Multimed Tools Appl (2017) 76:17479–17495
DOI 10.1007/s11042-016-4119-2

Low-rank feature selection for multi-view regression

Rongyao Hu[1] · Debo Cheng[1] · Wei He[1] · Guoqiu Wen[1] · Yonghua Zhu[2] · Jilian Zhang[3] · Shichao Zhang[1]

Received: 1 July 2016 / Revised: 10 September 2016 / Accepted: 1 November 2016 /
Published online: 23 November 2016

Abstract Real life data and information often have different ways to obtain. For example, in computer vision, we can describe an objective by different types, such as text, video and picture. And even from variety of angles. These different descriptors of the same object are usually called multi-view data. In ordinarily, dimensional reduction methods usually include feature selection and subspace learning, respectively, can have better interpretative capability and stabilizing performance, and now are very prevalent method for high-dimensional data. However, it is usually not considering the relationship among class indicators, so the performance of regression model is not very ideal. In this paper, we simultaneously consider feature selection, low-rank selection, and subspace learning into a unified framework. Specifically, under the framework of linear regression model, we first use the low-rank constraint to feature selection which considers two aspects of information inherent in data. The low-rank constraint takes the correlation of response variables into account, then embed an $\ell_{2,p}$-norm regularizer to consider the correlation among variety of class indicators, and feature vectors and their corresponding response variables. Meanwhile, we take LDA algorithm which belong to the subspace learning to further adjust relevant feature selection results into account. Lastly, we conducted experiments on several real multi-views image sets and corresponding experimental consequences also validated the furnished method outperformed all comparison algorithms.

✉ Shichao Zhang
zhangsc@mailbox.gxnu.edu.cn

[1] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi, 541004, People's Republic of China

[2] Guangxi University, Nanning, Guangxi, 530004, People's Republic of China

[3] Guangxi University of Finance and Economics, Nanning, Guangxi, 530003, People's Republic of China

Keywords Feature selection · Subspace learning · Multi-view dataset · Low-rank selection · Sparse coding technology

1 Introduction

In the actual of field, such as pattern recognition and machine learning, it always describe different kinds of data with high-dimensional data, and homologous treatment process greatly increases the time complexity and space complexity [3, 34, 37]. Therefore, there are a number of ways to reduce the dimensions of data, and to find out a smaller and a more representative subset of features. And generally that is called dimensional reduction, meanwhile, which has been become an important research field of machine learning. Furthermore, the blessing of dimensional usually indicated that high-dimensional data usually have a low-dimensional structure. For this reason, a large number of dimensional reduction methods (including feature selection methods and subspace learning methods) have produced to search for the low-dimensional structure [15, 33]. Dimensional reduction is not only reducing the processing time and the storage structure, but also can get a learning model which has more compact structure and more generalization ability.

As usual, dimensional reduction methods are divided into feature selection methods [27, 28] and subspace learning methods [20, 23]. Recently, due to varieties and different methods have been flourished with many types of applications. Therefore, Feature selection methods have been widely used for reducing the dimensions of different types of data, and outputting a set of fundamental and representative features matrix [9, 32]. Based on the availability of class labels, the previous feature selection methods usually have been categorized into the following types, i.e., supervised feature selection methods [23, 40], semi-supervised feature selection methods[9, 20], and unsupervised feature selection methods [22, 31]. In spite of these methods directly selecting a subset of features lead to interpretation, and they can explain the relationship of the corresponding internal data, so the actual performance of feature selection methods appear unstable. Furthermore, subspace learning methods [23, 35] have been established to map all of features into a lower dimensional subspace, and removed irrelevant inherent attributes by some algorithm regulations. In this way, the high-dimensional problem can be solved and achieved stable performance. For instance, Zhu et al. [35] conducted subspace learning and converted original data into low-dimensional Hamming subspace, and then considered the correlations between the original space and the group effect of the features in training data. It is that the proposed model can be applied for the fast multimedia search, and also could receive interpretation ability and more stable performance. In addition, it can further adjust the results of the feature selection by the subspace learning that in order to consummate the presented new model.

Motivated by the successful application of low-rank for subspace clustering [2, 10, 26] and aimed to improve the single view model by multi-view method. During this article, we present a novel method called Sparse Low-Rank Feature Selection (shorted for SLR_FS) for multi-view regression. We simultaneously consider low-rank subspace selection and row-sparsity based feature selection into a unified framework for feature selection. Specifically, we use a low-rank constraint to conduct subspace selection by thinking about the correlation variables. Then a feature selection model with a $\ell_{2,p}$-norm regularizer is constructed to consider the correlation between features and their corresponding data samples. Furthermore, a new effective iterative optimization method is proposed to solve the corresponding objective function. And the novel optimization method enables the proposed method to be used in the different kinds large-scale data sets.

This paper get achievements for the following two aspects:

- This work conducts low-rank feature selection [6, 7, 39] for multi-view by considering two aspects of correlations inherent in data, such as the correlation between response variables via an $\ell_{2,p}$-norm regularizer [36, 41] and the correlation between any pairs of samples via a low-rank constraint. Their common goals are to identify relevant features and discard irrelevant features.
- This paper integrates supervised feature selection with subspace learning into a framework. With the goal of outputting a stable feature selection and interpreter ability. It has always been a challenge to embed the two different conceptual topics (i.e., feature selection and subspace learning) in a uniform framework for data mining [29, 30, 38]. Consequently, this article embeds a low-rank subspace learning into a novel devised feature selection model, where the row-sparsity structure ensures the multi-view data matrix to remove the irrelevant and noisy data. This naturally leads to the two conceptual topic interact with each other and ends up with a better feature selection results.

The remainder parts of this article are revealed such as: Section 2 introduces related work with remarkable framework for feature selection and Section 3 shows specific details of our proposed model and proposes an effective and reasonable optimization by an iteration algorithm, in order to ensure the global convergence. Then in Section 4 and Section 5, respectively, reasonable and effective experimental results are showed and detailed analysis, and concluded on the whole article.

2 Related work

Over the past decade, plenty of feature selection methods already are presented to overcome and address high-dimensional issue. Feature selection methods select a subset of features in accordance with some criteria, such as distinguishing features with good characteristics and correlating to the predefined goal. In this way, some methods roughly categorized into the following types, such as filter methods [19], wrapper methods [12, 21] and embedded [17, 25] methods.

