Multimedia Systems (2017) 23:285–291
DOI 10.1007/s00530-015-0487-0

SPECIAL ISSUE PAPER

# Feature selection by combining subspace learning with sparse representation

**Debo Cheng[1] · Shichao Zhang[1] · Xingyi Liu[2] · Ke Sun[1] · Ming Zong[1]**

Published online: 17 October 2015


**Abstract** A novel feature selection algorithm is designed for high-dimensional data classification. The relevant features are selected with the least square loss function and $\ell_{2,1}$ -norm regularization term if the minimum representation error rate between the features and labels is approached with respect to only these features. Taking into account both the local and global structures of data distribution with subspace learning, an efficient optimization algorithm is proposed to solve the joint objective function, so as to select the most representative features and noise-resistant features to enhance the performance of classification. Sets of experiments are conducted on benchmark datasets, show that the proposed approach is more effective and robust than existing feature selection algorithms.

**Keywords** Feature selection · High-dimensional data classification · Subspace learning · Joint objective function

## 1 Introduction

In real applications, data are often of high dimension, i.e., with a large number of features [18, 22, 37]. And most features are hardly with contributions to data mining tasks. This kind of features is often attributed to irrelevant and redundant features that can affect the classification performance. In other words, high-dimensional datasets not only consume huge storage and computation resources, but also degrade the total performance of a learning algorithm, i.e., so-called the curse of dimensionality [6, 31]. Therefore, feature selection has been regarded as an important and effective technique to weaken the affection of high dimension to classification performance, as well as to significantly improve the comprehensibility of discovered results [17, 21, 30].

Feature selection aims to maintain the original features of the data, at the same time, seeks to identify the main features and weed out the irrelevant ones, so as to build an impactful learning model [1, 11, 23]. Traditional feature selection methods generally evaluate the features one by one without respect to the relationship between features, the local structure of the data, the local structure of the data, and the label distribution of the data. To deal with these issues, sparse learning has been applied to feature selection algorithms. However, there are still two limitations in these feature selection methods based on sparse learning algorithms [19]. One is that the sparse penalty terms, such as $\ell_1$-norm and $\ell_{2,1}$-norm, are not enough to regulate the sparseness of feature representation. Namely, they may not select the optimum subset features from the original data [33, 36]. Another is that these methods only consider the global structures of the data, but neglects the local information of the data which can also well impact on selecting the significant features.

Motivated by the above-observed points, we design a novel framework for efficiently integrating the sparse model and subspace learning, so as to select the important features for a given dataset [16, 34]. First of all, the least square loss function is applied to measure the correlation between the feature set and the label set. Furthermore, the $\ell_{2,1}$-norm regularization term is employed to shrink the least square loss function, to generate a sparse model.

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

[1] Guangxi Key Lab of Multi-source Information Mining and Security, College of CS and IT, Guangxi Normal University, Guilin 541004, China

[2] Qinzhou University, Qinzhou 535000, China

Because this model only takes into account the global structure of the data, the local structure of the data is combined with manifold constraints during selecting the main relevant features. While the locality preserving projection (LPP) [5] is sensitive to outliers and noise features, the LPP is adopted to keep the local structure of the data. Finally, the Fisher's LDA [4] is advocated to protect the global structure of the data. According to these measures, the proposed framework can select a subset of important relevant features for dataset which efficiently reduces the dimensionality and speeds up the learning process.

The primary contributions of the study are summarized as follows:

- A novel framework is designed for feature selection for combining the subspace learning with the sparse graph representation. This framework can select the main discriminative features due to taking both the local and global structures of the data into account.
- A novel iterative optimization algorithm is exploited to solve the joint objective function and obtain the optimum solution. It can also testify the optimization algorithm which can quickly yet efficiently converge to the optimum solution.
- Compared with the generic dimensionality reduction algorithms, the proposed algorithm demonstrates the best performed results, especially on high-dimensional data.
- The rest of this paper is organized as follows: Sect. 2 summarizes some recently works related to feature selection. Section 3 designs the feature selection algorithm and its improvement. Section 4 experimentally evaluates the proposed approach and compared it with existing feature selection methods. Finally, this paper is concluded in Sect. 5.

