ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Optimal Data Projection for Kernel Spectral Clustering

D. H. Peluffo[1], C. Alzate[2], J. A. K. Suykens[3]*, and G. Castellanos-Dominguez[4]

1- Université catholique de Louvain,
Machine Learning Group - ICTEAM
2- IBM Research - Ireland
3- Katholieke Universiteit Leuven, ESAT-STADIUS
4- Universidad Nacional de Colombia,
Signal Processing and Recognition Group

Abstract. Spectral clustering has taken an important place in the context of pattern recognition, being a good alternative to solve problems with non-linearly separable groups. Because of its unsupervised nature, clustering methods are often parametric, requiring then some initial parameters. Thus, clustering performance is greatly dependent on the selection of those initial parameters. Furthermore, tuning such parameters is not an easy task when the initial data representation is not adequate. Here, we propose a new projection for input data to improve the cluster identification within a kernel spectral clustering framework. The proposed projection is done from a feature extraction formulation, in which a generalized distance involving the kernel matrix is used. Data projection shows to be useful for improving the performance of kernel spectral clustering.

1 Introduction

Spectral clustering is a suitable technique to deal with grouping problems involving unlabeled-data, especially when clusters are hardly separable. Many approaches have been proposed, ranging from the basic methods that include binary cluster indicators heuristically resulting from a normalized cut-based formulation [1] to the more elaborated kernel-based methods employing least squares-support vector machines (LS-SVM) [2]. In particular, kernel methods are of great interest since they allow to incorporate prior knowledge into the clustering procedure [3]. Due to its unsupervised nature, clustering is very often a parametric procedure, and then a set of initial parameters should be properly selected to avoid any local optimum solution. Typically, the initial parameters are the kernel or similarity matrix and the number of groups [4]. Nonetheless, in some problems when data are represented in a high-dimensional space and/or data-sets are complex and linearly non-separable, a proper feature extraction may be an advisable alternative [5]. In particular, a projection generated by a proper feature extraction procedure may provide a new feature space wherein the clustering procedure can reach more accurate cluster indicators. In other words, data projection accomplishes a new representation space, where the clustering can be improved, in terms of a given mapping criterion, rather than performing the clustering procedure directly over the original input data.

*Johan Suykens acknowledges support by Research Council KUL, ERC AdG A-DATADRIVE-B, GOA/10/09MaNet, CoE EF/05/006, FWO G.0588.09, G.0377.12, SBO POM, IUAP P6/04 DYSCO.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The present work introduces a projection focusing on a better analysis of the structure of data that is devised for a concrete method, namely, the Kernel Spectral Clustering (KSC) [2]. Since data projection can be seen as a feature extraction process, we propose the M-inner product-based data projection [6], in which the similarity matrix is also considered within the projection framework, similarly as discussed in [7]. There are two main reasons for using data projection to improve the performance of kernel spectral clustering: firstly, the data global structure is taken into account during the projection process and, secondly, the kernel method exploits the information of local structures. The proposed method, termed Projected Kernel Spectral Clustering (PKSC) is compared with the baseline KSC, as well as multi-cluster spectral clustering [1], kernel k-means and min-cuts [3]. Clustering performance is tested on images extracted from the free access Berkeley Segmentation Data Set [8] as well as data sets from the UCI repository [9].

This paper is organized as follows: Section 2 describes the KSC method. In section 3, we introduce the data projection for KSC. In section 4 are shown some experimental results and discussion. Finally, section 5 holds the conclusions.