Filter methods rank feature in accordance with the intrinsic property of data without any learning algorithms. Then to choose features with high scores for the remaining selection tasks. Therefore, the specific selection process of filter method that is distinguished from the other models process. Such as, Tabakhi et al. [19] presented to select the representative feature subset with a learned iterative algorithm. And the obtained new unsupervised feature selection algorithm is optimized by ant colony. Cao et al. [4] utilized the value of false discovery rates method to further decrease the influence of redundant genes. Meanwhile, it also could obtain the statistical significant.

Wrapper methods found out a learning algorithm to gain the accuracy of feature subsets to the predicting the target, and usually have better performance than filter methods. For example, Cathy et al. [12] proposed to consider feature selection between the clustering variables and irrelevant variables, and used the Gaussian mixture models to enhance the clustering effect. Unler et al. [21] presented to combine the filter model and the particle swarm optimization based wrapper model to feature subset selection that extended the filter-wrapper algorithm. Chyzhyk et al. [5] mixed an extreme learning algorithm and a genetic algorithm to explore the feature combination space for an optimal subset of features.

However, the wrapper method is more expensive of computation complexity than filter method.

Embedded methods usually make feature selection as part of the operation in the learning model, and utilize the objective function and other operations to get optimal feature subset. For instance, You et al. [25] proposed a novel embedded feature selection method which used the multi-label instead of the single label. In this way, it added a prediction risk criterion to evaluation of features for search of feature subset. Shi et al. [17] integrated the embedded learning with sparse regression into a unified framework to conduct an effectively sparse regression model.

3 Method

Here, we put this section divide into three subsections. We first introduce some notations used for this article and describe the proposed SLR_FS method in detail, in Section 3.1 and Section 3.2, respectively. Moreover, the optimization method to the objection function is given in Section 3.3.

3.1 Notations

Throughout this article, we use boldface uppercase letters to denote matrices, and utilize boldface lowercase letters to indicate vectors. Let $\mathbf{X} = \left[x_1^v, ..., x_i^v, ..., x_n^v\right] \in \mathbb{R}^{C^v \times n}$ be the data matrix of view v, and $\mathbf{Y} = [y_1, ..., y_k] \in \mathbb{R}^{n \times k}$ denote the normalized class indicator matrix. Also we show the Frobenius norm, $\ell_{2,p}$-norm of a matrix $\mathbf{X}$ respectively as $||\mathbf{X}||_F = \sqrt{\sum_i ||\mathbf{x}_i||_2^2} = \sqrt{\sum_j ||\mathbf{x}_j||_2^2}$ and $||\mathbf{X}||_{2,p} = \left(\sum_i \sqrt{\sum_j x_{ij}^2}\right)^{\frac{1}{p}}$. We further denote the inverse operator and the transpose of a matrix $\mathbf{X}$ as $\mathbf{X}^{-1}$ and $\mathbf{X}^T$, respectively.

3.2 Multi-view low-rank feature selection

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ denotes the signal view of training samples, where m and n, respective, denote the dimensionality of features and the number of samples. And given a class indicator matrix $\mathbf{Y} \in \mathbb{R}^{n \times k}$, where k denote the number of class labels. In general, we construct a linear relationship with the following formulation:

$$\min_{\mathbf{Z}} g(\mathbf{Z}) = f(\mathbf{Z}) + \lambda\phi(\mathbf{Z}) \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{m \times k}$ denotes the reconstruction coefficient matrix, $f(\mathbf{Z})$ denotes the loss term imposed on $\mathbf{Z}$, and $\phi(\mathbf{Z})$ denotes the regularization term and λ denotes a positive constant. However, in the literature [17], $f(\mathbf{Z})$ is defined as $f(\mathbf{Y} - \mathbf{X}^T\mathbf{Z})$ which aim at obtaining the minimum regression error between corresponding class indicator matrix $\mathbf{Y}$ and their prediction $\mathbf{X}^T\mathbf{Z}$. In this way, we defined the least loss function between feature matrix and class indicator matrix is formulated as:

$$\min_{\mathbf{Z}} f(\mathbf{Y} - \mathbf{X}^T\mathbf{Z}) + \lambda\phi(\mathbf{Z}) \tag{2}$$

Obviously, (2) considers the sample similarity among samples to conduct regression model on the whole method.

Therefore, to discuss separately, due to the multi-view linear regression and we can get a specific loss function, such as:

$$\min_{\mathbf{Z}^{\mathbf{v}}} ||\mathbf{Y} - \mathbf{X}_v^T \mathbf{Z}_v||_F^2 \tag{3}$$

where $||.||_F$ denotes the Frobenius matrix norm. Obviously, the optimization function in (3) is convex and smooth, so the optimal coefficient matrix $\mathbf{Z}_v$ is obtained as $\mathbf{Z}_{\mathbf{v}} = (\mathbf{X}_v \mathbf{X}_v^T)^{-1} \mathbf{X}_v \mathbf{Y}$. However, $\mathbf{X}_v \mathbf{Y}$ is not always invertible in practical applications. Meanwhile the matrix of $\mathbf{Z}_v$ does not make use of the fact that multiple responses are likely correlated. To do this, one way of taking advantages of possible correlation between response variables may impose a constraint on the rank of $\mathbf{Z}_v$, such as:

$$rank(\mathbf{Z}_v) = r \leq \min(m, k) \tag{4}$$

A directly intuition on (4) is that there is a number of linear constraints on regression coefficient $\mathbf{Z}_v$, and hence the estimation efficiency is improved and the number of effective number of parameters is reduced.

In real applications, there is an underlying connection between different kinds of things. In particular, this relationship may become more obvious in the multi-view problem. To find out these hidden internal structure information and utilize such corresponding structures to make the learning model better, the low-rank constraint on (4) is reasonable for finding the low-dimensional structure in high-dimensional data. Moreover, due to the literature [1] which points out that, when we conduct linear regression model in the projected LDA space which is called for low-rank linear regression. And consequently, in order to further adjusts the result of the features and finally forms a low-rank linear regression model with discriminant characteristics. In this way, a low-rank constrain structure is introduced on $\mathbf{Z}_v$ can naturally be expressed as a product of two r-rank matrices as follows:

$$\mathbf{Z}_v = \mathbf{A}_v \mathbf{B}_v \tag{5}$$

where $\mathbf{A}_v \in \mathbb{R}^{C^v \times r}$ and $\mathbf{B}_v \in \mathbb{R}^{r \times k}$. Then, for the fixed r, we can conduct a low-rank multi-view regression model as follows:

