Interactive Data Visualization Using Dimensionality Reduction and Similarity-Based Representations

P. Rosero-Montalvo[1,2], P. Diaz[2,3], J.A. Salazar-Castro[4,5], D.F. Peña-Unigarro[5], A.J. Anaya-Isaza[6,7], J.C. Alvarado-Pérez[8,9], R. Therón[8], and D.H. Peluffo-Ordóñez[1(✉)]

[1] Universidad Técnica del Norte, Ibarra, Ecuador
dhpeluffo@utn.edu.ec
[2] Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
[3] Universidad Nacional de la Plata, Ensenada, Argentina
[4] Universidad Nacional Sede Manizales, Manizales, Colombia
[5] Universidad de Nariño, Pasto, Colombia
[6] Universidad Surcolombiana, Neiva, Colombia
[7] Universidad Tecnológica de Pereira, Pereira, Colombia
[8] Universidad de Salamanca, Salamanca, Spain
[9] Coorporación Universitaria Autónoma de Nariño, Pasto, Colombia

Abstract. This work presents a new interactive data visualization approach based on mixture of the outcomes of dimensionality reduction (DR) methods. Such a mixture is a weighted sum, whose weighting factors are defined by the user through a visual and intuitive interface. Additionally, the low-dimensional representation space produced by DR methods are graphically depicted using scatter plots powered via an interactive data-driven visualization. To do so, pairwise similarities are calculated and employed to define the graph to be drawn on the scatter plot. Our visualization approach enables the user to interactively combine DR methods while provided information about the structure of original data, making then the selection of a DR scheme more intuitive.

Keywords: Data visualization · Dimensionality reduction · Pairwise similarity

1 Introduction

The aim of dimensionality reduction (DR) is to obtain lower dimensional representations of high-dimensional input data keeping -under a pre-established criterion- the structure of data as well as possible. Reaching this aim, entails both the performance of a pattern recognition system and intelligible data representation can be improved [1]. Traditionally, DR methods are designed by following pre-established optimization criteria and design parameters. But they mostly lack of properties like interactivity and controllability, being important characteristics of the field of Information Visualization (InfoVis) [2]. InfoVis

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 334–342, 2017.
DOI: 10.1007/978-3-319-52277-7_41

provides interfaces and graphical ways of representing data making the available information more usable and intelligible for the user. However, it turns out that DR outcomes can be enhanced by taking advantages of some properties of InfoVis methods [3,4]. Following this premise, some approaches have proposed [5,6], making use of interactivity with equalizer-bar like interfaces or geometric interaction models. In general, such approaches implement interesting interactive models but their final visualization lacks the information about structure of the data from the original input space -at least in an easy to understand and/or visual way-.

In this work, we introduce a new visualization approach using an interactive mixture of data representations resultant from DR methods. After performing the DR methods on the input data, a set of lower-dimensional representation spaces are obtained. Particularly, the mixture is done via a weighted sum. In order to give users a sense of the structure of data, we implement a data-driven visualization in addition to the conventional scatter plot. Such a visualization captures the structure of the input data by using a similarity matrix (as well, affinity matrix from graph theory), which captures the degree of similarity or affinity between every pair of data points. The visualization consists of plotting lines (edges) between data points exhibiting the highest value of similarity. Additionally, to provide more sense of interactivity, user can control the number of edges by a varying parameter -working as a slider bar within an interface-. By design, affinity is selected as a Gaussian one so that the structure of local neighbor points can be taken into account. Particularly, low-dimensional spaces are obtained by the state of the art of methods such as: Classical Multidimensional Scaling (CMDS) [2], Laplacian Eigenmaps (LE) [7], Locally Linear Embedding (LLE) [8], Stochastic Neighbor Embedding (SNE), and t-Student-distributed-SNE (t-SNE) [1,7]. To perform the mixture, user can set the weighting factors by picking up values from a equalizer-bar-like interface. To test our visualization approach, we use a 3D artificial spherical shell data set. The quality of resultant representation spaces is quantified by a scaled version of the average agreement rate between K-ary neighborhoods [9]. The proposed mixture may represent every single dimensionality reduction approach as well as it helps users to find a suitable representation of input data within a visual and friendly user interface.

