Applications of Discriminative Dimensionality Reduction

Barbara Hammer, Andrej Gisbrecht and Alexander Schulz
CITEC Centre of Excellence, Bielefeld University, Bielefeld, Germany

Keywords: Dimensionality Reduction, Fisher Information Metric, Classifier Visualization, Evaluation.

Abstract: Discriminative nonlinear dimensionality reduction aims at a visualization of a given set of data such that the information contained in the data points which is of particular relevance for a given class labeling is displayed. We link this task to an integration of the Fisher information, and we discuss its difference from supervised classification. We present two potential application areas: speed-up of unsupervised nonlinear visualization by integration of prior knowledge, and visualization of a given classifier such as an SVM in low dimensions.

1 INTRODUCTION

Caused by a rapid digitalization of almost all areas of daily life, data sets and learning scenarios are increasing dramatically with respect to both, size and complexity. This fact poses new challenges for standard data analysis tools: on the one hand, methods have to deal with very large data sets such that many algorithms rely on sampling or approximation techniques to maintain feasibility (Bekkerman et al., 2011; Tsang et al., 2005). Hence valid results have to be guaranteed based on a small subset of the full data only. On the other hand, an exact objective is often not clear a priori; rather, the user specifies her interests and demands interactively when applying data mining techniques and inspecting the results (Ward et al., 2010). This places the human into the loop, causing the need for intuitive interfaces to the machine learning scenarios (Vellido et al., 2012; Rüping, 2006). In turn, this demand causes an additional need for fast and online machine learning technology since the user is usually not willing to wait for more than a few seconds until she gets (at least preliminary) results.

The visual system constitutes one of our most advanced senses, and humans display astonishing cognitive capabilities as concerns vision such as grouping of objects or instantaneous recognition of artifacts in visual scenes. In consequence, visualization plays an essential part in the context of interactive machine learning. This causes the need for reliable, fast and online visualization techniques of data and machine learning results when training on the given data.

Dimensionality reduction refers to the specific task to map high dimensional data points into low dimensions such that data can directly be displayed on the screen while as much information as possible is preserved. Classical techniques such as a simple principle component analysis (PCA) offer a linear projection only, thus their flexibility is limited. Nevertheless, they are widely used today due to their excellent generalization ability and scalability.

In recent years, a large variety of nonlinear alternatives has been proposed, formalizing the ill-posed objective of what means 'structure preservation' via different mathematical objectives. Popular examples include techniques such as maximum variance unfolding, non-parametric embedding, Isomap, locally linear embedding (LLE), stochastic neighbor embedding (SNE), and similar, see e.g. the overviews (Bunte et al., 2012a; Lee and Verleysen, 2007; Maaten and Hinton, 2008). These techniques, however, have several drawbacks such that many practitioners still rely on simpler linear techniques such as PCA (Biehl et al., 2011): many nonlinear techniques provide a mapping of the given data points only, requiring additional effort for out-of-sample extensions. Due to the inherent ill-posedness of dimensionality reduction, the results are not easily interpretable by humans and first formal evaluation measures for dimensionality reduction have just recently been proposed (Lee and Verleysen, 2010). Further, most techniques depend on pairwise distances of data such that they scale at least quadratically with the data set size, making the techniques infeasible for large data sets.

In this contribution, we consider a specific variant of dimensionality reduction: discriminative dimensionality reduction, i.e. the case where data are accompanied by additional labeling. In this setting, the

Hammer B., Gisbrecht A. and Schulz A. (2013).
Applications of Discriminative Dimensionality Reduction.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 33-41
DOI: 10.5220/0004245300330041

goal is to visualize those aspects of the data which are of particular relevance for the given labeling. A few approaches have been proposed in this context: classical Fisher's linear discriminant analysis (LDA) projects data such that within class distances are minimized while between class distances are maximized, still relying on a linear mapping. The objective of partial least squares regression (PLS) is to maximize the covariance of the projected data and the given auxiliary information. It is also suited if data dimensionality is larger than the number of data points. Informed projections (Cohn, 2003) extend PCA to minimize the sum squared error and the mean value of given classes, this way achieving a compromise of dimensionality reduction and clustering. In (Goldberger et al., 2004), the metric is adapted according to auxiliary class information prior to projection to yield a global linear matrix transform. Further, interesting extensions of multidimensional scaling to incorporate class information have recently been proposed (Witten and Tibshirani, 2011). Modern techniques extend these settings to general nonlinear projections of data. One way is offered by kernelization such as kernel LDA (Ma et al., 2007; Baudat and Anouar, 2000; Mika et al., 1999). Another principled way to extend dimensionality reducing data visualization to auxiliary information is offered by an adaptation of the underlying metric. The principle of learning metrics has been introduced in (Kaski et al., 2001; Peltonen et al., 2004): the standard Riemanian metric is substituted by a form which measures the information of the data for the given classification task (Kaski et al., 2001; Peltonen et al., 2004; Venna et al., 2010). A slightly different approach is taken in (Geng et al., 2005), relying on an ad hoc adaptation of the metric. Metric adaptation based on the classification margin and subsequent visualization has been proposed in (Bunte et al., 2012b), for example. Alternative approaches to incorporate auxiliary information modify the cost function of dimensionality reducing data visualization. The approaches introduced in (Iwata et al., 2007; Memisevic and Hinton, 2005) can both be understood as extensions of SNE. Multiple relational embedding (MRE) incorporates several dissimilarity structures in the data space induced by labeling, for example, into one latent space representation. Colored MVU incorporates auxiliary information into MVU by substituting the raw data by the combination of the data and the covariance matrix induced by the given auxiliary information.

