Post-traitement pour la classification probabiliste non supervisée sous contraintes

Nguyen-Viet-Dung Nghiem *, Christel Vrain*
Thi-Bich-Hanh Dao* Ian Davidson**

*Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, F-45067, Orléans, France
prenom.nom@univ-orleans.fr
**Department of Computer Science, University of California, Davis, USA
davidson@cs.ucdavis.edu

Résumé. Le clustering sous contraintes permet d'intégrer des connaissances expertes, que ce soit des contraintes classiques must-link ou cannot-link ou des contraintes plus complexes. La plupart des algorithmes de clustering probabilistes intègrent les contraintes en ajoutant un terme dans la fonction objectif représentant leur satisfaction. Ils intègrent donc difficilement différents types de contraintes et ne garantissent pas la satisfaction de toutes les contraintes. Nous proposons une méthode qui, à partir du résultat d'un algorithme de clustering sous la forme d'une matrice de probabilité d'affectation de points aux clusters, trouve la meilleure affectation qui satisfasse toutes les contraintes. Cette méthode peut s'appliquer à tout algorithme probabiliste y compris ceux utilisant l'apprentissage profond. Les expérimentations montrent que notre méthode est compétitive avec des méthodes parmi les plus récentes.

1 Introduction

Le clustering semi-supervisé est un domaine déjà longuement étudié. Généralement, il consiste à introduire des connaissances initiales sur les étiquettes des points sous la forme de contraintes *must-link* ou *cannot-link* entre paires de points spécifiant si ces points doivent apparaître ou non dans le même cluster (Wagstaff et Cardie, 2000; Wagstaff et al., 2001; Bilenko et al., 2004; Davidson et Ravi, 2005; Wang et Davidson, 2010). Intégrer des connaissances plus générales dans un processus de clustering nécessite souvent d'adapter les algorithmes existants, conduisant à des systèmes dédiés à certains types de contraintes. Récemment il a été montré que des formalismes déclaratifs comme la Programmation Linéaire en Nombres Entiers (PLNE) ou la Programmation par Contraintes permettaient d'intégrer facilement des connaissances de formes variées, cependant au prix de la complexité (Davidson et al., 2010; Babaki et al., 2014; Ouali et al., 2016; Dao et al., 2017). De plus, comme la plupart des approches en clustering, elles reposent sur une distance calculée dans l'espace des entrées, et l'on sait que l'utilisation d'une distance en grande dimension pose de nombreux problèmes.

Ces dernières années, les avancées en apprentissage profond permettent de transformer les données d'entrée dans des espaces de dimension plus faible grâce à des plongements non linéaires. Cela a conduit au développement du *deep clustering* (Song et al., 2013; Xie et al.,