Identification de sous espaces caractéristiques des classes issues de K-means parcimonieuse

Abdoul Wahab Diallo & Mory Ouattara & Ndèye Niang

AIMS, Mbour, Sénégal
abdoul.w.m.diallo@aims-senegal.org

CEDRIC CNAM Paris, France
ndeye.niang_keita@cnam.fr

Université de San pedro, Côte d'ivoire
ouattara.mory@usp.edu.ci

September 15, 2022

Plan

Introduction

La complexité croissante des données a conduit à de nombreuses avancées méthodologiques et algorithmiques dans le domaine du clustering.

Bi-clustering or co-clustering

Kriegel and al (2009)

Clustering Ensemble Multi-block Clustering

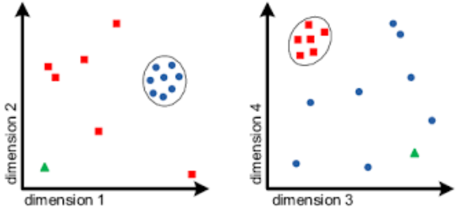
Strehl (2002), Vega-Pons (2010)

Subspace Clustering

Parsons and al (2004)

Subspace Clustering

- Identifier simultanément
  - Les groupes d'individus homogènes
  - Les sous-groupes de variables pertinentes pour l'explication de chaque classe



Source: researchgate.net/Thomas-Seidl/publication

- Hard Soft subspace CLIQUE(Agrawal and al)(1998)
- Soft subspace clustering EWKM (Jing and al) (2007).
- En grande dimension, l'interprétation des poids peut être fastidieuse.

Sparse K-means

- Les auteurs maximisent l'inertie interclasse pondérée avec des contraintes sur les poids.

$$\max_{C_1,\dots,C_k,w} \left\{ \sum_{j=1}^{p} w_j \Big(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_i, x_i' \in C_k} d(x_i^j, x_{i'}^j) \Big) \right\}$$

où $w \in \mathbb{R}^p$ sous les contraintes : $||w||_2 \leq 1$, $||w||_1 \leq s$, $w_j \geq 0$, $\forall j$.

Le paramètre s est fixé soit :

- En suivant l'évolution de la fonction objectif
- En utilisant un critère tel que la Gap statistique

[D.M. Witten and R.Tibshirani, 2010]

Optimisation des poids

- Les poids des variables sont déterminés à l'aide du processus suivant :

$$w = \frac{S(a_+, \Delta)}{||S(a_+, \Delta)||_2}$$

où

$$a_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{x_i, x_i' \in C_k} d(x_i^j, x_{i'}^j)$$

- et $\Delta = 0$ si $||w||_1 < s$;
- sinon choisir $\Delta > 0$ si $||w||_1 = s$.
- S est un opérateur défini par $S(x, c) = sign(x)(|x| - c)_+$.

Sparse Subspace K-means

- SSKM est une extension de la Sparse K-means pour déterminer le sous-espace caractéristique des classes.
- Le critère associé à l'algorithme SSKM est le suivant :

$$\max_{C_1,\ldots,C_k,w}\left\{\sum_{j=1}^{p} w_j\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d(x_i^j,x_{i'}^j)-\sum_{k=1}^{K}\frac{1}{n_k}\sum_{x_i,x_i'\in C_k} d(x_i^j,x_{i'}^j)\right)\right\}$$

$$\max_{C_1,\ldots,C_K,w^k}\left\{\sum_{j=1}^{p}\sum_{k=1}^{K} w_j^k\left(\frac{1}{nK}\sum_{i=1}^{n}\sum_{i'=1}^{n} d(x_i^j,x_{i'}^j)-\frac{1}{n_k}\sum_{x_i,x_{i'}\in c_k} d(x_i^j,x_{i'}^j)\right)\right\}$$

sous les contraintes : $\| w^k \|_2 \leq 1$, $\| w^k \|_1 \leq s$ et $w_j^k \geq 0, \forall j = 1, \ldots, p \; et \; k = 1, \ldots, K$

Optimisation des poids

- L'optimisation des poids ce fait en utilisant la formule suivante:

$$w^k = \frac{S((a^k)_+,\Delta)}{||S((a^k)_+,\Delta)||_2}$$

où

$$a_j^k = \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d(x_i^j, x_{i'}^j) - \frac{1}{n_k}\sum_{x_i, x_{i'} \in c_k} d(x_i^j, x_{i'}^j)$$

- avec $\Delta = 0$ si $||w^k||_1 < s$;
- sinon choisir $\Delta > 0$ si $||w^k||_1 = s$.