## 2 Related works

This section briefly recalls the recent studies concerning both feature selection algorithm and sparse representation which have been successfully applied in high-dimensional data. Nie et al. employed joint $\ell_{2,1}$-norm minimization on both loss function and regularization to select the most relevant features. Moreover, they proposed an efficient method to solve the objective function of the feature selection [9]. With the graph-theoretic clustering technique, the original features are divided into different clusters. Features in different clusters are relatively independent. And then, the most representative features from each cluster are selected to makeup a subset of features. Based on the above two ideas, Song et al. proposed a fast feature selection algorithm [12]. Liu et al. proposed a novel feature selection algorithm that considered not only both maximal relevance to the class labels and minimal redundancy to the selected features, but also an agglomerative way to enhance the performance of pattern classification [6]. Qiao et al. advocated of preserving the sparse reconstructive structures of the data via minimizing a $\ell_1$-norm regularization-related objective function, called sparsity preserving projection (SPP). Moreover, SPP can automatically choose the neighborhood of the features by the natural discriminating information [10]. Wang et al. integrated multiple kernel learning, sparse coding and graph regularization for feature selection, and developed a novel data representation algorithm to optimize the feature selection objective function [13]. Hence, this algorithm could consider the local manifold structure of the data. Zhu et al. utilized the relationship between indicator vectors, as well as considers the canonical correlations between features of different modalities via projecting them into a canonical space [35]. The framework has successfully been applied in Alzheimer's disease diagnosis. Zhou et al. proposed to recover the true sparse centroid from the data and advocated the approximation set coding approach [24].

## 3 Approach

Throughout the paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its $i$th row and $j$th column are denoted as $x^i$ and $x_j$, respectively. We denote $\ell_2$-norm and $\ell_{2,1}$-norm as $\left\|x_j\right\|_2 = \sqrt{\sum_{i=1}^{n} x_{i,j}^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} x_{i,j}^2}$, respectively. We further denote the transpose operator, the trace operator and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $\mathrm{tr}(\mathbf{X})$ and $\mathbf{X}^{-1}$, respectively.

### 3.1 Proposed algorithm

In this section, we present an efficient supervised feature selection algorithm to improve the classification performance. Given $\mathbf{X} \in R^{p \times s}$, where $p$ is the number of feature variables and $s$ is the number of samples. The samples corresponding to label matrix is $\mathbf{Y} \in \mathbf{R}^{c \times s}$, where $c$ is the number of classes, i.e., 0–1 encoding. The proposed feature selection algorithm aims to seek a functional dependence between $\mathbf{X}$ and $\mathbf{Y}$. First of all, we use the linear representation method to describe the relation between the feature sets $\mathbf{X}$ and the indicator matrix $\mathbf{Y}$ [28, 34]. Moreover, the least squares method is a classical estimator in which the method is chosen to minimize the reconstruction error. The objective function of least square is defined as follows:

$$\mathbf{A} = \arg\min_{\mathbf{A}} \left\|\mathbf{A}^T\mathbf{X} - \mathbf{Y}\right\|_{2,1} \tag{1}$$

where $\mathbf{A} \in \mathbf{R}^{p\times c}$ denotes the coefficient representation matrix, namely, the relation between the feature sets and the labels of class. The solution $\mathbf{A}$ cannot respond to the information, features of which are main relevant features and the others are the irrelevant or low-relevant features. Moreover, a multi-task learning formula with a sparse least square regression model has been successfully applied for a binary classification [20, 29]. Therefore, we use the follow formula instead of Eq. (1).

$$\mathbf{A} = \arg\min_{\mathbf{A}} \left\|\mathbf{A}^T\mathbf{X} - \mathbf{Y}\right\|_{2,1} + \rho\|\mathbf{A}\|_{2,1} \tag{2}$$

where $\rho$ is a positive parameter and controls the sparsity. Therefore, we can assign a large weight to the main relevant features and a small or zero weight to the weak-relevant features via Eq. (2). Equation (2) has been proved efficient for binary classification [25, 26]. In this paper, we utilize the correlation of different classes via regarding each class as one task to make it deal with the multi-class classification. Nevertheless, it cannot guarantee to be selected which is conductive to better classification performance due to that Eq. (2) cannot ensure that the neighborhood structure of the selected features is preserved.

For the purpose of realizing multi-class classification, we consider the global and the local topological structures of the data according to the distribution of the data features into the proposed framework. Firstly, we employ a Fisher's LDA [4, 27], which considers the global data distribution based between within-class variance and between-class variance to find the main relevant class. But the Fisher's LDA penalize term $\frac{\mathbf{A}^T\sum_g\mathbf{A}}{\mathbf{A}^T\sum_h\mathbf{A}}$ is the non-convexity, where $\sum_g$ denotes the within-class variance and $\sum_h$ denotes the between-class variance. Fortunately, Ye [20] proposed to utilize a multivariate linear regression model that defines the class label matrix $\mathbf{Y} = [y_{i,k}]$ to replace the Fisher's LDA penalized term.

