Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction

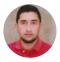
Conference Paper · August 2016

DOI: 10.1109/STSIVA.2016.7743318

Diego Peña
University of Nariño
10 PUBLICATIONS **24** CITATIONS

Jose Alejandro Salazar Castro
University of Nariño
16 PUBLICATIONS **25** CITATIONS

Diego Peluffo
Yachay Tech
169 PUBLICATIONS **359** CITATIONS



Paul Rosero
Universidad Técnica del Norte
44 PUBLICATIONS **46** CITATIONS

Interactive Visualization Methodology of High-Dimensional Data with a Color-Based Model for Dimensionality Reduction

Diego F. Peña-Unigarro,
Jose A. Salazar-Castro
Universidad de Nariño
Pasto, Colombia
Universidad Nacional de Colombia sede Manizales
Manizales,Colombia
diferpun@gmail.com,
alejo26st@udenar.edu.co

Diego H. Peluffo-Ordóñez,
Paul D. Rosero-Montalvo,
Omar R. Oña-Rocha,
Andrés A. Isaza
Universidad Técnica del Norte,
Instituto Tecnologico 17 de Julio
Ibarra,Ecuador
Universidad Surcolombiana,
Universidad Tecnológica
de Pereira
Pereira,Colombia
dhpeluffo@utn.edu.ec,
pdrosero@utn.edu.ec,
oronia@utn.edu.ec,
andres.anaya@usco.edu.co

Juan C. Alvarado-Pérez,
Roberto Theron
Universidad de Salamanca
Salamanca, España
Corporación Universitaria Autónoma de Nariño
Pasto, Colombia
jcalvarado@usal.es,theron@usal.es

Abstract

Nowadays, a consequence of data overload is that world's technology capacity to collect, communicate, and store large volumes of data is increasing faster than human analysis skills. Such an issue has motivated the development of graphic ways to visually represent and analyze high-dimensional data. Particularly, in this work, we propose a graphical interface that allow the combination of dimensionality reduction (DR) methods using a chromatic model to make data visualization more intelligible for humans. This interface is designed for an easy and interactive use, so that input parameters are given by the user via the selection of RGB values inside a given surface. Proposed interface enables (even non-expert) users to intuitively either select a concrete DR method or carry out a mixture of methods. Experimental results proves the usability of our interface making the selection or configuration of a DR-based visualization an intuitive and interactive task for the user.

1. Introduction

The transformation of high-dimensional data into a lower-dimensional version that preserves as much information as possible from the original data is a research area widely studied [17, 18], given its ability to reduce the computational cost and/or improve the performance of both pattern recognition and information visualization systems [12, 13]. In spite of the existence of tools reaching efficiency indicators in terms of computational performance, exploration and representation of high dimensional data, they lack of properties like interactivity and controllability. Therefore, it is required an expert intervention providing prior knowledge to the system for testing DR techniques as well as interpreting their results being no always readily understood [12,15]. In consequence there is a gap between the users knowledge and the database to be analyzed [16, 18]. The reduction of this gap is the premise that this research is based on.

