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/.

Interactive dimensionality reduction of large datasets using interpolation

Ignacio Díaz[1], Daniel Pérez[1], Abel A. Cuadrado [1], Diego García[1], Manuel Domínguez [2] *

1- Electrical Engineering Dept. University of Oviedo
Edif. Dept., Campus de Viesques s/n 33204, Gijón - SPAIN
2- SUPPRESS Research Group, University of Leon,
Escuela de Ingenierías, Campus de Vegazana, 24007 León, Spain

Abstract. In this work we present an approach to achieve interactive dimensionality reduction (iDR) on large datasets. The main idea of the paper relies on using *generalized regression neural network* (GRNN) interpolation to obtain massive out of sample projections from iDR projections obtained on a reduced sample of the original dataset. The proposed method allows to achieve fluid iDR interaction on datasets between 45× and 100× larger than with the original DR method for similar latencies, yet achieving good distance preservation. The paper includes a rank-based comparison between the proposed method and the DR method used alone for different datasets and parameter values.

1 Introduction

Visual analytics (VA) allows the user to get insight in data analysis problems by means of an efficient combination of machine learning, data visualization and interaction. One approach in VA consists in embedding steerable machine learning algorithms into the visualization [1, 2]. In such approach, the user can modulate the behavior of the ML algorihtm by re-tuning its parameters to explore the problem in different ways. This kind of interaction goes beyond the classical mechanisms of zoom, pan, etc. since it implies a reconfiguration of the visualization according to more general goals, such as, for instance, modifying the metrics used in the analysis. In *interactive dimensionality reduction* (iDR) a DR algorithm is run under user-driven changes on the parameters or the input data, and its projections are represented *during* convergence thereby providing an advanced visual feedback on the structure of data [3]. Despite iDR applications have been described [3], it has serious limitations on the input data sizes, since it requires projection updates at video framerate speeds for effective low-latency interaction, making it viable only for sample sizes of a few hundreds of points. In [4], kernel based out of sample methods are proposed for visualization of large datasets, but not for an interactive task. We propose here an approximation for DR based on *generalized regression neural network*(GRNN) interpolation that can be used in iDR algorithms, allowing speed-ups of as much as 100× the classical DR approaches, for some configurations using a t-SNE algorithm [5]. We make a comparison of the full-DR approach vs the grnn-DR approach

*The authors would like to thank financial support from the Spanish Ministry of Economy (MINECO)and FEDER funds from the EU under grants DPI2015-69891-C2-1/2-R and from the Principado de Asturias government through the predoctoral grant "Severo Ochoa".

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/.

using rank-based methods (co-ranking), revealing that the quality of the grnn-DR approach is competitive and present an example iDR application where 4000 points are explored with framerates in the order of 15 frames per second (fps). The paper is organized as follows. In section 2 we describe the GRNN approximation for fast DR. In Section 3 we compare the quality of the proposed method vs the original DR, first describing the evaluation measure (3.1) and the methodology employed for the comparative analysis (3.2), and then discussing the results of the comparison (3.3) both in terms of the quality of the projection and in terms of speed. Finally, section 4 concludes the paper.

2 Method description

Let's consider a set of points in the input data space $\{\mathbf{x}_i\}_{i=1,\dots,Q}$, being $\mathbf{x}_i \in \mathbb{R}^D$. We propose the following method for fast DR in interactive applications:

