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 Parameterless t-SNE for Faithful Cluster Embeddings from Prototype-based Learning and CONN Similarity

Josh Taylor[1] and Erzsébet Merényi[2]

1- Rice University - Department of Statistics
6100 Main St, Houston, Texas, USA

2- Rice University - Departments of Statistics & Electrical and Computer Engineering

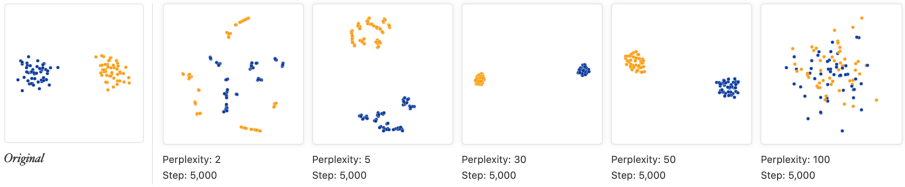
Abstract. We propose an improvement to t-SNE which allows automated specification of its perplexity parameter using topological information about a data manifold revealed through neural prototype-based learning. This information is contained in the CONN (CONNectivity) similarity of neural prototypes, which expresses the strength (weakness) of topological connectivity at various points within the manifold. Experiments show that improvements, collectively called **CONNt-SNE**, are capable of producing meaningful and trustworthy low-dimensional embeddings without the need to heuristically optimize over (i.e., grid search) t-SNE's perplexity space. Data-driven perplexity determination improves our confidence that any structure appearing in the embeddings is valid and not merely an artifact of spurious parameterization.

1 Background

t-SNE [1] has attracted wide attention both within and outside the machine learning community as a tool for producing low-dimensional non-linear embeddings $T = \{t_s\}_{s=1}^{N} \in \mathbb{R}^{d'}$ of high-dimensional point clouds $X = \{x_s\}_{s=1}^{N} \in \mathbb{R}^{d}$, where $d' << d$, for exploratory (visual) data analysis. Typically $d' \in \{2, 3\}$. The appetite for such analysis across disciplines is strong, but many questions have been raised about what, exactly, can (should) be inferred from a t-SNE embedding. t-SNE's introduction subtly stresses its distinction as a technique only for visualization (vs. feature engineering/extraction), yet its embeddings are often clustered either informally (via visual assessment) or formally (applying a clustering algorithm to T). Some [2] have noticed relative deficiencies in t-SNE's ability to faithfully indicate separation in complex manifolds. [3] offers a list of various misinterpretations that can be made from a t-SNE embedding due to its unfaithful representation of cluster sizes, shapes, densities, compactness and separability. Most of these issues arise because t-SNE is designed to preserve conditional probabilities between points instead of distance. We believe these issues are not severe impediments to successful cluster discovery from low-d representations; indeed, over the last three decades the lattice representations of data learned by Self-Organizing Maps [4] have produced many successful clusterings without explicit preservation of, e.g. scatter, between the high- and low-d spaces. However, [3] does raise one issue we feel fundamentally impacts the fidelity of a t-SNE representation: that of selecting its main "perplexity" parameter, which

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/.

we abbreviate px. px indirectly controls the number of neighbor similarities that t-SNE attempts to preserve. An example taken from [3] of the various t-SNE embeddings which can arise from different px specifications is given below.



Here, the original data (left-most panel) is very simple–two dimensional with two well-defined clusters–yet inspection of the embeddings resulting from some perplexity values (2, 5, 100) would yield a different conclusion. [1] suggests that t-SNE is relatively insensitive to different values but in practice an optimal value is obviously data-dependent and should be data-driven. CONNt-SNE provides a mechanism for such, using information freely available and commonly invoked during prototype-based clustering.

1.1 The t-SNE Algorithm

The t-SNE algorithm begins by defining Gaussian similarities between two points in $X \in \mathbb{R}^d$ as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_k - x_i||^2/2\sigma_i^2)} \tag{1}$$

where $p_{\cdot|i}$ is the conditional distribution of all other x_j given x_i and, by convention, $p_{i|i} = 0$. We let $P = \{p_{ij}\}$ be the $N \times N$ matrix of such (symmetrized) similarities and denote its i-th row by P_i. Each Gaussian bandwidth σ_i is controlled by the (global) perplexity parameter px, set (through iterative search) such that following relationship holds:

$$px = 2^{H(P_i)}, \quad H(P_i) = -\sum_j p_{j|i} \log_2(p_{j|i}) \tag{2}$$

Pointwise similarities in $\mathbb{R}^{d'}$ are derived from the pdf of the Student's t-distribution with one degree of freedom: $q_{ij} = \frac{(1+||t_i - t_j||^2)^{-1}}{\sum_{k \neq l}(1+||t_k - t_l||^2)^{-1}}$, where again we let $Q = \{q_{ij}\}$. Coordinates t_i are determined through minimization of the Kullback-Leibler divergence as cost, $C = KL(P||Q)$.

