



Comparative Analysis Between Embedded-Spaces-Based and Kernel-Based Approaches for Interactive Data Representation

C. K. Basante-Villota^{1,2,3,4}, C. M. Ortega-Castillo^{1,2,3,4}(✉),
D. F. Peña-Unigarro^{1,2,3,4}, J. E. Revelo-Fuelagán¹, J. A. Salazar-Castro^{2,3},
and D. H. Peluffo-Ordóñez^{3,4}

¹ Universidad de Nariño, Pasto, Colombia
karlosmaor@hotmail.com

² Universidad Nacional, sede Manizales, Manizales, Colombia

³ Corporación Universitaria Autónoma de Nariño, Pasto, Colombia

⁴ Yachay Tech, Urcuquí, Ecuador

Abstract. This work presents a comparative analysis between the linear combination of embedded spaces resulting from two approaches: (1) The application of dimensional reduction methods (DR) in their standard implementations, and (2) Their corresponding kernel-based approximations. Namely, considered DR methods are: CMDS (Classical Multi-Dimensional Scaling), LE (Laplacian Eigenmaps) and LLE (Locally Linear Embedding). This study aims at determining -through objective criteria- what approach obtains the best performance of DR task for data visualization. The experimental validation was performed using four databases from the UC Irvine Machine Learning Repository. The quality of the obtained embedded spaces is evaluated regarding the $R_{NX}(K)$ criterion. The $R_{NX}(K)$ allows for evaluating the area under the curve, which indicates the performance of the technique in a global or local topology. Additionally, we measure the computational cost for every comparing experiment. A main contribution of this work is the provided discussion on the selection of an interactivity model when mixing DR methods, which is a crucial aspect for information visualization purposes.

Keywords: Artificial intelligence
Dimensionality reduction methods · Kernel · Kernel PCA
CMDS · LLE · LE

1 Introduction

Nowadays, the large volumes of data are accompanied by the need of powerful tools for analysis and representation, as, you could have a dense repository of data, but without the appropriate tools the information obtained may not

be very useful [1]. The need arises to find different techniques and tools that help researchers or analysts in tasks such as obtaining useful patterns for large volumes of data, these tools are the subject of an emerging field of research known as Knowledge Discovery in Bases of Data (KDD). Dimension reduction (DR) is considered within the KDD process as a pre-processing stage because it projects the data to a space where the original data is represented with fewer attributes or characteristics, preserving the greater intrinsic information of the original data to enhance tasks such as data mining and machine learning. For example, in classification tasks knowing the representation of the data as well as knowing whether these have separability characteristics, make easier to engage and interpret by the user [2,3].

We have two method PCA (Principal Component Analysis) and the CMDS (Classical Multi-Dimensional Scaling) which are part of those classic RD methods whose objective is to preserve variance or distance [4]. Recently, the focus of DR methods is based on criteria aimed at preserving the data topology. A topology of this type could be represented in an undirected and weighted graph based on data constructed whose points represent the nodes, and their edge's weights are contained in an affinity and non-negative similarity matrix. This representation is leveraged by methods based on spectral and divergence approaches, for the spectral approach we can represent the weights of the distances in a similarity matrix, such as with the LE (Laplacian Eigenmaps) method [5] and using a matrix of unsymmetrical similarity and focusing on the local structure of the data, the method called LLE (Locally Linear Embedding) arises [6]. There is also the possibility of working on the high-dimensional space with the advantage of greatly enhancing the representation and the embedded data visualization of the original space mapped to the high-dimensional space, from the calculation of the eigen decomposition. An estimate of the inner product (kernel) can be designed based on the function and application which one wants to develop [7], in this work the kernel matrices will represent distance or similarity functions associated with a dimension reduction method.

In this research three spectral dimension reduction methods are considered, trying to encompass different criteria which CMDS, LLE and LE are based on, these are used under two approaches, one of them is the representation of their embedded spaces obtained from their standard algorithms widely explained in [5,6,8], and the second is based on the kernel approaches of the same methods. After obtaining each of the embedded spaces, a linear weighting is performed for combine the different approaches leveraging each of the RD methods, the same is done for the kernel matrices obtained from the approximations of the spectral methods. Subsequently the Kernel PCA technique is applied to reduce the dimension to obtain the embedded space from the combination of the kernel-based approach. The combination of embedded spaces already obtained from the RD methods is not clear and intuitive mathematically, on the other hand, the linear combination of kernel or similarity matrices which are represented in the same infinite space is more intuitive and concise mathematically. Nevertheless, in tasks such as visualization of information, choosing any of the two interaction

methods for dimension reduction is a crucial task on which the representation of the data and also the interpretation by the user will depend, therefore this research proposes the quantitative and qualitative comparison in addition to the demonstration of the previous assumption in order to contribute to machine learning tasks, visualization data, data mining where dimension reduction execute an imperative role, For example, perform tasks of classification of high dimension data, it is necessary to visualize them in such a way that they are understandable for non-expert users who want to know he topology of the data and characteristics such as separability which aid to determine which classifier could be adequate for determinate data record.

