ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Out-of-Sample Kernel Extensions for Nonparametric Dimensionality Reduction

Andrej Gisbrecht[1], Wouter Lueks[2], Bassam Mokbel[1], Barbara Hammer[1]

1- University of Bielefeld - CITEC centre of excellence, Germany
2- University of Nijmegen - Faculty of Science, The Netherlands

Abstract. Nonparametric dimensionality reduction (DR) techniques such as locally linear embedding or t-distributed stochastic neighbor (t-SNE) embedding constitute standard tools to visualize high dimensional and complex data in the Euclidean plane. With increasing data volumes and streaming applications, it is often no longer possible to project all data points at once. Rather, out-of-sample extensions (OOS) derived from a small subset of all data points are used. In this contribution, we propose a kernel mapping for OOS in contrast to direct techniques based on the DR method. This can be trained based on a given example set, or it can be trained indirectly based on the cost function of the DR technique. Considering t-SNE as an example and several benchmarks, we show that a kernel mapping outperforms direct OOS as provided by t-SNE.

1 Introduction

In the last years, many nonlinear dimensionality reduction techniques have been developed which allow to project high-dimensional data points to the Euclidean plane preserving as much structure of the original data as possible, see e.g. [2, 8, 12]. This way, humans can directly inspect the data, thereby relying on their astonishing cognitive capabilities in visual perception. An increasing size and complexity of the data sets makes interactive techniques more and more common, where humans focus on relevant parts or add additional information on demand, see e.g. [7]. In these cases, the full data set is often not available a priori, moreover, very fast visualization of the selected information is required.

Some of the most powerful techniques such as t-distributed stochastic neighbor embedding (t-SNE) [12] scale quadratically with the size of the data, thus they are infeasible for large data sets. Due to this fact, visualization has to rely on subsamples to arrive at a mapping in reasonable time. Unlike classical techniques such as principal component analysis, however, many modern nonlinear dimensionality reduction techniques do not offer an explicit embedding function [8]. Rather, to extend cost function based DR methods to new data points, for example, requires the minimization of a cost function [2]. Such out-of-sample extensions are computational costly and their quality, i.e. the generalization ability of such projections to novel data points, is not clear at all.

Due to this fact, several approaches have been proposed to extend nonlinear DR to explicit mappings. The approach [1] proposes interpolation techniques for multidimensional scaling and the generative topographic mapping. The contribution [10] considers visualization by means of a kernel function optimized according to a quadratic cost function. In the approach [11] t-SNE is combined with deep feedforward networks to arrive at a highly nonlinear embedding function. Recently, a variety of well-known non-parametric nonlinear DR techniques has been put into a general framework based on the notion of cost functions [2]. Based on this formulation, an extension of the techniques to non-parametric out-of-sample extensions (OOS) by means of optimization as well as an extension to explicit OOS mappings by means of the integration of a parameterized function has been proposed.

In this contribution, we propose explicit kernel embeddings for OOS of DR techniques. Relying on t-SNE as an exemplary DR technique, we propose three different ways to train the mapping: an explicit solution can be derived from a simple interpolation given a fixed sample set of projected points. Alternatively, a kernel mapping can be obtained based on a sample set by incorporating regularization in the form of support vector machine regression. Finally, the parameters of the kernel function can be optimized according to the cost function of the DR method based on a sample set. We show that the proposed techniques arrive at reasonable kernel mappings which outperform direct OOS in most cases.

