ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Class-aware *t*-SNE: ca*t*-SNE

Cyril de Bodt[1], Dounia Mulders[1], Daniel López-Sánchez[3],
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

3- Universidad de Salamanca - BISITE Research Group
C/ Espejo S/N, 37007 Salamanca, Spain

Abstract. Stochastic Neighbor Embedding (SNE) and variants like *t*-distributed SNE are popular methods of unsupervised dimensionality reduction (DR) that deliver outstanding experimental results. Regular *t*-SNE is often used to visualize data with class labels in colored scatterplots, even if those labels are actually not involved in the DR process. This paper proposes a modification of *t*-SNE that employs class labels to adjust the widths of the Gaussian neighborhoods around each datum, instead of deriving those from a perplexity set by the user. The widths are fixed to concentrate a major fraction of the probability distribution around a datum on neighbors with the same class. This tends to shrink the bulk of the classes and to stretch their low-dimensional separation. Experimental results show that the proposed class-aware *t*-SNE (ca*t*-SNE) outperforms regular *t*-SNE in KNN classification tasks carried out in the embedding.

1 Introduction

Dimensionality reduction (DR) [1] aims at producing relevant low-dimensional (LD) representations of high-dimensional (HD) data sets. Relevance can cover several aspects of data, such as the preservation of variance (in principal component analysis, PCA), of pairwise distances (in stress-based multidimensional scaling), of pairwise inner products between data mapped in a feature space induced by a kernel which is either user-defined (kernel PCA) or data-driven (locally linear embedding; Laplacian eigenmaps; maximum variance unfolding).

Recently, developments in DR have focused on the preservation of small neighborhoods, with proxies like stochastic neighbor embedding (SNE) [2] and its variants [3, 4, 5], including *t*-SNE [6]. The latter has become very popular due to remarkable performances, especially on data exhibiting a clustered structure. In experiments, the meaningfulness of the clusters in the LD space is often visually assessed by considering data sets with class labels, which are expected to correlate with the perceived clusters. Similarly, DR quality is sometimes indirectly evaluated by measuring how accurately an embedding performs in KNN classification tasks [3, 7]. In practice, though, *t*-SNE and most other SNE variants reduce dimensionality in an unsupervised way, ignoring class labels that might be available.

*DM and CdB are FNRS Research Fellows. JAL is a Senior FNRS Research Associate.

This paper aims at explicitly accounting for class labels in t-SNE to improve KNN accuracy in the LD embedding. For this purpose, we modify the t-SNE adjustment of the individual radius of the normalized Gaussian neighborhood around each datum. Instead of targeting a fixed neighborhood entropy, provided by the user through the perplexity, we adjust the neighborhood radius for neighbors with the same class to cumulate a dominant fraction of the probability distribution. This results in smaller HD neighborhoods near class boundaries than in their bulk, and therefore tends to stretch the former and shrink the latter.

The rest of this paper is organized as follows. Section 2 briefly summarizes related works. Section 3 is a reminder of regular t-SNE, while Section 4 details ca$t$-SNE, our proposed class-aware variant of t-SNE. Next, Section 5 describes the quality assessment of DR, in both unsupervised and supervised settings. Section 6 presents the experiments and discusses their results. Finally, Section 7 draws the conclusions and sketches some perspectives for future works.

2 Related works

Different approaches refine unsupervised DR algorithms to account for class labels [8]. In particular, some studies feed DR methods with supervised HD distances, leading to supervised versions of Isomap [9, 10] and NeRV [3]. A preprocessing step hence deals with the classes, before DR. In contrast, linear projections of HD data may maximize the accuracy either of a KNN classifier in the LD space [11], or of a generative model of the labels given the LD points [12]. The Hilbert-Schmidt independence criterion enables better relating the LD coordinates with the classes in maximum variance unfolding [13]. Other studies require class probabilities for each HD sample and seek for the LD space minimizing their Kullback-Leibler (KL) divergence with LD probabilities induced by an isotropic Gaussian mixture with one component per class [14]. Alternatively, a SNE extension [15] suggests defining several HD neighborhood distributions derived from both the HD coordinates and class information. Minimizing their mismatch with LD neighborhood distributions enables tuning the embedding.

3 SNE and t-SNE

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{i=1}^N$ denote a set of N points in a HD space (HDS) with M features. Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ represent it in a P-dimensional 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 $j \in \mathcal{I}\backslash\{i\}$ [2]:

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

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 HD and LD similarity distributions.

