Spectral Clustering Using Compactly Supported Graph Building

A.M. Álvarez-Meza, A.E. Castro-Ospina, and Germán Castellanos-Domínguez

Signal Processing and Recognition G., Universidad Nacional de Colombia, Manizales
{amalvarezme,aecastroo,cgcastellanosd}@unal.edu.co

Abstract. Clustering techniques demand on suitable models of data structures to infer the main samples patterns. Nonetheless, detection of data structures becomes a difficult task when dealing with nonlinear data relationships and complex distributions. Here, to support clustering tasks, we introduce a new graph building strategy based on a compactly supported kernel technique. Thus, our approach makes relevant pairwise sample relationships by finding a sparse kernel matrix that codes the main sample connections. Clustering performance is assessed on synthetic and real-world data sets. Obtained results show that the proposed method enhances the data interpretability and separability by revealing relevant data relationships into a graph-based representation.

1 Introduction

In disciplines like data mining and machine learning, the presence of available unlabeled data gives rise to develop unsupervised inference techniques, which commonly make use of many measures of proximity (dissimilarity/similarity) for associating unlabeled samples. Nevertheless, to get a suitable metric-based representation that allows encoding complex data structures still poses an open issue. Specifically, the baseline k-means algorithm produces data clusters by imposing Euclidean-based similarities among samples. However, this technique demands on fixing the number of data groups and is sensitive to its initialization, mostly, in non-linear distributed data problems [6].

In order to exploit more powerful similarities among samples, spectral-based clustering techniques are developed, which build a weighted graph to model complex data structures. Moreover, some alternatives of graph construction have been proposed to encode non-linear similarities among samples, such as ε-neighborhood graph, k-nearest neighbors graph, fully connected graph, and graph building based on local scaling [7,9]. Nevertheless, the performance of these graph-like versions is highly dependent on the prior knowledge about the needed free parameters, tending to fail when dealing with noisy distributions and complex data structures suffering discontinuities.

Here, we introduce a graph building representation methodology based on compactly supported radial basis functions to highlight relevant sample similarity relationships. We aim to search for a sparse sample representation that better unfolds the main data structure. To this end, we employ a compactly supported

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 327–334, 2014.

kernel-based function to build a graph that codes relevant sample relationships while avoiding irrelevant connections [10]. Carried out testing over synthetic and real-world data sets shows that the proposed methodology improves clustering performance in comparison to the considered baseline algorithms due to the built graph omits those connections not related to the data structure. Furthermore, the proposed graph building strategy can be useful to enhance the data interpretability when dealing with complex structures.

2 Graph-Based Data Structure Identification by Compactly Supported Criterion

Let $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ be an input data matrix holding N samples and P features, where each row $\{\boldsymbol{x}_i\in\mathbb{R}^P : i=1,\ldots,N\}$ represents a data sample. To discover relevant input data structure, relationships among samples can be highlighted by means of a complete, weighted, undirected graph representation $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{\Omega})$, which contains a set of nodes $\boldsymbol{V}=\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N\}$. Edge weights for connecting nodes i to j $(i\neq j)$ are defined through the affinity matrix $\boldsymbol{\Omega}\in\mathbb{R}^{N\times N}$, holding elemnts $\Omega_{ij}=\mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)$, being $\mathcal{K}(\cdot,\cdot)$ a positive definite kernel function, mostly, assumed as the Gaussian kernel [3]. Among many others kernels such as the Laplacian and polynomial, the Gaussian function is the only one able to find a Hilbert space with universal approximating capability [5].

