ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Perplexity-free Parametric t-SNE

Francesco Crecchi[1], Cyril de Bodt[2], Michel Verleysen[2], John A. Lee[2] and Davide Bacciu[1] *

1- Universitá di Pisa - Dipartimento di Informatica
Largo Bruno Pontecorvo, 3, 56127 Pisa - Italy

2 - Université catholique de Louvain - ICTEAM
Place du Levant 3 L5.03.02, 1348 Louvain-la-Neuve - Belgium

Abstract. The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm is a ubiquitously employed dimensionality reduction (DR) method. Its non-parametric nature and impressive efficacy motivated its parametric extension. It is however bounded to a user-defined perplexity parameter, restricting its DR quality compared to recently developed multi-scale perplexity-free approaches. This paper hence proposes a multi-scale parametric t-SNE scheme, relieved from the perplexity tuning and with a deep neural network implementing the mapping. It produces reliable embeddings with out-of-sample extensions, competitive with the best perplexity adjustments in terms of neighborhood preservation on multiple data sets.

1 Introduction

Dimensionality reduction (DR) aims at appropriately mapping high-dimensional (HD) data sets into low-dimensional (LD) spaces. The HD neighborhoods preservation typically characterizes the LD embedding relevance. Different paradigms have been studied in DR, from linear projections as principal component analysis (PCA) to nonlinear distance preservation schemes [1]. These being strongly limited by the norm concentration phenomenon [2], neighbor embedding approaches as Stochastic Neighbor Embedding (SNE) [3] and extensions [4, 5] have been developed, reaching outstanding DR performances. Remarkably, the t-SNE algorithm [6] acquired tremendous popularity among the DR community. This motivated the design of t-SNE variants, such as its parametric version [7]. Parametric schemes can indeed easily deal with very large-scale databases, by learning the projection on a random subset only and mapping the remaining data samples afterward [8]. They can also nicely adapt to online settings, by adding new data points to an existing embedding.

However, t-SNE requires the user to choose a perplexity to tune the widths of its HD Gaussian neighborhoods. While such a single-scale method well retains neighborhood sizes near to the perplexity but without attaining similar performances for the other neighborhoods, multi-scale approaches typically much better recover both local and global HD structures [9], and present the key advantage to avoid relying on a perplexity parameter. Nevertheless, no parametric extensions of these schemes have been proposed yet.

*CdB is a FNRS Research Fellow. JAL is a FNRS Senior Research Associate.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

This paper hence presents a parametric version of multi-scale t-SNE [10], generalizing parametric t-SNE [7] in a perplexity-free method with more reliable neighborhood preservation. The neural network architecture implementing the HD-LD mapping in [7] has been refined according to recent deep learning insights. Experiments highlight competitive results with parametric t-SNE and show more efficient management of datasets of heterogeneous structure.

This paper is structured as follows: Section 2 first summarizes SNE, t-SNE and multi-scale t-SNE, while Section 3 reviews parametric t-SNE and introduces its proposed multi-scale, perplexity-free version. Section 4 describes DR quality assessment and its application in the parametric setting. Section 5 presents the experiments, while Section 6 sketches conclusions and future works.

2 SNE, t-SNE and multi-scale t-SNE

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{i=1}^N$ be a set with N points in a HD space (HDS) with dimension M. Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ model it in a P-dimensional space (LDS), $P \leq M$. The HD (LD) distance between the ith and j^{th} points is noted δ_{ij} (d_{ij}). SNE introduces HD and LD similarities, for $i \in \mathcal{I} = \{1, \ldots, N\}$ and $j \in \mathcal{I} \backslash \{i\}$ [3]:

$$\sigma_{ij} = \frac{\exp\left(-\pi_i \delta_{ij}^2/2\right)}{\sum_{k \in \mathcal{I}\backslash\{i\}} \exp\left(-\pi_i \delta_{ik}^2/2\right)}, \; s_{ij} = \frac{\exp\left(-d_{ij}^2/2\right)}{\sum_{k \in \mathcal{I}\backslash\{i\}} \exp\left(-d_{ik}^2/2\right)}, \; \sigma_{ii} = s_{ii} = 0.$$

A binary search tunes the precision π_i to set the perplexity of the distribution $[\sigma_{ij}; j \in \mathcal{I}\backslash\{i\}]$ to a user-specified soft neighborhood size $K_\star$: π_i such that $\log K_\star = -\sum_{j \in \mathcal{I}\backslash\{i\}} \sigma_{ij} \log \sigma_{ij}$. SNE then optimizes the LDS by minimizing the sum of the KL divergences between the HD and LD similarity distributions.

