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

sert à partager des données, des expériences et des résultats, mais il ne permet pas d'exécuter, ni même de télécharger les programmes.

La situation est encore plus difficile en fouille de données relationnelles. Pour mémoire, la plupart des logiciels de fouille de données s'appliquent à des données attributs-valeurs, c'est-à-dire qui peuvent naturellement être disponible dans un tableur, ou une seule table d'une base de données relationnelle. La fouille de données relationnelles, comme son nom l'indique, considère le cas de données qui seraient naturellement représentées par plusieurs tables dans une base de données relationnelle, par exemple des clients et leurs achats, des patients et leurs examens médicaux, etc. Ces données ne sont pas forcément stockées dans un Système de Gestion de Bases de Données (SGBD). Elles peuvent être présentées sous la forme de faits prolog, en logique des prédicats du premier ordre. D'ailleurs les techniques de fouille de données relationnelles se rattachent en général à la programmation logique inductive. Nous constatons malheureusement que la plupart des algorithmes proposés par les chercheurs de ce domaine sont difficiles à se procurer et à mettre en oeuvre.

Ainsi nous constatons le manque de services web pour faciliter la fouille de données, en particulier relationnelles. Dans cet article, nous considérons la plateforme clowdflows.org. Nous passerons en revue ses caractéristiques principales dans la prochaine section. Dans la section suivante, nous présenterons trois familles d'algorithmes de fouille de données relationnelles que nous avons ajoutés à cette plateforme. Nous conclurons en proposant quelques perspectives pour rendre la fouille de données plus facile à mettre en oeuvre grâce à des plateformes web telles que clowdflows.

2 Clowdflows

Clowdflows est un logiciel de fouille de données proposé et développé par l'équipe de Nada lavrac au Jozef Stefan Institute à Ljubljana, en Slovénie (Kranjc et al., 2012). Le site original est clowdflows.org. Un miroir est disponible au laboratoire ICube à Strasbourg, clowdflows.u-strasbg.fr La mise à disposition de miroirs permet de répartir la charge, mais surtout de rapprocher les algorithmes des données. En effet à l'ère des données massives, il semble préférable d'éviter de déplacer de telles données et de permettre aux algorithmes de fouille de travailler avec les données où elles sont stockées. Le nom de la plateforme, clowdflows, indique qu'elle concerne les flux de traitement, par le terme flows, d'une part. D'autre part, le terme clowd, est l'ancienne orthographe de cloud et met l'accent sur le fait que la fouille s'exécute dans le nuage, et rappelle aussi par sa proximité avec le terme crowd, que clowdflows permet à une foule, au moins une communauté d'utilisateurs ou contributeurs, de partager des flux de traitements.

Clowdflows nécessite un accès au réseau internet. Son interface graphique s'exécute sur un navigateur disposant de javascript. Il n'y a rien à installer. Il suffit de se créer un compte sur le site [1]. On peut ainsi travailler depuis n'importe où. Tout est sauvegardé sur le serveur, à part les données si elles viennent d'un SGBD. Les données peuvent être chargées depuis des fichiers ou depuis un SGBD. Les SGBD MySQL et PostgreSQL sont déjà pris en charge. D'autres pourront facilement être ajoutés. Les traitements s'exécutent sur le serveur. En fait, clowdflows

1. Dans un premier temps, les données ne sont pas partagées entre les miroirs clowdflows. Il faut donc créer un compte sur chaque miroir que l'on utilise et les flux ne sont accessibles que sur ce miroir. Cependant il est possible d'exporter des flux d'un miroir pour les importer sur un autre.

