*Computers, Materials & Continua* CMC, vol.60, no.2, pp.741-756, 2019

# Robust Re-Weighted Multi-View Feature Selection

**Yiming Xue[1], Nan Wang[2], Yan Niu[1], Ping Zhong[2,*], Shaozhang Niu[3] and Yuntao Song[4]**

**Abstract:** In practical application, many objects are described by multi-view features because multiple views can provide a more informative representation than the single view. When dealing with the multi-view data, the high dimensionality is often an obstacle as it can bring the expensive time consumption and an increased chance of over-fitting. So how to identify the relevant views and features is an important issue. The matrix-based multi-view feature selection that can integrate multiple views to select relevant feature subset has aroused widely concern in recent years. The existing supervised multi-view feature selection methods usually concatenate all views into the long vectors to design the models. However, this concatenation has no physical meaning and indicates that different views play the similar roles for a specific task. In this paper, we propose a robust re-weighted multi-view feature selection method by constructing the penalty term based on the low-dimensional subspaces of each view through the least-absolute criterion. The proposed model can fully consider the complementary property of multiple views and the specificity of each view. It can not only induce robustness to mitigate the impacts of outliers, but also learn the corresponding weights adaptively for different views without any presetting parameter. In the process of optimization, the proposed model can be splitted to several small scale sub-problems. An iterative algorithm based on the iteratively re-weighted least squares is proposed to efficiently solve these sub-problems. Furthermore, the convergence of the iterative algorithm is theoretical analyzed. Extensive comparable experiments with several state-of-the-art feature selection methods verify the effectiveness of the proposed method.

**Keywords:** Supervised feature selection, multi-view, robustness, re-weighted.

[1]College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China.

[2]College of Science, China Agricultural University, Beijing, 100083, China.

[3]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[4]Viterbi School of Engineering, University of Southern California, Los Angeles, 90089, USA.

*Corresponding Author: Ping Zhong. Email: zping@cau.edu.cn.

## 1 Introduction

In many applications, we need to deal with a large number of data which have high dimensionality. Handling high dimensional data (such as image and video data) may bring many challenges, including the added computational complexity and the increased chance of over-fitting. So how to effectively reduce the dimensionality has become an important issue. As an effective method of selecting representative features, feature selection has attracted many attentions. Feature selection methods [Fang, Cai, Sun et al. (2018)] can be grouped into filter methods, wrapper methods, and embedded methods. Filter methods select features according to the general characteristics of data without taking the learning model into consideration. Wrapper methods select features by taking the performance of some model as criterion. Embedded methods incorporate feature selection and classification process into a single optimization problem, which can achieve reasonable computational cost and good classification performance. Thus, embedded methods are in the dominant position in machine learning, and Least absolute shrinkage and selection operator (Lasso) [Tibshirani(2011)] is one of the most important representatives.

Recently, unlike the previous vector-based feature selection methods (such as Lasso) that are only used for binary classification, many of matrix-based structured sparsity-inducing feature selection (SSFS) methods have been proposed to solve multi-class classification [Gui, Sun, Ji et al. (2016) ]. Obozinski et al. [Obozinski, Taskar and Jordan (2006)] first introduced $l_{2,1}$-norm regularization term which is an extension of $l_1$-norm in Lasso for multi-task feature selection. The $l_{2,1}$-norm regularizer can obtain a joint sparsity matrix because minimizing $l_{2,1}$-norm will make the rows of transformation matrix corresponding to the nonessential features become zeros or close to zeros. Thanks to its good performance, many SSFS methods based on $l_{2,1}$-norm regularizer have been proposed [Yang, Ma, Hauptman et al. (2013); Chen, Zhou and Ye (2011); Wang, Nie, Huang et al. (2011); Wang, Nie, Huang et al. (2011); Jebara (2011)]. In addition, Nie et al. [Nie, Huang, Cai et al. (2011)] utilized $l_{2,1}$-norm penalty to construct a robust SSFS method called RFS to deal with bioinformatics tasks. Unlike the frequently-used least squared penalty, the residual in RFS is not squared, and thus the outliers have less influence. With the aid of $l_{2,1}$-norm penalty, several robust SSFS methods have been proposed [Zhu, Zuo, Zhang et al. (2015); Du, Ma, Li et al. (2017)].

