Expert Systems with Applications 40 (2013) 2189–2197

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

# A comparative study of nonlinear manifold learning methods for cancer microarray data classification

Carlotta Orsenigo *, Carlo Vercellis

*Dept. of Management, Economics and Industrial Engineering, Politecnico di Milano Via Lambruschini 4b, 20156 Milano, Italy*

ARTICLE INFO

*Keywords:*
Nonlinear dimensionality reduction
Manifold learning
Classification
Cancer microarray data

ABSTRACT

The paper presents an empirical comparison of the most prominent nonlinear manifold learning techniques for dimensionality reduction in the context of high-dimensional microarray data classification. In particular, we assessed the performance of six methods: isometric feature mapping, locally linear embedding, Laplacian eigenmaps, Hessian eigenmaps, local tangent space alignment and maximum variance unfolding. Unlike previous studies on the subject, the experimental framework adopted in this work properly extends to dimensionality reduction the supervised learning paradigm, by regarding the test set as an out-of-sample set of new points which are excluded from the manifold learning process. This in order to avoid a possible overestimate of the classification accuracy which may yield misleading comparative results. The different empirical approach requires the use of a fast and effective out-of-sample embedding method for mapping new high-dimensional data points into an existing reduced space. To this aim we propose to apply multi-output kernel ridge regression, an extension of linear ridge regression based on kernel functions which has been recently presented as a powerful method for out-of-sample projection when combined with a variant of isometric feature mapping. Computational experiments on a wide collection of cancer microarray data sets show that classifiers based on Isomap, LLE and LE were consistently more accurate than those relying on HE, LTSA and MVU. In particular, under different experimental conditions LLE-based classifier emerged as the most effective method whereas Isomap algorithm turned out to be the second best alternative for dimensionality reduction.



## 1. Introduction

Dimensionality reduction methods attempt to provide meaningful low-dimensional representations of high-dimensional data. They may prove quite useful to deal with both unsupervised tasks, such as clustering or data visualization, and supervised learning, where they can speed up the training time and increase the prediction accuracy.

Microarray data classification is a rather attractive domain for dimensionality reduction. The collection of gene expression profiles from microarray experiments usually results in high-dimensional data sets composed by a huge number of features (genes) and a relatively few number of distinct examples (patients). The classification of these data sets represents a computationally intensive task often affected by the curse of dimensionality (Bellman, 1961), that may deeply benefit in terms of efficiency and accuracy from dimensionality reduction as a preliminary step.

Beside linear methods, such as principal component analysis (Jolliffe, 1986) and metric multidimensional scaling (Cox & Cox, 1994), nonlinear dimensionality reduction techniques, able to tackle data sets with underlying nonlinear structures, have been more recently proposed. Within the family of nonlinear algorithms manifold learning methods have drawn great interest, by providing considerable results on artificial and real world data sets especially for data visualization. They include, among others, isometric feature mapping (Isomap) (de Silva & Tenenbaum, 2002; Tenenbaum, de Silva, & Langford, 2000), locally linear embedding (LLE) (Roweis & Saul, 2000; Saul & Roweis, 2003), Laplacian eigenmaps (LE) (Belkin & Niyogi, 2000), Hessian eigenmaps (HE) (Donoho & Grimes, 2003), local tangent space alignment (LTSA) (Zhang & Zha, 2004) and maximum variance unfolding (MVU) (Weinberger & Saul, 2004).

Manifold learning methods attempt to recover the low dimensional manifold along which data are supposed to lie. Formally, given a set of data points $\mathcal{S}_m = \{\mathbf{x}_i, i \in \mathcal{M} = \{1, 2, \ldots, m\}\} \subset \Re^n$ arranged along a smooth nonlinear manifold $M$ of intrinsic dimension $d$, with $d \ll n$, they aim at finding a function $f : M \to \Re^d$ mapping $\mathcal{S}_m$ into $\mathcal{D}_m = \{\mathbf{z}_i, i \in \mathcal{M} = \{1, 2, \ldots, m\}\} \subset \Re^d$ such that some geometrical properties of the data in the input space are preserved in the reduced space. These techniques are often referred to as spectral embedding methods (Weinberger & Saul, 2004) since they convert the dimensionality reduction problem into the eigendecomposition of a symmetric positive semidefinite matrix.

* Corresponding author. Tel.: +39 02 23993970; fax: +39 02 3993978.
*E-mail addresses:* carlotta.orsenigo@polimi.it (C. Orsenigo), carlo.vercellis@polimi.it (C. Vercellis).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.044

In this paper we perform a systematic empirical comparison of the most prominent nonlinear manifold learning methods in the context of high-dimensional data classification. The experimental framework entailed in our work differs significantly from the one considered in previous studies (Bartenhagen, Klein, Ruckert, Jiang, & Dugas, 2010; van der Maaten, Postma, & Herik, 2009), in which every data set is entirely projected in the low-dimensional space before being separated into training and test set. This approach may cause an overestimate of the classification accuracy on the test set whose points have been unfairly used for the manifold reconstruction, and may yield misleading results in the comparative evaluation. Our computational framework more properly extends to dimensionality reduction the supervised learning paradigm, by regarding the test set as an out-of-sample set of new points which are excluded from both the manifold learning process and the training phase. More precisely, the evaluation scheme is based on a three-step procedure: each data set is first divided into pairs of training and test sets and, for each pair, the $d$-dimensional embedding of the training points is computed. Then, the function mapping the training set into its projection is inductively approximated and is used to find the low-dimensional representation of the corresponding test set. Finally, a generic classifier is trained on the projected training set and its accuracy is estimated on the embedded test set.