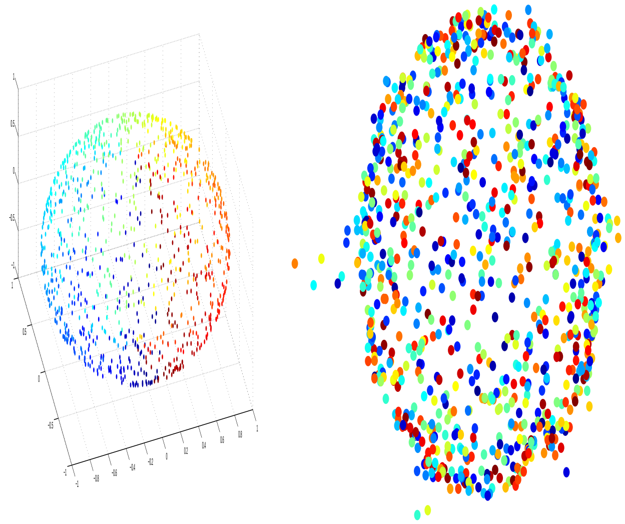
This paper attempts to jointly take advantage of techniques from the field of DR and concepts from the information visualization aiming to enable the user (not necessarily expert) to directly interact with the database. Doing so, users can get an overview of the data in order to draw conclusions and make decisions [16]. This paper presents an intuitive model that allows the combination of three DR methods providing both interactivity and controllability. Proposed model is based on the RGB color space, where every primary color (red (R), green (G),and blue (B)) represents a particular DR method while the whole range of colors derived from the combination will be reflected in the mixture of DR methods. To do so, conventional DR methods are implementing through kernel approximations [10, 11, 15], which are combined to reach a final kernel matrix. Finally, such a kernel matrix feeds a generalized algorithm of kernel principal component analysis (KPCA) [10]. The benefit of this approach is that user may utilize DR methods over the data, even with no knowledge about the theoretical foundations behind them. The user control the results by just exploring an intuitive, color-based interface. This chromatic model uses the color points within a surface, defining the degree or level at which the DR methods (Kernel matrices) are used, that is, the set of weighting factors. Such surface is a superposition of channels to form the full range of colors and a point on the surface is translated into an RGB value, which defines the mixture of the kernels. This approach allows to evaluate visually the behavior of the low-dimensional data regarding the kernel mixture. The chromatic model proposed in this paper is evaluated using three DR methods, namely: locally linear embedding (LLE) [14], multidimensional classical scaling (CMD) [3] and laplacian eigenmaps (LE) [2]. The experiments are performed over two real databases (images of objects - COIL 20 digits - MNIST) and two artificial databases (spherical shell and Swiss roll) [1]. The DR performance is quantified by a scaled version of the average agreement rate between K-ary neighborhoods explained in [8].

The rest of this paper is organized as follows: Section 2 outlines data visualization via dimensionality reduction. Section 3 describes the Proposed color-based model for the combination of DR methods. Experimental setup and results are presented in Sections 4 and 5, respectively. Finally, some final remarks are drawn in section 6.

2. Data visualization via dimensionality reduction

Visualization is the first stage of data analysis where the goal is to make sense of the data before proceeding with others steps like modeling, classification and analysis [18]. Given a large set of measured variables, an obvious idea is to reduce the attributes or features in the measurements by representing them with a smaller set of more condensed variables [16]. Dimensionality reduction allows the extraction of lower dimensional, relevant information from big collections of data aimed at improving the performance of a pattern recognition system or allowing for intelligible data visualization. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$ [12, 13]. In Figure 1 the effect of one DR method is shown.

Classical DR approaches were conceived following an intuitive criterion, such as variance preservation (principal component analysis - PCA) or distance preservation (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aimed at preserving the data topology. Such a topology is often given by a data-related graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian eigenmaps [2]. On the other hand, once normalized, it can also represent a probability distribution. The latter is the case of the methods based on divergences such as stochastic neighbor embedding [13].



(a) 3D sphere (b) Embedded data from sphere

Figure 1. Dimensionality reduction effect over an artificial 3d sphere, when The DR method called LLE is applied.