As is known, the description of an object from multiple views is more informative than the one from single view, and a large amount of multi-view data have been collected. In order to describe this kind of data in a better way, a lot of features are extracted by different feature extractors. How to integrate these views to select more relevant feature subset is important for the subsequent classification model. Xiao et al. [Xiao, Sun, He et al. (2013)] firstly proposed the two-view feature selection method. However, many objects are described from more than two views. Wang et al. [Wang, Nie, Huang et al. (2012); Wang, Nie and Huang (2013); Wang, Nie, Huang et al. (2013)] proposed the improved multi-view feature selection methods to handle the general case. In Wang et al. [Wang, Nie and Huang (2012)], they established a multi-view feature selection framework by adopting $G_1$-norm regularizer and $l_{2,1}$-norm regularizer to make both views and features sparsity. In Wang et al. [Wang, Nie and Huang (2013); Wang, Nie, Huang et al. (2013)], SSMVFS and SMML

were proposed by using the same framework to induce the structured sparsity. Specifically, SSMVFS employed the discriminative K-means loss for clustering, and SMML employed the hinge loss for classification. Zhang et al. [Zhang, Tian, Yang et al. (2014)] proposed a multi-view feature selection method based on $G_{2,1}$-norm regularizer by incorporating the view-wise structure information. Gui et al. [Gui, Rao, Sun et al. (2014)] proposed a joint feature extraction and feature selection method by considering both complementary property and consistency of different views.

Multi-view feature selection methods have achieved good performance. Concatenating multiple views into new vectors is a common way when establishing the multi-view feature selection schemes. However, the concatenated vectors have no physical meaning, and the concatenation implies that the different views have similar effects on a specific class. In addition, the concatenated vectors are always high dimensional, which increases the chance of over-fitting. Noticing these limitations, some multi-view clustering methods [Xu, Wang and Lai (2016); Xu, Han, Nie et al. (2017)] have been proposed to learn the corresponding distribution of different views.

Inspired by the above work, we propose a robust re-weighted multi-view feature selection method (RRMVFS) without concatenation. For each view, we make the predictive values close to the real labels, and construct the penalty term by using the least-absolute criterion, which can not only induce robustness but also learn the corresponding view weights adaptively without any presetting parameter. Based on the proposed penalty, the scheme is established by adding $G_1$-norm and $l_{2,1}$-norm regularization terms for the structured sparsity. In the procedure of optimization, the proposed model can be decomposed into several small scale subproblems, and an iterative algorithm based on Iterative Re-weighted Least Squares (IRLS)[Daubechies, Devore, Fornasier et al. (2008)] is proposed to solve the new model. Furthermore, the theoretical analysis of convergence is also presented.

In a nutshell, the proposed multi-view feature selection method has the following advantages:

• It can fully consider the complementary property of multiple views as well as the specificity of each view, since it assigns all views of each sample to the same class while separately imposes penalty for each view.

• It can reduce the influence of outliers effectively because the least-absolute residuals of each view are combined as penalty.

• It can learn the view weights adaptively by a re-weighted way, where the weights are updated according to the current weights and bias matrix without any presetting parameter.

• It can be efficiently solved due to the following two reasons. One is that the objective function can be decomposed into several small scale optimization subproblems, and the other one is that IRLS can solve the least-absolute residual problem within finite iterations.

• Extensive comparison experiments with several state-of-the-art feature selection methods show the effectiveness of the proposed method.

The paper is organized as follows. In Section 2, we present our feature selection model and algorithm in detail, and the convergence of the proposed algorithm is analysed. After

presenting the extensive experiments in Section 3, we draw the conclusions in Section 4.

