CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 22/11/2015

Internship proposal

Sujet : Practical Interpolation on Partially Ordered Datasets

Can be followed up with a PhD thesis

Encadrement

Antoine Amarilli, Pierre Senellart

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : février/mars 2016

Équipe(s) d'accueil de la thèse

département INFRES, équipe DBWeb

Mots clés

Interpolation, partial orders, crowdsourcing, computational geometry, taxonomy, treewidth, data mining

Sujet détaillé

In crowdsourcing and other settings, one must determine unknown numerical values using the crowd users or other expensive means. For instance, one may wish to determine the support of association rules when performing *crowd mining*, namely, data mining on a crowd of users [4]. Other tasks include classify an item in a taxonomy by determining its compatibility score with the taxonomy classes, classifying the performance of a candidate on multiple test questions, or determining the performance of a system on configurations within a large parameter space.

In all these cases, the values that we wish to identify have *structure*. For instance, the support of association rules and frequent itemsets is *monotone* in the Boolean lattice of itemsets; the compatibility of an item in a taxonomy is monotone with respect to the taxonomy itself, and the performance of a candidate or system may be constrained by the tasks, e.g., candidates will make less errors on strict subsets of a task. This means that the values to identify are not independent, but are *related by order constraints*: we know that some values must be less than others.

Our goal is to answer a query about the values that we are acquiring. For instance, "What is the best category in which we can classify the product?" Alternatively, "On which tasks does the

candidate seem proficient?", or "Among configurations that have performance ≤ 95%, which one is the fastest?" As it is expensive to query the values, we must answer such questions even when we have only acquired a small number of values, namely, we must use the order constraints to *interpolate* the missing values from the known ones. This is a complex task, however, as the relation between these values is an arbitrary partial order.

In collaboration with Yael Amsterdamer and Tova Milo from Tel Aviv University, we have investigated how to define a principled interpolation scheme. Our current draft [2] studies partial orders with some known exact values, and defines interpolation as computing the center of mass of the convex polytope defined by the constraints. We propose a brute-force algorithm for this task, show that the task is intractable in general, but identify simple cases where it is tractable, e.g., tree-shaped posets, whose Hasse diagram (DAG of compatibility relations) is a tree.

The main goal of this internship is to study practical applications of our current work. First, at a theoretical level, by studying more general classes of posets for which the interpolation problem can be solved. For instance, we suspect that our tractability result extends from tree-shaped posets to *tree-like posets*, where the Hasse diagram has *bounded treewidth*. This may follow from the applicability of message-passing techniques for inference in bounded-treewidth graphical models [5], and may relate to our earlier work [3] on bounded-treewidth probabilistic structures.

Second, at a practical level, the goal would be to measure experimentally how well missing data can be fitted by interpolation. A benchmark could compare our scheme to simpler schemes that ignore the missing data values or complete them in a straightforward way. Another interesting question is how much the choice of *prior* impacts the performance, and whether we can do better than a uniform prior in some cases. The evaluation can be performed, e.g., on datasets collected from crowd users by our collaborators in Tel Aviv University.

Other possible directions for the internship include the study of alternative principled interpolation models, e.g., those ensuring that interpolation is *stable*, meaning that fixing an unknown item to its interpolated value does not change the other values. More ambitiously, one could study situations where the observed values are themselves uncertain or inconsistent, or study the more general problem of deciding which missing values should be acquired next [4, 1].

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

Student who is currently pursuing a Master (M2) in computer science

- Familiarity with theoretical computer science, complexity theory and algorithmics
- Experience in programming to implement practical benchmarks
- Fluent level of English

Candidatures

Interested candidates can apply by emailing pierre@senellart.com and a3nm@a3nm.net. Please indicate which master you are currently pursuing.

Référence

[1] A. Amarilli, Y. Amsterdamer, and T. Milo. On the complexity of mining itemsets from the crowd using taxonomies. In Proc. ICDT, 2014. http://arxiv.org/abs/1312.3248.

[2] A. Amarilli, Y. Amsterdamer, T. Milo, and P. Senellart. Top-k querying for incomplete data under order constraints. Draft: http://pierre.senellart.com/publications/amarilli2016top.pdf, 2015.

[3] A. Amarilli, P. Bourhis, and P. Senellart. Provenance circuits for trees and treelike instances. In Proc. ICALP, 2015. http://a3nm.net/publications/amarilli2015provenance.pdf.

[4] Y. Amsterdamer, Y. Grossman, T. Milo, and P. Senellart. Crowd mining. In Proc. SIGMOD, 2013. http://pierre.senellart.com/publications/amsterdamer2013crowd.pdf.

[5] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. J. Royal Statistical Society. Series B, 1988. http://intersci.ss.uci.edu/wiki/pdf/Lauritzen1988.pdf.