Utilisation des règles d'association pour la prédiction de valeurs manquantes

Tao-Yuan Jen*, Dominique Laurent*, Gorgoumack Sambe* **

*ETIS-CNRS, Université Cergy Pontoise,
F-95000 Cergy Pontoise
jen@u-cergy.fr, dlaurent@u-cergy.fr
**Université de Ziguinchor
BP : 523 Ziguinchor Sénégal
gsambe@univ-zig.sn

Résumé. Le traitement des valeurs manquantes est une problématique importante dans le domaine des entrepôts de données. Plusieurs solutions ont été proposées pour la prédiction de valeurs manquantes, présentant les caractéristiques suivantes : (i) la prédiction traite soit des valeurs continues soit des valeurs discrètes, et (ii) la prédiction est approximative (soit elle est associée à une probabilité soit elle concerne un ensemble de valeurs). Récemment, une méthode de prédiction permettant de traiter indépendamment les cas continu et discret a été proposée, en se basant sur les règles d'association. Cette méthode permet de prédire, avec une confiance *toujours égale à* 1, soit un ensemble de valeurs dans le cas discret, soit un intervalle de valeurs dans le cas continu.

Dans cet article, nous reprenons cette approche basée sur l'extraction de règles d'association et nous montrons comment générer des règles de prédictions portant sur une *unique valeur* et dont la confiance est toujours égale à 1. Afin d'obtenir de telles règles, notre méthode suppose que l'on dispose d'une hiérarchie décrivant des concepts généralisant les valeurs qui peuvent être prédites.

1 Introduction

Avec la mondialisation, la croissance et la compétition effrénée, les entreprises ont vu s'accroître le volume de leurs données. Ces données dispersées sur plusieurs sites de l'entreprise, sont regroupées et fédérées sur un seul support de données appelé *entrepôt de données*, à des fins de consultation et d'analyse.

La présence de *valeurs manquantes* dans les entrepôts de données est un problème crucial car les méthodes utilisées pour l'analyse de ces entrepôts produisent généralement des résultats erronés ou incomplets. Dans la mise en place d'un entrepôt de données, la phase de nettoyage des données est estimée entre 30 et 80% du temps de développement. Pour remédier au problème des valeurs manquantes, plusieurs solutions sont proposées (Kamber et Han (2005)) :

- ignorer les données comportant des valeurs manquantes,
- les remplacer manuellement,