Sujet de Master 2 Recherche

Apprentissage incrémental/anytime de modèles relationnels probabilistes

Introduction

De par la multiplication croissante des données produites, le traitement des masses de données, souvent évoqué sous le terme ambigu de Big Data, joue un rôle primordial dans la société de demain [Hamel et Marguerit, 2013]. Fouille de données et Apprentissage sont au coeur de ce processus de traitement des masses de données, avec pour objectif de transformer des données en des modèles informatiques ensuite embarqués dans des applications destinées à des utilisateurs finaux. Cependant, comme le note [Amershi et al. 2014], les utilisateurs potentiels de ces applications, souvent experts du domaine d'application, ont une participation limitée dans le processus de leur développement. Le processus classique de construction de tels modèles est centré sur un analyste des données et l'usager final intervient peu, se limitant à la fourniture de données, à la réponse à des questions relatives au domaine, ou à donner un avis général sur le modèle appris.

Pour faire le parallèle avec les méthodologies Agile pour le développement logiciel, de nouvelles approches de Fouille/Apprentissage font leur apparition, avec pour objectif la mise à jour de modèles incrémentaux, permettant aux usagers d'explorer interactivement le modèle, de conduire le système vers un comportement attendu, et plus généralement, toujours d'après S. Amershi, de faciliter la démocratisation de l'apprentissage ou de la fouille et de permettre la prise en main de tels algorithmes par des utilisateurs finaux pour créer des systèmes basés sur leurs propres besoins.

Cette interaction effective avec l'utilisateur introduit de nouveaux défis en Apprentissage et en Fouille de données, défis qui sont au cœur du projet de l'équipe DUKe (Data User Knowledge) du Laboratoire d'Informatique de Nantes (LINA UMR CNRS 6241).

L'ambition générale de l'équipe est de proposer des algorithmes de publication des données, d'apprentissage interactif, de fouille exploratoire, dans lesquels l'utilisateur joue un rôle actif. Pour l'un des enjeux qui nous intéresse consiste à considérer l'utilisateur comme un acteur du processus de fouille de données ou d'apprentissage. L'idée est de développer des mécanismes de fouille ou d'apprentissage reposant sur une structure itérative, incrémentale et adaptative impliquant l'utilisateur final. Notre objectif est de proposer des algorithmes possédant de bonnes propriétés en terme d'interactivité : interruptibles à tout moment (anytime), incrémentaux, rapides, pouvant incorporer des connaissances externes fournies par l'utilisateur, ou disposant de métaphores graphiques permettant une interaction riche.

Les modèles graphiques probabilistes tels que les réseaux bayésiens [Pearl 88] sont un outil de modélisation de connaissances et de raisonnement utilisés de manière croissante dans de nombreux domaines. Les travaux théoriques initiés par J. Pearl dans les années 80 ont ainsi récemment été récompensés par le prix Turing 2013, équivalent au prix Nobel d'Informatique.

Ces modèles ont donné lieu à de nombreuses extensions comme les PRM, Probabilistic Relational Models [Friedman et al. 99, Getoor et al. 07] pour la prise en compte de données structurées stockées dans des bases de données relationnelles.

L'équipe DUKe du LINA est une des spécialistes européennes dans ce domaine, avec des compétences en terme d'apprentissage des réseaux bayésiens à partir de données [Naim et al. 07] et leur utilisation dans le domaine de la recommandation. [Ben Ishak et al. 13, Coutant et al. 13, Chulyadyo and Leray, 2015]. Ces travaux se sont aussi traduit par une collaboration avec le LARODEC (ISG Tunis) et deux thèses en cotutelles soutenues, et deux autres en cours.

Objectif du stage

L'objectif est de réfléchir sur l'apprentissage incrémental et anytime des modèles relationnels probabilistes, et de proposer une première solution s'appuyant sur l'existant.

Travail à réaliser

- Se familiariser avec le formalisme des PRM
- Se familiariser avec les algorithmes d'apprentissage incrémentaux/anytime existants pour

les réseaux bayésiens « simples »
- Se familiariser avec les algorithmes d'apprentissage classique des PRM
- Proposer une première solution pour l'apprentissage incrémental/anytime des PRM
- Implémenter un prototype à l'aide de la plate-forme PILGRIM développée par l'"équipe DUKe.

Ce travail sera supervisé par P. Leray (LINA / DUKe, Nantes) et M. Ben Messaoud (LARODEC / ISG Sousse, Tunisie). Le stagiaire sera intégré à une équipe de plusieurs doctorants et stagiaires travaillant sur les PRM.

Période : Février-Juillet

Compétences
- Concepts de probabilité, statistiques et bases de données relationnelles
- Programmation C++

Références

[Amershi et al., 2014]	S. Amershi et al., Power to the people: The role of humans in interactive machine learning. AI Magazine, 35, 4 (Winter 2014), pp. 105-120.
[Ben Ishak et al., 2013]	Ben Ishak, M., Ben Amor, N., and Leray, P. (2013). A relational bayesian network-based recommender system architecture. In Proceedings of the 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO 2013), pages 1-6, Hammamet, Tunisia.
[Chulyadyo et Leray, 2015]	Chulyadyo, R. and Leray, P. (2015). Integrating spatial information into probabilistic relational model. In Proceedings of 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), pages ?-?, Paris, France.
[Coutant et al., 2013]	Coutant, A., Leray, P., and Le Capitaine, H. (2013). Learning probabilistic relational models using co-clustering methods. In Structured Learning: Inferring Graphs from Structured and Unstructured Inputs (SLG 2013) ICML Workshop, pages ?-?, Atlanta, USA.
[Friedman et al. 1999]	Friedman N, Getoor L, Koller D, Pfeffer A. (1999) "Learning probabilistic relational models". In International joint conferences on artificial intelligence, 1300–09
[Getoor et al. 2007]	L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Taskar (2007). "Probabilistic Relational Models." In L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning. MIT Press.
[Hamel et Marguerit, 2013]	M.-P. Hamel et D. Marguerit, Analyse des big data - quels usages, quels défis ? Notes d'analyses n°08, 11/2013, Commissariat général à la stratégie et à la prospective.
[Naïm et al., 2007]	Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O., and Becker, A. (2007). Réseaux bayésiens. Eyrolles, Paris, 3 edition.
[Pearl 1988]	Judea Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988