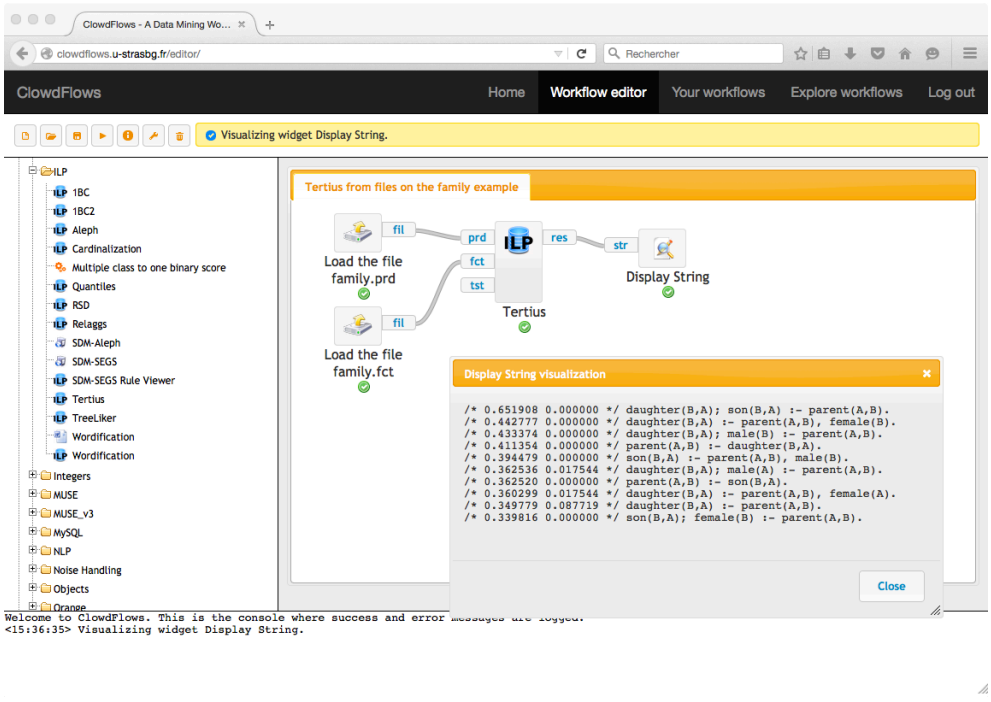

FIG. 1 – *Exemple de flux de traitement avec Tertius sur clowdflows*

peut appeler d'autres serveurs pour exécuter des algorithmes de fouille mis à disposition sous la forme de services web. La charge de calcul peut ainsi être répartie entre plusieurs serveurs afin de permettre une montée en charge, transparente à l'utilisateur, typique du nuage, cloud. Actuellement il y a une répartition des traitements. Il n'y a pas encore de répartition de charge entre les serveurs.

L'interface graphique est présentée par la figure 1. Une fois que l'utilisateur s'est familiarisé avec l'environnement de travail, l'interface graphique se révèle conviviale et facile à prendre en main. Le flux de traitement est composé graphiquement dans la fenêtre principale, en sélectionnant des composants et en les reliant dans l'ordre des traitements à effectuer. L'ensemble des composants est disponible dans la barre de gauche, organisée sous la forme d'une arborescence. Il est facile de réaliser un processus complet de fouille de données, du chargement des données à l'affichage des résultats, en passant par les prétraitements, séparation des données d'entraînement et de test, construction des modèles, et même comparaison de plusieurs approches. De nombreuses fonctionnalités, prétraitements et algorithmes d'apprentissage, de plusieurs librairies connues (weka, orange, scikit) y sont déjà intégrées. De plus on peut aisément intégrer graphiquement des web services, sans modifier le code source de clowdflows, et les utiliser directement dans des flux de traitements. Enfin il est possible de contribuer au code source python, qui devra ensuite être déployé sur les différents miroirs.

Une fonctionnalité importante de clowdflows est la possibilité de rendre public un flux de

traitement à travers une URL . Cette URL peut être incluse dans un article, comme nous le ferons dans les sections suivantes. Cela permet à d'autres utilisateurs de reproduire exactement l'expérience réalisée par les auteurs. De plus ces flux de traitements publics servent de tutoriels. Ils sont regroupés et facilement accessibles sur la plateforme. Ils permettent de voir à travers des exemples les différents problèmes de fouille qui peuvent être résolus et de s'en inspirer pour résoudre de nouveaux problèmes.

