

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