Optimization of Fuzzy C-Means with Alternating Direction Method of Multipliers

Benoit Albert, Violaine Antoine, and Jonas Koko

LIMOS, Université Clermont Auvergne, France
{benoit.albert,violaine.antoine,jonas.koko}@uca.fr

1 Informations

In this file you will find the codes associated with the work presented in French at EGC 2023[1] and OLA 2023.

The idea of this work is to use a more rebosted optimization method than the simple alternating optimization (file: FCM_AO). Optimization of fuzzy c-means by the alternating directions method (file: FCM_ADMM). An important note in these functions, the data matrix dimensions are : number of attributes by number of objects!
The main file allows to compare these two methods according to the data sets (file: Data) (see next section).
Finally, the folder DisplayIndex gathers the files necessary to display the partitioning and to evaluate it by the ARI score.

2 Datasets

I used a simpler normalisation than that used for the articles to make the results more reproducible.
We used 11 data sets. The first five corresponds to real data from the UCI library [1] : IRIS, WINE, SEEDS, WDBC, and DRYBEAN. We also used six synthetic data sets [2] : A1, A3, DIM32, DIM64, S1, and S3. We have referenced Table 1 their characteristics, i.e. the number of classes c, objects n and attributes p, as well as the optimal penalty parameter r^*, and by default r_d.
The data has been normalised between 0 and 1. Let x be the data variable, let $x(:,l)$ be the column l, this is the data set restricted to the lth attribute (p attributs).

$$x_n(:,l) = \frac{x - mean(x(:,l))}{max(x(:,l)) - mean(x(:,l))}, \quad l \in [1,p]. \tag{1}$$

The files storing the data have 4 variables: the name of the dataset, the normalised data, the labels and the number of objects in each cluster.

[1] https://archive.ics.uci.edu/ml/datasets.php
[2] https://cs.joensuu.fi/sipu/datasets/

Table 1: Characteristics of data sets.

	IRIS	WINE	SEEDS	WDBC	DRYBEAN	A1	A3	DIM32	DIM64	S1	S3
c	3	3	3	2	7	20	50	16	16	15	15
n	150	178	210	569	13611	3000	7500	1024	1024	5000	5000
p	4	13	7	30	16	2	2	32	64	2	2
r^*	30	30	25	5000	5.10^4	10	10	500	500	10	10
r_d	7200	27768	17640	136560	6.097.728	$4, 8.10^4$	3.10^6	2^{21}	2^{22}	6.10^5	6.10^5

For insensitivity of the results to the initialization for every algoithms, we first ran ADMM with the Euclidean distance ($\mathbf{ADMM}_{eu}$) with $r = 2.5$ and set a maximum number of iterations to 50 starting with random $\boldsymbol{U}^0$. The following results were obtained with an inner loop for ADMM equal to 5.

Table 2: ARI score (UCI).

	IRIS	WINE	SEEDS	WDBC	DRYBEAN
FCM-GK	0.76	0.36	0.72	0.41	**0.70**
ADMM_{r^*}	**0.78**	0.80	**0.73**	**0.76**	0.32
ADMM_{r_d}	0.72	**0.94**	**0.73**	**0.76**	0.32

Table 3: ARI score (Synthetic data).

	A1	A3	DIM032	DIM064	S1	S3
FCM-GK	**0.90**	**0.90**	0.40	0.108	**0.99**	**0.66**
ADMM_{r^*}	0.23	0.14	**0.55**	**0.61**	0.31	0.24
ADMM_{r_d}	0.24	0.14	**0.55**	**0.68**	0.31	0.27

References

1. Albert, B., Antoine, V., Koko, J.: Optimisation de fuzzy c-means (fcm) clustering par la méthode des directions alternées (admm). In: Extraction et Gestion des Connaissances: Actes de la conférence EGC'2023. vol. 39. BoD-Books on Demand (2023)