ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Relevance Learning for Dimensionality Reduction

Alexander Schulz, Andrej Gisbrecht and Barbara Hammer *

CITEC center of excellence, Bielefeld University, Germany

Abstract. Nonlinear dimensionality reduction (NLDR) techniques offer powerful data visualization schemes capturing nonlinear effects of the data at the costs of a decreased interpretability of the projection: Unlike for linear counterparts such as principal component analysis, the relevance of the original feature dimensions for the NLDR projection is not clear. In this contribution we propose relevance learning schemes for NLDR which enable to judge the relevance of a feature dimension for the projection. This technique can be extended to a metric learning scheme which opens a way to imprint the information as provided by a given visualization on the data representation in the original feature space.

1 Introduction

Nonlinear dimensionality reduction (NLDR) has been pioneered in approaches such as Isomap, locally linear embedding, t-distributed stochastic neighbor embedding (t-SNE), or neighbor retrieval visualizer (NeRV), to name just a few popular techniques [8, 14, 16, 2]. Unlike linear counterparts such as principal component analysis (PCA), nonparametric projection schemes are capable of a reliable representation of dominant structural elements such as cluster formation. Still practitioners often prefer linear techniques over more flexible nonparametric methods, one of the reasons being their direct interpretability: for linear techniques, the relation between the original feature dimensions and the projections is explicit and the relevance of features for the visualization can be quantified by the size of the parameters in the linear mapping. In contrast, the original features are hidden in a NLDR projection and their relevance is not clear.

In this contribution, we propose relevance learning schemes for NLDR projections which enhance a given visualization by a ranking scheme indicating the relevance of the input features for the data projection. This approach is in line with recent techniques to enhance machine learning models by interpretable components [15, 9, 12]. Note that there exist approaches which simultaneously perform data analysis and feature selection [6, 10], but they cannot be utilized for giving insight into existing NLDR methods such as t-SNE. These methods are typically fully unsupervised. Conversely, there exist purely supervised feature selection techniques for classification [3]. In contrast, our proposal rather acts like a wrapper approach to add interpretable components to existing NLDR methods. We investigate two methods for relevance determination: on the one hand, we apply feature selection techniques to quality evaluation measures for NLDR as proposed in [7]. This enables a reliable ranking of the feature dimensions. On the other hand, based on a smooth quality evaluation for NLDR as proposed in [16], we develop a metric adaptation scheme which adjusts relevance terms in

*Funding by DFG (HA 2719/7-1) and by the CITEC center of excellence is acknowledged.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

the original feature space. Besides resulting in a similar qualitative ranking, this enables an explicit quantitative estimation of the feature relevance and a corresponding change of the feature representation in the original space as imprinted by the given visualization. Thus, this offers a first step towards techniques to interactively change the data representation based on a given data visualization.

2 Dimensionality Reduction

Dimensionality reduction (DR) maps data points $X = \{\vec{x}^i \in \mathbb{R}^n \mid i = 1 \dots N\}$ to projections $Y = \{\vec{y}^i \in \mathbb{R}^2 \mid i = 1 \dots N\}$ such that as much structure as possible is preserved. Techniques differ in the way how this is formalized, see e.g. [2] for a unifying presentation of popular DR schemes. Linear methods such as PCA offer an explicit mapping $\vec{y}^i = \mathbf{W}\vec{x}^i$ while many NLDR schemes are nonparametric. We will exemplarily consider t-SNE [14] which, as an objective, optimizes the Kullback Leibler divergence of probabilities as induced by data pairs in the original space and the visualization space, respectively. We will also consider an extension of t-SNE to a discriminative DR method, Fisher t-SNE (F-t-SNE)[5]. For the latter, data are labeled, and the euclidean metric for X is exchanged by the Fisher metric to account for the auxiliary class information.

Since DR is essentially unsupervised it is not clear how to quantitatively evaluate a given data visualization. The co-ranking framework as proposed in [7] offers a very popular evaluation scheme in dependence of a given neighborhood range k: given k, it evaluates the average overlap of neighborhoods of size k in the projection space and the original data space, i.e. it measures the quality

$$Q_k(X,Y) = \sum_i \left(N_k(\vec{x}^i) \cap N_k(\vec{y}^i)\right)/(Nk)$$

where $N_k(\vec{x}^i)$ (resp. $N_k(\vec{y}^i)$) are the indices of the k closest points of $\vec{x}^i$ in the data space (resp. projection space). Interestingly, the quality summarizes various popular alternative evaluation measures [7]. A principled alternative has been proposed in [16] based on an information retrieval model, which contains the quality as one special case. It also introduces a smooth extension

$$Q_k^{\mathrm{NeRV}}(X,Y) = \gamma \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} + (1-\gamma) \sum_i \sum_{j \neq i} q_{j|i} \log \frac{q_{j|i}}{p_{j|i}}$$

