ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Impact of data subsamplings in Fast Multi-Scale Neighbor Embedding

Pierre Lambert[1], John A. Lee[1,2], Michel Verleysen[1], Cyril de Bodt[1]

1- UCLouvain.be - ICTEAM/ELEN
Place du Levant 3 L5.03.02, 1348 Louvain-la-Neuve - Belgium

2- UCLouvain.be - IREC/MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium

Abstract. Fast multi-scale neighbor embedding (f-ms-NE) is an algorithm that maps high-dimensional data to a low-dimensional space by preserving the multi-scale data neighborhoods. To lower its time complexity, f-ms-NE uses random subsamplings to estimate the data properties at multiple scales. To improve this estimation and study the f-ms-NE sensitivity to randomness, this paper generalizes the f-ms-NE cost function by averaging several subsamplings. Experiments reveal that this can slightly improve the quality of the embeddings while maintaining reasonable computation times. Codes are available at `https://github.com/cdebodt/Fast_Multi-scale_NE`.

1 Introduction

Dimension reduction (DR) maps high-dimensional (HD) data to a low-dimensional (LD) space such that the LD embedding faithfully represents the HD data. Some information is typically lost in the DR process and, therefore, the faithfulness of a projection is considered with respect to some criterion. The main uses of DR are in machine learning, to curb the curse of dimensionality, and in visualisation.

When HD dimension is high, the Euclidean distances between points tend to concentrate towards similar values [1]. Because of this phenomenon, using distance preservation as a direct criterion for DR can become problematic. Neighbor embedding (NE) techniques such as stochastic neighbor embedding (SNE) [2] alleviate the effects of distance concentration by defining neighbor probability distributions in both spaces to embed points in LD [3]. These methods are frequently used in visualization as they tend to produce embeddings that show well separated, readable clusters and nicely preserve local neighborhoods [4].

Originally, NE algorithms required the user to specify a scale for the neighborhood preservation. By capturing the structures on a single scale, embeddings produced by these methods can lead the user to erroneous interpretations. More recently, multi-scale NE approaches were introduced by combining neighborhood probabilities tuned across various scales [5, 6]. Multi-scale approaches often yield better results by preserving the distant structures, while still ensuring good local neighborhood preservation. However, multi-scale NE methods have a time complexity of $\mathcal{O}(N^2 \log N)$ with N data points, restricting their use to data sets of moderate size. A fast version of multi-scale NE (f-ms-NE) [7] has a time complexity of $\mathcal{O}(N \log^2 N)$, with a slight decrease in neighbor preservation as

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

a trade-off. To achieve such an acceleration, f-ms-NE relies on random data subsamplings to capture both the local and global structures in the HD data. However, the stability of the f-ms-NE performances with respect to the number of these subsamplings has not been characterized previously.

This paper hence generalizes f-ms-NE by defining cost functions averaged over multiple random subsamplings, enabling to study the effect of their number on the LD embeddings. In addition to aiming at increasing DR quality while preserving manageable computation times, this work further assesses the sensitivity of f-ms-NE to randomness. Public codes are freely available at `https://github.com/cdebodt/Fast_Multi-scale_NE`.

This paper is structured as follows: Section 2 summarizes NE algorithms. Section 3 explains how f-ms-NE is adapted to consider multiple random subsamplings. Section 4 details experimental results by comparing the performances of the f-ms-NE algorithm across an increasing number of subsamplings, in both terms of DR quality and speed. Final conclusions are drawn in Section 5.

2 Neighbor embedding algorithms

The f-ms-NE algorithm accelerates multi-scale NE, which is based on single-scale NE. Section 2.1 first sketches single-scale NE, Section 2.2 then describes its multi-scale extension, and Section 2.3 finally details the f-ms-NE acceleration.

2.1 Single-scale neighbor embedding

Let $\mathbf{\Xi} = [\xi_i]_{i=1}^N$ denote a set of N points in a HD space with M features. Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ represent these data in a P-dimensional LD space, $P \leq M$. The HD and LD distances between the i^{th} and j^{th} points are δ_{ij} and d_{ij}, respectively, for $i \in \mathcal{I} = \{1, ..., N\}$ and $j \in \mathcal{I} \setminus \{i\}$. SNE [2] aims at preserving pairwise similarities from HD to LD, which are respectively defined as

$$\sigma_{ij} = \frac{\exp(-\pi_i \delta_{ij}^2/2)}{\sum_{k\in\mathcal{I}\setminus\{i\}} \exp(-\pi_i \delta_{ik}^2/2)},\ s_{ij} = \frac{\exp(-d_{ij}^2/2)}{\sum_{k\in\mathcal{I}\setminus\{i\}} \exp(-d_{ik}^2/2)},\ \sigma_{ii} = s_{ii} = 0\ . \quad (1)$$

Precision π_i is adapted to the local density in order to tune the HD similarities to a target scale. The scale is provided by the user in the form of a perplexity K for the distribution $[\sigma_{ij}; j \in \mathcal{I} \setminus \{i\}]$ such that $\log K = -\sum_{j\in\mathcal{I}\setminus\{i\}} \sigma_{ij} \log \sigma_{ij}$.

