TANAGRA : un logiciel gratuit pour l'enseignement et la recherche

Ricco Rakotomalala

ERIC – Université Lumière Lyon 2
5, av Mendès France
69676 Bron
rakotoma@univ-lyon2.fr
http://eric.univ-lyon2.fr/~ricco

Résumé. TANAGRA est un logiciel « open source » librement accessible sur le web, il tente de concilier deux types d'utilisation. D'une part, en proposant une interface suffisamment conviviale, il est accessible aux utilisateurs non-spécialistes qui veulent effectuer des études sur des données réelles. D'autre part, en définissant une architecture simplifiée à l'extrême, les efforts de développement portent sur l'essentiel, à savoir la mise au point et l'intégration d'algorithmes de fouille de données, les chercheurs peuvent ainsi mener des expérimentations sur les méthodes. Dans cet article, nous présentons les principales fonctionnalités du logiciel en essayant de le positionner sur l'échiquier des (très) nombreux logiciels diffusés actuellement.

1 Introduction

TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche, diffusé sur internet (http://eric.univ-lyon2.fr/~ricco/tanagra). Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données. Sa principale originalité est qu'il tente de concilier une utilisation « néophyte » et « experte ».

Son premier objectif est d'offrir aux étudiants et aux experts d'autres domaines (médecine, bio-informatique, marketing, etc.) une plate-forme facile d'accès, respectant les standards des logiciels actuels, notamment en matière d'interface et de mode de fonctionnement, il doit être possible d'utiliser le logiciel pour mener des études sur des données réelles. Le second objectif est de proposer aux chercheurs une architecture leur facilitant l'implémentation des techniques qu'ils veulent étudier, de comparer les performances de ces algorithmes. TANAGRA se comporte alors plus comme une plate-forme d'expérimentation qui leur permettrait d'aller à l'essentiel en leur épargnant toute la partie ingrate de la programmation de ce type d'outil, notamment la gestion des données. Point très important à nos yeux, la disponibilité du code source est un gage de crédibilité scientifique, elle assure la reproductibilité des expérimentations publiées par d'autres chercheurs et, surtout, elle permet la comparaison et la vérification des implémentations.

TANAGRA n'intègre pas en revanche tout ce qui fait la puissance des outils commerciaux : multiplicité des sources de données, accès direct aux entrepôts de données et autres datamarts, interactivité des traitements avec des outils de visualisation sophistiqués. Ces outils, aussi séduisants et utiles soient-ils dans le cadre d'études sur des données réelles,