$$y_{i,k} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}} & \text{if } l(x_i) = k \\ -\sqrt{\frac{n_k}{n}} & \text{otherwise} \end{cases} \tag{3}$$

where $n_k$ denotes the sample size of the class $k$ and $l(x_i)$ is a class label of $x_i$. So we can efficiently utilize the global structure of the data via the class indicator matrix $\mathbf{Y}$, and cannot transform the original features space into a low-dimensional space. Secondly, we employ Locality preserving projection (LPP) [5], which is a subspace learning method to preserve the local structure of the features. The purpose of LPP is to seek an embedding space where the preserved local structure, i.e., maintain the local structure of the data. We briefly describe the definition of LPP as follows:

$$\min_{\mathbf{A}} \sum_{i,j} (y_i - y_j)^2 w_{i,j} \tag{4}$$

where the $y = \mathbf{A}^T x_i$ , $i = 1, 2, \ldots, p.$ and $w_{i,j}$ is a heat kernel $w_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\delta}\right)$, $\delta$ is a positive parameter. Substituting the $y = \mathbf{A}^T x_i$ into Eq. (2), we obtain:

$$\begin{aligned} &\sum_{i,j} (\mathbf{A}^T x_i - \mathbf{A}^T x_j)^2 w_{i,j} \\ &= \sum_i \mathbf{A}^T x_i d_{i,i} x_i{}^T \mathbf{A} - \sum_{ij} \mathbf{A}^T x_i w_{i,j} x_i{}^T \mathbf{A} \\ &= \mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A}) - \mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{A}) \\ &= \mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) \end{aligned} \tag{5}$$

where $\mathbf{D} = [d_{i,i} = \sum_j w_{i,j}]$ is a diagonal matrix, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is a Laplacian matrix. Therefore, we use the $\mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})$ term to describe the topological relation of the data [32]. Hence, we present an efficient supervised feature selection algorithm and the objective function as follows.

$$\arg\min_{\mathbf{A}} \left\|\mathbf{A}^T\mathbf{X} - \mathbf{Y}\right\|_{2,1} + \rho_1\mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) + \rho_2\|\mathbf{A}\|_{2,1} \tag{6}$$

where $\rho_1$ and $\rho_2$ are the positive tuning parameters, and the class indicator matrix $\mathbf{Y}$ is defined in Eq. (3). Therefore, according to the objective function, we can know that the proposed algorithm is the integration of subspace learning (i.e., LDA and LPP) and feature selection as a consolidated framework. In addition, we develop an optimal algorithm to solve the optimum solution of the objective function. The proposed optimal algorithm could very efficiently and quickly converge to the global optimum solution and the experiments also testify this.

### 3.2 Optimization

Note that Eq. (6) is a convex function and the last two terms are non-smooth, we cannot solve it straightforwardly. Therefore, we propose an efficient optimization algorithm to solve the objective function.

Denote $\mathbf{A}^T\mathbf{X} - \mathbf{Y} = \left[v^1, \ldots, v^s\right]$, and we also define the diagonal matrixes as $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{D}}$ with the $k$th diagonal element as $\tilde{\mathbf{D}}_i = \frac{1}{2||v^k||_2}$ and $\tilde{\mathbf{D}}_i = \frac{1}{2||a^k||_2}$, respectively. The objective function in Eq. (6) is equivalent to

$$\begin{aligned} &\arg\min_{\mathbf{A}} \mathrm{tr}((\mathbf{A}^T\mathbf{X} - \mathbf{Y})^T\tilde{\mathbf{D}}(\mathbf{A}^T\mathbf{X} - \mathbf{Y})) \\ &\quad + \rho_1\mathrm{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) + \rho_2\|\mathbf{A}\|_{2,1} \end{aligned} \tag{7}$$

Specifically, we take the derivative about each row $a_i (1 \le i \le s)$ and setting it to zero, we can obtain:

$$\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T a_i - \mathbf{X}\bar{\mathbf{D}}y^i + \rho_1 \mathbf{L}a_i + \rho_2 \tilde{\mathbf{D}}a_i = 0 \tag{8}$$

Thus, we can know the following formula:

$$a_i = (\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T + \rho_1 \mathbf{L} + \rho_2 \tilde{\mathbf{D}})^{-1}\mathbf{X}\bar{\mathbf{D}}y^i \tag{9}$$

We can know that the $\bar{\mathbf{D}}$ and $\widetilde{\mathbf{D}}$ are unknown and depend on $\mathbf{A}$. For the purpose of solving Eq. (9), we present an iteration algorithm as follows, the pseudo of the optimal algorithm is as shown in Algorithm 1.

**Algorithm 1:** The pseudo of the Algorithm 1.

**Input**: $\mathbf{X}$
**Output**: $a^{(t)} \in \mathbf{R}^{p\times c}$
1 Initialize $\mathbf{A}^1 \in \mathbf{R}^{p\times c}$, $t$=1;
2 **while** Eq.(7) is not converge **do**.
3 Compute the diagonal matrices $\bar{\mathbf{D}}_i$ and $\widetilde{\mathbf{D}}_i (1 \le i \le s)$.
4 For each $i(1 \le i \le s)$,
$a_i^{t+1} = (\mathbf{X}\bar{\mathbf{D}}_i^{(t)}\mathbf{X}^T + \rho_1\mathbf{L} + \rho_2\tilde{\mathbf{D}}_i^{(t)})^{-1}\mathbf{X}\bar{\mathbf{D}}_i^{(t)}y^i$.
5 $t$=$t$+1;
6 **end**