- Obtain a representative sample $\{\mathbf{m}_i\}_{i=1,\dots,S}$, being $\mathbf{m}_i \in \mathbb{R}^D$, with a reduced size $S \ll Q$, that is distributed with a similar pdf to the input dataset. Well known methods can be used, such as k-means, SOM or neural gas, allowing to obtain a set of codebooks $\mathbf{m}_i$ that preserve the structure of the original dataset.
- Using a DR algorithm, obtain low-dimensional projections $\{\mathbf{g}_i\}_{i=1,\dots,S}$, being $\mathbf{g}_i \in \mathbb{R}^d$, of the reduced sample.
- Obtain out-of-sample projections by means of a continuous mapping obtained from $\{\mathbf{m}_i, \mathbf{g}_i\}$ as input-output samples. We propose in this paper to use *generalized regression neural network* (GRNN) as a simple, non-parametric and fast method

$$\mathbf{y}'_i = \mathrm{grnn}_{\mathbf{m}_j \to \mathbf{g}_j}(\mathbf{x}_i) \stackrel{\mathrm{def}}{=} \frac{\sum_j \mathbf{g}_j \phi(\mathbf{x}_i - \mathbf{m}_j)}{\sum_j \phi(\mathbf{x}_i - \mathbf{m}_j)}, \qquad \text{being} \qquad \phi(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}$$

3 Comparative analysis of the proposed method

3.1 Evaluation measure

We consider *rank-based criteria* [6] to compare both approaches. Considering the co-ranking matrix $Q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|$, being ρ_{ij} and r_{ij} the ranks in the high and low dimensionality space, respectively, as described in [6], compute the quality measure

$$Q_{NX}(K) = \tfrac{1}{KN} \textstyle\sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl}$$

that describes the number of points that remain in the K-neighborhood [7]. To summarize the neighborhood preservation, we consider aggregated values of the $Q_{NX}(K)$ curve taking its average between k_1 and k_2

$$Q_{NX}(k_1, k_2) = \tfrac{1}{k_2 - k_1 + 1} \textstyle\sum_{K=k_1}^{k_2} Q_{NX}(K).$$

In particular, we shall consider three descriptors to evaluate the local, medium and global neighborhood preservation

$$Q_{NX}^{\mathrm{loc}} = Q_{NX}(1, k_1), \quad Q_{NX}^{\mathrm{med}} = Q_{NX}(k_1, k_2), \quad Q_{NX}^{\mathrm{glo}} = Q_{NX}(k_2, N)$$

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/.

3.2 Comparison methodology

We compared the interpolation method described in section 2 (grnn-DR) with the original DR (full-DR) according to neigborhood preservation, using rank-based criteria to compare the input dataset $\mathbf{x}_i$ against its projections $\mathbf{y}_i$ and $\mathbf{y}'_i$ obtained with full-DR and grnn-DR, respectively. To do the evaluation in broad conditions we used a grid approach considering different datasets and parameters of the method, including: dataset (*spiral*, *surface*, *swiss roll*), the width factor σ of the GRNN interpolation, the perplexity p of the t-SNE algorithm, and the level of added noise to the original data ν. The evaluation method has the following steps:

- Add some noise to the original dataset $\mathbf{x}_i \leftarrow \mathbf{x}_i + \nu\boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i \rightarrow N(\mathbf{0}, \mathbf{I})$.
- Use k-means to obtain a set of S representative codebook vectors $\mathbf{m}_i$ of the original dataset $\mathbf{x}_i$.
- Obtain the projections $\mathbf{y}_i$ using a t-SNE on $\mathbf{x}_i$ with perplexity p.
- To avoid the need to run a different t-SNE on the $\mathbf{m}_i$ (and hence avoid the effect of different initializations) take $\mathbf{g}_i = \mathbf{y}_{c(i)}$, where $c(i) = \arg\min_k\{\|\mathbf{x}_i - \mathbf{m}_k\|^2\}$ is the index of the best matching unit for point $\mathbf{x}_i$.
- Obtain the projections using the interpolation $\mathbf{y}'_i = \mathrm{grnn}_{\mathbf{m}_i \rightarrow \mathbf{g}_i}(\mathbf{x}_i)$, as described above.
- Compute $Q_{NX}^{\mathrm{loc}}, Q_{NX}^{\mathrm{med}}, Q_{NX}^{\mathrm{glo}}$ for $k_1 = 10$ and $k_2 = 50$.
- Repeat for all 4-tuples (dataset, σ, p, ν) resulting from all combinations of the 4 elements.

