Neurocomputing 112 (2013) 92–108

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation

John A. Lee [a,*,1], Emilie Renard [b], Guillaume Bernard [a], Pierre Dupont [b], Michel Verleysen [b]

[a] Université catholique de Louvain, Molecular Imaging, Radiotherapy, and Oncology—IREC, Avenue Hippocrate 55, B-1200 Bruxelles, Belgium
[b] Université catholique de Louvain, Machine Learning Group—ICTEAM, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Available online 7 March 2013

Keywords:
Dimensionality reduction
Manifold learning
Stochastic neighbor embedding
Divergence

ABSTRACT

Stochastic neighbor embedding (SNE) and its variants are methods of dimensionality reduction (DR) that involve normalized softmax similarities derived from pairwise distances. These methods try to reproduce in the low-dimensional embedding space the similarities observed in the high-dimensional data space. Their outstanding experimental results, compared to previous state-of-the-art methods, originate from their capability to foil the curse of dimensionality. Previous work has shown that this immunity stems partly from a property of shift invariance that allows appropriately normalized softmax similarities to mitigate the phenomenon of norm concentration. This paper investigates a complementary aspect, namely, the cost function that quantifies the mismatch between similarities computed in the high- and low-dimensional spaces. Stochastic neighbor embedding and its variant *t*-SNE rely on a single Kullback–Leibler divergence, whereas a weighted mixture of two dual KL divergences is used in neighborhood retrieval and visualization (NeRV). We propose in this paper a different mixture of KL divergences, which is a scaled version of the generalized Jensen–Shannon divergence. We show experimentally that this divergence produces embeddings that better preserve small *K*-ary neighborhoods, as compared to both the single KL divergence used in SNE and *t*-SNE and the mixture used in NeRV. These results allow us to conclude that future improvements in similarity-based DR will likely emerge from better definitions of the cost function.

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 (HD) space to the low-dimensional (LD) one. For instance, principal component analysis (PCA) [1–3] and classical multidimensional scaling (MDS) [4–6] rely on linear projections that maximize variance preservation and dot product preservation, respectively. Nonlinear variants of metric MDS [7] are based on (weighted) distance preservation: they build a low-dimensional embedding that reproduce as faithfully as possible the pairwise distances measured in the data space. These distances can be Euclidean or approximation of geodesic lengths [8–12]. The use of similarities in DR is quite recent and emerged with methods based on spectral optimization. Among many other examples, Laplacian eigenmaps [13], locally linear embedding [14], and diffusion maps [15] involve sparse matrices of similarities, also called affinity matrices. In spite of a sound theoretical framework, these methods fail to outperform older techniques in typical visualization tasks [16–18]. A possible explanation is that these methods can be reformulated into classical MDS achieved in an unknown feature space [19,20]. In this case, the definition of the similarities merely determines the implicit, arbitrary nonlinear mapping from the data space to the feature space [21,22].

Genuine similarity preservation appeared later with a technique called stochastic neighbor embedding (SNE) [23]. In contrast with spectral methods that directly convert the pairwise similarities defined in the HD space into inner products, SNE matches similarities that are computed both in the HD and LD spaces. To some extent, the set of normalized similarities between a given datum and all others can be seen as an a priori distribution of neighbors, which justifies the term 'stochastic' in the method name. Interest in the new paradigm developed in SNE grew

* Corresponding author. Tel.: +32 2 7649528.
E-mail address: john.lee@uclouvain.be (J.A. Lee).
[1] J.A.L. is a Research Associate with the Belgian National Fund of Scientific Research (F.R.S.-FNRS).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.036

significantly after the publication of variants such as Student t-distributed SNE (t-SNE) [18] and NeRV [24], standing for neighborhood retrieval and visualization. These variants led to breakthroughs in terms of DR quality, with outstanding experimental results [18,24]. Nevertheless, the reasons of this performance leap remain partly unknown. The role played by SNE's very specific similarities has been investigated in [25], where it was shown that similarities defined as softmax ratios with an appropriate normalization benefit from a shift invariance property. This normalization allows the similarities to alleviate the phenomenon of norm concentration [26], which has been identified as the main cause of the poor performance of DR techniques based on distance preservation [25]. In that perspective, SNE and its variants are among the seldom nonlinear DR methods that effectively defeat the curse of dimensionality [27,28]. Another approach is followed in [29,30], where shift invariance is gained in a MDS variant by maximizing the Pearson correlation between the (vectorized) distance matrices in the HD and LD spaces, instead of the norm of their difference.

This paper focuses on a complementary aspect of similarity-based DR, namely, the definition of pertinent cost functions [31]. In SNE and its t-distributed variant t-SNE, the cost function is a sum of Kullback–Leibler (KL) divergences. For each datum, a divergence measures the mismatch between an a priori distribution of its neighbors in the HD space and the corresponding distribution computed in the LD space. As the KL divergence is asymmetric with respect to the two distributions it compares, NeRV also involves the 'dual' KL divergence, where the two distributions are swapped. A metaparameter controls the weight of two dual divergences in the cost function. In an information retrieval perspective, it has been shown that this metaparameter allows NeRV to reach different tradeoffs between *precision* and *recall* [24]. According to the nomenclature developed in [32], NeRV entails a type 1 mixture of KL divergences, that is, a linear mixture of two dual divergences. For equal mixture weights, the resulting divergence is symmetric with respect to the two compared distributions. In this paper, we investigate a type 2 mixture of KL divergences [32], which involves a composite distribution and also a nonlinear mixture of two divergences. This second type of mixture is closely related to the generalized Jensen–Shannon divergence [33,34] and, like the type 1 mixture, it is also symmetric when both weights are equal to one half. Using a criterion of K-ary neighborhood preservation, we show experimentally that the type 2 mixture outperforms both the type 1 mixture and the usual non-blended divergence. A careful examination of the gradient of each mixture reveals some clues to justify this better behavior.