**Lemma 1** *For any nonzero vectors $a, b \in R^m$, the following inequality is always true.*

$$\|a\|_2 - \frac{\|a\|_2^2}{\|b\|_2} \le \|b\|_2 - \frac{\|b\|_2^2}{\|b\|_2} \tag{10}$$

*Proof* There is an obviously right inequality: $(\sqrt{l} - \sqrt{m})^2 \ge 0$, then we have

$$\begin{aligned}&(\sqrt{l} - \sqrt{m})^2 \ge 0 \Rightarrow l - 2\sqrt{lm} + m \ge 0\\ &\Rightarrow \sqrt{l} - \frac{l}{2\sqrt{m}} \le \frac{\sqrt{m}}{2} \Rightarrow \sqrt{l} - \frac{l}{2\sqrt{m}} \le \sqrt{m} - \frac{m}{2\sqrt{m}}\end{aligned} \tag{11}$$

And we utilize the vectors a *and* b *to replace the vectors* l *and* s *in Eq.* (11) respectively. Hence, we can arrive at Eq. (10).

**Theorem 1** *Algorithm 1 decreases the objective value in Eq.* (9) *in each of iterations.*

*Proof* In terms of Step 2 in Algorithm 1, we have

$$\begin{aligned}\mathbf{A}^{(\mathbf{t+1})} = \min_{\mathbf{A}} &\operatorname{tr}((\mathbf{A}^T\mathbf{X} - \mathbf{Y})^T\bar{\mathbf{D}}^{(t)}(\mathbf{A}^T\mathbf{X} - \mathbf{Y})) + \rho_1\mathbf{L}\\ &+ \rho_2 tr(\mathbf{A}^T\tilde{\mathbf{D}}^{(t)}\mathbf{A})\end{aligned} \tag{12}$$

Therefore, we have

$$\begin{aligned}&\operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t+1)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2\operatorname{tr}((\mathbf{A}^{(\mathbf{t+1})})^T\tilde{\mathbf{D}}^{(t+1)}\mathbf{A}^{(\mathbf{t+1})})\\ &\le \operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2 tr(\mathbf{A}^{(\mathbf{t})})^T\tilde{\mathbf{D}}^{(t)}\mathbf{A}^{(\mathbf{t})}\\ &\Rightarrow \operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t+1)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2\sum_{k=1}^{d}\left(\frac{||(a^{(t+1)})^k||_2^2}{2||(a^{(t)})^k||_2} - ||(a^{(t+1)})^k||_2 + ||(a^{(t+1)})^k||_2\right)\\ &\le \operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2\sum_{k=1}^{d}\left(\frac{||(a^{(t)})^k||_2^2}{2||(a^{(t)})^k||_2} - ||(a^{(t)})^k||_2 + ||(a^{(t)})^k||_2\right)\\ &\Rightarrow \operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t+1)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t+1})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2\sum_{k=1}^{d}||(a^{(t+1)})^k||_2\\ &\le \operatorname{tr}((\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)\bar{\mathbf{D}}^{(t)}(\mathbf{X}^T\mathbf{A}^{(\mathbf{t})} - \mathbf{Y}^T)^T) + \rho_1\mathbf{L}\\ &\quad + \rho_2\sum_{k=1}^{d}||(a^{(t)})^k||_2\end{aligned}$$

According to Lemma 1, the term $||a||_2 - \frac{||a||_2^2}{2||a_0||_2} \le ||a_0||_2 - \frac{||a_0||_2^2}{2||a_0||_2}$ is always true for any nonzero vectors $a$ and $a_0$. Therefore, the objective value can be decreased in each of iteration in Algorithm 1. Moreover, $\mathbf{A}^{(\mathbf{t})}$ and $\mathbf{D}_i^{(t)}$ will be satisfied when Eq. (12) converges. Thanks to Eq. (6) which is a convex function, so the matrix $\mathbf{A}$ is a global optimal solution of Eq. (6) when it satisfies Eq. (12). Thus, Algorithm 1 will converge to the global optimum of Eq. (6), and the experiment also demonstrates that the optimization method can quickly converge to the optimized solution.

## 4 Experiments

In this section, we adopt extensive experiments to test the performance of our feature selection algorithm for pattern classification. Furthermore, we compare our algorithm with the other dimension reduction methods from MATLAB toolbox.[1] The compared methods are used to getting the subset of attributions for pattern classification. The compared methods follow principal component analysis (PCA), multidimensional scaling (MDS), Sammon mapping (Sammon), Laplacian eigenmaps (Laplacian), diffusion maps (D-maps), kernel PCA (KPCA), stochastic neighbor embedding (SNE), symmetric stochastic neighbor embedding (SymSNE), t distributed stochastic neighbor embedding (tSNE), neighborhood components analysis (NCA), maximally collapsing metric learning (MCML). To evaluate the validity of the proposed algorithm, we employ the

[1] http://lvdmaaten.github.io/drtoolbox/.

