Supervised dimension reduction mappings

Kerstin Bunte[1], Michael Biehl[1] and Barbara Hammer[2]

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

Abstract. We propose a general principle to extend dimension reduction tools to explicit dimension reduction mappings and we show that this can serve as an interface to incorporate prior knowledge in the form of class labels. We explicitly demonstrate this technique by combining locally linear mappings which result from matrix learning vector quantization schemes with the t-distributed stochastic neighbor embedding cost function. The technique is tested on several benchmark data sets.

1 Introduction

In many areas such as robotics, medicine, biology, etc. electronic data sets are increasing rapidly with respect to size and complexity. On the one hand these data can provide additional useful information for human users in the respective field. On the other hand, it becomes more and more difficult to directly access the information. As a consequence, many data visualization and dimensionality reduction techniques have emerged in the last years which help humans to rapidly scan through large volumes of data relying on their cognitive capabilities for visual perception [6, 11, 12]. Dimension reduction techniques can be decomposed into classical linear techniques such as principle component analysis (PCA) or linear discriminant analysis (LDA), and modern nonlinear tools involving, for example, locally linear embedding (LLE), Isomap, t-distributed stochastic neighbor embedding (t-SNE), maximum variance unfolding (MVU), etc. [6]. Most of the latter techniques, however, provide a mapping of the given data points only, rather than an explicit embedding function. As a consequence, additional effort has to be taken for out-of-sample extensions. In addition, the outcome of the models heavily relies on the chosen criterion, consequently dimensionality reduction is an inherently ill-posed problem. In this contribution we propose a general principle to extend dimension reduction tools to obtain an explicit mapping with fixed prior shape. This has two consequences: It allows immediate out-of-sample extensions and one can directly access the generalization ability of the models. We show in examples that the latter is excellent, implying that the techniques can drastically be accelerated by reducing training to only a small subset. In addition, the integration of prior knowledge in the form of class labels is easily possible by biasing the dimensionality reduction mapping towards this auxiliary information. We show the feasibility and efficiency of this approach by a direct comparison to recent alternatives as proposed e.g. in [12].

2 Learning Mappings for Dimension Reduction

Most dimension reduction methods produce a mapping of data points $\mathbb{R}^N \ni \boldsymbol{x}^i \to \boldsymbol{y}^i \in \mathbb{R}^2$, only. The embedding of new points requires additional computation, often the optimization is run again keeping all known points fixed. Besides the additional effort, this method has the drawback that it is difficult to investigate the generalization ability of these mappings and some effort has

to be done to integrate auxiliary information into the mapping prescription. We can avoid these problems by the definition of an explicit dimension reduction mapping function $f : \mathbb{R}^N \to \mathbb{R}^2$, $\boldsymbol{x}^i \to \boldsymbol{y}^i = f(\boldsymbol{x}^i)$ for the projection of the points. Essentially, we propose to fix a parameterized form of f prior to training. Then the parameters are optimized according to the objective as specified by the respective dimension reduction method.

In the literature, a few dimension reduction technologies provide an explicit mapping of the data: linear methods such as PCA provide an explicit linear function [1]. Nonlinear extensions thereof can be realized by autoencoder networks Manifold charting starts from locally linear embeddings given by local PCAs and glues these pieces together by minimizing the error on the overlaps [3, 9]. Topographic maps such as the self-organizing map (SOM) or generative topographic mapping (GTM) characterize data in terms of prototypes which are visualized in low dimensions [2, 5]. Due to the clustering, new data can directly be visualized by mapping these data to their closest prototype.

A few dimension reduction mappings which give coordinates per default have been extended to global mappings. Locally linear coordination (LLC) [9] extends locally linear embedding (LLE) by assuming that locally linear methods, such as local PCAs, are available, and by glueing them together adding affine transformations. The additional parameters are optimized using the LLE cost function. Parameterized t-distributed stochastic neighbor embedding (t-SNE) [10] extends t-SNE towards an embedding given by a multilayer neural network. The network parameters are determined using back propagation, where, instead of the mean squared error, the t-SNE cost function is taken as objective.

