Une approche en programmation par contraintes pour la classification non supervisée

Thi-Bich-Hanh Dao, Khanh-Chuong Duong, Christel Vrain

LIFO, Université d'Orléans, 45067 Orléans cedex 02, France
{thi-bich-hanh.dao, khanh-chuong.duong, christel.vrain}@univ-orleans.fr

Résumé. Dans cet article, nous abordons le problème de classification non supervisée sous contraintes fondé sur la programmation par contraintes (PPC). Nous considérons comme critère d'optimisation la minimisation du diamètre maximal des clusters. Nous proposons un modèle pour cette tâche en PPC et nous montrons aussi l'importance des stratégies de recherche pour améliorer son efficacité. Notre modèle basé sur la distance entre les objets permet de traiter des données qualitatives et quantitatives. Des contraintes supplémentaires sur les clusters et les instances peuvent directement être ajoutées. Des expériences sur des ensembles de données classiques montrent l'intérêt de notre approche.

1 Introduction

La problématique de classification non supervisée (aussi appelée clustering) a été longuement étudiée pendant de nombreuses années avec des approches comme k-means et k-médoïdes. En général, le problème consiste à partitionner un ensemble de n objets en k classes non vides et deux à deux disjointes. C'est un champ de recherche difficile pour plusieurs raisons : le choix de la mesure de dissimilarité entre les objets dépendant principalement de l'application mais influant fortement sur les résultats, la définition du critère à optimiser, la taille de l'espace de recherche avec pour conséquence la nécessité de définir des heuristiques conduisant souvent à un optimum local. Poser des contraintes sur la solution recherchée permet d'une part, de modéliser plus finement les applications réelles et d'autre part de restreindre la taille de l'espace de recherche. Néanmoins, la plupart des algorithmes classiques n'ont pas été développés pour la classification non supervisée sous contraintes et doivent être adaptés, si possible, pour prendre en compte les contraintes posées par l'utilisateur. Développer des solveurs généraux applicables à une grande variété de problèmes pose de nouveaux défis.

D'autre part, des avancées récentes en Programmation par Contraintes (PPC) ont rendu ce paradigme beaucoup plus puissant. Plusieurs travaux (De Raedt et al. (2008, 2010)) (Boizumault et al. (2011)) ont étudié l'intérêt de la PPC pour modéliser des problèmes de fouille de données et ont montré l'apport de la déclarativité inhérente à la PPC.

Dans ce papier, nous proposons un cadre pour modéliser la classification non supervisée sous contraintes en Programmation par Contraintes. L'intérêt de notre approche est de fournir un modèle déclaratif permettant de spécifier le problème de classification non supervisée et d'intégrer facilement des contraintes. Dans notre modèle, nous faisons l'hypothèse que nous