Besides symmetrizing the similarities, t-SNE employs a Student t-distribution with one degree of freedom in the LDS, mitigating the crowding problem [6]:

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

The t-SNE cost function $C_{t-SNE} = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \sigma_{ij,t} \log\left(\sigma_{ij,t} / s_{ij,t}\right)$ remains as in SNE. It is minimized by gradient descent, with iterates scaling in $\mathcal{O}\left(N^2\right)$ time.

4 Accounting for class labels in t-SNE

Let c_i be the class associated with $\boldsymbol{\xi}_i$. Let us next define a condition on the weighted proportion t_i of neighbors $\boldsymbol{\xi}_j$ sharing the same class as $\boldsymbol{\xi}_i$, i.e.,

$$t_i = \sum_{j \in \mathcal{I} \setminus \{i\} | c_j = c_i} \sigma_{ij} > \theta \;, \tag{1}$$

where $t_i \in [0, 1]$ as $\sum_{j \in \mathcal{I} \setminus \{i\}} \sigma_{ij} = 1$, and hyper-parameter θ lies in $[0.5, 1[$ to ensure the majority of class c_i. Precision π_i is then minimized under condition (1), ensuing in the largest neighborhood of $\boldsymbol{\xi}_i$ in which class c_i remains dominant. If no precision π_i fulfills condition (1), for instance if $\boldsymbol{\xi}_i$ is an outlier drown in another class than c_i, then π_i is set to maximize t_i. Although a hyper-parameter θ is introduced in (1), maximizing t_i for all $i \in \mathcal{I}$ would induce unnecessary large π_i for i in class bulks, leading to class burstings in the LD space.

No other change is brought to t-SNE. The perplexity meta-parameter in t-SNE is hence replaced with the threshold θ, between 0.5 (simple majority) and 1 (unanimity). This **c**lass-**a**ware variant of t-SNE is coined as ca$t$-SNE.

5 Quality assessment of dimensionality reduction

Some studies developed quality criteria for unsupervised DR, measuring the HD neighborhood preservation in the LDS [16]. This principle is adopted in several publications [3, 5]. Let ν_i^K and n_i^K denote the K nearest neighbor sets of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ in the HDS and LDS, with $Q_{\text{NX}}(K) = \sum_{i \in \mathcal{I}} \left|\nu_i^K \cap n_i^K\right| / (KN) \in [0, 1]$ measuring their average normalized agreement. As $\mathbb{E}\left[Q_{\text{NX}}(K)\right] = K / (N-1)$ for random LD points, $R_{\text{NX}}(K) = \left((N-1) Q_{\text{NX}}(K) - K\right) / (N - 1 - K)$ allows comparing different neighborhood sizes [4]. It is often displayed with a log-scale for K as closer neighbors typically prevail. The area under the resulting curve, $\text{AUC}\left[R_{\text{NX}}(K)\right] = \left(\sum_{K=1}^{N-2} R_{\text{NX}}(K)/K\right) \Big/ \left(\sum_{K=1}^{N-2} K^{-1}\right)$, lying in $[-1, 1]$, grows with DR quality, quantified at all scales with an emphasis on small ones.

When data come with class labels, unsupervised DR can also be assessed by its performances in classification tasks, by reporting the accuracy of a KNN classifier in the LD space [3, 7]. Following this line, we define the KNN gain as

$$G_{\text{NN}}(K) = \frac{1}{N} \sum_{i \in \mathcal{I}} \frac{\left|\left\{j \in n_i^K \text{ s.t. } c_i = c_j\right\}\right| - \left|\left\{j \in \nu_i^K \text{ s.t. } c_i = c_j\right\}\right|}{K} \;. \tag{2}$$

It averages the gain (or loss, if negative) of neighbors of the same class around each point, after DR. Hence, a positive value correlates with likely improved KNN classification performances. The KNN gain $G_{\mathrm{NN}}(K)$ can also be displayed w.r.t. K, with a log-scale for K. A global score summarizing the curve is provided by its area $\mathrm{AUC}\left[G_{\mathrm{NN}}(K)\right] = \left(\sum_{K=1}^{N-2} \mathrm{G}_{\mathrm{NN}}(K)/K\right) \Big/ \left(\sum_{K=1}^{N-2} K^{-1}\right) \in [-1, 1]$.