The remaining of the paper is organized as follows: In Sect. 2, Data visualization via dimensionality reduction is outlined. Section 3 introduces the proposed interactive data visualization scheme. Experimental setup and results are presented in Sects. 4 and 5, respectively. Finally, Sect. 6 gathers some final remarks as conclusions and future work.

2 Data Visualization via Dimensionality Reduction

Perhaps, one of the most intuitive ways of visualizing numerical data is through a 2- or 3-dimensional representation of original data, which can be readily represented using a scatter plot. In consequence, dimensionality reduction arises as an Correspondingly, DR is aiming at reaching a low-dimensional data representation, upon which both the classification task performance is improved in terms of

accuracy, as well as the intrinsic nature of data is properly represented [10]. So, when performing a DR method, a more realistic and intelligible visualization for the user is expected [1]. More technically, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1 \leq 1 \leq N}$ such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1 \leq 1 \leq N}$ being $\boldsymbol{y}_i \in \mathbb{R}^d$, where $d < D$ [1,11]. Figure 1 depicts an instance where a manifold, so-called 3D spherical shield, is embedded into a 2D representation, which resembles to an unfolded version of the original manifold.

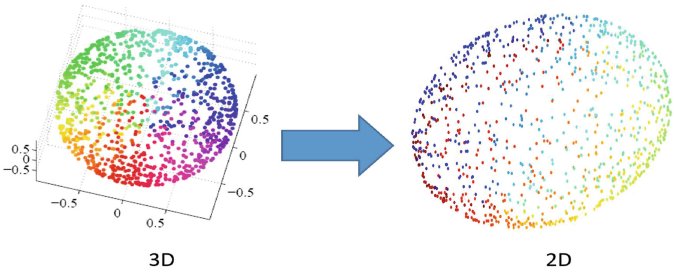


Fig. 1. Dimensionality reduction effect over an artificial (3-dimensional) spherical shell manifold. Resultant embedded (2-dimensional) data is an attempt to unfolding the original data.

3 Interactive Data Visualization Scheme

The proposed visualization approach, here called DataVisSim, involves three main stages: mixture of DR outcomes, interaction, and visualization, as depicted in the block diagram of Fig. 2. One of the most important contributions of this work is that information on the structure of the input high-dimensional space is added to the visual final representation, by using a pairwise-similarity-based scheme.

3.1 Mixture

Let us suppose that the input matrix $\boldsymbol{Y}$ is reduced by using M different DR methods, yielding then a set of lower-dimensional representations: $\{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(M)}\}$. Herein, we propose to perform a weighted sum in the form:

$$\bar{\boldsymbol{X}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{X}^m, \tag{1}$$

where $\{\alpha_1, \cdots, \alpha_M\}$ are the weighting factors. To make the selection of weighting factors intuitive, we use probability values so that $0 \leq \alpha_m \leq 1$ and $\sum_{m1=1}^{M} \alpha_m = 1$, and therefore all matrices $\boldsymbol{X}^{(m)}$ should be normalized to rely within a hypersphere of ratios.

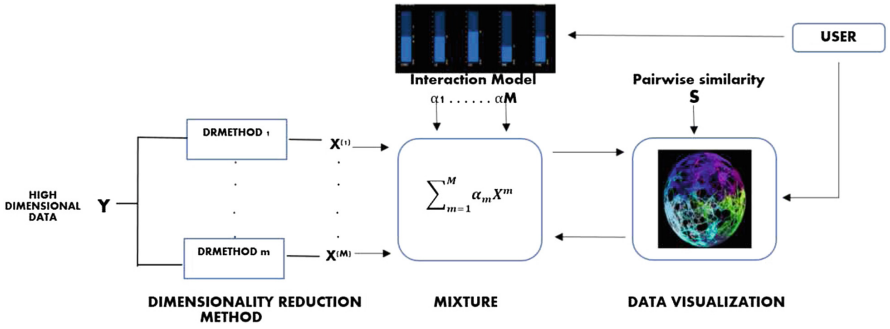


Fig. 2. Block diagram of proposed interactive data visualization using dimensionality reduction and similarity-based representations (DataVisSim). Roughly speaking, it works as follows: first performs a mixture of resultant lower-dimensional representation spaces by taking advantage of conventional implementations of traditional DR methods. The interaction is provided through a interface that enables user to dynamically input the weighting factors for the aforementioned mixture. For visualization, a novel similarity-based approach is used.

