ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction

John A. Lee[1,2], Diego H. Peluffo-Ordóñez[2], and Michel Verleysen[2] *

1 - Molecular Imaging, Radiotherapy, and Oncology – SSS/IREC/MIRO
Université catholique de Louvain, Belgium

2 - Machine Learning Group – SST/ICTM/ELEN
Université catholique de Louvain, Belgium

Abstract. Stochastic neighbor embedding (SNE) is a method of dimensionality reduction that involves softmax similarities measured between all pairs of data points. To build a suitable embedding, SNE tries to reproduce in a low-dimensional space the similarities that are observed in the high-dimensional data space. Previous work has investigated the immunity of such similarities to norm concentration, as well as enhanced cost functions. This paper proposes an additional refinement, in the form of multiscale similarities, namely averages of softmax ratios with decreasing bandwidths. The objective is to maximize the embedding quality at all scales, with a better preservation of both local and global neighborhoods, and also to exempt the user from having to fix a scale arbitrarily. Experiments on several data sets show that this multiscale version of SNE, combined with an appropriate cost function (sum of Jensen-Shannon divergences), outperforms all previous variants of SNE.

1 Introduction

Dimensionality reduction (DR) aims at producing faithful and meaningful representations of high-dimensional data into a lower-dimensional space. The general intuition that drives DR is that close or similar data items should be represented near each other, whereas dissimilar ones should be represented far from each other. Through the history of DR, authors have formalized this idea of neighborhood preservation in various ways, using several models for the mapping or embedding of data from the high-dimensional space (HD) to the low-dimensional one (LD). For instance, principal component analysis (PCA) and classical metric multidimensional scaling (MDS) [1] rely on linear projections that maximize variance preservation and dot product preservation, respectively. Nonlinear variants of metric MDS [1, 2, 3] are based on (weighted) distance preservation. The use of similarities in DR is quite recent and appeared with methods expressed as eigenproblems, like Laplacian eigenmaps [4] and locally linear embedding [5]. These involve sparse matrices of similarities, also called affinity matrices, defined in the HD space. Genuine similarity preservation, with similarities in both HD and LD spaces, appeared later with stochastic neighbor embedding [6] (SNE). Interest in this new paradigm grew after the publication of variants such as t-SNE

*J.A.L. is a Research Associate with the Belgian fund of scientific research F.R.S.-FNRS. D.H.P.-O. is funded by a grant of the F.R.S.-FNRS (project T.0175.13 'DRedVis').

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

[7] and NeRV [8]. These methods significantly outperform older ones in terms of DR quality. The role played by SNE's specific similarities has already been investigated [9], as well as the contribution of improved cost functions [8, 10].

This paper builds upon these previous improvements and proposes a refined similarity definition, that allows for a multiscale approach of DR. The rationale is that SNE's similarities define soft Gaussian neighborhoods and therefore involve a bandwidth that is indirectly chosen by the user, through a perplexity parameter, while DR should actually deliver optimal results at all scales. To address this shortcoming of SNE-like methods, this paper introduces generalized similarities that are averages of softmax ratios with decreasing bandwidths, covering all neighborhood sizes, from large to small. At the expense of a moderate computational complexity increase, these new multiscale similarities are parameter-free and also perform better in experiments.

The rest of this paper is organized as follows. Section 2 defines the proposed multiscale similarities. Section 3 deals with cost functions and their optimization. After describing how DR quality is assessed, Section 4 presents and discusses the experimental results. Finally, Section 5 draws the conclusions.

2 Multiscale softmax similarities

Let $\mathbf{\Xi} = [\boldsymbol{\xi}_i]_{1\leq i\leq N}$ denote a set of N points in some M-dimensional space. Similarly, let $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$ be its representation in a P-dimensional space, with $P \leq M$. The squared Euclidean distances between the ith and jth points are given by $\delta_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2^2$ and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ in the HD and LD spaces respectively. Starting from pairwise distances, similarities with L scales can be defined for $i \neq j$ in the M- and P-dimensional spaces as $\sigma_{ij} = \frac{1}{L}\sum_{l=1}^{L}\sigma_{ijl}$ and $s_{ij} = \frac{1}{L}\sum_{l=1}^{L} s_{ijl}$, where the single-scale similarities are given by

$$\sigma_{ijl} = \frac{\exp(-\pi_{il}\delta_{ij}/2)}{\sum_{k,k\neq i}\exp(-\pi_{il}\delta_{ik}/2)} \quad \text{and} \quad s_{ijl} = \frac{\exp(-p_{il}d_{ij}/2)}{\sum_{k,k\neq i}\exp(-p_{il}d_{ik}/2)} \ . \qquad (1)$$

Symbols π_{il} and p_{il} denote the precisions (inverse of the squared bandwidths) of the ith datum on the lth scale. If $i = j$, then $\sigma_{ijl} = s_{ijl} = 0$ by convention. In case of a single scale ($L = 1$), the similarities reduces to those used in SNE. An important feature of similarities defined as softmax ratios such as above is their normalization, which grants them a property of shift invariance with respect to δ_{ij} and d_{ij} [9]. This property is essential to overcome distance concentration.

