Expert Systems with Applications 39 (2012) 7757–7762

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Kernel ridge regression for out-of-sample mapping in supervised manifold learning

Carlotta Orsenigo *, Carlo Vercellis

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

ARTICLE INFO

Keywords:
Supervised manifold learning
Nonlinear dimensionality reduction
Supervised isomap
Kernel ridge regression
Classification

ABSTRACT

Manifold learning methods for unsupervised nonlinear dimensionality reduction have proven effective in the visualization of high dimensional data sets. When dealing with classification tasks, supervised extensions of manifold learning techniques, in which class labels are used to improve the embedding of the training points, require an appropriate method for out-of-sample mapping.

In this paper we propose multi-output kernel ridge regression (KRR) for out-of-sample mapping in supervised manifold learning, in place of general regression neural networks (GRNN) that have been adopted by previous studies on the subject. Specifically, we consider a supervised agglomerative variant of Isomap and compare the performance of classification methods when the out-of-sample embedding is based on KRR and GRNN, respectively. Extensive computational experiments, using support vector machines and *k*-nearest neighbors as base classifiers, provide statistical evidence that out-of-sample mapping based on KRR consistently dominates its GRNN counterpart, and that supervised agglomerative Isomap with KRR achieves a higher accuracy than direct classification methods on most data sets.

1. Introduction

Dimensionality reduction techniques are aimed at discovering a low-dimensional structure hidden in the original data. This can be useful as a preprocessing step for data visualization, clustering, classification or regression, to reduce training time and achieve better results particularly when dealing with high dimensional data sets.

Beside linear methods, such as principal component analysis (PCA) (Jolliffe, 1986) and metric multidimensional scaling (MDS) (Cox & Cox, 1994), nonlinear dimensionality reduction techniques have been more recently proposed to uncover a lower dimensional manifold along which data lie. Manifold learning methods include, among others, isometric feature mapping (Isomap) (Tenenbaum, de Silva, & Langford, 2000), locally linear embedding (LLE) (Roweis & Saul, 2000) and Laplacian eigenmaps (Belkin & Niyogi, 2003), which are based on a common framework that first determines the neighbors of each point in the input space and then computes the spectral embedding using the eigenvectors of an appropriate weight matrix defined over the input data.

Unlike linear PCA and MDS, most manifold learning methods do not derive an explicit mapping function for embedding the data into the reduced space. Therefore, when new data points need to be projected into the lower dimensional space, a fast and accurate method for out-of-sample mapping has to be devised. This happens in particular for classification problems, where supervised extensions of manifold learning methods are considered in which the class labels are used to improve the embedding of the training points (Geng, Zhan, & Zhou, 2005; Vlachos, Domeniconi, Gunopulos, Kollios, & Koudas, 2002). It is apparent that the effectiveness of the method adopted for out-of-sample projection is critical for achieving a high classification accuracy.

To accomplish the out-of-sample mapping in supervised manifold learning, we propose to use multi-output kernel ridge regression (KRR) in place of general regression neural networks (GRNN) that have been adopted by previous studies on the subject (Geng et al., 2005; Gu, Xu, & Ye, 2007; Ribeiro et al., 2009; Vlachos et al., 2002). GRNN is a universal technique for learning a map between vector spaces, based on a normalized radial basis function network in which a hidden unit is introduced for every training point (Specht, 1991). GRNN has the advantage of being faster to train than other neural networks, since the weights are calculated in a single step instead of being computed by a backpropagation descent algorithm.

In turn, ridge regression is an alternative method for learning functions, based on a regularized extension of least squared techniques (Hastie, Tibshirani, & Friedman, 2009). Although it is usually formulated for learning real-valued maps, it can be modified to deal with vector-valued regression (Cortes, Mohri, & Weston, 2005). Furthermore, nonlinear maps can be approximated by means of kernel ridge regression, an extension of linear ridge regression based on kernel functions. Likewise GRNN, KRR offers the advantage of being fast to evaluate, requiring only a single matrix inversion which depends on the number of points but is

* Corresponding author. Tel.: +39 02 23993970.
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.
doi:10.1016/j.eswa.2012.01.060

independent of the number of attributes in the input space. The rationale behind resorting to KRR for out-of-sample embedding is that it can achieve an ideal trade-off between bias and variance of the estimates by means of the regularization term, thus resulting in a more precise approximation of the mapping aimed at out-of-sample data projection.

In this paper, we consider a supervised extension of Isomap for dimensionality reduction, in which the distances between points are modified according to their labels as in Geng et al. (2005). Furthermore, to obtain a connected neighborhood graph without increasing too much the number of neighbors, we introduce a variant of Isomap based on a hierarchical agglomeration of the components. The performances of KRR for out-of-sample mapping are finally evaluated in terms of classification accuracy. To this aim, the projected training data are fed into a generic classifier whose accuracy is estimated on the embedding of the out-of-sample test set.

Extensive computational experiments have been performed on a collection of real word data sets, using support vector machines (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995) and k-nearest neighbors (Duda & Hart, 1973) as base classifiers. Our tests provide statistical evidence that out-of-sample mapping based on KRR consistently dominates its GRNN counterpart, and that supervised agglomerative Isomap with KRR achieves a higher accuracy than direct classification methods for most data sets.

