An Improved Multi-Class Spectral Clustering Based on Normalized Cuts

Diego Hernán Peluffo-Ordóñez, Carlos Daniel Acosta-Medina, and César Germáan Castellanos-Domínguez

Signal Processing and Recognition Group, Universidad Nacional de Colombia, Km. 9, Vía al aeropuerto, Campus la Nubia, Caldas, Manizales, Colombia
{dhpeluffoo,cdacostam,cgcastellanosd}@unal.edu.co

Abstract. In this work, we present an improved multi-class spectral clustering (MCSC) that represents an alternative to the standard k-way normalized clustering, avoiding the use of an iterative algorithm for tuning the orthogonal matrix rotation. The performance of proposed method is compared with the conventional MCSC and k-means in terms of different clustering quality indicators. Results are accomplished on commonly used toy data sets with hardly separable classes, as well as on an image segmentation database. In addition, as a clustering indicator, a novel unsupervised measure is introduced to quantify the performance of the proposed method. The proposed method spends lower processing time than conventional spectral clustering approaches.

1 Introduction

Spectral clustering has taken an important place within the context of pattern recognition, mainly, because this technique represents a very suitable alternative to solve problems when data are not labeled and classes are hardly separable. Derived from the normalized cuts-based clustering, described in detail in [1], many enhancing approaches have been proposed. For instance, kernel-based methods employing support vector machines (SVM) are discussed in [2–4], and methods with improved affinity or similarity matrices, proposed in [5]. Spectral clustering technique has been successfully applied in several applications such as image segmentation [6, 7] and has shown to be a powerful tool to determine information about initial data, namely, estimation of the group number [5, 8] and local scaling [9]. Commonly, the application of spectral clustering methods involves the determination of a new representation space, whose resultant dimension is lower than that from original data, and then a dimensionality reduction procedure should be accomplished. In that way, the relation among elements are conserved as well as possible. Thus, eigenvectors and eigenvalues based analysis takes place. This is because the information given by eigen-space (i.e, space generated by eigenvectors) is directly associated with the clustering quality. The computation of such eigen-space is usually a high time consuming computational procedure. Therefore, a computation of spectral clustering method with reasonable computational load, but keeping high clustering performance still remains an open issue.

In this work, an improved alternative to conventional k-way normalized cuts-based clustering is presented, which improves the computational cost avoiding the iterative

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 130–137, 2012.

search for tuning. Improved method also provides a better estimation of the initial parameter from the information given by the proposed solution. We solve the classical problem of spectral clustering without using any heuristical searching approach, instead, we accomplish a deterministic solution by means of solving an equation matrix of the form $\boldsymbol{ARB} = \boldsymbol{C}$, as discussed in [10]. This equation allows to determine the rotation matrix $\boldsymbol{R}$, which generates an infinite number of solutions and is then chosen as that shows better convergence. Solving such equation matrix yields a solution that satisfies the condition given by the objective function but not the orthogonality condition. Therefore, we introduce a regularized form in order to obtain an orthonormal feasible solution. For assessing the performance of proposed method, we carry out experiments on toy data sets as well as the Berkeley Segmentation Dataset [11] to evaluate our method in terms of segmentation. As a clustering performance measure, we apply the total clustering performance taking advantage of the labels and segmented reference images and introduce an unsupervised measure that takes into consideration the quality clustering in terms of spectral information. Also, we include stages for estimating the number of groups and computing the affinity matrix as described in [5]. We compared our method with a conventional *K*-means and a *K*-way normalized-based clustering, as explained in [1].

2 Clustering Method

2.1 Multi-Class Spectral Clustering (MCSC)

A weighted graph can be represented as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \boldsymbol{W})$, where $\mathbb{V}$ is the set of either nodes or vertices, $\mathbb{E}$ is the edge set, and $\boldsymbol{W}$ represents the relationship among nodes, named, *affinity matrix*. Given that, each affinity matrix entry w_{ij} of $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ represents the weight of the edge between i-th and j-th element, it must be a non-negative value. Value N is the number of considered nodes. In addition, for a non-directed graph, it holds that $w_{ij} = w_{ji}$. Therefore, affinity matrix must be chosen as symmetric and positive semi-definite. After clustering procedure, a binary indicator matrix $\boldsymbol{M} = [\boldsymbol{m}_1| \ldots |\boldsymbol{m}_K]$ is accomplished, where each vector set $\boldsymbol{m}_k$ is a column vector formed by data point membership regarding cluster $k = 1, \ldots, K$ and K is the number of groups. Each entry ik from the $N \times K$ dimensional matrix $\boldsymbol{M}$ is defined as $m_{ik} = \lfloor i \in \mathbb{V}_k \rfloor, \quad i \in \mathbb{V}, \quad k = 1, \ldots, K$, where notation $\lfloor \cdot \rfloor$ stands for a binary indicator - it equals to 1 if its argument is true and, otherwise, 0. Also, because each node can only belong into one partition, the condition $\boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N$ must be satisfied, where $\mathbf{1}_d$ is a d-dimensional ones vector.

