Comparaisons structurelles de grandes bases de données par apprentissage non-supervisé

Guénaël Cabanes, Younès Bennani

LIPN-CNRS, UMR 7030,
99 Avenue J-B. Clément, 93430 Villetaneuse, France

Résumé. Dans le domaine de la fouille de données, mesurer les similitudes entre différents sous-ensembles est une question importante qui a été peu étudiée jusqu'à présent. Dans cet article, nous proposons une nouvelle méthode basée sur l'apprentissage non-supervisé. Les différents sous-ensembles à comparer sont caractérisés au moyen d'un modèle à base de prototypes. Ensuite, les différences entre les modèles sont détectées en utilisant une mesure de similarité.

1 Introduction

La croissance exponentielle des données engendre des volumétries de bases de données très importantes. Toutefois, la capacité à analyser les données reste insuffisante. Dans de nombreux cas, la capacité à mesurer des similitudes entre différents ensembles de données devient un élément important de l'analyse.

Les principaux enjeux pour l'étude de ce type de données sont, d'une part, l'obtention d'une description condensée des propriétés des données (Gehrke et al., 2001; Manku et Motwani, 2002), mais aussi la possibilité de détecter des variations ou des changements dans la structure des données (Cao et al., 2006; Aggarwal et Yu, 2007). Nous proposons dans cet article un algorithme capable de réaliser ces deux tâches [1]. Cet algorithme apprend d'abord une representation abstraite des sous-ensembles à comparer, puis évalue leur similarité en se basant sur cette représentation. La représentation abstraite est calculée par l'apprentissage d'une variante des SOM (Self-Organising Map, Kohonen (2001)), enrichie d'informations structurelles extraites des données. Nous proposons une méthode pour estimer, à partir de la SOM enrichie, une fonction de densité représentative des données. La mesure de dissimilarité entre sous-ensembles est alors une mesure de la divergence entre deux fonctions de densité.

L'avantage de cette méthode est la comparaison de structures par l'intermédiaire des modèles qui les décrivent, ce qui permet des comparaisons à n'importe quelle échelle sans surcharge de la mémoire de stockage. De plus, l'algorithme est très efficace en terme de temps d'exécution et de mémoire requise. Il est donc bien adapté pour la comparaison de grandes bases de données ou pour la détection de changement de structure d'un flux de données.

Le reste de cet article est organisé comme suit. La section 2 présente le nouvel algorithme. La section 3 décrit les tests de validation effectués et les résultats obtenus. Enfin, une conclusion est donnée dans la section 4.

[1] Ce travail a été soutenu en partie par le projet CADI (N°ANR - 07 TLOG 003), financé par l'ANR (Agence Nationale de la Recherche).