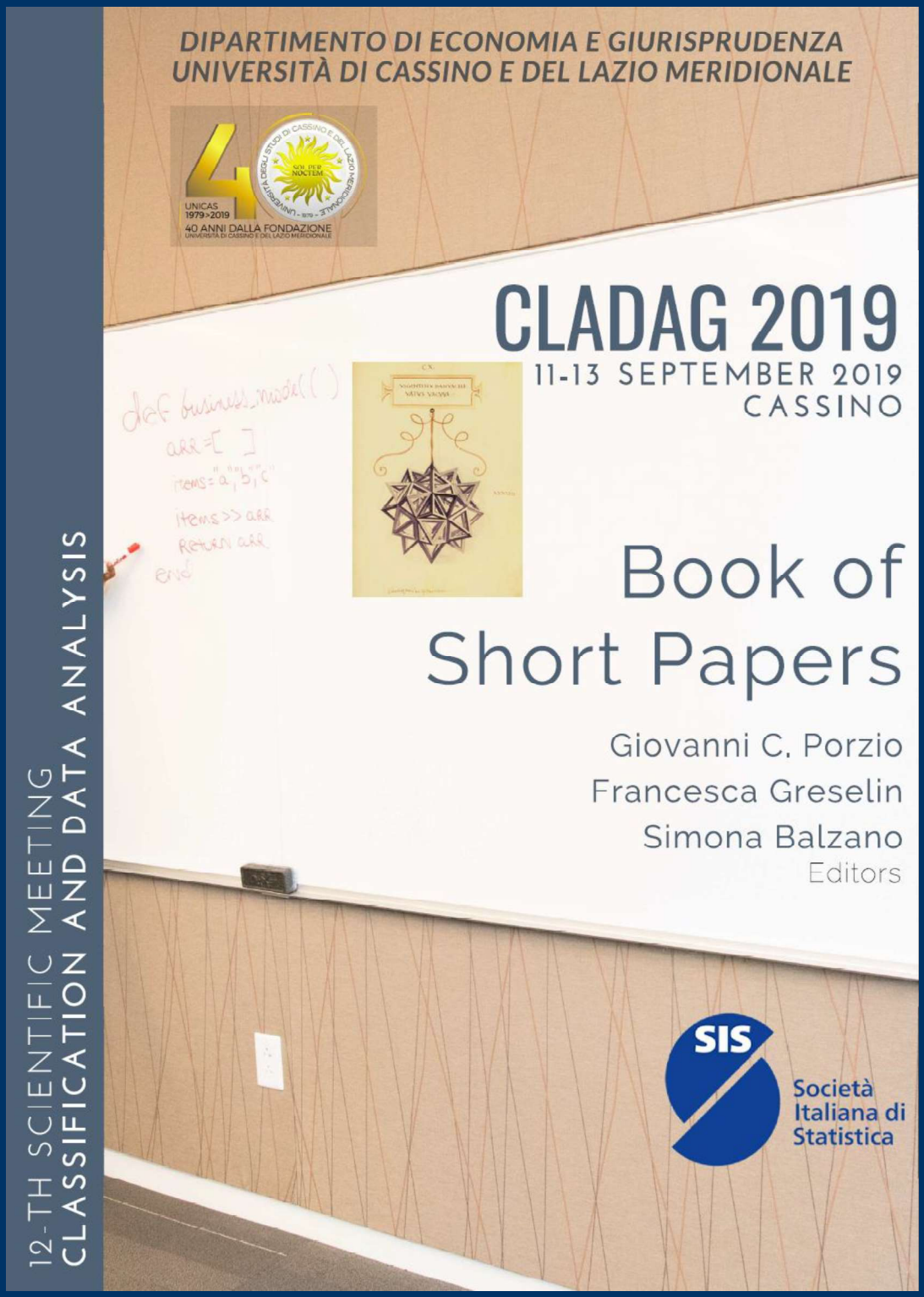
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6

CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

Contents

Keynotes lectures

Invited and contributed sessions

CONSENSUS CLUSTERING VIA PIVOTAL METHODS

Leonardo Egidi[1], Roberta Pappadà[1], Francesco Pauli[1] and Nicola Torelli[1]

[1] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste, (e-mail: legidi@units.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it)

ABSTRACT: We propose an approach to the cluster ensemble problem based on pivotal units extracted from a co-association matrix. It can be seen as a modified version of K-means method, which utilizes pivots for careful seeding. Different criteria for identifying the pivots are discussed, as well as preliminary results concerning the comparison with alternative ensemble methods.