What are the differences of a supervised visualization as compared to a direct classification of the data, i.e. a simple projection of the data points to their corresponding class labels? What are potential applications of such techniques? These questions are in the focus of this contribution. We will argue that auxiliary information in the form of class labeling can play a crucial role when addressing dimensionality reduction: on the one hand, it offers a natural way to shape the inherently ill-posed problem of dimensionality reduction by explicitly specifying which aspects of the data are relevant and, in consequence, which aspects should be emphasized – those aspects of the data which are relevant for the given auxiliary class labeling. In addition, the integration of auxiliary information can help to solve the problem of the computational complexity of dimensionality reduction. In this contribution, we will show that discriminative dimensionality reduction can be used to infer a mapping of points based on a small subsample of data only, thus reducing the complexity by an order of magnitude. We will use this technique in a general framework which allows us to visualize not only a given labeled data set, rather full classification models can be displayed this way, as we will demonstrate for the case of SVM classifiers.

Now we will first introduce the Fisher metric as a general way to include auxiliary class labels into a non-linear dimensionality reduction technique. We show the difference of the result from a direct classification in the context of discriminative t-SNE. Afterwards, we address two applications of this setting: integration of auxiliary information into kernel t-SNE mapping to obtain valid results from a small subset of data only, and vizualisation of a given SVM classifier.

2 SUPERVISED VISUALIZATION BASED ON THE FISHER INFORMATION

In the following, we will consider only one prototypical dimensionality reduction technique and emphasize the role of *discriminative* visualization rather than a comparison of the underlying dimensionality reduction technique: we restrict to t-distributed stochastic neighbor embedding (t-SNE), which constitutes one of the most successful nonlinear dimensionality reduction techniques used today (Maaten and Hinton, 2008). All arguments as given below could also be based on alternatives such as LLE or Isomap.

Given a set of data points $\mathbf{x}_i$ in some high-dimensional data space X, t-SNE finds projections $\mathbf{y}_i$ for these points in the two dimensional plane $Y = \mathbb{R}^2$ such that the probabilities of data pairs in the original space and the projection space are preserved as much

as possible. More precisely, probabilities in the original space are defined as $p_{ij} = (p_{(i|j)} + p_{(j|i)})/(2N)$ where N is the number of data points and

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}.$$

depends on the pairwise distance; σ_i is automatically determined by the method such that the effective number of neighbors coincides with a priorly specified parameter, the perplexity. In the projection space, probabilities are induced by the student-t distribution rather than Gaussians

$$q_{ij} = \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}.$$

to avoid the crowding problem by means of using a long tail distribution. The goal is to find projection points $\mathbf{y}_i$ such that the difference between p_{ij} and q_{ij} becomes small as measured by the Kullback-Leibler divergence. Usually, a gradient based optimization technique is used to minimize these costs.

As mentioned already above, the goal of dimensionality reduction is inherently ill-posed: in general, there does not exist a loss-free representation of data in two-dimensions, such that information loss is inevitable. Thereby, it depends on the users need which type of information is relevant for the application. A chosen dimensionality reduction technique implicitly specifies which type of information is preserved by means of specifying an abstract mathematical objective which is optimized while mapping. Such an abstract cost function, however, is hardly accessible by a user, and it cannot easily be altered according to the users needs. Due to this fact, it has been proposed e.g. in (Kaski et al., 2001; Peltonen et al., 2004; Venna et al., 2010) to enhance data by auxiliary information specified by the user which should be taken into account while projecting. Formally, we assume that every data point $\mathbf{x}_i$ is equipped with a class label c_i which are instances of a finite number of possible classes c. Now projection points $\mathbf{y}_i$ should be found such that the aspects of $\mathbf{x}_i$ which are relevant for c_i are displayed.

