Clustering par apprentissage de distance guidé par des préférences sur les attributs

Adnan El Moussawi^{*,**}, Ahmed Cheriat^{**},
Arnaud Giacometti^{*}, Nicolas Labroche^{*}, Arnaud Soulet^{*}

^{*}Université François Rabelais de Tours, 3 place Jean Jaures, 41000 Blois
{Prénom}.{Nom}@univ-tours.fr,

^{**}Groupe Kalidea, 47 rue de l'Est, 92774 Boulogne-Billancourt Cedex
{aelmoussawi, acheriat}@kalidea.com

Résumé. Ces dernières années de nombreuses méthodes semi-supervisées de clustering ont intégré des contraintes entre paires d'objets ou d'étiquettes de classe, afin que le partitionnement final soit en accord avec les besoins de l'utilisateur. Pourtant dans certains cas où les dimensions d'études sont clairement définies, il semble opportun de pouvoir directement exprimer des contraintes sur les attributs pour explorer des données. De plus, une telle formulation permettrait d'éviter les écueils classiques de la malédiction de la dimensionnalité et de l'interprétation des clusters. Cet article propose de prendre en compte les préférences de l'utilisateur sur les attributs afin de guider l'apprentissage de la distance pendant le clustering. Plus précisément, nous montrons comment paramétrer la distance euclidienne par une matrice diagonale dont les coefficients doivent être au plus proche des poids fixés par l'utilisateur. Cette approche permet d'ajuster le clustering pour obtenir un compromis entre les approches guidées par les données et par l'utilisateur. Nous observons que l'ajout des préférences est parfois essentiel pour atteindre un clustering de meilleure qualité.

1 Introduction

Les méthodes de segmentation ou de clustering sont particulièrement adaptées à la recherche des différents profils des clients, par exemple pour leur proposer des programmes spécifiques de fidélisation. Néanmoins, des cas d'usages au sein de l'entreprise Kalidea¹ démontrent que de nombreux problèmes restent posés. Tout d'abord, par croisement des différentes mesures (chiffre d'affaire, quantité de produits vendus ou invendus, etc) et dimensions d'analyse possibles (temporelles, géographiques, gammes de produits, etc), les attributs candidats pour segmenter les données sont extrêmement nombreux. Ainsi, il s'avère rapidement illusoire d'effectuer une segmentation sur l'ensemble de tous les attributs d'analyse possibles, que ce soit du fait de la malédiction de la dimensionnalité ou de la difficulté à interpréter et analyser les clusters extraits. Par ailleurs, on constate souvent que les clusters produits ne sont pas toujours en accord avec les besoins ou intuitions des utilisateurs des outils de segmentation.

1. <http://www.kalidea-ce.com/>