ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

VisCoDeR: A Tool for Visually Comparing Dimensionality Reduction Algorithms

Rene Cutura[1], Stefan Holzer[1], Michaël Aupetit[2], and Michael Sedlmair[1,3]

1- University of Vienna, Department of Computer Science, Austria

2- Qatar Computing Research Institute, HBKU, Doha, Qatar

3- Jacobs University, Computer Science & Electrical Engineering, Bremen, Germany

Abstract. We propose VisCoDeR, a tool that leverages comparative visualization to support learning and analyzing different dimensionality reduction (DR) methods. VisCoDeR fosters two modes. The *Discover mode* allows qualitatively comparing several DR results by juxtaposing and linking the resulting scatterplots. The *Explore mode* allows for analyzing hundreds of differently parameterized DR results in a quantitative way. We present use cases that show that our approach helps to understand similarities and differences between DR algorithms.

1 Introduction

Dimensionality reduction (DR) is a widely used approach to visualize high-dimensional data. Most commonly, a high-dimensional dataset gets projected to a 2D scatterplot so users can learn about clusters or distributions in the data. The process of DR for visualization, however, does not come without challenges and threats. The low-dimensional projection contains errors and a good understanding of the DR process is needed to correctly interpret the results. Unfortunately, in many cases end-users are not aware of these distortions [4].

Our goal is to help overcome this problem by providing easier and more intuitive tools for users to learn and analyze DR algorithms. State-of-the-art approaches in this area use interactive visualization to explain individual algorithms, such as t-SNE [12], or enrich scatterplots to reveal potential misinterpretations [1, 10]. These approaches focus on understanding and using one algorithm at a time. The goal of our work goes beyond in that we try to better understand and characterize the value of **comparative visualization**. The basic idea is to show results of multiple different algorithms and different parameterizations in parallel and allow the user to compare among them. In doing so, users can learn about similarities and differences between DR algorithms.

Towards this goal, we contribute a tool called VisCoDeR[1] that illustrates the potential of using interactive comparative visualization to support learning and analyzing DR algorithms. We designed the tool with two user groups in mind: junior data scientists who seek to learn and better understand DR algorithms, as well as DR designers, who seek to evaluate and analyze DR methods. We show that several important use cases linked with understanding DR are supported and can help users to learn about DR algorithms and their behaviors.

[1] Visual Comparison of Dimensionality Reduction Algorithms, anagram of *discover*

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

2 Background and Related Work

Most people learn DR algorithms in courses, tutorials, or books which focus on explaining the mathematical foundations of DR algorithms using some toy datasets. To actually use DR algorithms, they need to utilize R, python, or similar to call a (parameterized) DR algorithm and create static scatterplots. With these standard processes, crucial learning and analysis tasks remain unsupported or cumbersome though [8]: What do these DR results exactly mean? Which DR algorithm should be used for a specific dataset and problem at hand? How do results from different algorithms compare among each other? What is the influence of different parameters and how should they be set?

To overcome these issues, researchers have investigated how *interactive visualization* can help to tackle this problem [7]. For example, Wattenberg and Viega [12] provide a website with the goal that users gain intuition about the parameters of the t-SNE algorithm [5]. Other researchers contributed approaches to enrich DR visualizations to better understand assumptions and protect from potential misinterpretations [2, 10]. Probing projection [10], for instance, is a tool that allows learning about the mechanics of the used DR algorithm by interaction techniques that show mapping distortions directly in the projection.

All these approaches focus on one DR algorithm at a time. Although identified as a core challenge of interactive DR [7], *comparing* and picking among different algorithms and parameterizations has gained very little attention so far. The Dimstiller tool is a notable exception in that it offers users guidance for picking a good DR algorithm [4]. Still it does not support a direct comparison between different algorithms and their parameterizations.