How can this be realized? A Riemannian manifold can easily be defined which is based on the information of $\mathbf{x}_i$ for the class labels as metric tensor. The tangent space at $\mathbf{x}_i$ is equipped with the quadratic form

$$d_{\mathbf{x}_i}(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{J}(\mathbf{x}_i)\mathbf{y}$$

where $\mathbf{J}(\mathbf{x})$ denotes the Fisher information matrix

$$\mathbf{J}(\mathbf{x}) = E_{p(c|\mathbf{x})}\left\{\left(\frac{\partial}{\partial\mathbf{x}}\log p(c|\mathbf{x})\right)\left(\frac{\partial}{\partial\mathbf{x}}\log p(c|\mathbf{x})\right)^T\right\}.$$

A Riemannian metric is induced by minimum path integrals using this quadratic form locally, i.e.

$$d(\mathbf{x},\mathbf{y}) = \inf_p \int_0^1 \sqrt{d_{p(t)}(p',p')}dt$$

where $p : [0,1] \to X$ ranges over all smooth curves from $p(0) = \mathbf{x}$ to $p(1) = \mathbf{y}$ in X. We refer to this metric as the Fisher metric in the following. Thus, auxiliary information can be integrated into t-SNE or any other dimensionality reduction technique which relies on distances by substituting the Euclidean metric by the Fisher metric.

In how far is this technique different from a simple classification of data, i.e. in how far does a projection carry more information than a simple projection of the data to their distinct class labels? A very simple example as shown in Fig. 1 illustrates the difference: Three classes which consist of two clusters each are generated in two dimensions. Thereby, the classes of two modes overlap (see arrow). We measure pairwise distances of these data using the Fisher metric. These values are displayed using metric multidimensional scaling. As can be seen, the following effects occur:

- the distance of data within a single mode belonging to one class becomes smaller by scaling dimensions which are unimportant for a given labeling at a smaller scale. Thus, data points in one clearly separated mode have the tendency to be mapped on top of each other, and these cluster structures become more apparent.
- the number of modes of the classes is preserved, emphasizing the overall structure of the class distribution in space – unlike a simple mapping of data to class labels which would map all modes of one class on top of each other.
- overlapping classes are displayed as such (see arrow) and directions which cause this conflict are preserved since they have an influence on the class labeling. In contrast, a direct mapping of such data to their class labels (if possible) would resolve such conflicts in the data.

In practice, the Fisher distance has to be estimated based on the given data only. The conditional probabilities $p(c|\mathbf{x})$ can be estimated from the data using the Parzen nonparametric estimator

$$\hat{p}(c|\mathbf{x}) = \frac{\sum_i \delta_{c=c_i}\exp(-\|\mathbf{x}-\mathbf{x}_i\|^2/2\sigma^2)}{\sum_j \exp(-\|\mathbf{x}-\mathbf{x}_j\|^2/2\sigma^2)}.$$

The Fisher information matrix becomes

$$\mathbf{J}(\mathbf{x}) = \frac{1}{\sigma^4}E_{\hat{p}(c|\mathbf{x})}\left\{\mathbf{b}(\mathbf{x},c)\mathbf{b}(\mathbf{x},c)^T\right\}$$

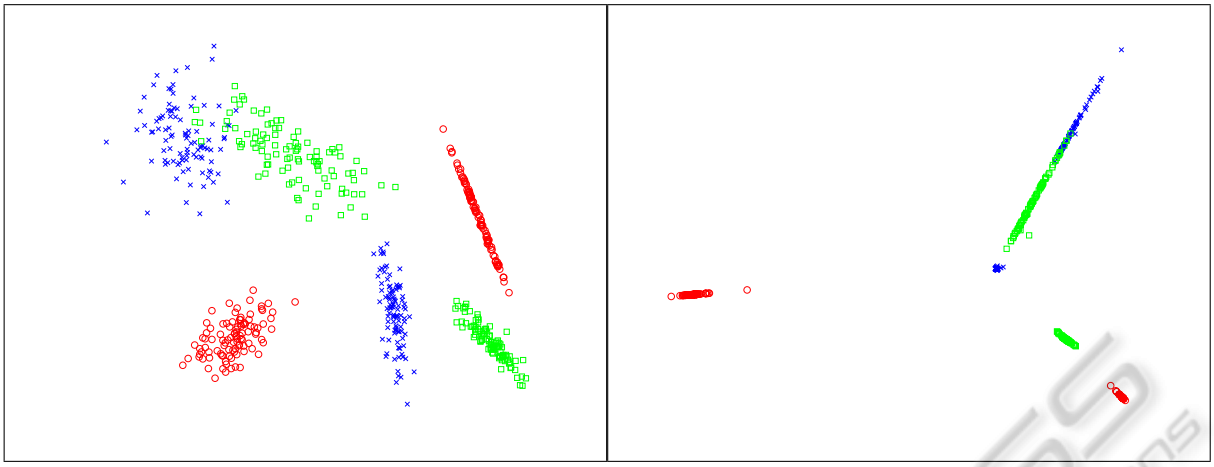

Figure 1: A simple example which demonstrates important properties of the Fisher Riemannian tensor: multi-modality as well as class overlaps are preserved. The original data are displayed at the left, a plot of the data equipped with the Fisher metric displayed using metric multidimensional scaling is shown on the right, the arrows point to regions of overlap of the classes, which are preserved by the metric.