where

$$p_{j|i} = \frac{\exp(-d(\vec{x}^i, \vec{x}^j)^2/(\sigma_i^x(k))^2)}{\sum_{l \neq i} \exp(-d(\vec{x}^i, \vec{x}^l)^2/(\sigma_i^x(k))^2)} \;,\; q_{j|i} = \frac{\exp(-\|\vec{y}^i - \vec{y}^j\|^2/(\sigma_i^y(k))^2)}{\sum_{l \neq i} \exp(-\|\vec{y}^i - \vec{y}^l\|^2/(\sigma_i^y(k))^2)}$$

with weighting $\gamma \in [0,1]$, d referring to the distance in the data space X, and the standard deviation chosen such that the actual number of neighbors is k.[1] These costs have been used in [16] as objective for the NLDR technique NeRV.

[1] In [16], the standard deviations σ_i^x and σ_i^y are the same. For our purposes, we will consider a varying neighborhood size.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

3 Relevance Learning for DR

Note that NLDR techniques such as t-SNE provide a nonparametric mapping $\vec{x}^i$ to $\vec{y}^i$ for which an interpretation is not clear. In particular, it is not clear how relevant a given feature x^i_l is for the mapping. We are interested in ways to enhance NLDR by a relevance weighting for the features $l \in \{1, \ldots, n\}$ of X.

Evaluation functions for NLDR allows us to directly transfer classical feature selection techniques [3]: we can apply one forward or backward selection step regarding one feature for these evaluation functions. This yields our first two relevance determination techniques. Assume a NLDR $X \to Y$ is given.

- $\lambda^k_{\text{forward}}(l) := Q_k(X|_l, Y)$ where $X|_l$ considers only feature l, i.e. the points $(x^i_l) \in \mathbb{R}$. This induces an ascending relevance ranking of the features.
- $\lambda^k_{\text{backward}}(l) := Q_k(X|_{\neg l}, Y)$ where points $(x^i_1, \ldots, x^i_{l-1}, x^i_{l+1}, \ldots, x^i_n) \in \mathbb{R}^{n-1}$ are considered. This induces a relevance ranking in descending order.

These measures yield a qualitative evaluation of the relevance. For quantitative measures, we consider the smooth quality Q^{NeRV}_k. The idea is to change the metric in X such that it takes into account the relevance of the dimension l: $d(\vec{x}^i, \vec{x}^j)^2 = \sum_l (x^i_l - x^j_l)^2$ becomes $\sum_l \lambda^2_l (x^i_l - x^j_l)^2$. This corresponds to a feature transformation $X_\lambda = \{(\lambda_1 x^i_1, \ldots, \lambda_n x^i_n) \mid i\}$ of X. We are interested in relevance terms λ such that the transformed feature space X_λ is as close as possible to the projection Y as measured by quality evaluation measures:

- $\lambda^k_{\text{NeRV}}(l) := \lambda^2_l$ where λ optimizes $Q^{\text{NeRV}}_k(X_\lambda, Y) + \delta \sum_l \lambda^2_l$.

$\delta > 0$ weights the sparsity constraint. To compute $\lambda^k_{\text{NeRV}}(l)$, we optimize the objective L1 regularized quality $Q^{\text{NeRV}}_k(X_\lambda, Y) + \delta \sum \lambda^2_l$ with respect to λ^2_l. We use a gradient technique similar to well known algorithms from neural network optimization [11]. Strictly speaking, the result is not necessarily unique due to possible local optima; in practice, we did not observe problems.

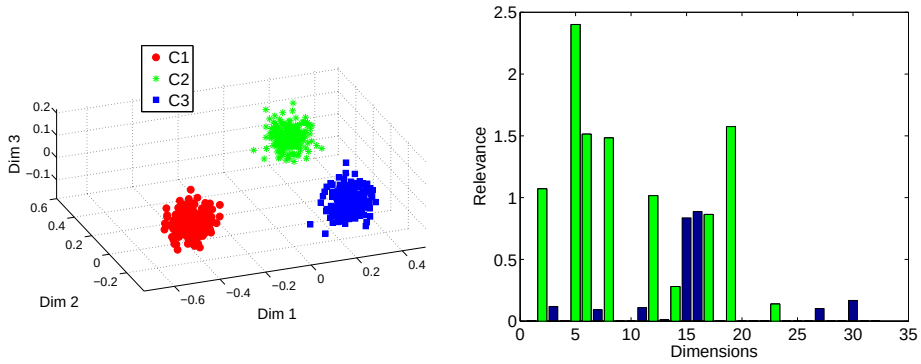

Fig. 1: Left: Data set1. Right: Relevance profile of the Adrenal data set. Green marks indicate that these 9 dimensions are also the top ones in [1].

