POWER-LAW STOCHASTIC NEIGHBOR EMBEDDING

Huan-Hsin Tseng[⋆] *Issam El Naqa*[⋆] *Jen-Tzung Chien*[†]

[⋆]Department of Radiation Oncology, University of Michigan Health System, Ann Arbor, MI
[†]Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

Stochastic neighbor embedding (SNE) aims to transform the observations in high-dimensional space into a low-dimensional space which preserves neighbor identities by minimizing the Kullback-Leibler divergence of the pairwise distributions between two spaces where Gaussian distributions are assumed. Data visualization could be improved by adopting the t-SNE where Student t distribution is used in the low-dimensional space. However, data pairs in the latent space are forced to be squeezed due to the loss of dimensions. This study incorporates the *power-law* distribution into construction of the p-SNE. Such an unsupervised p-SNE increases the physical forces in neighbor embedding so that the neighbors in the low-dimensional space can be adjusted flexibly to reflect the neighboring in the high-dimensional space. The experiments on three learning tasks illustrate that the manifold or data structure using the proposed p-SNE is preserved in better shape than that using SNE and t-SNE.

Index Terms— Manifold learning, dimensionality reduction, power law, stochastic neighbor embedding, visualization

1. INTRODUCTION

In real-world applications, we face signal processing problems on high-dimensional data such as audio, music, text, images, videos and social networks. The "curse of dimensionality" becomes serious when the information systems are operated in high-dimensional space. To deal with this issue, a popular approach is to transform the high-dimensional data into the low-dimensional one. We aim to learn a low-dimensional representation which is efficient to extract useful information for classification and prediction [1]. Basically, algorithms for learning representation range from the *linear* transformations, such as the principal component analysis and the linear discriminant analysis, to *nonlinear* mappings, such as the locally linear embedding [2] and the stochastic neighbor embedding (SNE) [3, 4, 5, 6] which are regarded as nonparametric mappings. A parametric mapping based on deep neural network [7, 8] was learned to handle the unseen data in manifold learning. Nevertheless, SNE has been extensively developed for probabilistic dimensionality reduction and data visualization.

In general, SNE [3] is carried out by optimizing the Kullback-Leibler (KL) divergence for distributions of different neighbors in high-dimensional space and low-dimensional space. The distributions are characterized by Gaussian using the distance between two samples. However, SNE suffers from the crowding problem [4] so that the pairwise distances in the low-dimensional space cannot faithfully reflect those in high-dimensional observation space. To avoid this problem, the prior distribution in the low-dimensional map was modified as the heavy-tailed distributions [5], e.g. Student t distribution [4]. Also, in [9], an exit distribution [10] was employed in SNE subject to a spherical constraint. This paper investigates the effect of attractive force and repulsive force in different variants of SNE and proposes a new power-law SNE (p-SNE) where the low-dimensional neighbor representation is characterized by the power-law distribution. Using this p-SNE, the attractive and repulsive forces are increased for nearby samples. The activation of samples in the high-dimensional map can be effectively expressed in the low-dimensional map. The solution to crowding problem is further strengthened.

2. STOCHASTIC NEIGHBOR EMBEDDING

SNE is known as a nonlinear manifold learning. Suppose we are given a set of N high-dimensional data $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^D\}$. SNE attempts to find the low-dimensional representation $\mathcal{Y} = \{\mathbf{y}_i \in \mathbb{R}^d\}$ such that $\mathbf{y}_i$ preserves the pairwise similarity of $\mathbf{x}_i$ with $d < D$. The pairwise similarity is measured by the conditional probability $p_{j|i}$ that $\mathbf{x}_j$ is a neighbor of $\mathbf{x}_i$. This probability is modeled by a Gaussian distribution with a variance σ^2 expressed by

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2\right)}. \tag{1}$$

Correspondingly, we need to measure the conditional probability $q_{j|i}$ for a pair of neighbors $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional space. The aim of neighbor embedding is to match these two sets of distributions $P_i = \{p_{j|i}\}$ and $Q_i = \{q_{j|i}\}$ for individual sample i as well as possible. To do so, we minimize a cost function $\mathcal{L}$ which is the sum of KL divergences between P_i and Q_i from all samples

$$\mathcal{L} \triangleq \sum_i \mathcal{D}_{\mathrm{KL}}(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\left(\frac{p_{j|i}}{q_{j|i}}\right). \tag{2}$$

SNE has different variants which depend on the definition of conditional distribution $q_{j|i}$ in the low-dimensional space.