where

$$\mathbf{b}(\mathbf{x},c) = E_{\xi(i|\mathbf{x},c)}\{\mathbf{x}_i\} - E_{\xi(i|\mathbf{x})}\{\mathbf{x}_i\}$$

$$\xi(i|\mathbf{x},c) = \frac{\delta_{c,c_i}\exp(-\|\mathbf{x}-\mathbf{x}_i\|^2/2\sigma^2)}{\sum_j \delta_{c,c_j}\exp(-\|\mathbf{x}-\mathbf{x}_j\|^2/2\sigma^2)}$$

$$\xi(i|\mathbf{x}) = \frac{\exp(-\|\mathbf{x}-\mathbf{x}_i\|^2/2\sigma^2)}{\sum_j \exp(-\|\mathbf{x}-\mathbf{x}_j\|^2/2\sigma^2)}$$

E denotes the empirical expectation, i.e. weighted sums with weights depicted in the subscripts. If large data sets or out-of-sample extensions are dealt with, a subset of the data only is usually sufficient for the estimation of $\mathbf{J}(\mathbf{x})$.

There exist different ways to approximate the path integrals based on the Fisher matrix as discussed in (Peltonen et al., 2004). An efficient way which preserves locally relevant information is offered by T-approximations: T equidistant points on the line from $\mathbf{x}_i$ to $\mathbf{x}_j$ are sampled, and the Riemannian distance on the manifold is approximated by $d_T(\mathbf{x}_i,\mathbf{x}_j) =$

$$\sum_{t=1}^{T} d_1\left(\mathbf{x}_i + \frac{t-1}{T}(\mathbf{x}_j-\mathbf{x}_i), \mathbf{x}_i + \frac{t}{T}(\mathbf{x}_j-\mathbf{x}_i)\right)$$

where $d_1(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i-\mathbf{x}_j)^T J(\mathbf{x}_i)(\mathbf{x}_i-\mathbf{x}_j)$ is the standard distance as evaluated in the tangent space of $\mathbf{x}_i$. Locally, this approximation gives good results such that a faithful dimensionality reduction of data can be based thereon.

Now the question occurs what are benefits of an integration of such knowledge. Here we present two potential applications. Thereby, we restrict to one typical real-life benchmark data set, the USPS data, only due to space limitations, results for alternative benchmarks being similar.

3 APPLICATION (I): TRAINING A VISUALIZATION MAPPING

Similar to many other nonlinear projection techniques, t-SNE has the severe drawback that it scales quadratically with the size of the training set making it infeasible for large data sets. In addition, it does not provide an explicit mapping of the points; rather, out-of-sample extensions have to be implemented by means of an additional optimization. Because of this fact, it has been proposed in (Gisbrecht et al., 2013) to extend t-SNE towards a mapping in the following way: an explicit functional form is defined as

$$\mathbf{x} \mapsto \mathbf{y}(\mathbf{x}) = \sum_j \alpha_j \cdot \frac{k(\mathbf{x},\mathbf{x}_j)}{\sum_l k(\mathbf{x},\mathbf{x}_l)}$$

where $\alpha_j \in Y$ are points in the projection space and the points $\mathbf{x}_j$ are taken from a fixed sample of data points used to train the mapping. k is the Gaussian kernel. This mapping is parameterized by α_j. Due to its form as a generalized linear mapping, these parameters can analytically be determined as the least squares solution of an exemplary set of points $\mathbf{x}_i$ and projections $\mathbf{y}_i$ obtained by standard t-SNE (or any other dimensionality reduction technique). Then the matrix $\mathbf{A}$ of parameters α_j is given by

$$\mathbf{A} = \mathbf{Y}\cdot\mathbf{K}^{-1}$$

where $\mathbf{K}$ is the normalized matrix with entries $k(\mathbf{x}_i,\mathbf{x}_j)/\sum_j k(\mathbf{x}_i,\mathbf{x}_j)$. $\mathbf{Y}$ denotes the matrix of projections $\mathbf{y}_i$, and $\mathbf{K}^{-1}$ refers to the pseudo-inverse.

This technology, referred to as kernel t-SNE, has the benefit that training can be done on a small sub-

set of data only, extending the mapping to the full data set by means of the explicit mapping prescription. Thus, a considerable speed up can be obtained, provided a small subsample of points is sufficient to train the mapping. However, here occurs a problem: often, the structure of the data such as clusters is not yet pronounced if only a small sample of data is used for training kernel t-SNE. In consequence, kernel t-SNE when trained on a subsample does not clearly emphasize an underlying class structure as compared to t-SNE when trained on the full data set.