**Table 1** Dataset description

| Dataset | Samples | Features | Classes |
|---|---|---|---|
| Arcene | 100 | 9920 | 2 |
| Breast cancer1 | 49 | 2166 | 2 |
| Breast cancer2 | 78 | 24,481 | 2 |
| Breast cancer3 | 295 | 24,496 | 2 |
| Breast cancer4 | 286 | 22,283 | 2 |
| Colon cancer | 62 | 2000 | 2 |
| GDS531 | 173 | 12,651 | 2 |
| GDS1027 | 154 | 26,923 | 4 |
| GDS1319 | 123 | 22,625 | 4 |
| GDS1454 | 180 | 54,613 | 4 |
| Ovarian cancer | 54 | 22,283 | 2 |
| Train | 168 | 147 | 9 |

SVM classifier from the LIBSVM toolbox[2] with the parameter spaces of $c$ and $g$ as $\{2^{-5}, \ldots, 2^{5}\}$. Moreover, we also utilize the original data to be classified via the SVM classifier as the baseline algorithms (Original for short).

### 4.1 Experiments verification and setup

Datasets in our experiments are detailed as follows:

Arcene and Train sets were downloaded from UCI[3] Breast cancer1, Breast cancer2, Breast cancer3, Breast cancer4 were obtained from works by West et al. [15], Laura J. van't Veer et al. [7], Mj van de Vijver et al. [8] and Wang et al. [14], respectively. GDS531, GDS1027, GDS1319, GDS1-454 are publicly available[4] Colon Cancer and Ovarian Cancer adopted came from the reports by Alon et al. [2] and Berchuck et al. [3], respectively.

Hence, we utilize the above twelve datasets to validate our feature selection algorithm. The detailed description of all datasets concerning samples, features and the number of classes is summarized in Table 1. The compared algorithm is also performed in the same experimental environment with the proposed algorithm. The original data are randomly divided into ten subsets. A single subset is regarded as the test sample and the rest subsets are used as training samples. Hence, we utilize tenfold cross validation in all experiments and repeat the whole process ten times to avoid the sample.

The parameters $\rho_1$ and $\rho_2$ are set to 0.001, 0.01, 0.1, 1, 10, 100 in Eq. (6). We are also setting the appropriate parameters for the other compared algorithms. The number of selected features is obtained by the maximum likelihood estimator (MLE) to select the optimum feature dimension for compared algorithms. We employ the classification accuracy to evaluate all the algorithms and report the best results of all algorithms using different optimal parameters. Finally, we summarize the results with the average value of 10 times plus or minus the standard deviation.

### 4.2 Experimental results

We summarize the classification accuracy results of all algorithms both binary classification and multi-class classification problems in the Tables 2 and 3. As shown in Tables 2 and 3, compared with the other algorithms, our proposed algorithm could yield the best performance. Our proposed algorithm had averagely improved at least 11.12 % of accuracy performance than simply classifying without any feature selection method (i.e., Original). Our proposed algorithm gained the best precision on gene dataset GDS1319, and improved about 34.18 % accuracy than SNE; moreover, the performance is best than the other algorithms. The reason is that our proposed algorithm takes into account the global and local structure of the original data space, the similarity of the data is preserved by our proposed framework during feature selection.

About compared algorithms, most of the results with feature selection algorithms are better than the results without feature selection. The reason is that feature selection algorithms not only wipe off the redundancy, noise samples and accelerated computing, but also obtained a better performance. However, there are some data sets with feature selection methods that obtain a worse performance than without feature selection algorithm, which is not surprising; many articles also said that feature selection algorithm could speed up computation and reduced storage cost.

The compared algorithms could be divided into two categories, i.e., manifold learning and others. Manifold learning contains MDS, Sammon, Laplacian, Dmaps et al. These algorithms recover the low-dimensional manifold structure from high-dimensional data, namely the high-dimensional space to find low-dimensional manifold, and to find the corresponding embedding maps. Analyzing these results, we can know that the proposed algorithm is best than these algorithms due to which these algorithms may ignore the global structure of the data.

Our algorithm has been successfully validated and can be applied in binary-class and multi-class classification problems. For example, the proposed algorithm can obtain a best performance than Original, PCA, MDS, Sammon, Laplacian, DMaps, KPCA, SNE, SymSNE, tSNE, NCA, MCML on multi-class datasets such as GDS1027, GDS1319, GDS1454, Train. And the performance of the proposed algorithm is outstanding. Moreover, our proposed algorithm could obtain some more stable results than the other compared algorithms.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

[3] http://archive.ics.uci.edu/ml/.