2 Methodology

Mathematically, the objective of dimension reduction is to map or project (linear transformation) data from a high-dimensional space $\mathbf{Y} \in \mathbb{R}^{D \times N}$ a low-dimensional space $\mathbf{X} \in \mathbb{R}^{d \times n}$, where $d < D$, therefore, The original data and the embedded data will consist of N points or registers, denoted respectively by $\mathbf{y}_i \in \mathbb{R}^D$ and $\mathbf{X}_i \in \mathbb{R}^d$ with $\{\mathbf{K}^{(1)}, \dots, \mathbf{K}^{(M)}\}$ [5,6]. It means that the number of samples in the high-dimensional data matrix would not be affected when the number of attributes or characteristics is reduced. In order to represent the resulting embedded space in a two-dimensional Cartesian plane, this research takes into account only the two main characteristics in the kernel matrix, which represent most of the information in the original space.

2.1 Kernel Based Approaches

The RD method known as principal component analysis (PCA) is a linear projection that tries to preserve the variance from the values and eigenvectors of the covariance matrix [9, 10]. Moreover, when a data matrix is centered, which means that the average value of the rows (characteristics) is equal to zero, the preservation of variance could be named as a preservation of the Euclidean internal product [9].

Kernel PCA method is as similar as PCA method which maximizes the variance criterion, but in this case of a kernel matrix, which is basically an internal product of an unknown space of high dimension. We define $\phi \in \mathbb{R}^{D \times N}$ a high-dimensional space with $D_h \gg D$, which is completely unknown except for its internal product that can be estimated [9]. To use the properties of this new high-dimensional space and its internal product, it is necessary to define a function $\phi(\cdot)$ that can map the data from the original space to the high-dimension (ϕ) as follows:

$$\begin{aligned} \phi(\cdot) &: \mathbb{R}^D \mathbb{R}^{D_h} \\ \mathbf{y}_i &\Rightarrow \phi(\mathbf{y}_i), \end{aligned} \tag{1}$$

where the i-th vector column of the matrix $\phi = \phi(\mathbf{y}_i)$.

Considering the conditions of Mercer [11], and the matrix \mathbf{f} is centered, the internal product of the kernel function $\mathbf{K}(\cdot, \cdot)$ can be calculated as follows: $\phi(y_i)^T \phi(y_i) = \mathbf{K}(y_i, y_j)$. In short, the kernel function can be understood as a composition of the mapping generated by $\phi(\cdot)$ and its scalar product as follows: $\phi(y_i)^T \phi(y_i)$, so for each pair of elements of the set Y its scalar product is directly assigned without going through the mapping (ϕ). Organizing all possible internal products in a $\mathbf{K}_{N \times N}$ array will result in a kernel matrix:

$$\mathbf{K}_{N \times N} = \boldsymbol{\varphi}^T_{D_h \times N} \boldsymbol{\varphi}_{D_h \times N}. \quad (2)$$

The advantage of working with the high-dimensional space (ϕ) is that it can greatly improve the representation and visualization of the embedded data from the original space mapped to the high-dimensional space, from the calculation of the eigenvalues and eigenvectors of its product internal. An estimation of the internal product (kernel) can be designed based on the function and application that the user wants to develop [12], in this case the kernel matrices will represent distance functions associated with a dimension reduction method, approximations kernels presented below are widely explained in [13]. The kernel representation for the CMDS reduction method is defined as the distance matrix $\mathbf{D} \in \mathbb{R}^{R \times N}$ doubly centered, that is, making the mean of the rows and columns zero, as follows:

$$\mathbf{K}_{CMDS} = -\frac{1}{2}(\mathbf{I}_N - \mathbf{1}_N \mathbf{1}_N^T) \mathbf{D} (\mathbf{I}_N - \mathbf{1}_N \mathbf{1}_N^T), \quad (3)$$

where the ij entry of \mathbf{D} is given by the Euclidean distance:

$$d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2. \quad (4)$$

A kernel for LLE can be approximated from a quadratic form in terms of the matrix \mathbf{W} holding linear coefficients that sum to 1 and optimally reconstruct observed data. Define a matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ as $\mathbf{M} = (\mathbf{I}_N - \mathbf{W})(\mathbf{I}_N - \mathbf{W}^T)$ and λ_{max} as the largest eigenvalue of \mathbf{M} . Kernel matrix for LLE is in the form:

$$\mathbf{K}_{LLE} = \lambda_{max} \mathbf{I}_N - \mathbf{M}. \quad (5)$$

Considering that kernel PCA is a maximization problem in the high-dimensional covariance represented by a kernel, LE can be represented as the pseudo-inverse matrix of the graph \mathbf{L} , as shown in the following expression:

$$\mathbf{K}_{LE} = \mathbf{L}^\dagger, \quad (6)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$, \mathbf{S} , such that \mathbf{S} is a dissimilarity matrix and $\mathbf{D} = \text{Diag}(\mathbf{S} \mathbf{1}_N)$ is the degree matrix is the matrix of the degree of \mathbf{S} . The similarity matrix \mathbf{S} is organized in such a way that the relative width parameter is estimated by maintaining the entropy of the distribution with the nearest neighbor with approximately $\log K$, where K is the given number of neighbors as explained in [14]. For this investigation the number of neighbors was established as the integer closest to 10% of the amount of data.

Finally, to project the data matrix $\mathbf{Y} \in \mathbb{R}^{D \times N}$ into an embedded space $\mathbf{X} \in \mathbb{R}^{d \times N}$ we use the PCA dimension reduction method. In PCA, the embedded space is obtained by selecting the most representative eigenvectors of the covariance matrix [6, 10]. Therefore, we obtain the d most representative eigenvectors of the kernel matrix $\mathbf{K}_{N \times N}$ obtained previously, constructing the embedded space \mathbf{X} . As it was said for this research, the embedded space with two dimensions that represents most of the characteristics of the data is established.

2.2 DR-Methods Mixturing

In terms of data visualization through RD methods, the parameters to be combined are the kernel matrices and the embedded spaces obtained in each method, each matrix corresponds to each of the M RD methods considered, that is $\{\mathbf{K}^{(1)}, \dots, \mathbf{K}^{(M)}\}$. Consequently, a matrix is obtained depending on the kernel approach or final embedded space \mathbf{K} resulting from the mixing of the M matrices, such that:

$$\widehat{\mathbf{K}} = \sum_{m=1}^M \alpha_m \mathbf{K}^{(m)}, \quad (7)$$

Defining α_m as the weighting factor corresponding to the method M and $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_m\}$ as the weighting vector. In this research these parameters will be defined as 0.333 for each of the three methods used, so the sum of the three will be 1 in order to provide to each method equal priority, since the aim of this research is to present a comparison of each proposed approach in a equal conditions scenario, Each $\mathbf{K}^{(M)}$ will represent the kernel matrices obtained after applying the approximations presented in Eqs. (3), (5) and (6) or the embedded spaces obtained by applying the RD methods in their classical algorithm.

3 Results

Data-Sets: Experiments are carried out over four conventional data sets. The first data set (Fig. 1(a)) is an artificial spherical shell ($N = 1500$ data points and $D = 3$). The second data set (Fig. 1(c)) is a toy set here called Swiss roll ($N = 3000$ data points and $D = 3$). The third data set (Fig. 1(d)) is `Coil_20` is a database of gray-scale images of 20 objects. Images of the objects were taken at pose intervals of 5 degrees. This corresponds to 72 images per object ($N = 1440$ data points 20 and $D = 1282$ -number of pixels) [15]. The fourth data set (Fig. 1(b)) is a randomly selected subset of the MNIST image bank [11], which is formed by 6000 gray-level images of each of the 10 digits ($N = 1500$ data points 150 instances for all 10 digits and $D = 242$). Figure 1 depicts examples of the considered data sets.

Performance Measure: In dimensionality reduction, the most significant aspect, which defines why a RD method is more efficiency, is the capability of preserve the data topology in low-dimensional space regarding the high-dimension.