Since any kernel function must satisfy the Mercer conditions, its use should ensure a stable spectral decomposition. Therefore, aiming to estimate existing relevant sample similarities, a compactly supported kernel-based representation is computed as a sparse version of matrix $\boldsymbol{\Omega}$, yielding:

$$\boldsymbol{\Omega}_\gamma = \boldsymbol{\Phi}_\gamma \circ \boldsymbol{\Omega} \tag{1}$$

being $\gamma\in\mathbb{R}^+$ a given threshold parameter controlling the sparsity degree of the kernel matrix $\boldsymbol{\Omega}_\gamma\in\mathbb{R}^{N\times N}$, notation $\circ$ stands for the Hadamard product, and $\boldsymbol{\Phi}_\gamma\in\mathbb{R}^{N\times N}$ is an sparsification matrix with elements, $\phi_\gamma(ij)$, calculated as:

$$\phi_\gamma(ij) = \Big([1-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2/\gamma]_+\Big)^\nu, \tag{2}$$

where function $[\cdot]_+$ computes the positive part of its argument and with $\nu\geq(P+1)/2$, and notation $\|\cdot\|$ stands for the Euclidian norm.

To achieve a suitable local and global data structure representation, the value of γ must be properly computed by finding a sparse matrix $\boldsymbol{\Omega}_\gamma$ properly encoding relevant node connections. Hence, we propose to exploit the correlation between $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_\gamma$ matrices using the following empirical kernel alignment function:

$$\mathcal{A}(\gamma) = \frac{\left\langle \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}_\gamma \right\rangle_F}{\left\|\tilde{\boldsymbol{\Omega}}\right\|_F \left\|\tilde{\boldsymbol{\Omega}}_\gamma\right\|_F}, \quad \mathcal{A}(\gamma)\in\mathbb{R}[0,1] \tag{3}$$

where notations $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ stand for the Frobenius inner product and norm, respectively; both matrices, $\tilde{\boldsymbol{\Omega}} = \boldsymbol{H}\boldsymbol{\Omega}\boldsymbol{H}$, $\tilde{\boldsymbol{\Omega}}_\gamma = \boldsymbol{H}\boldsymbol{\Omega}_\gamma\boldsymbol{H}$, are the centralized version of the considered kernels. Here, matrix $\boldsymbol{H} = \boldsymbol{I} - N^{-1}\mathbf{1}\mathbf{1}^\top$ is a centralization matrix, $\boldsymbol{I}$ is the N-dimension identity matrix, and $\mathbf{1}$ is an all-ones N-dimensional vector. It is worth noting that centered alignment-based functions have been demonstrated to correlate better than the uncentered one according to previous theoretical and experimental analysis [2].

With regard to quantify matrix sparsity of $\boldsymbol{\Omega}_\gamma$, we make use of the following rate:

$$s_\gamma = N_0/N^2, \quad s_\gamma \in \mathbb{R}[0,1] \tag{4}$$

where $N_0 \in \mathbb{N}$ stands for the number of zero-entries in $\boldsymbol{\Omega}_\gamma$.

To properly fix the values of empirical alignment $\mathcal{A}(\gamma)$ and sparsity s_γ, both valued within the range $[0, 1]$, we must take into account the following statements: a higher $\mathcal{A}(\gamma)$ value – a lower information loss due to the sparsification process. In contrast, a higher s_γ value – a higher degree of sparsity.

Provided the above statements, we introduce a regularization-based criterion to fix γ optimizing the trade-off between $\mathcal{A}(\gamma)$ and s_γ values, that is:

$$\gamma^* = \underset{\gamma}{\operatorname{argmin}} \left\{ \sqrt{(1-\lambda)(\log \mathcal{A}(\gamma))^2 + \lambda(\log s_\gamma)^2} \right\}, \tag{5}$$

being $\lambda \in [0, 1]$ the regularization parameter.