3.3 Results

Neighborhood preservation. In the application of the former evaluation methodology, we considered the following datasets:

$$\begin{aligned}
\textit{helix} &: \mathbf{x}_i = [\cos(2\pi u_i), \sin(2\pi u_i), 2u_i]^T \\
\textit{6d-surface} &: \mathbf{x}_i = [u_i, v_i, u_i v_i, \cos(u_i), \sin(u_i), 0.5(u_i + v_i)]^T \\
\textit{swiss roll} &: \mathbf{x}_i = [(3 + u_i)\cos(\pi u_i), 2\pi v_i, (3 + u_i)\sin(\pi u_i)]^T
\end{aligned}$$

being u_i, v_i uniform random variables in the interval $(-1, 1)$ and $Q = 200$ the number of points in all datasets. We applied the evaluation method for the three previous datasets, using the $t-$SNE for projection, $S = 40$ codebooks obtained using $k-$means, and considered the following variations in the parameter sets: 3 levels of added noise $\nu = \{0, 0.1, 0.3\}$; 3 perplexities $p = \{5, 10, 15\}$; 3 width factors $\sigma = \{0.2, 1, 5\}$.

Visual comparison. A visual comparison case showing the projections using both methods and the whole $Q_{NX}(K)$ curve is shown in Fig. 1, describing the performance of the full-DR and grnn-DR methods for the *swiss roll* dataset on four runs, with $p = 20$, $\nu = 0$ and $\sigma = \{0.2, 1, 2, 5\}$. It can be seen that the grnn-DR method performs better than the full-DR method in global neighborhood preservation, while showing competitive results also in local neigborhoods for small values of σ.

Analysis of the results. The grid combination of the previous parameter sets leads to a total of 81 runs. To summarize the analysis, we aggregated the results

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/.

perplexity	noise	sigma	q (loc)	q_ (loc)	q (med)	q_ (med)	q (glo)	q_ (glo)
5.0	0.0	0.2	0.805	0.866	0.8	0.876	0.885	0.911
		1.0	0.805	0.83	0.8	0.891	0.885	0.932
		5.0	0.805	0.765	0.8	0.878	0.885	0.921
	0.1	0.2	0.76	0.709	0.809	0.869	0.889	0.912
		1.0	0.76	0.637	0.809	0.869	0.889	0.927
		5.0	0.76	0.581	0.809	0.848	0.889	0.918
	0.3	0.2	0.664	0.545	0.75	0.796	0.875	0.892
		1.0	0.664	0.48	0.75	0.782	0.875	0.911
		5.0	0.664	0.429	0.75	0.754	0.875	0.902
10.0	0.0	0.2	0.871	0.865	0.869	0.895	0.903	0.914
		1.0	0.871	0.814	0.869	0.893	0.903	0.914
		5.0	0.871	0.754	0.869	0.876	0.903	0.908
	0.1	0.2	0.805	0.701	0.852	0.881	0.899	0.915
		1.0	0.805	0.635	0.852	0.87	0.899	0.915
		5.0	0.805	0.564	0.852	0.84	0.899	0.911
	0.3	0.2	0.713	0.537	0.78	0.797	0.88	0.891
		1.0	0.713	0.461	0.78	0.77	0.88	0.902
		5.0	0.713	0.415	0.78	0.737	0.88	0.898
20.0	0.0	0.2	0.853	0.857	0.903	0.906	0.902	0.91
		1.0	0.853	0.795	0.903	0.894	0.902	0.908
		5.0	0.853	0.752	0.903	0.875	0.902	0.903
	0.1	0.2	0.799	0.705	0.895	0.889	0.911	0.912
		1.0	0.799	0.583	0.895	0.854	0.911	0.908
		5.0	0.799	0.528	0.895	0.825	0.911	0.907
	0.3	0.2	0.702	0.542	0.803	0.798	0.882	0.889
		1.0	0.702	0.424	0.803	0.739	0.882	0.894
		5.0	0.702	0.387	0.803	0.708	0.882	0.894