$$\min_{\mathbf{A}_v, \mathbf{B}_v} ||\mathbf{Y} - \mathbf{X}_v^T \mathbf{A}_v \mathbf{B}_v||_F^2 \tag{6}$$

where the matrix $\mathbf{A}_v$ and $\mathbf{B}_v$ have the low-rank constraint simultaneously. Thereinto, $\mathbf{A}_v$ denotes the one view of the multi-view structure to the original matrix $\mathbf{X}_v$, and $\mathbf{B}_v$ denotes the one view of the multi-view structure to the matrix of $\mathbf{Y}_v$. But here all $\mathbf{Y}_v$ are exactly the same so that we take $\mathbb{Y}$ instead of $\mathbf{Y}_v$. Then the reduced $\mathbf{X}_v^T \mathbf{A}_v \in \mathbb{R}^{n \times r}$, which is then multiplied by $\mathbf{B}_v$ to represent the feature matrix $\mathbf{X}_v$.

Inspired by the above mentions, due to the regularization term based on sparsity as we know that it has been widely utilized to exploit the correlation information of structures among different kinds of features. This inspired us that it can effectively discovery the correlated weight of coefficient matrix by sparse coding technology. Specifically, when embedding the sparsity-based term into the model, it will make the corresponding coefficient matrix (or called the weight matrix) shrink to zeros. Hence, we can select representative subset of original feature matrix, which are corresponding to those features with non-zero coefficients. In this way, we can remove those noisy and redundant features. Actually, there already are several of methods have been proposed based on sparsity regularization method and applied for all kinds of fields. Inspired by the novel idea, in this paper a novel sparsity model is conducted to address related problems. So a new $\ell_{2,p}$-norm method is presented to minimize. It is need to note that, the parameter of p can control the degree of features with correlated structures. Meanwhile, when p is lower, and different kinds of

features will shared more structures. So we, during this article, embed the $\ell_{2,p}$-norm regularization term into our model, then we re-write the (6) with a matrix representation as follows:

$$\min_{\mathbf{A}_v,\mathbf{B}_v} ||\mathbf{Y} - \mathbf{X}_v^T \mathbf{A}_v \mathbf{B}_v||_F^2 + \lambda_v ||\mathbf{A}_v \mathbf{B}_v||_{2,p} \tag{7}$$

where $\mathbf{X}_v \in \mathbb{R}^{C^v \times n}$ denotes the multi-view data matrix, and $\mathbf{Y} \in \mathbb{R}^{n \times k}$ denotes the normalized class indicator matrix. Through the parameter λ_v control the penalty residual value of view v. And the $\ell_{2,p}$-norm regularizer $||\mathbf{A}_v \mathbf{B}_v||_{2,p}$ can be utilized to penalize all of coefficients in the same row of $\mathbf{A}_v \mathbf{B}_v$, and take them for selection get together. Our proposed method combines feature selection with subspace learning together for a joint framework. The furnished method is through combined low-rank linear regression and sparse regularization for better select representative subset among features. Then embedding LDA subspace learning algorithm to further adjust the final results of feature selection.

3.3 Optimization

After introduced the details of proposed method, then we present a novel solution method to solve the objective function of (7). While the $\ell_{2,p}$-norm generally is utilized to find out the information of sparse structures, and it leads to the objective function usually unable to be solved with a appropriate closed form. Furthermore, the objective function also unable to make two variables convergence simultaneously, i.e., $\mathbf{A}_v$ and $\mathbf{B}_v$. Hence, we can solve this problem such as:

First of all, We have definition on the diagonal matrix $\mathbf{D}_v$, and in this matrix that each of element can be defined as:

$$d_{ii} = \frac{1}{(2/p)(||\mathbf{z}_v^i||_2)^{2-p}} \quad s.t.\ i = 1, 2, ..., m\ ,\ 0 < p < 2 \tag{8}$$

where $\mathbf{z}_v^i$ denotes the i-th of the $\mathbf{Z}_v = \mathbf{A}_v \mathbf{B}_v$ matrix. The parameter p usually control the degree of features with correlated structures.

Therefore, the objective formulation in (7) which is equivalent to:

$$\min_{\mathbf{A}_v,\mathbf{B}_v} ||\mathbf{Y} - \mathbf{X}_v^T \mathbf{A}_v B_v||_F^2 + \lambda_v Tr\left(\mathbf{B}_v^T \mathbf{A}_v^T \mathbf{D}_v \mathbf{A}_v \mathbf{B}_v\right) \tag{9}$$

Then let the (9) equal to $J(\mathbf{A}_v, \mathbf{B}_v)$ and calculate the partial derivative of the matrix $\mathbf{B}_v$. Meanwhile, let the result equation to zero.

$$\frac{\partial J(\mathbf{A}_v, \mathbf{B}_v)}{\partial \mathbf{B}_v} = -2\mathbf{A}_v^T \mathbf{X}_v \mathbf{Y}_v + 2\mathbf{A}_v^T \mathbf{X}_v \mathbf{X}_v^T \mathbf{A}_v \mathbf{B}_v + 2\lambda \mathbf{A}_v^T \mathbf{D}_v \mathbf{A}_v \mathbf{B}_v = 0 \tag{10}$$

From the above equation that we can obtain the following equation.

$$\mathbf{B}_v = \left(\mathbf{A}_v^T (\mathbf{X}_v \mathbf{X}_v^T + \lambda_v \mathbf{D}_v) \mathbf{A}_v\right)^{-1} \mathbf{A}_v^T \mathbf{X}_v \mathbf{Y} \tag{11}$$

Then we bring the matrix $\mathbf{B}_v$ expression into the (9) that can be obtained.

$$\max_{\mathbf{A}_v} Tr\left(\left(\mathbf{A}_v^T \left(\mathbf{X}_v \mathbf{X}_v^T + \lambda_v \mathbf{D}_v\right) \mathbf{A}_v\right)^{-1} \mathbf{A}_v^T \mathbf{X}_v \mathbf{Y}_v \mathbf{Y}_v^T \mathbf{X}_v^T \mathbf{A}_v\right) \tag{12}$$

Here we need to pay attention to that.

$$\mathbf{S}_t = \mathbf{X}_v \mathbf{X}_v^T, \qquad \mathbf{S}_b = \mathbf{X}_v \mathbf{Y}_v \mathbf{Y}_v^T \mathbf{X}_v^T \tag{13}$$

where the between-class and total-scatter matrices of data in the LDA are denoted with $\mathbf{S}_b$ and $\mathbf{S}_t$, respectively. Hence, the ultimately solution of (9) that can be re-written as:

$$\mathbf{A}_v = \arg\max_{\mathbf{A}_v} \left\{ Tr\left(\left(\mathbf{A}_v^T (\mathbf{S}_t + \lambda_v \mathbf{D}_v) \mathbf{A}_v \right)^{-1} \mathbf{A}_v^T \mathbf{S}_b \mathbf{A}_v \right) \right\} \tag{14}$$

The above mentioned is called the problem of LDA, and in order to achieve global optimization solution to (14) is that, through the top r eigenvectors of $\mathbf{S}_t^{-1}\mathbf{S}_b$ with respect to the nonzero eigenvalues.