2 The t-Distributed Stochastic Neighbor Embedding and Out-of-Sample Extension

The t-distributed stochastic neighbor embedding (t-SNE) has been proposed in [12] as a highly flexible DR technique which tries to preserve probabilities as induced by pairwise distances in the data and projection space. Assume data points $\mathbf{x}_i \in \mathbb{R}^n$ are given. The Euclidean distance induces pairwise probabilities p_{ij} which are given by Gaussian probabilities. The data correspond to projections $\mathbf{y}_i \in \mathbb{R}^d$, where, typically $d = 2$, which also induce pairwise probabilities q_{ij} which are given by the student-t distribution of the distances of projected points. The t-SNE aims at finding projections $\mathbf{y}_i$ such that the difference of these probabilities as measured by the Kullback-Leibler divergence of these two probability distributions $\mathrm{KL}(P||Q)$ is minimized, whereby typically a gradient technique is used for optimization, see [12] for details. Obviously, this technique does not provide an explicit DR mapping $\mathbf{x} \to \mathbf{y} = \mathbf{y}(\mathbf{x})$. However, a direct OOS extension to new points is possible provided a fixed set of data $\{(\mathbf{x}_i, \mathbf{y}_i) \,|\, i = 1, \dots, N\}$ has been trained. These projections are fixed, and a new data point $\mathbf{x}$ is mapped to the coefficient vector $\mathbf{y}$ such that the costs $\mathrm{KL}(P||Q)$ are minimized where P and Q include the novel pair $(\mathbf{x}, \mathbf{y})$. A gradient method yields coefficients $\mathbf{y}$ as a local optimum of these costs. See [2] for an explicit formalization of the resulting formulas. We refer to this method by direct OOS (refer to as *oos* in Tab. 1) in the following.

3 Kernel t-Distributed Stochastic Neighbor Embedding

Instead of a direct OOS, we propose an explicit OOS mapping given by a kernel function:

$$\mathbf{x} \mapsto \mathbf{y}(\mathbf{x}) = \sum_l \alpha_l k(\mathbf{x}_i, \mathbf{x})$$

where $\alpha_l \in \mathbb{R}^d$ and $\mathbf{x}_i$ is a fixed sample of data points. k is an appropriate kernel function such as the Gaussian kernel (refer to as *rbf*) or the recently proposed 'nearly' parameterless ELM kernel (refer to as *elm*) [5]. The parameters α_l have to be determined based on a given training sample $\mathbf{x}_i$ of points. We propose three different techniques to determine these mapping parameters:

Mean square interpolation (refer to as *itp*): First, the sample points $\mathbf{x}_i$ are mapped to projections $\mathbf{y}_i$ by means of t-SNE. Then, mapping parameters α_l are determined to minimize the sum squared error of these projections $\mathbf{y}_i$ and the function values $\mathbf{y}(\mathbf{x}_i)$. This allows an explicit solution of the matrix $\mathbf{A}$ of parameters α_l as

$$\mathbf{A} = \mathbf{K}^{-1}\mathbf{y}$$

where $\mathbf{K}$ is the Gram matrix with entries $k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{Y}$ denotes the matrix of projections $\mathbf{y}_i$, and $\mathbf{K}^{-1}$ refers to the pseudo-inverse. Note that fast iterative alternatives would be possible.

Regularized regression (refer to as *svm*): Similar to mean square interpolation, sample points $\mathbf{x}_i$ are mapped to projections $\mathbf{y}_i$ by means of t-SNE. Mapping parameters α_l are determined by approxmating the regression task provided by the set $\{(\mathbf{x}_i, \mathbf{y}_i) \,|\, i = 1, \ldots, N\}$ with a standard SVM (we use an SMO implementation provided by [3] and the standard ϵ-tube.)

Kernel mapping by means of the t-SNE cost function (refer to as *k*): We optimize the parameters α_l such that the projections $\mathbf{y}(\mathbf{x}_i)$ of the given sample set are optimum in the sense of the t-SNE cost function. Hence we consider the costs $E := \mathrm{KL}(P||Q)$ where Q is determined based on the projections $\mathbf{y}(\mathbf{x}_i)$ provided by the kernel mapping. Optimization can be performed by gradient techniques where the gradient with respect to α_l is

$$\frac{\partial E}{\partial \alpha_l} = 4 \sum_i \sum_j (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j)k(\mathbf{x}_i, \mathbf{x}_l).$$

See [2] for a more detailed description of the general technique.