2 Kernel Spectral Clustering

The aim of clustering is to split an input data matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, such that $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top]$, into K disjoint subsets, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th d dimensional data point, N is the number of data points, and K is the number of desired groups. Employed method, herein termed *Kernel Spectral Clustering* (KSC) [2], is based on a weighted kernel principal component analysis (WKPCA) interpretation of spectral clustering with primal-dual least-squares SVM formulations, for which the following vector clustering model is introduced: Let $\boldsymbol{e}^{(l)} \in \mathbb{R}^N$ be the l-th projection vector, which is assumed in the following latent variable form: $\boldsymbol{e}^{(l)} = \boldsymbol{\Phi}\boldsymbol{w}^{(l)} + b_l\mathbf{1}_N$, $l \in \{1, \ldots, n_e\}$, where $\boldsymbol{w}^{(l)} \in \mathbb{R}^{d_h}$ is the l-th weighting vector, b_l is a bias term, and n_e is the number of considered latent variables. Notation $\mathbf{1}_N$ stands for a N-dimensional all-ones vector, and the matrix $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1)^\top, \ldots, \phi(\boldsymbol{x}_N)^\top]$, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times d_h}$, is a high dimensional representation of data. The function $\phi(\cdot)$ maps data from the original dimension to a higher one d_h ($\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_h}$). Therefore, $\boldsymbol{e}^{(l)}$ represents the latent variables from a set of n_e binary cluster indicators obtained with $\text{sign}(\boldsymbol{e}^{(l)})$, which are encoded to obtain the K resulting groups. Grounded on the least-squares SVM formulation of the model, the following optimization problem can be stated:

$$\max_{\boldsymbol{E},\boldsymbol{W},\boldsymbol{b}} \quad \frac{1}{2N}\,\text{tr}(\boldsymbol{E}^\top\boldsymbol{V}\boldsymbol{E}\boldsymbol{\Gamma}) - \frac{1}{2}\,\text{tr}(\boldsymbol{W}^\top\boldsymbol{W}) \quad \text{s. t.} \quad \boldsymbol{E} = \boldsymbol{\Phi}\boldsymbol{W} + \mathbf{1}_N \otimes \boldsymbol{b}^\top, \qquad (1)$$

where $\boldsymbol{E} = [\boldsymbol{e}^{(1)}, \cdots, \boldsymbol{e}^{(n_e)}]$, $\boldsymbol{E} \in \mathbb{R}^{N\times n_e}$, $\boldsymbol{V} \in \mathbb{R}^{N\times N}$ is the weight matrix for projections, $\boldsymbol{\Gamma} = \text{Diag}([\gamma_1, \ldots, \gamma_{n_e}])$, $\gamma_l \in \mathbb{R}^+$ is the l-th introduced regularization parameter, $\boldsymbol{W} = [\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(n_e)}]$, $\boldsymbol{W} \in \mathbb{R}^{d_h\times n_e}$, and $\boldsymbol{b} = [b_1, \ldots, b_{n_e}]$, $\boldsymbol{b} \in \mathbb{R}^{n_e}$. Notations $\text{tr}(\cdot)$ and $\otimes$ denote the trace and the Kronecker product, respectively. Also, taking into account that the kernel matrix represents the similarity matrix of a graph with K connected components as well as $\boldsymbol{V} = \boldsymbol{D}^{-1}$ where $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ is the degree matrix defined as $\boldsymbol{D} = \text{Diag}(\boldsymbol{\Omega}\mathbf{1}_N)$; then the $K-1$ eigenvectors contained in $\boldsymbol{A}$, associated to the largest

