Neurocomputing 000 (2018) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

# Robust jointly sparse embedding for dimensionality reduction

Zhihui Lai[a,*], Yudong Chen[a], Dongmei Mo[a], Jiajun Wen[a], Heng Kong[b]

[a] The College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] Shenzhen University General Hospital, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

*Article history:*
Received 13 August 2016
Revised 1 March 2018
Accepted 19 June 2018
Available online xxx

Communicated by Feiping Nie

*Keywords:*
Dimensionality reduction
Manifold learning
Robustness

ABSTRACT

As a famous linear manifold learning method, orthogonal neighborhood preserving projections (ONPP) is able to provide a set of orthogonal projections for dimensionality reduction. However, a problem of ONPP is that it takes the $L_2$-norm as the basic measurement and therefore tends to be sensitive to the outliers or the variations of the data. Aiming at strengthening the robustness of the conventional method ONPP, in this paper, a robust and sparse dimensionality reduction method based on linear reconstruction, called Robust Jointly Sparse Embedding (RJSE), is proposed by introducing $L_{2,1}$-norm as the basic measurement and regularization term. We design a simple iterative algorithm to obtain the optimal solution of the proposed robust and sparse dimensionality reduction model. Experiments on four benchmark data sets demonstrate the competitive performance of the proposed method compared with the state-of-the-art methods.



## 1. Introduction

Dimensionality reduction is a well-known data processing technique that can simultaneously reduce the redundant information and preserve important information from the high-dimensional data. The conventional dimensionality reduction methods including principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2–5] and sparse discriminant analysis (SDA) [6]. Unlike these methods, the non-linear manifold learning method, locally linear embedding (LLE) [7,8] can preserve the local geometry structure of the data set. The related techniques have been well studied in the fields of data mining and pattern recognition. To preserve the local geometry structure of the data with an explicit linear mapping, the linearization of the LLE and the other manifold learning methods were developed, among which the orthogonal neighborhood preserving projections (ONPP) [9] was one of the most well-known methods. ONPP is a linear approximation to the nonlinear method LLE. This scheme allows ONPP to preserve the manifold structure in a linear subspace. In [10], the authors proposed the sparse extension of ONPP. In [11,12], the authors further extended ONPP to process tensor data. What is more, [13] combined the supervised learning and manifold learning together for dimensionality reduction. Since these manifold learning methods showed promising performance in pattern recognition, they have been widely used in face recognition [14–16], disease classification [17], finger vein recognition [18], financial data analysis [19], clustering [20,21] and human age estimation [22].

However, the traditional manifold learning methods still have some drawbacks. First, these methods use the $L_2$-norm as measurement, which are sensitive to the outliers of data. Second, these methods lack the function of sparse feature selection for obtaining more reliable low-dimensional features. An effective way to enhance the robustness is to introduce an $L_{2,1}$-norm as the measurement instead of the $L_2$-norm in the traditional manifold learning methods. The $L_{2,1}$-norm metric has been a popular technique in recent years because of its competitive robustness to the outliers compared with the $L_2$-norm. Since the square operation does not need to be performed during the optimization of the $L_{2,1}$-nom-based model, the noise or reconstructive error will not be over-emphasized, which reduces the sensitiveness of the model. Therefore, many $L_{2,1}$-norm-based methods were proposed for improving the model's robustness including the robust feature selection [23–25], the robust PCA [26,27] and the robust classification [28]. In addition, the usage of the $L_{2,1}$-norm regularization is an effective way to obtain joint sparsity to improve the classification performance. In recent years, sparse learning has become more and more popular [29–33]. The so-called sparsity is to generate a sparse representation matrix or projection matrix so that most of the elements of the matrix become zero, thereby the main features of the images are emphasized and the feature extraction results are more explanatory.

In summary, the conventional methods, i.e., PCA, LDA and their extensions ignore the local geometry structure of the data set. And the manifold learning methods such as LLE and ONPP lack

* Corresponding author.
*E-mail addresses:* lai_zhi_hui@163.com (Z. Lai), dongmei_mo@qq.com (D. Mo).

https://doi.org/10.1016/j.neucom.2018.06.051
0925-2312/

of robustness or do not obtain joint sparsity for discriminative feature selection. Therefore, in this paper, we propose a robust and sparse subspace learning method by utilizing the $L_{2,1}$-norm as main metric and the regularization term to consider the local geometry structure information for robust sparse subspace learning. The proposed $L_{2,1}$-norm based model can be solved by a simple iterative algorithm.

The contributions of this paper are stated as follows:

(1) We construct a robust regression model based on ONPP and prove that the optimal solution space of the model is the same as the solution of ONPP in some special cases.
(2) By adding the $L_{2,1}$-norm as the regularization term, the joint sparsity can be easily obtained so as to improve the performance of feature selection. The proposed $L_{2,1}$-norm based model is proved to be more robust than other $L_1$-norm or $L_2$-norm based methods on four benchmark data sets.
(3) We provide the convergence proof of the proposed iterative algorithm. Extensive experimental results with figures and tables illustrate the effectiveness of our method.

The rest of the paper consists of five sections. In Section 2, we will introduce some notations used in this paper and then review LLE and its linear extension ONPP. In Section 3 we will introduce the regression form of ONPP. In addition, a robust and sparse subspace learning method, including the designed iterative algorithm and its convergence proof will be introduced in Section 4. Experimental results on four different data sets will be shown in Section 5. Section 6 summaries the paper.