3.2 Interaction Model

For the sake of interactivity, the values of every α_m, required to calculate $\bar{\boldsymbol{X}}$ according to Eq. (1), are to be defined by the users using an equalizer-bar available in the interface. Within a friendly-user and intuitive environment, weighting factors can be readily inputted by just picking up values from bars. In order to provide quick views of resultant representation space, as soon as a point is picked up the remaining ones are automatically completed following a uniform density probability function. The same is done in case than more than one value is selected.

3.3 Similarity-Based Visualization

The most used method to visualize 2- or 3-dimensional data is the scatter plot. In this work, we introduce a similarity-based visualization approach with the aim to provide a visual hint about the structure of the high-dimensional input data matrix $\boldsymbol{Y}$ into the scatter plot of its representation in a lower-dimensional space To do so, we use a pairwise similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$, such that $\boldsymbol{S} = [\boldsymbol{s}_{ij}]$. In terms of graph theory, entries $\boldsymbol{s}_{ij}$ defines the similarity or affinity between the i-th and j-th data point from $\boldsymbol{Y}$. Doing so, we can hold the structure of original input space in a topological fashion, specifically in terms of pairwise relationships. For visualization purposes, such a similarity is used to define graphically the relationship between data points by plotting edges. In order to control the amount of edges and make an appealing visual representations, the value of $\boldsymbol{s}_{ij}$ is constrained as $\boldsymbol{s}_{ij} > \boldsymbol{s}_{max}$, being $\boldsymbol{s}_{max}$ a maximum admissible similarity value to be given by the users as well. In other words, our visualization approach consists of building a graph with constrained affinity values.

4 Experimental Setup

Database: In order to visually evaluate the performance of the DataVisSim approach, we use an artificial spherical shell (N = 1500 data points and D = 3), as depicted in Fig. 1.

Parameter Settings and Methods: In order to capture the local structure for visualization, i.e. data points being neighbors, we utilize the Gaussian similarity given by: $\boldsymbol{s}_{ij} = -exp(-0.5\|\boldsymbol{y}_{(i)} - \boldsymbol{y}_{(j)}\|^2/\sigma^2)$. The parameter is a bandwidth value set as 0.1, being the 10% of the hypersphere ratio (applicable once matrices are normalized as discussed in Sect. 3.1. To perform the dimensionality reduction we consider $M = 5$ DR methods, namely: CMDS, LE, LLE, SNE, and t-SNE. All of them are intended to obtain spaces in dimension $d = 2$.

Performance Measure: To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [9] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N-1$, 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.

5 Results and Discussion

Figure 3 shows the scatter plots for the resultant low-dimensional spaces obtained by the considered dimensionality reduction methods, as well as the performed mixture. Quality curves and corresponding scatter of each mixture are shown

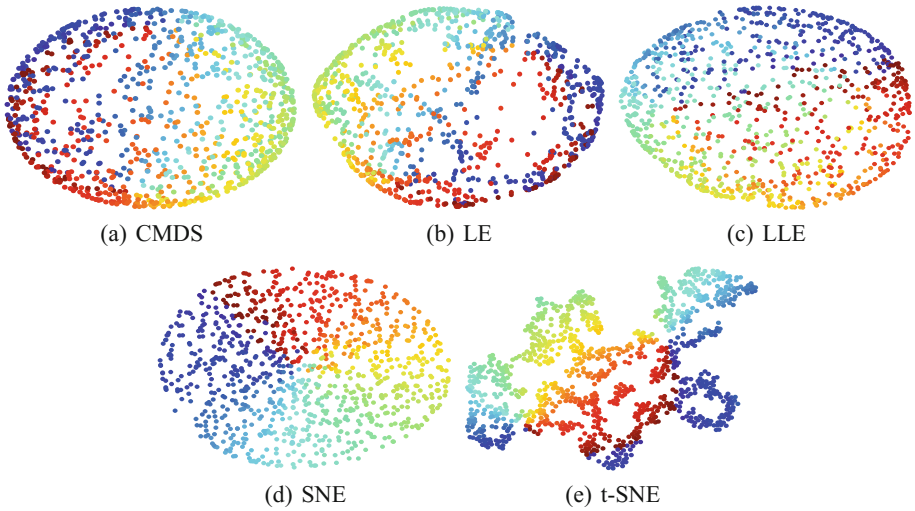
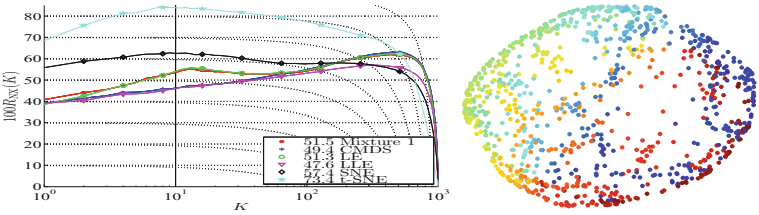