Algorithm 1 The Pseudo code of solving (7)

Input: $\mathbf{X} \in \mathbb{R}^{C^v \times n}$, $\mathbf{Y} \in \mathbb{R}^{n \times k}$, r, λ_v, ρ;
Output: $\mathbf{A}_v \in \mathbb{R}^{C^v \times r}$, $\mathbf{B}_v \in \mathbb{R}^{r \times k}$;

1. Initialize $t = 0$;
2. Initialize $\mathbf{A}_v(0)$ and $\mathbf{B}_v(0)$ as two random matrices;
3. Initialize $\mathbf{D}_v^{(t)} = \mathbf{I} \in \mathbb{R}^{C^v \times C^v}$
4. **Repeat**
5. Update $\mathbf{A}_v(t+1)$ via (14);
6. Update $\mathbf{B}_v(t+1)$ via (11);
7. Update the diagonal matrix $\mathbf{D}_v(t+1)$, where the i-th diagonal element through the (8) to calculate. And $\mathbf{z}_v^i$ is denoted as $\mathbf{z}_v^i = [\mathbf{A}_v(t+1)\mathbf{B}_v(t+1)]^i$.
8. $t = t+1$;
9. **Until** The difference between the objective function value of (7) in two sequential iterations less than 10^{-5}.

From Algorithm 1, we can get the necessity optimal solution $\mathbf{A}_v$, and then to solve (7) by the equation of (9), which denotes the classical regression problem. Furthermore, since the objective formulation in (7) is nontrivial, and there are still two variables $\mathbf{A}_v$ and $\mathbf{B}_v$ need to optimize. By this way, it is difficult to solve the problem that the regularization term is non-smooth. For this reason, we will prove the proposed algorithm is convergent as follows.

Proposition 1 *Equation* (7) *will lead to convergent when Algorithm 1 monotonically decreases which corresponding to objective function.*

Proof It is assuming that, in the step of (t+1)-th iteration, and we get

$$< \mathbf{A}_v(t+1), \mathbf{B}_v(t+1) >= \arg\min_{\mathbf{A}_v, \mathbf{B}_v} ||\mathbf{Y}_v - \mathbf{X}_v^T \mathbf{A}_v \mathbf{B}_v||_F^2 + \lambda_v Tr\left(\mathbf{B}_v^T \mathbf{A}_v^T \mathbf{D}_v(t) \mathbf{A}_v \mathbf{B}_v \right) \tag{15}$$

In other words, in the step of (t+1)-th iteration, we can obtain the following inequality:

$$\begin{aligned} &||\mathbf{Y} - \mathbf{X}_v^T \mathbf{A}_v(t+1)\mathbf{B}_v(t+1)||_F^2 \\ &+\lambda_v Tr(\mathbf{B}_v(t+1)^T \mathbf{A}_v(t+1)^T \mathbf{D}_v(t) \mathbf{A}_v(t+1)\mathbf{B}_v(t+1)) \\ &\leq ||\mathbf{Y} - \mathbf{X}_v \mathbf{A}_v(t)\mathbf{B}_v(t)||_F^2 + \lambda_v\ Tr(\mathbf{B}_v(t)^T \mathbf{A}_v(t)^T \mathbf{D}_v(t) \mathbf{A}_v(t)\mathbf{B}_v(t)) \end{aligned} \tag{16}$$

Note that $\mathbf{Z}_v(t) = \mathbf{A}_v(t)\mathbf{B}_v(t)$ and $\mathbf{Z}_v(t+1) = \mathbf{A}_v(t+1)\mathbf{B}_v(t+1)$. Then we take the matrix $\mathbf{D}_v$ which definition in (8) into the inequality.

$$\begin{aligned} &||\mathbf{Y} - \mathbf{X}_v^T\mathbf{Z}_v(t+1)||_F^2 + \lambda_v \sum_{i=1}^{m} \frac{(||\mathbf{z}_v^i(t+1)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \\ &\leq ||\mathbf{Y} - \mathbf{X}_v^T\mathbf{Z}_v(t)||_F^2 + \lambda_v \sum_{i=1}^{m} \frac{(||\mathbf{z}_v^i(t)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \end{aligned} \tag{17}$$

where $\mathbf{z}_v^i(t)$ and $\mathbf{z}_v^i(t+1)$ denote the i-th row of the matrix $\mathbf{Z}_v(t)$ and $\mathbf{Z}_v(t+1)$, respectively. Due to every of i, we get

$$(||\mathbf{z}_v^i(t+1)||_2)^{2-p} - \frac{(||\mathbf{z}_v^i(t+1)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \leq (||\mathbf{z}_v^i(t)||_2)^{2-p} - \frac{(||\mathbf{z}_v^i(t)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \tag{18}$$

Thus, summarizing the above m inequalities and multiplying the summation with respect to the regularization parameter λ, we get:

$$\begin{aligned} &\lambda \sum_{i=1}^{m} (||\mathbf{z}_v^i(t+1)||_2)^{2-p} - \frac{(||\mathbf{z}_v^i(t+1)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \\ &\leq \lambda \sum_{i=1}^{m} (||\mathbf{z}_v^i(t)||_2)^{2-p} - \frac{(||\mathbf{z}_v^i(t)||_2)^{2(2-p)}}{(2/p)(||\mathbf{z}_v^i(t)||_2)^{2-p}} \end{aligned} \tag{19}$$

After combining (17) with (19) together, so we have as follows:

$$||\mathbf{Y} - \mathbf{X}_v^T\mathbf{Z}_v(t+1)||_F^2 + \lambda||\mathbf{Z}_v(t+1)||_{2,p} \leq ||\mathbf{Y} - \mathbf{X}_v^T\mathbf{Z}_v(t)||_F^2 + \lambda||\mathbf{Z}_v(t)||_{2,p} \tag{20}$$

