Neurocomputing 73 (2010) 1074–1092

Contents lists available at ScienceDirect

Neurocomputing

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

Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data

Kerstin Bunte [a,*], Barbara Hammer [d], Axel Wismüller [b,c], Michael Biehl [a]

[a] University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, P.O. Box 407, 9700 AK Groningen, The Netherlands
[b] Department of Radiology, University of Rochester, 601 Elmwood Avenue, Rochester, NY 14642-648, USA
[c] Department of Biomedical Engineering, University of Rochester, 601 Elmwood Avenue, Rochester, NY 14642-648, USA
[d] Bielefeld University, CITEC, Universitätsstraße 23, 33615 Bielefeld, Germany

ARTICLE INFO

Available online 18 January 2010

Keywords:
Dimension reduction
Learning vector quantization
Visualization

ABSTRACT

Due to the tremendous increase of electronic information with respect to the size of data sets as well as their dimension, dimension reduction and visualization of high-dimensional data has become one of the key problems of data mining. Since embedding in lower dimensions necessarily includes a loss of information, methods to explicitly control the information kept by a specific dimension reduction technique are highly desirable. The incorporation of supervised class information constitutes an important specific case. The aim is to preserve and potentially enhance the discrimination of classes in lower dimensions. In this contribution we use an extension of prototype-based local distance learning, which results in a nonlinear discriminative dissimilarity measure for a given labeled data manifold. The learned local distance measure can be used as basis for other unsupervised dimension reduction techniques, which take into account neighborhood information. We show the combination of different dimension reduction techniques with a discriminative similarity measure learned by an extension of learning vector quantization (LVQ) and their behavior with different parameter settings. The methods are introduced and discussed in terms of artificial and real world data sets.

1. Introduction

The amount of electronic data doubles roughly every 20 months [1], and its sheer size makes it impossible for humans to manually scan through the available information. At the same time, rapid technological developments cause an increase of data dimension, e.g. due to the increased sensitivity of sensor technology (such as mass spectrometry) or the improved resolution of imaging techniques. This causes the need for reliable dimension reduction and data visualization techniques to allow humans to rapidly inspect large portions of data using their impressive and highly sensitive visual perception capabilities.

Dimension reduction and visualization constitutes an active field of research, see, e.g. [2–4] for recent overviews. The embedding of high-dimensional data into lower dimension is necessarily linked to loss of information. In the last decades an enormous number of unsupervised dimension reduction methods has been proposed. In general, unsupervised dimension reduction is an ill-posed problem since a clear specification which properties of the data should be preserved, is missing. Standard criteria, for instance the distance measure employed for neighborhood assignment, may turn out unsuitable for a given data set, and relevant information often depends on the situation at hand.

If data labeling is available, the aim of dimension reduction can be defined clearly: the preservation of the classification accuracy in a reduced feature space. Supervised linear dimension reducers are for example the generalized matrix learning vector quantization (GMLVQ) [5], linear discriminant analysis (LDA) [6], targeted projection pursuit [7], and discriminative component analysis [8]. Often, however, the classes cannot be separated by a linear classifier while a nonlinear data projection better preserves the relevant information. Examples for nonlinear discriminative visualization techniques include, extensions of the self-organizing map (SOM) incorporating class labels [9] or more general auxiliary information [10]. In both cases, the metric of SOM is adjusted such that it emphasizes the given auxiliary information and, consequently, SOM displays the aspects relevant for the given labeling. Further supervised dimension reduction techniques are model-based visualization [11] and parametric embedding [12]. In addition, linear schemes such as LDA can be kernelized yielding a nonlinear supervised dimension reduction scheme [13]. These models have the drawback that they are often very costly (squared or cubic with respect to the number of data points). Recent approaches provide scalable alternatives,

* Corresponding author. Tel.: +31 (0)50 363 7049.
E-mail address: k.bunte@rug.nl (K. Bunte).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.11.017

sometimes at the cost of non-convexity of the problem [14–16]. However, in most methods, the kernel has to be chosen prior to training and no metric adaptation according to the given label information takes place.

The aim of this paper is to identify and investigate principled possibilities to combine an adaptive metric and recent visualization techniques towards a discriminative approach. We will exploit the discriminative scheme exemplary for different types of visualization, necessarily restricting the number of possible combinations to exemplary cases. A number of alternative combinations of metric learning and data visualization as well as principled alternatives to arrive at discriminative visualization techniques (such as, e.g. colored maximum variance unfolding [17]) will not be tackled in this paper.

In this contribution we combine prototype-based matrix learning schemes, which result in local discriminative dissimilarity measures and local linear projections of the data, with different neighborhood based nonlinear dimension reduction techniques and a charting technique. The complexity of the matrix learning technique is only linear in the number of points S, their dimension N and can be controlled by the number of the prototypes m and sweep through the training set t, leading to an overall algorithm complexity of only $\mathcal{O}(S \cdot N \cdot m \cdot t)$. In the second step unsupervised techniques like manifold charting [18], Isomap [19], locally linear embedding (LLE) [20], the exploration observation machine (XOM) [21] and stochastic neighbor embedding (SNE) [22] are performed incorporating the supervised information from the LVQ approach. This leads to supervised nonlinear dimension reduction and visualization techniques. Note, for one presented training sample, the matrix learning scheme only needs to compute the distances to all prototypes. And the number of prototypes is usually much smaller than the number of data points. However, the combination with another dimension reduction technique may make the computation of the distances of all data points necessary, e.g. with Isomap or SNE. This is at least a quadratic problem but can be moderated by approximations [23–25].