Here, discriminative dimensionality reduction offers a possibility to substitute the loss of information due to a small training set by prior information as given by an explicit class labeling. On the one hand, it is possible to generate the training set of points $\mathbf{x}_i$ and its projections $\mathbf{y}_i$ for kernel t-SNE based on the Fisher metric provided class labeling c_i is available. In addition, kernel t-SNE can be extended to a discriminative mapping by using the Fisher metric also in the kernel mapping prescription $k(\mathbf{x}, \mathbf{x}_j)$.

Fig. 2 and Fig. 3 show example mappings of the USPS data set consisting of 11.000 points with 256 dimensions representing handwritten digits from 0 to 9 (Hastie et al., 2001). For training and the representation of the kernel mapping, 10% of the data are used. For the estimation of the Fisher information, 1% of the data are used. Clearly, the original kernel t-SNE mapping does not contain enough information to emphasize the cluster structure when trained on 10% of the data only, while t-SNE when trained on the full data set clearly displays the classes, as can be seen e.g. in (Maaten and Hinton, 2008). The resulting kernel t-SNE mapping and its out of sample extension are displayed in Fig. 2. In contrast, the cluster structure is clearly visible if auxiliary information is taken into account, Fisher kernel t-SNE and its extension to the full data set being displayed in Fig. 3.

4 APPLICATION (II): VISUALIZATION OF CLASSIFIERS

Classification constitutes one of the standard tasks in data analysis. At present, the major way to display the result of a classifier and to judge its suitability is by means of the classification accuracy. Visualization is used in only a few places when inspecting a classifier: If data live in a low dimensional space, a direct visualization of the data points and classification boundaries in 2D or 3D can be done. For high dimensional data, which constitutes the standard case, a direct visualization of the classifier is not possible. One line of research addresses visualization techniques to accompany the accuracy by an intuitive interface to set certain parameters of the classification procedure, such as e.g. ROC curves to set the desired specificity, or more general interfaces to optimize parameters connected to the accuracy (Hernandez-Orallo et al., 2011). Surprisingly, there exists relatively little work to visualize the underlying classifier itself for high dimensional settings. For the popular support vector machine, for examples, only some specific approaches have been proposed: one possibility is to let the user decide an appropriate linear projection dimension by means of tour methods (Caragea et al., 2008). As an alternative, some techniques rely on the distance of the data points to the class boundary and present this information using e.g. nomograms (Jakulin et al., 2005) or by using linear projection techniques on top of this distance (Poulet, 2005). A few nonlinear techniques exist such as SVMV (Wang et al., 2006), which visualizes the given data by means of a self-organizing map and displays the class boundaries by means of sampling. Further, very interesting nonlinear dimensionality reduction, albeit not for the primary aim of classifier visualization, has been introduced in (Braun et al., 2008). These techniques offer first steps to visually inspect an SVM solution such that the user can judge e.g. remaining error regions, the modes of the given classes, outliers, or the smoothness of the separation boundary based on a visual impression.

However, so far, these techniques are often only linear, they require additional parameters, and they provide combinations of a very specific classifier such as SVM and a specific visualization technique. Discriminative dimensionality reduction constitutes an important technique based on which a given classifier can be visualized. Here, we propose a principled alternative based on discriminative t-SNE with the Fisher metric. We assume a classification mapping $f : X \to \{1, \ldots, c\}$ is present, which can be given by a support vector machine, for example. This mapping has been trained using some points $\mathbf{x}_i$ and their label c_i. We assume that the label prediction $f(\mathbf{x}_i)$ of a point $\mathbf{x}_i$ can be accompanied by a real value $r(\mathbf{x}_i) \in \mathbb{R}$ which indicates the (signed) strength of class-membership association. This can be given by the class probability or the distance from the decision boundary, for example. Now the task is to map the data points $\mathbf{x}_i$ as well as the classification boundary induced by f to two dimensions.

A very simple approach consists in a sampling of the original space X and a projection of these data $\mathbf{x}$ colored by class labels $f(\mathbf{x})$ using a standard di-

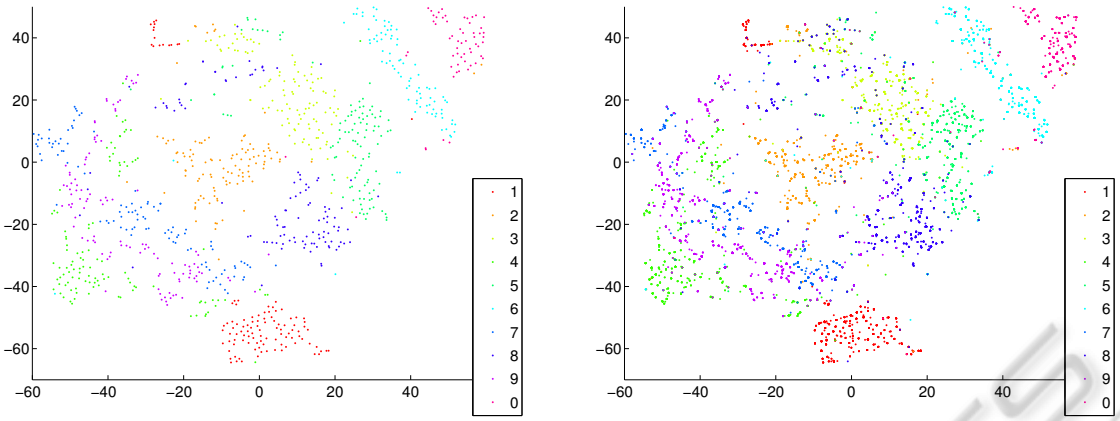

Figure 2: Visualization of the USPS data set using kernel t-SNE for the training set (top) and out of sample extension (bottom).