Then, the well-known *k*-way normalized cuts-based clustering, described in [1], can be written as:

$$\max \varepsilon(\boldsymbol{M}) = \frac{1}{K}\frac{\text{tr}(\boldsymbol{M}^\top \boldsymbol{W}\boldsymbol{M})}{\text{tr}(\boldsymbol{M}^\top \boldsymbol{D}\boldsymbol{M})} \tag{1a}$$

$$\text{s. t.: } \boldsymbol{M} \in \{0,1\}^{N\times K}, \quad \boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N \tag{1b}$$

where $\boldsymbol{D} = \text{Diag}(\boldsymbol{W}\mathbf{1}_N)$ is the degree matrix related to weights or affinity matrix. Notation $\text{Diag}(\cdot)$ denotes a diagonal matrix formed by its argument vector. Expressions (1a) and (1b) are the formulation of normalized cuts optimization problem, named (*NCPM*).

In order to guarantee that $\boldsymbol{M}$ becomes binary it is needed that $||\boldsymbol{M}||_F^2 = \text{tr}(\boldsymbol{M}^\top \boldsymbol{M}) = \text{tr}(\boldsymbol{E}) = n$, where $\boldsymbol{E} = \text{Diag}(\sum_{i=1}^n m_{i1}, \cdots, \sum_{i=1}^n m_{iK})$. Therefore, the *NCPM* optimization problem can be expressed as:

$$\max \varepsilon(\boldsymbol{M}) = \text{tr}(\boldsymbol{M}^\top \boldsymbol{W} \boldsymbol{M}) \quad \text{s. t.: } \text{tr}(\boldsymbol{M}^\top \boldsymbol{D} \boldsymbol{M}) = \text{const.}, \quad \boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N \tag{2}$$

2.2 Proposed *NCPM* Solution Based on an One Iteration Tuning

The *NCPM* optimization problem can be relaxed as follows. Let $\boldsymbol{P} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ be a normalized affinity matrix. Then, a relaxed *NCPM* version can be expressed as:

$$\max \text{ tr}(\boldsymbol{L}^\top \boldsymbol{P} \boldsymbol{L}), \quad \text{s. t.} \quad \boldsymbol{L}^\top \boldsymbol{L} = \boldsymbol{I}_K \tag{3}$$

where $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{M}$.

A feasible solution to this problem is $\boldsymbol{L}^* = \boldsymbol{V}_K \boldsymbol{R}$, where $\boldsymbol{V}_K$ is any orthonormal basis of the K-dimensional principal subspace of $\boldsymbol{P}$, and $\boldsymbol{R}$ is an arbitrary rotation matrix. At the end, a binarization process must be applied, e.g., by employing the sign function. Thus, there exists an infinite number of solutions. To overcome this issue with the aim to reach a deterministic solution without using an iterative algorithm, the following mathematical development is done.

According to the constraint given in (1b), we have:

$$\boldsymbol{M}\mathbf{1}_K = \boldsymbol{D}^{1/2}\boldsymbol{L}\mathbf{1}_K = \boldsymbol{D}^{1/2}\boldsymbol{V}_K\boldsymbol{R}\mathbf{1}_K = \mathbf{1}_N \tag{4}$$

Therefore, a possible rotation matrix $\boldsymbol{R}$ can be chosen as $\boldsymbol{R} = 1/k^2 \boldsymbol{V}_K^\top \boldsymbol{D}^{-1/2}\mathbf{1}_N\mathbf{1}_K^\top$. Yet, the previous solution do not satisfy the orthonormal condition given by (3), since it holds that $\boldsymbol{R}^\top\boldsymbol{R} = \frac{K}{N}\boldsymbol{D}^{-1} \neq \boldsymbol{I}_K$, and thus, $\boldsymbol{R}^\top \neq \boldsymbol{R}^{-1}$. So, as a way to avoid this drawback, a constrained optimization problem is introduced:

$$\min ||\boldsymbol{V}_K\boldsymbol{R}\mathbf{1}_K - \boldsymbol{D}^{-1/2}\mathbf{1}_N||_F^2, \quad \text{s. t. } \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}_K \tag{5}$$

where $||\cdot||_F$ stands for Frobenius norm.