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

eigenvalues, are piecewise constant and become indicators of the corresponding connected parts of the graph. Therefore, value n_e is fixed to be $k-1$ [2]. With the aim of achieving a dual formulation, but satisfying the condition $\boldsymbol{b}^\top \mathbf{1}_N = 0$ by centering vector $\boldsymbol{b}$ (i.e. with zero mean), the bias term is in the form $b_l = -1/(\mathbf{1}_N^\top \boldsymbol{V} \mathbf{1}_N)\mathbf{1}_N^\top \boldsymbol{V} \boldsymbol{\Omega} \boldsymbol{\alpha}^{(l)}$. Thus, the solution of problem of Eq. (1), via Karush-Kuhn-Tucker (KKT) conditions over its Lagrangian, is reduced to the following eigenvector-related problem: $\boldsymbol{A}\boldsymbol{\Lambda} = \boldsymbol{V}\boldsymbol{H}\boldsymbol{\Omega}\boldsymbol{A}$, where matrix $\boldsymbol{H} \in \mathbb{R}^{N\times N}$ is the centering matrix that is defined as $\boldsymbol{H} = \boldsymbol{I}_N - 1/(\mathbf{1}_N^\top \boldsymbol{V} \mathbf{1}_N)\mathbf{1}_N\mathbf{1}_N^\top \boldsymbol{V}$, ($\boldsymbol{I}_N$ denotes a N-dimensional identity matrix) and $\boldsymbol{\Omega} = [\Omega_{ij}]$, $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$, being $\Omega_{ij} = \mathsf{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i), i, j \in 1, \ldots, N$. Notation $\mathsf{K}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ stands for the introduced kernel function. As a result, the set of projections can be calculated as follows: $\boldsymbol{E} = \boldsymbol{\Omega}\boldsymbol{A} + \mathbf{1}_N \otimes \boldsymbol{b}^\top$. Once the projections are calculated, we proceed to carry out the cluster assignment by following an encoding procedure applied on the same projections. Because each cluster is represented by a single point in the $K-1$-dimensional eigenspace, such that those single points are always in different orthants due also to the KKT conditions, we can encode the eigenvectors considering that two points belong to the same cluster if they are in the same orthant in the corresponding eigenspace [2]. Then, a codebook can be obtained from the rows of the matrix containing the $K-1$ binarized leading eigenvectors in the columns, by using $\mathrm{sgn}(\boldsymbol{e}^{(l)})$. Then, matrix $\widetilde{\boldsymbol{E}} = \mathrm{sgn}(\boldsymbol{E})$ becomes the codebook being each row a codeword.

3 Data Projection improving Kernel Spectral Clustering

The proposed improving data projection is summarized as follows: Given both an input data matrix $\boldsymbol{X}$, as well as any orthonormal rotation matrix, $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_d$, $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$, then, we can introduce a linear data projection, ruled by the expression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Q}$, where $\boldsymbol{Y} \in \mathbb{R}^{N\times d}$. Furthermore, to accomplish the dimension reduction, we consider the approximated rotation matrix $\widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} = \boldsymbol{I}_p$, $\widehat{\boldsymbol{Q}} \in \mathbb{R}^{d\times p}$, which is a truncated representation of $\boldsymbol{Q}$, where $p < d$. Likewise, a truncated linearly projected data $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{N\times p}$ is introduced, such that $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{Q}}$. Consequently, we can yield an expression for the reconstructed data matrix $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{Q}}^\top$, $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N\times d}$. Since $\widehat{\boldsymbol{Q}}$ is p-dimensional, $\widehat{\boldsymbol{X}}$ becomes a lower rank matrix representing the original data $\boldsymbol{X}$. In order to obtain a rotation matrix $\widehat{\boldsymbol{Q}}$, such that $\widehat{\boldsymbol{Y}}$ holds the projected vectors mostly contributing to the explained variance regarding matrix $\boldsymbol{\Sigma}$, and using the M-inner [6] norm as a distance measure to quantify the quality of provided data projection, it is possible to device the following optimization problem:

$$\min_{\widehat{\boldsymbol{Q}}} \|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2_{\boldsymbol{\Sigma}} = \max_{\widehat{\boldsymbol{Q}}} \ \mathrm{tr}(\widehat{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{\Sigma} \boldsymbol{X} \widehat{\boldsymbol{Q}}) \quad \text{s. t.} \quad \widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} = \boldsymbol{I}_d, \tag{2}$$