3. Proposed model for interactive dimensionality reduction using a color-based approach

This Section describes the proposed model, here so-called, color-based or chromatic model is based on RGB color space and enables an interactive combination of three different spectral unsupervised DR methods, for an (even inexperienced) user, allowing the improvement of the data visualization procedure. A suitable and versatile approximation for spectral DR methods are kernel matrices because they make a linear combination feasible [10, 11, 15]. Our approach works as follows: A normalized image can be defined as a matrix array described by the function $I : \mathbb{N}^3 \rightarrow [0, 1]$, where every pair of numbers $x, y : \mathbb{N}^2$ are known as pixels and each value of $I(x, y, c)$ is associated with the pixel intensity (x, y) of channel c [5]. Decompositions are associated with channels c, whose values are between 0 and 1 if they are normalized, where 0 indicates complete absence of that channel (black color) and 1 is related with the maximum intensity (white color) [5]. This model takes advantage of two properties of RGB images: spacial resolution and intensity resolution. On one hand, spacial resolution is defined as the number of pixels that an image contains and can be calculated by the $N_p = m * n$ where m and n are the number of rows and the number of columns, respectively [5]. On the other hand, intensity resolution is the intensity values that each pixel can take. For this model a 8-bit intensity resolution is taken, this means that there are $2^8 - 1 = 255$ intensity values since 0 value is considered [5].

In Figure 2, a two-channel image is considered where the spacial resolution is $m = 256$ rows by $n = 100$ columns, with this in mind it can be seen that m is equal to the intensity resolution, then each row can have 256 different intensity values from 0 to 255, nevertheless each channel has a different direction of change, c_1 has a decreasing change whereas c_2 has a increasing change. If an intensity value is taken from the image, two values of intensity are obtained due that there are two channels. Also if the intensity values are added the result is always equal to 225 (1 if a normalization is made).

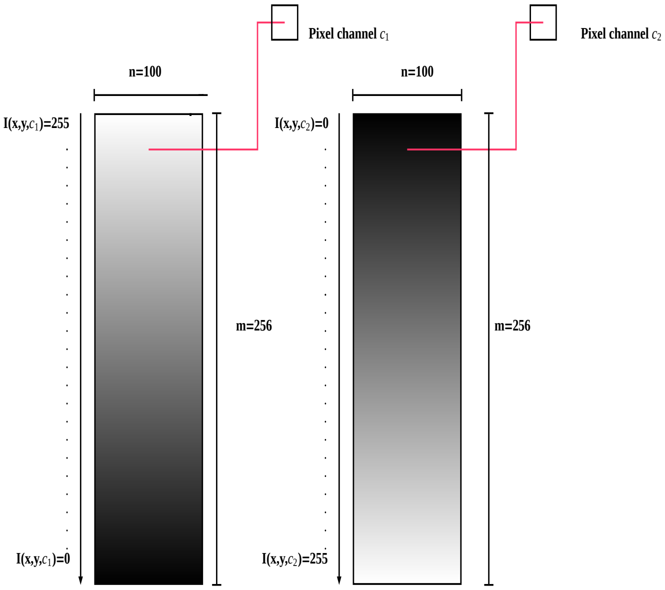


Figure 2. A two channels image.

This paper works with a RGB space, therefore the images have three channels $c_1 = R$, $c_2 = G$, $c_3 = B$ and each channel represents a DR method that in turn has been represented by kernel matrices. Red channel represents the first DR method (DR_1), green channel represents the second DR method (DR_2) and finally blue channel represents the third DR method (DR_3). The proposed interface enables the user to choose multiple combinations of DR methods which will be reflected in the range of colors from the chosen combination. Firstly, a combination of two methods is made with an RGB image which has one of the three channels with intensity value equal to 0 for all pixels, in consequence this image can be considered as a two-channel image and the property explained above can then be applied. For instance, with the combination of two DR methods there are three possible combinations as seen in Figure 3.

Finally, for the combination of three DR methods the same methodology has to be applied, nevertheless three directions of change have to be founded. A triangle was chosen because its three vertexes can contain the primary colors that represent the three different DR methods. In Figure 4a three directions of change can be seen to build the RGB space. However, the sum of the intensity of the three channels (red,green,blue) must be normalized, this means that it must be equal to 1 for any point in the triangle.

