Proposition de recherche doctorale

Clustering semi-supervisé et apprentissage actif

Mots clés :

- **Directeur de thèse** : Bernadette Bouchon-Meunier
- **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)

Le clustering est une tâche centrale du processus d'exploration de données et de découverte de connaissances. De nos jours, l'abondance de données et l'augmentation continue de leur volume imposent aux algorithmes de clustering de s'améliorer et de s'adapter selon les aspects suivants : qualité, vitesse, passage à échelle. Pour toutes ces raisons, le domaine du clustering est toujours extrêmement actif. Le clustering semi-supervisé est ainsi devenu depuis une dizaine d'années une piste de recherche très intéressante dont le but est de développer des algorithmes de clustering qui permettent à un expert humain d'intégrer des connaissances du domaine pour améliorer la pertinence des analyses. Ces connaissances peuvent être exprimées soit par un ensemble de données étiquetées (des seeds) ou soit par un ensemble de contraintes. Dans ce dernier cas, on distingue deux principaux types de contraintes : les must-link (ML) qui indiquent que deux points de l'ensemble de données doivent être dans le même groupe et les cannot-link (CL), qui inversement imposent que deux points appartiennent à deux clusters différents. Bien que les travaux actuels s'intéressent plus particulièrement à l'adaptation de méthodes de clustering existantes pour la prise en charge de contraintes ou de données étiquetées, ils conservent les mêmes limitations que les méthodes dont ils s'inspirent et reposent sur une sélection aléatoire des connaissances qui peut conduire à de mauvaises performances. Pour répondre à ces problèmes, cette thèse s'articule autour de deux contributions principales : (1) des méthodes intelligentes pour la sélection de contraintes ou de données étiquetées (les seeds) intégrées à des algorithmes actifs et (2) de nouveaux algorithmes de clustering semi-supervisé qui améliorent les méthodes décrites dans la littérature Dans le cadre de la collecte intelligente de contraintes, nous proposons une première mesure d'utilité d'une contrainte qui repose sur un graphe des k-plus proches voisins pour identifier les zones de transition entre clusters où les algorithmes font traditionnellement le plus d'erreurs. Cette mesure forme la base de notre algorithme actif de sélection de contraintes qui a été validé sur des jeux de données issus du UCI Machine Learning Repository ainsi que dans le cadre d'un prototype logiciel appliqué à l'analyse de bases d'images. Similairement, nous proposons trois nouvelles méthodes pour la sélection de données à étiqueter qui ont été évaluées également sur des données réelles et dans le cadre d'un prototype logiciel sur des bases d'images. Enfin cette thèse décrit deux nouveaux algorithmes de clustering : SSGC basé sur les seeds et MCLA basé sur les contraintes qui possèdent des complexités plus réduites, un paramétrage plus aisé et des performances comparables voire meilleures que les algorithmes de référence en clustering semi-supervisé. Mots-clés : algorithme de clustering, clustering semi-supervisé, apprentissage actif, contraintes, seeds, graphe des k-plus proches voisins.