where $\|\boldsymbol{X}\|^2_{\boldsymbol{\Sigma}}$ denotes the squared M-norm of $\boldsymbol{X}$ regarding any positive semi-definite matrix $\boldsymbol{\Sigma}$, such that it holds that $\|\boldsymbol{X}\|^2_{\boldsymbol{\Sigma}} = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{\Sigma} \boldsymbol{X})$. Previous formulation given by Eq. (2) takes place, since the following expression holds [7]: $\|\boldsymbol{X}\|^2_{\boldsymbol{\Sigma}} = \|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2_{\boldsymbol{\Sigma}} + \mathrm{tr}(\widehat{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{\Sigma} \boldsymbol{X} \widehat{\boldsymbol{Q}})$. Then, because $\|\boldsymbol{X}\|^2_{\boldsymbol{\Sigma}}$ remains constant, the aim of minimizing $\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2_{\boldsymbol{\Sigma}}$ and that of maximizing $\mathrm{tr}(\widehat{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{\Sigma} \boldsymbol{X} \widehat{\boldsymbol{Q}})$ can be reached simultaneously. By design, to incorporate the information given by the assumed similarity into the data

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

projection process, we employ the kernel matrix $\boldsymbol{\Omega}$ that is given as positive semidefinite matrix, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$. Next, by considering the maximization problem in Eq. (2), we can write its Lagrangian as $\mathsf{L}(\widehat{\boldsymbol{Q}}, \boldsymbol{\Delta}) = \text{tr}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}) - \text{tr}(\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} - \boldsymbol{I}_d))$. Then, equating the partial derivatives in the form: $\partial/\partial\widehat{\boldsymbol{Q}}\ \text{tr}(\widehat{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \widehat{\boldsymbol{Q}}) = \partial/\partial\widehat{\boldsymbol{Q}}\ \text{tr}(\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} - \boldsymbol{I}_d))$, we device the following dual problem: $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \boldsymbol{Q} = \boldsymbol{Q} \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} = \text{Diag}(\boldsymbol{\delta})$, $\boldsymbol{\Delta} \in \mathbb{R}^{d \times d}$, and vector $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_d]$ holds the Lagrangian multipliers. Then, we can infer that a feasible solution of this problem can be accomplished by selecting the Lagrange multipliers as the eigenvalues, as well as, making the matrix $\widehat{\boldsymbol{Q}}$ as the largest p eigenvectors of $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$. Dimension p can be established by means of the well-known explained variance criterion. Finally, the output projected data can be computed as $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{Q}}$. Assuming $\widehat{\boldsymbol{y}}_i$ as the i-th row vector of $\widehat{\boldsymbol{Y}}$, the projections are: $\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{\Omega}}\boldsymbol{A} + \boldsymbol{1}_N \otimes \boldsymbol{b}^\top$, being $\widehat{\Omega}_{ij} = \mathsf{K}(\widehat{\boldsymbol{y}}_i, \widehat{\boldsymbol{y}}_j)$. Again, a subsequent encoding process is needed to determine the cluster assignments as explained in previous section.

4 Results and discussion

The proposed PKSC method is compared with KSC [2], kernel k-means (KKM) [4], min cuts (Min-cuts) [10] and multi-cluster spectral clustering (MCSC) [1]. They are performed over the same conditions: kernel matrix and number of clusters. In case of PKSC, another aspect to take into consideration is the estimation of the value p. Here, we use an accumulated variance by setting a m% of accumulated variance to be captured by the p selected eigenvectors. Methods are assessed regarding image segmentation performance as shown in Fig. 1. The segmentation performance is quantified by a supervised index noted as Probabilistic Rand Index (PR) explained in [11], such that $PR \in [0, 1]$, being 1 when regions are properly segmented. Images are drawn from the free access Berkeley Segmentation Data Set [8]. To represent each image as a data matrix, we characterize the images by color spaces (RGB, YCbCr, LAbB, LUV) and the xy position of each pixel. To run the experiment, we resize the images at 20% of the original size due to memory usage restrictions. All the methods are performed with a given number of clusters K manually set as shown in shown in Fig. 1 and using the scaled exponential similarity matrix as described in [4], setting the number of neighbors to be 9. To determine p, we set $m = 95\%$. Also, another real databases collection is considered that is taken from the UCI repository [9]. For quantitative evaluation of compared clustering methods in terms of performance and stability, the estimated mean value of considered measures are shown in Table 1, which are computed after running algorithms 50 times. Methods are performed by setting the number of groups as the original number of classes. Again, a scaled similarity matrix is used with the number of neighbors equals to 15. In this case, for PKSC, dimension p is fixed by setting $m = 99\%$. We use two well-known clustering measures: Fisher's criterion (J) and Silhouette (S).