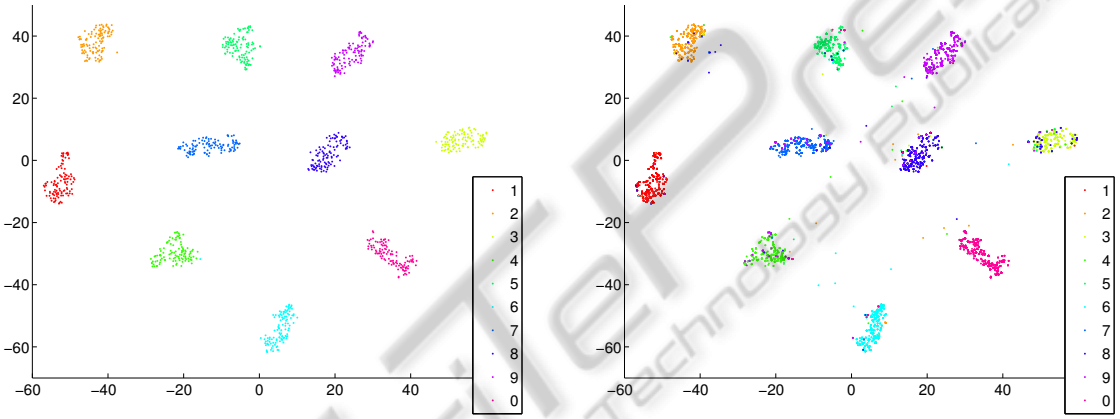

Figure 3: Visualization of the USPS data set using discriminative Fisher kernel t-SNE for the training set (left) and out of sample extension (right). Fisher kernel t-SNE provides clear class structures on these data unlike simple kernel t-SNE.

mensionality reduction technique. Since smooth values $r(\mathbf{x})$ are present, isobars corresponding to the classifier can then be displayed in the plane. This naive approach encounters two problems: (i) sampling the original data space X is infeasible due to a usually high dimensionality and (ii) projecting exhaustive samples from high dimensions to 2D necessarily encounters loss of possibly relevant information.

These two problems can be avoided if label information is taken into account already at the dimensionality reduction step. We propose the following procedure as displayed in Fig. 4:

- Project the data $\mathbf{x}_i$ using a nonlinear discriminative visualization technique leading to points $p(\mathbf{x}_i) \in Y = \mathbb{R}^2$.
- Sample the projection space Y leading to points $\mathbf{z}'_i$. Determine points $\mathbf{z}_i$ in the data space X which are projected to these points $p(\mathbf{z}_i) \approx \mathbf{z}'_i$.
- Visualize the training points $\mathbf{x}_i$ together with the contours induced by the sampled function $(\mathbf{z}'_i, r(\mathbf{z}_i))$.

This procedure avoids the problems of the naive approach: on the one hand, a discriminative dimensionality reduction technique focusses on the aspects which are particularly relevant for the class labels and thus emphasizes the important characteristics of the classification function. On the other hand, sampling takes place in the projection space only, which is low dimensional.

One question remains: how can we find points $\mathbf{z}_i \in X$ which correpond to the projections $\mathbf{z}'_i \in Y$? For this purpose, we take an approach similar to kernel t-SNE: we define a mapping

$$p^{-1} : Y \to X, \mathbf{y} \mapsto \sum_i \alpha_i \cdot \frac{k_i(\mathbf{y}_i, \mathbf{y})}{\sum_i k_i(\mathbf{y}_i, \mathbf{y})} = \mathbf{A} \cdot [\mathbf{K}]_i$$

of the projection space to the original space which is trained based on the given samples $\mathbf{x}_i$, its projections $\mathbf{y}_i$, and its labels c_i. As before, k is the Gaussian kernel, $\mathbf{K}$ the kernel matrix applied to the points $\mathbf{y}_i$ which