For the sake of simplicity, let us define $\boldsymbol{A} = \boldsymbol{V}_K$, $\boldsymbol{B} = \mathbf{1}_K$, $\boldsymbol{C} = \boldsymbol{D}^{-1/2}$. Also, we introduce a regularized orthonormal matrix $\widehat{\boldsymbol{R}} = \boldsymbol{R} + \alpha\boldsymbol{I}_K$ to be determined that guarantees the orthogonality condition. Then, the optimization problem is rewritten as:

$$\min ||\boldsymbol{A}\widehat{\boldsymbol{R}}\boldsymbol{B} - \boldsymbol{C}||_F^2, \quad \text{s. t. } \widehat{\boldsymbol{R}}^\top\widehat{\boldsymbol{R}} = \boldsymbol{I}_K \therefore ||\widehat{\boldsymbol{R}}^\top\widehat{\boldsymbol{R}} - \boldsymbol{I}_K||_F^2 = 0 \tag{6}$$

By only considering the objective function to be minimized, the two following solutions are possible [10]: $\text{vec}(\boldsymbol{R}) = [\boldsymbol{B}^\top \otimes \boldsymbol{A}]^\dagger \text{vec}(\boldsymbol{C})$ where $\boldsymbol{B}^\dagger$ represents the pseudo-inverse matrix of $\boldsymbol{B}$, and a real non-negative definite solution.

The latter solution, in its easiest form, is given by $\boldsymbol{R} = \boldsymbol{A}^-\boldsymbol{C}\boldsymbol{B}^- + \boldsymbol{Y} - \boldsymbol{A}^-\boldsymbol{A}\boldsymbol{Y}\boldsymbol{B}\boldsymbol{B}^-$ requiring that the *Ben-Israel and Geville* condition be satisfied [12], where the $K \times K$ dimensional matrix $\boldsymbol{Y}$ is arbitrary and $\boldsymbol{A}^-$ denotes the inner inverse of $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. Moreover, both solutions turn out to be no orthogonal and cannot be directly applied. Then, the *Gram-Schmidt* orthogonalization procedure is, perhaps, the most intuitive solution. But, despite the orthogonal condition be satisfied, however, the

original problem remains unsolved, i.e, there is no a solution that satisfies the relaxed problem. Instead, we propose a solution based on Lagrange multipliers, as follows.

The Lagrangian corresponding to the problem (6) is written as follows:

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{R}) = f(\alpha, \boldsymbol{R}) + \sum_{k=1}^{K} \lambda_k g_k(\alpha, \boldsymbol{R}) \tag{7}$$

where

$$f(\alpha, \boldsymbol{R}) = \text{ tr}\left((\boldsymbol{A}\widehat{\boldsymbol{R}}\boldsymbol{B} - \boldsymbol{C})^\top (\boldsymbol{A}\widehat{\boldsymbol{R}}\boldsymbol{B} - \boldsymbol{C})\right)$$

$$\sum_{k=1}^{K} \lambda_k g_k(\alpha, \boldsymbol{R}) = \text{ tr}\left((\boldsymbol{R}^\intercal \boldsymbol{R} - \boldsymbol{I}_K)^\top (\boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{I}_K)\boldsymbol{\Delta}\right)$$

Then, by solving $\frac{\partial f}{\partial \boldsymbol{R}} = 0$ and $\frac{\partial \sum_{k=1}^{K} \lambda_k g_k(\alpha, \boldsymbol{R})}{\partial \boldsymbol{R}} = \text{ tr}(\frac{\partial f}{\partial \boldsymbol{R}})$, we obtain:

$$\boldsymbol{R} = 2\alpha \boldsymbol{A}^\top \boldsymbol{C}\boldsymbol{B}^\dagger - \boldsymbol{I}_K \tag{8a}$$

$$\boldsymbol{\lambda} = \frac{1}{2} \text{ diag}\left((\boldsymbol{R} + (\alpha - 1)\boldsymbol{I}_K)^{-1} \frac{\partial f}{\partial \boldsymbol{R}}\right) \tag{8b}$$