4 Experiments

We test the techniques using four benchmark sets with different characteristics:

car: The car data from the UCI repository [4] comprises 1728 points described by six discrete features with ranges in $1, \ldots k$ (k is up to 6) such as purchase costs, maintenance costs, etc.

coil: The COIL-20 data set [12] consists of 72 gray scaled images (128x128 pixels) of 20 objects taken from different angels.

olivetti: The Olivetti faces data set from AT&T [12] consists of 400 images (64x64 pixels) of 40 different persons taking under different conditions.

henon: The Henon data set consists of 5000 data points in three dimensions forming a chaotic attractor as described by the corresponding dynamical equations which constitute a generalization of the popular Henon attractor to three dimensions [6].

We compare the result of the different techniques when training on different sized subsets of the data for OOS extensions. The width of the Gaussian kernel is taken as mean distance of a data point to its fifth closest neighbor.

Evaluation of the projection is done by means of the local quality measure as proposed in the approach [9]. The results of a repeated ten-fold cross-validation with five repeats are reported in Tab. 1. A corresponding quality graph is depicted in Fig. 1 for one example.

Interestingly, the direct out-of-sample extension provided by t-SNE is inferior in all but three cases. Hence, in addition to the direct availability of an explicit mapping, kernel dimensionality reduction mapping seems to improve the projection quality. A simple interpolation on a given training set is surprisingly good, the same holds for the training of a kernel mapping by means of a direct integration into the t-SNE cost function. In contrast, SVM approximation yields good results in only three out of twelve cases. This picture is also confirmed when inspecting the full quality graph as compared to the single value provided by the local quality, as exemplarily shown in Fig. 1, whereby the precise ordering is a bit more diverse depending on which neighborhood size k is considered.

5 Discussion

We have proposed a method to obtain an explicit nonlinear kernel mapping for dimensionality reduction based on t-SNE. Interestingly, including a kernel mapping

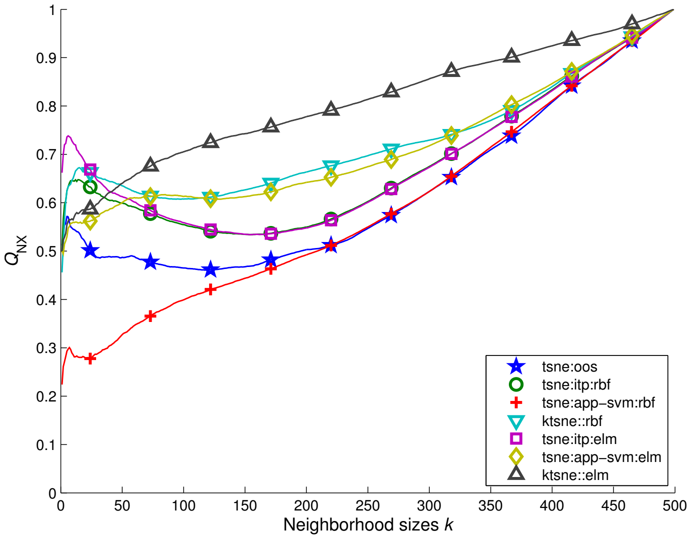

Fig. 1: Quality graph for the Henon data on 90% training data.

can improve upon direct out-of-sample extension in almost all cases. Further, a very simple strategy to train the map by simple least squares minimization on a given training set gives surprisingly good results. Naturally, the techniques can be extended in similar form to alternative nonlinear non-parametric dimensionality reduction techniques such as locally linear embedding or Isomap [8]. Case studies for different approaches will be the subject of future work.

Alternative techniques to extend nonlinear dimensionality reduction to large data sets include, among others, techniques which naturally provide an embedding mapping (such as [13, 14]), which are based on landmark techniques [12], or approximations [15]. Unlike our approach, these techniques have been proposed in combination with specific non-linear dimensionality reduction methods only.

References

[1] Bae, S.-H., Choi, J. Y., Qiu, J., and Fox, G. C. (2010). Dimension reduction and visualization of large high-dimensional data via interpolation. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, pages 203–214, New York, NY, USA. ACM.

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

[3] Chih-Chung Chang, Chih-Jen Lin (2001), *LIBSVM: a library for support vector machines*, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[4] A. Frank, A. Asuncion (2010), UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, available at `http://archive.ics.uci.edu/ml/`

[5] Freénay B, Verleysen M (2011). Parameter-insensitive kernel in extreme learning for nonlinear support vector regression. Neurocomputing (in press).