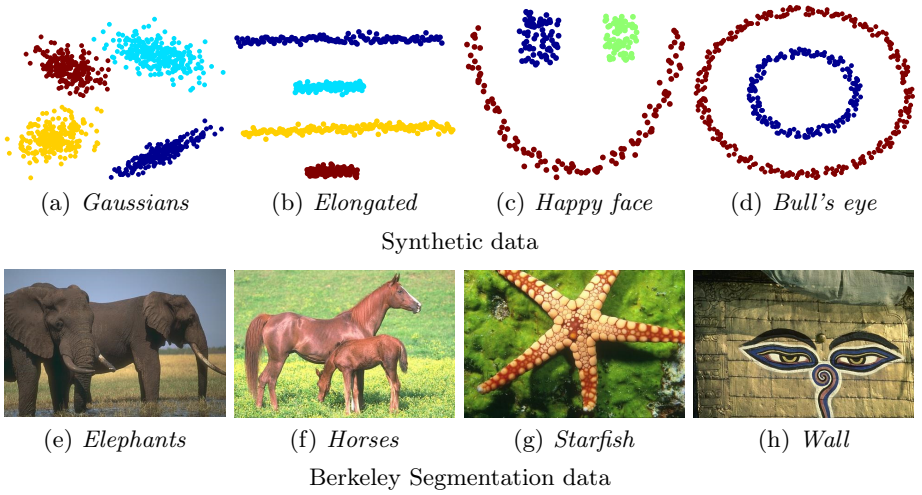
In consequence, the cost function given in Eq. (5) allows finding such a γ^* value that better makes prominent the main data structures during calculation of the sparse matrix $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$, given in Eq. (1), that is further used to build a suitable input data graph representation, $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}^*})$, to be employed in the next learning stages. We termed the proposed approach as Compactly Supported Graph (CSG). AS a result, we enhance the graph representation by holding the most relevant relationships among data points, therefore, revealing data structure and improving interpretability.

3 Experimental Set-up

3.1 Databases and Preprocessing

We provide validation of the proposed methodology for Spectral Clustering using Compactly Supported Graph Building over both synthetic and real data. In the former case, we employ three well-known synthetic collections that are publicly available at[1] representing a challenging clustering task due to their complex structures. In the latter case of real-world experimentation, we randomly pick out a collection of 30 images from the Berkeley Segmentation data set that is also publicly available at[2]. Fig. 1 shows an exemplary of both used data sets during testing.

[1] `http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html`
[2] `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`



(a) *Gaussians* (b) *Elongated* (c) *Happy face* (d) *Bull's eye*

Synthetic data

(e) *Elephants* (f) *Horses* (g) *Starfish* (h) *Wall*

Berkeley Segmentation data

Fig. 1. Exemplary of used data sets during testing

For preprocessing the tested real data sets, we resize every image at 15% to be further characterized by the widely used color spaces, namely: RGB, normalized RGB, HSV, and YCbCr. In addition, the spatial position of each pixel is also taken as another feature. Therefore, each image is represented by the corresponding input matrix $\boldsymbol{X} \in \mathbb{R}^{N \times 14}$, with N the number of pixels per image.

3.2 Kernel Parameter Tuning and Clustering Performance Measure

As stated earlier, we estimate the similarity matrix $\boldsymbol{\Omega}$ employing the Gaussian kernel. To tune the needed bandwidth value, we employ the novel approach introduced in [1] that takes into account both local and global properties to identify the main relationships between a given point and neighboring samples around it. Afterwards, the optimal parameter γ^* is computed by solving the cost function presented in Eq. (5), in the concrete case, by using the Particle Swarm Optimization-based solver, where the regularization parameter λ, is heuristically set to 0.5. Also, searching for γ^* is constrained with the interval ranging from the maximum to the minimum pairwise input sample distances estimated by the Euclidean metric. Lastly, we compute the sparse matrix $\boldsymbol{\Omega}_{\gamma^*}$ to perform the well-known spectral clustering algorithm [9]. Fig. 2 summarizes the main sketch of the proposed spectral clustering methodology based on the proposed CSG.

