Mathematical Foundations of the Self Organized Neighbor Embedding (SONE) for Dimension Reduction and Visualization

Kerstin Bunte[1], Frank-Michael Schleif[2], Sven Haase[3] and Thomas Villmann[3]

1- University of Groningen - Johann Bernoulli Institute for Mathematics and Computer Science, Nijenborgh 9, Groningen - The Netherlands

2- University of Bielefeld - CITEC Center of Excellence, Bielefeld - Germany

3- University of Apllied Sciences Mittweida, Mittweida - Germany

Abstract. In this paper we propose the generalization of the recently introduced Neighbor Embedding Exploratory Observation Machine (NE-XOM) for dimension reduction and visualization. We provide a general mathematical framework called Self Organized Neighbor Embedding (SONE). It treats the components, like data similarity measures and neighborhood functions, independently and easily changeable. And it enables the utilization of different divergences, based on the theory of Fréchet derivatives. In this way we propose a new dimension reduction and visualization algorithm, which can be easily adapted to the user specific request and the actual problem.

1 Introduction

Various dimension reduction techniques have been introduced based on different properties of the original data to be preserved. The spectrum ranges from linear projections of original data, such as in Principal Component Analysis (PCA) or classical Multidimensional Scaling (MDS) to a wide range of locally linear and non-linear approaches, such as Isomap, Locally Linear Embedding (LLE), Local Linear Coordination (LLC), or charting. Stochastic Neighbor Embedding (SNE)approximates the probability distribution in the high-dimensional space, defined by neighboring points, with their probability distribution in a lower-dimensional space. A technique called t-SNE was proposed in [10]. It is a variation of SNE considering another statistical model assumption for data distributions. 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) [5] and generalizations thereof and extensions of the Self Organizing Map (SOM) incorporating class labels. For a comprehensive review on nonlinear dimensionality reduction methods, we refer to [7]. Recently, the idea of fast and efficient online learning was combined with the high-quality of divergence based optimization, resulting in a new dimension reduction algorithm called Neighbor Embedding XOM (NE-XOM). Its usefulness and comparison with other methods is shown in [3]. The authors connected a computational approach to topology learning, the Exploration Observation Machine (XOM) as introduced in [12], with the divergence optimization of SNE. In this contribution, we extend the approach proposed in [3], with a mathematical foundation for the generalization of the principle to

arbitrary divergences based on Fréchet derivatives. This generalized framework is called Self Organized Neighbor Embedding (SONE) in the following. In this way we propose a new dimension reduction and visualization algorithm, which can easily adapted to the user specific request and the actual problem. We will describe the NE-XOM extension in section 2, describe the new generalized framework SONE in section 3, show the extension for some famous families of divergences and conclude in section 5.

2 The Neighbor Embedding XOM (NE-XOM)

In this section we review the combination of direct divergence inspired by SNE with fast sequential online learning resulting in a new algorithm called Neighbor Embedding XOM (NE-XOM) introduced in [3]. The original XOM algorithm maps a finite number of high-dimensional data points $\boldsymbol{x}^i \in \mathcal{X}$ in the observation space $\mathcal{X}$ to low-dimensional image vectors $\boldsymbol{y}^i \in \mathcal{E}$ in the embedding space $\mathcal{E}$. The embedding space is associated with a structure hypothesis, given by a number of sampling vectors $\boldsymbol{s} \in \mathcal{E}$, which corresponds to the final structure in which the data is embedded. Reasonable choices for the sampling vectors $\boldsymbol{s}$ are: the location on a regular lattice structure in $\mathcal{E}$, discrete positions in $\mathcal{E}$ as representation of a finite number of class centers, drawn from a mixture of Gaussian to represent a finite number of clusters, or uniformly sampled in a region of $\mathcal{E}$ to indicate that the visualization of the data should occupy the full projection space. For the extension, let $h_\sigma(d_\mathcal{X}(\Psi_{\mathrm{GKL}}(\boldsymbol{s}), \boldsymbol{x}^k))$ and $g_\varsigma(d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k))$ $\Big($abbreviated by $h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})}(k)$ and $g_\varsigma^{\boldsymbol{s}}(k)\Big)$ be any positive integrable measures denoting the neighborhood cooperation in the observation and the embedding space respectively. Following the ideas of SNE, NE-XOM tries to minimize the difference between these two neighborhood functions measured by the Kullback-Leibler (KL) divergence. Note, that in contrast to SNE, which is originally defined for probability densities $p(r)$ with scalar r, the constraint $\int p(r)\ dr = 1$ is not imposed here. The neighborhood function $h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})}$ of the observation space $\mathcal{X}$ might be a Gaussian: $h_\sigma^{ij} = \exp\left(\frac{-d_\mathcal{X}(\boldsymbol{x}^i, \boldsymbol{x}^j)}{2\sigma^2}\right)$ with $\sigma > 0$. Depending on the choice for the neighborhood cooperation g_ς in the embedding space with variance ς the learning rule and thus the final embedding may vary a lot. We will provide in the following the learning rules for the case of a Gaussian neighborhood cooperation:

