Nouveaux algorithmes de fouilles de données relationnelles de clowdflows

Nicolas Lachiche*, Alain Shakour*

*ICube 300 bd Brant 67400 Illkirch
{nicolas.lachiche,ashakour}@unistra.fr,
http://clowdflows.unistra.fr

Résumé. Clowdflows est un logiciel open source qui permet à un utilisateur de réaliser des processus entiers de fouille de données à partir d'un navigateur et d'une connexion internet. Les calculs sont réalisés dans le "nuage", c'est-à-dire de façon transparente sur plusieurs serveurs exécutant les calculs ou hébergeant les données. Dans cet article, nous rappelons les points forts de clowdflows et nous présentons trois familles d'algorithmes de fouille de données relationnelles que nous venons d'y intégrer. En effet clowdflows est la seule plateforme web permettant d'exécuter, voire comparer, plusieurs techniques de fouille de données relationnelles, souvent appelée programmation logique inductive.

1 Introduction et état de l'art

Un grand nombre d'algorithmes sont conçus et implémentés par les chercheurs en fouille de données. Peu d'entre eux font l'objet d'une valorisation et conduisent à un logiciel destiné aux utilisateurs. Il est souvent possible de se procurer le code ou un exécutable auprès des auteurs, mais ces logiciels sont souvent difficiles à mettre en oeuvre pour de nombreuses raisons. En effet, peu de documentation est disponible. Ils utilisent en général des formats d'entrée et sortie qui leur sont spécifiques. Le rôle et le réglage des paramètres relèvent plutôt de l'expertise des concepteurs du logiciel.

Néanmoins quelques logiciels de fouille de données sont plus connus et plus accessibles. Nous pouvons déjà citer la nébuleuse de bibliothèques en python ou R mises à disposition par leurs développeurs. Pour autant, elles n'échappent pas aux défauts cités précédemment. Et malgré leurs langages de programmation communs, elles ne sont pas intégrées en un seul environnement facile à utiliser pour les mettre en oeuvre et les comparer. A l'inverse, des logiciels comme knime ou rapidminer offrent une bonne intégration et facilité d'utilisation, mais sont gérés par des entreprises commerciales. Peu d'algorithmes produits en recherche ont une chance d'y être intégré. Il existe quelques environnements de fouille de données, libres et gratuits, comme weka. L'installation est facile mais il est nécessaire de les télécharger et de les installer sur la machine de l'utilisateur.

D'autres domaines, par exemple la bioinformatique, montrent l'exemple de services disponibles sur internet, sans installation, et permettent facilement à un utilisateur d'appliquer des algorithmes existants à ses données. Dans le domaine de la fouille de données, peu de services sont proposés sur le web. Nous pouvons citer OpenML.org. Il cible le travail collaboratif et