2.1. Gaussian distribution

The original SNE [3] was constructed by assuming the conditional distribution $q_{j|i}$, that picks latent sample $\mathbf{y}_j$ as the neighbor of sample $\mathbf{y}_i$, to be Gaussian with a shared variance $\sigma_i^2 = 1/2$ yielded by

$$q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)}. \tag{3}$$

In a symmetric SNE [11], the pairwise similarities encoded in P_i and Q_i are measured by using the *joint* probabilities

$$\begin{aligned} p_{ij} &= \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/2\sigma^2\right)} \\ q_{ij} &= \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k\neq l}\exp\left(-\|\mathbf{y}_k - \mathbf{y}_l\|^2\right)} \end{aligned} \tag{4}$$

where $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$. The objective of symmetric SNE is then obtained by substituting p_{ij} and q_{ij} in Eq. (4) into KL divergence in Eq. (2). This objective is minimized to find the optimal latent sample $\mathbf{y}_i$ in low-dimensional space according to the gradient

$$\frac{\partial \mathcal{L}_{\text{sne}}}{\partial \mathbf{y}_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(\mathbf{y}_i - \mathbf{y}_j\right). \tag{5}$$

Using SNE, the *repulsive force* is equipped to blow away those data points of less similarity and such effect soothes the so-called crowding problem, caused by dimensionality reduction such that the accommodation space of data points is lost during the transformation.

2.2. Student t distribution

To alleviate more of this problem, the heavy-tailed distribution based on Student t distribution was employed to characterize the joint probability of a data pair in the low-dimensional space [4]

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}}. \tag{6}$$

The resulting t-SNE is implemented by minimizing KL divergence between p_{ij} in Eq. (4) and q_{ij} in Eq. (6) using all data pairs. The gradient with respect to $\mathbf{y}_i$ is calculated by

$$\frac{\partial \mathcal{L}_{\text{t-sne}}}{\partial \mathbf{y}_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(\mathbf{y}_i - \mathbf{y}_j\right)\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}. \tag{7}$$

This t-SNE improves the crowding problem by increasing repulsion to the lower-dimensional *particles* $\mathcal{Y} = \{\mathbf{y}_i\}$.

2.3. Exit distribution

In [9], a spherical SNE (e-sSNE) was proposed for visualizing the hyperspectral data by minimizing the KL divergence between $\{p_{ij}\}$ and $\{q_{ij}\}$ where a spherical constraint was imposed and the *exit* distribution [10] is adopted as the neighbor probability in the low-dimensional space

$$q_{ij} = \frac{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^{-d}}{\sum_{k \neq l} \|\mathbf{y}_k - \rho \mathbf{y}_l\|^{-d}} \tag{8}$$

where $\rho \in [0, 1)$ and $q_{ij} \neq q_{ji}$. Exit distribution is known as a distribution of the exit place for the iterated Brownian motion in a hypersphere [10]. The embedding on a $(d-1)$- dimensional sphere $\mathbb{S}^{d-1}$ in e-sSNE is governed by solving the constrained optimization problem

$$\mathcal{L}_{\text{e-ssne}} = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + \sum_i \lambda_i \left(1 - \|\mathbf{y}_i\|^2\right) \tag{9}$$

where the Lagrange multiplier λ_i is introduced to restrict $\mathbf{y}_i$ on $\mathbb{S}^{d-1}$. This optimization is solved by using the gradient

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{e-ssne}}}{\partial \mathbf{y}_i} &= d \sum_j p_{ij} \left(\frac{\mathbf{y}_i - \rho \mathbf{y}_j}{\|\mathbf{y}_i - \rho \mathbf{y}_j\|^2} + \rho \frac{\rho \mathbf{y}_i - \mathbf{y}_j}{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^2}\right) \\ &+ d \left(\rho \sum_j q_{ij} \frac{\mathbf{y}_j - \rho \mathbf{y}_i}{\|\mathbf{y}_i - \rho \mathbf{y}_j\|^2} - \sum_j q_{ij} \frac{\mathbf{y}_i - \rho \mathbf{y}_j}{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^2}\right) \\ &- 2\lambda_i \mathbf{y}_i. \end{aligned} \tag{10}$$

The effects of neighbor embedding comprised of the spherical constraint and the exit distribution are coupled in Eq. (10). The individual effect is hard to analyze.

3. POWER-LAW MANIFOLD LEARNING

This paper presents a general framework of SNE with different realizations of joint distribution q_{ij} and divergence measure $\mathcal{D}$ or loss function $\mathcal{L}$. A variant of heavy-tailed SNE using power-law distribution is developed.