1.2 CONN Similarity

CONNt-SNE provides a framework for embedding the *prototypes* $W = \{w_i\}_{i=1}^M \in \mathbb{R}^d$, $M << N$, of a vector quantizer (VQ) trained on data X. While the prototypes of any VQ would suffice for this purpose we prefer neural variants such 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/.

the SOM and Neural Gas (NG, [5]) as the iterative stages of competition and cooperation during training result in better prototype placement in the data cloud [6]. Previous work [7] utilized t-SNE as a means to visualize Neural Gas prototypes but, contrary to this work, did not explore any ways by which t-SNE could be influenced by the VQ. The CONN similarity [8] between trained prototypes w_i and w_j, $\text{CONN}_{ij} = CADJ_{ij} + CADJ_{ji}$, can be calculated from a recall of the entire dataset, where $CADJ_{ij} = \sum_s I(BMU1(x_s) = i \wedge BMU2(x_s) = j)$, $BMU\{1,2\}$ are the index of the 1^{st} and 2^{nd} **B**est **M**atching **U**nits (prototypes) and $I()$ is the indicator function. CADJ_{ij} (the Cumulative ADJacency of i and j) reports the number of data vectors observed in the second-order Voronoi cell V_{ij} generated by W, and CONN is its symmetrized version. CONN is thus a weighted version of the Masked Delaunay Triangulation [9] whose entries reflect local data densities within the manifold.

2 CONNt-SNE

CONNt-SNE methodology arises from two key modifications to the original t-SNE algorithm. The first of these permits a varying perplexity px_i when setting each conditional distribution $p_{.|i}$. We now have M different (local) perplexities to specify but CONN provides a data-driven way of determining these parameters as the number of CONN neighbors of prototype i:

$$px_i = max\left(\sum_j I(CONN_{ij} > 0), 5\right).$$

We retained the lower bound $px_i \geq 5$ as suggested in [1] as our experiments show t-SNE can be unstable at low perplexity values. With px_i intelligently and automatically specified, the same procedure of (2) sets each local σ_i (and, consequently, P_i). We denote by P_{var} the matrix of prototype similarities defined using CONN-derived variable perplexity values px_i.

The second modification to t-SNE appends a term to its cost function to allow the CONN similarities to more directly influence the t-SNE embedding, which we do in an additive manner, yielding CONNt-SNE's cost function:

$$C^* = KL(P_{var} \,||\, Q) + KL(P_{CONN} \,||\, Q). \tag{3}$$

$P_{CONN} = CONN$, but normalized to have (1) unity row sums and then (2) unity grand sum, as required of a t-SNE similarity. This new cost term imparts information about the local data densities and is unique to vector quantizers. The additional term in CONNt-SNE's cost does increase runtimes ($\approx 50\%$ longer than t-SNE for the experiments of section 3). We argue this is a minor issue as 1) embedding prototypes (vs. data X) is already orders of magnitude faster due to drastically decreased sample size and 2) CONN is very sparse (most $CONN_{ij}$ values $= 0$), which could be exploited to expedite the additional gradient calculations required during minimization of (3).

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/.

3 Data Experiments

To demonstrate the effectiveness of CONNt-SNE we compare its two-dimensional embeddings to those of t-SNE for two real datasets:

- **MNIST**: 28x28 pixel grayscale images of handwritten digits 0-9. N=70,000; $d = 784$; # classes = 10. 2,000 NG prototypes trained for embedding.
- **Flow18**: Flow cytometry measurements of human peripheral blood mononuclear cells labeled by phenotype, subsampled as in [10]. $N = 946{,}915$; $d = 11$; # classes = 12. 1,225 NG prototypes trained for embedding.