SNE interprets these normalized similarities as neighborhood probabilities around each point, in order to produce a LD embedding that minimises the sum of Kullback-Leibler (KL) divergences $C_{SNE} = \sum_{i\in\mathcal{I}, j\in\mathcal{I}\setminus\{i\}} \sigma_{ij} \log(\sigma_{ij}/s_{ij})$.

The t-SNE extension [4] symmetrizes the similarities and considers a Student t function with one degree of freedom in LD space, to cope with 'crowding' problems in LD. The HD similarities τ_{ij} and LD ones t_{ij} are now defined as

$$\tau_{ij} = (\sigma_{ij} + \sigma_{ji})/(2N),\ t_{ij} = \frac{(1 + d_{ij}^2)^{-1}}{\sum_{k\in\mathcal{I}, l\in\mathcal{I}\setminus\{k\}} (1 + d_{kl}^2)^{-1}},\ \tau_{ii} = t_{ii} = 0\ . \quad (2)$$

2.2 Multi-scale neighbor embedding

The multi-scale SNE method [5] combines SNE similarities computed with exponentially increasing perplexities. Single-scale similarities are indexed by a scale counter h for $i \in \mathcal{I}$ and $j \in \mathcal{I} \setminus \{i\}$:

$$\sigma_{ijh} = \frac{\exp(-\pi_{ih}\delta_{ij}^2/2)}{\sum_{k\in\mathcal{I}\setminus\{i\}} \exp(-\pi_{ih}\delta_{ik}^2/2)},\ s_{ijh} = \frac{\exp(-p_{ih}d_{ij}^2/2)}{\sum_{k\in\mathcal{I}\setminus\{i\}} \exp(-p_{ih}d_{ik}^2/2)},\ \sigma_{iih} = s_{iih} = 0\ .$$

HD precisions π_{ih} are set using perplexities growing as $K_h = 2^{h-1}K_1$, with a small base perplexity $K_1 = 2$ and $1 \leq h \leq H = \lfloor \log_2(N/K_1) \rceil$, where $\lfloor \cdot \rceil$ denotes rounding. LD precisions $p_{ih} = p_h = K_h^{-2/P}$ follow the exponential growth of the HD precisions. Multi-scale similarities average single-scale ones as

$$\sigma_{ij} = H^{-1} \sum\nolimits_{h=1}^{H} \sigma_{ijh},\ \ s_{ij} = H^{-1} \sum\nolimits_{h=1}^{H} s_{ijh}\ . \tag{3}$$

The authors recommend using L-BFGS [8] to optimise the cost function.

2.3 Fast multi-scale neighbor embedding

Single-scale and multi-scale NE methods have time complexities of $\mathcal{O}(N^2)$ and $\mathcal{O}(N^2 \log N)$, respectively. The typical smallness of perplexity K with respect to N enables single-scale methods to consider that sufficiently distant points have a null HD similarity. Sparse HD similarities can be computed by finding the sets of nearest neighbors of each point using vantage-point trees [9]. This reduces their computation time to $\mathcal{O}(KN \log N)$. On the LD side, a Barnes-Hut (BH) algorithm [10, 11] approximates efficiently the similarities by relying on tree structures to compound far-away points, dispensing with computing all pairwise interactions, reducing the time complexity to $\mathcal{O}(N \log N)$. The BH method uses a threshold parameter $\theta \in [0, 1]$ that determines whether to use the approximation or not, with higher values meaning rougher but faster estimations.

The largest scale used to compute multi-scale similarities has a perplexity K_H in the $\mathcal{O}(N)$ range. At such a scale, the above sparse HD similarity approximation would require neighbor sets of $\mathcal{O}(N)$ size, defeating the purpose of accelerating the computations. For this reason, reducing the time complexity of multi-scale NE requires a different approach for the HD side. In [7], the authors tackle this problem by defining small multi-scale neighbor sets $\widetilde{\mathcal{I}}_i$ for $i \in \mathcal{I}$.