2.1 A General Principle

Popular dimension reduction techniques include methods which try to preserve distances such as multi dimensional scaling (MDS) or Isomap, or they preserve more general information connected to the neighborhood graph such as Laplacian eigenmaps. Furthermore some techniques try to preserve locally linear relationships such as LLE, pairwise distributions such as SNE and t-SNE or neighborhood distances giving maximum variance such as maximum variance unfolding, etc. Many of these approaches can be put into a common framework: characteristics of the original data points $\boldsymbol{x}^i$ are computed (such as pairwise distances, pairwise geodesic distances, locally linear relationship, etc.) and the same or similar characteristics are induced by the projected points $\boldsymbol{y}^i$. The goal is to find coefficients of the projections such that these two characteristics match as far as possible as measured by some cost function. Possibly additional constraints or objectives are formalized to achieve uniqueness. The methods differ in the choice of the data characteristics, the choice of the error measure, and the way in which optimization takes place. Thus, considering dimension reduction as optimization task allows us to formalize many different dimension reduction methods in a common framework as detailed in Table 1.

method	characteristics of data	characteristics of projections	error measure
MDS	Euclidean distance $d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j)$	Euclidean distance $d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)$	minimize weighted least squared error
Isomap	Geodesic distance $d_{\text{geodesic}}(\boldsymbol{x}^i, \boldsymbol{x}^j)$	Euclidean distance $d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)$	minimize weighted least squared error
LLE	reconstruction weights w_{ij} such that $\sum(\boldsymbol{x}^i - \sum_{i\to j} w_{ij}\boldsymbol{x}^j)^2$ is minimum with constraints $\sum_j w_{ij} = 1$	reconstruction weights $\tilde{w}_{ij}$ such that $\sum(\boldsymbol{y}^i - \sum_{i\to j} \tilde{w}_{ij}\boldsymbol{y}^j)^2$ is minimum with constraints $\sum \boldsymbol{y}^i = 0$, $\mathbf{Y}^t\mathbf{Y} = \mathbf{n}$	enforce identity $w_{ij} = \tilde{w}_{ij}$
Laplacian eigenmap	negative heat kernel weights $-w_{ij} = \exp(-d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j)^2/t)$ for $i \to j$	squared Euclidean distance $d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)^2$ for $i \to j$ with constraints $\mathbf{Y}^t D\mathbf{Y} = \mathbf{1}$, $\mathbf{Y}^t D\mathbf{1} = \mathbf{0}$	maximize correlation
MVU	Euclidean distance $d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j)$ for $i \to j$	Euclidean distance $d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)$ for $i \to j$ such that $\sum_{ij} d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)^2$ is maximum and $\sum_i \boldsymbol{y}^i = 0$.	enforce identitiy (introducing slack variables if necessary)
SNE	probab. $p_{j\|i} = \frac{\exp\left(-d_{\mathcal{X}}(\boldsymbol{x}^i,\boldsymbol{x}^j)^2/2\sigma_i\right)}{\sum_{k\neq i}\exp\left(-d_{\mathcal{X}}(\boldsymbol{x}^i,\boldsymbol{x}^k)^2/2\sigma_i\right)}$	probab. $q_{j\|i} = \frac{\exp\left(-d_{\mathcal{E}}(\boldsymbol{y}^i,\boldsymbol{y}^j)^2\right.}{\sum_{k\neq i}\exp\left(-d_{\mathcal{E}}(\boldsymbol{y}^i,\boldsymbol{y}^k)^2\right)}$	minimize Kullback-Leibler divergences
t-SNE	probab. $p_{ij} = \frac{p_{j\|i}+p_{i\|j}}{2n}$	probab. $q_{ij} = \frac{(1+d_{\mathcal{E}}(\boldsymbol{y}^i,\boldsymbol{y}^j)/\varsigma)^{-\frac{\varsigma+1}{2}}}{\sum_{k\neq l}(1+d_{\mathcal{E}}(\boldsymbol{y}^k,\boldsymbol{y}^l)/\varsigma)^{-\frac{\varsigma+1}{2}}}$	minimize Kullback-Leibler divergence

Table 1: Many dimensionality reduction methods can be put into a general framework: characteristics of the data are extracted. Projections lead to corresponding characteristics depending on the coefficients. These coefficients are determined such that an error measure of the characteristics is minimized, fulfilling probably additional constraints.

data set	description
letter recognition	contains 20,000 samples with 16 dimensions, separated into 26 classes, which are 4×4 images of the 26 capital letters
phoneme	consists of 3656 phoneme samples represented by a 20-dimensional vector with labels corresponding to 13 classes
landsat satellite	contains 6435 samples of 3×3 satellite images measured in four spectral bands resulting in 36 dimensions classes: red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, very damp grey soil

Table 2: Information about the data sets evaluated in the experimental section.