[4] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser.

**Table 2** Experimental results

| Method | Arcene | Breast cancer 1 | Breast cancer 2 | Breast cancer 3 | Breast cancer 4 | Colon cancer |
|---|---|---|---|---|---|---|
| Original | 0.7533 ± 3.4e−3 | 0.8932 ± 4.6e−3 | 0.9217 ± 1.8e−3 | 0.7399 ± 1.4e−3 | 0.7348 ± 1.4e−3 | 0.9010 ± 9.1e−5 |
| PCA | 0.8330 ± 9.2e−3 | 0.9183 ± 1.2e−2 | 0.9482 ± 4.5e−3 | 0.7407 ± 2.1e−3 | 0.7425 ± 2.1e−3 | 0.9174 ± 4.5e−4 |
| MDS | 0.8541 ± 8.6e−3 | 0.9233 ± 1.3e−2 | 0.9357 ± 4.3e−3 | 0.7607 ± 2.5e−3 | 0.7563 ± 2.2e−3 | 0.9307 ± 1.3e−3 |
| Sammon | 0.8503 ± 5.2e−3 | 0.9142 ± 1.2e−2 | 0.9228 ± 4.5e−3 | 0.7543 ± 2.8e−3 | 0.7605 ± 2.8e−3 | 0.9121 ± 7.2e−4 |
| Laplacian | 0.7928 ± 7.3e−3 | 0.9083 ± 9.9e−3 | 0.9273 ± 8.8e−3 | 0.7418 ± 2.0e−3 | 0.7655 ± 1.9e−3 | 0.9045 ± 1.2e−4 |
| DMaps | 0.7562 ± 6.7e−3 | 0.9533 ± 2.7e−3 | 0.9403 ± 4.0e−3 | 0.7544 ± 4.4e−3 | 0.7390 ± 4.3e−3 | 0.9176 ± 6.2e−3 |
| KPCA | 0.8441 ± 1.1e−2 | 0.9183 ± 1.1e−2 | 0.9482 ± 4.5e−3 | 0.7544 ± 1.2e−3 | 0.7416 ± 1.2e−3 | 0.9281 ± 1.0e−3 |
| SNE | 0.7832 ± 1.6e−2 | 0.9350 ± 1.1e−2 | 0.8496 ± 1.3e−2 | 0.7438 ± 2.0e−3 | 0.7380 ± 2.0e−3 | 0.9138 ± 6.7e−4 |
| SymSNE | 0.7921 ± 7.2e−3 | 0.8933 ± 1.3e−2 | 0.6952 ± 1.4e−2 | 0.7667 ± 7.9e−3 | 0.7371 ± 7.8e−3 | 0.9160 ± 1.1e−3 |
| tSNE | 0.8732 ± 3.7e−3 | 0.9183 ± 1.2e−2 | 0.9464 ± 4.8e−3 | 0.7432 ± 8.0e−3 | 0.7452 ± 8.0e−3 | 0.9141 ± 4.7e−4 |
| NCA | 0.8432 ± 6.6e−3 | 0.9183 ± 1.2e−2 | 0.9260 ± 4.1e−3 | 0.7491 ± 1.9e−3 | 0.7495 ± 1.9e−3 | 0.9131 ± 9.4e−4 |
| MCML | 0.8534 ± 9.1e−3 | 0.9267 ± 9.1e−3 | 0.9228 ± 4.5e−3 | 0.7708 ± 4.3e−3 | 0.6554 ± 4.3e−3 | 0.9305 ± 1.3e−3 |
| **Proposed** | **0.9716 ± 3.5e−3** | **0.9577 ± 4.1e−3** | **0.9707 ± 1.3e−3** | **0.9080 ± 2.2e−4** | **0.9668 ± 2.2e−4** | **0.9669 ± 1.2e−4** |