To compute $\widetilde{\mathcal{I}}_i$, the authors define a hierarchy of subsampled HD data sets $\{\Xi_h\}_{h=1}^{H}$, where Ξ_h is a random sample of Ξ with $\lfloor 2^{1-h}N \rceil$ elements drawn without replacement. A vantage-point tree is created on Ξ_h for each scale $h \in \{1, ..., H\}$; the trees are generated in $\mathcal{O}(N \log N)$ time. For each scale, the corresponding tree is used to compute the neighborhood $\widetilde{\mathcal{I}}_{ih}$ within the sample Ξ_h for $i \in \{1, ..., N\}$; this takes a total of $\mathcal{O}(N \log^2 N)$ time. By having samples $\{\Xi_h\}_{h=1}^{H}$ that decrease in size when the scale grows, the accounted neighbors $\widetilde{\mathcal{I}}_{ih}$ are more and more dispersed in the data cloud, hence capturing larger scale properties. The multi-scale neighbor sets are then defined as $\widetilde{\mathcal{I}}_i := \cup_{h=1}^{H} \widetilde{\mathcal{I}}_{ih}$.

Like the non-accelerated version, f-ms-NE uses multi-scale similarities averaged over sparse single-scale similarities,

$$\widetilde{\sigma}_{ij} = H^{-1}\sum_{h=1}^{H}\widetilde{\sigma}_{ijh}, \quad \text{with} \quad \widetilde{\sigma}_{ijh} = \begin{cases} \frac{\exp(-\widetilde{\pi}_{ih}\delta_{ij}^2/2)}{\sum_{k\in\widetilde{\mathcal{I}}_i}\exp(-\widetilde{\pi}_{ih}\delta_{ik}^2/2)} & \text{if } j\in\widetilde{\mathcal{I}}_i \\ 0 \text{ otherwise} & \end{cases} .$$

Precision $\widetilde{\pi}_{ih}$ is fixed such that

$$K_1 = -\sum_{j\in\widetilde{\mathcal{I}}_{ih}}\widetilde{\sigma}_{ijh}^{\pi}\log\widetilde{\sigma}_{ijh}^{\pi} \quad \text{where} \quad \widetilde{\sigma}_{ijh}^{\pi} = \begin{cases} \frac{\exp(-\widetilde{\pi}_{ih}\delta_{ij}^2/2)}{\sum_{k\in\widetilde{\mathcal{I}}_{ih}}\exp(-\widetilde{\pi}_{ih}\delta_{ik}^2/2)} & \text{if } j\in\widetilde{\mathcal{I}}_{ih} \\ 0 \text{ otherwise} & \end{cases} .$$

In the $t-$distributed version, the authors use symmetrised HD similarities $\widetilde{\tau}_{ij} = (\widetilde{\sigma}_{ij}+\widetilde{\sigma}_{ji})/(2N)$ and optimise the cost function $\widetilde{C}_t = -\sum_{i\in\mathcal{I},j\in\mathcal{I}\setminus\{i\}}\widetilde{\tau}_{ij}\log t_{ij}$.

3 Multiple subsamplings in fast multi-scale NE

In f-ms-NE, the larger-scale structures of the HD data are captured by computing neighborhoods on samples of decreasing sizes. Therefore, f-ms-NE estimates the larger-scale data properties with some sensitivity to randomness, as an unlucky sampling could bypass some important structures within the data. We hence propose to adapt f-ms-NE by using multiple subsamplings of the HD data, aiming at reducing the impact of randomness when modeling large-scale structures.

The proposed adaptation is to perform the whole sampling process R times, to compute sets $\widetilde{\mathcal{I}}_{ir}$ for $i\in\{1,...,N\}$ and $r\in\{1,...,R\}$. The sparse similarities are computed independently for each batch and the cost function becomes

$$\widetilde{C}_t = R^{-1}\sum_{r=1}^{R}\left(-\sum_{i\in\mathcal{I},j\in\mathcal{I}\setminus\{i\}}\widetilde{\tau}_{ijr}\log t_{ij}\right) . \tag{4}$$

As R is set to a small value independent of N, the time complexity with regards to N remains unchanged. But as the original cost function is computed R times, a computation time increase is expected. However, each repetition being independent, a parallel implementation with R threads can easily be done.

4 Experimental results and discussion

The method is tested on seven data sets of size N and dimensionality M [12]. Anuran: (N,M) = (7195, 22); Plant: (N,M) = (9568, 4); Gestures: (N,M) = (9901, 18); Satellite: (N,M) = (6434, 36); Waveform: (N,M) = (5000, 40); Theorem: (N,M) = (6118, 51); and Musk: (N,M) = (6598, 166). On each set and for each $R\in\{1,...,5\}$, f-ms-NE with multiple subsamplings is applied 30 times using a random initialisation, and 30 times with a PCA initialisation. The t-distributed version of the algorithm is used and the target dimension P is 2 for all experiments. Each optimization consists of 30 iterations of the L-BFGS algorithm and uses a BH threshold $\theta = 0.75$, as these values tend to produce good DR quality in reasonable time for the original f-ms-NE [7].

