

Une méthode supervisée pour initialiser les centres des K-moyennes

Oumaima Alaoui Ismaili^{*,**}, Vincent Lemaire^{*}, Antoine Cornuèjols^{**}

^{*}Orange Labs, AV. Pierre Marzin 22307 Lannion cedex France
(oumaima.alaouiismaili, vincent.lemaire)@orange.com

^{**}AgroParisTech 16, rue Claude Bernard 75005 Paris
antoine.cornuejols@agroparistech.fr

Résumé. Au cours des dernières années, la classification à base de clustering s’est imposée comme un sujet de recherche important. Cette approche vise à décrire et à prédire un concept cible d’une manière simultanée. Partant du fait que le choix des centres pour l’algorithme des K-moyennes standard a un impact direct sur la qualité des résultats obtenus, cet article vise alors à tester à quel point une méthode d’initialisation supervisée pourrait aider l’algorithme des K-moyennes standard à remplir la tâche de la classification à base des K-moyennes.

1 Introduction

Depuis quelques années, l’étude d’un nouvel aspect de l’apprentissage connu sous le nom de la *classification à base de clustering* (ou *Supervised clustering en anglais*) a suscité beaucoup d’intérêt (e.g., (Eick et al., 2004) et (Cevikalp et al., 2007)). Les approches appartenant à ce type d’apprentissage visent à décrire et à prédire d’une manière simultanée (Alaoui Ismaili et al., 2015a). Dans ce cadre d’étude, on suppose que la classification à base de clustering est étroitement liée à l’estimation de la distribution des données conditionnellement à une variable cible. A partir d’une base de données étiquetée, ces approches cherchent à découvrir la structure interne de la variable cible afin de pouvoir prédire ultérieurement la classe des nouvelles instances.

La figure 1 illustre la différence entre les trois types d’apprentissage : le clustering standard (a), la classification supervisée (b) et la classification à base de clustering (c). Dans la classification supervisée, la compacité des classes apprises (dans la phase d’apprentissage) n’est pas une condition importante (e.g, le groupe B2 de (b)). Le clustering regroupe les instances homogènes sans tenir compte de leur étiquetage (e.g, le groupe A4 de la a)). La classification à base de clustering vise à former des groupes compacts et purs en termes de classes (e.g, les 6 groupes de (c)).

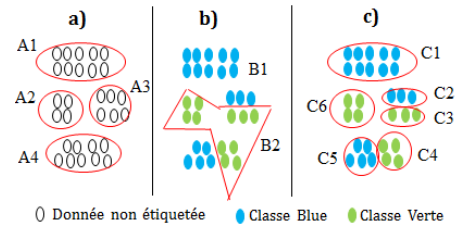


FIG. 1: Types d’apprentissage