# Sélection de mesures de similarité pour la classification de données catégorielles

Guilherme Alves, Miguel Couceiro, Amedeo Napoli

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy
{guilherme.alves-da-silva,miguel.couceiro,amedeo.napoli}@loria.fr

**Résumé.** Le partitionnement de données est une opération très utilisée dans l'exploration et l'analyse de données, en particulier pour traiter des tableaux de données qui comprennent des attributs catégoriels. Une telle opération repose sur des mesures de similarité, qui sont proposées en nombre dans la littérature. Cependant, le choix d'une mesure est complexe et dépend du contexte et des données en cours d'étude. Dans cet article, nous cherchons à caractériser de façon automatique la "meilleure" mesure de similarité pour partitionner un jeu de données particulier. Nous présentons les bases de notre approche et une étude empirique qui porte sur des données catégorielles ainsi qu'une évaluation de cette approche.

## 1 Introduction

Beaucoup d'opérations du monde réel engendrent de très grandes masses de données, relatives par exemple à la consommation, la santé, le temps, l'espace, les réseaux sociaux ... Ces données sont généralement hétérogènes, signifiant entre autres qu'elles sont décrites par des attributs de types différents, numériques ou catégoriels (Šulc et Řezanková, 2019).

Diverses méthodes de fouille de données permettent de tenir compte de cette hétérogénéité. Ainsi, les méthodes de partitionnement [1] sont souvent employées pour analyser de telles données, découvrir des classes ou des profils, ou résumer et organiser des données (Barioni et al., 2014). Un processus de partitionnement repose sur des mesures de similarité, entre les objets et entre les classes, qui jouent un rôle fondamental. Il existe de nombreuses mesures pour les données catégorielles (Boriah et al., 2008), sans qu'il n'en existe une qui soit universelle et meilleure qu'une autre dans toutes circonstances. De fait, le choix d'une mesure de similarité dépend le plus souvent des caractéristiques des données et du contexte de l'étude.

Plusieurs approches ont été proposées pour automatiser le choix d'un algorithme de partitionnement (Pimentel et de Carvalho, 2019; Abdulrahman et al., 2018). Dans ce cadre, le "méta-apprentissage" ("meta-learning") peut guider la construction de classifieurs en fonction des caractéristiques des données et de celles des algorithmes (Brazdil et al., 2008). Par exemple, il est possible, à l'image du raisonnement à partir de cas, de s'appuyer sur une base d'épisodes de résolution de problèmes et d'adapter un des épisodes (le plus proche) au traitement des données courantes. Dans notre cas, c'est plutôt le choix d'une mesure de similarité

1. Nous parlons ici de "classification non supervisée" ou de "clustering".