Pattern Recognition 48 (2015) 1563–1570

Contents lists available at ScienceDirect

Pattern Recognition

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

Greedy approaches to semi-supervised subspace learning

Minyoung Kim

Department of Electronics & IT Media Engineering, Seoul National University of Science & Technology, Seoul 139-743, Republic of Korea

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form
20 September 2014
Accepted 19 October 2014
Available online 31 October 2014

Keywords:
Dimensionality reduction
Infinite-dim greedy search
Semi-supervised learning

ABSTRACT

Subspace estimation is of paramount importance in dealing with high-dimensional data with noise. In this paper we consider a semi-supervised learning setup where certain supervised information (e.g., class labels) is available for only a part of data samples. First we formulate a unifying optimization problem that subsumes the well-known principal component analysis in unsupervised scenarios as a special case, while exploiting labeled data effectively. To circumvent difficult matrix rank constraints in the original problem, we propose a nuclear norm based relaxation that ends up with convex optimization. We then provide an infinite-dimensional greedy search algorithm that solves the optimization problem efficiently. An extension to nonlinear dimensionality reduction is also introduced, which is as efficient as the linear model via dual representation with kernel trick. The effectiveness of the proposed approach is demonstrated experimentally on several semi-supervised learning problems.

1. Introduction

In dealing with high-dimensional data, it is a central problem to estimate a low-dimensional subspace that represents the underlying data distribution compactly and faithfully. Also known as *dimensionality reduction*, it allows computationally tractable solutions, avoids the curse of dimensionality [1], and has a filtering effect for potentially noisy data. The efficacy of subspace learning has been verified in several applications in diverse fields including pattern recognition [2–5].

Previous approaches have attempted to discover a low-dim structure of data either by extracting global statistical information such as principal components (i.e., directions of the largest possible variance) of PCA [6], or by exploiting a geometric nature of data such as geodesic distances of ISOMAP [3]. Besides these *unsupervised* scenarios, often the data samples are annotated with additional label information that affects the formation of the low-dim subspace. In typical *supervised* setups, we have additional label information (e.g., class labels) that indicates grouping/separation of data points in the intrinsic subspace. The idea has been formulated in different ways as the linear discriminant analysis (LDA) [1] and related models [7–9].

However, relatively few research has been conducted to deal with dimension reduction with partially labeled data. The *semi-supervised* setups are more common in practice, and naturally unifies two extreme scenarios mentioned above. In this paper we propose a novel semi-supervised subspace learning algorithm. We first formulate an optimization problem that finds a subspace with minimal data reconstruction error (i.e., faithfully preserving the variation in data), and at the same time, exploits the class supervision for correct formation of the labeled data points in the embedded subspace. Our subspace optimization problem subsumes the unsupervised PCA as a special case, while extending the graph Laplacian based regularization [10] to supervised data.

To circumvent difficult matrix rank constraints in the optimization problem, we introduce a nuclear norm based relaxation that yields convex optimization. However, solving the relaxed problem is still challenging due to the non-differentiable objective originating from the nuclear norm. We provide an efficient infinite-dimensional greedy search algorithm: an over-complete basis set of symmetric dyads is considered to represent a family of subspace mappings, where at each stage a new basis is selected in a greedy fashion. The effectiveness of the proposed approach is demonstrated empirically in several semi-supervised subspace learning scenarios.

We contrast our approach with some of the existing semi-supervised embedding or classification approaches recently introduced and closely related to ours. In [11], fairly reasonable extensions of the existing (unsupervised) nonlinear dimension reduction algorithms like LLE and ISOMAP have been introduced. Unlike ours and other commonly assumed setups, however, they rather focus on the semi-supervised setups where the low-dim coordinates are known for some of the data points.

In [12], the affinity/dissimilarity of low-dim embeddings in accordance with available label information has been exploited within the graph Laplacian framework similar in spirit to our objective (especially, the cost term related to labels). But they only consider the binary affinity relationship (either must-link or cannot-link), which may not be appropriate for the problems with

E-mail address: mikim21@gmail.com

http://dx.doi.org/10.1016/j.patcog.2014.10.031
0031-3203/

non-binary labels or real-valued affinity measures. On the other hand, our framework admits arbitrary label structures through similarity scores properly chosen.