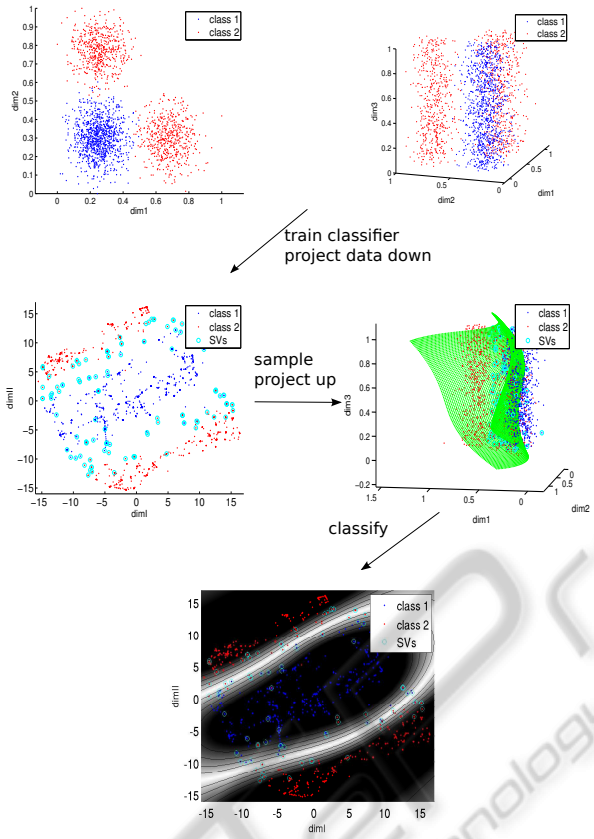

Figure 4: Principled procedure how to visualize a given data set and a trained classifier. The example displays a SVM trained in 3D.

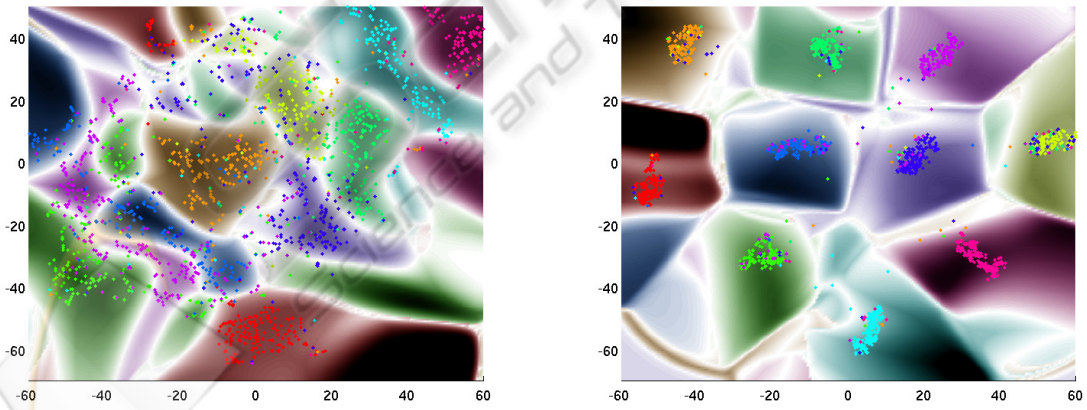

Figure 5: Visualization of an SVM classifier trained on the USPS data set by means of kernel t-SNE (top) and Fisher kernel t-SNE (bottom).

are projections of $\mathbf{x}_i$ and $[\mathbf{K}]_i$ the ith column. $\mathbf{A}$ is the matrix of parameters α_i. These parameters α_i are determined by means of a numeric optimization technique such that the following error is minimized:

$$\lambda_1 \cdot \|\mathbf{X} - \mathbf{A} \cdot \mathbf{K}\|^2 + \lambda_2 \cdot \|r(\mathbf{X}) - r(\mathbf{A} \cdot \mathbf{K})\|^2$$

Thereby, $\mathbf{X}$ denotes the points $\mathbf{x}_i$ used to train the discriminative mapping. $r(\cdot)$ denotes real values associated to the classification f indicating the strength of the class-membership association. λ_1 and λ_2 are positive weights which balance the two objectives formalized by this functional form: a correct inverse mapping of the data $\mathbf{x}_i$ and its projections $\mathbf{y}_i$ on the one side and a correct match of the induced classifications via the given classifier f on the other side.

An example application of this procedure for the USPS data set is based on the k t-SNE projections as specified in the last paragraph. An SVM with Gaussian kernel is trained on a subset of the data which is not used to train the subsequent kernel t-SNE or Fisher kernel t-SNE, respectively. A classification accuracy of 99% on the training set and 97% on the test set arises. We use two different kernel t-SNE mappings to obtain a training set for the inverse mapping p^{-1}: kernel t-SNE and Fisher kernel t-SNE, respectively. The weights of the cost function has been chosen as $\lambda_1 = 0.1$ and $\lambda_2 = 10000$, respectively. The resulting visualization of the SVM classification is displayed in Fig. 5(top) if the procedure is based on kernel t-SNE and Fig. 5(bottom) if the procedure is based on Fisher kernel t-SNE.