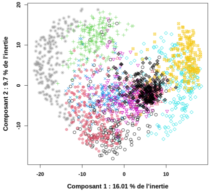
Les données

Données	n	p	k
X_1	200	60	4
X_2	200	60	4
DMU	2000	649	10
IS	2310	18	7

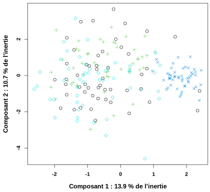
Table: Les données

- critères d'évaluation :
    - **Normalized Mutual Information** (NMI)
    - **Adjusted Rand Index** (ARI)

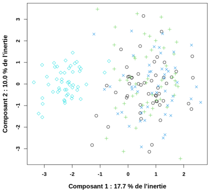
Data



(a) Projection of DMU clusters



(b) Projection of X_2 clusters on Bloc 3



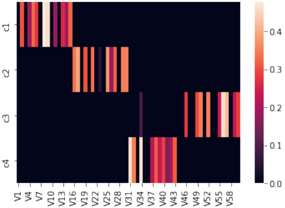
(c) Projection of X_2 clusters on Bloc 4

Comparaison des performances

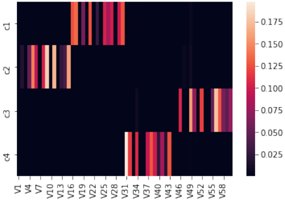
Données	Indices	K-means	EWKM	Sparse K-means	SSKM
X_1	NMI	0.95 (0.1)	0.52 (0.1)	**1** (0)	**1** (0)
	ARI	0.93 (0.13)	0.45 (0.21)	**1** (0)	**1** (0)
X_2	NMI	0.76 (0.08)	0.47 (0.17)	0.89 (0)	**0.92** (0.02)
	ARI	0.77 (0.12)	0.45 (0.17)	0.92 (0)	**0.92** (0.03)
DMU	NMI	0.73 (0.04)	0.49 (0.05)	0.78 (0)	**0.81** (0.02)
	ARI	0.63 (0.06)	0.38 (0.07)	0.71 (0)	**0.75** (0.05)
IS	NMI	**0.57** (0.02)	0.47 (0.08)	0.56 (0)	**0.57** (0.01)
	ARI	**0.47** (0.03)	0.34 (0.09)	0.46 (0)	**0.46** (0.01)

Table: Performances des méthodes : K-means, EWKM, Sparse K-means et SSKM sur X_1, X_2, DMU et IS

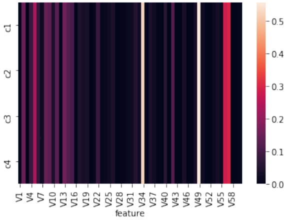
Evaluation des variables pertinentes pour la table X_2



(a) SSKM

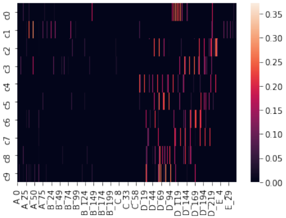


(b) EWKM

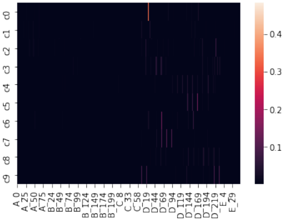


(c) SparseKM

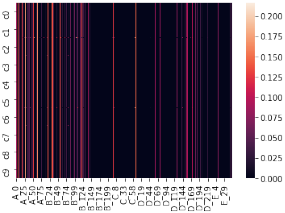
Évaluation des variables pertinentes pour la table DMU



(a) SSKM



(b) EWKM



(c) SparseKM

Conclusion & Perspectives

- Sur les données utilisées, la méthode SSKM fournit de bonnes performances.
  - Partition performantes au sens des indices utilisés.
  - Identification des sous espaces .

- La méthode SSKM facilite la description des classes à travers l'utilisation des poids

- Nous envisageons d'étendre la définition des poids aux blocs lorsque les données sont structurées en blocs comme dans FGKM, Chen and al, 2012.

Merci pour votre attention !!!

Réferences

- Daniela M. Witten and Robert Tibshirani. *A framework for feature selection in clustering*. Journal of the American Statistical Association, 713–726, 2010.
- X Chen and Y Ye and X Xu and J. Z Huang. *A feature group weighting method for subspace clustering of high-dimensional data*. Pattern Recognition, 434-446, 2012.
- L. Jing and M. Ng and J. Huang. *An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data*. Knowledge and Data Engineering, IEEE Transactions, 1026-1041, 2007.
- R. Agrawal and J. Gehrk and D. Gunopulos, and P. Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data. ACM Press, 94–105, 1998.