K-means seeding via MUS algorithm

Inizializzazione del K-means tramite l'algoritmo MUS

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Abstract K-means algorithm is one of the most popular procedures in data clustering. Despite its large use, one major criticism is the impact of the initial seeding on the final solution. We propose a modification of the K-means algorithm, based on a suitable choice of the initial centers. Similarly to clustering ensemble methods, our approach takes advantage of the information contained in a co-association matrix. Such matrix is given as input for the MUS algorithm that allows to define a pivot-based initialization step. Preliminary results concerning the comparison with the classical approach are discussed.

Abstract *L'algoritmo K-medie è una delle procedure di raggruppamento più utilizzate. Tuttavia, una delle maggiori criticità di tale metodo riguarda l'impatto della scelta dei semi iniziali sulla configurazione finale. In questo lavoro viene proposta una variante del K-medie basata su una scelta opportuna dei semi iniziali. In linea con i cosiddetti 'metodi di insieme', l'approccio considerato sfrutta l'informazione contenuta in una matrice di co-associazione. Tale matrice viene utilizzata dall'algoritmo MUS per definire i semi iniziali dei gruppi sulla base di unità pivotali. Vengono discussi alcuni risultati preliminari riguardanti il confronto con l'approccio classico.*

Key words: Clustering, pivotal unit, seeding

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: leoegidi@hotmail.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it,

1 Introduction

The goal of cluster analysis is to group a given collection of objects in such a way that instances in the same cluster are as similar as possible, according to a suitable "similarity" criterion (see, e.g., [5]). One of the most popular and widely used clustering techniques is K-means algorithm (see [7] and the references therein). For a given dataset of n observations $\mathbf{Y} = (y_1, \ldots, y_n)$, with $y_i \in \mathscr{Y} \subset \mathbb{R}^d,\ i = 1, \ldots, n$, K-means seeks to find a partition of the data into K clusters $\mathscr{S} = \{S_1, \ldots, S_K\}$ so as to minimize

$$\phi(\mathscr{S}) = \sum_{k=1}^{K} \sum_{y_i \in S_k} ||y_i - \mu_k||^2,$$

where μ_k is the k-th cluster center. The algorithm begins with K randomly initialized centers, and assigns each point to the nearest center; then, the clusters' centers are recomputed and the partition updated; such process is iterated until a stable configuration is reached. In many instances, the number of clusters K is specified in advance, and the optimal solution is sought conditional on such value. Also the distance adopted is set in advance, depending on the nature of the data and user subjective preferences. Initial seeding is a more technical issue, usually less discussed, whose impact on the final result is often neglected. Nonetheless, the choice of the initial group centers may strongly affect the clustering solution. As already mentioned, the classical approach in K-means clustering uses a random seeding in the first step of the procedure. In practice, multiple random seeds are considered, and the final K-means solution is chosen as the one that minimizes the objective function. An alternative seeding technique is proposed by [1], among others. Many extensions of the classical approach have been explored in the literature (for a complete review see [7]). As it is discussed in [7], classical approaches have been challenged by using ensembles methods.

Clustering ensembles methods ([9, 10]) explore the idea of the so-called evidence accumulation in order to summarize the information coming from multiple clusterings into a pairwise *co-association* matrix, regarded to as a similarity matrix ([8, 6]). Such matrix is constructed by taking the co-occurrences of pairs of units in the same cluster among the total number of partitions, and then used to deliver the final consensus clustering. This kind of matrix has been estimated in [4] and used in the context of the label switching problem in Bayesian estimation of finite mixture models. In particular, the algorithm of Maxima Units Search (MUS), introduced in [4] and further developed in [3], has proved to be useful in extracting some specific units–one for each mixture component–called pivots, from a large and sparse similarity matrix representing an estimate of the probability that pairs of units belong to the same group.

Here the main idea is to exploit the use of the pivots detected by the MUS algorithm, which are determined as the observations that are "as far away from each other as possible" according to the co-association matrix, as group centers in the initialization step of K-means procedure, where such units.

Section 2 briefly reviews the MUS procedure for the identification of pivotal units from a given set of data, and outlines the proposed approach. In Section 3 the performance of the presented algorithm is preliminarily investigated by means of a small simulation study. Section 4 presents some final remarks.

2 Seeding via MUS algorithm

Consider H distinct partitions of a set of n d-dimensional statistical units into K groups determined by some clustering technique. It is possible to map them into a $n \times n$ co-association matrix C with generic element $c_{i,j} = n_{i,j}/H$, where $n_{i,j}$ is the number of times the pair (y_i, y_j) is assigned to the same cluster with respect to the clustering ensemble. Units which are very distant from each other are likely to have zero co-occurrences; as a consequence, C is a square symmetric matrix expected to contain a non-negligible number of zeros.