Our focus is to fill this gap and investigate the comparative visualization [3, 9] of many DR results at the same time. In particular VisCoDeR enables learning about important aspects of DR techniques like: **linearity**, **explanatory dimensions**, **iterative optimization** and **stability**. It also supports teaching about **locating and identifying distortions**, understand mapping **sensitivity to parameters** if any, **connections between projections and original dimensions**, and **connections between DR techniques**.

3 VisCoDeR

In the following, we explain the two modes of VisCoDeR, the qualitative *Discover mode*, and the quantitative *Explore mode*.

3.1 Discover Mode

The main component of the *Discover mode* shows juxtaposed DR-results (Fig. 1b). Points are color-coded according to their classes[2], helping users to visually connect the different scatterplots.

[2]Without loss of generality, we focus on labeled data for our illustrations here

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

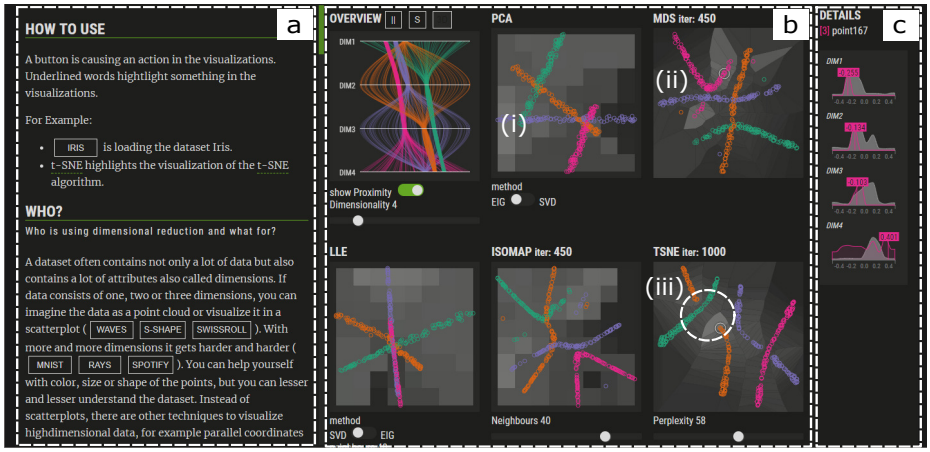

Fig. 1: Screenshot of the *Discover mode* using the *Rays* data—(a): linked textual description; (b): juxtaposed DR results, (i) *Component Plane* view in the background, (ii) *Proximity* view in the background, (iii) 2 crossed rays; (c): distribution of all data (grey) and a selected class (pink) along all input dimensions.

A *textual description* (Fig. 1a) offers a verbal explanation of DR algorithms, just as in a text book. In addition, text passages are interactively connected with the visualization of the DR results though. In doing so, learners can, for instance, explore different datasets and DR algorithms in context of the text.

We also show a *direct representation* of the original data, at the top left corner of the main view (Fig. 1b). For 3D data we show an interactive 3D plot. For higher dimensions, we use a parallel coordinate plot, a scatterplot matrix, or a density map [6]. The idea is to help learners to make an easier mental **connection between projections and original dimensions** by giving them a more direct representation of the high-dimensional space and link it to the DR results and its **distortions** [12]. The distribution of the data and of selected classes is shown in an additional histogram view in Fig. 1c.

Two additional interactive views help foster the learning process. Upon selecting a point, the *Proximity* view [1] indicates other points that are close in the high-dimensional space but not in the projection using a bright background color (Fig. 1b-ii). This view reveals *false* and *missed neighbors* of the selected point and helps to **locate and identify distortions**. The *Component Plane* view [11] helps to explore how dimensions are mapped in the DR results. The background of the points is colored brighter the higher their value along the selected dimension is (Fig. 1b-i). This view reveals **explanatory dimensions** of clusters and can demonstrate the **linearity** of the projection.