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

4 Experiments

All relevance measures yield a ranking of the dimensions according to their relevance for the visualization at hand, but only λ_{NeRV} can also be employed as a metric for the original data. In the following, we i) demonstrate how the methods work for a simple toy scenario, ii) we compare the rankings of the dimensions qualitatively for different projection types and iii) we show that the metric induced by $\lambda^k_{\mathrm{NeRV}}$ improves the similarity of the original and projected data. We also compare one of our relevance profiles to one from the literature and observe a large accordance. In all experiments we set $\gamma = 0.5$ and $\delta = 1$.

Proof of Concept: Data *set1* contains three clusters with 20 points each in three dimensions, see Fig. 1. The third dimensions does not contain cluster information. A t-SNE projection yields well-separated clusters in two dimensions.

The relevance profiles λ are shown in Fig. 2 for varying k. The ranking of the dimensions induced by λ is identical for all techniques. λ_{NeRV} mirrors the irrelevance of dimension 3 as soon as the neighborhood exceeds the cluster size. Locally, all dimensions carry information because of the isotropic cluster shapes.

While the baseline for irrelevant dimensions for λ_{NeRV} is zero, the baseline for $\lambda_{\mathrm{forward}}$ is given by the diagonal, and the baseline for $\lambda_{\mathrm{backward}}$ depends on the data and is given by the quality of the projection. As can be seen from Fig. 2 (middle and right) also the forward and backward relevance selection methods clearly mark dimension 3 as unimportant. Unlike λ_{NeRV}, these relevance schemes do not indicate the local importance of all dimensions.

Qualitative comparison for different mapping characteristics: We compare the relevance ranks induced by the three schemes using different data and projection characteristics: Data *set2* contains three two-dimensional Gaussians arranged above each other along dimension 3. Although this dimension has small variance, it is relevant for cluster separation. Correspondingly, PCA and t-SNE lead to different map characteristics PCA ignoring dimension 3 while t-SNE emphasizes it. Data *set3* consists of ten features with three classes in the first two dimensions. The other dimensions contain increasingly noisy copies. Projections of PCA, t-SNE and F-t-SNE look similar here, with F-t-SNE better emphasizing the cluster structure which is mostly apparent in the first two dimensions.

We report the feature ranking of the most relevant features for different neighborhood sizes k (medium $\widehat{=} 0.8 \cdot$ cluster size, large $\widehat{=} 1.2 \cdot$ cluster size) in Table 1. The rankings induced by the different techniques coincide in most

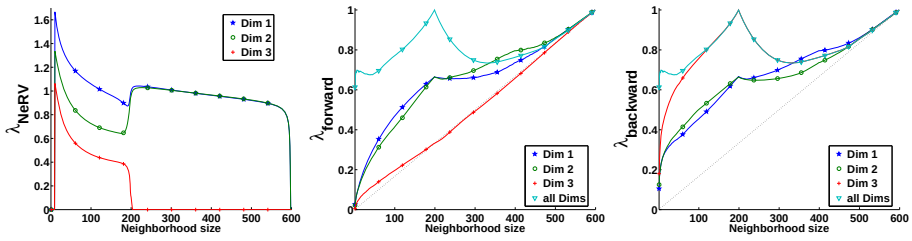

Fig. 2: Relevance determination for data set1 using λ_{NeRV} (left), $\lambda_{\mathrm{forward}}$ (middle) and $\lambda_{\mathrm{backward}}$ (right).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Table 1: Feature ranking induced by the different techniques for set2 and set3

neighb.	medium	large	medium	large	medium	large
set2	λ_{NeRV}		$\lambda_{\mathrm{forward}}$		$\lambda_{\mathrm{backward}}$	
PCA	$(1,2) \gg 3$	$(1,2) \gg 3$	$(1,2) \gg 3$	$(1,2) \gg 3$	$(1,2) \gg 3$	$(1,2) \gg 3$
t-SNE	$3 \gg (1,2)$	$3 \gg (1,2)$	$3 \gg (1,2)$	$3 \gg (1,2)$	$3 \gg (1,2)$	$3 \gg (1,2)$
set3	λ_{NeRV}		$\lambda_{\mathrm{forward}}$		$\lambda_{\mathrm{backward}}$	
PCA	$3 > 2 \gg 1$	$(2,3) \gg 1$	$1 > 3 > 2$	$1 > 3 > 2$	$2 > (1,3,4)$	$2 > (1,3,4)$
t-SNE	$1 > 2 \gg 9$	$1 \gg 2 \gg 9$	$1 > (2,3)$	$1 > (2,3)$	$2 > 9 > 3$	$2 > (1,3,9)$
F-t-SNE	$1 > 2 \gg 3$	$1 > 2 \gg 3$	$1 > (2,3)$	$1 > (2,3)$	$2 > (1,3,4)$	$2 > (1,3,4)$

settings and clearly indicate dimensions which one expect as relevant for the data set and projection as such, with a few notable exceptions: For set3 high redundancy is present. This causes a clearer emphasize of dimensions 1 and 2 for F-t-SNE. Further, this redundancy cannot be accounted for by forward or backward selection, while λ_{NeRV}, optimizing simultaneously for all features, breaks ties in favor of the less noisy features 1 and 2.