6 Experiments

The performances of t-SNE and ca$t$-SNE are compared in terms of both HD neighborhood preservation and KNN gain. The employed data sets include (1) an origin-centered, unit-radius, spherical shell ($N = 1500$, $M = 3$) with class labels defined as $c_i = 3 + \sum_{d=1}^{3} \mathrm{sign}(\xi_{di})$, (2) the COIL-20 database ($N = 1440$, $M = 128^2$) containing images of 20 objects, interpreted as classes, under 72 pose angles [17], (3) a subset of the MNIST handwritten digits ($N = 1500$, $M = 28^2$) [18], with the 10 digits as classes, and (4) UCI Abalone database ($N = 4177$, $M = 8$) in which the abalone sex defines classes [19]. Typical perplexities among $\{8, 16, 32, 64\}$, from small to large, are used with t-SNE, while θ in ca$t$-SNE ranges from 0.5 to 0.9 with 0.1 step. Target dimension P is two for all data sets.

Figure 1 illustrates the results on all but Abalone database, due to space limits. On all data sets, ca$t$-SNE improves the KNN gain over t-SNE, especially with large threshold θ. In particular, ca$t$-SNE with $\theta = 0.9$ is superior to t-SNE according to $\mathrm{G}_{\mathrm{NN}}(K)$ for all K and data sets. Astonishing performances of ca$t$-SNE are also observed on Abalone data set w.r.t. t-SNE. Besides, the perplexity maximizing $\mathrm{AUC}\left[\mathrm{G}_{\mathrm{NN}}(K)\right]$ in t-SNE depends on the considered database.

In the LD embeddings, t-SNE tends to exaggerate clusters, including those due to sampling, irrespective of class labels, like in the spherical manifold. In contrast, ca$t$-SNE keeps tight class bulks, magnifying only regions near or across their separation, especially in the sphere and MNIST data sets. Indeed, condition (1) is easily satisfied in the homogeneous center of the classes, enabling to decrease the precisions π_i of the corresponding HD data points, leading to larger Gaussian neighborhoods. These are readily rendered in the LD space by concentrating the classes, which slightly deteriorates the preservation of small within-class neighborhoods, compared to t-SNE, as indicated by the $R_{\mathrm{NX}}(K)$ curves. On the other hand, class boundaries are magnified in LD, thanks to larger precisions π_i of the concerned $\boldsymbol{\xi}_i$, with tighter Gaussian neighborhoods in HD that get stretched in LD. This behavior enhances KNN classification performances in the LD space, with high $G_{\mathrm{NN}}(K)$ scores, and improves the reproduction of large neighborhoods, as suggested by the $R_{\mathrm{NX}}(K)$ curves. Finally, since the marker size in the embeddings is proportional to t_i, small markers refer to outliers with many HD neighbors from another class. They hence lie near the LD class borders computed by ca$t$-SNE, for instance in the MNIST data set.

7 Conclusion

This paper shows that regular, unsupervised t-SNE can be turned into a class-aware embedding method, coined as ca$t$-SNE. Class labels are accounted in the fitting of the bandwidths of the Gaussian neighborhoods around each datum, to reach a specified proportion of neighbors with the same class as the considered datum. Smaller bandwidths occur near class boundaries, which are thus magnified in the embeddings. Experiments show that ca$t$-SNE outperforms t-SNE in KNN classification tasks based on the embeddings. Visually, ca$t$-SNE tightens the bulk of the classes and loosens their boundaries. Future perspectives aim at extending these developments with learning anisotropic Mahalanobis distances [11], and comparing ca$t$-SNE with state-of-the-art supervised DR methods.

References

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4(Jun):119–155, 2003.

[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] X. Geng, Z. De-Chuan, and Z.-H. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst., Man, Cybern. B*, 35(6):1098–1107, 2005.

[10] C.-G. Li and J. Guo. Supervised isomap with explicit mapping. In *ICICIC'06*, volume 3, pages 345–348. IEEE, 2006.

[11] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, pages 513–520, 2005.

[12] J. Peltonen and S. Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16(1):68–83, 2005.

[13] L. Song, A. Smola, K. Borgwardt, and A. Gretton. Colored maximum variance unfolding. In *NIPS*, pages 1385–1392, 2008.

[14] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. Griffiths, and J. Tenenbaum. Parametric embedding for class visualization. In *NIPS*, pages 617–624, 2005.

[15] R. Memisevic and G. Hinton. Multiple relational embedding. In *NIPS*, pages 913–920, 2005.

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

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

[18] 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.

