Forage distribué des données : une comparaison entre l'agrégation d'échantillons et l'agrégation de règles

M. Aounallah*, S. Quirion*** et G. Mineau**

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

Résumé. Pour nous attaquer au problème du forage de très grandes bases de données distribuées, nous proposons d'étudier deux approches. La première est de télécharger seulement un échantillon de chaque base de données puis d'y effectuer le forage. La deuxième approche est de miner à distance chaque base de données indépendamment, puis de télécharger les modèles résultants, sous forme de règles de classification, dans un site central où l'agrégation de ces derniers est réalisée. Dans cet article, nous présentons une vue d'ensemble des techniques d'échantillonnage les plus communes. Nous présentons ensuite cette nouvelle technique de forage distribué des données où la mécanique d'agrégation est basée sur un coefficient de confiance attribué à chaque règle et sur de très petits échantillons de chaque base de données. Le coefficient de confiance d'une règle est calculé par des moyens statistiques en utilisant le théorème limite centrale. En conclusion, nous présentons une comparaison entre les meilleures techniques d'échantillonnage que nous avons trouvées dans la littérature, et notre approche de forage distribué des données (FDD) basée sur l'agrégation de modèles.

1 Introduction

Ce papier traite du problème de forage de plusieurs bases de données gigantesques et géographiquement distribuées, en présentant et en comparant deux techniques de forage de données. La première technique que nous avons examinée utilise un échantillon de taille raisonnable de chaque base de données, auxquels, une fois agrégés, nous appliquons une technique de forage de données. Cette technique relève de l'agrégation de données. Dans cette perspective, nous avons étudié les techniques d'échantillonnage existantes. Une description de ces dernières ainsi qu'une comparaison empirique sont présentées plus loin dans cet article.

La deuxième technique de forage de données, que nous introduisons (basée sur l'agrégation de modèles), se propose d'appliquer individuellement sur chaque base de