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/.

Extensive assessment of Barnes-Hut *t*-SNE

Cyril de Bodt[1], Dounia Mulders[1], Michel Verleysen[1] and John A. Lee[2] *

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

2- Université catholique de Louvain - IREC/MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium

Abstract. Stochastic Neighbor Embedding (SNE) and variants are dimensionality reduction (DR) methods able to foil the curse of dimensionality to deliver outstanding experimental results. Mitigating the crowding problem, *t*-SNE became an extremely popular DR scheme. Its quadratic time complexity in the number of samples is nevertheless unaffordable for big data sets. This motivates its Barnes-Hut (BH) acceleration for large-scale use. Although the latter is faster by orders of magnitude, few studies quantify its DR quality with respect to *t*-SNE. Extensive comparisons between *t*-SNE and its BH version are conducted using neighborhood preservation-based criteria. Both methods perform very similarly, suggesting the BH scheme superiority thanks to its reduced time complexity.

1 Introduction

Dimensionality reduction (DR) aims to compute relevant low-dimensional (LD) representations of high-dimensional (HD) data sets. The embedding relevance typically relies on the HD neighborhoods preservation. Different paradigms formalized this principle [1], from early linear models such as principal component analysis (PCA) [1] to nonlinear distance preservation approaches [2]. The latter being strongly affected by the norm concentration phenomenon [3], similarity-based methods such as Stochastic Neighbor Embedding (SNE) [4] and variants [5, 6, 7] emerged with remarkable DR performances. In particular, *t*-SNE [8] became one of the most widely used DR method. Its quadratic time complexity in the number of samples N however limits its applicability to moderate-size data sets. Some studies hence proposed fast schemes for SNE-like methods [9]. The Barnes-Hut (BH) acceleration of *t*-SNE [10] approximates *t*-SNE gradient in $\mathcal{O}\left(N \log N\right)$ time, even though it introduces a new threshold hyper-parameter. Although some works reported impressive computation time gains over *t*-SNE [10], surprisingly few studies analyzed its DR quality. Apart from depicting some data sets LD representations, [10] only measures the 1-nearest neighbor errors of the embeddings of a single labeled data set, for several thresholds and one perplexity. Being solely applicable to labeled data set, this measure at best

*Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11. DM and CdB are Research Fellows of the FNRS. JL is a Senior Research Associate with the FNRS.

reveals the nearest HD neighbors retrieval and is at worst unrelated to neighborhood preservation. Thus it hardly quantifies the HD clusters preservation in the LD space and the HD manifolds unrolling. These are instead well described by analyzing whether larger than 1-size neighborhoods are preserved as well [11].

This work examines the BH t-SNE quality on numerous real-world and artificial data sets, using various perplexities and BH thresholds, thanks to neighborhood-based DR performance criteria [11]. The experimental results suggest that t-SNE and its BH acceleration perform equally well, the latter being sometimes slightly superior to the former. Moreover the DR quality is almost independent of the BH threshold. Fixing it to an adequate value then saves considerable model selection time. These results combined with its reduced time complexity allow the BH acceleration to outperform t-SNE in all aspects.

This paper is organized as follows: Section 2 first reviews SNE, t-SNE and BH t-SNE. Section 3 introduces the DR performance criteria used in this study. Section 4 details the experimental comparison results and Section 5 concludes.

2 SNE, t-SNE, and Barnes-Hut t-SNE

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{i=1}^N$ denote a set of N points in a HD metric space with M features. Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ represent it in a P-dimensional LD metric space (LDS), $P \leq M$. The HD (LD) distance between the i^{th} and j^{th} points is denoted by δ_{ij} (d_{ij}). SNE defines HD and LD similarities, for $i \in \mathcal{I} = \{1, \ldots, N\}$ and $j \in \mathcal{I}\backslash\{i\}$ [4]:

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

The precision π_i is set by binary search to fix the perplexity of the distribution $[\sigma_{ij}; j \in \mathcal{I}\backslash\{i\}]$ to a user-defined 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 finds the LD positions by minimizing the sum of the KL divergences between the H- and LD similarity distributions.

Besides symmetrizing the similarities, t-SNE employs a Student t-distribution with one degree of freedom in the LDS, circumventing the crowding problem [8]:

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

The t-SNE cost function $C_{t-SNE} = \sum_{i\in\mathcal{I}, j\in\mathcal{I}\backslash\{i\}} \sigma_{ij,t} \log\left(\sigma_{ij,t}/s_{ij,t}\right)$ remains as in SNE. It is minimized by gradient descent, with the gradient being

$$\partial C_{t-SNE}/\partial \mathbf{x}_v = 4 \sum_{i\in\mathcal{I}\backslash\{v\}} \left(\sigma_{vi,t} - s_{vi,t}\right)\left(\mathbf{x}_v - \mathbf{x}_i\right)\left(1 + d_{vi}^2\right)^{-1} \qquad (1)$$

when using the Euclidean LD distance. The latter is evaluated in $\mathcal{O}\left(N^2\right)$ time.