The main task of the MUS algorithm is to detect submatrices of small rank from the co-association matrix and extract those units $y_{i_1}, \ldots, y_{i_K}$ such that the $K \times K$ submatrix of C with only the $i_1, \ldots, i_K$ rows and columns has few, possibly none, nonzero elements off the diagonal (that is, this submatrix is identical or nearly identical). Practically, the resulting units—hereafter pivots—have the desirable property to be representative of the group they belong to. From a computational point of view, the issue is non-trivial and involves a global search row by row; as n, K and the number of zeros within C increase, the procedure becomes computationally demanding. Given that the pivots correspond to well separated units in the data space, they can represent an alternative approach to the random seeding in K-means setting. A similar idea has been discussed in [1], where the initial centers are chosen on the basis of suitable weights assigned to data points.

Although K-means clustering is one of the most popular algorithms due to its simplicity and low computational burden, one major criticism is the impact of the choice of the initial centers on the final solution. However, limited work has been developed for improving the seeding of the centers. A modified version of K-means could benefit from a pivot-based initialization step. In particular, the starting point is performing multiple runs of the classical K-means with K fixed, and build the co-association matrix of data units. Such matrix is given as input for the MUS procedure, yielding the pivots regarded to as cluster centers. Intuitively, such approach represents a careful seeding which may improve the validity of the final configuration. According to the general K-means method, steps 1a—1c of the `MUS`K-means algorithm summarized below collapse in a single step, where the initial centers are chosen uniformly at random from the data space $\mathscr{Y}$. The remaining steps coincide with those of the classical K-means version.

MUSK-means:

1a Perform H classical K-means algorithms, and obtain then H distinct data partitions, with initial centers chosen uniformly at random.

1b Build the co-association matrix C, where $c_{i,j} = n_{i,j}/H$, with $n_{i,j}$ the number of times the pair (y_i, y_j) is assigned to the same cluster among the H partitions.

1c Apply the MUS algorithm to the matrix C and find the pivots $y_{i_1}, \ldots, y_{i_K}$. For each group, set the initial center $\mu_k = y_{i_k}$.

2 For each k, $k = 1, \ldots, K$, set the cluster S_k to be the set of points in $\mathscr{Y}$ that are closer to μ_k than they are to μ_j for all $j \neq k$.

3 For each k, $k = 1, \ldots, K$, set μ_k to be the center of mass of all points in S_k: $\mu_k = \frac{1}{|S_k|} \sum_{y \in S_k} y$, where $|S_k|$ is the cardinality of S_k.

4 Repeat Steps 2 and 3 until $\mathscr{S}$ no longer changes.

3 Simulation results

A preliminary simulation study is carried out in order to explore the performance of the methodology proposed in Sect. 2. One of the drawbacks of K-means is its inefficiency in distinguishing between groups of unbalanced sizes. For this reason, two different scenarios in which the classical approach may fail to identify the 'natural' groups are considered in the following. In particular, the two simulated datasets reproduce two clusterings in two dimensions, with three and two groups, respectively. For illustration purposes, the results from a single simulation are shown in Fig. 1. The left panel (top) displays the first simulated scenario, where the input data consist of three clusters drawn from bivariate Gaussian distributions with 20, 100 and 500 observations, respectively. The partitions obtained from the classical K-means algorithm using multiple random seeds and from MUSK-means are plotted in the top central and right panel of Fig. 1, respectively. As can be seen, classical K-means tends to split the cluster with the highest density in two separate clusters; conversely, the cluster composition identified by MUSK-means shows a greater agreement with the true partition, and the final centers are close to the pivotal units used for the seeding. The second configuration (see the bottom panel of Fig. 1) consists of data with 'two-sticks' shaped groups of 30 and 370 observations, respectively. Classical K-means fails in recognizing the true pattern, and the final centers both belong to the largest cluster. Clustering based on our pivotal units seems to correctly identify the simulated groups, since two well separated pivots are identified and set as initial group centers.

In order to evaluate and compare the performance of classical K-means and MUSK-means, a common measure of the similarity between two partitions, namely,