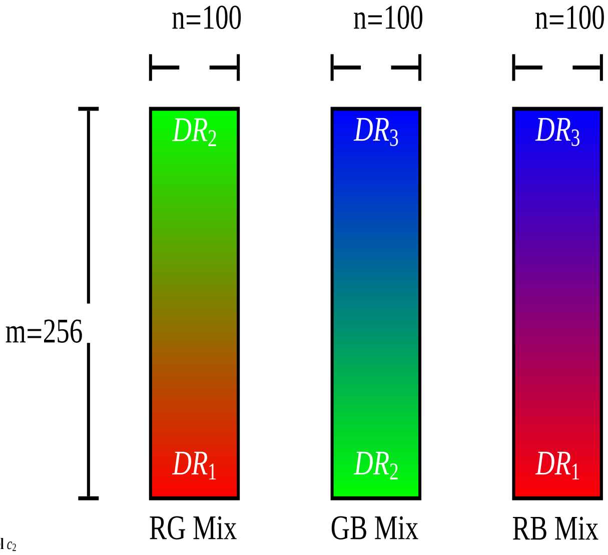
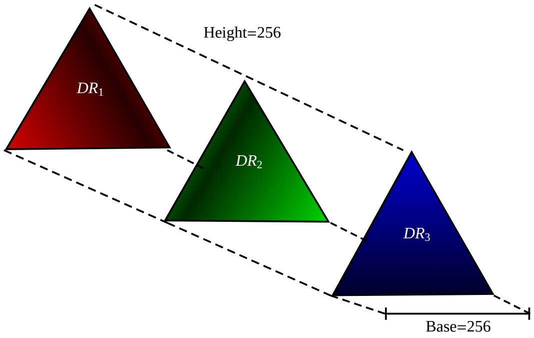
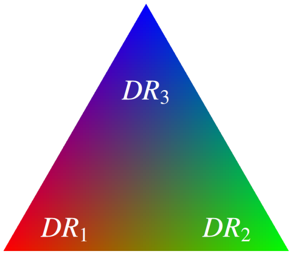


Figure 3. Possible combinations of two DR methods.



(a) Three channels of the RGB space.



(b) Chromatic model.

Figure 4. In (a) the three Chanel of RGB space are represented and the final model in (b) is the superposition of the three channels

For data visualization purposes through DR methods, the terms to be combined are the kernel matrices corresponding to the considered DR methods. Therefore, we obtain a resultant kernel matrix $\widehat{\mathbf{K}}$ as the mixture of M kernel matrices $\{\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(M)}\}$ so:

$$\widehat{\mathbf{K}} = \sum_{m=1}^{m} \alpha_m \mathbf{K}^{(m)} \tag{1}$$

where α_m is the coefficient or weighting factor corresponding to method m and $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_m\}$ is weighting vector. These coefficients are to be associated with intensity values of each pixel inside the image. In this work $m = 3$ and the relationship between the points inside the surface and the coefficients of linear combination are given by the pixel intensity (x, y) inside the color surface (see Figure 4b) associated with RGB channels using (2). An image is a two-dimensional discrete signal this means that α exhibits non-continuous changes. Therefore a resolution can be defined as $\alpha_{res} = \frac{1}{255}$, where α_{res} is the smallest change value of α.

$$\boldsymbol{\alpha} = \{I(x, y, R), (x, y, G), I(x, y, B)\} \tag{2}$$

4. Experimental setup

4.1. Data-sets

Experiments are carried out over three conventional data sets. The first data set is an artificial spherical shell ($N = 1500$ data points and $D = 3$). The second data set is a toy set here called Swiss roll ($N = 3000$ data points and $D = 3$). The third is the `COIL-20` image bank [9], which contains 72 gray-level images representing 20 different objects ($N = 1440$data points –20 objects in 72 poses/angles– with $D = 128^2$). The fourth data set is a randomly selected subset of the `MNIST` image bank [7], 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 $D = 24^2$). Figure 5 depicts examples of the considered data sets.