The BH t-SNE acceleration approximates (1) with an $\mathcal{O}\left(N \log N\right)$ time complexity [10]. As σ_{ij} vanishes as δ_{ij} grows, sparse HD similarities are defined:

$$\sigma_{ij,s} = \begin{cases} \frac{\exp\left(-\pi_{i,s}\delta_{ij}^2/2\right)}{\sum_{k\in\mathcal{S}^i} \exp\left(-\pi_{i,s}\delta_{ik}^2/2\right)} \text{ if } j \in \mathcal{S}^i \\ 0 \text{ otherwise} \end{cases},\ \pi_{i,s} \text{ s.t. } \log K_\star = \sum_{j\in\mathcal{S}^i} \sigma_{ij,s} \log \frac{1}{\sigma_{ij,s}},$$

where $\mathcal{S}^i$ indexes the $\lfloor 3K_\star \rceil$ nearest neighbors of $\boldsymbol{\xi}_i$ in $\boldsymbol{\Xi}\backslash\{\boldsymbol{\xi}_i\}$. These are derived for all $i \in \mathcal{I}$ in $\mathcal{O}\left(K_\star N \log N\right)$ time by creating a vantage-point tree on $\boldsymbol{\Xi}$ [10]. The sparse t-SNE HD similarities then develop as $\sigma_{ij,ts} = \left(\sigma_{ij,s} + \sigma_{ji,s}\right)/\left(2N\right)$.

After replacing $\sigma_{vi,t}$ with $\sigma_{vi,ts}$ in (1), a BH algorithm estimates the remaining non-sparse sums $\sum_{i\in\mathcal{I}\backslash\{v\}} s_{vi,t}\left(\mathbf{x}_v - \mathbf{x}_i\right)\left(1 + d_{vi}^2\right)^{-1} =: F_v$ for all $v \in \mathcal{I}$ in $\mathcal{O}\left(N \log N\right)$ average time [10]. Note that if $d_{ij} \ll d_{vi} \approx d_{vj}$, then $s_{vi,t} \approx s_{vj,t}$. Furthermore P is usually small in visualization tasks. A quad- ($P = 2$) or octree ($P = 3$) can hence be created on $\mathbf{X}$. Performing a depth-first search then approximates F_v in $\mathcal{O}\left(\log N\right)$ average time [10]. At each node during the traversal, one determines whether the LD points contained in the corresponding cell can be represented by their center of mass $\mathbf{x}_{cell}$, using the condition $r_{cell}\,/d_{v,cell} < \theta$, where r_{cell} is the cell diagonal length, $d_{v,cell}$ is the LD distance between $\mathbf{x}_v$ and $\mathbf{x}_{cell}$ and $\theta \in [0,1]$ is a user-defined threshold hyper-parameter trading off accuracy and speed. When the latter is satisfied, the depth-first search stops and $\mathbf{x}_{cell}$ summarizes the contributions to F_v of the LD points in the current cell.

The BH algorithm, along with the $\sigma_{ij,s}$, approximates (1) in $\mathcal{O}\left(N \log N\right)$ average time, as the perplexity $K_\star$ is usually small compared to N. The estimated gradient can in turn be used in gradient-based minimization schemes.

3 Dimensionality reduction quality assessment

Some studies developed DR quality criteria measuring the HD neighborhoods preservation in the LDS [11], becoming generally adopted in several publications [5, 7]. Let ν_i^K and n_i^K index the K nearest neighbors of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ in the HD and LDS, with $Q_{NX}\left(K\right) = \sum_{i\in\mathcal{I}} \left|\nu_i^K \cap n_i^K\right| / (KN) \in [0,1]$ measuring their average normalized agreement. Its expectation being $K/\left(N-1\right)$ for random LD points, $R_{NX}\left(K\right) = \left(\left(N-1\right) Q_{NX}\left(K\right) - K\right)/\left(N - 1 - K\right)$ allows comparing different neighborhood sizes [6]. The curve is often displayed with a log-scale for K as closer neighbors typically prevail. Evaluated in $\mathcal{O}\left(N^2 \log N\right)$ time [11], its area

$$AUC = \left(\sum_{K=1}^{N-2} R_{NX}\left(K\right)/K\right) \Bigg/ \left(\sum_{K=1}^{N-2} K^{-1}\right) \in [-1,1]$$

grows with the DR quality, quantified at all scales with an emphasis on small ones.

