

Développement de méthodes de classification basées sur l'Analyse de Concepts Formels sous la plateforme WEKA

KHALFI Besma*, CHERIF Rahma**, MEDDOURI Nida**, MADDOURI Mondher***

* khalfi.besma@hotmail.com,

** cherifrahma,nmeddouri@gmail.com

***mondher.maddouri@fst.rnu.tn

1 Introduction

Le travail présenté dans ce papier se place dans le contexte d'intégration de nouvelles méthodes de classification basée sur l'Analyse de Concepts Formels sous WEKA et propose une nouvelle famille nommée "Lattice" qui contient six méthodes basées toutes sur la construction d'un treillis de concepts. Le but de la classification est d'organiser un ensemble d'objets en classes homogènes. Les objets sont représentés par des enregistrements qui sont constitués d'un ensemble de champs (attributs) prenant leurs valeurs dans un domaine (contexte). Une classe désigne un ensemble défini de données et d'objets semblables. Ces données et objets, éléments d'une classe, sont des instances. Le treillis de concepts offre un cadre intéressant en classification. Deux phases sont indispensables dans un processus de classification, à savoir la phase d'apprentissage et la phase de classification. La phase d'apprentissage consiste à organiser l'information extraite d'un ensemble d'objets sous forme d'un treillis. La phase de classification consiste à déterminer la classe de nouveaux objets plus ou moins détériorés par la navigation dans le treillis.

2 Classification par ACF

L'Analyse de Concepts Formels est un formalisme de la notion philosophique de concepts défini comme étant un couple d'extension (ensemble des objets) et d'intention (attributs nécessaires et suffisants) permettant de caractériser et apprendre le concept [Ganter (1999)]. La classification basée sur l'Analyse de Concepts Formels est une approche symbolique de fouille de données qui permet d'extraire des corrélations, des motifs et des règles, selon les concepts générés à partir des données [Carpineto (2004)]. Plusieurs méthodes de classification basées sur l'Analyse de Concepts Formels ont été développées. Nous citons : GRAND [Oosthuizen (1988)] et RULEARNER [Sahami (1995)] qui se basent sur la classification à partir d'un treillis complet de concepts. La limite commune entre ces systèmes est la complexité exponentielle temporelle et spatiale pour la génération du treillis. Les systèmes comme LEGAL [Liquiere (1990)] et CIBLE [Njiwoua (2005)] ont la particularité de construire uniquement un demi treillis ce qui réduit leur complexité théorique, leur temps d'exécution et le nombre de concepts générés. La couverture des concepts fait partie des treillis qui ne contiennent que les concepts formels les plus pertinents. Le calcul du concept pertinent se base sur une fonction de mesure telle que l'entropie, le gain, la fermeture, etc. La construction d'une couverture de

Développement de méthodes de classification basées sur l'Analyse de Concepts Formels sous la plateforme WEKA

concepts se base sur des heuristiques qui réduisent la complexité de l'apprentissage. Parmi les méthodes qui se basent sur la notion de couverture, nous citons IPR [Maddouri (2005)] qui se base sur les pseudo-concepts et BFC qui est une approche basée sur le Boosting de Concepts Formels. Cette nouvelle approche a une complexité polynômiale meilleure que les autres méthodes basées sur ACF, ce qui explique sa rapidité.

3 Méthodes intégrées à WEKA

WEKA est une collection de classes JAVA traitant les problèmes de fouille de données. WEKA possède toute une palette d'outils pour la pré-analyse et la visualisation, la classification, la régression et des algorithmes construisants des règles d'association. La famille "Lattice" que nous avons intégré sous WEKA présente GRAND et RULEARNER (Génération d'un treillis complet), LEGAL et CIBLe (Génération d'un semi treillis) et IPR et BFC [[Meddouri (2009b),[Meddouri (2009a)] (Génération d'une couverture de concepts).

Références

- Carpineto, C., G. R. (2004). Concept data analysis: Theory and applications.
- Ganter, G., W. R. (1999). Formal concept analysis- mathematical foundations. springer.
- Liquiere, M., n. N. E. (1990). Legal: learning with galois lattice. *In Proceeding of the Sieme Journée sur l'Apprentissage (JFA'90), Lannion, FRANCE.*
- Maddouri, M. (2005). Towards a machine learning approach based on incremental concept formation. *intelligent data analysis. 8(3), 267–280.*
- Meddouri, N., M. M. (2009a). Boosting formal concepts to discover classification rules. *In Proceeding of the 22rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE'09), Tainan ,TAIWAN.*
- Meddouri, N., M. M. (2009b). Générer des règles de classification par dopage de concepts formels. *In Proceedings of the 9mes Journes Francophones en Extraction et Gestion des Connaissances (EGC'09), Strasbourg, FRANCE.*
- Njiwoua, P., M. N. P. (2005). Treillis de concepts et classification supervisée. *Revue des sciences et technologies de l'information.*, 449–488.
- Oosthuizen, G. D. (1988). The use of a lattice in knowledge processing. *PhD thesis, Glasgow, Scotland, UK.*
- Sahami, M. (1995). Learning classification rules using lattices (extended abstract). *In proceeding of the 8th European on Machine Learning [ECML'95], Heraclion, GREECE.*

Summary

To overcome the shortcomings of conventional techniques for data analysis, data mining is an effective tool to exploit and extract knowledge from data. Many free platforms of data mining provide various techniques and classification methods. These platforms note the absence of methods based on formal concept analysis. This paper presents the development and integration of supervised classification methods based on the use of a particular graph called concept lattice or Galois lattice as classifier in the WEKA platform.