This general view allows us to simultaneously extend all methods to dimension reduction mappings in a general way. In a first step, the principled form and complexity of the mapping is fixed by a parameterized function $f_W : \mathbb{R}^N \to \mathbb{R}^2$ with parameters W. This function can be given by a linear function, a locally linear function, a feedforward neural network, etc.. Then, instead of coefficients $\boldsymbol{y}^i$, the images of the map $f_W(\boldsymbol{x}^i)$ are considered and the map parameters W are optimized according to the characteristics of the data and corresponding error measure, respectively. This principle leads to a well defined mathematical objective for the mapping parameters W for every dimensionality reduction method as summarized above. The way in which optimization takes place is possibly different from the original method: while numerical methods such as gradient descent can still be used, it is frequently no longer possible to find closed form solutions for spectral methods. However, numerical optimization can be used as a default in all cases. We exemplify the above in terms of locally linear mappings built on top of locally linear projections, whereby we combine these functions with the t-SNE cost term. We emphasize the possibility to integrate auxiliary information into the process, and we conduct corresponding experiments later.

2.2 Supervised Locally Linear t-SNE Mapping

We can impose a global nonlinear embedding function on top of locally linear projections obtained, e.g., using prototype based methods [7, 8]. Since we want to obtain a supervised visualization of data which emphasizes the aspects relevant for a given labeling of the data, we take locally linear projections which are biased according to the given auxiliary information, i.e. the projections are given by supervised prototype based methods such as matrix learning vector quantization [8, 4]. We assume that locally linear projections have the form:

$$\boldsymbol{x}^l \mapsto p_k(\boldsymbol{x}^l) = \Omega_k \boldsymbol{x}^l - \boldsymbol{w}^k \tag{1}$$

with local matrices Ω_k and prototypes $\boldsymbol{w}^k$. Further, we assume the existence of responsibilities r_{lk} of mapping p_k for $\boldsymbol{x}^l$, which can be given by the receptive fields of the locally linear maps centered around $\boldsymbol{w}^k$ or Gaussians centered around these points, for example. We assume $\sum_k r_{lk} = 1$. Then a global mapping which combines these linear pieces can be defined as

$$f_W : \boldsymbol{x}^l \mapsto \boldsymbol{y}^l = \sum_k r_{lk}(L_k \cdot p_k(\boldsymbol{x}^l) + l_k) \ , \tag{2}$$

using locally linear projections L_k and local offsets l_k to align the local pieces. Note that the dimensionality of the weights W which have to be determined depends on the number of pieces k and the dimensionality of the local projections. Usually, it is much smaller than the number of coefficients when projecting all points $\boldsymbol{y}^l$ directly to the Euclidean plane. These parameters can be determined by a stochastic gradient descent. The derivative of the t-SNE cost function yields

$$\begin{aligned}
\frac{\partial E_{\text{t-SNE}}}{\partial L_k} &= \frac{\varsigma+1}{\varsigma} \sum_{ij} \frac{(p_{ij} - q_{ji})}{1 + d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)/\varsigma} \cdot (\boldsymbol{y}^i - \boldsymbol{y}^j)(r_{ik} p_k(\boldsymbol{x}^i) - r_{jk} p_k(\boldsymbol{x}^j)) \\
\frac{\partial E_{\text{t-SNE}}}{\partial l_k} &= \frac{\varsigma+1}{\varsigma} \sum_{ij} \frac{(p_{ij} - q_{ji})}{1 + d_{\mathcal{E}}(\boldsymbol{y}^i, \boldsymbol{y}^j)^2/\varsigma} \cdot (\boldsymbol{y}^i - \boldsymbol{y}^j)(r_{ik} - r_{jk})
\end{aligned}$$

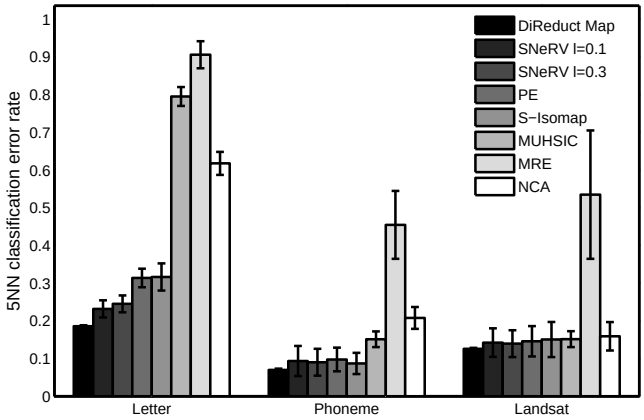

Fig. 1: Comparison of the 5 nearest neighbor errors for all data sets.

assuming Euclidean distance in the projection space. In the following section we evaluate the proposed method using the learning rule based on the t-SNE cost function defined above and locally linear projections obtained by *limited rank matrix learning vector quantization* (LiRaM LVQ) [4] in a supervised manner.

