Neurocomputing 90 (2012) 23–45

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences

Kerstin Bunte [a,b,*], Sven Haase [c], Michael Biehl [a], Thomas Villmann [c]

[a] Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, P.O. Box 407, 9700AK Groningen, The Netherlands
[b] University of Bielefeld, CITEC Center of Excellence, D-33501 Bielefeld, Germany
[c] Department of Mathematics, University of Applied Sciences Mittweida, Germany

ARTICLE INFO

Available online 20 March 2012

Keywords:
Dimension reduction
Visualization
Divergence optimization
Nonlinear embedding
Stochastic neighbor embedding

ABSTRACT

We present a systematic approach to the mathematical treatment of the t-distributed stochastic neighbor embedding (t-SNE) and the stochastic neighbor embedding (SNE) method. This allows an easy adaptation of the methods or exchange of their respective modules. In particular, the divergence which measures the difference between probability distributions in the original and the embedding space can be treated independently from other components like, e.g. the similarity of data points or the data distribution. We focus on the extension for different divergences and propose a general framework based on the consideration of Fréchet-derivatives. This way the general approach can be adapted to the user specific needs.

1. Introduction

Various dimension reduction techniques have been introduced based on the aim of preserving specific properties of the original data. The spectrum ranges from linear projections of original data, such as principal component analysis (PCA) or classical multi-dimensional scaling (MDS) [1] to a variety of locally linear and nonlinear approaches, such as isomap [2,3], locally linear embedding (LLE) [4], local linear coordination (LLC) [5], or charting [6,7].

Other methods aim at the preservation of the classification accuracy in lower dimensions and incorporate the available label information for the embedding, e.g. linear discriminant analysis (LDA) [8] and generalizations thereof [9], extensions of the self-organizing map (SOM) [10], incorporating class labels [11], and limited rank matrix learning vector quantization (LiRaM LVQ) [12,13]. For a comprehensive review on nonlinear dimensionality reduction methods, we refer to [14].

Recently, the stochastic neighbor embedding (SNE) [15] and extensions thereof have become popular for visualization. SNE approximates the probability distribution in the high-dimensional space, defined by neighboring points, with their probability distribution in a lower-dimensional space. In [16] the authors proposed a technique called t-SNE, which is a variation of SNE considering a particular statistical model assumption for data distributions. The similarity of the distributions is quantified in terms of the Kullback–Leibler divergence. In [17] it is argued that the preservation of shift-invariant similarities as employed by SNE and its variants is superior in comparison to distance preservation as performed by many traditional dimension reduction techniques.

Functional metrics like Sobolev distances, kernel-based dissimilarity measures and divergences have attracted attention recently for the processing of data showing a functional structure. These dissimilarity measures were for example investigated as alternatives to the most common choice, the Euclidean distance [18–22]. The application of divergences for Vector Quantization and Learning Vector Quantization schemes have been investigated in [23,24].

This work bases on [25], where the self-organized neighbor embedding (SONE), which can be seen as a hybrid between the self-organizing map (SOM) and SNE, has been extended to the use of arbitrary divergences. In this contribution, we formulate a mathematical framework based on Fréchet derivatives which allows to generalize the concept of SNE and t-SNE to arbitrary divergences. This leads to a new dimension reduction and visualization scheme, which can be adapted to the user specific requirements in an actual problem. We summarize the general classes of divergences following the scheme introduced by [26] and extended in [23]. The mathematical framework for functional derivatives of continuous divergences is given by the functional-analytic generalization of common derivatives, known as Fréchet derivatives [27,28]. It is the generalization of partial derivatives for the discrete variants of the divergences.

* Corresponding author at: CITEC Center of Excellence - Cognitive Interaction Technology, Bielefeld University, Universitätsstraße 21–23, D-33615 Bielefeld, Germany. Tel.: +49 521106 12130.
E-mail address: kerstin.bunte@googlemail.com (K. Bunte).
URL: http://www.cit-ec.de/de/tcs/kerstin (K. Bunte).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.034

We introduce a general mathematical framework for the extension of SNE and t-SNE for arbitrary divergences. The different classes of divergences are characterized and for various examples the Fréchet derivatives are identified. We demonstrate the proposed framework for the example case of the Gamma divergence. The behavior of different divergences stemming from the identified divergence families are shown on several examples in the image analysis domain.

2. Review of SNE and t-SNE

Generally, dimensionality reduction methods convert a high dimensional data set $\{x_i\}_{i=1}^n \in \mathbb{R}^N$ into low dimensional data $\{\xi_i\}_{i=1}^n \in \mathbb{R}^M$. A probabilistic approach to visualize the structure of complex data sets, preserving neighbor similarities is stochastic neighbor embedding (SNE), proposed by Hinton and Roweis [15]. SNE converts high-dimensional Euclidean distances between data points into probabilities that represent similarities. The conditional probabilities $p_{j|i}$ that a data point x_i would pick x_j as its neighbor is given by

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

with $p_{i|i} = 0$. The variance σ_i of the Gaussians centered around x_i is determined by a binary search procedure [16]. The density of the data is likely to vary. In dense regions a smaller value of σ is more appropriate than in sparse regions. Let P_i be the conditional probability distribution over all other data points given point x_i. This distribution has an entropy which increases as σ_i increases. SNE performs a binary search for the value of σ_i which produces a P_i with a fixed perplexity specified by the user. The perplexity is defined as

$$\mathrm{perpl}(P_i) = 2^{H(P_i)}, \tag{2}$$

where $H(P_i)$ is the Shannon entropy of P_i measured in bits: $H(P_i) = -\sum_j p_{j|i}\log_2 p_{j|i}$. It can be interpreted as a smooth measure of the effective number of neighbors and typical values ranges between 5 and 50 dependent on the data set size.