KEYWORDS: cluster ensemble, pivot, K-means.

1 Introduction

Ensembles methods have recently emerged as a valid alternative to conventional clustering techniques and have shown to effectively improve the quality of clustering results and achieve robustness (see, e.g., Strehl & Ghosh, 2002, Jain, 2010). Such methods require a strategy to generate multiple clusterings of the same data set (the ensemble) and then combine them into a *consensus* partition (presumably superior), by following the idea of evidence accumulation, i.e., by viewing each clustering result as an independent evidence of data structure. A common way to do this is to obtain a new pairwise similarity matrix, or co-association matrix, by taking the co-occurrences of pairs of points in the same group across all partitions (Fred & Jain, 2005). Then, a similarity-based clustering algorithm can be applied to this matrix to yield the final partition.

We propose to use the co-association matrix to find some specific units (hereafter, pivots) which are representative of the group they belong to (because they never or very rarely co-occur with members of other groups). Various criteria for detecting the pivots are proposed in Section 2. Section 3 illustrates the use of pivotal methods for data clustering, and compare the proposed approach with classical K-means and other common ensemble methods.

Pivotal methods and related clustering procedures are implemented via the R package `pivmet`, available from the Comprehensive R Archive Network at

http://CRAN.R-project.org/package=pivmet.

2 Pivotal methods based on co-association

Let $\mathbf{Y} = (y_1, \ldots, y_n)$ be a set of n observations, where $y_i \in \mathbb{R}^d$. Consider a set $\mathcal{P} = \{P^1, P^2, \ldots, P^H\}$ of H partitions of the data points into K disjoint clusters, derived from an arbitrary clustering algorithm. Note that the number of groups is pre-specified and equal for all P^h. $\mathcal{P}$ can be summarized via the $n \times n$ co-association matrix C with generic element

$$c_{i,j} = \frac{1}{H} \sum_{h=1}^{H} |P^h(y_i) = P^h(y_j)|, \tag{1}$$

where $|\cdot|$ denotes the indicator function, and $P^h(y_i)$, $P^h(y_j)$, represent the clusters of the objects y_i and y_j in P^h, respectively. Clearly, units which are very dissimilar from each other are likely to have zero co-occurrences; as a consequence, C is expected to contain a non-negligible number of zeros. Given a large and sparse 0-1 matrix, the Maxima Units Search (MUS) algorithm seeks those elements, among a pre-specified number of candidate pivots, whose corresponding rows contain more zeros compared to all other units (Egidi *et al.*, 2018c). Define a reference partition, $G_1, \ldots, G_K$ of $y_1, \ldots, y_n$ obtained by applying, for instance, an agglomerative hierarchical algorithm into K groups. The MUS procedure takes C as input and outputs a set of K units–one for each group of the reference partition–that exhibit the highest degree of separation (Egidi *et al.*, 2018b). As an alternative approach, the pivot y_{i_k} for group G_k can be chosen so that it is as far as possible from units that might belong to other groups and/or as close as possible to units that belong to the same group, according to one of the following objective functions

$$(a) \max_{i_k} \sum_{j \in G_k} c_{i_k,j} \quad (b) \min_{i_k} \sum_{j \notin G_k} c_{i_k,j} \quad (c) \max_{i_k} \sum_{j \in G_k} c_{i_k,j} - \sum_{j \notin G_k} c_{i_k,j}, \tag{2}$$

where $c_{i,j}$ is defined as in (1). Ideally, the $K \times K$ submatrix of C with only the rows and columns corresponding to $i_1, \ldots, i_K$ will be the identity matrix. In practice, it may contain few nonzero elements off the diagonal.