Table 1: Neighborhood preservation (q = full-DR, q_ = grnn-DR, higher is better)

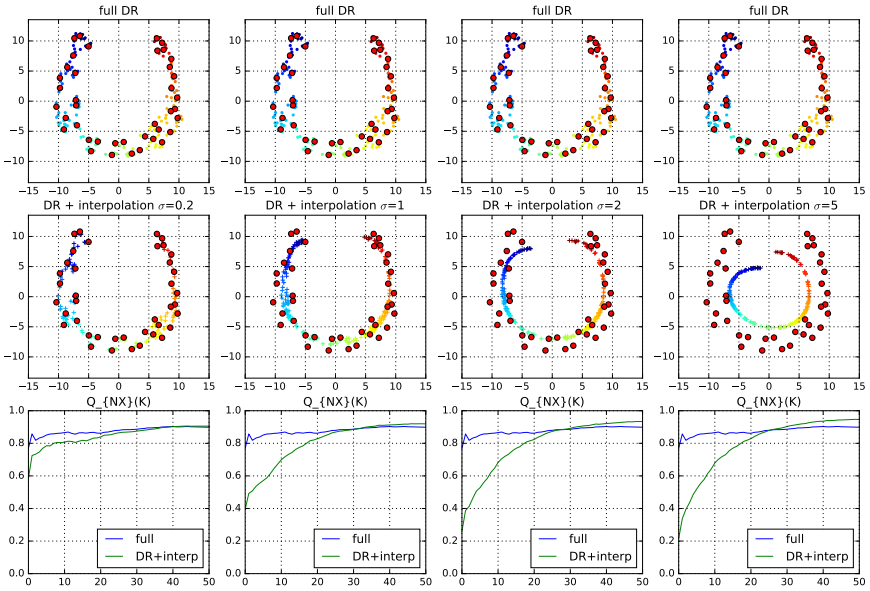

Fig. 1: Co-ranking based $Q_{NX}(K)$ curve for full-DR vs grnn-DR for $\sigma \in \{0.2, 1, 2, 5\}$ on the *swiss roll* dataset (large markers: codebooks, small markers: interpolated points).

computing the average for the three datasets in all combinations, resulting in Table 1 with the 27 combinations of (σ, ν, p). Some interesting conclusions can be drawn from Table 1 and from Fig. 2, which presents the same results in

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/.

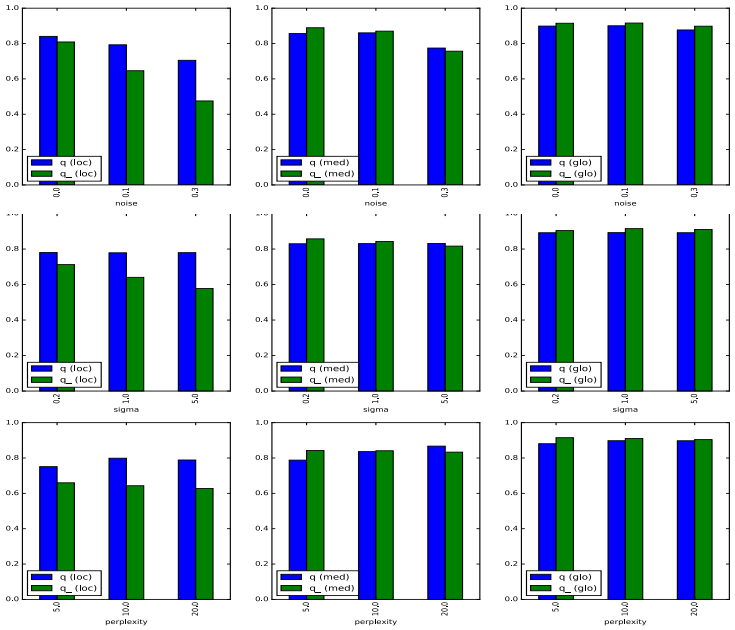

Fig. 2: Global comparison of neigborhood preservation based on the rank-based quality criterion $Q_{NX}(K)$. Local, medium and global metrics are shown on the left, center and right columns, respectively.