The low-dimensional counterparts ξ_i and ξ_j of the high-dimensional data points x_i and x_j are modeled by similar probabilities

$$q_{j|i} = \frac{\exp(-\|\xi_i - \xi_j\|^2)}{\sum_{j\neq i}\exp(-\|\xi_i - \xi_j\|^2)}, \tag{3}$$

with again $q_{i|i} = 0$. SNE tries to find a low-dimensional data representation which minimizes the mismatch between the conditional probabilities $p_{j|i}$ and $q_{j|i}$. As a measure of mismatch the Kullback–Leibler divergence $\mathrm{D_{KL}}$ is used such that the cost function SNE is given by

$$C = \sum_i \mathrm{D_{KL}}(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}, \tag{4}$$

where Q_i is defined similar to P_i as the conditional probability distribution over all other points given ξ_i. The cost function is not symmetric and focuses on retaining the local structure of the data in the mapping. Large costs appear for mapping nearby data points widely separated in the embedding, but there is only small cost for mapping widely separated data points close together. The minimization of the cost function equation (4) is performed using a gradient descent approach. For details we refer to [15].

The so-called "crowding problem" may be observed in SNE and other local techniques, like for example Sammon mapping [16]. The (even very small) attractive forces might crush together moderately dissimilar points in the center of the map. Therefore, in [16] van der Maaten and Hinton presented a technique called t-SNE, which is a variation of SNE considering another statistical model assumption for the data distribution to avoid that problem. Instead of using the conditional probabilities $p_{j|i}$ and $q_{j|i}$ the joint probability distributions P and Q are used to optimize a symmetric version of SNE with the cost function

$$C = \mathrm{D_{KL}}(P\|Q) = \sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{5}$$

with $p_{ii} = q_{ii} = 0$. Here, the pairwise similarities in the high-dimensional space are defined by the conditional probabilities

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{6}$$

and the low-dimensional similarities are given by

$$q_{ij} = \frac{(1+\|\xi_i - \xi_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|\xi_k - \xi_l\|^2)^{-1}}. \tag{7}$$

The application of the heavy-tailed Student t-distribution with one degree of freedom allows to model moderate distances in the high-dimensional space by much larger distances in the embedding. Therefore, the unwanted attractive forces between map points that represent moderately dissimilar data points is eliminated. See [16] for further details.

3. A generalized framework

In this paper we provide the mathematical framework for the generalization of t-SNE and SNE, with respect to the use of arbitrary divergences in the cost-function for the gradient descent. We generalize the definitions towards continuous measures in the high-dimensional space $\mathcal{X} = \{x,y\}$ and a low-dimensional space $\mathcal{E} = \{\xi,\zeta\} \in \mathbb{R}^M$. The pairwise similarities in the high-dimensional original data space are set to

$$p = p_{xy} = \frac{p_{y|x} + p_{x|y}}{2\cdot\int 1\, dy'}, \tag{8}$$

with conditional probabilities

$$p_{y|x} = \frac{\exp(-\|x-y\|^2/2\sigma_x^2)}{\int\exp(-\|x-y'\|^2/2\sigma_x^2)\, dy'}.$$

3.1. The generalized t-SNE gradient

Let $\mathrm{D}(p\|q)$ be a divergence for non-negative integrable measure functions $p = p(r)$ and $q = q(r)$ with a domain V and $\xi,\zeta \in \mathcal{E}$ distributed according to $\Pi_\mathcal{E}$ [26]. Further, let $r(\xi,\zeta) : \mathcal{E}\times\mathcal{E}\to\mathbb{R}$ with the distribution $\Pi_r = \phi(r,\Pi_\mathcal{E})$. Let us use the squared Euclidean distance in the low dimensional space:

$$r = r(\xi,\zeta) = \|\xi - \zeta\|^2. \tag{9}$$

For t-SNE, q is obtained by means of a Student t-distribution, such that

$$q(r(\xi',\zeta')) = \frac{(1+r(\xi',\zeta'))^{-1}}{\iint(1+r(\xi,\zeta))^{-1}\, d\xi\, d\zeta}, \tag{10}$$

which we will abbreviate below for reasons of clarity as

$$q(r') = \frac{(1+r')^{-1}}{\int\int(1+r)^{-1}\, d\xi\, d\zeta} = f(r')\cdot I^{-1}. \tag{11}$$

Now let us consider the derivative of D with respect to ξ:

$$\begin{aligned}\frac{\partial \mathrm{D}}{\partial\xi} &= \frac{\partial \mathrm{D}(p,q(r(\xi,\zeta)))}{\partial\xi} = \iint\frac{\delta \mathrm{D}}{\delta r'}\frac{\partial r'}{\partial\xi}\, d\xi'\, d\zeta' \\ &= \iint\frac{\delta \mathrm{D}}{\delta r(\xi',\zeta')}\frac{\partial r(\xi',\zeta')}{\partial\xi}\, d\xi'\, d\zeta' = 4\int\frac{\delta \mathrm{D}}{\delta r(\xi,\zeta)}(\xi-\zeta)\, d\zeta.\end{aligned} \tag{12}$$