Each softmax ratio in σ_{ij} and s_{ij} can be interpreted as a probabilistic or stochastic membership to Gaussian neighborhoods with radii $\pi_{il}^{-1/2}$ and $p_{il}^{-1/2}$. These radii can reformulated into sizes of soft K-ary neighborhoods by computing entropies and perplexities, namely, $H_{il} = -\sum_{j=1}^{N}\sigma_{ijl}\log\sigma_{ijl}$ and $K_{il} = \exp H_{il}$. In previous SNE-like methods, the user chooses a unique perplexity value K_0 that is then used to adjust the precision π_{il} of each neighborhood. In practice, these methods solve $\log K_0 = H_{il}$ for $1 \leq i \leq N$ (and $l = 1$), in order to have soft neighborhoods with size K_0 around each datum. In the proposed mul-

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

tiscale approach, several perplexities are fixed beforehand, without any user input. They are given by $K_l = 2^{L_{\max}-l+1}$, with $1 \leq l \leq L \leq L_{\max} = \lfloor \log_2(N/4) \rceil$. Upper bound $L_{\max}$ prevents σ_{ijL} from getting nearly constant.

In the LD space, the coordinates in $\mathbf{X}$ and therefore distances d_{ij} are not determined yet, which forbids identifying precisions p_{ij} in the same way as π_{ij} in the HD space. Regular SNE easily circumvent this issue by using unit precisions ($p_{il} = 1$ for all i, $l = 1$). In the multiscale approach, we set $p_{il} = K_l^{-2/P}$, knowing that in uniform distribution in a P-dimensional space, the number of neighbors grows like R^P, where R is the neighborhood radius.

3 Divergences to measure similarity mismatch

Due to normalization, softmax similarities add up to one, that is, $\sum_j \sigma_{ij} = \sum_j s_{ij} = 1$. Therefore, $\boldsymbol{\sigma}_i = [\sigma_{ij}]_{1 \leq j \leq N}$ and $\mathbf{s}_i = [s_{ij}]_{1 \leq j \leq N}$ can be seen as discrete probability distributions and divergences can be used to assess their mismatch. In SNE, the Kullback-Leibler divergence is used. It is defined as $D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = \sum_j \sigma_{ij} \log(\sigma_{ij}/s_{ij})$. The cost function of SNE [6] can then be written as $E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}) = \sum_i D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$. The variant of SNE called NeRV [8] blends two dual KL divergences. Such a mixture is written as $D^{\beta}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = (1-\beta) D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) + \beta D_{\mathrm{KL}}(\mathbf{s}_i \| \boldsymbol{\sigma}_i)$, where parameter β balances both terms. The cost function is then $E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}, \beta) = \sum_i D^{\beta}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$. Another way to combine KL divergences is given by

$$D^{\beta}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = (1-\beta) D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{z}_i) + \beta D_{\mathrm{KL}}(\mathbf{s}_i \| \mathbf{z}_i) \ , \tag{2}$$

where $\mathbf{z}_i = (1-\beta)\boldsymbol{\sigma}_i + \beta \mathbf{s}_i$. For $\beta = 1/2$, $D^{1/2}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is known as the type 2 symmetric KL divergence, or symmetric Jensen-Shannon divergence. This mixture of divergences has been shown to be an effective cost function in a DR method called Jensen-Shannon embedding (JSE, or 'Jessie') [10].

Like JSE, the proposed multiscale approach minimizes (2), with $\beta = 1/2$. The gradient of (2) being complicated, it is not shown here, due to space limitations. In practice, the limited-memory version of the Broyden-Fletcher-Goldfarb-Shanno technique (L-BFGS) can minimize (2) very efficiently. Each step performs an iterative line search using Wolfe conditions. In order to avoid poor initial guesses of the step size, the search direction is the gradient scaled with a good approximation of the diagonal of the Hessian matrix.