$$g_\varsigma^{\boldsymbol{s}}(k) = \exp\left(\frac{-d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k)}{2\varsigma^2}\right) \text{ derivative: } \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} = \left(-\frac{g_\varsigma^{\boldsymbol{s}}(k)}{2\varsigma^2}\right) \frac{\partial d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \quad (1)$$

and a t-Distribution-like cooperation function:

$$g_\varsigma^{\boldsymbol{s}}(k) = \left(1 + \frac{d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k)}{\varsigma}\right)^{\left(-\frac{\varsigma+1}{2}\right)} \text{ deriv.: } \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} = \frac{\left(-\frac{\varsigma+1}{2\varsigma}\right) g_\varsigma^{\boldsymbol{s}}(k)}{(1 + \frac{d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k)}{\varsigma})} \frac{\partial d_\mathcal{E}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \ . \quad (2)$$

For positive measures p and q the Generalized Kullback-Leibler (GKL) divergence: $\mathrm{D}_{\mathrm{GKL}}(p\|q) = \int p(r) \ln\left(\frac{p(r)}{q(r)}\right)\, dr - \int [p(r) - q(r)]\, dr$ is used. We are able to define a cost function using the neighborhood functions from the original and the embedding space and the GKL divergence $\mathrm{D}_{\mathrm{GKL}}$:

$$E_{\text{NE-XOM}} \sim \int \sum_i \delta_{\Psi_{\mathrm{GKL}}(\boldsymbol{s}),\boldsymbol{x}^i} \cdot \sum_j \mathrm{D}_{\mathrm{GKL}}\left(h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})}(j) | g_\varsigma^{\boldsymbol{s}}(j)\right)\, p(\boldsymbol{s}) d\boldsymbol{s} \tag{3}$$

where $h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})} = h_\sigma(d_{\mathcal{X}}(\Psi_{\mathrm{GKL}}(\boldsymbol{s}), \boldsymbol{x}^k))$ and the best match data point $\Psi_{\mathrm{GKL}}(\boldsymbol{s})$ for a given sampling vector $\boldsymbol{s}$ is given by

$$\Psi_{\mathrm{GKL}}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \mathrm{D}_{\mathrm{GKL}}\left(h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})}(j) \| g_\varsigma^{\boldsymbol{s}}(j)\right) \text{ is minimum.} \tag{4}$$

This results in the learning rule for the NE-XOM:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \tau \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}_k} \left(1 - \frac{h_\sigma^{\Psi_{\mathrm{GKL}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}\right)\ , \tag{5}$$

In the following sections we will generalize this concept for arbitrary divergences.

3 A Generalized Framework for Dimension Reduction

Divergences can be an alternative to the most frequently used Euclidean distance and may lead to improved classification accuracy. Furthermore divergences can be applied in the field of dimension reduction: for example in Stochastic Neighbor Embedding (SNE), t-distributed SNE (t-SNE) and Multidimensional Scaling (MDS) [10, 6]. In [11] the mathematical foundation to extend SNE and t-SNE for use with arbitrary divergences is given. We will use this concept to generalize the algorithm explained in section 2.

Divergences are functionals $\mathrm{D}(p\|q)$ designed as dissimilarity measures between two nonnegative integrable functions p and q [4]. In practice, usually p corresponds to the observed data and q denotes the estimated or expected data. We call p and q positive measures defined on r in the domain V. The weight of the functional p is defined as $W(p) = \int_V p(r)\, dr$. Positive measures with the additional constraint $W(p) = 1$ are denoted as probability density functions. Generally speaking, divergences measure a quasi-distance or directed difference. In contrast to a metric, a divergence must not be symmetric in the sense $\mathrm{D}(p\|q) = \mathrm{D}(q\|p)$ and does not necessarily satisfy the triangular inequality $\mathrm{D}(p\|q) \leq \mathrm{D}(p\|z) + \mathrm{D}(z\|q)$. Note, that the definition of the considered divergences for non-normalized positive measures has an important property. It allows the analysis of patterns of different size to be weighted differently, e. g. images with different size or documents of variable length. Following [4] one can distinguish at least three main families of divergences with the same consistent properties: Bregman-divergences, Csiszár's f-divergences and γ-divergences. Note that all these families contain the Kullback-Leibler (KL) divergence as special case, so the KL-divergence can be seen as the non empty intersection.