**Table 3** Experimental results

| Method | GDS531 | GDS1027 | GDS1319 | GDS1454 | Ovarian cancer | Train |
|---|---|---|---|---|---|---|
| Original | 0.8197 ± 1.8e−4 | 0.8237 ± 2.2e−3 | 0.9399 ± 8.4e−4 | 0.8189 ± 1.9e−4 | 0.8773 ± 5.6e−4 | 0.8314 ± 1.6e−4 |
| PCA | 0.8193 ± 8.2e−4 | 0.8370 ± 8.7e−3 | 0.9750 ± 3.2e−3 | 0.9010 ± 2.4e−3 | 0.8933 ± 8.6e−3 | 0.8327 ± 1.1e−4 |
| MDS | 0.7923 ± 5.9e−4 | 0.8667 ± 4.4e−3 | 0.9755 ± 3.1e−3 | 0.9173 ± 1.2e−3 | 0.8900 ± 9.1e−3 | 0.8394 ± 2.1e−4 |
| Sammon | 0.8141 ± 6.1e−3 | 0.8593 ± 4.4e−3 | 0.9500 ± 1.8e−3 | 0.9307 ± 2.2e−3 | 0.8977 ± 8.2e−3 | 0.8413 ± 5.4e−4 |
| Laplacian | 0.8211 ± 6.9e−5 | 0.8667 ± 5.4e−3 | 0.9442 ± 1.5e−3 | 0.9121 ± 1.4e−3 | 0.8967 ± 8.0e−3 | 0.8400 ± 1.0e−4 |
| DMaps | 0.8214 ± 1.9e−3 | 0.8845 ± 4.4e−3 | 0.9434 ± 2.9e−3 | 0.9045 ± 2.3e−3 | 0.8967 ± 1.4e−3 | 0.8383 ± 2.3e−4 |
| KPCA | 0.8343 ± 5.1e−3 | 0.8993 ± 3.4e−3 | 0.9820 ± 1.2e−3 | 0.9176 ± 3.0e−3 | 0.9000 ± 1.3e−2 | 0.8439 ± 1.9e−4 |
| SNE | 0.8284 ± 2.2e−3 | 0.8444 ± 5.4e−3 | 0.6413 ± 6.4e−3 | 0.9281 ± 2.2e−3 | 0.8833 ± 4.8e−3 | 0.8388 ± 7.4e−5 |
| SymSNE | 0.8204 ± 2.0e−4 | 0.8741 ± 6.0e−3 | 0.6565 ± 9.3e−3 | 0.9138 ± 1.0e−3 | 0.8863 ± 2.6e−3 | 0.8403 ± 1.7e−4 |
| tSNE | 0.8402 ± 2.8e−3 | 0.8519 ± 5.0e−3 | 0.9417 ± 4.7e−3 | 0.9160 ± 8.4e−4 | 0.8800 ± 7.0e−3 | 0.8498 ± 5.4e−4 |
| NCA | 0.8214 ± 1.7e−3 | 0.8704 ± 3.8e−3 | 0.9404 ± 1.1e−3 | 0.9141 ± 2.4e−3 | 0.9177 ± 2.8e−3 | 0.8470 ± 3.8e−4 |
| MCML | 0.8396 ± 9.6e−4 | 0.8733 ± 6.1e−3 | 0.9403 ± 2.7e−3 | 0.9131 ± 2.3e−3 | 0.9100 ± 9.1e−3 | 0.8510 ± 8.1e−4 |
| **Proposed** | **0.9272 ± 7.9e−5** | **0.9356 ± 1.1e−3** | **0.9831 ± 2.1e−4** | **0.9305 ± 8.9e−4** | **0.9830 ± 3.4e−4** | **0.8882 ± 2.4e−4** |

## 5 Conclusions

In this paper, we have formulated a novel feature selection algorithm and focused on integration the subspace learning and feature selection method as a unified framework. The proposed algorithm is called as sparse graph representation feature selection for supervised classification. The proposed algorithm has two advances compared with the state-of-the-art methods: (1) the framework is based on sparse learning model and use an efficient optimization algorithm to solve it; (2) take into account both global and local structures of the data. The proposed feature selection can efficiently select the important features and apply the selected features for SVM classification, the performance is much better than other compared algorithms. Therefore, we could validate the performance of the proposed algorithm by analyzing the classification accuracies of both binary classification and multi-class classification problems.

**Acknowledgments** This work is supported in part by the China "1000-Plan" National Distinguished Professorship; the China 973 Program under Grant 2013CB329404; the Natural Science Foundation of China under Grants 61170131, 61450001, 61363009, 61263035 and 61573270; the China Postdoctoral Science Foundation under Grant 2015M570837; the Guangxi Natural Science Foundation (Grant No: 2015GXNSFCB139011); the funding of Guangxi "100-Plan"; the Guangxi Natural Science Foundation for Teams of Innovation and Research under Grant 2012GXNSFGA060004; and the Guangxi "Bagui" Teams for Innovation and Research; Innovation Project of Guangxi Graduate Education YCSZ2015095, YCSZ2015096.