Since $\mathbf{A}_v$ and $\mathbf{B}_v$ are updated according to the gradient, it demonstrates the presented method will monotonically decrease the corresponding objective function during every iteration. Moreover, our proposed method also will avoid to local convergence by rapid, and it will take converge incline to the global optimal solution. □

4 Experiments

We compare our method with one baseline method and six approaches over four multi-view data sets. Specifically, we first utilize each dimensionality reduction method to project original high-dimensional data into the lower dimensional subspace for each view, and to integrate each view into the $\mathbf{X}_v$ matrix to carry out the sparse learning and then to conduct regression with Support Vector Machine (SVM) via the LIBSVM toolbox.[1]

4.1 Experimental Setup

We have verification of the proposed SLR_FS method on four publicly available multi-view data sets. Specifically as follows:

– Tem-Dino [18]: It contains 675 images of 2 objects, one object has 312 images and the other has 363 images. Each model is executed more that 30 times by 2 models and is

[1] http://www.csie.ntu.edu.tw/cjlin/libsvm/.

recorded with 2 cameras to observe with different angles. And each image displayed in grayscale and deal with the same size of 640 × 480 pixels.
- tripod [13]: This dataset collects 2299 images for 20 sequences of cars. Each car is photographed more that 20 times and is recorded with 2 cameras to observe with different angles. All images are reshaped into 250 × 376 pixels in grayscale.
- rand_wiki [16]: This dataset includes 10 semantic of classes (e.g., Biology, History, Media and Music). Then a random split was used to produce 2149 documents with a training set, and 717 documents belong to a test set.
- imaxs [24]: This video dataset involves 12 different action classes (e.g., check watch, scratch head and cross arms). Thereinto, every action is complied 3 times with 12 actors and is observed the subjects by 5 cameras to record with different kinds of perspectives.

We also give more details of all data sets in Table 1.

We have compared our method with the following representative feature selection methods:

- NFS: Non Feature Selection(NFS) uses original variable to conduct classification by the SVM classifier directly. Here we use NFS to indicate if dimensionality reduction methods make sense in real applications.
- LDA: Linear Discriminant Analysis (LDA) [11], as one of subspace learning methods, preserves the neighborhood relation of each class sample to conduct subspace learning. Hence, LDA is a global subspace learning method.
- LR: Linear Regression (LR) [8] is a regression method which utilizes minimum square function to model analysis between one or more independent variables and dependent variables.
- LGR: Logistic Regression (LGR) [1] is an biased estimator regression for linear data analysis. Its character is able to give up the unbiased of the least square method.
- MI: The method of Mutual Information (MI) [14] takes the logarithmic compression and selects a mapping function to map processing for the original data. Finally, the important features were selected according to their scores.
- SD: This method of Statistical Dependency (SD) [14] selects a mapping function to process the original data. The resulting scores correspond to the level of important of the strength features. Then feature selection for the top-ranking scores.
- RSR: This unsupervised approach [31] chooses a representative response matrix through the self-representation method. Then embedded into the sparse learning model for feature selection. At the same time, the size of the coefficient matrix means that the importance of the strength of the corresponding features.

NFS does not conduct dimensionality reduction and can be regarded as the baseline of all dimensionality reduction methods, such as LDA, LR, LGR, MI, SD, RSR and our

Table 1 The details of data sets

Data sets	Instances	Targets	View	Types
Tem-Dino	675	2	2	Image, Model
tripod	2299	20	20	Image, Car
rand_wiki	128	2149	10	Text, Image
imaxs	180	2500	12	Human action

Table 2 The average correlation coefficient (aCC ± STD) for all of multi-view data sets. The best performance are emphasized by boldface in each column

	tripod	Tem-Dino	rand_wiki	imaxs
NFS	0.6249 ± 0.1631	0.7527 ± 0.0514	0.2741 ± 0.0245	0.7102 ± 0.0352
LDA	0.6302 ± 0.1420	0.7662 ± 0.0480	0.2402 ± 0.0267	0.6620 ± 0.0316
LR	0.6918 ± 0.1821	0.8010 ± 0.0468	0.2731 ± 0.0200	0.7272 ± 0.0353
LGR	0.7082 ± 0.1051	0.8462 ± 0.0422	0.2502 ± 0.0258	0.7195 ± 0.0353
MI	0.6526 ± 0.1240	0.8180 ± 0.0482	0.2960 ± 0.0313	0.7325 ± 0.0327
SD	0.6659 ± 0.1445	0.8162 ± 0.0469	0.2842 ± 0.0258	0.7401 ± 0.0271
RSR	0.6784 ± 0.1045	0.8023 ± 0.0429	0.2774 ± 0.0474	0.7455 ± 0.0298
SLR_FS	**0.7526 ± 0.0745**	**0.8962 ± 0.0379**	**0.2964 ± 0.0290**	**0.7469 ± 0.0327**

SLR_FS. In dimensionality reduction methods, LDA belong belong to subspace learning, while feature selection methods include LR, LGR and RSR. Meanwhile, both MI and SD belong to feature projection method. Our proposed method includes subspace learning, thus LDA can be regarded as the other baseline method. RSR does not consider subspace learning in the unsupervised feature selection and RSR method denoted the supervised general have better result than unsupervised method. In contrast, our method embedded a regularization term in the feature selection model and utilized the subspace learning method to adjust the finally selection results.

In the following experimental setting in [38] that we compared all methods with a 10-fold cross-validation. Specifically, we first partitioned the entire of original data set into 10 subsets by random. Then we used, at here, nine subsets for training and selected the remaining one subset for testing. Meanwhile, we have a repeat the entire of process with 10 times, in order to avoid the possible bias during original dataset partitioning for cross-validation. Lastly, the ultimate results were computed by averaging results for all experiments.

Due to all experiments, we will tune the parameters λ among the values $\lambda \in \{10^0, ...10^5\}$ to select the best parameters automatically, and we set $[c, g] \in \{2^{-5}, ...2^5\}$ in the SVM by a 5-fold inner cross-validation to distinguish different types of samples. Moreover, to select the best performance of a group and to return its coefficient as the coefficient final result with the 10-fold cross-validation of the model.

