Abstract

Dimensionality reduction (DR) methods represent a suitable alternative to visualizing data. Nonetheless, most of them still lack the properties of interactivity and controllability. In this work, we propose a data visualization interface that allows for user interaction within an interactive framework. Specifically, our interface is based on a mathematic geometric model, which combines DR methods through a weighted sum. Interactivity is provided in the sense that weighting factors are given by the user via the selection of points inside a geometric surface. Then, (even non-expert) users can intuitively either select a concrete DR method or carry out a mixture of methods. Experimental results are obtained using artificial and real datasets, demonstrating the usability and applicability of our interface in DR-based data visualization.

1. Introduction

Dimensionality reduction (DR) is a key stage for designing pattern recognition and data mining systems when dealing with high-dimensional data sets [1]. The aim of DR methods is to extract lower dimensional, relevant information (called embedded data) from high-dimensional input data, so that both the performance of a pattern recognition system can be improved and data representation becomes more intelligible [2]. Since DR methods are often developed under determined design parameters and pre-established optimization criteria, they still lack the properties of user interaction and controllability, which are characteristic of information visualization procedures [3]. The field of information visualization (Info Vis) is aimed at developing graphical ways of representing data so that information can be more usable and intelligible for the user. Then, one can intuit that DR can be improved by importing some properties of Info Vis methods. This is in fact the premise on which this research is based [4].

This paper presents an attempt to link the field of dimensionality reduction with that of information visualization, in order to harness the special properties of the latter within DR frameworks. In particular, the properties of controllability and interactivity are of interest, which should make the DR outcomes significantly more understandable and tractable for the (no-necessarily-expert) user [5]. These two properties allow the user to have freedom to select the best way for representing data. Specifically, we propose a geometrical strategy to set the weighting factors for linearly combining DR methods. This is done from kernel approximations [6, 7] of conventional methods (Classical Multidimensional Scaling - CMDS [3], Laplacian Eigenmaps – LE, and Locally Linear Embedding - LLE), which are combined to reach a mixture of kernels. To involve the user in the selection of a method, we use a polygonal approach so the points inside the polygon surface defines the degree or level that a kernel is used, that is, the set of weighting factors. Such polygon has as many edges as the number of considered kernels. This approach allows to evaluating visually the behavior of the embedding data regarding the kernel mixture.

For experiments, we use publicly available databases from the UCI Machine Learning Repository [8] as well as a subset of images from Columbia University Image Library [9]. To assess the performance of the kernel mixture, we consider conventional methods of spectral dimensionality reduction such as multidimensional scaling, locally linear embedding and laplacian eigenmaps [10]. The quality of obtained embedded data is quantified by a scaled version of the average agreement rate between K-ary neighborhoods [19]. Provided mixture represents every single dimensionality reduction approach as well as it helps users to find a suitable representation of embedded data within a visual and intuitive framework.

The remaining of the paper is organized as follows: In section 2, data visualization using DR is outlined. Section 3 introduces a novel mathematical geometric model based on a polygonal approach aimed at performing customized DR tasks. Experimental setup and results are shown in Sections 4 and 5, respectively. Finally, Section 6 gathers some final remarks as conclusions and future work.
2. Data visualization

An intuitive way of visualizing numerical data is via a 2D or 3D scatter plot, which is a natural and intelligible visualization fashion for human beings. Therefore, it entails that the initial data should be represented into a lower-dimensional space. In this sense, dimensionality reduction takes places, being an important stage within both the pattern recognition and data visualization systems. 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 [1]. So, a more realistic and intelligible visualization for the user is obtained [2]. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $Y = [y_i]_{i=1}^N$, such that $y_i \in \mathbb{R}^D$, into a low-dimensional, latent data matrix $X = [x_i]_{i=1}^N$, being $x_i \in \mathbb{R}^d$, where $d < D$. Figure 1 depicts an instance where a manifold, so-called Swiss roll, is embedded into a 2D representation, which resembles to an unfolded version of the original manifold.

![Dimensionality reduction effect over a Swiss roll manifold. Resultant embedded data is an attempt to unfolding the original data.](image)

Classical DR approaches aims to preserve variance (principal component analysis - PCA) or distance (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aiming at preserving the data topology. Such a topology can be represented by a data-driven 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 (LE) [10]. As well, using a non-symmetric similarity matrix and focusing on data local structure, the Locally Linear Embedding (LLE) method arose [18]. On the other hand, once normalized, similarity matrix can also represent probability distributions, as do the methods based on divergences such as stochastic neighbor embedding [11].