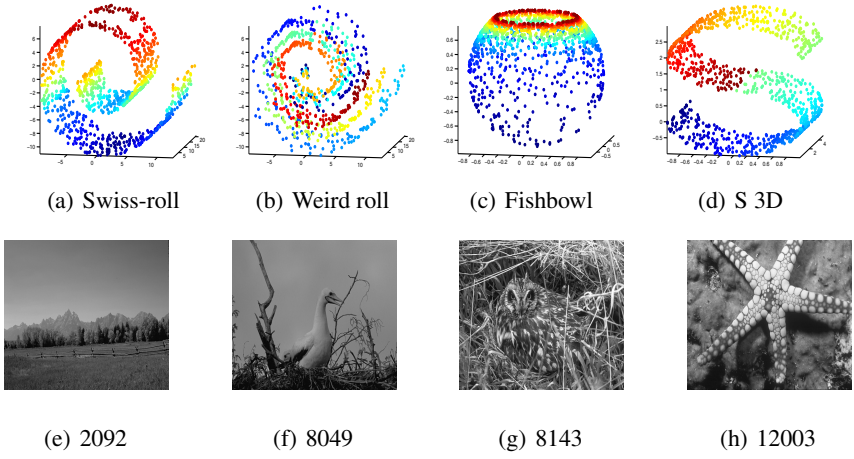
where $\boldsymbol{\Delta} = \text{ Diag}(\boldsymbol{\lambda})$, $\boldsymbol{B}^\dagger$ denotes the pseudo-inverse matrix of $\boldsymbol{B}$ and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. At the end, a continuous estimation of matrix $\boldsymbol{M}$ can be written as $\widehat{\boldsymbol{M}} = \boldsymbol{D}^{1/2}\boldsymbol{L} = \boldsymbol{D}^{1/2}\boldsymbol{V}_K\widehat{\boldsymbol{R}}$, which is further binarized to determine the initial clustering problem solution $\boldsymbol{M}$. Parameter α is chosen at the minimal point of $\mathcal{L}$, i.e., where both f and g are minimum.

3 Experimental Setup

Experiments are carried out on two well-known database collections: Firstly, a toy data comprising the following several data sets (`Swiss-roll`, `weird roll`, `fishbowl`, and `S-3D`) shown in upper row of Fig. 1). Secondly, an image collection extracted from the free access Berkeley Segmentation Dataset [11]. In this work, we considered the first 100 train images from 2092 until 66039 (in ascendant order). In bottom row of Fig. 1, some samples from image database are shown. All considered images are size scaled at 20% and characterized per pixel by means of standard and normalized RGB color space, and XY position. Estimation of the number of groups, k, is based on calculation of the eigenvector set of the affinity matrix. In particular, the scaled exponential affinity matrix $\boldsymbol{W} = \{w_{ij}\}$ is employed that holds elements defined as follows [5]:

$$w_{ij} = \begin{cases} \exp(-\frac{\mathrm{d}^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sigma_i \sigma_j}), & i \neq j \\ 0, & i = j \end{cases} \tag{9}$$

where $\boldsymbol{X} \in \Re^{N \times p} = (\boldsymbol{x}_1^\intercal, \ldots, \boldsymbol{x}_n^\intercal)^\intercal$ is the data matrix to be clustered, $\boldsymbol{x}_i \in \Re^p$ is its corresponding i-th data point, $\sigma_i = \mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_n)$, $\boldsymbol{x}_n$ denotes the n-th nearest neighbor, and $d(\cdot, \cdot)$ stands for Euclidean distance. The value of n is experimentally set to be 7.



(a) Swiss-roll (b) Weird roll (c) Fishbowl (d) S 3D

(e) 2092 (f) 8049 (g) 8143 (h) 12003

Fig. 1. Employed database collections for testing of discussed approach of multi-class spectral clustering. In upper row, exemplary of toy data comprising data sets (`Swiss-roll`, `weird roll`, `fishbowl`, and `S-3D`). In the bottom row, some numbered samples of image database.

Performance of discussed method is assessed by means of three considered clustering indicators: clustering performance, estimated number of groups, and the introduced cluster coherence as an unsupervised measure. The last measure is inferred from the optimization problem given in (1a), and its maximal value is 1, due to its normalization with respect to the affinity matrix degree. Then, after selecting a proper affinity matrix, cluster coherence measure indicates an adequate clustering whenever its value is close to 1. In addition, because of its descent monotonicity property, this measure automatically penalizes the group number. Table 1 shows the clustering performance measures with their description.