The rest of this paper is organized as follows. Section 2 describes the normalized similarities used in SNE and its variants to define a priori distributions of neighbors. Section 3 deals with the two different types of divergence mixtures that measure the mismatch between these distributions. Section 4 focuses on optimization issues and analyzes the gradient of the considered divergence mixtures. Section 5 presents and discusses the experimental results. Finally, Section 6 draws the conclusions and sketches some perspectives.

2. Shift-invariant softmax similarities

Let $\mathbf{\Xi}=[\boldsymbol{\xi}_i]_{1\le i\le N}$ denote a set of N points in some M-dimensional space. Similarly, let $\mathbf{X}=[\mathbf{x}_i]_{1\le i\le N}$ be its representation in a P-dimensional space, with $P\le M$. The Euclidean distances between the ith and jth points are given by $\delta_{ij}=\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|_2$ and $d_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|_2$ in the HD and LD spaces respectively. The term similarity generally refers to a quantity that decreases as the distance grows. In SNE, the similarities associated with δ_{ij} and d_{ij} are defined for $i\neq j$ by

$$\sigma_{ij}=\frac{\exp(-\gamma_{ij})}{\sum_{k,k\neq i}\exp(-\gamma_{ik})} \quad \text{and} \quad s_{ij}=\frac{\exp(-g_{ij})}{\sum_{k,k\neq i}\exp(-g_{ik})}, \tag{1}$$

where $\gamma_{ij}=\gamma(\delta_{ij}/\lambda_i)$ and $g_{ij}=g(d_{ij})$. Functions γ and g are both non-negative with a non-negative derivative. Parameter λ_i is a bandwidth that can be seen as a soft neighborhood radius. By convention, $\sigma_{ij}=s_{ij}=0$ if $i=j$.

In SNE and NeRV, the similarities are Gaussian, with γ and g being defined as

$$\gamma(u)=g(u)=u^2/2. \tag{2}$$

In t-SNE, the similarities in the LD space are defined in a different way than in the HD space, by using an unnormalized probability mass function of a Student t distribution with m degrees of freedom[2]

$$s_{ij}=\frac{(1+s_{ij}^2/m)^{-(m+1)/2}}{\sum_k(1+s_{ij}^2/m)^{-(m+1)/2}}. \tag{3}$$

This amounts to opting for

$$g(u)=\frac{m+1}{2}\ln(1+u^2/m). \tag{4}$$

in (1). The heavier tail of the Student t function, as compared to the Gaussian, induces an exponential transformation between the HD and LD distances [35]. The longer the distance is in the HD space, the stronger it is stretched in the LD space.

An important feature of similarities defined as softmax exponential ratios such as above is their normalization, that is, $\sum_j\sigma_{ij}=\sum_j s_{ij}=1$. Combined with positivity, it allows the similarities σ_{ij} and s_{ij} to be interpreted as a priori probabilities for $\boldsymbol{\xi}_j$ and $\mathbf{x}_j$ to be neighbors of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, respectively. But more importantly, normalization implies scale invariance with respect to $\exp(-\gamma_{ij})$ in σ_{ij}, which in turn translates into shift invariance with respect to $\gamma(\delta_{ij}/\lambda_i)=\delta_{ij}^2/(2\lambda_i^2)$ [25]. Since null distances have a trivial distribution that differs from that of non-zero distances, they are excluded from the sum in the normalization denominators in (1). As a direct result, the shift applicable to δ_{ij}^2 can range from $-\min_{j,j\neq i}\delta_{ij}^2$ to ∞. The lower end of this interval ensures that the shifted distances remain positive. A negative shift close to this lower bound is particularly interesting to mitigate the phenomenon of norm concentration [26]. One manifestation of this phenomenon is the following: for a finite sample of points $\mathbf{\Xi}$, $\min_{j,j\neq i}\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|$ grows faster with M than $\max_j\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|$. In other words, the relative variance of a discrete distribution of Euclidean distances (namely, its variance divided by the square of its mean) vanishes when M tends to ∞. The changing shape of distance distributions, depending on the dimensionality, partly explains the failure of DR methods based on distance preservation. The distances in LD spaces are always and systematically 'too scattered' to match those observed in HD spaces. Invariance to shifts in similarities circumvents this problem [25].

3. Divergences to measure the similarity mismatch

Thanks to positivity and normalization, vectors $\boldsymbol{\sigma}_i=[\sigma_{ij}]_{1\le j\le N}$ and $\mathbf{s}_i=[s_{ij}]_{1\le j\le N}$ can be seen as discrete probability

[2] The exact definition of s_{ij} in [18] also entails a slightly different normalization of the similarity, with a sum in the denominator that runs over both indices instead of the second one only. In practice, doing so simplifies the gradient of t-SNE but has no significant effect on the method results. Our definitions of σ_{ij} and s_{ij} in (1) reproduce those of SNE in [23] and have the advantage of instantiating twice the very same template.