# Mesure de qualité de la classification à base de clustering

Oumaima Alaoui Ismaili*,**, Vincent Lemaire*, Antoine Cornuéjols**

*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.** La classification à base de clustering est une technique d'apprentissage hybride. Elle vise à décrire et à prédire simultanément. Pour ce genre de technique, la difficulté réside essentiellement dans le choix du critère capable de mesurer la qualité des résultats obtenus.

## 1 Introduction

Depuis quelques années, on assiste à une augmentation significative du volume de données. Pour pouvoir en extraire les connaissances utiles, de nombreuses approches d'apprentissage ont été développées. Récemment, une nouvelle technique d'apprentissage hybride a été proposée. Elle inclut à la fois le principe de la classification supervisée (Cornuéjols et Miclet (2010)) (Figure 1b)) et celui du clustering (Figure 1a). Cette technique est appelée «*classification à base de clustering*» (Eick et al. (2004)). Elle a comme objectif de décrire et prédire d'une manière simultanée (Figure 1c)). En effet, elle permet de prédire la classe des nouvelles instances tout en se basant sur la structure interne de la variable cible découverte lors de la phase d'apprentissage. Chaque groupe construit durant cette phase est alors constitué d'une collection d'instances similaires et de même classes. Pour plus de détail sur ce type d'apprentissage voir par exemple (Alaoui Ismaili et al. (2015))

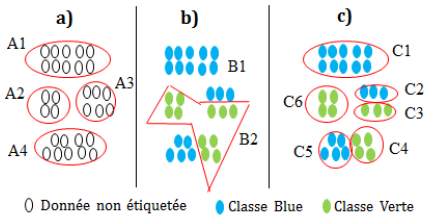


FIG. 1 – *La différence entre les différentes types de classification :* **a)** *clustering,* **b)** *classification supervisée et* **c)** *classification à base de clustering.*

L'évaluation de la qualité de la partition générée par un modèle d'apprentissage (que ce soit supervisé ou non supervisé) est une étape primordiale. Dans le cadre de la classification à base de clustering, on définit une bonne partition comme étant la partition qui permet de

réaliser un bon compromis entre l'homogénéité et la pureté des clusters en termes de classes (Figure 1c)). A notre connaissance, jusqu'à présent, il n'existe pas dans la littérature un critère global (associant à la fois une partie "interne" pour évaluer la description des groupes et une partie "externe" pour évaluer la prédiction) qui permet d'évaluer la qualité des résultats issus d'un algorithme de classification à base de clustering.

Notre objectif est donc de chercher un critère d'évaluation qui permet de mesurer le compromis homogénéité/pureté. Ce résumé est a pour but donc de présenter notre problématique afin de pouvoir en discuter lors de la journée du clustering.

## 2 Mesure de qualité de la classification à base de clustering

Le critère d'évaluation recherché est un critère global permettant d'identifier les partitions ayant : $1)$ une forte similarité intra-cluster, $2)$ une faible similarité inter-cluster et $3)$ un taux de bonne classification élevé.

A titre d'exemple, dans la figure 1c), on aimerait s'apercevoir que les clusters C3 et C6 sont certes de même classe mais relativement distants. De plus, les clusters C4 et C5 sont homogènes mais de classe différentes. En respectant ce principe, la partition ayant 6 clusters est celle qui permette de mieux décrire la structure interne de la variable cible.

Un bon critère d'évaluation dans le cadre de la classification à base de clustering est définit comme étant le critère qui respecte les points suivants :

- Facile à interpréter
- Faiblement biaisé par le nombre de clusters, le nombre d'individus et le nombre d'individus dans chaque cluster
- Relativement stable en cas de perturbations aléatoires.
- Avoir une complexité acceptable, vis-à-vis à celle de l'algorithme de clustering

## Références

Alaoui Ismaili, O., V. Lemaire, et A. Cornuéjols (2015). Classification à base de clustering ou comment décrire et prédire simultanément. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)*.

Cornuéjols, A. et L. Miclet (2010). *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

Eick, C. F., N. Zeidat, et Z. Zhao (2004). Supervised clustering-algorithms and benefits. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pp. 774–776. IEEE.

## Summary

Clustering-based classification is a hybrid learning approach allowing to describe and predict simultaneously. Nevertheless, choosing the suitable criterion to evaluate its obtained results is a complex task.