Fig. 3. The effects of dimensionality reduction of RD methods considered on the 3D sphere. The results are embedded data represented in a bidimensional space.

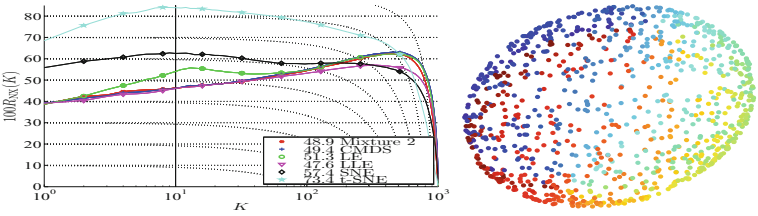
in Figs. 4, 5, 6, 7 and 8. As seen, $R_{NX}(K)$ measure allows for assessing both the different mixtures and the RD methods independently. Since the area under its curve represents a quality measure of the low-dimensional space, is in turn a visual and intuitive indicator that helps the user to find the best either a single DR method or the proper mixture.

To test the DataVis approach, we implement an interface on Processing software, which allows to easily code visual arts. Then, it results appealing for creating visual analytics interfaces. Figure 9 shows a view of the implemented interface. For the sake of easily handling so that (even non-expert) users may interact with DR methods and their feasible combinations in an intuitive manner



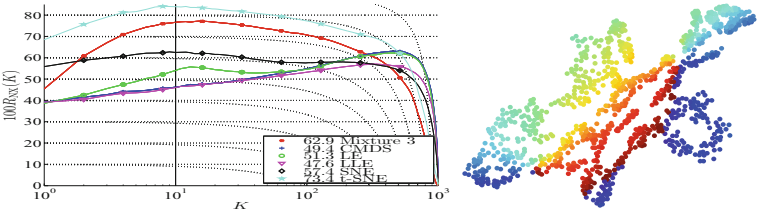
(a) R_NX(K) for all considered methods. (b) Mixture 1

Fig. 4. (a) Performance of the mixture 1 and all methods deemed RD. In (b) the embedded data resulting from mixture 1 are indicated.



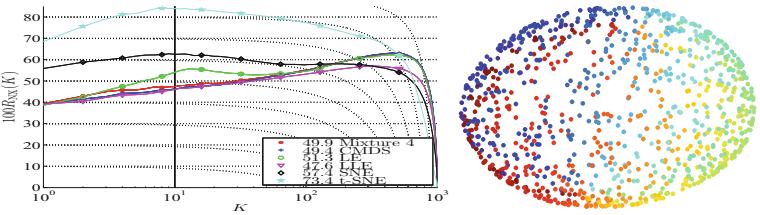
(a) R_NX(K) for all considered methods. (b) Mixture 2

Fig. 5. (a) Performance of the mixture 2 and all methods deemed RD. In (b) the embedded data resulting from mixture 2 are indicated.



(a) R_NX(K) for all considered methods. (b) Mixture 3

Fig. 6. (a) Performance of the mixture 3 and all methods deemed RD. In (b) the embedded data resulting from mixture 3 are indicated.



(a) $R_N X$(K) for all considered methods. (b) Mixture 4

Fig. 7. (a) Performance of the mixture 4 and all methods deemed RD. In (b) the embedded data resulting from mixture 4 are indicated.