Table 3 The average root mean squared error (aRMSE ± STD) for all of multi-view data sets. The best performance are emphasized by boldface in each column

	tripod	Tem-Dino	rand_wiki	imaxs
NFS	3.3100 ± 0.6704	5.8552 ± 1.0935	0.1285 ± 0.0189	0.0427 ± 0.0511
LDA	3.4020 ± 0.6821	5.6820 ± 1.2001	0.1302 ± 0.0194	0.0762 ± 0.0440
LR	3.1500 ± 0.7028	5.8240 ± 1.2830	0.0685 ± 0.0216	0.0477 ± 0.0522
LGR	3.2502 ± 0.7639	4.8895 ± 1.0351	0.0802 ± 0.0231	0.0437 ± 0.0320
MI	3.0602 ± 0.6810	5.1125 ± 1.4802	0.0522 ± 0.0216	0.0580 ± 0.0410
SD	2.9422 ± 0.6620	5.3751 ± 1.3481	0.0629 ± 0.0264	0.0762 ± 0.0402
RSR	2.8540 ± 0.5216	5.1330 ± 1.1250	0.0290 ± 0.0088	0.0424 ± 0.0429
SLR_FS	**2.3093 ± 0.3761**	**4.6865 ± 1.0151**	**0.0191 ± 0.0096**	**0.0415 ± 0.0379**

In the section, we used average Correlation Coefficient (aCC) and average Root Mean Squared Error (aRMSE) as corresponding evaluation metric, and to evaluate the regression performance of all multi-view data sets and comparison methods.

We defined average Correlation Coefficient (aCC) as follows:

$$aCC = \frac{1}{n}\sum_{i=1}^{d} CC = \frac{1}{n}\sum_{i=1}^{d} \frac{\sum_{l}^{N_{test}} \left(y_i^{(l)} - \bar{y}_i\right)\left(\hat{y}_i^{(l)} - \bar{\hat{y}}_i\right)}{\sqrt{\sum_{l=1}^{N_{test}} \left(y_i^{(l)} - \bar{y}_i\right)^2 \sum_{l=1}^{N_{test}} \left(\hat{y}_i^{(l)} - \bar{\hat{y}}_i\right)^2}} \tag{21}$$

Then, we defined average Root Mean Squared Error (aRMSE) as follows:

$$aRMSE = \frac{1}{n}\sum_{i=1}^{d} RMSE = \frac{1}{n}\sum_{i=1}^{d} \sqrt{\frac{\sum_{l}^{N_{test}} \left(y_i^{(l)} - \hat{y}_i^{(l)}\right)^2}{N_{test}}} \tag{22}$$

where N_{test} means the size of test data set, then $\hat{y}^{(l)}$ and $y^{(l)}$ be the vectors of the predicted and actual targets for $x^{(l)}$, respectively. Besides, $\bar{\hat{y}}$ and $\bar{y}$ be the vectors of averages of the predicted and actual targets, respectively. A larger aCC shows better correlation coefficient results, while a smaller aRMSE means better robust.

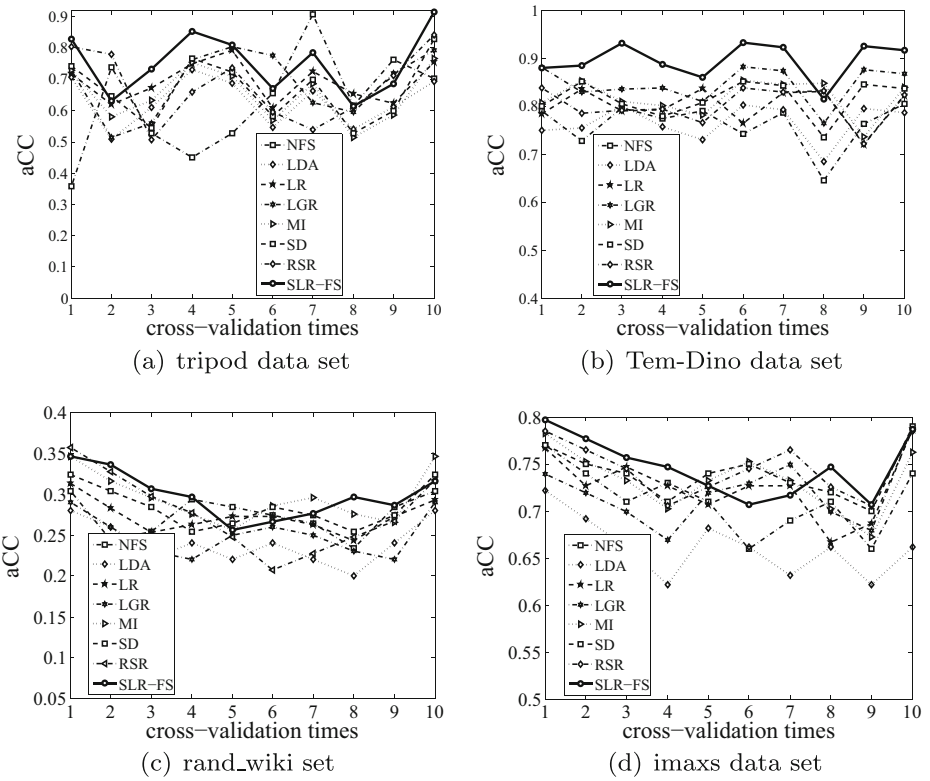

Fig. 1 The average correlation coefficient (aCC) of all methods for multi-view regression tasks

4.2 Regression experiment results

At here, we reported the result of average correlation coefficient (aCC) in Table 2 and average root mean squared error (aRMSE) in Table 3 for all multi-view data sets. We also indicated the results of correlation coefficient and root mean squared error of each fold in 10-fold cross-validation for all data sets in Figs. 1 and 2, respectively.