Suitability of induced feature transformation: Finally, we demonstrate the suitability of the metric induced by λ_{NeRV} to imprint the information of the projection Y to X. We evaluate this property by a comparison of the nearest neighbor (NN) error of the data in the projection space and the original data space X or its transformation, respectively. Thereby, we learn λ_{NeRV} based on a Fisher t-SNE mapping which also takes the available label information into account. We expect that the NN error improves in the latter setting for the transformed representation X_λ of X. Further, we expect that the classification is also improved if standard t-SNE is applied to the data X_λ.

We use two data sets: The *USPS* data set [4] contains images of size 16×16 of the handwritten digits 0 to 9 where we randomly select 200 images per class. The *Adrenal* data set [1] contains 147 patients characterized by 32 features. The data describe two different kinds of adrenal tumors.

The results using λ_{NeRV}, which is learned on the F-t-SNE mapping, are reported in Table 2. The classification error reduces, if X is projected to two dimensions using t-SNE because of the elimination of noise, and even more so if F-t-SNE is used, i.e. the class information directs what is considered as noise. Interestingly, the classification improves when transforming the data according to the learned relevance from λ_{NeRV}, albeit only a linear transformation of the data takes place this way. This behavior is also preserved if a standard t-SNE projection is used on top of the feature transformation. Hence the results substantiate the possibility to change the data representation based on visual information this way, albeit the method is still limited to a global linear weighting.

For the Adrenal data, we compare the relevance profile λ_{NeRV} with relevances from [1], obtained differently. Interestingly, there is a large overlap of these two results as shown in Fig. 1. Unlike [1] we can obtain this profile in one run of the algorithm making repetitions and thresholding as used in [1] superfluous.

Table 2: 1-NN errors in various data spaces of the data sets USPS and Adrenal.

neighb.				medium	large	medium	large
data sets	X	t-SNE(X)	F-t-SNE(X)	X_λ		t-SNE(X_λ)	
USPS	7.3%	6.7%	0.0%	2.2%	2.7%	3.1%	3.5%
Adrenal	10.9%	8.8%	0.7%	7.5%	6.8%	7.5%	6.8%

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

5 Discussion

We have presented relevance determination schemes for dimensionality reduction, thus offering a first step to shed some light on the interpretability of a given nonparametric data visualization. Interestingly, one technique also yields an explicit feature transformation such that it opens the way towards an interactive data transformation based on a visualization of a given data set only.

So far, feature ranking is global and we have restricted the learned metrics to a global diagonal form. Extensions to local schemes and more powerful quadratic forms are the subject of ongoing work. Further, improvements of the computational complexity using techniques as presented in [17, 13] are possible.

References

[1] M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In *ESANN'12*, pages 423–428. d-side publishing, 2012.

[2] K. Bunte, M. Biehl, and B. Hammer. A general framework for dimensionality reducing data visualization mapping. *Neural Computation*, 24(3):771–804, 2012.

[3] G. Doquire and M. Verleysen. Mutual information-based feature selection for multilabel classification. *Neurocomputing*, 122:148–155, 2013.

[4] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[5] A. Gisbrecht, A. Schulz, and B. Hammer. Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, accepted, 2013.

[6] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, Sept. 2004.

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

[8] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 2007.

[9] S. Lespinats and M. Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. *Comput. Graph. Forum*, 30(1):113–125, 2011.

[10] D. M. Maniyar and I. T. Nabney. Data visualization with simultaneous feature selection. In *CIBCB*, pages 1–8. IEEE, 2006.

[11] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 586–591. IEEE Press, 1993.

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

[13] L. van der Maaten. Barnes-Hut-SNE. *arXiv:1301.3342 , CoRR*, abs/1301.3342, 2013.

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

[15] A. Vellido, J. Martin-Guerrero, and P. Lisboa. Making machine learning models interpretable. In *ESANN'12*, 2012.

[16] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR-10*, 11:451–490, 2010.

[17] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In S. Dasgupta and D. Mcallester, editors, *Proceedings of the 30th ICML-13*, volume 28, pages 127–135. JMLR Workshop and Conference Proceedings, May 2013.