Some other recent approaches [13,14] aim to extend the discriminative dimension reduction algorithms like LDA to unlabeled data points. In particular, the regularized semi-supervised learning framework has been employed in [14], where the graph-based manifold structure is considered to impose a penalty term for unlabeled data points. Although these approaches achieve accurate label prediction in many situations, one potential drawback is that they do not explicitly consider the goodness of data reconstruction (e.g., the objective of PCA). On the contrary, our approach takes into account not only the overall variation of the original data, but also consistency with available labels, enjoying the merits from both generative and discriminative learning.

Another interesting line of research is devoted for regarding the unlabeled data as a regularizer for learning the classification function. The main idea is to exploit the unlabeled data to enforce the classifier output to be consistent with the original data space. Specifically, the LapSVM [15] utilizes the SVM-like hinge loss for the labeled data, while the unlabeled data are exploited as a regularizer for the class prediction function via the graph Laplacian based constraint to preserve similarity in the original data. Later in [16], the functional structure of the predictor is further extended to a deep (multi-layer) neural network architecture (rather than a shallow linear functional in the LapSVM), which is often believed to uncover the useful but unknown feature structures in the raw data.

Compared to our approach, however, their objectives take less tractable forms of non-differentiable hinge functions. Moreover, the optimizations involve quite complicated steps of general (sub) gradient descent where the classifiers are highly coupled with the data samples in complex ways. As demonstrated in the experiments, our approach is computationally far more efficient than these approaches. They often suffer from the overhead in computing the gradients that go through the entire training samples especially when the function architecture becomes more complex and deeper.

1.1. Problem setups and notations

A data (feature) point is denoted by $\mathbf{x} \in \mathbb{R}^p$. One may have label $y \in \mathcal{Y}$ for each data point, where y can be, but not restricted to, a class label $\mathcal{Y} = \{1, \ldots, K\}$. We are given the labeled data $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$ and the unlabeled data $\mathcal{U} = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\}$, which are i.i.d. samples[1] from an unknown distribution $P(\mathbf{x}, y)$. The norm $\|\cdot\|$ (without subscript) indicates the L_2 norm or Euclidean norm. $\|\cdot\|_F$ is the Frobenius norm. The task of subspace learning is to find a subspace projection matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ for $\mathbf{b}_j \in \mathbb{R}^p$, $j = 1, \ldots, q$, such that $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$ ($\in \mathbb{R}^q$) is a faithful low-dim representation for $\mathbf{x}$. The subspace dimension q ($\ll p$) is assumed to be estimated or known a priori. The label y (available only for $\mathcal{L}$) typically guides grouping/separation of the data points in the subspace, specifically, for $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ it is preferred to have $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$ and $\mathbf{z}' = \mathbf{B}^\top \mathbf{x}'$ lie close to (far apart from) each other when certain similarity score measure $w(y, y')$ on $\mathcal{Y} \times \mathcal{Y}$ is large (resp., small).

2. Greedy semi-supervised subspace learning

We begin with formulating an optimization problem over the subspace projection matrix $\mathbf{B}$. First, the subspace needs to preserve the salient information in the original data as much as possible. This can be achieved by forcing the reconstruction of the data point $\mathbf{x}$, that is, $\mathbf{B}\mathbf{B}^\top \mathbf{x}$, to be close to $\mathbf{x}$. This is done for all available data (i.e., $\mathcal{L} \cup \mathcal{U}$), hence minimizing the reconstruction error, $\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{B}\mathbf{B}^\top \mathbf{x}_i\|^2$.

To reflect affinity/separation supervision in the labeled data, we penalize any inconsistency between the labels and the distances of data points within the subspace. We let w_{ij} be the similarity score between y_i and y_j ($i, j = 1, \ldots, l$). For discrete class labels, a usual choice is $w_{ij} = 1$ (-1) if $y_i = y_j$ ($y_i \neq y_j$), while for real-valued (vector) y, one typically defines $w_{ij} = \exp(-\kappa \|y_i - y_j\|^2)$ for some $\kappa > 0$. Then the label regularization for two labeled points $\mathbf{x}_i, \mathbf{x}_j$ ($\in \mathcal{L}$) can be expressed as $w_{ij}\|\mathbf{B}^\top \mathbf{x}_i - \mathbf{B}^\top \mathbf{x}_j\|^2$, which enforces equi-labeled points to lie in the vicinity, and vice versa.

