

# Apprentissage non-supervisé relationnel dans l'espace des coordonnées barycentriques

Parisa Rastin, Basarab Matei, Guénaél Cabanes

LIPN-CNRS, UMR 7030, Université Paris 13  
rastin@lipn.univ-paris13.fr

**Résumé.** Les approches basées sur les prototypes sont très populaires en apprentissage non supervisé, en raison de la compacité du modèle résultant (les prototypes), de la puissance descriptive de ces prototypes et de la faible complexité de calcul du modèle. Habituellement, le meilleur choix de prototype est le barycentre du cluster. Le prototype est alors défini comme l'objet minimisant la somme des distances carrées avec tous les objets du cluster. Cependant, dans de nombreux cas, les objets ne peuvent pas être facilement définis dans un espace euclidien sans perte d'information et/ou un pré-traitement coûteux, ce qui limite la construction des prototypes. Dans cet article, nous proposons une approche de K-moyennes relationnelle utilisant un ensemble unique de points de support basé sur le formalisme des coordonnées barycentriques, afin d'unifier la représentation des objets et des prototypes et permettant un processus d'apprentissage incrémental simple pour le clustering relationnel.

## 1 Introduction

L'apprentissage non supervisé (ou clustering) permet de calculer un modèle de la structure de données lorsqu'aucune autre information n'est connue. Les objets sont regroupés en "clusters", en fonction de leur similarité. Cette classification est une représentation compacte de la distribution sous-jacente des données. De nombreux algorithmes ont été proposés, qui peuvent être classés en plusieurs familles (Bishop et al., 1998). Dans cet article, nous nous intéressons à la famille des algorithmes à base de prototypes, dans laquelle chaque cluster est représenté par un prototype : un nouvel objet dans l'espace de représentation. Les approches basées sur des prototypes sont très populaires en raison de la compacité du modèle obtenu (les prototypes), du pouvoir descriptif de ces prototypes et de la faible complexité de calcul du modèle (chaque objet est comparé à un ensemble de prototypes généralement réduit). Cette faible complexité explique à elle seule la popularité des approches à base de prototypes pour des applications réelles. Habituellement, le meilleur choix de prototype est le barycentre du cluster. Le prototype est ainsi défini comme étant l'objet minimisant la somme des distances carrées avec tous les objets du cluster. Si les objets sont décrits comme des vecteurs numériques dans un espace euclidien, la définition des prototypes du cluster est simple. Dans ce cas, un prototype est un vecteur défini dans le même espace, calculé comme la moyenne vectorielle des objets appartenant à son cluster. En fait, la plupart des algorithmes basés sur des prototypes ne sont