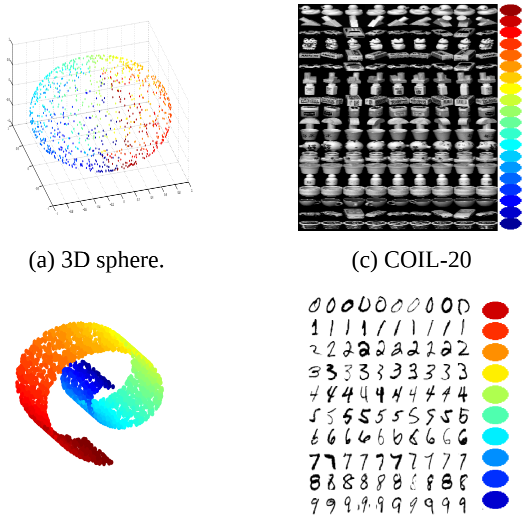
4.2. Methods

Three spectral approaches are considered, namely: classical multidimensional scalling (CMDS) [3], locally linear embedding (LLE) [14], and graph Laplacian eigenmaps (LE) [2]. They are all performed in their standard algorithms. Also, in order to evaluate our framework, kernel approximations are also considered. CMDS kernel is the double centered distance matrix $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ so

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

where the ij entry of $\boldsymbol{D}$ is given by $d_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2$.

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


(a) 3D sphere.

(c) COIL-20

(b) Swiss Roll.

(d) MNIST

Figure 5. The four considered data-sets.

$$\boldsymbol{K}_{LLE} = \lambda_{max} \boldsymbol{I}_N - \boldsymbol{M}. \tag{4}$$

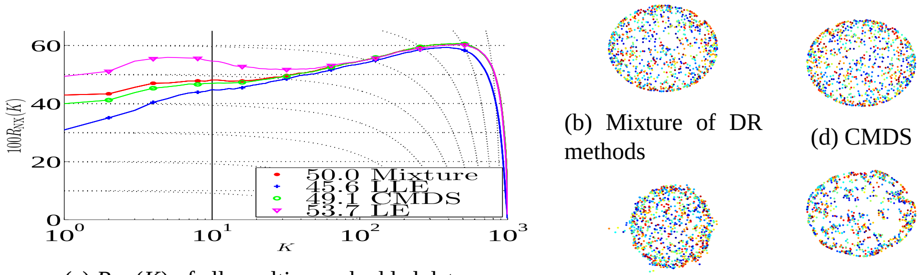
Since kernel PCA is a maximization of the high-dimensional covariance represented by a kernel, LE can be represented as the pseudo-inverse of the graph Laplacian $\boldsymbol{L}$:

$$\boldsymbol{K}_{LE} = \boldsymbol{L}^\dagger, \tag{5}$$

where $\boldsymbol{L} = \boldsymbol{\mathcal{D}} - \boldsymbol{S}$, $\boldsymbol{S}$ is a similarity matrix and $\boldsymbol{\mathcal{D}} = \mathrm{Diag}(\boldsymbol{S}\mathbf{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [6]. The similarity matrices are formed in such a way that the relative bandwidth parameter is estimated keeping the entropy over neighbor distribution as roughly $\log K$ where K is the given number of neighbors as explained in [4]. For all methods, input data is embedded into a 2-dimensional space, then $d = 2$. The number of neighbors is established as $K = 30$ for all considered data sets.

4.3. Quality measures

To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [8] is used, which is ranged within the interval $[0, 1]$. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.



(a) $R_{NX}(K)$ of all resulting embedded data.

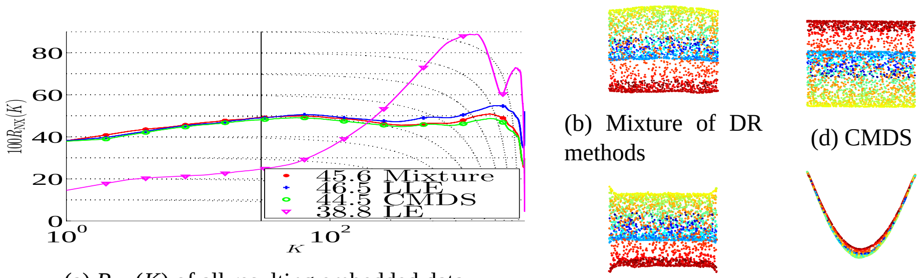
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 6. Results for the 3D sphere data-set.