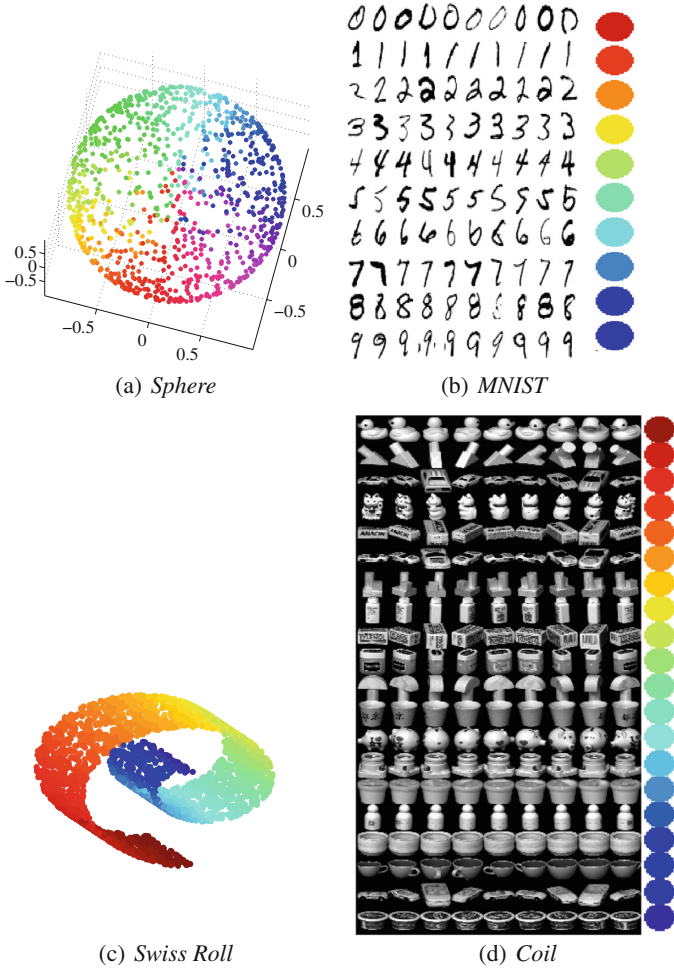


Fig. 1. The fourth considered datasets, source: <https://archive.ics.uci.edu/ml/datasets.html>.

Therefore, we apply a quality criterion used by conserving the k -th closest neighbors developed in [16], as efficiency measure for each approach proposed for the interactive RD methods mixture. This criterion is widely accepted as an adequate unsupervised measure [14, 17], which allows the embedded space to assess in the following way: The rank of ε_j with respect to ε_i in high-dimensional space is denoted as:

$$p_{ij} = |\{k: \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}|. \quad (8)$$

In Eq. (8) $|\cdot|$ denotes the set cardinality. Similarly, in [13] is defined that the range of \mathbf{x}_j with respect to \mathbf{x}_i in the low-dimensional space is:

$$\mathbf{r}_{ij} = |\{\mathbf{k}: \mathbf{d}_{ik} < \mathbf{d}_{ij} \text{ or } (\mathbf{d}_{ik} = \mathbf{d}_{ij} \text{ and } 1 \leq k < j \leq N)\}|. \quad (9)$$

The k -th neighbors of $\zeta \mathbf{i}$ and \mathbf{x}_i are the sets defined by (10) and (11), respectively.

$$\mathbf{v}_i^k = \{\mathbf{j}: \mathbf{1} \geq \mathbf{p}_{ij} < K\}, \quad (10)$$

$$\mathbf{n}_i^k = \{\mathbf{j}: \mathbf{1} \geq \mathbf{r}_{ij} < K\}. \quad (11)$$

A first performance index can be defined as:

$$\mathbf{Q}_{NX}(\mathbf{K}) = \sum_{i=1}^N \frac{|\mathbf{v}_i^k \cap \mathbf{n}_i^k|}{KN} = 1. \quad (12)$$

Equation (12) results in values between 0 and 1 and measures the normalized average according to the corresponding k -th neighbors between the high-dimensional and low-dimensional spaces. Defining in this way a coclassification matrix:

$$[\mathbf{Q} = \mathbf{q}_{NX}] \text{ for } j \geq N - 1, \quad (13)$$

whit $\mathbf{q}_{kl} = |\{(\mathbf{i}, \mathbf{j}): p_{ij} = k \text{ and } p_{ij} = l\}|$.

