Le forage distribué des données : une méthode simple, rapide et efficace

Mohamed Aounallah et Guy Mineau

Département d'informatique et de génie logiciel
Pavillon Adrien-Pouliot, Université Laval
G1K 7P4, Canada
{Mohamed.Aoun-Allah, Guy.Mineau}@ift.ulaval.ca,
http://w3.ift.ulaval.ca/∼moaoa
http://www.ift.ulaval.ca/Personnel/prof/mineau.htm

Résumé. Dans cet article nous nous attaquons au problème du forage de très grandes bases de données distribuées. Le résultat visé est un modèle qui soit et prédictif et descriptif, appelé méta-classificateur. Pour ce faire, nous proposons de miner à distance chaque base de données indépendamment. Puis, il s'agit de regrouper les modèles produits (appelés classificateurs de base), sachant que chaque forage produira un modèle prédictif et descriptif, représenté pour nos besoins par un ensemble de règles de classification. Afin de guider l'assemblage de l'ensemble final de règles, qui sera l'union des ensembles individuels de règles, un coefficient de confiance est attribué à chaque règle de chaque ensemble. Ce coefficient, calculé par des moyens statistiques, représente la confiance que nous pouvons avoir dans chaque règle en fonction de sa couverture et de son taux d'erreur face à sa capacité d'être appliquée correctement sur de nouvelles données. Nous démontrons dans cet article que, grâce à ce coefficient de confiance, l'agrégation pure et simple de tous les classificateurs de base pour obtenir un agrégat de règles produit un méta-classificateur rapide et efficace par rapport aux techniques existantes.

1 Introduction

Ce papier traite du problème de forage de plusieurs bases de données gigantesques et géographiquement distribuées dans le but de produire un ensemble de règles de classification qui expliquent les groupements de données observés. Le résultat de ce forage sera donc un méta-classificateur aussi bien prédictif que descriptif. En d'autres termes, nous visons à produire un modèle qui permet non seulement de prédire la classe de nouveaux objets, mais qui permet aussi d'expliquer les choix de ses prédictions. Nous croyons que ce genre de modèles, basés sur des règles de classification, devrait aussi être facile à comprendre par des humains, ce qui est également l'un de nos objectifs. Il faut dire toutefois que nous nous plaçons dans le contexte où il est impossible de rapatrier toutes ces bases dans un même site, et ce, soit à cause du temps de téléchargement, soit à cause de l'impossibilité de traiter la base ainsi agrégée.

Dans la littérature, les techniques de forage distribué de données à la fois prédictives et descriptives sont malheureusement peu nombreuses. La plupart d'entre elles tentent de produire