Classification de Données Complexes par Globalisation de Mesures de Similarité via les Moyennes Quasi-Arithmétiques

Étienne-Cuvelier*, Marie-Aude-Aufaure**

*ICHEC - Brussels Management School
etienne.cuvelier@ichec.be

**Datarvest
marie-aude.aufaure@datarvest.com

Résumé. La plupart des méthodes de classification sont conçues pour des types particuliers de données: données numériques, textuelles, catégoriques, fonctionnelles, probabilistes ou encore de type graphes. Cependant, les données générées dans notre quotidien sont en général composées de données de types mixtes. Par exemple, si nous considérons la prévention cardiaque dans le domaine de la santé, les applications vont combiner des données issues de capteurs avec d'autres données telles que l'âge, le niveau d'effort, la fréquence cardiaque maximale, des histogrammes de fréquences cardiaques moyennes lors de précédents efforts, etc. Ceci nous amène à la problématique de construire des classes en tenant compte de ces différentes données, et de définir une mesure de similarité à partir des similarités de paires d'objets sur les différents types de variables. Dans cet article nous proposons une méthode de classification basée sur la fusion des matrices de similarité à l'aide des moyennes quasi-arithmétiques qui permet de choisir les différentes "dimensions" des données à considérer, et ce quel que soit le type de données, pour autant qu'une mesure, de similarité ou de dissimilarité existe pour chacun des types de données, ce qui est très souvent le cas.

1 Introduction

A l'ère des Big Data, le volume et la diversité des données disponibles de manière numérique ne cesse de croître. Ces données proviennent de capteurs, de réseaux sociaux, du web, des traces laissées par nos appareils mobiles, de nos achats, des données ouvertes, etc. Cette diversité est d'une grande richesse et est valorisée par les entreprises à travers des applications toujours plus personnalisées. Si les architectures informatiques passent bien à l'échelle, avec notamment du stockage dans de nouveaux outils de type NoSQL, l'utilisation des technologies distribuées avec Hadoop et son écosystème ou encore le traitement en temps contraint avec Spark et les bibliothèques associées, le croisement de toutes ces données restent encore un défi à l'heure actuelle. De nombreuses méthodes de classification ont été conçues pour un type de données particulier. Or, il peut être utile dans de nombreuses applications, de pouvoir résumer et synthétiser un ensemble de données de types divers provenant de sources hétérogènes. Par