Sujet de thèse doctorat en informatique

Approches déclaratives efficaces pour l'extraction de motifs d'intervalles clos

Normandie Univ., UNICAEN, CNRS – GREYC UMR 6072, France

1 Contexte

En fouille de données, l'extraction de motifs [7] vise à identifier des régularités dans des ensembles de données (datasets). Ces motifs permettent de faire émerger des relations implicites dans une grande masse de données. Il appartient ensuite aux data scientists et aux experts sur les données de déterminer si un motif est le résultat d'une simple corrélation, ou s'il est le fruit d'une relation directe entre ses composants, par exemple un lien de causalité entre deux composants. De nombreuses applications en fouille de données sont amenées à traiter des données numériques. Dans ce contexte, il est souvent nécessaire de passer par une étape de binarisation pour profiter des méthodes symboliques. Cette binarisation s'effectue soit en créant des attributs pour constituer des intervalles sur les valeurs des données [4], soit en partitionnant les valeurs dans le dataset en catégories selon différents paramètres : taille, composition de blocs au regard de leur classe, tests statistiques, entropie, etc. [2]. La binarisation entraîne une perte d'information par rapport aux données d'origine. Une technique comme l' "interordinal scaling" permet de préserver l'information d'origine mais conduit à des données de grande taille. Afin d'avoir une méthode capable de traiter directement les données numériques, Kaytoue et al. [4] ont proposé la méthode *MintIntChange* capable d'extraire des motifs d'intervalles fermés réduisant ainsi la chaîne de traitement de données. Cependant, étendre *MintIntChange* pour être capable de traiter d'autres tâches de fouille avec de nouvelles contraintes s'avère une tâche non triviale.

2 Sujet

La synergie entre la fouille de données et les paradigmes déclaratifs tels que SAT, la Programmation Par Contraintes (PPC) et la Programmation Linéaire en Nombre Entier (PLNE) a connu un grand essor au cours de la dernière décennie à travers différents travaux [5, 1, 8, 6]. L'avantage principal de ces approches réside dans leur côté déclaratif qui offre une large flexibilité pour s'adapter à des tâches variées en fouille de données. Ces approches incluent la possibilité d'intégrer de nouvelles contraintes spécifiées par l'utilisateur sans besoin de développer de nouveaux algorithmes spécifiques de résolution.

L'objectif de ce travail de thèse est de définir des méthodes et de concevoir des outils permettant l'extraction de motifs d'intervalles dans un cadre déclaratif. En effet, travailler directement sur les données numériques est un enjeu majeur pour réduire les étapes dans la chaîne de traitement des données. De plus, les approches déclaratives sont particulièrement adaptées pour cette tâche. Dans le cadre de la programmation par contraintes, différents types de contraintes sont à la disposition de l'utilisateur pour définir le réseau de contraintes à résoudre : des contraintes prédéfinies dans le solveur ou des contraintes définies en extension par un ensemble de valeurs autorisées ou interdites. En outre, un utilisateur expert peut définir ses propres contraintes en établissant la sémantique de la contrainte, ainsi que l'algorithme de filtrage associé [3, 5, 1].

Les principales contributions attendues sont :

- Proposition d’une approche déclarative offrant un bon compromis entre efficacité et flexibilité pour l’extraction de motifs d’intervalles.
- Les approches proposées doivent être capables d’enrichir l’expérience de l’utilisateur en mettant en place des moyens (contraintes ou techniques d’apprentissage) pour éviter d’inonder l’utilisateur avec des motifs inintéressants.
- L’avantage principal des méthodes déclarative réside dans leur aspect générique et tout domaine applicatif générant des données numériques peut être considéré comme un objet d’étude. Néanmoins, nous visons dans cette thèse l’extraction de motifs dans des bases de données moléculaires où les molécules sont décrites par des descripteurs numériques.

3 Contexte et encadrement

L’étudiant(e) sera accueilli(e) au sein de l’équipe CODAG du laboratoire GREYC de l’université de Caen Normandie. Les travaux seront dirigés conjointement par Patrice Boizumault (Pr), Bruno Crémilleux (Pr) et Abdelkader Ouali (MC).

4 Financement

ANR HAISCoDe (Contrats doctoraux en intelligence artificielle).

5 Profil du candidat

Le(a) candidat(e) doit être titulaire d’un diplôme master ou équivalent en informatique avec des compétences dans le domaine de la programmation par contraintes, la programmation linéaire et la fouille de motifs. Le candidat(e) recherché(e) devra avoir de solides compétences en programmation C++, JAVA, et Python3. La maîtrise des solveurs comme OR-tools, SCIP, Choco serait un vrai plus.

6 Candidature

Pour candidater, merci d’envoyer à l’adresse patrice.boizumault@unicaen.fr, bruno.cremilleux@unicaen.fr, et abdelkader.ouali@unicaen.fr les documents suivants au format PDF :

- Curriculum Vitae;
- lettre de motivation détaillant vos expériences, enseignants ou personnes avec qui vous avez travaillé, et votre intérêt par rapport à la thèse;
- relevés de notes de L3 et de Master;
- relevé de notes de licence 3, de master ou équivalent pour les écoles d’ingénieurs;
- si possible, noms de personnes (enseignants ou autre personne) pouvant fournir des informations sur vos compétences et votre travail.

Références

- [1] Mohamed-Bachir Belaid, Christian Bessiere, and Nadjib Lazaar. Constraint programming for mining borders of frequent itemsets. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1064–1070. ijcai.org, 2019.
- [2] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In Armand Prieditis and Stuart J. Russell, editors, *Machine Learning, Proceedings*

- of the *Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 194–202. Morgan Kaufmann, 1995.
- [3] Arnold Hien, Samir Loudni, Noureddine Aribi, Yahia Lebbah, Mohammed El Amine Laghzaoui, Abdelkader Ouali, and Albrecht Zimmermann. A relaxation-based approach for mining diverse closed patterns. In Frank Hutter, Kristian Kersting, Jeffrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part I*, volume 12457 of *Lecture Notes in Computer Science*, pages 36–54. Springer, 2020.
 - [4] Mehdi Kaytoue, Sergei O. Kuznetsov, and Amedeo Napoli. Revisiting numerical pattern mining with formal concept analysis. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1342–1347. IJCAI/AAAI, 2011.
 - [5] Nadjib Lazaar, Yahia Lebbah, Samir Loudni, Mehdi Maamar, Valentin Lemière, Christian Bessiere, and Patrice Boizumault. A global constraint for closed frequent pattern mining. In Michel Rueher, editor, *Principles and Practice of Constraint Programming - 22nd International Conference, CP 2016, Toulouse, France, September 5-9, 2016, Proceedings*, volume 9892 of *Lecture Notes in Computer Science*, pages 333–349. Springer, 2016.
 - [6] Abdelkader Ouali, Samir Loudni, Yahia Lebbah, Patrice Boizumault, Albrecht Zimmermann, and Lakhdar Loukil. Efficiently finding conceptual clustering models with integer linear programming. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 647–654. IJCAI/AAAI Press, 2016.
 - [7] Luc De Raedt. A perspective on inductive databases. *SIGKDD Explor.*, 4(2) :69–77, 2002.
 - [8] Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining : From pattern condensed representations to dynamic constraint satisfaction problems. *Artif. Intell.*, 244 :48–69, 2017.