The following section gives a short overview over the techniques. We focus on the question in how far local linear discriminative data transformations as provided by GMLVQ offer principled possibilities to extend standard unsupervised visualization tools to discriminative visualization. Section 3 discusses the different approaches for one artificial and three real world data sets and compares the results to popular supervised as well as unsupervised dimension reduction techniques. Finally we conclude in Section 4.

2. Supervised nonlinear dimension reduction

For general data sets a global linear reduction to lower dimensions may not be sufficient to preserve the information relevant for classification. In [3] it is argued that the combination of several local linear projections to a nonlinear mapping can yield promising results. We use this concept and learn discriminative local linear low-dimensional projections from labeled data using an efficient prototype based learning scheme, generalized matrix learning vector quantization (GMLVQ). Locally linear projections which result from this first step provide, on the one hand, local transformations of the data points which preserve the information relevant for the classification as much as possible. Instead of the local coordinates, local distances induced by these local representation of data can be considered. As a consequence, visualization techniques which rely on local coordinate systems or local distances, respectively, can be combined with this first step to arrive at a discriminative global nonlinear projection method. This way, an incorporation into techniques such as manifold charting [18], Isomap [19], locally linear embedding (LLE) [20], stochastic neighbor embedding (SNE) [22], maximum variance unfolding (MVU) [26] and the exploration observation machine (XOM) [21] becomes possible.

The following subsections give a short overview over the initial prototype based matrix learning scheme and the different visualization algorithms.

2.1. Localized LiRaM LVQ

Learning vector quantization (LVQ) [27] constitutes a particularly intuitive classification algorithm which represents data by means of prototypes. LVQ itself constitutes a heuristic algorithm, hence extensions have been proposed for which convergence and learnability can be guaranteed [28,29]. One particularly crucial aspect of LVQ schemes is the dependency on the underlying metric, usually the Euclidean metric, which may not suit the underlying data structure. Therefore, general metric adaptation has been introduced into LVQ schemes [29,30]. Recent extensions parameterize the distance measure in terms of a relevance matrix, the rank of which may be controlled explicitly. The algorithm suggested in [5] can be employed for linear dimension reduction and visualization of labeled data. The local linear version presented here provides the ability to learn local low-dimensional projections and combine them into a nonlinear global embedding using charting techniques or projection methods based on local data topologies or local distances. Several schemes for adaptive distance learning exist, for example large margin nearest neighbor (LMNN) [31] to name just one. We compared the LMNN technique with the LVQ based approach on the basis of a content based image retrieval application in an earlier publication (see [32]). Furthermore it should be mentioned that LMNN learns a global distance measure. More powerful, local distance learning would presumably involve higher computational complexity and should be feasible for small dimensionality N only.

We consider training data $\boldsymbol{x}_i \in \mathbb{R}^N$, $i = 1 \ldots S$ with labels y_i corresponding to one of C classes, respectively. The aim of LVQ is to find m prototypes $\boldsymbol{w}_j \in \mathbb{R}^N$ with class labels $c(\boldsymbol{w}_j) \in \{1, \ldots, C\}$ such that they represent the classification as accurately as possible. A data point $\boldsymbol{x}_i$ is assigned to the class of its closest prototype $\boldsymbol{w}_j$ where $d(\boldsymbol{x}_i, \boldsymbol{w}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{w}_l)$ for all $j \neq l$. d usually denotes the squared Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{w}_j) = (\boldsymbol{x}_i - \boldsymbol{w}_j)^\top (\boldsymbol{x}_i - \boldsymbol{w}_j)$. Generalized LVQ (GLVQ) [33] adapts prototype locations by minimizing the cost function

$$E_{\mathrm{GLVQ}} = \sum_{i=1}^{S} \Phi\left(\frac{d(\boldsymbol{w}_J, \boldsymbol{x}_i) - d(\boldsymbol{w}_K, \boldsymbol{x}_i)}{d(\boldsymbol{w}_J, \boldsymbol{x}_i) + d(\boldsymbol{w}_K, \boldsymbol{x}_i)}\right), \tag{1}$$

where $\boldsymbol{w}_J$ denotes the closest prototype with the same class label as $\boldsymbol{x}_i$, and $\boldsymbol{w}_K$ is the closest prototype with a different class label. Φ is a monotonic function, e.g. the logistic function or the identity. In this work we use the identity. This cost function aims at an adaptation of the prototypes such that a large hypothesis margin is obtained, this way achieving correct classification and, at the same time, robustness of the classification, see [34]. A learning algorithm can be derived from the cost function E_{GLVQ} by means of a stochastic gradient descent as shown in [29,28].

Matrix learning in GLVQ (GMLVQ) [30,34] substitutes the usual squared Euclidean distance d by a more advanced dissimilarity measure which contains adaptive parameters, thus resulting in a more complex and better adaptable classifier. In [5], it was proposed to choose the dissimilarity as

$$d^{\Lambda_j}(\boldsymbol{w}_j, \boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{w}_j)^\top \Lambda_j (\boldsymbol{x}_i - \boldsymbol{w}_j), \tag{2}$$