3. Mathematical geometric approach

In this Section, we introduce a novel method for interactive data visualization within a white box environment, which consists of the combination of different, spectral unsupervised DR methods. For the sake of versatility and since spectral methods are susceptible to be represented by kernels, combination is carried out taking into account the corresponding kernel matrices. Our method is based on a mathematical geometric approach that allows for performing the mixture of DR methods in an interactive fashion, so that kernel matrices are linearly combined and respective coefficients are related to geometric coordinates inside a geometric surface. So, users –even non-expert ones – might easily and intuitively select a single method or combine methods fulfilling their needs by just de exploring the geometric shape and picking up points from the surface thereof. Figure 2 shows graphically a possible mathematic-geometric model regarding a polygonal approach. In general, any set of methods can be represented by a collection of functions $\{f_1, \ldots, f_M\}$, where $M$ is the number of considered functions. A fashion to perform a pair-wise mixture is the continuous deformation of one function onto another using homotopy basic [12, 13]. Then, a simple model of homotopy can be written as $h(f_1, f_2, \lambda) = \lambda f_1 + (1 - \lambda) f_2$, where $\lambda$ is the homotopy parameter. In terms of an interactive interface, such a parameter would become a sliding bar. So far, this approach can be graphically represented by a line with length 1 drawn between two points representing the two functions, as seen in Figure 2(a).

![Polygonal approach to perform the mixture of a set of functions $\{f_1, \ldots, f_M\}$. Parameter $\lambda$ controls the pairwise mixture.](image)
Notwithstanding, this model can be naturally extended to more than two methods within a polygonal framework, in such a way that three functions are to be represented with a triangle (Figure 2(b)), four functions with a rhombus (Figure 2(c)), and so forth.

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 \( \mathbf{R} \) as the mixture of \( M \) kernel matrices \( \{\mathbf{K}^{(1)},...,\mathbf{K}^{(M)}\} \) so:

\[
\mathbf{R} = \sum_{m=1}^{M} \alpha_m \mathbf{K}^{(m)},
\]

where \( \alpha_m \) is the coefficient or weighting factor corresponding to method \( m \) and \( \mathbf{\alpha} = [\alpha_1,...,\alpha_M] \) is weighting vector. As mentioned above, these coefficients are to be associated with geometric coordinates of points inside the surface.

In this work, the relationship between the points inside the surface and the coefficients of linear combination is given by the distance from every single vertex (representing considering methods) to the selected point, as can be appreciated in Figure 3. Then, these distances as the ratios \( \{r_1,...,r_M\} \) of circles with areas \( \{A_1,...,A_M\} \) centered at each vertex. Therefore, it is evident that the area of the \( m \)-th circle is \( A_m = \pi r_m^2 \).

Then, once such areas are normalized to sum to 1, the complement value of them becomes a proper estimation of the weighting factors. In this connection, the values of \( \mathbf{\alpha} \) are given by:

\[
\alpha_m = \text{sinc} \left( 1 - \frac{A_m}{\sum_{m=1}^{M} A_m} \right).
\]

Additionally, in order to assigning a higher value to that weighting factor corresponding to the vertex closer to the selected point, function \( \text{sinc}(\cdot) \) is used, which also adds an equalization effect over the values.

**4. Experimental setup**

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 the COIL-20 image bank [14], which contains 72 gray-level images representing 20 different objects (\( N = 1440 \) data points –20 objects in 72 poses/angles– with \( D = 1282 \)). The third data set is a randomly selected subset of the MNIST image bank [15], 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 \( D = 242 \)). The fourth data set is a toy set here called Swiss roll (\( N = 3000 \) data points and \( D = 3 \)). Figure 4 depicts examples of the considered data sets.
Three kernel approximations for spectral DR methods [6] are considered. Namely, Classical Multidimensional Scalling (CMDS), locally linear embedding (LLE), and graph Laplacian Eigenmaps (LE). CMDS kernel is the double centered distance matrix $D \in \mathbb{R}^{N \times N}$ so:

$$K^{(1)} = K_{\text{CMDS}} = -\frac{1}{2}(I_N - 1_N 1_N^T)D(I_N - 1_N 1_N^T),$$

(3)

