Comparaison de critères de pureté pour l'intégration de connaissances en clustering semi-supervisé

Germain Forestier, Cédric Wemmert, Pierre Gançarski

Université de Strasbourg - LSIIT - CNRS - UMR 7005
Pôle API, Bd Sébastien Brant - 67412 Illkirch, France
{forestier,wemmert,gancarski}@unistra.fr

Résumé. L'utilisation de connaissances pour améliorer les processus de fouille de données a mobilisé un important effort de recherche ces dernières années. Il est cependant souvent difficile de formaliser ce type de connaissances, comme celles-ci sont souvent dépendantes du domaine. Dans cet article, nous nous intéressons à l'intégration de connaissances sous la forme d'objets étiquetés dans les algorithmes de clustering. Plusieurs critères permettant d'évaluer la pureté des clusters sont présentés et leur comportement est comparé sur des jeux de données artificiels. Les avantages et les inconvénients de chaque critère sont analysés pour aider l'utilisateur à faire un choix.

1 Introduction

L'intégration de connaissances dans les algorithmes de *clustering* a connu un fort intérêt ces dernières années, des connaissances dites du domaine (*background knowledge*) étant souvent disponibles. Celles-ci peuvent se présenter sous des formes très différentes et sont difficiles à formaliser de manière générique car elles dépendent souvent du domaine d'application. Plusieurs travaux (Wagstaff et al., 2001; Bilenko et al., 2004) se sont intéressés à l'utilisation de connaissances sous la forme de contraintes entre paires d'objets. À l'instar d'un algorithme supervisé qui va apprendre une fonction de classification, cette information peut être utilisée pendant le processus de clustering pour guider l'algorithme vers une solution en accord avec ces connaissances.

Un concept récurrent dans les méthodes utilisant ce type de connaissances est la pureté des clusters qui consiste à évaluer la qualité des clusters par rapport à ces objets étiquetés. Un cluster pur sera un cluster dans lequel tous les objets, dont la classe est connue, appartiennent à une et une seule classe. Un cluster impur présentera des objets de classes différentes.

L'objectif de cet article est de présenter et de comparer différentes méthodes d'évaluation de la pureté des clusters. Dans la section 2, nous présentons un état de l'art de l'utilisation de connaissances en clustering. Dans la section 3, différentes mesures de pureté sont présentées et comparées. Enfin, la section 4 présente les conclusions et les futures pistes de recherche de ces travaux.