In addition to symmetrizing the similarities, t-SNE uses a LD Student t-distribution with one degree of freedom, overcoming the crowding problem [6]:

$$\sigma_{ij,t} = \frac{\sigma_{ij} + \sigma_{ji}}{2N}, \; s_{ij,t} = \frac{1}{\left(1 + d_{ij}^2\right) \sum_{k \in \mathcal{I}, l \in \mathcal{I}\backslash\{k\}} \left(1 + d_{kl}^2\right)^{-1}}, \; s_{ii,t} = 0.$$

Gradient descent then minimizes the same cost function as SNE, $C_{t-SNE} = \sum_{i \in \mathcal{I}, j \in \mathcal{I}\backslash\{i\}} \sigma_{ij,t} \log\left(\sigma_{ij,t} / s_{ij,t}\right)$. In practice, the $K_\star$ parameter is frequently arbitrarily set. Small values shape the LDS by mainly neglecting mid- and large-scale data interactions, due to the quickly vanishing HD similarities. Large perplexities lead to close to uniform HD similarities, impairing the small neighborhood reproduction. These difficulties are solved by multi-scale methods [9], which preserve both local and global structures at once in the LDS by accounting for various perplexities. Multi-scale t-SNE [10] refines the HD similarities as

$$\sigma_{hij} = \frac{\exp\left(-\pi_{hi} \delta_{ij}^2/2\right)}{\sum_{k \in \mathcal{I}\backslash\{i\}} \exp\left(-\pi_{hi} \delta_{ik}^2/2\right)}, \; \overline{\sigma}_{ij} = \frac{1}{H} \sum_{h=1}^{H} \frac{\sigma_{hij} + \sigma_{hji}}{2N}, \; \sigma_{hii} = \overline{\sigma}_{ii} = 0,$$

where π_{hi} is fixed as in SNE using perplexity $K_h = 2^h$ and $H = \lfloor \log_2\left(N / K_1\right) \rceil$ is the number of scales. The $[\overline{\sigma}_{ij}]_{i,j=1}^N$ similarities are then matched with $[s_{ij,t}]_{i,j=1}^N$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

by optimizing SNE cost function. Conveniently, no more perplexity parameter is needed, the HD structure being faithfully quantified in a data-driven way.

3 Parametric t-SNE and multi-scale parametric t-SNE

In its original formulation, t-SNE [6] is a non-parametric manifold learner. The main limitation of the non-parametric manifold learners is that they do not provide a parametric mapping between the high-dimensional data space and the low-dimensional latent space, making it impossible to embed new data points without the need to re-train the model. The out-of-sample extension of t-SNE, namely parametric t-SNE [7], parametrizes the non-linear mapping between the data space and the latent space through a feed-forward neural network.

As a neural network with sufficient hidden layers (with non-linear activation functions) is known to be capable of approximating arbitrarily complex non-linear functions [11], it is used here for learning the parametric mapping $f : \mathbf{\Xi} \rightarrow \mathbf{X}$ from the data space $\mathbf{\Xi}$ to the low-dimensional latent space $\mathbf{X}$. The weights of the neural network are learned to minimize the Kullback-Leibler (KL) divergence C_{t-SNE} between the HD and LD probability distributions[1]. The asymmetric nature of the KL divergence leads the minimization to focus on modeling large σ_{ij}'s by large s_{ij}'s. Hence the objective function focuses on modeling similar data points close together in the latent space. As a result of this, parametric t-SNE focuses on preserving the local structure of the data.

The *locality* of the method is retained according to the main hyper-parameter in t-SNE: the perplexity $K_\star$. It can be interpreted as the size of the soft Gaussian neighborhood in the HD space, and it has a strong impact on the modeled representation. The optimal tuning of this parameter is the key to obtain LD embeddings retaining most of the local structure of the data. The user can be relieved from this duty by employing a multi-scale approach [10], which automatically takes into account various perplexities to preserve both local and global structures at once.

The main contribution of this work is to introduce the multi-scale approach into parametric t-SNE, leading to a parametric manifold-learner capable to retain both the local and the global structure of the data, without the need of tuning the perplexity parameter. To achieve this, we updated the original parametric t-SNE neural network by using $\overline{\sigma}_{ij}$ to compute HD similarities, in a multi-scale fashion. Moreover, we replaced logistic activation functions with piecewise-linear ones (i.e. *ReLU*s) which do not saturate during training. This simple architectural choice allowed us to ease the training procedure by dismissing the unsupervised pre-training step introduced in the original parametric t-SNE paper [7]. The source code is publicly available at `https://github.com/FrancescoCrecchi/Multiscale-Parametric-t-SNE`.