Proposed data projection transforms data from the original d-dimensional feature space into a reduced p-dimensional one. However, since the cluster structure is the most important characteristic describing the input data, the main goal of data projection is to find a lower dimensionality representation maximally preserving the original cluster structure. This work introduces a data projection that focuses on the analysis of the local

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

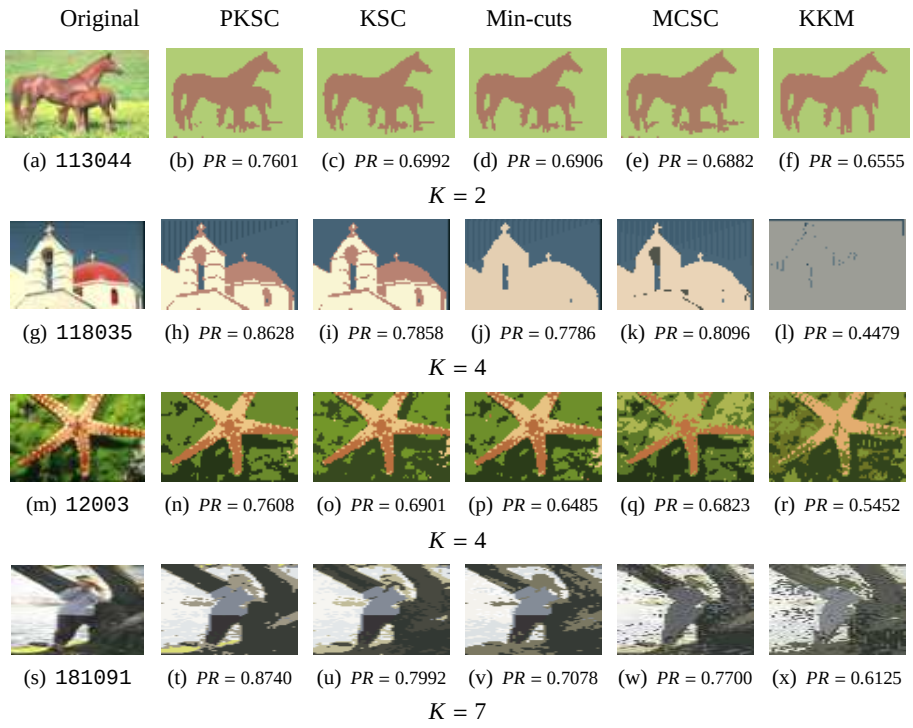

Fig. 1: Clustering performance on image segmentation along 10 iterations

data structure to improve the performance of KSC clustering method. Proposed scheme improves the performance of kernel spectral clustering since, firstly, the data global structure is taken into account in the projection process and, secondly, the kernel method exploit the local structure information. Due to these properties, our method outperforms the remaining considered methods in terms of the considered clustering performance measures. As can be noticed, our method works well on image segmentation, which means that complex data can be rightly modeled by PKSC. In addition, since global structure is also considered, PKSC is also able to deal with real databases where some compactness is guaranteed. Then, our approach is a more flexible and versatile method.