3 Experiments

In this section we evaluate the proposed method on three data sets also investigated in [12] and described in table 2. For learning locally linear projections, we use LiRaM LVQ with the rank of matrices limited to 10, 10 and 30 for the data sets respectively and the number of prototypes equal to the number of classes. For the coordination, we use crisp responsibilities given by the receptive fields. All other parameters are set as default values. We train the mapping using only a small subset of the full data set (7%-18%) and evaluate the results for the full data set by using the mapping.

For evaluation we measure the classification error of the resulting visualization using a 5-nearest neighbor evaluation (5NN error). We compare our results with the results taken from [12] on the same data sets where six state-of-the art supervised nonlinear embeddings are tested (*Supervised neighbor retrieval visualizer* (SNeRV), *Multiple relational embedding* (MRE), *Colored maximum variance unfolding* (MUHSIC), *Supervised isomap* (S-Isomap), *Parametric embedding* (PE), *Neighbourhood component analysis* (NCA)). Note that these methods use only a small subpart of the dataset within an evaluation, while we evaluate our approach on the full data set by means of the explicit mapping. The comparison of the error rates are shown in Figure 1. Interestingly, the classification error obtained by the proposed method is smaller than the alternatives for all three data sets. This is particularly remarkable since we used only a fraction of the data to obtain the map, i.e. the proposed method displays excellent and very efficient generalization to large data sets. The corresponding visualizations display a clear class structure as shown in Figure 2.

4 Conclusion

We have proposed a general way to extend arbitrary dimension reduction techniques, based on cost optimization, to explicit mappings which take into account

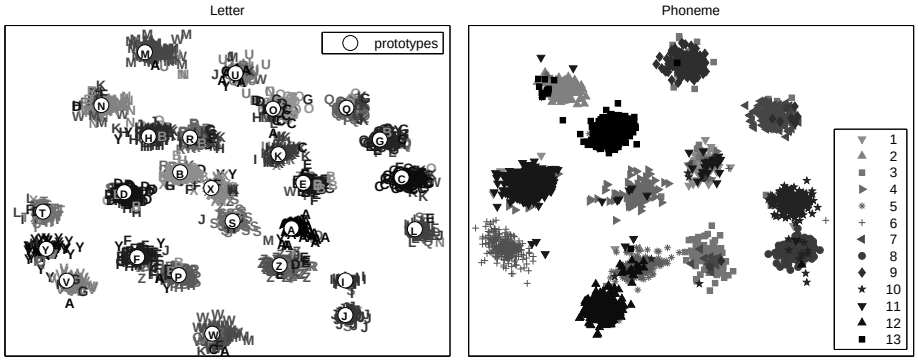

Fig. 2: Visualization of the Letter and Phoneme data set in two dimensions as obtained by the proposed supervised dimension reduction mapping.

prior class labeling. We demonstrated the feasibility of the approach for locally linear maps obtained from matrix learning vector quantization and the t-SNE cost function for their coordination, yielding excellent results for three benchmarks. This technique offers the possibility of very flexible and very efficient visualization, since a bias towards given information can easily be integrated into the form of the mapping function and initial solutions, on the one hand, and a small number of data is sufficient to obtain a visualization of the full data set due to the excellent generalization ability of the technique.

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] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[2] C. M. Bishop and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[3] M. Brand. Charting a manifold. In *Advances in Neural Information Processing Systems NIPS 15*, pages 961–968. MIT Press, 2003.

[4] K. Bunte, B. Hammer, A. Wismüller, and M. Biehl. Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data. *Neurocomputing*, 73(7-9):1074–1092, 2010.

[5] T. Kohonen. *Self-organizing Maps*. Springer, 1995.

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

[7] R. Möller and H. Hoffmann. An extension of neural gas to local PCA. *Neurocomputing*, 62(305-326), 2004.

[8] P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[9] Y. W. Teh and S. Roweis. Automatic alignment of local representations. In *Advances in Neural Information Processing Systems 15*, pages 841–848. MIT Press, 2003.

[10] L. J. P. van der Maaten. Learning a parametric embedding by preserving local structure. In *12th AI-STATS*, number 5, pages 384–391. JMLR W&CP, 2009.

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

[12] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, March 2010.