[1] Refer to [7] for training procedure details.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

4 Assessing the quality of dimensionality reduction

Quality criteria for unsupervised DR usually measure the HD neighborhood preservation in the LDS [12, 4]. The K nearest neighbor sets of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ in the HDS and LDS are noted ν_i^K and n_i^K. Their average normalized agreement is $Q_{\mathrm{NX}}(K) = \sum_{i\in\mathcal{I}} \left|\nu_i^K \cap n_i^K\right| / (KN) \in [0,1]$. As random LD points yield $\mathbb{E}\left[Q_{\mathrm{NX}}(K)\right] = K/(N-1)$, distinct neighborhood sizes are confronted using $R_{\mathrm{NX}}(K) = \left((N-1)\, Q_{\mathrm{NX}}(K) - K\right)/(N-1-K)$, displayed with a log-scale for K to favor the typically prevailing closer neighbors [13]. The area under this curve, $\mathrm{AUC} = \left(\sum_{K=1}^{N-2} R_{\mathrm{NX}}(K)/K\right) \Big/ \left(\sum_{K=1}^{N-2} K^{-1}\right) \in [-1,1]$ increases with DR quality, assessed at all scales with an emphasis on smaller ones [9].

In the context of parametric mappings, R_{NX} curves are computed on both the training set and the union of the training and test sets in Section 5, i.e. before and after projecting the test samples. Parametric maps are indeed commonly used to add new points to an existing LDS. Thus, after the test samples projection, one should assess the quality of the whole augmented LDS and not only of the test set, as the data neighborhoods involve both training and test points.

5 Experiments

We compared the generalization capability of parametric t-SNE (p.t-SNE) for various perplexities and multi-scale parametric t-SNE (Ms.p.t-SNE) in terms of DR quality, for multiple data sets, measured by the R_{NX} and AUC metrics. We measured the model performances in the so-called *extended* scenario which comprises mapping new points into an existing LDS. For a given data set, we held-out a separate test set (30% of the data). The number of hidden layers is set to four for all datasets, matching the original t-SNE paper NN architecture, and a grid-search is performed to choose the number of neurons per layer. Each model is then trained on training samples and performances are measured on the set obtained by merging training and test splits, as mentioned in Section 4. The used data sets are a subset of MNIST (N = 1000, M = 784) [14], COIL-20 (N = 1440, M = 16384) [15], ECOLI (N = 336, M = 8) [16] and Helix data set (N = 1000, M = 3) [9]. Target dimensionality P is 2 for all data sets. Figure 1 reports the quality curves R_{NX} for all data sets and all the compared methods. The proposed multi-scale approach outperforms its single-scale variants in almost all data sets for the training set and performs at least as well as the best single-scale variant on the extended set. More importantly, the curves show that setting the wrong perplexity value has a severe impact on the produced mapping, confirming the usefulness of the multi-scale approach as it reliably approximates the best possible perplexity choice.

6 Conclusions, limitations and future works

In this paper, we combined the benefits of the multi-scale approaches with the parametric t-SNE method, to obtain a perplexity-free parametric scheme that

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

reliably approximates both local and global structures of the data manifold without the need of tuning the perplexity hyper-parameter. This is demonstrated in the experimental assessment: the proposed multi-scale method approaches, and sometimes outperforms, the best single-scale variant on all data sets in projecting new data points into an existing embedding (*extended* scenario).

When embedding new points, all compared methods better retain global structures than local ones. The performance gap between the training and the extended scenarios for small neighborhood sizes may be due to the small sizes of our training data sets [7]. In future works, we intend to investigate this phenomenon by using larger data sets and other parametric manifold learning methods (e.g. PCA, NCA, Autoencoders). Moreover, we plan to mitigate such phenomenon by introducing a regularization term in the loss to better retain the local structure of the data when generalizing to new data points.

References

[1] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 2007.

[2] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Proc. Computer Science*, 4:538–547, 2011.

[3] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840, 2002.

[4] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb):451–490, 2010.

[5] K. Bunte, S. Haase, M. Biehl, and T. Villmann. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

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

[7] Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391, 2009.

[8] K. Bunte, M. Biehl, and B. Hammer. A general framework for dimensionality reducing data visualization mapping. *Neural Computation*, 24(3):771–804, 2012.