The two objectives are combined as (using constant $\gamma \geq 0$):

$$\min_{\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{B}\mathbf{B}^\top \mathbf{x}_i\|^2 + \gamma \sum_{i,j=1}^{l} w_{ij} \|\mathbf{B}^\top \mathbf{x}_i - \mathbf{B}^\top \mathbf{x}_j\|^2. \tag{1}$$

The second term can be seen as an extension of the Laplacian eigenmap [10] to the label-driven affinity graph. Also, the first term is the well-known objective of the unsupervised PCA[2] whose solution coincides with the leading q eigenvectors of the sample covariance matrix, $(1/n)\sum_i \mathbf{x}_i \mathbf{x}_i^\top$ [6]. Hence having $\gamma = 0$ essentially reduces to the PCA.

Note that $\|\mathbf{B}^\top \mathbf{x}_i - \mathbf{B}^\top \mathbf{x}_j\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{B}\mathbf{B}^\top (\mathbf{x}_i - \mathbf{x}_j)$. We introduce a new matrix variable $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ imposing $\mathbf{A}$ to be symmetric positive semi-definite with rank q. Then (1) can be equivalently re-written as (letting $\mathbf{d}_{ij} \triangleq \mathbf{x}_i - \mathbf{x}_j$):

$$\begin{aligned} \min_{\mathbf{A}} \quad & \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{x}_i\|^2 + \gamma \sum_{i,j=1}^{l} w_{ij} \mathbf{d}_{ij}^\top \mathbf{A} \mathbf{d}_{ij} \\ \text{s.t.} \quad & \mathbf{A} \succcurlyeq 0, \ \text{rank}(\mathbf{A}) = q. \end{aligned} \tag{2}$$

So we do optimization over $\mathbf{A}$, and later $\mathbf{B}$ can be obtained by factorizing $\mathbf{A}$ (e.g., spectral decomposition[3]).

Although the objective becomes convex quadratic in $\mathbf{A}$, the rank constraint makes the optimization non-convex, which is difficult to solve in general. We consider a slight relaxation of (2) by introducing the so-called *nuclear norm*, $\|\mathbf{A}\|_n \triangleq \sum_k \sigma_k(\mathbf{A})$, where $\sigma_k(\mathbf{A})(>0)$ are the singular values of $\mathbf{A}$. As the number of the singular values equals the rank of $\mathbf{A}$, minimizing the nuclear norm has an effect of imposing the low-rank constraint. By denoting $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{D} = \sum_{ij} w_{ij} \mathbf{d}_{ij} \mathbf{d}_{ij}^\top$, the relaxed optimization is written as

$$\min_{\mathbf{A} \succcurlyeq 0} \|\mathbf{X} - \mathbf{A}\mathbf{X}\|_F^2 + \gamma \, \text{Tr}(\mathbf{D}\mathbf{A}) + \eta \|\mathbf{A}\|_n, \tag{3}$$

where $\text{Tr}(\cdot)$ is the trace. Note that estimating the subspace dimension q is translated into choosing the constant $\eta \geq 0$ properly.

Now, (3) is an instance of convex optimization, however, it is still challenging to solve due to non-differentiability of the nuclear norm. We address this issue by representing $\mathbf{A}$ as a conic combination of symmetric dyadic products, namely, $\mathbf{A} = \sum_k \theta_k \mathbf{u}_k \mathbf{u}_k^\top$ for $\theta_k > 0$ and $\|\mathbf{u}_k\| = 1$. This looks similar to spectral decomposition, however, we do not explicitly enforce orthogonality of $\{\mathbf{u}_k\}$. Then it follows that

$$\|\mathbf{A}\|_n = \left\| \sum_k \theta_k \mathbf{u}_k \mathbf{u}_k^\top \right\|_n \leq \sum_k \theta_k \|\mathbf{u}_k \mathbf{u}_k^\top\|_n = \sum_k \theta_k, \tag{4}$$

where we use the triangle inequality of the norm function and $\|\mathbf{u}_k \mathbf{u}_k^\top\|_n = 1$. We replace $\|\mathbf{A}\|_n$ by the upper bound $\sum_k \theta_k$.

For our representation of $\mathbf{A}$ as $\{(\theta_k, \mathbf{u}_k)\}$, we do optimization via infinite-dimensional greedy search motivated by [17]. The idea is,

[1] The samples in $\mathcal{U}$ are hence generated from the marginal $P(\mathbf{x})$.

[2] To be precise, orthonormal constraints ($\mathbf{B}^\top \mathbf{B} = \mathbf{I}$) need to be imposed.

[3] Hence, the subspace basis vectors $\mathbf{b}_j$ can be orthogonal, but not necessarily unit-norm. This can be considered as stretching out or shrinking some basis directions, which does not change the subspace itself.