In the following we will briefly review the concept of Fréchet derivatives and we will define the mathematical framework for Self Organized Neighbor embedding (SONE) using arbitrary divergences.

We use the concept of Fréchet derivatives of a function f defined on Banach spaces:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(L[f + \epsilon h] - L[f] \right) =: \frac{\delta L[f]}{\delta f}[h] \ . \tag{6}$$

for the generalization of the definition given in Eq. (5). Detailed descriptions and formulas can be found in [2].

3.1 Self Organized Neighbor Embedding (SONE)

We define a cost function for arbitrary Divergences $\mathrm{D}(p\|q)$:

$$E_{\mathrm{SONE}} = \int \sum_i \delta_{\Psi^{\mathrm{D}}(\boldsymbol{s}),\boldsymbol{x}^i} \cdot \sum_j \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})}(j) \middle\| g_\varsigma^{\boldsymbol{s}}(j) \right) \ p(\boldsymbol{s}) d\boldsymbol{s} \ , \tag{7}$$

where the best matching data point $\Psi^{\mathrm{D}}(\boldsymbol{s})$ for $\boldsymbol{s}$ is defined as:

$$\Psi_{\mathrm{D}}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})}(j) \middle\| g_\varsigma^{\boldsymbol{s}}(j) \right) \text{ is minimum.} \tag{8}$$

Let V be a Banach space and $U \subset V$ an open subset of V. The divergence $\mathrm{D} : U \to \mathbb{R}$ is defined as a mapping from U to $\mathbb{R}$. Further D uses a bounded linear operator: the integral $\int : V \to \mathbb{R}$. So the derivative of the cost function (7) with respect to the image vectors $\boldsymbol{y}^k$ can be done using the Fréchet derivative Eq. (6):

$$\frac{\partial E_{\mathrm{SONE}}}{\partial \boldsymbol{y}^k} = \int \left[\frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}}[l] \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}}{\partial \boldsymbol{y}^k} \right] \ dl = \left. \frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}} \right|_k \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} \ . \tag{9}$$

This yields the online learning update rule for $\boldsymbol{s}$ and learning rate τ:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \tau \Delta \boldsymbol{y}^k \quad \text{with} \quad \Delta \boldsymbol{y}^k = \left. \frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}} \right|_k \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} \tag{10}$$

The explicit formulas for the special learning rules in case of Gaussian and t-distribution (Eq. (1) and (2)) and different divergences can be found in [2].

4 Example

The identification of bacteria is an important task in medicine or biology and is often done using large databases with reference signatures [9]. The reference spectra of the different bacteria species are in parts very similar and multi-modal

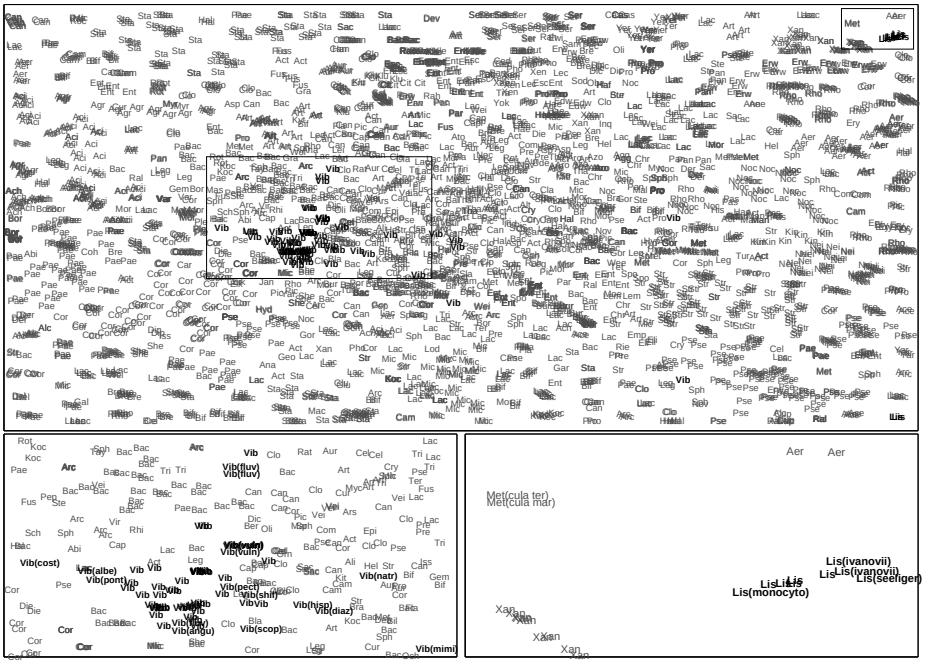

Fig. 1: Similarity map of the bacteria data set and two selected zoomed regions.

as an additional challenge for the identification methods. To maintain these databases efficient exploration and visualization tools are necessary. Common tasks are the identification of outliers, strong overlapping and therefore hard to distinguish data clusters or erroneous measurements.

