Neurocomputing 74 (2011) 1351–1358

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Relevance learning in generative topographic mapping

Andrej Gisbrecht, Barbara Hammer *

University of Bielefeld, CITEC Cluster of Excellence, Germany

ARTICLE INFO

Available online 22 February 2011

Keywords:
GTM
Topographic mapping
Relevance learning
Supervised visualization

ABSTRACT

Generative topographic mapping (GTM) provides a flexible statistical model for unsupervised data inspection and topographic mapping. Since it yields to an explicit mapping of a low-dimensional latent space to the observation space and an explicit formula for a constrained Gaussian mixture model induced thereof, it offers diverse functionalities including clustering, dimensionality reduction, topographic mapping, and the like. However, it shares the property of most unsupervised tools that noise in the data cannot be recognized as such and, in consequence, is visualized in the map. The framework of visualization based on auxiliary information and, more specifically, the framework of learning metrics as introduced in [14,21] constitutes an elegant way to shape the metric according to auxiliary information at hand such that only those aspects are displayed in distance-based approaches which are relevant for a given classification task. Here we introduce the concept of relevance learning into GTM such that the metric is shaped according to auxiliary class labels. Relying on the prototype-based nature of GTM, efficient realizations of this paradigm are developed and compared on a couple of benchmarks to state-of-the-art supervised dimensionality reduction techniques.

1. Introduction

Generative topographic mapping (GTM) has been introduced as a generative statistical model corresponding to the classical self-organizing map for unsupervised data inspection and topographic mapping [1]. An explicit statistical model has the benefit of great flexibility and easy adaptability to complex situations by means of appropriate statistical assumptions. Further, by offering an explicit mapping of latent space to observation space and a constrained Gaussian mixture model based thereof, GTM offers diverse functionality including visualization, clustering, topographic mapping, and various forms of data inspection. Like standard unsupervised machine learning and data inspection methods, however, GTM shares the garbage in - garbage out problem: the information inherent in the data is displayed independent of the specific user intention. Hence, if 'garbage' is present in the data, this noise is presented to the user since the statistical model has no way to identify the noise as such.

The domain of data visualization by means of dimensionality reduction techniques constitutes a matured field of research, many powerful nonlinear reduction techniques as well as a Matlab implementation being readily available, see e.g. [27,28,16,35,12,4,22,36,26]. However, researchers in the community start to appreciate that the inherently ill-posed problem of unsupervised data visualization and dimensionality reduction has to be shaped according to the user's needs to arrive at optimum results. This is particularly pronounced for real-life data sets which frequently do not allow a widely loss-free embedding into low dimensionality. Therefore, it has to be specified which parts of the available information should be preserved while embedding.

On the one hand, formal evaluation measures have been developed which allow an explicit formulation and evaluation based on the desired result, see e.g. [31–33,17]. On the other hand, researchers start to develop methods which can take auxiliary information into account. This way, the user can specify which information in the data is interesting for the current situation at hand by means of e.g. labeled data.

There exist a few classical mechanisms which take class labeling into account to reduce the data dimensionality: Feature selection constitutes one specific type of dimensionality reduction. Feature selection constitutes a well investigated research topic with numerous proposals based on general principles such as information theory or dedicated approaches developed for specific classifiers; see e.g. [13] for an overview. However, this way, the dimensionality reduction is restricted to very simple projections to coordinate axes.

More variable albeit still linear projection methods are in the focus of several classical discriminative dimensionality reduction tools: Fisher's linear discriminant analysis (LDA) projects data such that within class distances are minimized while between class distances are maximized. One important restriction of LDA is given by the fact that, this way, a meaningful projection to

* Corresponding author. Tel.: +49 521 106 12115; fax: +49 521 106 12181.
E-mail address: bhammer@techfak.uni-bielefeld.de (B. Hammer).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.12.015

dimensionality at most $c-1$, c being the number of classes, can be obtained. Hence, for two class problems only a linear visualization is found. Partial least squares regression (PLS) constitutes another classical method which objective is to maximize the covariance of the projected data and the given auxiliary information. It is also suited for situations where data dimensionality is larger than the number of data points; in such cases a linear projection is often sufficient and the problem is to find good regularizations to adjust the parameters accordingly. Informed projection [9] extends principal component analysis (PCA) to also minimize the sum squared error of data projections and the mean value of given classes, this way achieving a compromise of dimensionality reduction and clustering in the projection space. Another technique relies on metric learning according to auxiliary class information. For a metric which corresponds to a global linear matrix transform to low dimensionality this results in a linear discriminative projection of data, as proposed e.g. in [11,7].

Modern techniques extend these settings to general nonlinear projection of data into low dimensionality such that the given auxiliary information is taken into account. One way to extend linear approaches to nonlinear settings is offered by kernelization. This incorporates an implicit nonlinear mapping to a high-dimensional feature space together with the linear low-dimensional mapping. It can be used for every linear approach which relies on dot products in the feature space only such that an efficient computation is possible, such as several variants of kernel LDA [18,3]. However, it is not clear how to choose the kernel since its form severely influences the final shape of the visualization. In addition, the method has quadratic complexity with respect to the number of data due to its dependency on the full Gram matrix.