3.1. General objective function

Let $\mathcal{M}^D$ and $\mathcal{N}^d$ denote two manifolds in high-dimensional and low-dimensional spaces, respectively. A general manifold learning is to find a mapping $\varphi : \mathcal{M} \to \mathcal{N}$ such that an objective $\mathcal{L}\left(P(\mathcal{X}), Q(\mathcal{Y})\right)$ is optimized to obtain $\mathcal{Y} = \varphi(\mathcal{X})$. The joint distribution can be expressed in a basic form of

$$q_{ij} = \frac{Q(r_{ij})}{\sum_{k \neq l} Q(r_{kl})} \tag{11}$$

where $r_{ij} \triangleq \|\mathbf{y}_i - \mathbf{y}_j\|$ is a distance measure and Q is an arbitrary but decreasing function ($\dot{Q} < 0$). This general SNE allows different choices of symmetric probability distributions P and Q equipped in $\mathcal{M}$ and $\mathcal{N}$. Notably, we adopt an arbitrary f divergence [12] to build up a general objective function

$$\mathcal{L} \triangleq \sum_i \mathcal{D}_f(P_i \| Q_i) = \sum_i \sum_j p_{ij} f\left(\frac{q_{ij}}{p_{ij}}\right) \tag{12}$$

where $f(\cdot)$ denotes a convex function with $f(1) = 0$. Under this objective, the gradient for optimization is computed as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} &= \frac{2}{Z}\left[\sum_j \dot{Q}(r_{ij}) f'\left(\frac{q_{ij}}{p_{ij}}\right)\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right)\right] \\ &- \frac{2}{Z^2}\left(\sum_{k \neq l} f'\left(\frac{q_{kl}}{p_{kl}}\right) Q(r_{kl})\right)\left[\sum_j \dot{Q}(r_{ij})\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right)\right]\end{aligned} \tag{13}$$

where $\dot{Q}(r) = \frac{d}{dr} Q(r)$, $f'(t) = \frac{d}{dt} f(t)$ and $Z = \sum_{k \neq l} Q(r_{kl})$ is the normalization term. KL divergence is then a special case of f divergence when $f(t) = -\log t$. Also, χ^2-divergence is the case when $f(t) = (t-1)^2$. Basically, Eq. (13) tells us that the origin of *attraction* in the first term comes from the decay rate of target similarity $\dot{Q}$ and the differential of f-divergence f' while the *repulsion* in the second term is related to $\dot{Q}$. Considering the case of KL divergence, the gradient in Eq. (13) is realized as

$$2 \sum_j p_{ij} \frac{\dot{Q}(r_{ij})}{Q(r_{ij})}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) - \frac{2}{Z}\left[\sum_j \dot{Q}(r_{ij})\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right)\right]. \tag{14}$$

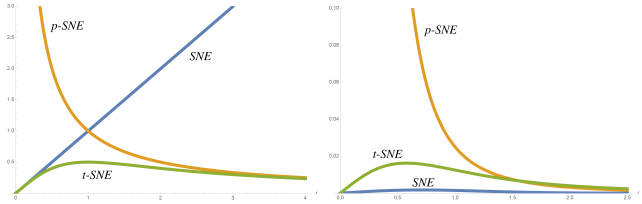
3.2. Attractive and repulsive forces

The gradients in Eqs. (5) and (7) are viewed as the sum of forces for SNE and t-SNE, respectively [3, 11]. In general, how the neighbor embedding performs and how the nonlinear embedding behaves can be interpreted by the physical properties from mechanics. Typically, a mechanical system of N particles with mass m and positions $\mathbf{y}_i$ subject to the potential energy $V(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ has the Lagrangian [13]

$$L(\mathbf{y}_1, \ldots, \mathbf{y}_N, \dot{\mathbf{y}}_1, \ldots, \dot{\mathbf{y}}_N) = T - V \tag{15}$$

	SNE	t-SNE	p-SNE
$Q(r)$	e^{-r^2}	$(1+r^2)^{-1}$	$r^{-\alpha}$
attraction	$4pr$	$\frac{4pr}{1+r^2}$	$\frac{2\alpha p}{r}$
repulsion	$\frac{1}{Z}4re^{-r^2}$	$\frac{1}{Z}\frac{4r}{(1+r^2)^2}$	$\frac{1}{Z}\frac{2\alpha}{r^{\alpha+1}}$

Table 1. $Q(r)$, attractive and repulsive forces for three SNEs.