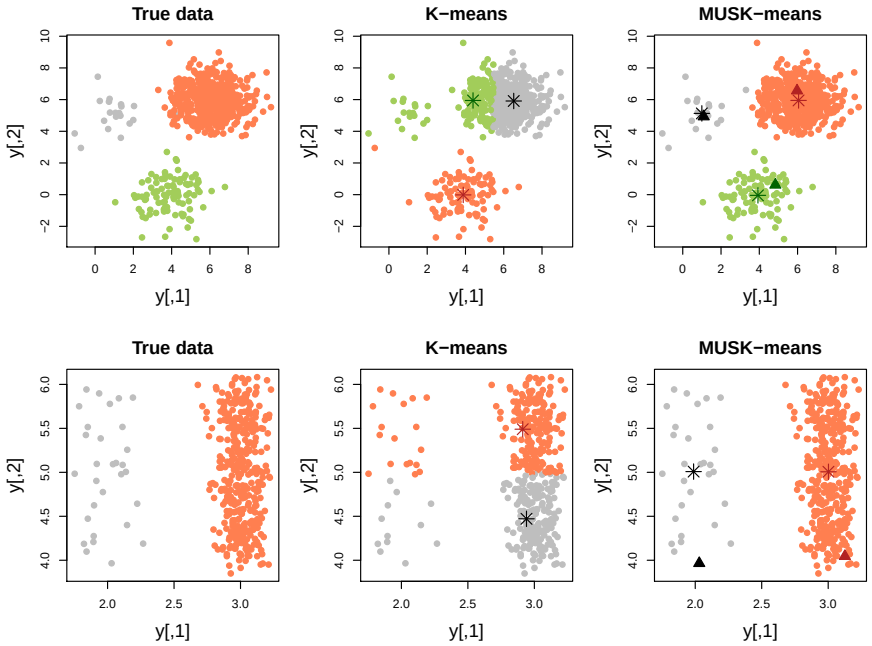

Fig. 1: From left to right. Input data generated from a mixture of three Gaussian distributions (620 samples) (top) and 'stick' data (400 samples) with two groups (bottom) of unequal sample sizes; clustering solutions obtained via classical K-means and MUSK-means algorithms. Each cluster identified is shown in a different color, with final group centers and pivots marked via asterisks and triangles symbols, respectively.

the Adjusted Rand Index (ARI), is computed at each iteration between the resulting clustering and the true data partition. The number of replications in the simulation study is set equal to 1000. Fig. 2 shows the comparison in terms of ARI, for the first scenario considered. As may be noted, MUSK-means gives overall good results; in fact, it yields higher values for 60% of the replications, whereas the two procedures yield the same value of the index in 36% of cases. Concerning the second scenario characterized by 'two-sticks' data, the ARI for K-means is always approximately equal to 0; the same index for MUSK-means is about zero in 43% of cases, while for the remaining 57% it outperforms classical K-means giving an ARI equal to 1, denoting a perfect agreement with the true partition.

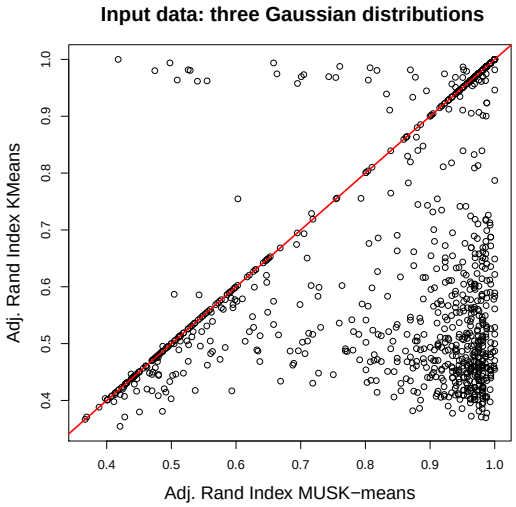

Fig. 2: Comparison between the ARI obtained via classical K-means and MUSK-means algorithms for the input Gaussian data over 1000 replications.

4 Discussion

We propose a modified K-means algorithm which exploits a pivotal-based phase seeding. Despite the limited study, preliminary results seem to be promising in terms of clusters' identification. It is worth noting that the proposed algorithm is in general computationally more demanding that the standard procedure, and the complexity grows with the size of the dataset. On the other hand, similarly to clustering ensemble, our method takes advantage of the construction of a co-association matrix, whose information has been only partially exploited so far. Further work is needed to investigate the use of such matrix and the pivotal units for inferring the optimal number of groups.

References

1. Arthur, D., Vassilvitskii, S.: k−means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 1027-1035 (2007)
2. Dongkuan, X., Yingjie, T.: A comprehensive survey of clustering algorithms. Annals of Data Science **2**(2), 165–193 (2015)
3. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Maxima Units Search (MUS) algorithm: methodology and applications. In: Perna, C. , Pratesi, M., Ruiz-Gazen A. (eds.) Studies in Theoretical and Applied Statistics, Springer Proceedings in Mathematics & Statistics 227, pp. 71–81 (2018)

4. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Relabelling in Bayesian mixture models by pivotal units. Stat. Comput. **28**(4), 957–969 (2018)
5. Everitt, B.S.: Cluster Analysis. Heinemann, London, United Kingdom (1981)
6. Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6), 835–850 (2005)
7. Jain, A.: Data clustering: 50 years beyond K-means. Patt. Recog. Lett. **31**(8), 651–666 (2010)
8. Lourenco, A., Bulò, S. R., Fred, A., Pelillo, M.: Consensus clustering with robust evidence accumulation. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 307-320. Springer, Berlin, Heidelberg (2013)
9. Strehl, A., Joydeep G.X: Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
10. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence **25**(3), 337–372 (2011)