We have omitted experiments on simple synthetic data, which elucidated few differences. The top panel of Figure 1 displays the resulting embeddings using a Principal Components (PCA) initialization and $px = 30$ for t-SNE, which is a widely used default. Points are colored by class (prototypes inherit their class label by plurality vote of the labeled data mapped to them). Both methods produced well organized visual groupings of known classes, with CONNt-SNE exhibiting slightly better cluster retention and separation (note the axis scales– CONNt-SNE repeatedly utilizes larger areas of the t-SNE plane). For MNIST, CONNt-SNE delineates digits 4 and 9 more cleanly and shows less confusion between 3 and 5. For Flow18, CONNt-SNE has more meaningfully separated the light blue B-cells from salmon-colored Lin- cells; Lin- stem cells differentiate to form B-cells, so they are biologically related but distinct. CONNt-SNE has also retained a visual "bridge" between CD14 expressed monocytes (red and mint green clusters); as the mint green CD14var monocytes comprise mixed marker expression (CD14+/-), we believe this is a more faithful representation of the underlying data relationships. From these observations we conclude CONNt-SNE more sensitive to the subtleties of complex data.

Beyond visual inspection we have also attempted a quantitative assessment of the embedding by measuring various cluster criteria on the prototype clusters *in the embedded space* (not in $\mathbb{R}^d$). Internal measures (**D**avies-**B**ouldin **Index**, **G**eneralized **D**unn **I**ndex, **SIL**houette Index) compare a known partition to ratios of within-cluster scatter and between-cluster separation to suggest how well point placement agrees with the partition. External measures (**A**djusted **R**and **I**ndex, **JAC**card Similarity, **N**ormalized **M**utual **I**nformation) report the concordance of truth and predicted labels. For the latter, we produced four different clusterings of each embedding using Hierarchical Agglomerative Clustering (complete, average, and single linkage) and k-means. The true # of clusters was used to guide dendrogram cutting (for HAC) and centroid specification (k-means).

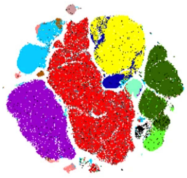
As layout initialization affects t-SNE's quality we have repeated these measurements for 11 different t-SNE initializations (PCA + 10 randomly seeded) for each dataset and method. The boxplots in the bottom panel of Figure 1 compare summary statistics of CONNt-SNE's cluster criteria to t-SNE with perplexities $\in \{10, 30, 50\}$ (the general range suggested in [1]). For all criteria a higher value is preferable except DBI where a lower value indicates better performance. According to almost all criteria summarized in the boxplots, CONNt-SNE's median

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/.

performance is at least as good as the best-performing t-SNE parameterization, mimicking $px = 30$ for MNIST and $px = 10$ for Flow18 (notable exceptions are the GDI and SIL indexes for Flow18).

4 Conclusions and Future Work

Prototype representations of data simultaneously reduce sample size and boost signal-to-noise ratios. This not only alleviates the computational burden of applying t-SNE directly to data, but also results in higher quality embeddings, as is visible when comparing the well separated structure of Figure 1 to a less separated, previously published t-SNE embedding of the entire Flow18 dataset from [10], inset right. Further, our data experiments give confidence that CONNt-SNE's automatic parameterization produces embeddings which faithfully and reliably represent cluster structure as well as the best parameterized t-SNE, without the need to heuristically grid search for (visually subjective) optimality. CONNt-SNE's ability to recognize and respond to structural subtleties in real data facilitates more meaningful inference from its embeddings. As CONNt-SNE is new we have many ideas for its further use, including extensions of this framework to other dimensionality reduction techniques and sensible methods to harness the VQ mapping to permit embeddings of new data points without further t-SNE training.



References

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

[2] Kadim Taşdemir and Erzsébet Merényi. SOM-based topology visualisation for interactive analysis of high-dimensional large datasets. *Machine Learning Reports*, 1:13–15, 2012.

[3] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[4] Teuvo Kohonen. *Self-Organizing Maps*. Springer, 2000.

[5] Thomas M. Martinetz and Klaus J. Schulten. A "neural gas" network learns topologies. In Teuvo Kohonen, Kai Mäkisara, Olli Simula, and Jari Kangas, editors, *Proceedings of the International Conference on Artificial Neural Networks 1991* (Espoo, Finland), pages 397–402. Amsterdam; New York: North-Holland, 1991.

[6] Marie Cottrell, Barbara Hammer, Alexander Hasenfuß, and Thomas Villmann. Batch and median neural gas. *Neural Networks*, 19(6-7):762–771, 2006.

[7] Kadim Taşdemir. Dimensionality reduction based similarity visualization for neural gas. In *2014 IEEE International Conference on Data Mining Workshop*, pages 668–675, 2014.

