Proposition de recherche doctorale

Apprentissage automatique et inférence dans les grands réseaux collaboratifs.

Mots clés :

- **Directeur de thèse** : Patrick Gallinari
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

La thèse vise à explorer l'apprentissage statistique et l'inférence dans les grands réseaux collaboratifs liés au web. Nous cherchons à développer des procédures d'apprentissage et d'inférence pour répondre aux problèmes soulevés par de nombreuses applications comme par exemple la détection de fraude, de spam sur le web ou dans les blogs, la découverte de relations entre individus, la recommandation, et toutes sortes d'applications de la classification ou de la prévision sur ces grands graphes. Le sujet s'inscrit dans un courant de recherche extrêmement récent, qui vise à réinventer la fouille de données pour le domaine des grands réseaux collaboratifs et particulièrement les réseaux sociaux. Pour cela, nous nous appuierons sur des méthodes de l'apprentissage statistique. Aspects algorithmiques On explorera plusieurs familles de méthodes de l'apprentissage dans le but d'une part d'apprendre et de découvrir les relations présentes implicitement dans les données et d'autre part d'effectuer de l'inférence dans des grands réseaux. Le sujet conduira à revisiter dans ce cadre coopératif, des problématiques générales de l'apprentissage. Un aspect important concerne le développement de modèles d'apprentissage permettant d'analyser des données du type graphe. Deux grandes problématiques seront explorées : La découverte de relations Un des objectifs de la thèse est de proposer des modèles permettant la découverte de relations sur les contenus et les individus. La découverte de relations entre objets, entre individus ou entre objets et individus peut se formuler naturellement comme un problème de découverte de variables latentes. Ces dernières représentent alors les relations sémantiques entre données, les thématiques présentes dans les données de contenu, les relations entre individus, etc. Ces modèles permettent de faire de l'inférence sur les différentes variables du problème par le calcul de probabilités marginales. Ils permettent également de faire de l'inférence sur plusieurs variables simultanément et d'identifier par exemple les groupes d'individus constituant une communauté thématique. L'inférence collective L'inférence collective consiste à calculer des scores en chacun des nœuds de ce réseau, correspondant par exemple à une classe, une valeur de pertinence, un score d'ordonnancement. Sur le plan de la fouille de données, les deux premiers cas correspondent à des problèmes de classification collective où il s'agit d'affecter une étiquette aux nœuds, le troisième cas correspond à une problématique d'ordonnancement où il s'agit d'ordonner des informations (documents ou images) en fonction de leur pertinence pour un besoin d'information. Les problèmes d'inférence collective sont en général NP-complets. Les algorithmes développés doivent permettre de calculer de façon efficace des solutions approchées au problème. Les algorithmes d'inférence collective exploitent soit des extensions des techniques de relaxation, de classification itérative, et plus généralement des techniques inspirées des champs markoviens, soit des méthodes qui combinent l'optimisation d'une fonction de score portant sur les caractéristiques des nœuds et des contraintes liées à la structure relationnelle. Toutes ces méthodes ont été développées dans un cadre de classification bi-classe ou multi-classe. Ces travaux sont souvent extrêmement récents (2006-2007 - voir les références ci dessous) et leur potentiel opérationnel est encore limité. Un premier travail sera de sélectionner quelques représentants de ces principales familles de méthodes et de tester le passage à l'échelle des principales familles d'algorithmes univariés sur le cas des problèmes de classification (filtrage, découverte de communautés). Dans un second temps il s'agira de développer des extensions de ces méthodes qui concernent : o L'apprentissage et l'inférence pour des problèmes autres que la classification. o La prise en compte de relations multiples correspondant à différentes sources d'informations relationnelles. o Le passage à l'échelle. Les applications Parallèlement au développement des méthodes, nous étudierons une sélection d'applications clé sur qui nous permettrons de tester sur des cas réels et en grande dimension les algorithmes développés. Ces applications seront liés à l'analyse des réseaux sociaux, avec des problèmes comme comme la découverte de liens ou relations entre individus et entre données, la découverte du ou des rôles des individus dans une communauté.

Résumé du projet de recherche (Langue 2)

La thèse vise à explorer l’apprentissage statistique et l’inférence dans les grands réseaux liés au web. L’apprentissage statistique a pour l’instant été développé sur des hypothèses classiques d’indépendance des données qui sont erronées dans le cas des données en réseau. Il s’agit d’explorer cette nouvelle thématique en développant des méthodes d’apprentissage adaptées au cas des données relationnelles. Cette nouvelle problématique constitue une véritable rupture dans le domaine du traitement des données. D’un point de vue technique tout reste à inventer. Quelques références Références [Cook 96] Diane J. Cook, Lawrence B. Holder, Surnjani Djoko. Scalable Discovery of Informative Structural Concepts Using Domain Knowledge. IEEE Expert: Intelligent Systems and Their Applications. Volume 11 , Issue 5 (October 1996). 59-68 [Lu 03] Lu Q., Getoorr L., Link based classification, ICML, 496-503, 2003. [Macskassy 2007] Sofus A. Macskassy, Foster Provost. Classification in Networked Data: A Toolkit and a Univariate Case Study. The Journal of Machine Learning Research, Volume 8, 2007. [McCallum 07] Andrew McCallum, Xuerui Wang and Andres Corrada-Emmanuel Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of Artificial Intelligence Research (JAIR), 2007. [Zhou 06] Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, Hongyuan Zha, Probabilistic Models for Discovering E-Communities ,In proceedings of the 15th ACM International World Wide Web Conference, Scotland, 2006. [Zhou07] D. Zhou, J. Huang and B. Schölkopf. Learning with Hypergraphs, Clustering , Classification and Embedding. Proceeding of the NIPS. 2007

Informations complémentaires (Langue 1)

Ces travaux feront l'objet d'échanges internationaux dans le cadre du réseau d'excellence Pascal. Nous avons co-organisé dans ce cadre des workshops internationaux et développé des relation avec des équipes dans différents pays. Des échanges sont possible dans le cadre du développement de projets communs au sein de ce réseau.