Notice that this different perspective requires a fast and accurate method for embedding new high-dimensional data points into an existing reduced space. To this aim we resort to multi-output kernel ridge regression (KRR), an extension of linear ridge regression based on kernel functions which has been recently proposed for out-of-sample mapping in (Orsenigo & Vercellis, 2012b). KRR has proven to be very effective when combined with a supervised variant of Isomap in the classification of a broad collection of real world data sets. Indeed, it outperformed generalized regression neural networks which have been generally adopted by previous studies.

The remainder of the paper is organized as follows. Section 2 offers an overview of the nonlinear manifold learning methods involved in the empirical comparison. Section 3 describes multi-output kernel ridge regression for out-of sample mapping. Section 4 illustrates the experimental settings. Section 5 presents the comparative results concerning the classification of a wide range of cancer microarray data sets. Conclusions are discussed in Section 6.

## 2. Nonlinear manifold learning methods

This section provides a brief description of the manifold learning techniques considered in this study. Our attention was devoted to those methods which find a low-dimensional representation by preserving global or local geometrical properties of the data. Note that almost all algorithms are based on a common framework. A sparse matrix is first derived from a weighted neighborhood graph whose nodes correspond to data points and edges represent neighborhood relations. The embedding in the low $d$-dimensional space is then obtained by computing the $d$ eigenvectors associated with the $d$ largest or smallest nonzero eigenvalues of this matrix.

In what follows, let $\mathbf{X}$ denote the $m \times n$ matrix whose rows represent the input vectors $\mathbf{x}_i$, $\mathbf{Z}$ the corresponding $m \times d$ matrix whose rows are the projected vectors $\mathbf{z}_i$, $\mathbf{\Lambda}$ a square $d$-dimensional diagonal matrix of eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_d)$, $\mathbf{V}$ the $m \times d$ matrix of associated eigenvectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)$, $\mathbf{I}_d$ the identity matrix of size $d$ and $\mathbf{1}$ a $m$-dimensional vector of ones.

### 2.1. Isometric feature mapping

Isometric feature mapping (Isomap) represents a generalization of metric multidimensional scaling (MDS) to nonlinear manifolds. Unlike classical MDS which attempts to preserve the Euclidean distance between data points, Isomap finds an embedding in which the geodesic distance between two points in the input space is as close as possible to the Euclidean distance between their projections in the target space. The geodesic distance is defined by the length of the shortest curve connecting two points on the underlying manifold. It is estimated by the shortest path computed between the corresponding vertices in a weighted neighborhood graph, in which every data point is connected to its $k$ nearest neighbors and the weight of an edge equals the Euclidean distance of its endpoints.

Let $\mathbf{D}_G$ denote the matrix of the geodesic distances between the points in the neighborhood graph. The embedding into the $d$-dimensional space is computed by minimizing the function

$$\|\tau(\mathbf{D}_G) - \tau(\mathbf{D}_Z)\|_F, \tag{1}$$

where $\mathbf{D}_Z = [d_{ij}]$ is the matrix of pairwise Euclidean distances $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|$ of the data projections in $\Re^d$, the $\tau$ operator converts distances to inner products and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The global minimum of (1) is achieved by computing the $d$ eigenvectors associated to the $d$ largest eigenvalues of the geodesic distances matrix $\tau(\mathbf{D}_G)$, and by setting the projections $\mathbf{Z} = \mathbf{V}\mathbf{\Lambda}^{1/2}$.

Within the original Isomap algorithm two alternative criteria for defining the neighborhood of each point $\mathbf{x}_i$ were proposed: searching for the $k$ nearest neighbors in terms of the Euclidean distance or choosing all points lying within a fixed-radius hypersphere centered on $\mathbf{x}_i$. Both methods suffer from drawbacks, resulting in instability and low robustness for data sets affected by noise, outliers or scarcity of examples, since these conditions may generate short-circuits within the manifold that distort the low-dimensional embedding (Balasubramanian, Schwartz, Tenenbaum, de Silva, & Langford, 2002). Alternative approaches have been proposed to overcome this problem, by taking into account the density of each point to properly modify the distances in the input space (de Silva & Tenenbaum, 2002), by using adaptive procedures for generating the neighborhoods (Wei, Peng, Lin, Huang, & Wang, 2008; Zhan, Yin, Liu, & Zhang, 2009) or by resorting to double-bounding rules capable of preventing short-circuits (Orsenigo & Vercellis, 2012a). Furthermore, it has been shown that isometric feature mapping may fail in case of nonconvex manifolds (Tipping, 2000).

Despite these weaknesses Isomap has been successfully applied to the analysis of high-dimensional biomedical data (Dawson, Rodriguez, & Malyj, 2005; Park, 2012; Weng, Zhang, Lin, & Zhang, 2005). In the context of classification a supervised extension has been also proposed in which the distances between points are modified according to their labels (Geng, Zhan, & Zhou, 2005).

### 2.2. Locally linear embedding

Locally linear embedding (LLE) attempts to recover the global structure of nonlinear manifolds from locally linear fits, so to preserve the local geometry of the input data in the low-dimensional space.

Once the neighborhood graph is constructed based on the Euclidean distance, LLE represents each point $\mathbf{x}_i$ as a linear combination of its neighbors

$$\mathbf{x}_i = \sum_{j \in \mathcal{K}_i} w_{ij}\mathbf{x}_j, \quad i \in \mathcal{M}, \tag{2}$$

where $\mathcal{K}_i$ is the set of indices of the $k$ nearest neighbors of $\mathbf{x}_i$, and the generic weight $w_{ij}$ highlights the role of neighbor $j$ in the reconstruction of point $i$. The weight coefficients for all data points are then computed by minimizing the function