Figure 1 shows the quality of the resulting embeddings. As in [7], their quality is measured by the area under the curve (AUC) of the relative neighborhood preservation R_{NX} [13], with a logarithmic scale for the neighborhood size. The AUC of R_{NX} reaches 1 when the preservation is perfect for all data scales.

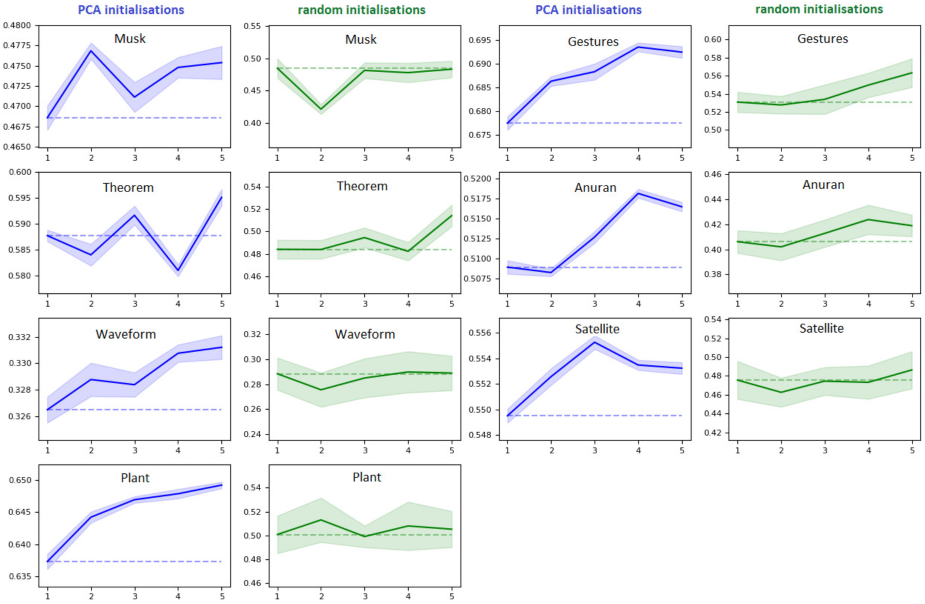

Fig. 1: Evolution of the AUC of the R_{NX} curves with the number of subsamplings. The y-axis is the AUC, and the x-axis is the number R of subsamplings. The solid lines are the mean values over 30 trials, the shaded areas correspond to one standard deviation on each side. The dashed line corresponds to the mean AUC for the original version of f-ms-NE, which uses one sampling.

We observe that subsampling multiple times has an effect on the DR quality of the embeddings that f-ms-NE yields. However, the change is often slight when compared to embeddings produced from a single sampling and, surprisingly, the change is not always positive. Using multiple subsamplings in conjunction with a PCA initialisation tends to produce a more consistent increase in quality, but the magnitude of the increase depends on the data set. The modest benefits of using multiple subsamplings can suggest that the construction of $\widetilde{\mathcal{I}}_i$ in the original f-ms-NE algorithm is already quite robust to randomness.

In terms of computation time, we obtain a consistent linear increase in time with respect to R. With a single thread implementation on the tested computer, using $R = 5$ requires approximately twice the time that the algorithm would take in its original version.

5 Conclusion

Experimental results indicate that using multiple subsamplings in f-ms-NE can provide improvements to the DR quality. The moderate nature of these improvements hints that the sampling process used by f-ms-NE enables to model the multi-scale structures of the HD data with a remarkable robustness with respect to the randomness.

To further study the influence of randomness on f-ms-NE, alternate subsampling strategies will be envisioned. For instance, using a biased sampling to increase the diversity of the sampled points in the hierarchy of subsampled HD data sets $\{\Xi_h\}_{h=1}^{H}$ is likely to improve the estimation of the global structure of the HD data.

References

[1] D. François, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19:873–886, 2007.

[2] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[3] John A. Lee and Michel Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4:538–547, 2011.

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

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

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

[7] Cyril de Bodt, Dounia Mulders, Michel Verleysen, and John Aldo Lee. Fast multiscale neighbor embedding. *IEEE T NEUR NET LEAR*, pages 1–15, 2020.

[8] R. Byrd, Peihuang Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16:1190–1208, 1995.

[9] Peter Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. volume 93, 01 1993.

[10] Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014.

[11] Cyril de Bodt, Dounia Mulders, Michel Verleysen, and John A. Lee. Extensive assessment of Barnes-Hut t-SNE. In *ESANN*, pages 135–140, 2018.

[12] Moshe Lichman. UCI Machine Learning repository, 2013.

[13] John Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72:1431–1443, 03 2009.