## 2. Related work

In this section, we first introduce the notations and definitions used in this paper. Then we will review the non-linear manifold learning method LLE and the linearization one, i.e., ONPP.

### 2.1. Notations and definition of $L_{2,1}$-norm

We first introduce some notations used in this paper. All matrices are represented by the uppercase like $X$ and $Y$. Vectors are represented by the lowercase like $a$ and $p$. In this paper, for data matrix $X \in R^{m\times n}$, $m$ denotes the dimensionality of data points and $n$ denotes the number of data points. Besides, $d$ is the desired dimensionality of the low-dimensional feature vectors.

The $L_{2,1}$-norm plays an important role in jointly sparse feature selection. For a general matrix $X$, $x^i$ and $x_j$ denote its $i$th row and $j$th column, respectively.

The $L_{2,1}$-norm of matrix $X$ is defined as

$$\|X\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} X_{ij}^2} = \sum_{i=1}^{m} \left\|x^i\right\|_2$$

where $X_{ij}$ denotes the element in the $i$th row and $j$th column of $X$. The $L_{2,1}$-norm meets the three requirements of valid norm [23].

### 2.2. Locally linear embedding

LLE algorithm supposes every data points can be reconstructed by its neighbors and is divided into three steps. First, find out the $k$ nearest neighbor points of $x_i$ based on the Euclidean distance. Second, compute the weight matrix $W$ from the $k$ nearest neighbor points. The optimization problem is as follow.

$$\min_W \sum_i \left\|x_i - \sum\nolimits_{j\in C_k(x_i)} W_{ij} x_j\right\|^2 \text{s.t.} \sum_{j\in C_k(x_i)} W_{ij} = 1 \tag{1}$$

where $W_{ij}$ is the weight coefficient and $C_k(x_i)$ is index set of the $k$ nearest neighbor points of $x_i$.

Third, all the sample points $x_i$ are projected to a low-dimensional space $Y = [y_1, y_2, \ldots y_n]$. The optimization problem is as follow.

$$\min_Y \sum_i \left\|y_i - \sum\nolimits_{j\in C_k(x_i)} W_{ij} y_j\right\|^2 \text{s.t.} Y^T Y = I \tag{2}$$

where $I$ is identity matrix. Finally, the optimal solution of (2) is provided by the following eigenfunction.

$$(I-W)^T(I-W)y = \lambda y \tag{3}$$

The solution spaces of LLE are the eigenvectors according to the first $d$ minimal eigenvalues $\lambda$ [34].

### 2.3. Orthogonal neighborhood preserving projections

ONPP is the linearization of LLE. It adds a linear projection $p \in R^m$ based on LLE. After dimensionality reduction by linear projection, the sample point and its neighboring points have the minimum reconstruction error. Therefore, ONPP aims to solve the following optimization problem:

$$\min_p \sum_i \left\|p^T x_i - \sum\nolimits_{j\in C_k(x_i)} W_{ij} p^T x_j\right\|^2 \qquad \text{s.t. } p^T p = 1 \tag{4}$$

Similar to LLE, problem (4) can be transformed into an eigenvalue problem.

$$X(I-W)^T(I-W)X^T p = \tilde{\lambda} p \tag{5}$$

Then the solutions of $P = [p_1, p_2, \ldots p_d]$ are the eigenvectors corresponding to the $d$ minimal eigenvalues $\tilde{\lambda}$ [34].

## 3. $L_{2,1}$-norm based regression model

In this section, we propose a $L_{2,1}$-norm based regression model for linear reconstruction which is easy to be solved for obtaining the optimal solution. What's more, we discuss the connection between the proposed model and traditional ONPP.

### 3.1. Robust regression model

In order to increase the robustness of ONPP, we use $L_{2,1}$-norm as the measurement of the loss function. Besides, different from ONPP which minimizes the information loss in low-dimensional feature spaces, we aim to reduce the reconstruction error of different neighbors. We have the following model by using $L_{2,1}$-norm for increasing the robustness.

$$\min_a \left\|MX^T a a^T - MX^T\right\|_{2,1} \qquad \text{s.t. } a^T a = 1 \tag{6}$$

where $a \in R^m$ is the projection and $M = (I - W)$. For the optimization problem (6), we have the following Theorem 1.

**Theorem 1.** *The optimization problem (6) derives the same solution spaces as the one of ONPP if the diagonal elements of matrix D derived by the $L_{2,1}$-norm optimization problem are the same nonzero constant.*

**Proof.** We use an iteratively reweight algorithm to solve the $L_{2,1}$-norm based model as in [23]. From (6), we can derive

$$\begin{aligned}
&\left\|MX^T a a^T - MX^T\right\|_{2,1} \\
&= tr\left[\left(MX^T a a^T - MX^T\right)^T D \left(MX^T a a^T - MX^T\right)\right] \\
&= tr\left[a^T X M^T D M X^T a - a^T X M^T D M X^T a - a^T X M^T D M X^T a \right.\\
&\quad \left. + X M^T D M X^T\right]
\end{aligned} \tag{7}$$

where $D$ is a diagonal matrix and defined as

$$D_{ii} = \frac{1}{2\left\|h^i\right\|_2} \tag{8}$$