4 Experimental comparison

To study diversified public databases, the performance comparison of t-SNE and its BH acceleration is conducted on the COIL-20 data set (COIL) [12], the B. Frey's images (B.Freys), the Olivetti faces (Oliv.), the MNIST test set (MNIST) [13], three three-dimensional artificial sets (Sphere, Clusters and Torus), respectively distributed on a unit-radius sphere, on eight equal-size Gaussian clusters and on a torus, as well as on various UCI data sets [14]: Glass Identification (Gl.id.), Breast Tissue (Br.Tis.), Ecoli, Wine, Parkinsons (Park.) [15], Yacht Hydrodynamics (Ya.Hyd.), Seeds, Concrete Slump Test (Conc.Sl.) [16], Servo, Optical Recognition of Handwritten Digits test set (Digits), Haberman's Survival (Hab.), Iris, Fertility (Fert.) [17], Planning Relax (Pl.Rel.) [18], Leaf [19], Wholesale customers (Whol.c.) [20], Istanbul Stock Exchange (Ist.St.) [21], User Knowledge Modeling (Us.kn.) [22], Indian Liver Patient (ILP), SPECTF Heart (SPECTF), Concrete Compressive Strength (CCS) [23], Ionosphere (Iono.), Connectionist Bench on Vowel Recognition and Deterding Data (Con.v.) and Breast Cancer Wisconsin Diagnostic (BCW). Figure 1 indicates the number of samples N and features M for all the data sets, being of moderate size to allow t-SNE application. Z-score standardization is applied to the UCI data sets when the ranges of their features differ. Potential data labels are ignored. The employed optimization schemes and experimental setup are as described in [8, 10]: the LD dimension P is set to 2 and Euclidean distances are used in both the HD and LDS. Unlike [8, 10], different perplexities are studied, expressed both absolutely and relatively with respect to N. Absolute perplexities are usually employed in the literature, while relative ones aim to ease comparison across data sets.

Figure 1 displays the AUC difference between BH t-SNE and t-SNE for the different data sets, perplexities and thresholds. Observing the y-axis scales and as the AUC ranges in $[-1, 1]$, it is first noteworthy that the BH acceleration performs extremely similarly to t-SNE, sometimes even slightly better, except occasionally for very large perplexities compared to N. This, however, hardly matters as rather small perplexities are employed in practice. Second, although a larger θ implies cruder BH estimations of (1), the embedding quality is almost independent of θ, most curves being close to horizontal except for some high perplexities. The same holds regarding the computation times of the BH acceleration, not provided due to space limitations, except for larger data sets for which, as reported in [10], they tend to be roughly constant from $\theta = 0.3$ to 1 and rapidly increasing for smaller thresholds. These combined results suggest that tuning θ may be avoided by setting it to a reasonable value, such as 0.5. This allows saving cumbersome model selection procedures in practice.

The above conclusions also hold when analyzing the nearest HD neighbors retrieval instead of the AUC, the $R_{NX}(K)$ curves or the LD embeddings, not shown due to space limits. The Zoutendijk condition further supports them [24]: as long as the dot product between (1) and its BH estimate remains positive, gradient descent with adequate step sizes leads to a C_{t-SNE} stationary point.

5 Conclusion

This paper studies the DR quality of the BH acceleration of t-SNE. Although it impressively reduces t-SNE computation time [10], the embedding quality resulting from its approximations has barely been quantified in the past. Thanks to neighborhood-based DR performance criteria, t-SNE and its fast BH version are extensively compared on real-world and artificial databases. The results tend to show that the methods behave extremely similarly. The new hyper-parameter in the BH scheme furthermore hardly influences the DR quality, which spares costly model selection. The BH method hence improves t-SNE in all respects.

