# Un critère d'évaluation pour les K-moyennes prédictives

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.** L'algorithme des K-moyennes prédictives est un des algorithmes de clustering prédictif visant à décrire et à prédire d'une manière simultanée. Contrairement à la classification supervisée et au clustering traditionnel, la performance de ce type d'algorithme est étroitement liée à sa capacité à réaliser un bon compromis entre la description et la prédiction. Or, à notre connaissance, il n'existe pas dans la littérature un critère analytique permettant de mesurer ce compromis. Cet article a pour objectif de proposer une version modifiée de l'indice Davies-Bouldin, nommée SDB, permettant ainsi d'évaluer la qualité des résultats issus de l'algorithme des K-moyennes prédictives. Cette modification se base sur l'intégration d'une nouvelle mesure de dissimilarité permettant d'établir une relation entre la proximité des observations en termes de distance et leur classe d'appartenance. Les résultats expérimentaux montrent que la version modifiée de l'indice DB parvient à mesurer la qualité des résultats issus de l'algorithme des K-moyennes prédictives.

## 1 Introduction

L'algorithme des K-moyennes prédictives (Ismaili et al., 2015, 2016; Dimitrovski et al., 2014) est une version modifiée de l'algorithme des K-moyennes traditionnel (MacQueen, 1967). L'objectif de ce type d'algorithme est de *décrire et de prédire simultanément*. Contrairement à l'algorithme des K-moyennes traditionnel, l'algorithme des K-moyennes prédictives cherche à discerner à partir d'une base de données étiquetées, des groupes d'instances compacts, éloignés les uns des autres et purs en termes de classe dans le but de prédire ultérieurement la classe des nouvelles instances (voir la figure 1).

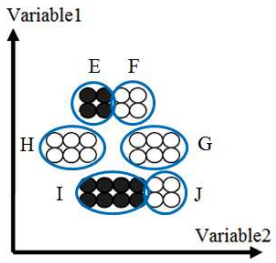


FIG. 1: Objectif du clustering prédictif

Pour mesurer la qualité des résultats issus de l'algorithme des K-moyennes prédictives, trois points doivent être pris en considération : $i$) le taux de bonnes prédiction, $ii$) la compacité et $iii$) la séparabilité des clusters. Il s'agit ici de réaliser un compromis entre la prédiction et la description.