(a) $R_{NX}(K)$ of all resulting embedded data.

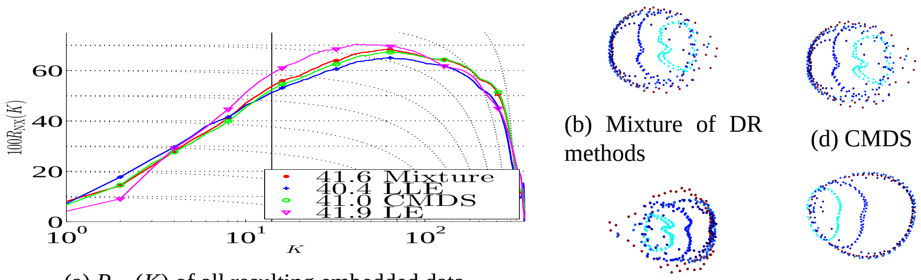
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 7. Results for the Swiss roll data-set.



(a) $R_{NX}(K)$ of all resulting embedded data.

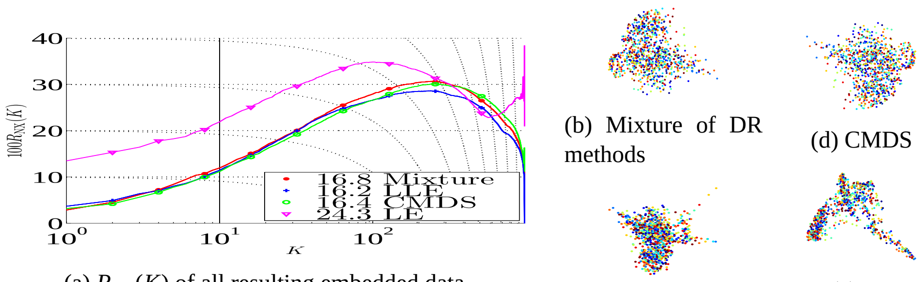
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 8. Results for COIL data-set.



(a) $R_{NX}(K)$ of all resulting embedded data.

(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 9. Results for MNIST dataset.

4.4. Experiment description

To assess the performance of the interactive visualization interface, a testings were done by clicking on the colored surface. Doing so a collection of weighting factors are established to consequently carry out the mixture. Here, particularly test the vertexes and a random point inside the surface Figure 10.

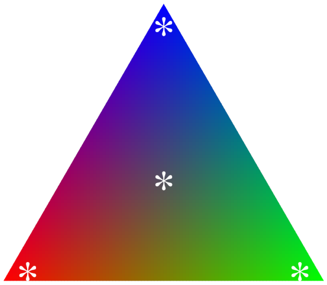


Figure 10. The chosen points for the experiment.

5. Results and discussion

The general interface's scheme is shown in Figure 11. The RGB color selected within the color surface defines the mixture of DR methods and the embedded data.

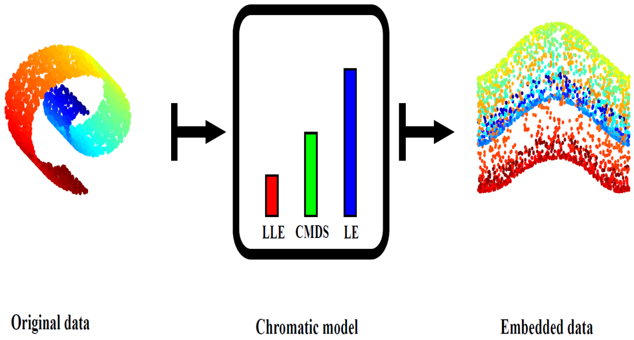


Figure 11. A general scheme of the proposed interface. This is an example for Swiss roll dataset.where the RGB value of the pixel define a new embedding space (right hand).