[6] Gonchenko, S.V., Ovsyannikov, I.I., Simo, C., Turaev, D. (2005), Three-dimensional Henon-like maps and wild Lorenz-like attractors, Technical Report MP-ARC-2005-111, University of Texas.

[7] Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). Visual analytics: Scope and challenges. In Simoff, S., Boehlen, M. H., and Mazeika, A., editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics.* Springer. Lecture Notes in Computer Science (LNCS).

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

[9] John Aldo Lee, Michel Verleysen (2010). *Scale-independent quality criteria for dimensionality reduction.* Pattern Recognition Letters 31(14): 2248-2257.

[10] Suykens, J. A. K. (2008). Data visualization and dimensionality reduction using kernel maps with a reference point. *Neural Networks, IEEE Transactions on*, 19(9):1501 –1517.

[11] van der Maaten, L. J. P. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AI-STATS)*, number 5, pages 384–391. JMLR W&CP.

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

[13] C. Bishop, M. Svensen, and C. Williams (1998). The generative topographic map. *Neural Computation* 10(1):215-234.

[14] Lowe, D. and Tipping, M. E. (1997). NeuroScale: Novel Topographic Feature Extraction using RBF Networks. *Advances in Neural Information Processing Systems*, 9:543-549.

[15] Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M. (2003). Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *Advances in Neural Information Processing Systems*, 16:177-184.

data set	car			coil			olivetti			henon		
% used for training	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
oos	0.18 (0.02)	0.29 (0.02)	0.35 (0.05)	0.24 (0.06)	**0.62** (0.02)	**0.65** (0.02)	0.19 (0.10)	**0.44** (0.05)	0.25 (0.03)	0.49 (0.03)	0.58 (0.02)	0.56 (0.03)
itp(rbf)	**0.30** (0.01)	**0.34** (0.02)	**0.39** (0.02)	0.44 (0.04)	**0.58** (0.01)	**0.61** (0.01)	0.29 (0.08)	**0.41** (0.03)	**0.41** (0.03)	0.72 (0.02)	0.62 (0.02)	0.59 (0.02)
svm(rbf)	0.28 (0.01)	0.31 (0.00)	0.32 (0.03)	0.43 (0.02)	0.51 (0.01)	0.55 (0.01)	0.31 (0.02)	0.40 (0.03)	0.37 (0.04)	0.57 (0.01)	0.35 (0.04)	0.29 (0.02)
k(rbf)	0.28 (0.02)	0.26 (0.01)	**0.40** (0.02)	**0.47** (0.01)	0.57 (0.02)	0.57 (0.03)	**0.37** (0.03)	**0.41** (0.03)	0.35 (0.05)	0.71 (0.04)	0.62 (0.02)	**0.60** (0.01)
itp(elm)	**0.30** (0.02)	**0.37** (0.01)	0.37 (0.03)	0.42 (0.05)	0.55 (0.01)	0.59 (0.01)	0.27 (0.08)	**0.41** (0.02)	**0.40** (0.05)	**0.80** (0.02)	**0.72** (0.02)	**0.70** (0.02)
svm(elm)	0.28 (0.02)	0.31 (0.01)	0.31 (0.01)	0.43 (0.03)	0.48 (0.01)	0.51 (0.02)	0.31 (0.02)	**0.41** (0.02)	0.39 (0.05)	**0.80** (0.02)	**0.63** (0.06)	0.55 (0.03)
k(elm)	0.26 (0.02)	0.28 (0.01)	0.24 (0.03)	**0.46** (0.01)	0.44 (0.03)	0.42 (0.02)	**0.40** (0.02)	0.37 (0.04)	0.37 (0.04)	0.50 (0.10)	0.58 (0.05)	0.58 (0.06)

Table 1: Local quality obtained by the different OOS techniques in a cross-validation for different training set sizes, the result is evaluated by means of a cross-validation, the standard deviation is given in parenthesis. For each setting, the best two results are depicted in boldface.