where the $ij$ entry of $D$ is given by $d_{ij} = ||y_i - y_j||^2_2$ and $|| \cdot ||^2$ stands for Euclidean norm. A kernel for LLE can be approximated from a quadratic form in terms of the matrix $W$ holding linear coefficients that sum to 1 and optimally reconstruct observed data. Define a matrix $M \in \mathbb{R}^{N \times N}$ as $M = (I_N - W)(I_N - W^T)$ and $\lambda_{\text{max}}$ as the largest eigenvalue of $M$. Kernel matrix for LLE is in the form:

$$K^{(3)} = K_{\text{LLE}} = \lambda_{\text{max}} I_N - M.$$  

(4)

Since kernel PCA is a maximization problem of the covariance of the high dimensional data represented by a kernel, LE can be expressed as the pseudo-inverse of the graph Laplacian $L$:

$$K^{(5)} = K_{\text{LE}} = L^T,$$

(5)

where $L = D - S$, $S$ is a similarity matrix and $D = \text{Diag}(S1_N)$ is the degree matrix. All previously mentioned kernels are widely described in [6]. The similarity matrix $S$ is 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 [16]. The number of neighbors is established as $K = 30$. As well, a RBF kernel is also considered: $K^{(4)} = K_{\text{RBF}}$ whose $ij$ entries are given by $\exp(-0.5||y_i - y_j||^2_2/\sigma^2)$ with $\sigma = 0.1$. For all methods, input data is embedded into a 2-dimensional space ($d = 2$).

Accordingly, our approach is performed considering $M = 4$ kernels. The resultant kernel provided $\tilde{K}$ here as well as the individual kernels $\{K^{(1)}, \ldots, K^{(4)}\}$ are tested by obtaining embedded data from kernel PCA, as explained in [17]. To quantify the performance of studied methods, the scaled version of the average agreement rate $R_N(K)$ introduced in [19] is used, which is ranged within the interval $[0, 1]$. Since $R_N(K)$ 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

Following are presented some experimental results aimed at testing all the considered datasets regarding the embedded data using $R_N(K)$ as a quality indicator. Resultant kernel matrices feed a kernel PCA algorithm to output 2D-dimensional data spaces. Since all the experiments are carried out considering four methods, the polygonal surface is then a rhombus as shown in Figure 5. Given that the mixture presented here is a linear combination, when coefficients are selected from the perimeter only two kernels are considered. Then, user can appreciate the deformation of the resulting embedding from a method onto that from another method by moving from one vertex to another on the respective edge.
Figure 6. Results for COIL dataset. Results are shown regarding the quality measure $R_{XX}(K)$. The curves and their AUC (a) for all considered methods are depicted, as well as the embedding data (b)-(f). Individual embedding data spaces are obtained by selecting the points rightly on the vertexes meanwhile the mixture is done with the coefficients associated to the central point.

Figure 7. Results for the spherical shell dataset.

Figure 8. Results for the Swiss roll dataset.

Figure 9. Results for MNIST dataset.
Indeed, selecting coefficients associated with the vertexes, the effect of a single method is performed. In addition, when selecting inner points the effects of every method is taken into account to calculate the resultant kernel. Therefore, the proposed approach enable users (even those not expert) to interact with the DR outcomes by intuitively selecting points from a polygonal surface. Overall obtained results are shown in Figures 6 to 9.

6. Conclusion and future work

The proposed interface represents an interactive approach to visualize the embedded data resulting from dimensionality reduction (DR) methods. This approach is based on a geometric perspective of homotopy allowing for dealing with more than two functions. In this case, kernel matrices representing spectral, unsupervised methods of DR. In particular, our approach relates the inner points location of a polygonal surface with the values of weighting factors used to carry out the linear combination of kernel matrices. Given this graphic and intuitive framework, even non-expert users might easily select a method or combination of methods by picking up points from a polygonal surface fulfill their specific needs.

As a future work, more developed and interactive models are to be explored. As well, ways to optimize and speed up algorithm routines are to be studied and developed.

Acknowledgments

This work is supported by the “Grupo de Investigación en Ingeniería Eléctrica y Electrónica – GIIEE” from Universidad de Nariño, as well as ESLINGA Research Group from Universidad Cooperativa de Colombia, sede Pasto.

As well, authors acknowledgment the research project “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.

References