(a) (b)

Fig. 1. (a) Attractive and (b) repulsive forces in three SNEs.

where $T = \frac{m}{2}\sum_{i=1}^{N}\|\dot{\mathbf{y}}_i\|^2$ is the sum of all kinetic energies. This equation of motion for N particles follows the Euler-Lagrange equation $\frac{\partial L}{\partial \mathbf{y}_i} = \frac{d}{dt}\frac{\partial L}{\partial \dot{\mathbf{y}}_i}$ so that we derive

$$F_i = m\ddot{\mathbf{y}}_i = -\frac{\partial V}{\partial \mathbf{y}_i} \tag{16}$$

which is regarded as the force exerted on particle $\mathbf{y}_i$. Because the potential energy is physically comparable with the learning objective $V = \mathcal{L}$, the gradients in Eq. (5) for SNE and Eq. (7) for t-SNE are seen as the negative forces which drive how particle or latent variable $\mathbf{y}_i$ is moving. The vector $\mathbf{y}_j - \mathbf{y}_i$ implies the *attractive force* on $\mathbf{y}_i$ while $\mathbf{y}_i - \mathbf{y}_j$ indicates the *repulsive force* on $\mathbf{y}_i$. Therefore, the forces of SNE and t-SNE on a particle $\mathbf{y}_i$ are expressed by

$$\begin{aligned} F_i^{\text{sne}} &= 4\sum_j (p_{ij}r_{ij} - q_{ij}r_{ij})\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) \\ F_i^{\text{t-sne}} &= 4\sum_j \frac{(p_{ij} - q_{ij})r_{ij}}{1 + r_{ij}^2}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) \end{aligned} \tag{17}$$

where $(\mathbf{y}_j - \mathbf{y}_i)/r_{ij}$ denotes the unit vector of force direction. The magnitude of repulsive force from Eq. (14) is found by $2|\sum_j \dot{Q}(r_{ij})/Z|$ which is proportional to the decay rate of target similarity. The attractive and repulsive forces in SNE and t-SNE are accordingly obtained and shown in Table 1. The corresponding function $Q(r)$ in Eq. (11) is also given. Importantly, the choice of $Q(r)$ is influential since the resulting forces are changed dramatically. But, the spherical constraint in e-sSNE only constrains the force with a lower bound.

3.3. Power-law distribution

Considering the mechanics of attractive and repulsive forces, we propose a new SNE, named as the p-SNE, by using the Pareto distribution [14] or the power-law distribution where $Q(r) = 1/r^{\alpha}$ with $\alpha > 0$. Power-law distribution is behaved as a heavy-tailed distribution. This p-SNE is implemented by using the gradient $\frac{\partial \mathcal{L}_{\text{p-sne}}}{\partial \mathbf{y}_i}$ given by

$$-2\alpha\sum_j \frac{p_{ij}}{r_{ij}}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) + \frac{2\alpha}{\sum_{k\neq l} r_{kl}^{-\alpha}}\left(\sum_j \frac{1}{r_{ij}^{\alpha+1}}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right)\right). \tag{18}$$

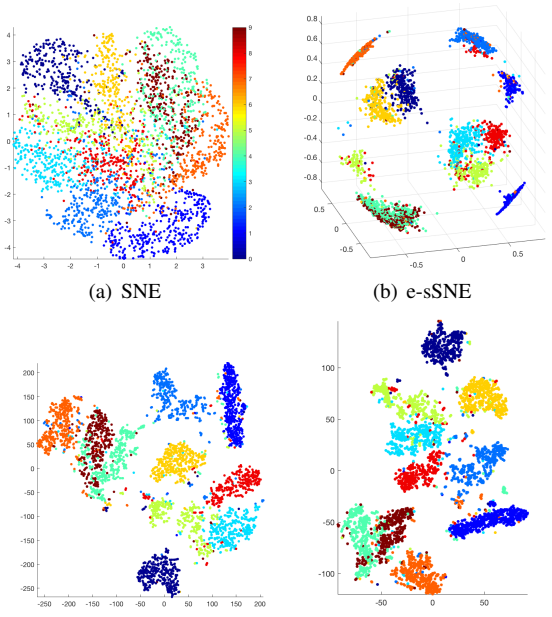
The attractive and repulsive forces are obtained as given in Table 1. Figure 1 compares the attractive and repulsive forces by using SNE (blue), t-SNE (green) and the proposed p-SNE (orange) with $\alpha = 2$. We can see that p-SNE produces strong forces even the neighbor distance r is small. The attractive and repulsive forces of p-SNE are stronger than those of t-SNE in most circumstances. This leads to widely separated particles of different affinities. The crowding problem in manifold learning is tackled due to the high-force motions for neighbor embedding in the low-dimensional space.