À notre connaissance clowdflows est un des rares logiciels intégrant plusieurs algorithmes de fouille de données relationnelles et le seul s'exécutant directement sur un navigateur, sans aucune installation. Dès ses premières versions, clowdflows permet d'exécuter aleph et son composant de générations de nouveaux traits, features en anglais, et des algorithmes de propositionalisation, RSD, RelF, et la wordification (Lavrac et al., 2014).

3 Trois nouvelles familles de techniques de fouille de données relationnelles

Comme clowdflows est un logiciel ouvert (open source), il est possible d'ajouter des fonctionnalités. Nous avons ainsi ajouté aisément plusieurs algorithmes de fouille de données relationnelles développés par des membres de ICube :

- un algorithme de découverte de règles : Tertius,
- deux classeurs bayésiens naïfs : 1BC et 1BC2,
- deux techniques de propositionalisation : la cardinalisation et les quantiles.

3.1 Découverte de règles : Tertius

Tertius (Flach et Lachiche, 2001) est un algorithme de découverte de règles. Il s'appuie sur une mesure de qualité, appelée confirmation, combinant support et confiance, et permet d'extraire les k meilleures règles ou toutes les règles de confirmation supérieure à un seuil choisi par l'utilisateur.

La figure 1 montre un exemple typique d'utilisation de tertius, à partir de fichiers de faits prolog, pour découvrir des règles entre les prédicats parent, fille, fils, male et femelle. Cet exemple est accessible et peut être reproduit à l'URL http ://clowdflows.unistra.fr/workflow/77/

3.2 Classeurs bayésiens : 1BC et 1BC2

1BC et 1BC2 (Flach et Lachiche, 2004) sont deux classeurs bayésiens naïfs qui s'appliquent à des données relationnelles. 1BC se sert du biais de langage pour réaliser une propositionalisation à la volée. 1BC2 estime de façon récursive les probabilités de listes ou ensembles d'éléments étant données leurs probabilités respectives.

La figure 2 donne un exemple de mise en oeuvre de 1BC à partir d'une base de données relationnelles, avec affichage d'une courbe ROC. Cet exemple d'application aux molécules mutagènes en validation croisée peut être reproduit et modifié à partir de l'URL http ://clowdflows.unistra.fr/workflow/14/