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

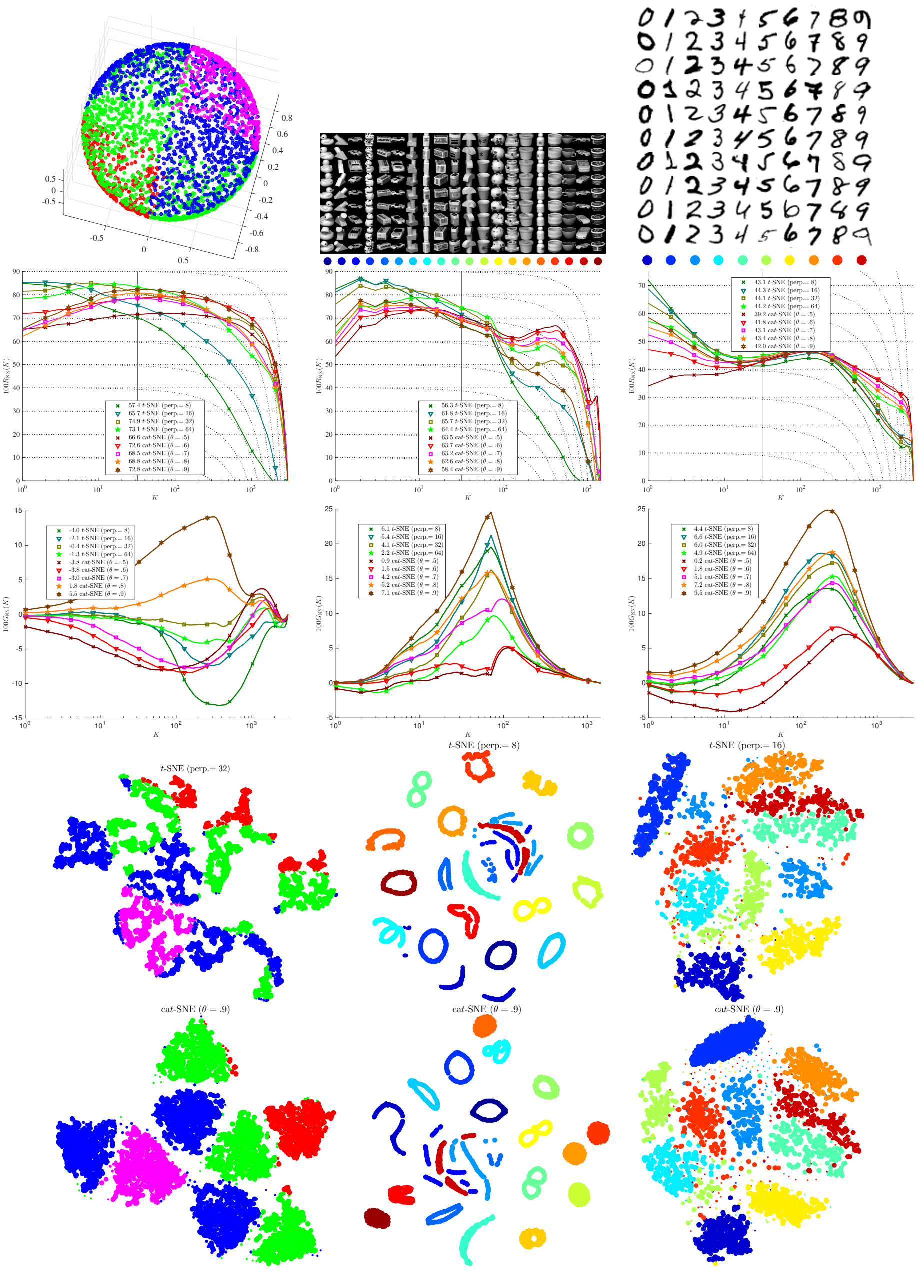

Fig. 1: Results for Sphere, COIL-20, and MNIST (left to right). From top to bottom: data illustrations, $R_{\mathrm{NX}}(K)$ curves and AUC (in legend), $\mathrm{G}_{\mathrm{NN}}(K)$ and AUC, embeddings with t-SNE and cat-SNE showing the highest AUC$[\mathrm{G}_{\mathrm{NN}}(K)]$. Marker sizes in embeddings reflect t_i in (1). For Abalone, AUC$[\mathrm{G}_{\mathrm{NN}}(K)] < 0$ for t-SNE (all perplexities), while it ranges from 1.2 to 4.3 with cat-SNE.