3 A simulation experiment

In order to illustrate the proposed algorithm, we simulate bivariate data from a mixture of three Gaussian distributions with mean vectors $\boldsymbol{\mu}_1 = (1, 5)$, $\boldsymbol{\mu}_2 =$

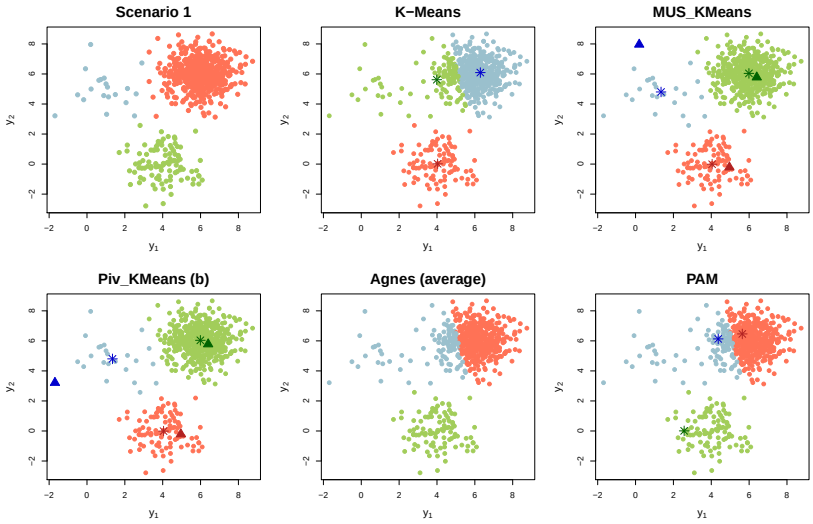


Figure 1. *Mixture of three Gaussian distributions (sample size n=620). Cluster centers and/or pivots for each method are marked via asterisks and triangles, respectively.*

$(4,0)$, $\boldsymbol{\mu}_3 = (6,6)$, and the 2×2 identity matrix as covariance matrix. The components have sample size 20, 100 and 500, respectively (see Figure 1, top-left panel). The K-means algorithm with random seeds is used to generate a cluster ensemble of $H = 1000$ partitions, and obtain the co-association matrix C. For each simulated dataset, we proceed as follows:

1. For a given number of clusters K, obtain a partition of the data $G_1, \ldots, G_K$ (reference partition);
2. Apply the MUS algorithm or one alternative criterion in (2) to the matrix C to find K (distinct) pivots $y_{i_1}, \ldots, y_{i_K}$;
3. Run the K-means algorithm using the pivots as initial cluster centers.

The proposed modification of the standard K-means technique introduces a pivot-based initialization step with the aim of reducing the effect of random seeding (see also Egidi *et al.*, 2018a). An alternative approach to careful seeding can be found in Arthur & Vassilvitskii, 2007. Figure 1 shows the solution from K-means, using $K = 3$, and by pivotal methods MUS and criterion (b) in Eq. (2), where Average-Linkage (AL) agglomerative clustering is used to obtain the reference partition. The results of consensus clustering using PAM (Partitioning Around Medoids) method and AL-agglomerative hierar-

chical clustering (agnes) are also shown (Single Linkage (SL) and Complete Linkage (CL) give similar results). Table 1 reports the comparison between the different methods in terms of Adjusted Rand Index (ARI), used to quantify the agreement between two partitions. The mean value is considered for 1000 simulations. Preliminary results suggest that the pivot-based approach outperforms the competing similarity-based ensemble methods and the standard K-means, which gives a mean ARI of 0.659.

Table 1. *2D Gaussian data: mean ARI (*1000 *simulations) between the final clustering and the true data partition. Ensemble methods use dissimilarities* $1 - c_{i,j}$.

Pivotal methods	MUS 0.857	(a) 0.865	(b) 0.883	(c) 0.779
Ensemble methods	agnes (*AL*) 0.512	agnes (*SL*) 0.535	agnes (*CL*) 0.514	PAM 0.506

References

ARTHUR, D., & VASSILVITSKII, S. 2007. k-means++: The advantages of careful seeding. *Pages 1027–1035 of: Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete algorithms*.

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018a. K-means seeding via MUS algorithm. *Pages 256–262 of: Book of Short Papers SIS 2018*.

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018b. Maxima Units Search (MUS) algorithm: methodology and applications. *Pages 71–81 of: Studies in Theoretical and Applied Statistics*.

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018c. Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, **28**, 957–969.

FRED, A. L. N., & JAIN, A. K. 2005. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 835–850.

JAIN, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651 – 666.

STREHL, A., & GHOSH, J. 2002. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, **3**, 583–617.