Obviously, the visualization based on Fisher kernel t-SNE displays much clearer class boundaries as compared to a visualization which does not take the class labeling into account. This visual impression is mirrored by a quantitative comparison of the projections. For the kernel t-SNE mapping, the classification induced in 2D as displayed in the map coincides with the original classification with a 85% accuracy only. If Fisher kernel t-SNE is used, the coincidence increases to 92%.

5 CONCLUSIONS

We have reviewed discriminative dimensionality reduction, its link to the Fisher information matrix, and we have discussed its difference to a direct classification. Based on Fisher kernel t-SNE, two applications have been proposed: a speed-up of dimensionality reduction on the one side and a visualization of a classifier such as SVM on the other side. So far, the applications have been demonstrated using one benchmark only, results for alternative benchmarks being similar. Note that the proposed techniques are not restricted to t-SNE, rather, similar techniques could be based on top of popular alternatives such as LLE or Isomap.

ACKNOWLEDGEMENTS

Funding by DFG under grants number HA 2719/7-1, HA 2719/4-1 and by the CITEC centre of excellence are gratefully acknowledged. We would like to thank the anonymous reviewers for helpful comments and suggestions.

REFERENCES

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404.

Bekkerman, R., Bilenko, M., and Langford, J., editors (2011). *Scaling up Machine Learning*. Cambridge University Press.

Biehl, M., Hammer, B., Merényi, E., Sperduti, A., and Villmann, T., editors (2011). *Learning in the context of very high dimensional data (Dagstuhl Seminar 11341)*, volume 1.

Braun, M. L., Buhmann, J. M., and Müller, K.-R. (2008). On relevant dimensions in kernel feature spaces. *J. Mach. Learn. Res.*, 9:1875–1908.

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

Bunte, K., Schneider, P., Hammer, B., Schleif, F.-M., Villmann, T., and Biehl, M. (2012b). Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26:159–173.

Caragea, D., Cook, D., Wickham, H., and Honavar, V. (2008). Visual methods for examining svm classifiers. In Simoff, S. J., Böhlen, M. H., and Mazeika, A., editors, *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 136–153. Springer.

Cohn, D. (2003). Informed projections. In Becker, S., Thrun, S., and Obermayer, K., editors, *NIPS*, pages 849–856. MIT Press.

Geng, X., Zhan, D.-C., and Zhou, Z.-H. (2005). Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(6):1098–1107.

Gisbrecht, A., Mokbel, B., and Hammer, B. (2013). Linear basis-function t-sne for fast nonlinear dimensionality reduction. In *IJCNN*.

Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. (2004). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hernandez-Orallo, J., Flach, P., and Ferri, C. (2011). Brier curves: a new cost-based visualisation of classifier performance. In *International Conference on Machine Learning*.

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T. L., and Tenenbaum, J. B. (2007). Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556.

Jakulin, A., Možina, M., Demšar, J., Bratko, I., and Zupan, B. (2005). Nomograms for visualizing support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 108–117, New York, NY, USA. ACM.

Kaski, S., Sinkkonen, J., and Peltonen, J. (2001). Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.

Lee, J. A. and Verleysen, M. (2010). Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31:2248–2257.

Ma, B., Qu, H., and Wong, H. (2007). Kernel clustering-based discriminant analysis. *Pattern Recognition*, 40(1):324–327.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Memisevic, R. and Hinton, G. (2005). Multiple relational embedding. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 913–920. MIT Press, Cambridge, MA.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48. IEEE.

Peltonen, J., Klami, A., and Kaski, S. (2004). Improved learning of riemannian metrics for exploratory analysis. *Neural Networks*, 17:1087–1100.

Poulet, F. (2005). Visual svm. In Chen, C.-S., Filipe, J., Seruca, I., and Cordeiro, J., editors, *ICEIS (2)*, pages 309–314.

Rüping, S. (2006). *Learning Interpretable Models*. PhD thesis, Dortmund University.

Tsang, I. W., Kwok, J. T., ming Cheung, P., and Cristianini, N. (2005). Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6:363–392.

Vellido, A., Martin-Guerroro, J., and Lisboa, P. (2012). Making machine learning models interpretable. In *ESANN'12*.

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

Wang, X., Wu, S., Wang, X., and Li, Q. (2006). Svmv - a novel algorithm for the visualization of svm classification results. In Wang, J., Yi, Z., Zurada, J., Lu, B.-L., and Yin, H., editors, *Advances in Neural Networks - ISNN 2006*, volume 3971 of *Lecture Notes in Computer Science*, pages 968–973. Springer Berlin / Heidelberg.

Ward, M., Grinstein, G., and Keim, D. A. (2010). *Interactive Data Visualization: Foundations, Techniques, and Application*. A. K. Peters, Ltd.

Witten, D. M. and Tibshirani, R. (2011). Supervised multidimensional scaling for visualization, classification, and bipartite ranking. *Computational Statistics and Data Analysis*, 55(1):789 – 801.