Prédiction des valeurs manquantes dans un entrepôt de données par combinaison de la programmation par contraintes et des KPPV

Fatiha Amanzougarene*, Karine Zeitouni*
Mohamed Chachoua**

*PRISM, 45, Avenue des États Unis, 78000 Versailles
(fatiha.amanzougarene, karine.zeitouni)@prism.uvsq.fr
**EIVP, 80, rue Rébeval, 75019 ParisAutre adresse
chachoua@eivp-paris.fr

Résumé. La présence de données manquantes dans les grandes bases de données scientifiques et statistiques est très courante. Dans cet article, nous proposons un modèle permettant la reconstruction des données manquantes dans le contexte des entrepôts de données. Notre approche de reconstruction de données manquantes consiste à combiner la programmation par contraintes et une technique d'apprentissage automatique, à savoir l'algorithme des k-plus proches voisins. La programmation par contraintes permet d'inférer les données manquantes à partir des contraintes sommaires définies sur les données de base. La méthode des k-plus proches voisins permet d'augmenter la précision de cette inférence. Outre l'application aux entrepôts de données classiques, nous avons étendu notre approche à des entrepôts de données dits qualitatifs correspondant à notre cas d'étude, à savoir l'évaluation des gênes des chantiers urbains.

1 Introduction et motivation

Les entrepôts de données sont particulièrement exposés à des données manquantes ou des valeurs aberrantes en raison de la qualité des sources ou de diverses incohérences lors de l'intégration des données (Ribeiro et al., 2011). Si les valeurs manquantes ne sont pas traitées, des faits importants peuvent ne pas être pris en considération dans le processus d'analyse.

Pour illustrer la problématique et la solution apportée, nous considérons tout au long de cet article un exemple lié à une application sur les gênes [1] des chantiers urbains [2]. Dans cette application, le sujet d'analyse correspond à la gêne. Ce sujet est étudié selon les dimensions suivantes : le temps, l'espace, les nuisances [3] et les catégories de population (par exemple, adulte cadre en bonne santé, mère au foyer, enfant avec problème de respiration). Les mesures associées au fait gêne sont l'effectif pour chaque catégorie de populations, la durée d'exposition ainsi que le degré de gêne. Les deux premières mesures sont numériques tandis que la

1. On définit la gêne comme étant une relation subjective entre un individu et un phénomène nuisible
2. Ce travail rentre dans le cadre du projet ANR, dénommé Furtivité Urbaine Réseaux et Travaux (FURET)
3. On désigne par nuisance une perturbation du milieu ayant un effet négatif comme le bruit, les odeurs, etc.