Therefore $\mathbf{Q}_{NX}(\mathbf{K})$ counts k -by- k blocks of \mathbf{Q} , the range preserved (in the main diagonal) and the permutations within the neighbors (on each side of the diagonal) [12]. This research employs an adjustment of the curve $\mathbf{Q}_{NX}(\mathbf{K})$ introduced in [12] in order that the area under the curve is an adequate indicator of the embedded data topology preservation, hence, the quality curve that is applied into the visualization methodology is given by:

$$\mathbf{R}_{NX}(\mathbf{K}) = \frac{(N - 1)\mathbf{Q}_{NX}(\mathbf{K}) - N}{N - 1 - K}. \quad (14)$$

When the equation in (14) is expressed logarithmically, errors in large neighborhoods are not proportionally as significant as small ones [14]. This logarithmic expression allows obtaining the area under the curve of $\mathbf{R}_{NX}(\mathbf{K})$ given by:

$$\mathbf{AUC} \log_K(\mathbf{R}_{NX}(\mathbf{K})) = \frac{\sum_{K=1}^{N-2} \frac{\mathbf{R}_{NX}(\mathbf{K})}{K}}{\sum_{K=1}^{N-2} \frac{1}{K}}. \quad (15)$$

The results obtained by applying the methodology proposed over four data bases described, are shown in Fig. 2, where the curve $\mathbf{R}_{NX}(\mathbf{K})$ of each approach is presented as well as the \mathbf{AUC} in (13) which assess the dimension reduction quality corresponding to each proposed combination. As a result, for RD procedure in terms of visualization we show the embedded space for each test performed. It is necessary to clarify that each combination was carried out same scenario with equal conditions which allows us to measure a computational cost in terms of execution time, which are shown in Table 1. This is an important

Table 1. Consumed time for performing each approach over the fourth dataset.

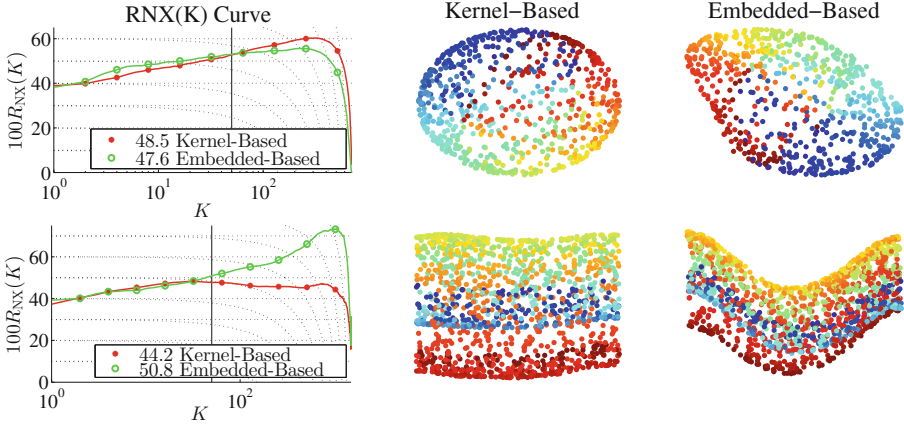
Based approach	Dataset	Computacional time (sec)
Kernel	3D sphere	6, 27
	Swiss Roll	6, 43
	Coil-20	28, 94
	MINST	37, 87
Embedded-spaces	3D sphere	2, 88
	Swiss Roll	3, 09
	Coil-20	15, 24
	MINST	16, 24

issue if users are seeking for an interactive RD methods mixture which has a satisfactory performance, as well as an efficient computational development.

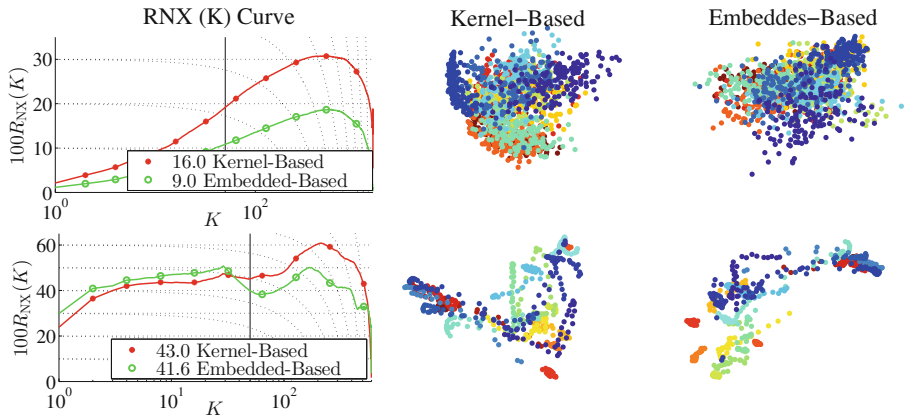
Nevertheless, results achieved in this research allows us to conclude that in data visualization terms performing an interactive mixture RD method based on kernel is more favorable than based on standard methods, mathematically combining a kernel approximations, which means that each kernel approximation is in the same high-dimensional space where all classes are separable before developing the mixture, is more appropriate than combining obtained embedded space from an unknown space which are the standard methods.