Table 1. Applied performance measures

Measure	Description
Clustering performance CP	Complement of standard error (e). $CP = 100 - e$
Cluster coherence $\varepsilon_{\boldsymbol{M}}$	It is near to 1 when clustering is properly done. $\varepsilon_{\boldsymbol{M}} = \frac{1}{k}\sum_{l=1}^{k} \frac{\boldsymbol{M}_l^{\mathsf{T}}\boldsymbol{W}\boldsymbol{M}_l}{\boldsymbol{M}_l^{\mathsf{T}}\boldsymbol{D}\boldsymbol{M}_l}$
Estimated number of groups ($\hat{k}$)	Eigenvectors-based estimation [5]

Lastly, testing within experimental framework is carried out by employing `MATLAB` Version 7.10.0.499 (R2010a) in a standard PC Intel(R) Core 2 Quad 2.8 GHz and 4 Gb RAM memory.

4 Results and Discussion

Table 2 shows the numerical results obtained for both toy data sets and image database. In general, we can note that the MCSC works significantly better than the conventional partitional clustering, showing the benefit of the spectral methods when clusters are not linearly separable. This fact can be appreciated in Fig. 2.

Table 2. Results for toy data sets

Method	Toy data sets			Image database		
	CP $(\mu - \sigma)$	ε_M $(\mu - \sigma)$	$\hat{k}$ $(\mu - \sigma)$	CP $(\mu - \sigma)$	ε_M $(\mu - \sigma)$	$\hat{k}$ $(\mu - \sigma)$
K-means	63.25 – 9.07	0.58 – 0.13		59.25 – 10.55	0.54 – 0.09	
MCSC	89.50 – 4.21	0.86 – 0.05	5 – 0.47	68.62 – 6.51	0.76 – 0.02	8 – 0.75
Improved MCSC	89.50 – 3.67	0.90 – 0.02		70.37 – 8.09	0.78 – 0.04	

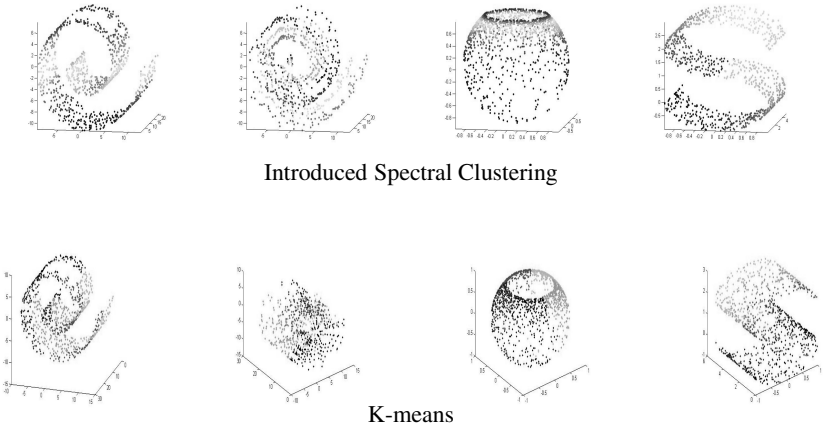
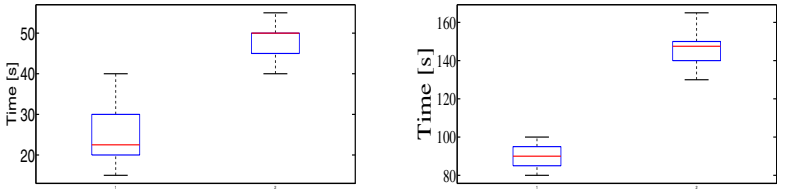


Fig. 2. Clustering results after testing of considered databases

Then, for some applications it might be of benefit to increase the computational burden to improve significatively the performance. Nonetheless, to overcome this issue, we introduce a free iterative algorithm approach, in which instead of applying a complex or time-consuming procedures, we only need to determine parameter α for calculating the indicator binary matrix. Therefore, required time for estimating α becomes considerably lower than that one needed to binarize the clustering solution iteratively, as shown in Fig. 3. Computational time reduces since tuning of parameter α is, mostly, carried out by means of an heuristical search having inexpensive computational burden. In contrast, the conventional MCSC involves calculation of eigenvalues and eigenvector per iteration; both procedures being high time consuming.