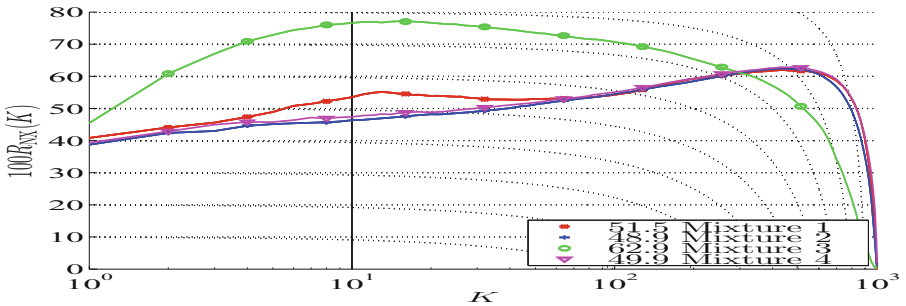


Fig. 8. Performance of all selected mixtures.

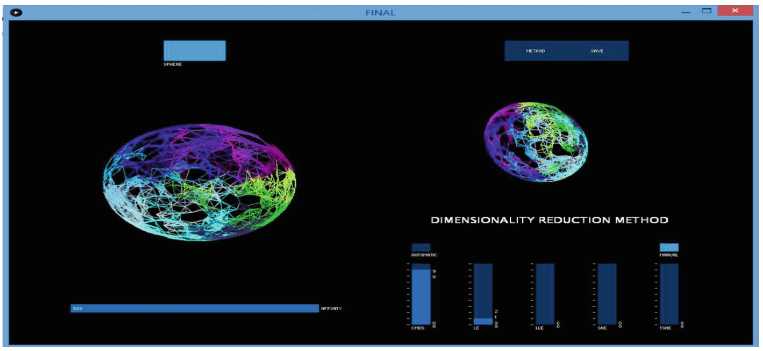


Fig. 9. View of the DataVisSim interface implemented on processing software (https://sites.google.com/site/intelligentsystemsrg/home/gallery).

using equalizer-like bars. This is possible because of resultant data representations are properly set according to the human perception. As well, the interface incorporates a slider bar to dynamically draw the edges between nodes. This is useful for visual analysis given that it allows to relate the structure of high-dimensional data (original data) within the visualization of the low-dimensional

representation space. Therefore, it is provided a powerful tool for making decisions of the most suitable representation of the original data, in other words, the most proper DR methods.

6 Conclusions and Future Work

This work presents a new interactive data visualization approach based on mixture of the outcomes of dimensionality reduction (DR) methods. The core of this approach consists of plotting lines (edges) between data points exhibiting the highest value using a similarity matrix which measure the degree of similarity or affinity between every pair of data points capturing the structure of the input data. Such visualization of a topology can be represented by a data-driven graph in addition to the conventional scatter plot, to provide more sense of interactivity to the user for selecting and/or combining DR methods while providing information about the structure of original data. Correspondingly, data points represent the nodes and an affinity matrix holds the pairwise edge weights. As a future work, other dimensionality reduction methods are to be integrated into data-driven graph, so that a good trade between preservation of data structure and intelligible data visualization can be reached. More mathematical properties will be explored to design data-driven schemes that best approximate the topology data.

References

1. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Short review of dimensionality reduction methods based on stochastic neighbour embedding. In: Villmann, T., Schleif, F.-M., Kaden, M., Lange, M. (eds.) Advances in Self-Organizing Maps and Learning Vector Quantization. AISC, vol. 295, pp. 65–74. Springer, Heidelberg (2014). doi:10.1007/978-3-319-07695-9_6
2. Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer Science & Business Media, New York (2005)
3. Dai, W., Hu, P.: Research on personalized behaviors recommendation system based on cloud computing. Indones. J. Electr. Eng. Comput. Sci. **12**, 1480–1486 (2013)
4. Ward, M.O., Grinstein, G., Keim, D.: Interactive Data Visualization: Foundations, Techniques, and Applications. CRC Press, Boca Raton (2010)
5. Peluffo-Ordónez, D.H., Alvarado-Pérez, J.C., Lee, J.A., Verleysen, M., et al.: Geometrical homotopy for data visualization. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015) (2015)
6. Díaz, I., Cuadrado, A.A., Pérez, D., García, F.J., Verleysen, M.: Interactive dimensionality reduction for visual analytics. In: Proceedings of 22th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014), pp. 183–188. Citeseer (2014)
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**, 1373–1396 (2003)
8. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)

9. 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)
10. Bertini, E., Lalanne, D.: Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In: Proceedings of ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, pp. 12–20. ACM (2009)
11. 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)