[8] K. Taşdemir and E. Merényi. Exploiting data topology in visualization and clustering of self-organizing maps. *IEEE Transactions on Neural Networks*, 20(4):549–562, April 2009.

[9] Thomas Martinetz and Klaus Schulten. Topology representing networks. *Neural Networks*, 7(3):507 – 522, 1994.

[10] Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):1–12, 2019.

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/.

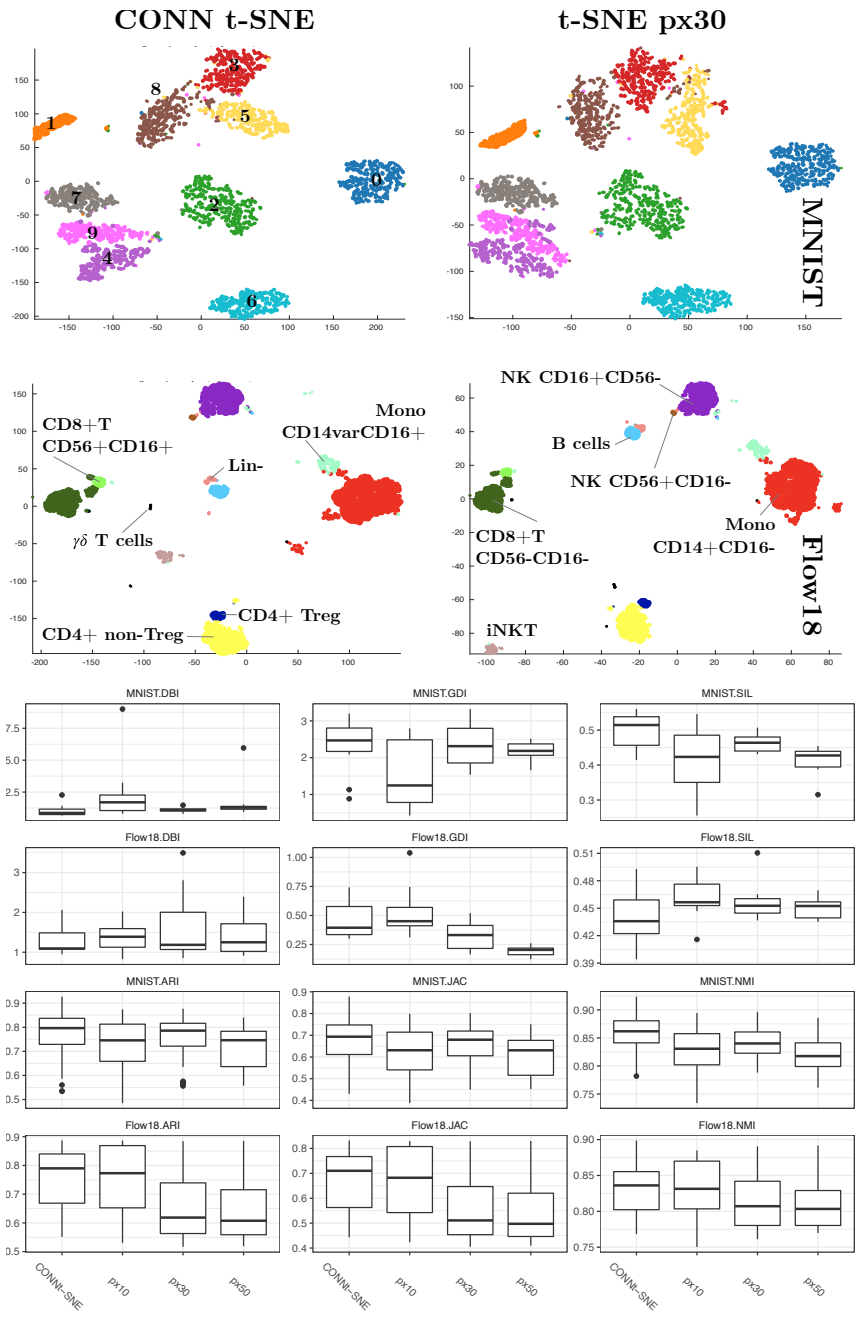


Fig. 1: **Top Panel:** Embeddings of the learned Neural Gas prototypes of the MNIST and Flow18 datasets by CONNt-SNE and standard t-SNE. **Bottom Panel:** Summary statistics of the internal (top two rows) and external (bottom two rows) cluster validity measures, as described in the text.