A multiscale approach slightly increases the computational complexity, compared to regular SNE. Recalling that $L_{\max} = \lfloor \log_2(N/4) \rfloor$, each cost function evaluation requires a number of operations proportional to $N^2 \log N$, instead of N^2. Another issue is the risk of getting stuck in poor local minima, due to components σ_{ijl} with low perplexity values (small neighborhoods). After initialization of $\mathbf{X}$ with the P first principal component of $\boldsymbol{\Xi}$, we address this issue by carrying out a few L-BFGS iterations with $L = 1$, then a few ones with $L = 2$, and so on until we reach $L = L_{\max}$ and take into account all scales on an equal footing. This allows us to embed data with a first focus on global structure, while more and more local details are introduced progressively.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

4 Quality assessment, experiments, and results

Several performance indicators can assess the quality of embeddings produced by DR methods. However, the consensus that emerges from several publications is to use the average agreement rate between K-ary neighborhoods in the high- and low-dimensional spaces [11, 8, 12]. If ν_i^K and n_i^K denote the K-ary neighborhoods of vectors $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, respectively, then the average agreement rate can be written as $Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} |\nu_i^K \cap n_i^K|/(KN)$. It varies between 0 (empty intersection) and 1 (perfect agreement). Knowing that random coordinates in $\mathbf{X}$ lead on average to $Q_{\mathrm{NX}}(K) \approx K/(K-1)$ [12], the useful range of $Q_{\mathrm{NX}}(K)$ is $N-1-K$, which depends on K. Therefore, in order to fairly compare or combine values of $Q_{\mathrm{NX}}(K)$ for different neighborhood sizes, the criterion can be rescaled to get $R_{\mathrm{NX}}(K) = \frac{(N-1)Q_{\mathrm{NX}}(K)-K}{N-1-K}$, for $1 \leq K \leq N-2$. This modified criterion indicates the improvement over a random embedding and has the same useful range between 0 and 1 for all K. In experiments, the whole curve $R_{\mathrm{NX}}(K)$ is shown, with a logarithmic scale for K. This choice is again justified by the fact that the size K and radius R of small neighborhoods in a P-dimensional space are (locally) related by $K \propto R^P$. A logarithmic axis also reflects that errors in large neighborhoods are proportionally less important than in small ones. Eventually, a scalar score is obtained by computing the area under the $R_{\mathrm{NX}}(K)$ curve in the log plot, given by $\mathrm{AUC}_{\log K}(R_{\mathrm{NX}}(K)) = (\sum_{K=1}^{N-2} R_{\mathrm{NX}}(K)/K)/(\sum_{K=1}^{N-2} 1/K)$. The AUC assesses DR quality at all scales, with the most appropriate weights.

The experiments involve three data sets, all to be re-embedded in two dimensions ($P = 2$). The first one is a toroidal string, looking like a circular coil spring (30 coils, $M = 3$, $N = 3000$). The second one is the COIL20 image bank ($M = 128^2$, $N = 1440$; 20 objects, 72 poses/angles). The third data set is a random subsample of the MNIST database of scanned handwritten digits ($M = 24^2$, $N = 3000$; approx. 300 images per digit). All gray-level images are vectorized and no PCA preprocessing is achieved.

The proposed multiscale JSE is compared to classical metric MDS (CMDS), non-metric MDS [1] (NMDS), Sammon's nonlinear mapping [2] (NLM), curvilinear component analysis [3] (CCA), SNE [6], t-SNE [7], NeRV [8], and single-scale JSE [10]. The target perplexity for SNE, t-SNE, NeRV, and single-scale JSE is 32. For JSE with multiscale similarities (mss), there is no target perplexity. Figures 1 to 3 report the quality curves and some embeddings.

Classical methods like CMDS, NMDS, and NLM favor the rendering of the global arrangement of the data sets (peak in the right part of the quality curves). CCA, SNE, NeRV, and single-scale JSE succeed in better preserving mid-sized to small neighborhoods, with a bump near the chosen perplexity value (vertical line in diagrams). Thanks to their specific similarities [9], only SNE and its variants work well with very HD data (COIL20 and MNIST). Among them, t-SNE delivers the best results for small neighborhoods, at expense of overlooking the global structure. Multiscale JSE provides the best tradeoff at all scales, from local to global, with systematically the highest $\mathrm{AUC}_{\log K}(R_{\mathrm{NX}}(K))$.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

