Expert Systems With Applications 93 (2018) 134–142

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Global and intrinsic geometric structure embedding for unsupervised feature selection

Yuan Wan, Xiaoli Chen*, Jinghui Zhang

School of Science, Wuhan University of Technology, Wuhan 430070, China

ARTICLE INFO

Article history:
Received 14 June 2017
Revised 14 September 2017
Accepted 3 October 2017
Available online 4 October 2017

Keywords:
Sparse preserve projection
Low rank representation
$l_{2,1/2}$-Matrix norm
Unsupervised feature selection

ABSTRACT

Dimensionality reduction becomes a significant problem due to the proliferation of high dimensional data. Sparse preserving projection (SPP) obtains the intrinsic geometric structure of the data, which contains natural discriminating information, and avoids the selection of parameters as well. However, SPP neglects the global structures since it computes the sparse representation of each data individually. Low rank representation (LRR), another commonly used dimensionality reduction method, finds the lowest rank representation of all data jointly, and is capable of capturing the global structures of data. Therefore in this paper, we propose a method, global and intrinsic geometric structure embedding for unsupervised feature selection (GGEFS), by constructing a low-rank-sparse graph. Our GGEFS method contains the loss of information, the preservation of structural information and the sparse regularization of projection matrix, on which we impose $l_{2,1/2}$-matrix norm to select sparser and discriminative features. An effective iterative algorithm based on Lagrange Multiplier method is described to solve GGEFS. Extensive experimental results demonstrate that the proposed algorithm outperform several state-of-the-art unsupervised feature selection methods.

1. Introduction

High dimension data is commonly encountered in many applicable fields, such as data mining (Agrawal et al., 1999), pattern recognition (Yu et al., 2001) and biomedical science (Clarke et al., 2008). Such kinds of data increase storage space, which is in need of well-performed hardware, and also introduce noise and redundancy. So dimensionality reduction becomes an urgent problem. Dimension reduction methods are of two main categories, feature selection and subspace learning. Feature selection is to select the most representative features from the original feature space under some certain criteria, and the collection of selected features is a subset of the original features. Different methods of subspace learning, however, aim to learn a transformation, which maps the original high dimensional feature space into a lower dimensional subspace, and thus new features are generated. A classical subspace learning method is Principal Component Analysis (PCA (Jiang et al., 2014)), which retains the variance of the data in the maximum extent to get the low dimensional representation of the data from a global perspective. Local structure of the data also contains important discriminative information (Bottou et al., 1992), therefore some dimensionality reduction algorithms by different methods of preserving local structure are proposed, such as Locality Preserving Projection (LPP (He & Niyogi, 2005)) and Local Linear Embedding (LLE (Roweis et al., 2001)).The core idea of these local structure preserving methods is embedding the neighborhood relationship, that is learned from the original data, into the lower dimensional subspace in different ways. Due to the high efficiency of local structure preservation of data, these methods are widely used in many feature selection methods (Cai, He, & Han et al, 2010; Zhou et al., 2016). Laplacian Score (He et al., 2005) selects features by evaluating features based on LPP, Li et al. (2008) propose discriminative locally linear embedding based on LLE. Wang et al.(2016) propose neighborhood embedding feature selection (NEFS), which learns the sparse representation by considering the nearest neighbors of each sample as a dictionary and then embeds the representation into the model of feature selection. For such graph-based locality preservation methods, there are still some challenges: (1) Using KNN to construct adjacency graph is not efficient enough to get discriminative information (Zhu, 2008). (2) The parameters of the neighborhood and heat kernel width are hard to set. (3) The eigen decomposition of dense matrix is time-consuming and in need of large storage. To address such challenges, Qiao et al. (2010) propose sparsity preserving projection (SPP) method, which aims to preserve the structure information by learning sparse reconstruction relationship between the original data, thus the intrinsic geometric structure of the original

* Corresponding author.
E-mail addresses: wanyuan@whut.edu.cn (Y. Wan), chenxiaoli@whut.edu.cn, 785535947@qq.com (X. Chen), Jingzhang@whut.edu.cn (J. Zhang).

https://doi.org/10.1016/j.eswa.2017.10.008
0957-4174/

data can be reflected, containing more natural discriminating information. And many SPP-based methods have been proposed (Lu et al., 2013; Wang et al., 2016). However, these SPP-based methods are not robust to the noise because the latent global structure of data is neglected. Low-rank representation (LRR (Liu et al., 2010)) is better at capturing the global structure of data by seeking the lowest rank representation among all the candidates and represents the data samples as linear combinations of the bases in a dictionary. Du et al. (2016) propose low rank sparse preserve projection for face recognition, which seeks the projective matrix by preserving both the global structure and locally linear structure of the data after constructing a low rank and sparse graph. As for the sparse regularization of projection matrix, being implemented by l_1-norm (lasso (Tibshirani, 2011)), although it is convenient to compute, it is not effective in selecting sufficient sparse features. Some researchers have extended it to l_p-norm $(0<p<1)$ (Foucart et al., 2009), and Xu et al. (2012) demonstrate that when p is 1/2, the performance of feature selection is the best. Due to the neglection of the correlationship between features, Nie et al. (2010) propose joint $l_{2,1}$-norm for feature selection and has been widely applied in many feature reduction methods (Zhou et al., 2016; Zhu et al., 2016). As l_p-norm can select sparser features than l_1-norm, Wang et al. (2013) propose $l_{2,p}$-matrix norm $(0<p<1)$ and empirically point out that when p equals 1/2, the regularization selects the sparsest and more robust features. However, less works based on $l_{2,p}$-matrix norm are proposed.