Another principled way to extend dimensionality reduction to auxiliary information is offered by an adaptation of the underlying metric which measures similarity in the original data space, see e.g., [6,34]. The principle of learning metrics has been introduced in [20,21]: the standard Riemanian metric of the given data manifold is substituted by a form which measures the information of the data for the given classification task. The Fisher information matrix induces the local structure of this metric and it can be expanded globally in terms of path integrals. This metric is integrated into self-organizing maps (SOM), multidimensional scaling (MDS), and a recent information theoretic model for data visualization which directly relies on the metric in the data space [20,21,33]. A drawback of the proposed method is its high computational complexity due to the dependency of the metric on path integrals or approximations thereof. A slightly different approach is taken in [10]: Instead of learning the metric, an ad hoc adaptation is used which also takes given class labeling into account. The corresponding metric induces a k-nearest neighbor graph which is shaped according to the given auxiliary information. This can directly be integrated into a supervised version of Isomap. The principle of discriminative visualization by means of a change of the metric is considered in more generality in the approach [8]. Here, a metric induced by prototype-based matrix adaptation as introduced e.g. in [24,25] is integrated in several popular visualization schemes including Isomap, manifold charting, locally linear embedding, etc.

Alternative approaches to incorporate auxiliary information modify the cost function of dimensionality reduction tools to include the given class information. The approaches introduced in [15,19] can both be understood as extensions of stochastic neighbor embedding (SNE). SNE tries to minimize the deviation of the distribution of data induced by pairwise distances in the original data space and projection space, respectively. Parametric embedding (PE) substitutes these distributions by conditional probabilities of classes, given a data point, this way mapping both, data points and class centers at the same time. For this procedure, however, an assignment of data to unimodal class centers needs to be known in advance. Multiple relational embedding (MRE) incorporates several dissimilarity structures in the data space induced by labeling, for example, into one latent space representation. For this purpose, the difference of the distribution of each dissimilarity matrix and the distribution of an appropriate transform of the latent space are accumulated, whereby the transform is adapted during training according to the given task. The weighting of the single components is taken according to the task at hand, whereby the authors report an only mild influence of the weighting on the final outcome. It is not clear, however, how to pick the form of the transformation to take into account multimodal classes.

Colored maximum variance unfolding (MVU) incorporates auxiliary information into MVU by substituting the raw data which is unfolded in MVU by the combination of the data and the covariance matrix induced by the given auxiliary information. This way, differences which should be emphasized in the visualization are weighted by the differences given by the prior labeling. Like MVU, however, the method depends on the full Gram matrix and is computationally demanding, such that approximations have to be used.

These approaches constitute promising candidates which emphasize the relevance of discriminative nonlinear dimensionality reduction. Only few of these methods allow an easy extension to new data points or approximate inverse mappings. Further, most methods suffer from high computational costs which make them infeasible for large data sets.

In this contribution, we extend GTM to the principle of learning metrics by combining the technique of relevance learning as introduced in supervised prototype-based classification schemes and the prototype-based unsupervised representation of data as provided by GTM. We propose two different ways to adapt the relevance terms which rely on different cost functions connected to prototype-based classification of data. Unlike [2], where a separate supervised model is trained to arrive at appropriate metrics for unsupervised data visualization, we can directly integrate the metric adaptation step into GTM due to the prototype-based nature of GTM. We test the ability of the model to visualize and cluster given data sets on a couple of benchmarks. It turns out that, this way, an efficient and flexible discriminative data mining and visualization technique arises.

2. The generative topographic mapping

The GTM as introduced in [1] models data $\mathbf{x} \in \mathbb{R}^D$ by means of a mixture of Gaussians which is induced by a lattice of points $\mathbf{w}$ in a low-dimensional latent space which can be used for visualization. The lattice points are mapped via $\mathbf{w} \mapsto \mathbf{t} = y(\mathbf{w},\mathbf{W})$ to the data space, where the function is parameterized by $\mathbf{W}$; one can, for example, pick a generalized linear regression model based on Gaussian base functions

$$y : \mathbf{w} \mapsto \Phi(\mathbf{w}) \cdot \mathbf{W} \tag{1}$$

where the base functions Φ are equally spaced Gaussians with variance σ^{-1}. Every latent point induces a Gaussian

$$p(\mathbf{x}|\mathbf{w},\mathbf{W},\beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|\mathbf{x} - y(\mathbf{w},\mathbf{W})\|^2\right) \tag{2}$$

with variance β^{-1}, which gives the data distribution as mixture of K modes

$$p(\mathbf{x}|\mathbf{W},\beta) = \sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}|\mathbf{w}^k,\mathbf{W},\beta) \tag{3}$$