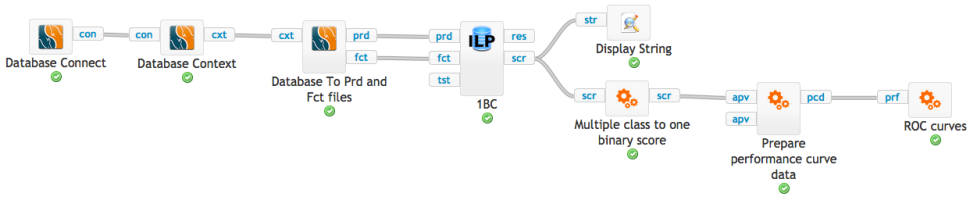

FIG. 2 – *Application de 1BC sur mutagenesis stocké en base de données*

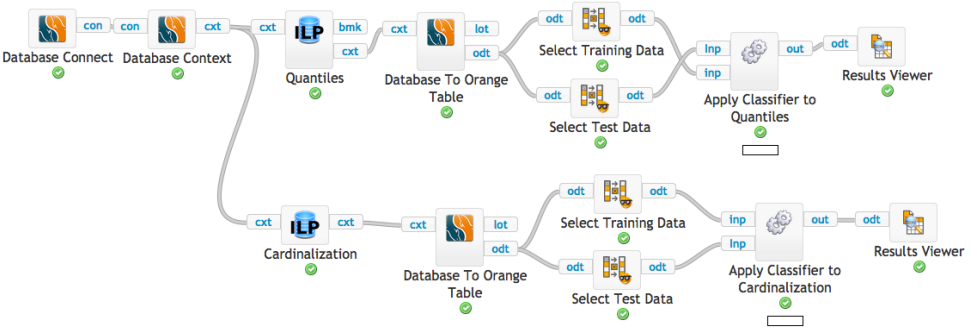

FIG. 3 – *Comparaison des prédictions de la cardinalisation et des quantiles*

3.3 Propositionalisation : cardinalisation et quantiles

La cardinalisation et les quantiles (Ahmed et al., 2015) sont deux approches de propositionalisation. Elles s'appliquent aux attributs numériques des tables secondaires, par exemple le montant des achats des clients, et calculent le montant tel qu'il y ait au moins n éléments de la table secondaire, par exemple au moins n achats par client, pour la cardinalisation, et le montant tel qu'il y ait au moins une certaine proportion des achats en dessous de ce montant, pour les quantiles.

La figure 3 montre un exemple où les prédictions de ces techniques de propositionalisation sur les ilots urbains sont comparées. Nous avons également intégré une approche plus simple de propositionalisation, Relaggs (Krogel et Wrobel, 2001), qui utilise les fonctions d'agrégation classiques de SQL : minimum, maximum, moyenne, etc. Cet exemple est disponible à l'URL http ://clowdflows.unistra.fr/workflow/78/

4 Conclusion

Clowdflows est une plateforme qui simplifie l'utilisation d'algorithmes existants de fouille de données en les rendant accessibles à travers un navigateur internet et en exécutant les calculs sur des serveurs dédiés. Les flux de traitements créés graphiquement peuvent être rendus

publics afin de reproduire les expériences présentées dans un article ou pour servir d'exemple à modifier pour résoudre des problèmes similaires. De plus la plateforme peut facilement être complétée. Des algorithmes accessibles par des web services peuvent être intégrés à chaud dans un flux de traitement, sans avoir à modifier le code source. Et d'autres composants graphiques peuvent être ajoutés en modifiant le code qui est open-source. Une plateforme telle que clowdflows permet de mettre à disposition un vrai service de fouille de données, nécessaire pour simplifier la pratique de la fouille de données à l'ère du big data.

Références

Ahmed, C. F., N. Lachiche, C. Charnay, S. E. Jelali, et A. Braud (2015). Flexible propositionalization of continuous attributes in relational data mining. *Expert Syst. Appl. 42*(21), 7698–7709.

Flach, P. A. et N. Lachiche (2001). Confirmation-guided discovery of first-order rules with tertius. *Machine Learning 42*(1/2), 61–95.

Flach, P. A. et N. Lachiche (2004). Naive bayesian classification of structured data. *Machine Learning 57*(3), 233–269.

Kranjc, J., V. Podpecan, et N. Lavrac (2012). Clowdflows : A cloud based scientific workflow platform. In P. A. Flach, T. D. Bie, et N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, Volume 7524 of *Lecture Notes in Computer Science*, pp. 816–819. Springer.

Krogel, M. et S. Wrobel (2001). Transformation-based learning using multirelational aggregation. In C. Rouveirol et M. Sebag (Eds.), *Inductive Logic Programming, 11th International Conference, ILP 2001, Strasbourg, France, September 9-11, 2001, Proceedings*, Volume 2157 of *Lecture Notes in Computer Science*, pp. 142–155. Springer.

Lavrac, N., M. Perovsek, et A. Vavpetic (2014). Propositionalization online. In T. Calders, F. Esposito, E. Hüllermeier, et R. Meo (Eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part III*, Volume 8726 of *Lecture Notes in Computer Science*, pp. 456–459. Springer.

Summary

Clowdflows is an open-source software that enable users to define and run entire data mining process from a web browser and an internet connection. Computations run in the cloud, that is to say transparently on several servers, sharing computations or hosting data. In this article, we remind the strengths of clowdflows and we present three families of relational data mining algorithms that we recently integrated into clowdflows. Indeed clowdflows is the only web platform able to run and compare several relational data mining techniques, also known as Inductive Logic Programming.