[9] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[10] C. de Bodt, D. Mulders, M. Verleysen, and J. A. Lee. Perplexity-free t-SNE and twice Student tt-SNE. In *ESANN*, pages 123–128, 2018.

[11] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[12] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009.

[13] 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*, 112:92–108, 2013.

[14] Yann LeCun, Corinna Cortes, and JC Burges. Christopher. the MNIST database of handwritten digits, 1998.

[15] S. Nene, S. Nayar, and H. Murase. Columbia object image library (COIL-20). 1996.

[16] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

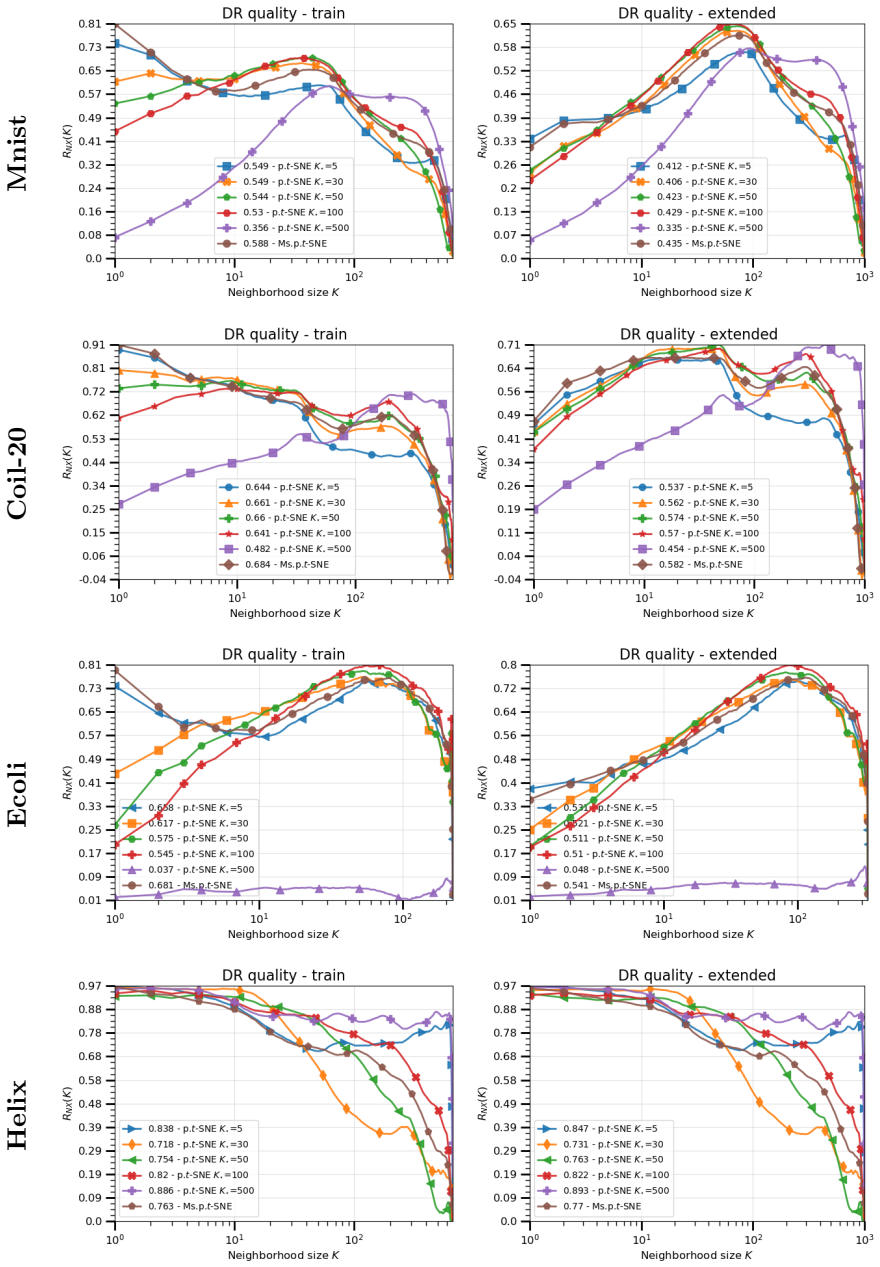

Fig. 1: Quality curves for all data sets and all compared methods for training and extended scenarios. Each curve reports R_{NX}. The higher the curve, the better. The corresponding AUC value for each method is reported in the legend, right before the name.