Some DR techniques, such as ISOMAP or t-SNE, require **parameters** to be set, and rely on **iterative optimization**. VisCoDeR allows to interactively change these parameters, and shows the iterative optimization process by animating the respective scatterplot. In doing so, the user can test the **stability** of the technique and learn about its **sensitivity to parameters**.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

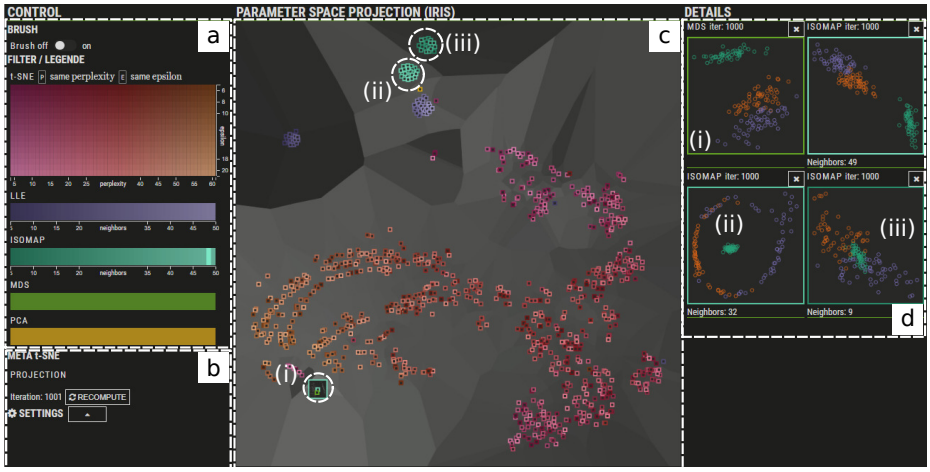

Fig. 2: Screenshot of the *Explore mode* using the *Iris* data—(a): interactive color legends of DR algorithms and their parameterizations; (b): settings/parameterization and control over the meta t-SNE; (c): meta-map with 1004 DR results and activated proximity visualization; (d): clicking a dot in the meta-map displays the DR result in detail.

3.2 Explore Mode

In the *Explore mode*, users can analyze hundreds of DR results at a glance. This visualization is specifically interesting for investigating different parameterizations [9] and further understand **parameter sensitivity**. The main idea is to display all DR results in a meta-map (Fig. 2c). Each dot represents a DR scatterplot result, so that close/far dots mean similar/different DR results. To build the meta-map, we first encode each precomputed DR scatterplot as a vector of the Euclidean distance of all of its N points to their center of gravity. Then we apply t-SNE to these N-dimensional vectors to get the meta-map.

We use 1D and 2D color-scales (Fig. 2a) to create a better understanding of DR parameterizations. Users can interactively brush the color scales and see the corresponding DR dots highlighted in the meta-map. Vice versa, brushing dots in the meta-map highlights them in the colored parameter space. Several other features exist, such as interactively drilling down into details of each DR (Fig. 2d), or recomputing the whole meta-map using other t-SNE parameters (Fig. 2b), and evaluating its quality with the *Proximity* view.

4 Use Cases

We built an online prototype of VisCoDeR[3]. For illustration, we integrated a set of default datasets and DR algorithms, so users can start right away. Users can also upload their own datasets and DR algorithms.

[3]The prototype, a video, and more can be found at `http://renecutura.eu/viscoder/`

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

4.1 Use Cases with the Discover Mode

We use the DR techniques PCA, LLE, MDS, ISOMAP, and t-SNE on several artificial and real datasets: *Iris*, *MNIST*, *Swiss roll*, *S-shape*, *Waves*, *Spotify*, and *Rays*/*Rays-touching*. Based on our own teaching experience and on commonly known stumbling blocks from the literature [2], we now report typical cases of DR education that could be further enriched with VisCoDeR.