4. EXPERIMENTS

4.1. Experimental setup

We conducted three sets of experiments on manifold learning and investigated the visualization by using different gray-scale image data including (1) the MNIST handwritten digits [15], (2) the COIL-20 objects [16], and (3) the Olivetti faces [4]. First, the MNIST dataset consists of 60,000 training images with 10 handwritten digits. Each digit is centered in a size of 28×28. Second, the COIL-20 dataset contains the images of 20 different objects. Each object has 72 angles sampled from a 360° view. Totally, there are 1,440 images. Each image has a resolution 128×128. In addition, the Olivetti faces dataset comprises the images from 40 distinct persons. Each individual has 10 facial variations from viewpoints or expressions. All 400 Olivetti faces are in a size 64×64.

In the implementation, the whitening process using PCA was applied to all datasets. The image data were downsized to 50 dimensions such that some noises were suppressed but data structure was still remained. Different SNEs were performed to reduce the dimensionality from $D = 50$ to $d = 2$. $\alpha = 2$ for p-SNE is set in all experiments. In MNIST, we randomly select 3,000 images of 10 digits for evaluation. Different classes of digits, objects and faces are shown in different colors. For comparison, we carry out the manifold learning using SNE [3], t-SNE [4], e-sSNE [9] and the proposed p-SNE and evaluated their two-dimensional data visualizations.



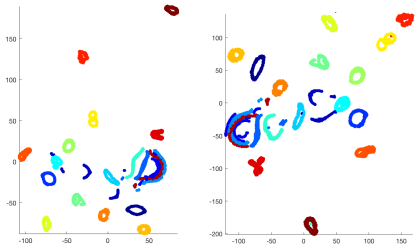
(a) SNE (b) e-sSNE

(c) t-SNE (d) p-SNE

Fig. 2. Visualization of MNIST digits using different SNEs.

4.2. Experimental results

Figure 2 compares the visualizations of using four SNEs where MNIST dataset is used. It is obvious that SNE does suffer from the crowding problem where 10 digits in different colors are confusing in two-dimensional space. Using e-sSNE, the digit images are projected to the surface of a sphere due to the spherical constraint. Crowding problem is considerably alleviated, but still several digits are confusing. Using t-SNE, those confusing digits could be separated. However, p-SNE obtained a better separation with a wider range for different digits with t-SNE. This is because that p-SNE introduces stronger forces in dimensionality reduction than other methods. The clustering performance can be also reflected by the Davies-Bouldin index (DBI) [17] as shown in Figures 5(a)(b) where the lowest DBI matches the true number of digits by using p-SNE.



(a) t-SNE (b) p-SNE

Fig. 3. Visualization of COIL-20 objects using t-SNE and p-SNE.

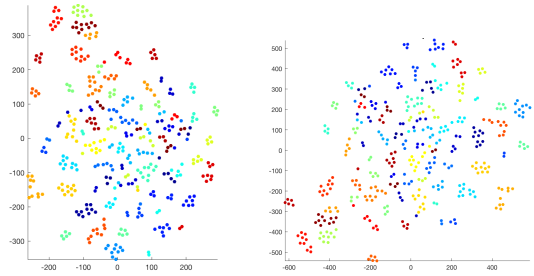
Figure 3 displays two-dimensional representations of 1,440 object images by using COIL-20 dataset. 20 distinct classes are shown by their colors. From Figures 5(c)(d), p-SNE demonstrates better grouping for samples from the same class than t-SNE in terms of DBI. We can see that p-SNE attains the clusters of samples which easily converge into a single component for individual classes while t-SNE may render the scattered pieces with disjoint components. One also observes that p-SNE can learn the well-separated objects with a large two-dimensional subspace in an unsupervised way.

We also compare the visualizations of 400 Olivetti faces from 40 distinct persons by using t-SNE and p-SNE as illustrated in Figures 4 and 5(e)(f). The corresponding faces can be found at the link below[1]. Again, compared with t-SNE, p-SNE has stronger forces to stimulate the movement of facial samples in the low-dimensional space. Using t-SNE, a small set of classes are mixing while p-SNE can produce well distributed and separated samples for 40 groups of facial images. Here, p-SNE shows a stronger grouping capability than t-SNE such that there are a smaller number of images of the same person scattered with other persons. This is reflected by DBI.