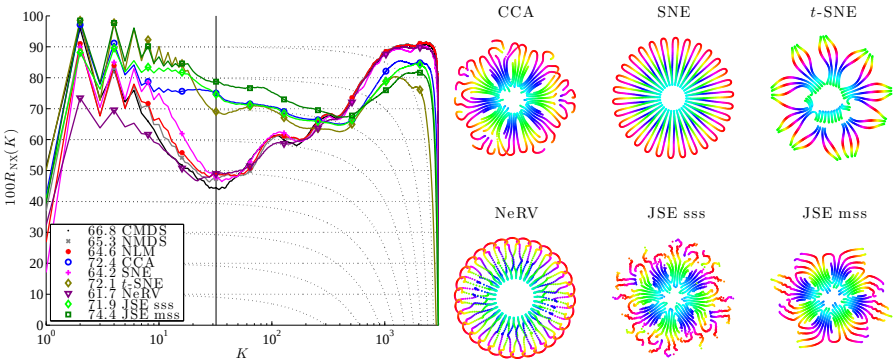

Fig. 1: The quality curves, AUCs, and some embeddings for the toroidal string. Each curve reports $R_{\mathrm{NX}}(K)$, the relative improvement over a random embedding. The higher the curve, the better; the AUCs stand in the legend. The vertical line indicates the target perplexity for SNE and all its single-scale variants. Dotted isolines are shown for both $Q_{\mathrm{NX}}(K)$ (curved) and $R_{\mathrm{NX}}(K)$ (straight).

5 Conclusion

Similarity preservation has been a game changer in nonlinear DR. Methods like SNE, t-SNE, NeRV, and JSE are almost immune to norm concentration and provide excellent embeddings. However, they all rely on single-scale similarities, that is, soft Gaussian neighborhoods whose single bandwidth is adjusted by the user, through a perplexity parameter. If DR quality is assessed with K-ary neighborhood agreement in the HD and LD space, then this design is expected to favor some scale at the expense of the others. This paper tackles this issue with multiscale similarities, having several bandwidths that span all neighborhood sizes. Combined with a well chosen cost function, like that of JSE, these new similarities outperform all single-scale variants of SNE. Moreover, they are parameter-free and therefore do not require the user to choose a perplexity.

References

[1] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, New York, 1997.

[2] J.W. Sammon. A nonlinear mapping algorithm for data structure analysis. *IEEE Transactions on Computers*, CC-18(5):401–409, 1969.

[3] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, January 1997.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *NIPS 2001*, volume 14. MIT Press, 2002.

[5] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

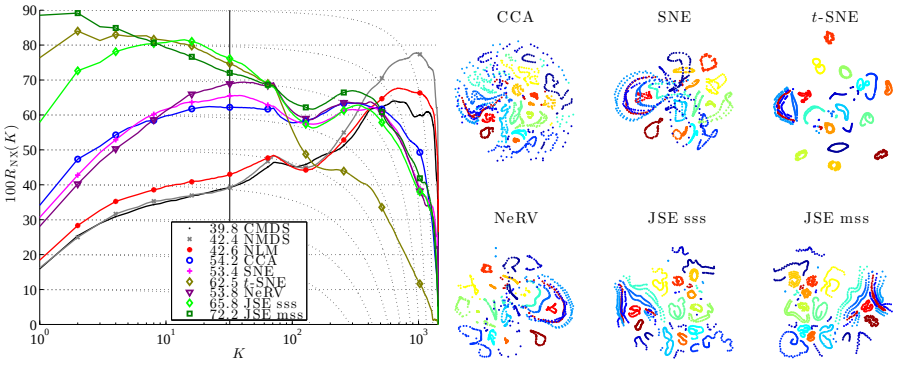

Fig. 2: The quality curves, AUCs, and some embeddings for the COIL20 subset.

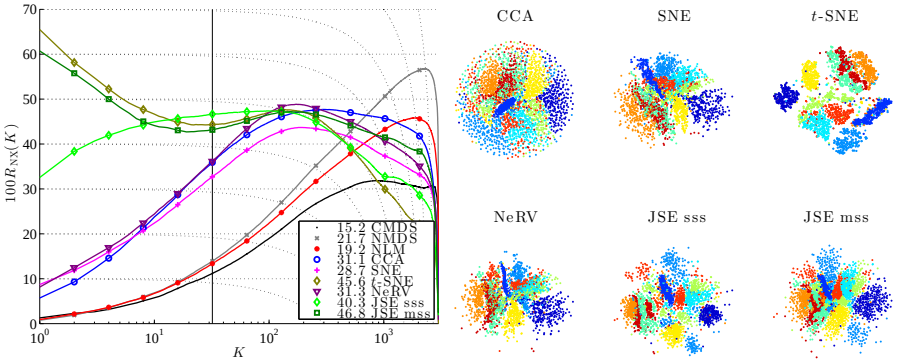

Fig. 3: The quality curves, AUCs, and some embeddings for the MNIST subset.

[6] G. Hinton and S.T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS 2002*, volume 15, pages 833–840. MIT Press, 2003.

[7] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[8] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR*, 11:451–490, 2010.

[9] J.A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. In *Proc. ICCS 2011*, pages 538–547, Singapore, 2011.

[10] 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.

[11] S.L. France and J.D. Carroll. Development of an agreement metric based upon the RAND index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. In *Proc. MLDM 2007*, pages 499–517. Springer-Verlag, 2007.

[12] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *JASA*, 104(485):209–219, 2009.