Here we consider a database of $N = 3048$ bacteria samples measured and prepared in accordance to [1, 9]. Each sample is given as a vector $\boldsymbol{x} \in \mathbb{R}^D$, with dimensionality D (number of peaks), considered as a function p. Overall the data contain around 200 species in accordance to the taxonomy of bacteria and are quite challenging for visualization. For each x a labeling is available shown in Fig. 1 by a three letter code. The map obtained with t-SNE is also able to separate some clusters of bacteria, but the center is more crowded then the SONE map SONE allows to influence the granularity and enforce spreading of the data samples on the given structure hypothesis, which enhances visibility of single samples. The quality of the both the SONE and t-SNE embedding measured with the method proposed in [8] behaves quite similar.

The SONE representation was already quite effective in representing the many bacteria spectra and similar samples are indeed plotted near to each other, which is in good agreement to the expectations of the experts [9]. The map also allows to identify isolated clusters like the one depicted in the right subplot. This plot contains most of the *Listeria* spectra from the database which are known to be very distinctive. For the second subplot (left) a large cohort of *Vibrio* spectra is shown. It is more diverse and very well represented, but we can also identify more distant Vibrio items which by closer inspection are indeed special cases.

The map allows the biochemical expert to navigate through the similarity space and to analyze spectra found to be (dis-)similar by the model.

5 Conclusion

In this article we provide the mathematical foundation for a generalization of Self Organized Neighbor Embedding (SONE) which can be applied in dimension reduction and visualization tasks. The framework allows for the use of a very broad class of divergences as cost function. In this context, we first present a general formulation of SONE as a gradient based optimization scheme. The use of a particular dissimilarity measure requires the availability of its Fréchet-derivative, which we present for a wide class of divergences. Detailed descriptions and formulas can be found in [2].

We showed the applicability in the experiment section on the example of a similarity map in the domain of Bacteria diversity. In forthcoming studies we will examine the role of the different divergence families and their advantages for data domains.

Acknowledgment

This work was supported by the "Nederlandse organisatie voor Wetenschappelijke Onderzoek (NWO)" under project code 612.066.620 and by the "German Science Foundation (DFG)" under grant number HA-2719/4-1. Further, financial support from the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative is gratefully acknowledged.

References

[1] S. B. Barbuddhe, T. Maier, G. Schwarz, M. Kostrzewa, H. Hof, E. Domann, T. Chakraborty, and T. Hain. Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology*, 74(17):5402–5407, 2008.

[2] K. Bunte, S. Haase, M. Biehl, and T. Villmann. Mathematical Foundations of Self Organized Neighbor Embedding (SONE) for Dimension Reduction and Visualization. Technical Report MLR-03-2010, Leipzig University, 2010. ISSN:1865-3960.

[3] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Neighbor embedding XOM for dimension reduction and visualization. Accepted in Neurocomputing, 2010.

[4] A. Cichocki, R. Zdunek, A. Phan, and S. I. Amari. *Non-negative matrix and tensor factorizations*. Wiley, Chichester, 2009.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd edn. (Computer Science and Scientific Computing Series)*. Academic Press, September 1990.

[6] P. L. Lai and C. Fyfe. Bregman divergences and multi-dimensional scaling. In *ICONIP, Revised Selected Papers, Part II*, pages 935–942. Springer-Verlag, 2008.

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

[8] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomput.*, 72(7-9):1431–1443, 2009.

[9] T. Maier, S. Klebel, U. Renner, and M. Kostrzewa. Fast and reliable maldi-tof ms-based microorganism identification. *Nature Methods*, (3), 2006.

[10] L. J. P. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

[11] T. Villmann and S. Haase. Mathematical foundations of th generalization of t-SNE amd SNE for arbitrary divergences. Technical Report MLR-02-2010, Leipzig University, 2010.

[12] A. Wismüller. Exploration-organized morphogenesis (XOM) – a general framework for learning by self-organization. In *FIPKM*, volume 37, pages 205–239, 2001.