(a) Toy data sets (b) Image database

Fig. 3. Box plots of time employed for clustering methods. Improved MCSC at left hand and classical MCSC at right hand.

For the Swiss-roll toy data set, Fig. 4 shows the penalization effect of measure ε_M when varying the group number. Testing is carried out computing the value ε_M for 10 iterations of the whole clustering procedure. As seen, the error bar corresponding to conventional is higher that the error achieved for the proposed MCSC method.

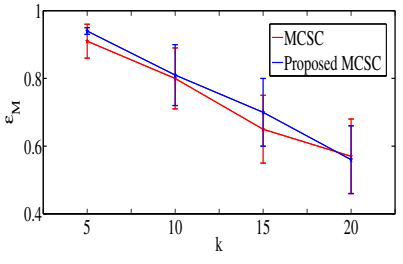


Fig. 4. Error bar comparing ε_M with the group number

5 Conclusions

This work introduces an improved multi-class spectral clustering method that is based on normalized cuts. Performance of discussed method is assessed by means of three considered clustering indicators: clustering performance, estimated number of groups, and the introduced cluster coherence. In terms of considered clustering indicators, the improved multi-class spectral clustering method exhibits a similar performance for tested databases including not linearly separable classes, because it employs the spectral information given by data and their transformations. Nonetheless, discussed method overperforms the conventional spectral methods, in terms of computational burden, since tuning of parameter α is, mostly, carried out by means of an heuristical search having inexpensive computational burden. In contrast, the conventional MCSC involves calculation of eigenvalues and eigenvector per iteration; both procedures being high time consuming.

Also, we introduced a non-supervised measure, associated with cluster coherence, that is inferred from a partition criterion, which showed to be a proper performance index. The cluster coherence measure, automatically, penalizes the number of groups

and generates a value close to 1, whenever the grouping is rightly performed and the affinity matrix is properly chosen.

As a future work, another properties of spectral clustering algorithms should be explored to develop a method less sensitive to initialization, which enhances the trade-off between performance and computational cost. Such method should include proper initialization, estimation of number of groups, feature selection and grouping stages based on spectral analysis.

Acknowledgments. This research is carried out within the projects: "Beca para estudiantes sobresalientes de posgrado de la Universidad Nacional de Colombia" and "Programa de financiación para Doctorados Nacionales de Colciencias".

References

1. Yu, S.X., Jianbo, S.: Multiclass spectral clustering. In: ICCV 2003: Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 313. IEEE Computer Society, Washington, DC (2003)
2. Alzate, C., Suykens, J.A.K.: Multiway spectral clustering with out-of-sample extensions through weighted kernel pca. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(2), 335–347 (2010)
3. Suykens, J.A.K., Alzate, C., Pelckmans, K.: Primal and dual model representations in kernel-based learning. Statistics Surveys 4, 148–183 (2010)
4. Dhillon, I., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–556. ACM (2004)
5. Zelnik-manor, L., Perona, P.: Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems 17, pp. 1601–1608. MIT Press (2004)
6. O'Callaghan, R.J., Bull, D.R.: Combined morphological-spectral unsupervised image segmentation. IEEE Transactions on Image Processing 14(1), 49–62 (2005)
7. Tung, F., Wong, A., Clausi, D.A.: Enabling scalable spectral clustering for image segmentation. Pattern Recognition 43(12), 4069–4076 (2010)
8. Lee, S., Hayes, M.: Properties of the singular value decomposition for efficient data clustering. IEEE Signal Processing Letters 11(11), 862–866 (2004), doi:10.1109/LSP.2004.833513
9. Álvarez-Meza, A., Valencia-Aguirre, J., Daza-Santacoloma, G., Castellanos-Domínguez, G.: Global and local choice of the number of nearest neighbors in locally linear embedding. Pattern Recognition Letters 32(16), 2171–2177 (2011)
10. Cvetkovic-Ilic, D.S.: Re-nnd solutions of the matrix equation AXB= C. Journal of the Australian Mathematical Society 84(1), 63–72 (2008)
11. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision, vol. 2, pp. 416–423 (July 2001)
12. Ben-Israel, A., Greville, T.N.E.: Generalized inverses: theory and applications, vol. 15. Springer (2003)