2. Unsupervised agglomerative Isomap

Given a set of data $\mathcal{S}_m = \{\mathbf{x}_i, i \in \mathcal{M} = \{1, 2, \ldots, m\}\} \subset \Re^n$, they may happen to be arranged along a manifold whose intrinsic dimension d is much lower than n. The aim of Isomap is to find an embedding of the given data set $\mathcal{S}_m$ into a lower dimensional space $\Re^d$, in such a way that the geodesic distance between a pair of points in the input space should be as close as possible to the Euclidean distance between their projections into the lower dimensional space. Since the manifold along which the points are supposed to lie is generally unknown in advance, it is not possible to compute the geodesic distances in an exact way so that they have to be evaluated approximately. The geodesic distance between each pair of points is estimated by the shortest path computed between the corresponding vertices in an appropriate neighborhood graph.

According to the original Isomap algorithm (Tenenbaum et al., 2000), two vertices are connected by an edge if and only if one of them is among the h nearest neighbors of the other, where h is a parameter to be specified. This leads to a weighted undirected neighborhood graph in which the length q_{ij} of an edge (i,j) equals the Euclidean distance e_{ij} between its endpoints. The Isomap algorithm further requires the neighborhood graph to be connected, in order to compute the shortest path between any pair of vertices. Consequently, the parameter h has to be chosen large enough to guarantee connectedness of the neighborhood graph. Unfortunately, large values of h may induce inappropriate connections between different folds of the manifold, resulting in an inaccurate embedding of the points in $\mathcal{S}_m$.

To avoid this drawback, we propose a two stage variant of Isomap: first, the neighborhood graph is built according to the original scheme; if this results in a disconnected graph, an iterative procedure is then applied that merges the two closest connected components as in a hierarchical agglomerative method for clustering.

The unsupervised agglomerative Isomap algorithm takes as input the dissimilarity matrix **Q**, together with the parameters d and h, and can be summarized as follows:

2.1. Procedure A-Isomap (**Q**, d, h)

1. Build the neighborhood graph by connecting each point to its h nearest neighbors.
2. If the neighborhood graph is not connected, iteratively agglomerate two components for which the interdistance is minimum, by adding an arc between the closest pair of points belonging to distinct components.
3. Approximate the geodesic distances with the shortest paths computed between every pair of vertices, obtaining the matrix $\mathbf{G} = [g_{ij}]$ of the corresponding lengths.
4. Find the embedding $\mathbf{z}_i, i \in \mathcal{M}$, of the points into the lower dimensional space $\Re^d$ using classical multidimensional scaling. To this aim, first compute the matrix $\mathbf{S} = -\mathbf{BKB}/2$, where $\mathbf{K} = [g_{ij}^2]$ is the matrix of the squared shortest path lengths and $\mathbf{B}$ is a centering matrix whose generic element is defined as $b_{ij} = \delta_{ij} - 1/m$, with δ_{ij} denoting the Kronecker's delta. Then take the first d eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ of the matrix $\mathbf{S}$ in non-increasing order and the corresponding eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$; the pth component of the ith point in $\Re^d$ is defined as $z_{i,p} = \sqrt{\lambda_p} v_{i,p}, p = 1, 2, \ldots, d$.

The dimension d of the manifold can be estimated by identifying the point at which the curve of residual variances versus the dimensionality of the embedding space flattens, as suggested in Tenenbaum et al. (2000).

3. Supervised agglomerative Isomap

A classification problem consists of determining a function which optimally describes the relationship between a set of points $\mathcal{S}_m = \{\mathbf{x}_i, i \in \mathcal{M}\} \subset \Re^n$ and their associated categorical labels $y_i \in \{1, 2, \ldots, D\}$.

When dealing with classification, unsupervised dimensionality reduction methods can be applied as a preprocessing step to the entire data set. As an alternative, supervised extensions of manifold learning can be devised by using class labels, so to achieve a projection of training data which is more suitable for the subsequent classification phase. In this case, however, an out-of-sample problem has to be solved in order to evaluate the accuracy of the classifier on a test set, whose class labels are kept hidden. This means that a method to project the test points into the embedding space has to be developed. This has been usually done on an inductive basis, by means of general regression neural networks. Here we propose an out-of-sample embedding method based on multi-output kernel ridge regression.

In particular, a supervised variant of Isomap can be obtained by using the training labels to increase inter-class dissimilarity while decreasing intra-class dissimilarity. More specifically, before applying algorithm A-Isomap the distance q_{ij} between the pair of points (i,j) is modified according to their labels, as suggested in (Geng et al., 2005):

$$q_{ij} = \begin{cases} \sqrt{1 - e^{\frac{-e_{ij}^2}{\beta}}} & \text{if } y_i = y_j, \\ \sqrt{e^{\frac{e_{ij}^2}{\beta}}} - \alpha & \text{if } y_i \neq y_j. \end{cases} \tag{1}$$

The parameter β is aimed at preventing an uncontrolled growth of q_{ij} when the Euclidean distance e_{ij} is large, and it is usually set equal to the average distance between all pairs of points. The parameter α is instead used to counteract overfitting during classification, by smoothing the gap between inter-class and intra-class dissimilarity.

The supervised agglomerative Isomap algorithm can be summarized as follows, where $\mathcal{S}_m$ denotes the training set with the associated vector of labels $\mathbf{y}_m$, and $\mathcal{T}_r$ is the out-of-sample set. In turn,