5. CONCLUSIONS

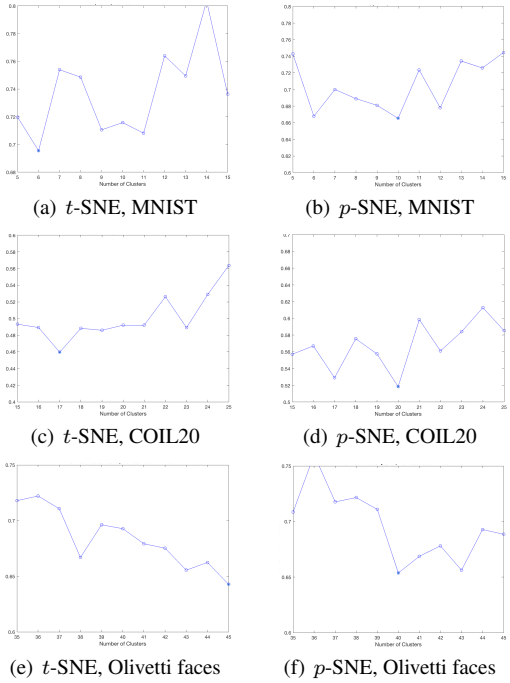
We have presented a new unsupervised learning approach to dimensionality reduction based on the stochastic neighbor embedding. To characterize the neighbor probability of a data pair in the low-dimensional subspace, this approach was evolved from SNE using Gaussian distribution to t-SNE using t distribution, e-sSNE using exit distribution, and then p-SNE using the power-law distribution. In addition, we developed a general solution to SNE where any form of decreasing function given by the distance measure of two low-dimensional samples could be applied. A general divergence measure based on f divergence was introduced to design different realizations of SNE with a variety of information measures. Importantly, such a solution was not only derived in mathematics but also interpreted in physics. We realized the SNE based on KL divergence and compared the attractive and repulsive forces of different SNEs in the low-dimensional spaces. It is found that the forces for neighbor embedding using p-SNE are stronger than those using other SNEs. This physical property provides an avenue to flexibly reflect the neighboring of low-dimensional samples as well as preserve the global structure or map for samples and clusters. The crowding problem could be resolved. From the experimental results, we know that the proposed p-SNE achieved a better performance in two-dimensional visualization with larger data range and class separation when compared with SNE, e-sSNE and t-SNE. This property is also reflected by DBI. Future works will include the extension of SNE by using other divergence measures and other decreasing functions for joint probability of data pairs. A mixture of using different measures or different functions and the effect of the corresponding forces for clustering and classification will be also investigated.

[1]https://github.com/HHTseng/Power-law-SNE



(a) t-SNE (b) p-SNE

Fig. 4. Visualization of Olivetti faces using t-SNE and p-SNE.



(a) t-SNE, MNIST (b) p-SNE, MNIST

(c) t-SNE, COIL20 (d) p-SNE, COIL20

(e) t-SNE, Olivetti faces (f) p-SNE, Olivetti faces

Fig. 5. Davies-Bouldin indices of t-SNE and p-SNE in three tasks.

6. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[3] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds., 2003, pp. 857–864.

[4] L. van der Maaten and G. E. Hinton, "Visualizing data using *t*-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[5] Z. Yang, I. King, Z. Xu, and E. Oja, "Heavy-tailed symmetric stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 2169–2177.

[6] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, 2012.

[7] M. R. Min, L. Maaten, Z. Yuan, A. J. Bonner, and Z. Zhang, "Deep supervised *t*-distributed embedding," in *Proc. of International Conference on Machine Learning*, 2010, pp. 791–798.

[8] J.-T. Chien and C.-H. Chen, "Deep discriminative manifold learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2672–2676.

[9] D. Lunga and O. Ersoy, "Spherical stochastic neighbor embedding of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 857–871, 2013.

[10] S. Kato, "A distribution for a pair of unit vectors generated by Brownian motion," *Bernoulli*, vol. 15, no. 3, pp. 898–921, 2009.

[11] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton, "Visualizing similarity data with a mixture of maps," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2007, pp. 67–74.

[12] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[13] H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, Addison Wesley, 2002.

[14] B. C. Arnold, *Pareto distribution*, Wiley Online Library, 2015.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, pp. 2278–2324, 1998.

[16] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Tech. Rep., CUCS-005-96, 1996.

[17] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.