Prédiction de valeurs manquantes dans les bases de données — Une première approche fondée sur la notion de proportion analogique

William Correa Beltran, Hélène Jaudoin, Olivier Pivert

Université de Rennes 1 – Irisa, Lannion, France
{William.Correa_Beltran@irisa.fr, Helene.Jaudoin@irisa.fr, Olivier.Pivert@irisa.fr}

Résumé. Cet article présente une méthode originale de prédiction de valeurs manquantes dans les bases de données relationnelles, fondée sur la notion de proportion analogique. Nous montrons en particulier comment un algorithme proposé dans le cadre de la classification automatique peut être adapté à cette fin. Deux cas sont considérés : celui d'une base de données transactionnelle (attributs booléens), et celui où les valeurs manquantes peuvent être de type numérique.

1 Introduction

Dans cet article, nous proposons une solution originale à un problème classique de bases de données qui consiste à prédire/estimer les valeurs manquantes dans une base de données relationnelle incomplète. De nombreuses approches ont été proposées pour traiter cette question, à la fois dans la communauté des bases de données et dans celle de l'apprentissage automatique, fondées sur des dépendances fonctionnelles (Atzeni et Morfuni (1986)), des règles d'association (Ragel (1998)), des règles de classification (Liu et al. (1997)), des techniques de clustering (Fujikawa et Ho (2002)), etc. Nous explorons quant à nous une nouvelle idée, issue de l'intelligence artificielle, qui consiste à exploiter les *proportions analogiques* (Prade et Richard (2012)) pouvant exister dans les données.

La suite de l'article est organisée comme suit. Dans la section 2, nous rappelons les notions de base concernant les proportions analogiques. La section 3 présente le principe général de l'approche que nous proposons pour estimer les valeurs manquantes, inspirée par la technique de classification proposée dans (Bayoudh et al. (2007); Miclet et al. (2008)). La section 4 est consacrée à une expérimentation visant à évaluer les performances de la méthode et à comparer cette dernière avec une technique classique d'estimation (kNN). Finalement, la section 5 rappelle les contributions principales de l'article et trace quelques perspectives de recherche.

2 Rappels sur les proportions analogiques

La présentation qui suit est tirée principalement de Miclet et Prade (2009). Une proportion analogique est une proposition de la forme « A est à B ce que C est à D », ce qui sera noté : $(A : B :: C : D)$. Dans la suite, les objets A, B, C, et D seront supposés être des n-uplets