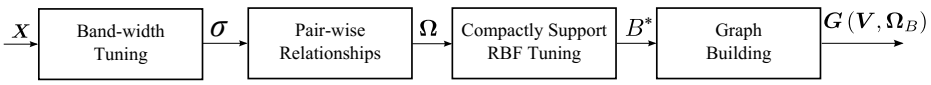


Fig. 2. Compactly Supported Graph representation scheme

For the sake of comparison, the proposed approach is contrasted against the baseline K-means clustering algorithm and the 7-Nearest Neighbor Spectral Clustering (7-NNSC) that is based on local scaling analysis [7,4]. To assess the clustering performance, we initially provide visually inspection of each considered clustering algorithm for synthetic data results. Moreover, for real-world data and taking advantage of hand-labeled ground-truth given by the data set, we compute the Normalized Probabilistic Rand (NPR) index allowing to measure the segmentation performance. NPR considers the label consistence of pixel pairs between the estimated label and a set of ground truth segmentations (see [8] for details).

4 Results and Discussion

As seen in Fig. 3 showing the graphs obtained by 7-NNSC and CSG approaches using the synthetic data sets, the former clustering approach provides a graph representation having more connections in comparison to the CSG one. Consequently, such a high connectivity behavior can mislead data interpretability and clustering procedures.

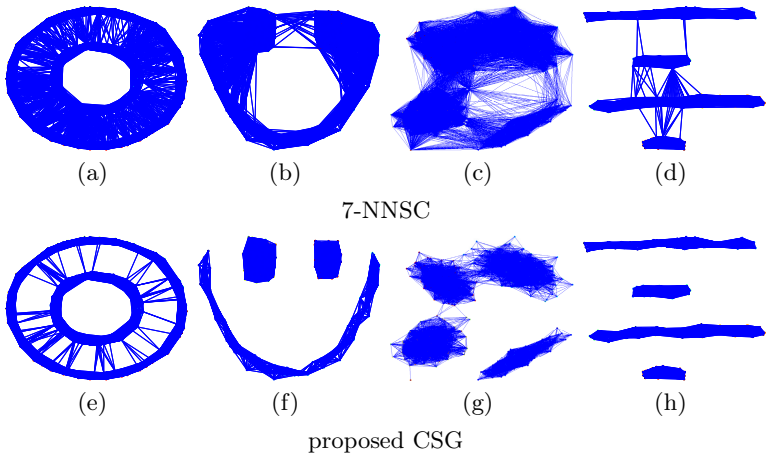


Fig. 3. Examples of accomplished graph representations for tested synthetic data sets

At the same time, Fig. 4 illustrates some clustering results obtained for k-means, 7-NNSC, and CSG approaches. The first algorithm achieves wrong assignments of clusters due to the sensitivity of the algorithm to its random centroid initialization that can be also biased by the Euclidean metric. As a result, the k-means clustering is not able to deal with data exhibiting complex structures.