aggregated form. In most of the cases, the grnn-DR method shows better performance than the full-DR in preservation of medium and global neighborhoods. The width factor σ has a strong impact in the local neighborhood preservation, with smaller values showing better results. This could be expected, since large values of σ provide regularized smooth manifolds, with low curvatures, that fail to describe the local variations in geometry. The local neighborhood preservation ability of grnn-DR is highly sensitive to noise, having its worst comparative performance with respect to full-DR for large noise levels.

Speed performance. To test the proposed approach for fast iDR on large datasets, the proposed method was implemented on a sample javascript interface that can be tested online in `http://isa.uniovi.es/~idiaz/demos/ESANN2018/`. The interface can be tested for versions of the three example datasets with a much larger number of points $Q = 4000$. For each dataset, the algorithm implements a t-SNE algorithm on $S = 100$ codebooks and then uses GRNN to project the Q points. It also implements a basic iDR functionality, allowing the user to turn on/off input features for distance metrics computation, thereby allowing to explore their efect in the data structure —see [3]. Our tests on a Debian GNU/Linux 9, 32 Gb RAM, Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz with NVIDIA GeForce GTX 745 for $Q = 4000$ using full-DR yielded, 0.2 fps, unfeasible for iDR operation. Using, however, the grnn-DR method for

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/.

$S = \{50, 100, 200\}$ resulted in framerates of $\{20, 16, 9\}$ fps, with speedups of $\{100\times, 80\times, 45\times\}$, respectively. These results are consistent since the computational complexity for full-DR is $\mathcal{O}(Q^2)$ while the grnn-DR combines the t-SNE, with $\mathcal{O}(S^2)$, and the GRNN, with $\mathcal{O}(Q \times S)$. Additional comparisons of our proposal yielded $21\times$, $26\times$, $56\times$ and $359\times$ computation time for *Barnes-Hut-SNE* (BHSNE) [8] with trade-off parameter values $\theta = 0.8$, 0.5, 0.2, and exact t-SNE respectively (note: these additional comparisons were done in Python, since no Javascript BHSNE implementations were available, and just for the projection computation, without plotting).

4 Conclusions

We have presented here a method for fast DR based on computing the projections on a reduced set of representative codebooks, followed by out-of-sample computation of the remaining points using a fast interpolation method (GRNN). A comparative analysis between the proposed method and an equivalent full-DR projection was done for three datasets under different conditions, using the co-ranking matrix to evaluate the neighborhood preservation. The results show that the interpolation method performs comparatively well, with worse —yet acceptable— results in local neighborhood preservation, while being superior in global preservation of the structure. An implementation of the method demonstrates that it can achieve significantly low latencies —20 fps— for as much as 4000 points, making it a good choice for iDR applications.

References

[1] Alexander Endert, William Ribarsky, Cagatay Turkay, William Wong, Ian Nabney, Ignacio Díaz, and Fabrice Rossi. The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum*, 2017.

[2] Jaegul Choo and Haesun Park. Customizing computational methods for visual analytics with big data. *Computer Graphics and Applications, IEEE*, 33(4):22–28, 2013.

[3] Ignacio Díaz, Abel A. Cuadrado, Daniel Pérez, Francisco J. García, and Michel Verleysen. Interactive dimensionality reduction for visual analytics. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, April 2014.

[4] Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, 147:71 – 82, 2015. Advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012).

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[6] John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431 – 1443, 2009. Advances in Machine Learning and Computational Intelligence.

[7] Wouter Lueks, Bassam Mokbel, Michael Biehl, and Barbara Hammer. How to evaluate dimensionality reduction? - improving the co-ranking matrix. *CoRR*, abs/1110.3917, 2011.

[8] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245, 2014.