## References


1. Zhu, X., Suk, H., Lee, S., Shen, D.: Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. IEEE Trans. Biomed. Eng. **PP**(99), 1 (2015)
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. **96**(12), 6745–6750 (1999)
3. Berchuck, A., Iversen, E.S., Lancaster, J.M., Pittman, J., Luo, J., Lee, P., Murphy, S., Dressman, H.K., Febbo, P.G., West, M.: Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. **11**(10), 3686–3696 (2005)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience, New York (2001)
5. He, X., Niyogi, P.: Locality preserving projections. In: NIPS, pp. 153–160 (2003)
6. Liu, H., Wu, X., Zhang, S.: A new supervised feature selection method for pattern classification. Comput. Intell. **30**(2), 342–361 (2014)
7. Lj, V.T.V., Dai, H., Mj, V.D.V., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van, D.K.K., Marton, M.J., Witteveen, A.T.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871), 530–536 (2002)
8. Mj, V.D.V., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., D. W. Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J.: A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. **347**(25), 1999–2009 (2002)
9. Nie, F., Huang, H., Cai, X., Ding, C.H.Q.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Advances in Neural Information Processing Systems, pp. 1813–1821 (2010)
10. Qiao, L., Chen, S., Tan, X.: Sparsity preserving projections with applications to face recognition. Pattern Recognit. **43**(1), 331–341 (2010)
11. Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C.: Semi-parametric optimization for missing data imputation. Appl. Intell. **27**(1), 79–88 (2007)
12. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. Knowl. Data Eng. **25**(1), 1–14 (2013)
13. Wang, J.J., Bensmail, H., Gao, X.: Feature selection and multi-kernel learning for sparse representation on a manifold. Neural Netw. Off. J. Int. Neural Netw. Soc. **51c**(3), 9C16 (2013)
14. Wang, Y., Klijn, J.G., Yi, Z., Sieuwerts, A.M., Look, M.P., Fei, Y., Talantov, D., Timmermans, M., Gelder, M.V., Yu, J.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**(9460), 671C679 (2005)
15. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. **98**(20), 11462–11467 (2001)
16. Wu, X., Zhang, C., Zhang, S.: Efficient mining of both positive and negative association rules. ACM Trans. Inf. Syst. (TOIS) **22**(3), 381–405 (2004)
17. Wu, X., Zhang, C., Zhang, S.: Database classification for multi-database mining. Inf. Syst. **30**(1), 71–88 (2005)
18. Wu, X., Zhang, S.: Synthesizing high-frequency rules from different data sources. IEEE Trans. Knowl. Data Eng. **15**(2), 353–367 (2003)
19. Yan, Y., Shen, H., Liu, G., Ma, Z., Gao, C., Sebe, N.: Glocal tells you more: Coupling glocal structural for feature selection with sparsity for image and video classification. Comput. Vis. Image Underst.124, 99–109 (2014)
20. Ye, J.: Least squares linear discriminant analysis. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1087–1093 (2007)
21. Zhang, S., Qin, Z., Ling, C.X., Sheng, S.: "Missing is useful": missing values in cost-sensitive decision trees. IEEE Trans. Knowl. Data Eng. **17**(12), 1689–1693 (2005)
22. Zhang, S., Zhang, C., Yan, X.: Post-mining: maintenance of association rules by weighting. Inf. Syst. **28**(7), 691–707 (2003)
23. Zhao, Y., Zhang, S.: Generalized dimension-reduction framework for recent-biased time series analysis. IEEE Trans. Knowl. Data Eng. **18**(2), 231–244 (2006)
24. Zhou, G., Geman, S., Buhmann, J.M.: Sparse feature selection by information theory. In: 2014 IEEE International Symposium on Information Theory (ISIT), pp. 926–930 (2014)
25. Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.T.: Sparse hashing for fast multimedia search. ACM Trans. Inf. Syst. (TOIS) **31**(2), 9 (2013)
26. Zhu, X., Huang, Z., Cui, J., Shen, H.T.: Video-to-shot tag propagation by graph sparse group lasso. IEEE Trans. Multimed. **15**(3), 633–646 (2013)
27. Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C.: Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognit. **45**(8), 3003–3016 (2012)
28. Zhu, X., Huang, Z., Shen, H.T., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 143–152 (2013)
29. Zhu, X., Huang, Z., Yang, Y., Shen, H.T., Xu, C., Luo, J.: Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognit. **46**(1), 215–229 (2013)
30. Zhu, X., Li, X., Zhang, S.: Block-row sparse multiview multilabel learning for image classification. IEEE Trans. Cybern. (2015)
31. Zhu, X., Suk, H., Lee, S., Shen, D.: Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain Imaging Behav. 1–11 (2015). doi:10.1007/s11682-015-9430-4
32. Zhu, X., Suk, H.-I., Shen, D.: Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3089–3096 (2014)
33. Zhu, X., Suk, H.-I., Shen, D.: Multi-modality canonical feature selection for Alzheimer's disease diagnosis. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014, pp. 162–169 (2014)
34. Zhu, X., Suk, H.-I., Shen, D.: A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis. NeuroImage **100**, 91–105 (2014)
35. Zhu, X., Suk, H.-I., Shen, D.: Sparse discriminative feature selection for multi-class Alzheimer's disease classification In: Machine Learning in Medical Imaging, pp. 157–164 (2014)
36. Zhu, X., Zhang, L., Huang, Z.: A sparse embedding and least variance encoding approach to hashing. IEEE Trans. Image Process. **23**(9), 3737–3750 (2014)
37. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z.: Missing value estimation for mixed-attribute data sets. IEEE Trans. Knowl. Data Eng. **23**(1), 110–121 (2011)