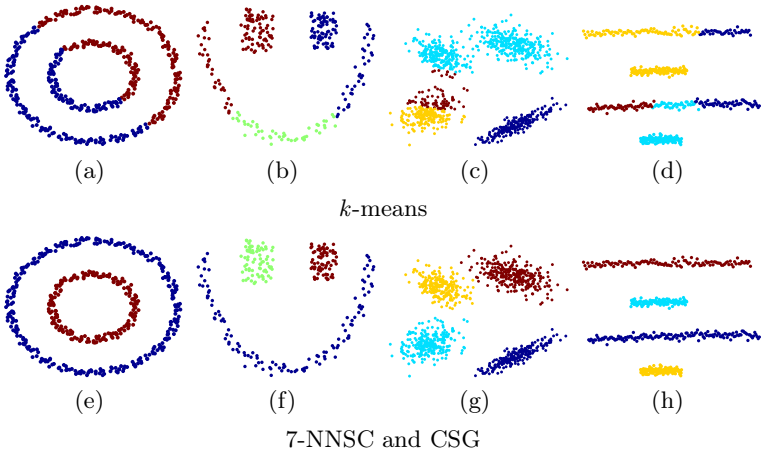
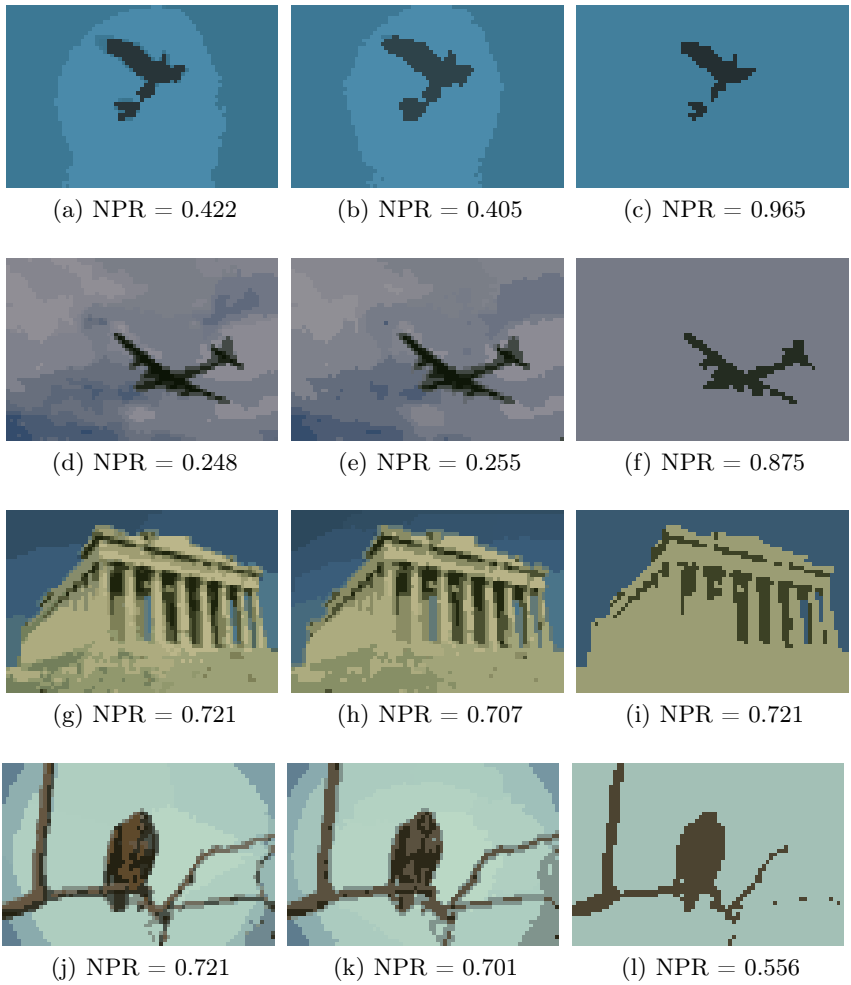


Fig. 4. Visual inspection of the resulting clustering (synthetic data sets)

In turn, the more elaborate 7-NNSC approach riches better clustering performance since it takes advantage of the graph-based representation as part of the achieved spectral clustering strategy. Nevertheless, the 7-NNSC requires to manually fix the number of neighbors for building the graph representation, whenever new input data are given. In practice, this tedious routine can be quite time consuming. With regard to the CSG clustering, the obtained results clearly evidence that the use of graph representations based on relevant sample relationships allows reaching suitable groups. Hence, we can infer that avoiding unnecessary connections among samples in the graph representation by means of the proposed CSG approach does not bias the clustering estimation.

To get visual inspection of the resulting clustering over real data, Fig. 5 shows an exemplary of the achieved image segmentation clustering. Particularly, Figs. 5(a) to 5(f) makes clear the advantage of the CSG in comparison to the other compared baseline methods by achieving higher NPR values. It should be quoted that the proposed method is less sensitive to false contours than k-means and 7-NNSC, as seen in Figs. 5(a) and 5(b). Moreover, the CSG-based clustering is able to deal with noisy environments by discarding not relevant sample relationships while retaining the main process information (see Figs. 5(f) and 5(i)). All above reached results can be explained by the fact that, in the our experiments, both the sparsification and centered alignment functions are equally weighted (i.e., λ=0.5) during searching the optimal value γ^*. Even so, in cases of data clusters having more richness of detail, the chosen value λ=0.5 may not be the best alternative, as seen in Fig. 5(l). Overall, proposed CSG method outperforms both k-means and 7-NNSC segmentations, being more stable and accurate in terms of NPR measure according to the statistical analysis shown in Fig. 6.