The *Rays* and *Rays-touching* datasets allow users to change dimensionality interactively. Points are generated along each dimension as a line segment with small Gaussian noise, crossing at the origin for *Rays-touching* and separated for *Rays*. Users can see PCA failing to map the rays as separate beams even for low dimensions while t-SNE always manages to separate them at the cost of splitting apart some of them too (Fig. 1b). The proximity coloring helps to further **locate distortions and identify them** as false-neighbors when two lines cross. For MDS (see Fig. 1b-ii), for instance, we can see both false and missed neighbors; t-SNE (Fig. 1b-iii) also reveals missed neighbors. The **linearity** of a projection, can be inspected by investigating the linearity of the projected line segments, and by checking whether the component plane of any dimension shows a linear gradient along each of these lines. Both demonstrate the linearity of the PCA mapping, while ISOMAP and t-SNE result in clearly non-linear projections. Also t-SNE is **iterative** while PCA comes at once being the result of a closed-form solution to an eigen-decomposition problem. The t-SNE map has no **stability** in contrary to PCA and LLE which always provide the same output.

For the *MNIST* dataset, t-SNE offers the most correct result regarding labels as ground truth. The different digits are clearly separated based on their image pattern, although all tested DR methods ignore their class labels [2]. Outlying digits can also be explained by referring to their snippet images showing the original digit data. Screenshots can be found on the supplemental webpage.

4.2 Use Cases with the Explore Mode

We now use the same DR techniques with the *Iris* datasets, which consists of two clusters (and three classes). We precomputed 1004 projections resulting from all the DR techniques varying their parameters if any. The meta-map supports various analytical tasks [9]. We discuss two examples; further details and tasks are illustrated in the video.

Sensitivity to Parameters— The meta-map shows similar scatterplot results as neighboring dots. A user might be interested in understanding the impact of t-SNE's two parameters: perplexity coded as hue, and epsilon as lightness. Inspecting the hue and lightness gradients in Fig. 2c, we can see that that perplexity (hue) follows a smooth gradient from left to right. In contrast, the dots with the same epsilon (lightness) are scattered over the meta-map, and seem to be overly sensitive to small changes of the parameter setting.

Connections between DR Techniques— The meta-map also reveals how different DR techniques are connected. ISOMAP, for instance, results from applying MDS to a distance matrix computed as the Euclidean length of all the

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

shortest-paths over the m-Nearest Neighbor graph of the data (m-NNG). If the number m of neighbors in ISOMAP is too low, the m-NNG is disconnected, resulting in maps with many distortions, all clustered in the meta-map (Fig. 2d-iii). If m is high enough the m-NNG is connected (Fig. 2d-ii). Increasing m, ISOMAP gets stepwise closer to the MDS result and is identical to MDS for the complete graph (Fig. 2d-i). Clusters form in the meta-map for a set of contiguous values of m, due to sudden creation of shortcuts in the neighborhood graph that abruptly change geodesic distances between data. Thus we can better understand the connection between ISOMAP and MDS techniques.

5 Conclusions and future work

Our work is meant as a first step to illustrate how visually comparing DR results can foster a better understanding of different behaviors of DR algorithms. There are many avenues for future work. One interesting idea would be to more closely couple the two different modes of VisCoDeR. Also, empirical studies can shed light on further usability improvements, and help understand the tool's performance "in the wild".

References

[1] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7):1304–1330, 2007.

[2] M. Aupetit. Sanity check for class-coloring-based evaluation of dimension reduction techniques. In *Proc. Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pages 134–141. ACM, 2014.

[3] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[4] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 3–10, 2010.

[5] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[6] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[7] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Trans. Visualization & Computer Graphics (TVCG)*, 23(1):241–250, 2017.

[8] 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.

[9] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE Trans. Visualization & Computer Graphics (TVCG)*, 20(12):2161–2170, 2014.

[10] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. Visualization & Computer Graphics (TVCG)*, 22(1):629–638, 2016.

[11] J. Vesanto. Som-based data visualization methods. *Intelligent data analysis*, 3(2):111–126, 1999.

[12] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 2016.