Due to Table 2, it demonstrates that the SLR_FS method has the best correlation coefficient. All methods get the lower result than other data sets since the rand_wiki data set has the variety of different forms aim to the same object. And it easier leads to describe the structure of same object has dissimilar data. However, it also denotes the proposed method combined low-rank constraint with sparsity learning, which could better remove the noisy and outlier data and select the more significant features. For instance, in the tripod data set, our method increased on 12.77 %, compared with the NFS which is used as the benchmark method, and increased on 4.44 %, compared with the LGR which got the second best performance among all methods. Then our method increased on 12.24 % and 10.00 %, 8.67 %, respectively, comparison with the subspace learning method and feature projection methods. Similarly, in the Tem-Dino data set, our method increased on 14.35 %, 13.00 %, 9.52 %,

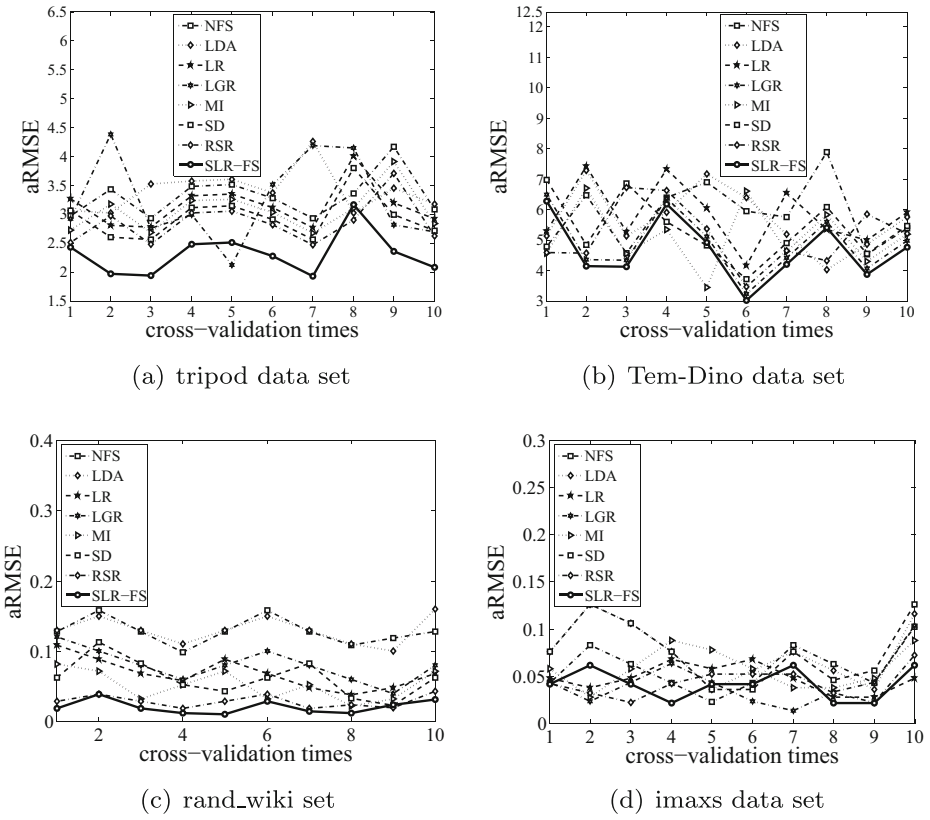

Fig. 2 The average root mean squared error (aRMSE) of of all methods for multi-view regression tasks

5.00 %, 7.82 %, 8.00 % and 9.39 %, respectively, compared with NFS, LDA, LR, LGR, MI, SD and RSR, in terms of average correlation coefficient.

Then analysis of Table 3, it shows that the SLR_FS method also has the minimum root mean squared error comparison with all comparison methods for multi-view regression tasks. With the same effect, our method has the minimum root mean squared error among comparison methods. For example, due to the least aRMSE results in the rand_wiki data that the proposed only get 1.91 %. Then directed to the imaxs data set, our method decreased on 0.12 %, 3.47 %, 0.62 %, 0.22 %, 1.65 %, 3.47 %, 0.09 %, respectively, compared with NFS, LDA, LR, LGR, MI, SD and RSR, in terms of average root mean squared error.

5 Conclusion

In this paper, we proposed a effective sparse low-rank method with feature selection to multi-view data sets. Specifically, the proposed method used a loss function based on linear model plus an $\ell_{2,p}$-norm sparse term to construct feature selection for achieving interpretation ability, and then employed LDA to conduct subspace learning for adjust the feature selection results. Experimental results on four multi-view data sets showed that the SLR-FS method outperformed the comparison method with two kinds of evaluation metric for the multi-view regression tasks.

In the future task, we will improve our model and apply for semi-supervised and unsupervised feature selection.

Acknowledgments This work was supported in part by the China "1000-Plan" National Distinguished Professorship; the Nation Natural Science Foundation of China (Grant No: 61263035, 61363009, 61573270 and 61672177), the China 973 Program (Grant No: 2013CB329404); the China Key Research Program (Grant No: 2016YFB1000905); the Guangxi Natural Science Foundation (Grant No: 2012GXNSFGA060004 and 2015GXNSFCB139011); the China Postdoctoral Science Foundation (Grant No: 2015M570837); the Innovation Project of Guangxi Graduate Education under grant YCSZ2016046; the Guangxi High Institutions' Program of Introducing 100 High-Level Overseas Talents; the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing; and the Guangxi "Bagui" Teams for Innovation and Research.

References

1. Cai X, Ding C, Nie F, Huang H (2013) On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1124–1132
2. Cao J, Wu Z, Wu J, Xiong H (2013) Sail: summation-based incremental learning for information-theoretic text clustering. IEEE Trans Cybern 43(2):570–584
3. Cao J, Wu Z, Wu J (2014) Scaling up cosine interesting pattern discovery: a depth-first method. Inf Sci 266(5):31–46
4. Cao Z, Wang Y, Sun Y, Du W, Liang Y (2015) A novel filter feature selection method for paired microarray expression data analysis. Int J Data Min Bioinforma 12(4):363–386
5. Chyzhyk D, Savio A, Graña M (2014) Evolutionary elm wrapper feature selection for alzheimer's disease cad on anatomical brain mri. Neurocomputing 128:73–80
6. Gao L, Song J, Nie F, Yan Y (2015a) Optimal graph learning with partial tags and multiple features for image and video annotation. In: CVPR
7. Gao L, Song J, Shao J, Zhu X, Shen H (2015b) Zero-shot image categorization by image correlation exploration. In: ICMR, pp 487–490

8. Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42(1):80–86
9. Liu H, Ma Z, Zhang S, Wu X (2015a) Penalized partial least square discriminant analysis with ℓ_1-norm for multi-label data. Pattern Recogn 48(5):1724–1733
10. Liu X, Guo T, He L, Yang X (2015b) A low-rank approximation-based transductive support tensor machine for semisupervised classification. IEEE Trans Image Process 24(6):1825–1838
11. Luo D, Ding CHQ, Huang H (2011) Linear discriminant analysis: new formulations and overfit analysis. In: Proceedings of the twenty-fifth AAAI conference on artificial intelligence, AAAI 2011, San Francisco, p 2011
12. Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with gaussian mixture models. Biometrics 65(3):701–709
13. Ozuysal M, Lepetit V, Fua P (2009) Pose estimation for category specific multiview object localization. In: IEEE conference on computer vision and pattern recognition, pp 778–785
14. Pohjalainen J, Rasanen O, Kadioglu S (2013) Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. Computer Speech & Language 29(1):145–171
15. Qin Y, Zhang S, Zhu X, Zhang J, Zhang C (2007) Semi-parametric optimization for missing data imputation. Appl Intell 27(1):79–88
16. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GRG, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: International conference on multimedia, pp 251–260
17. Shi X, Guo Z, Lai Z, Yang Y, Bao Z, Zhang D (2015) A framework of joint graph embedding and sparse regression for dimensionality reduction. IEEE Trans Image Process 24(4):1341–1355
18. Steven MS, Brian C, James D, Danniel S, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision & Pattern Recongnition, pp 519–528
19. Tabakhi S, Moradi P, Akhlaghian F (2014) An unsupervised feature selection algorithm based on ant colony optimization. Eng Appl Artif Intell 32(6):112–123
20. Tang Z, Zhang X, Li X, Zhang S (2016) Robust image hashing with ring partition and invariant vector distance. IEEE Trans Inf Forensics Secur 11(1):200–214
21. Unler A, Murat A, Chinnam RB (2011) mr 2 pso: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inf Sci 181(20):4625–4641
22. Wang D, Zhang H, Liu R, Liu X, Wang J (2016) Unsupervised feature selection through gram–schmidt orthogonalizationa word co-occurrence perspective. Neurocomputing 173:845–854
23. Wang T, Qin Z, Zhang S, Zhang C (2012) Cost-sensitive classification with inadequate labeled data. Inf Syst 37(5):508–516
24. Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3d exemplars. In: International conference on multimedia, pp 1–7
25. You M, Liu J, Li G, Chen Y (2012) Embedded feature selection for multi-label classification of music emotions. Int J Comput Intell Syst 5(4):668–678
26. Zhang C, Qin Y, Zhu X, Zhang J, Zhang S (2006) Clustering-based missing value imputation for data preprocessing. In: IEEE international conference on industrial informatics, pp 1081–1086
27. Zhang S (2012a) Decision tree classifiers sensitive to heterogeneous costs. J Syst Softw 85(4):771–779
28. Zhang S (2012b) Nearest neighbor selection for iteratively knn imputation. J Syst Softw 85(11):2541–2552
29. Zhang S, Cheng D, Zong M, Gao L (2016a) Self-representation nearest neighbor search for classification. Neurocomputing 195:137–142
30. Zhang S, Li X, Zong M, Cheng D, Gao L (2016b) Learning k for knn classification. In: ACM Transactions on Intelligent Systems and Technology, (Accepted)
31. Zhu P, Zuo W, Zhang L, Hu Q, Shiu SCK (2015) Unsupervised feature selection by regularized self-representation. Pattern Recogn 48(2):438–446
32. Zhu X, Zhang S, Zhang J, Zhang C (2007) Cost-sensitive imputing missing values with ordering. AAAI Press 2:1922–1923
33. Zhu X, Zhang S, Jin Z, Zhang Z, Xu Z (2011) Missing value estimation for mixed-attribute data sets. IEEE Trans Knowl Data Eng 23(1):110–121
34. Zhu X, Huang Z, Shen HT, Cheng J, Xu C (2012) Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recogn 45(8):3003–3016

35. Zhu X, Huang Z, Cheng H, Cui J, Shen HT (2013a) Sparse hashing for fast multimedia search. ACM Trans Inf Syst 31(2):9.1–9.24
36. Zhu X, Huang Z, Cui J, Shen HT (2013b) Video-to-shot tag propagation by graph sparse group lasso. IEEE Trans Multimed 15(3):633–646
37. Zhu X, Huang Z, Yang Y, Shen HT, Xu C, Luo J (2013c) Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recogn 46(1):215–229
38. Zhu X, Suk HI, Shen D (2014a) A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis. NeuroImage 100:91–105
39. Zhu X, Zhang L, Huang Z (2014b) A sparse embedding and least variance encoding approach to hashing. IEEE Trans Image Process 23(9):3737–3750
40. Zhu X, Li X, Zhang S (2016) Block-row sparse multiview multilabel learning for image classification 46(2):450–461
41. Zhu Y, Lucey S (2015) Convolutional sparse coding for trajectory reconstruction. IEEE Transactions on Pattern Analysis & Machine Intelligence 37(3):529–540

Rongyao Hu is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S., Guilin, Guangxi, P. R. China. Email: hu_No1@126.com.

Debo Cheng is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S., Guilin, Guangxi, P. R. China. Email: 15676209686@163.com.

Wei He is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S., Guilin, Guangxi, P. R. China. Email: risehnhew@163.com.

Guoqiu Wen is with College of Computer Science and Information Technology, Guangxi Normal University, Guiling, Guangxi, P.R. China. Her research interests include uncertainty mathematics, information system, and Decision making. E-mail: wenguoqiu2008@163.com.

Yonghua Zhu is with the Guangxi University, M. S., Nanning, Guangxi, P. R. China. Email: 1293234987@qq.com.

Jilian Zhang obtained his PhD degree in computer science at Singapore Management University. He is currently a professor with the School of Information and Statistics, Guangxi University of Finance and Economics, China. His research interests include data management, databases, query processing, and data privacy protection. He has published 26 papers on refereed journals and conferences, including IEEE TKDE, Information Systems, ACM SIGMOD, VLDB, IJCAI etc. He is a member of the ACM. E-mail: jilian.z.2007@phdis.smu.edu.sg.

Shichao Zhang is a Distinguished Professor and the director of Institute of School of Computer Science and Information Technology at the Guangxi Normal University, Guilin, P.R. China. He holds a Ph.D. degree in Computer Science from Deakin University, Australia. His research interests include data analysis and smart pattern discovery. He has published over 50 international journal papers and over 60 international conference papers. He has won over 10 nation-class grants, such as the China NSF, China 863 Program, China 973 Program, and Australia Large ARC. He is an Editor-in-Chief for International Journal of Information Quality and Computing, and is served as an associate editor for IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and IEEE Intelligent Informatics Bulletin. Email: zhangsc@mailbox.gxnu.edu.cn.