In section 4 the experiment is explained, three DR methods are considered and they are represented by the chromatic model. The experiment of Figure 10 is carried out to test the DR methods and the mixture, that it was taken approximately in a central point into the chromatic model. The results are shown on Figures 6 to 9. In the results can be appreciated the embedded data and several curves that gives a notion to the user about the performance of the low dimensional space and the preservation of neighbors. If the value of the area under the curve is greater, the performance of the embedded data will be better. An interesting fact can be seen in Figure 6a where the mixture have an area under the curve greater than some concrete DR methods.

6. Conclusions and future Work

The proposed chromatic model represents a suitable alternative to reduce the gap between users and data base because DR methods can be selected/mixed through a color-based framework. This approach results appealing since color is one of the first levels of human perception making naturally intuitive its use. Incorporating the chromatic model within an interface, the user can easily explore all the color surface to find the best representation of the input data into a lower-dimensional space. To do so, unsupervised DR methods are approximated by kernels matrices. Consequently, such matrices are linearly combined by means of weighted sum, whose coefficients are provided interactively by the user.

As a future work, more developed and interactive models are to be explored. As well, new kernel approaches from other dimensionality reduction methods that allow the arising of new DR approaches.

Acknowledgments

This work was supported by the Department of Electronics from Universidad de Nariño and Facultad de Ingeniera en Ciencias Aplicadas from Universidad Tcnica del Norte. Authors acknowledgment the research projects “Diseño e implementación de un prototipo electrónico de bajo costo para terapias de biofeedback en tratamientos de trastornos psicofisiológicos” funded by Fundación CEIBA and Gobernación de Nariño, as well as “Combinación interactiva de mtodos reducción de dimensión, a partir de una interfaz de visualización inteligente” from Corporación Universitaria Autónoma de Nariño.

References

[1] A. Asuncion and D. Newman. Uci machine learning repository. irvine, ca: University of california, school of information and computer science. *Available online a t http://www. ics. uci. edu/ mlearn/MLRepository. html*, 2007.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] I. Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[4] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton. Visualizing similarity data with a mixture of maps. In *International Conference on Artificial Intelligence and Statistics*, pages 67–74, 2007.

[5] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital image processing using MATLAB*. Pearson/Prentice Hall,, 2004.

[6] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 2013.

[9] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). *Dept. Comput. Sci., Columbia Univ., New York.[Online] http://www. cs. columbia. edu/CAVE/coil-20. html*, 62, 1996.

[10] D. H. Peluffo-Ordonez, J. Aldo Lee, and M. Verleysen. Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, pages 171–177. IEEE, 2014.

[11] D. H. Peluffo-Ordóñez, A. E. Castro-Ospina, J. C. Alvarado-Pérez, and E. J. Revelo-Fuelagán. Multiple kernel learning for spectral dimensionality reduction. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 626–634. Springer, 2015.

[12] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Recent methods for dimensionality reduction: A brief comparative analysis. In *European Symposium on Artificial Neural Networks (ESANN)*. Citeseer, 2014.

[13] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Short review of dimensionality reduction methods based on stochastic neighbour embedding. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 65–74. Springer, 2014.

[14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[15] J. Salazar-Castro, Y. Rosas-Narvaez, A. Pantoja, J. C. Alvarado-Perez, and D. H. Peluffo-Ordonez. Interactive interface for efficient data visualization via a geometric approach. In *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*, pages 1–6. IEEE, 2015.

[16] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. In *Computer Graphics Forum*, volume 34, pages 201–210. Wiley Online Library, 2015.

[17] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03*, 2012.

[18] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2634–2643, 2013.