5 Conclusions

This paper proposes a new data projection to improve the performance of clustering, in the concrete case, the Least-squares SVM-based Kernel Spectral Clustering is considered. Proposed data projection consists of a linear mapping based on the M-inner product approach, for which an orthonormal eigenvector basis is computed as the projection matrix. Moreover, the used projection matrix is computed over the spectrum of a weighted covariance matrix involving the information given by the similarity matrix. The strength of our approach is that local similarities and global structure are

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Data set	Measure	Method				
		Min-cuts	MCSC	KKM	KSC	PKSC
Iris	J	2.8 ± 0.15	2.01 ± 0.1	2.7 ± 0.12	2.9 ± 0.15	3.55 ± 0.2
	S	0.7 ± 0.009	0.45 ± 0.009	0.55 ± 0.01	0.65 ± 0.011	0.75 ± 0.014
Biomed	J	1.5 ± 0.09	0.95 ± 0.1	1.1 ± 0.11	1.1 ± 0.3	1.6 ± 0.3
	S	0.60 ± 0.01	0.25 ± 0.011	0.45 ± 0.014	0.54 ± 0.09	0.65 ±0.09
Auto mpg	J	0.69 ± 0.15	0.62 ± 0.13	1.04 ± 0.22	0.80 ± 0.11	0.89 ± 0.15
	S	0.35 ±0.009	0.30 ± 0.009	0.35 ± 0.009	0.42 ± 0.009	0.51 ± 0.009
Breast	J	0.85 ± 0.1	0.85 ± 0.12	0.85 ± 0.14	0.85 ± 0.23	1.25 ± 0.26
	S	0.75 ± 0.0091	0.61 ± 0.01	0.78 ± 0.014	0.78 ± 0.011	0.79 ± 0.031
Glass	J	0.54 ± 0.11	0.45 ± 0.12	0.50 ± 0.18	0.53 ± 0.22	0.55 ± 0.19
	S	0.61 ± 0.011	0.41 ± 0.011	0.53 ± 0.012	0.56 ± 0.023	0.61 ± 0.017
Diabetes	J	0.54 ± 0.091	0.45 ± 0.1	0.50 ± 0.11	0.54 ± 0.51	0.55 ± 0.21
	S	0.61 ± 0.0091	0.42 ± 0.0092	0.55 ± 0.011	0.59 ± 0.089	0.61 ± 0.013
Heart	J	0.11 ± 0.26	0.14 ± 0.26	0.14 ± 0.3	0.15 ± 0.54	0.16 ± 0.31
	S	0.32 ± 0.012	0.37 ± 0.012	0.39 ± 0.017	0.32 ± 0.056	0.42 ± 0.021

Table 1: Overall performance for clustering methods over real databases

employed to refine the projection procedure by preserving the most explained variance and reaching a projected space that improves the clustering performance within the studied framework.

As a future research, new optimal projections in terms of different clustering criteria can be considered. New works may focus on determining optimal basis within spectral analysis. As well, other kernels and applications can be considered.

References

[1] Yu Stella X. and Shi Jianbo. Multiclass spectral clustering. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 313, Washington, DC, USA, 2003. IEEE Computer Society.

[2] C. Alzate and J. A. K. Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel pca. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):335–347, 2010.

[3] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1):176–190, 2008.

[4] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.

[5] L. Wolf and S. Bileschi. Combining variable selection with dimensionality reduction. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 801 – 806 vol. 2, june 2005.

[6] Themistocles M Rassias. *Inner product spaces and applications*, volume 376. CRC Press, 1997.

[7] J.L. Rodríguez-Sotelo, D. Peluffo-Ordoñez, D. Cuesta-Frau, and G. Castellanos-Domínguez. Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering. *Computer Methods and Programs in Biomedicine*, 2012.

[8] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 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*, volume 2, pages 416–423, July 2001.

[9] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[10] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, Aug 2000.

[11] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):929–944, 2007.