P. and Yamanishi, Y. (2004). Supervised graph inference. In NIPS, pages 1433–1440.

[18] Biau G. and Bleakley, K. (2006). Statistical Inference on Graphs. Statistics & Decisions,

[19] Papa G., Bellet A., Clémençon S. (2016) On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability, NIPS