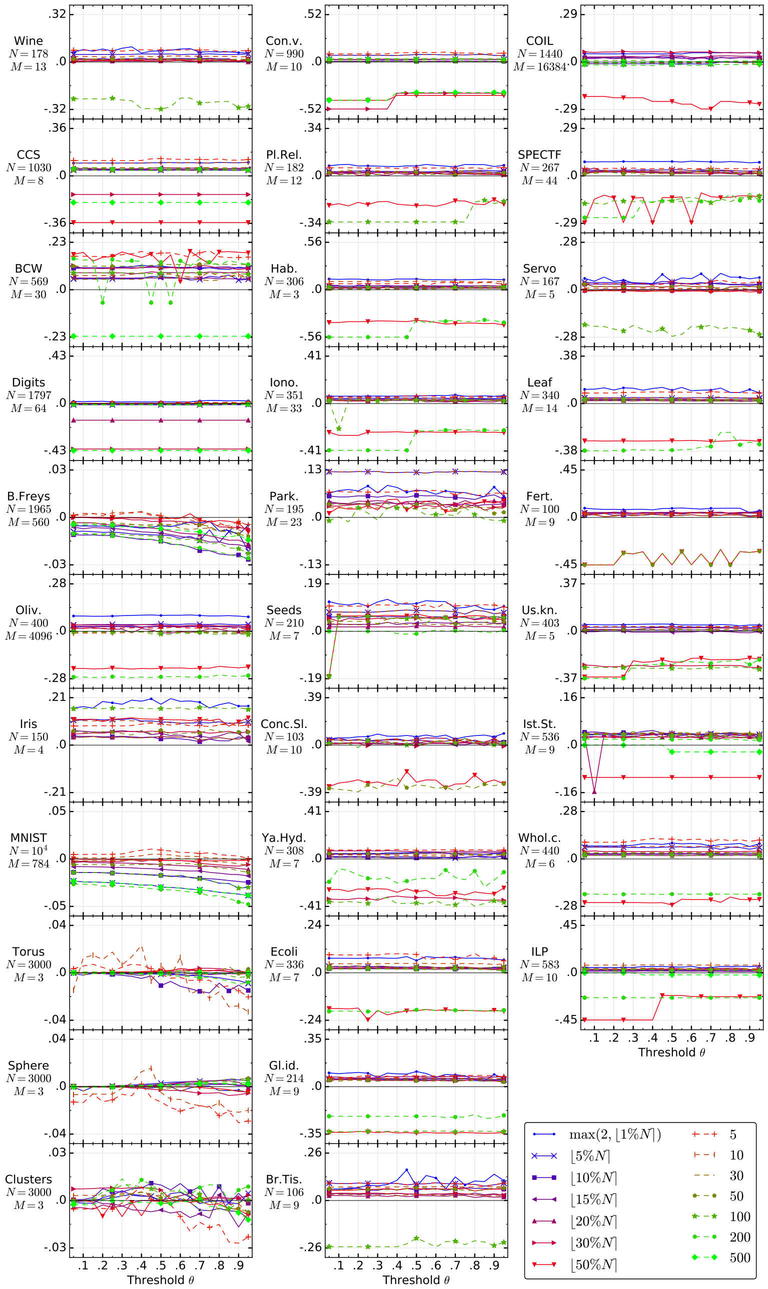

Fig. 1: AUC difference between BH t-SNE and t-SNE as a function of θ, ranging from .05 to .95 with a .05 step, on all data sets. Positive (negative) values show that the BH scheme out- (under-) performs t-SNE. The legend lists all the studied perplexities $K_\star$. Each $K_\star$ is used on the data sets with more than $K_\star$ samples.

References

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

[2] J. A. Lee, A. Lendasse, and M. Verleysen. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, 57:49–76, 2004.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] Z. Yang, J. Peltonen, and S. Kaski. Scalable Optimization of Neighbor Embedding for Visualization. In *ICML (2)*, pages 127–135, 2013.

[10] L. Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

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

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

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] M. Lichman. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2013. University of California, Irvine, School of Information and Computer Sciences.

[15] M. A. Little, P. E. McSharry, S. J. Roberts, D. A.E. Costello, and I. M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, 2007.

[16] I-C. Yeh. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6):474–480, 2007.

[17] D Gil, JL Girela, J De Juan, MJ Gomez-Torres, and M Johnsson. Predicting seminal quality with artificial intelligence methods. *EXPERT SYST APPL*, 39(16):12564–12573, 2012.

[18] R. Bhatt. Planning-Relax Dataset for Automatic Classification of EEG Signals'. UCI.

[19] PFB Silva, ARS Marcal, and RMA da Silva. Evaluation of features for leaf discrimination. *Lecture Notes in Computer Science*, 7950:197–204, 2013. Springer.

[20] N. Abreu. Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. ISCTE-IUL, Lisbon, 2011. Mestrado em Marketing.

[21] O. Akbilgic, H. Bozdogan, and M. E. Balaban. A novel Hybrid RBF Neural Networks model as a forecaster. *Statistics and Computing*, 24(3):365–375, 2014.

[22] HT Kahraman, S Sagiroglu, and I Colak. The development of intuitive knowledge classifier and the modeling of domain dependent data. *KNOWL-BASED SYST*, 37:283–295, 2013.

[23] I-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.

[24] G. Zoutendijk. *Methods of feasible directions: a study in linear and non-linear programming*. Elsevier, 1960.