The computational cost (Table 1) allows us to infer that the cost in executing kernel approaches and PCA kernel application for dimension reduction is a slightly more elevated in all cases. This is since the databases have a high number of registers, which means that acquiring the kernel matrices involves a lot of processing, as if the data base consists of n samples, the kernel matrix size will be $N \times N$.

Making a comparison of the $R_{NX}(K)$ curves for each database, there is a low performance in the dimension reduction process for the case of the Coil-20 database whose AUC is the lowest among all, which means that the data topology in the embedded space obtained is not as conserved as in the other studied cases. Evidently the best performance was accomplished for 3D spherical shell and Swiss roll which obtained the best AUC and preserve the data local structure, generally preserved local structure generates superior embedded spaces [13]. On the other hand, MNIST and spherical shell database preserved the global data structure in a preferable way as regards the other cases.



(a) Results for datasets: *Sphere 3D* and *Swiss Roll*.



(b) Results for datasets: *MNIST* and *Coil-20*.

Fig. 2. Results obtained for the four experimental databases

4 Conclusion

This work presented a comparative analysis of two different approaches for DRmethods mixing which are applied in an interactive. Results obtained in this research allows us to conclude that performing an interactive DR-methods mixture could be a tough task for a dataset with a great number of points and dimensions as it was proved that the computational cost is higher but also this approach gives to users a high-quality performance since, a greater area is obtained under the quality curve which indicates that the topology of the data can be preserved more. On the other hand, embedded-spaces-based approach has a slightly difference in the $R_{NX}(K)$ AUC curve, but it is not wide so if the user wants to carry out a quicker mixture, the embedded-spaces-based approach

will be more appropriate for data visualization where interactivity is the most important achievement seeking a better perception for the inexperienced users of their datasets.

Acknowledgements. This work is supported by the “Smart Data Analysis Systems - SDAS” group (<http://sdas-group.com>), as well as the “Grupo de Investigación en Ingeniería Eléctrica y Electrónica - GIIEE” from Universidad de Nariño. Also, the authors acknowledge to the research project supported by Agreement No. 095 November 20th, 2014 by VIPRI from Universidad de Nariño.

References

1. Sacha, D., et al.: Visual interaction with dimensionality reduction: a structured literature analysis. *IEEE Trans. Vis. Comput. Graph.* **23**(1), 241–250 (2017)
2. Peluffo Ordoñez, D.H., Lee, J.A., Verleysen, M.: Recent methods for dimensionality reduction: a brief comparative analysis. In: 2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014) (2014)
3. Peluffo-Ordóñez, D.H., Castro-Ospina, A.E., Alvarado-Pérez, J.C., Revelo-Fuelagán, E.J.: Multiple kernel learning for spectral dimensionality reduction. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. LNCS*, vol. 9423, pp. 626–634. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25751-8_75
4. Belanche Muñoz, L.A.: Developments in kernel design. In: ESANN 2013 Proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 24–26 April 2013, pp. 369–378 (2013)
5. Borg, I., Groenen, P.J.: *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York (2005). <https://doi.org/10.1007/0-387-28981-X>
6. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing* **72**(7–9), 1431–1443 (2009)
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)
8. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
9. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 171–177. IEEE (2014)
10. Gijón Gómez, J.: *Visualización bidimensional de problemas de clasificación en alta dimensión*. B.S. thesis (2013)
11. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: *Proceedings of the Twenty-First International Conference on Machine Learning*, p. 47. ACM (2004)
12. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge (2000)
13. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing* **112**, 92–108 (2013)

14. Cook, J., Sutskever, I., Mnih, A., Hinton, G.: Visualizing similarity data with a mixture of maps. In: *Artificial Intelligence and Statistics*, pp. 67–74 (2007)
15. Nene, S.A., Nayar, S.K., Murase, H., et al.: *Columbia object image library (coil-20)* (1996)
16. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *J. Am. Stat. Assoc.* **104**(485), 209–219 (2009)
17. France, S., Carroll, D.: Development of an agreement metric based upon the RAND index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. In: Perner, P. (ed.) *MLDM 2007. LNCS (LNAI)*, vol. 4571, pp. 499–517. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73499-4_38