(a) NPR = 0.422 (b) NPR = 0.405 (c) NPR = 0.965

(d) NPR = 0.248 (e) NPR = 0.255 (f) NPR = 0.875

(g) NPR = 0.721 (h) NPR = 0.707 (i) NPR = 0.721

(j) NPR = 0.721 (k) NPR = 0.701 (l) NPR = 0.556

Fig. 5. Some image segmentation results. Column 1: k-means. Column 2: 7-NNSC. Column 3: CSG.

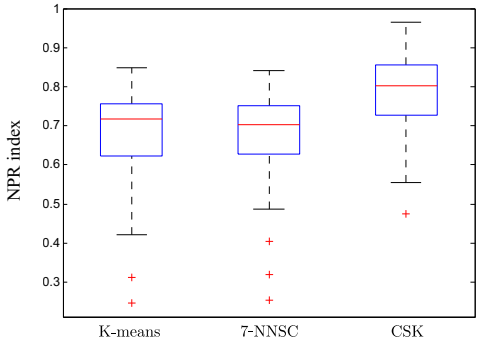


Fig. 6. Boxplots of the image segmentation results

5 Conclusions

An approach to build a suitable graph representation to support spectral clustering techniques was proposed. In this sense, a compactly supported kernel strategy is used to disconnect irrelevant connections or pairwise similarities in a given data set, while highlighting the main data structures. So, a regularization-based objective function was introduced to find a sparsification matrix that aims to preserve relevant connections or similarities into a graph. The proposed approach, CSG, is tested over synthetic and real-world data sets and compared against two benchmark approaches, namely, k-means and 7-NNSC. Obtained results show how a correct removal of irrelevant connections does not affect the clustering performance over synthetic data, and even enhances the data interpretability and separability. Moreover, CSG proposed approach outperforms, in most of the cases, image segmentation performances in comparison to baseline algorithms, by dealing with complex distributed data. As future work, it would be interesting to test different metrics in more complex clustering problems.

Acknowledgments. This work is carried out under grants provided by *Programa Nacional de Formación de Investigadores "Generación del Bicentenario"*, 2011/2012 and *Programa Nacional de Jóvenes Investigadores e Innovadores - 2012* (COLCIENCIAS).

References

1. Castro-Ospina, A.E., Álvarez-Meza, A.M., Castellanos-Domínguez, C.G.: Automatic graph building approach for spectral clustering. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) CIARP 2013, Part I. LNCS, vol. 8258, pp. 190–197. Springer, Heidelberg (2013)
2. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. The Journal of Machine Learning Research 13, 795–828 (2012)
3. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recognition 41(1), 176–190 (2008)
4. Liping, C., Xuchuan, Z., Jiancheng, S.: The approach of adaptive spectral clustering analyze on high dimensional data. In: ICCIS, pp. 160–162 (2010)
5. Liu, W., Principe, J.C., Haykin, S.: Kernel Adaptive Filtering: A Comprehensive Introduction, vol. 57. John Wiley & Sons (2011)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, California, USA, vol. 1, p. 14 (1967)
7. Perona, P., Zelnik-Manor, L.: Self-tuning spectral clustering. Advances in Neural Information Processing Systems 17, 1601–1608 (2004)
8. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 929–944 (2007)
9. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
10. Zhang, H., Genton, M., Liu, P.: Compactly supported radial basis function kernels (2004), `http://www4.stat.ncsu.edu/hzhang/research.html`