In the above research, we note that LRSPP owns better performance in structure learning. However, it lacks the measurement of information between the original data space and the learned subspace which is spanned by the selected features. Beside this, the using of$l_{2,1}$-norm fails to select sufficient sparse and discriminative features. To jointly address these two problems, we incorporate the loss of information, the embedding of low rank and sparse graph and $l_{2,1/2}$-matrix norm into a joint framework, named GGEFS, for dimensionality reduction. Now we state several characteristics of our algorithm as follow:

1. This approach considers both the information discrepancy between the original feature space and the lower dimensional subspace, which efficiently reduces the loss of information, and the structure preserving term is based on low rank sparse graph, which acquires adequate discriminative information and avoids problems of parameters selection.
2. We use $l_{2,1/2}$-matrix norm on the projection matrix, thus select sparser and discriminative features (Wang et al., 2013).
3. Lagrange Multiplier method is adopted to solve the optimization problem. Algorithm and convergence analysis in this paper are presented in Section 3.

The reminder of this paper is organized as follow. In Section 2, we present a generic select model and some background knowledge. In the following sections, our feature selection method is described in detail as well as the corresponding solution. Experimental results are reported in Section 4. And finally, we present our conclusion and the perspective of this work.

2. Related work

In this section, we briefly review the related research about our method, first we give a generic framework feature selection model, and low rank sparse representation is subsequently described.

2.1. A generic select model

Given n training samples$X=[x_1,x_2,..., x_n]\in R^{d\times n}$, each sample is located in d-dimensional space. A generic idea of feature selection is to consider three terms in a union model as:

$$\min_{W} Loss(X)+\alpha Loc(X)+\beta \mathrm{Reg}(W) \tag{1}$$

where α and β are regularization parameters to balance the local structural preservation and regularization, and W is the feature selection matrix. Implications of the three items are as follow:

1. The first term is the information discrepancy between the original data space and subspace spanned by the selected features, the usual measurement is Euclidean distance.
2. The second term is the information of structure preserving, which contains important discriminative information, and the usual methods are local linear embedding, linear preserve projection and so on.
3. The third term is sparse regularization, which controls the sparsity of projection matrix.

Group Lasso constructed with $l_{2,1}$-norm is used as the sparse regularization in recently research (Zhou et al., 2016). However, inspired by that l_p-norm $(0<p<1)$ is sparser than l_1-norm, Wang et al. (2013) extends the matrix norm to mixed$l_{2,p}$-norm and define

$$\|A\|_{2,p}=\left(\sum_{i=1}^{m}\left\|a^i\right\|_2^p\right)^{1/p} \tag{2}$$

where a^i is row vector of A. It's obvious that the noise magnitude of distant outlier with $l_{2,p}$-norm $(0<p<1)$ is less than that with $l_{2,1}$-norm, so $l_{2,p}$-norm based method is more robust. And Wang et al. empirically point that $l_{2,p}$-norm has the best performance in selecting sparse features when $p=0.5$.

2.2. Low rank sparse representation

Sparse representation aims to use fewer elements in dictionary to represent samples. Original data is usually treated as a dictionary to generate the sparse representation of each samplex_i, which can preserve intrinsic geometric information. Moreover, inspired by the property that low-rank can preserve the overall structural information, Du et al. (2016) integrates the above two ideas together to obtain the distribution of original data space. The formulation is given as:

$$\begin{aligned}&\min_{V,E}\|V\|_*+\lambda\|E\|_{2,1}+\gamma\|V\|_1\\&s.t.\ X=XV+E,\ diag(V)=0\end{aligned} \tag{3}$$

where V is the weight matrix reconstructed by dictionary X, l_1-norm is imposed on V for sparseness. $\|\bullet\|_*$ is nuclear norm, namely the sum of singular values of a matrix, and it is used to ensure its low rank property. E_i is the simulated noise matrix, λ and γ are non-negative equilibrium parameter. Since each x_i is excluded from reconstructing itself, the values of the diagonal elements of the matrix are restricted to 0.

3. Proposed method

In this section, we introduce GGEFS, which consist of the measurement of information discrepancy between the original data space and the lower dimensional subspace, structure preservation and sparse regularization of projection matrix, where the structure preserving term is shaped by embedding the weight matrix containing structure information of the data into the lower dimensional subspace. As follows, we describe our model of dimensionality reduction and the corresponding algorithm.