Now, we give the notation in this paper. Given $N$ samples which have $V$ views belonging to $P$ classes, the data matrix of the $v$th view is denoted as $X^{(v)} = [\boldsymbol{x}_1^{(v)}, \boldsymbol{x}_2^{(v)}, \cdots, \boldsymbol{x}_N^{(v)}] \in \mathcal{R}^{d_v \times N}$, where $\boldsymbol{x}_n^{(v)} \in \mathcal{R}^{d_v}$ is the $v$th view of the $n$th sample, and $d_v$ is the feature number of the $v$th view, $v = 1, \cdots, V$. The data matrix of all views can be represented by $X = [X^{(1)}; X^{(2)}; \cdots; X^{(V)}] \in \mathcal{R}^{d \times N}$ ( $[A; B]$ represents $\begin{bmatrix} A \\ B \end{bmatrix}$), where $d = \sum_{v=1}^{V} d_v$. Each sample is denoted by $\boldsymbol{x}_n = [\boldsymbol{x}_n^{(1)}; \boldsymbol{x}_n^{(2)}; \cdots; \boldsymbol{x}_n^{(V)}] \in \mathcal{R}^d$ (that is, the $n$th column of $X$), and its label vector is denoted as $\boldsymbol{y}_n = (y_{1n}, y_{2n}, \cdots, y_{Pn})^T \in \mathcal{R}^P$. If a sample belongs to the $p$th class, then the $p$th element of $\boldsymbol{y}_n$ is 1, i.e, $y_{pn} = 1$, and others are $-1$. The label matrix is denoted as $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N] \in \mathcal{R}^{P \times N}$. For each class and each view, there is a feature transformation vector denoted by $\boldsymbol{w}_p^{(v)} \in \mathcal{R}^{d_v}$ ($p = 1, \cdots, P; v = 1, \cdots, V$). Then the feature transformation matrix is represented as $W = [\boldsymbol{w}_1^{(1)}, \cdots, \boldsymbol{w}_P^{(1)}; \cdots, \cdots, \cdots; \boldsymbol{w}_1^{(V)}, \cdots, \boldsymbol{w}_P^{(V)}] \in \mathcal{R}^{d \times P}$, and the feature transformation matrix of the $v$th view is $W^{(v)} = [\boldsymbol{w}_1^{(v)}, \boldsymbol{w}_2^{(v)} \cdots, \boldsymbol{w}_P^{(v)}] \in \mathcal{R}^{d_v \times P}$.

## 2 Robust re-weighted multi-view feature selection method

### *2.1 Model formulation*

In order to select the relevant views and feature subset from the original ones, we first use label information of the multi-view data to build the penalty term through the loss minimization principle. We calculate the penalty term by least-absolute criterion which can induce robustness. Instead of concatenating all views as long vectors, the penalty term is established as the sum of the residuals calculated on the latent subspace of each view:

$$\sum_{v=1}^{V} \|{W^{(v)}}^T X^{(v)} + \boldsymbol{b}^{(v)} \mathbf{1}_N^T - Y\|_F \tag{1}$$

where $\boldsymbol{b^{(v)}} \in \mathcal{R}^P$ is a bias vector of the $v$th view, and $\mathbf{1}_N \in \mathcal{R}^N$ is the vector of all ones. The residuals are not squared and thus outliers have less effect compared with the squared residuals. Since different views have different effects for a specific class, we use $G_1$-norm regularizer to enforce the sparsity between views. Meanwhile, we use $l_{2,1}$-norm regularizer which has the ability of imposing the transformation matrix sparse between rows to select the representative features. The formulation of the proposed multi-view feature selection method can be described as follows:

$$\min_{W} J(W, B) = \sum_{v=1}^{V} \|{W^{(v)}}^T X^{(v)} + \boldsymbol{b}^{(v)} \mathbf{1}_N^T - Y\|_F + \gamma_1 \|W\|_{G_1} + \gamma_2 \|W\|_{2,1} \tag{2}$$

where $B = [\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \cdots, \boldsymbol{b}^{(V)}] \in \mathcal{R}^{P \times V}$, $\gamma_1$ and $\gamma_2 > 0$ are two non-negative trade-off parameters.

Formulation (2) assigns all views of each sample to the same class, and imposes the penalty separately on each view. It simultaneously considers the complementary property of different views and the specificity of each view.

### *2.2 Optimization algorithm*

Next we give an iterative optimization algorithm to solve the problem (2). Firstly, we transform the objective function of (2) as:

$$\sum_{v=1}^{V}\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F+\gamma_1\sum_{v=1}^{V}\sum_{p=1}^{P}\|\boldsymbol{w}_p^{(v)}\|_2+\gamma_2\sum_{v=1}^{V}\|W^{(v)}\|_{2,1}$$
$$=\sum_{v=1}^{V}J^v(W^{(v)},\boldsymbol{b}^{(v)}) \tag{3}$$

where $J^v(W^{(v)},\boldsymbol{b}^{(v)})=\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F+\gamma_1\sum_{p=1}^{P}\|\boldsymbol{w}_p^{(v)}\|_2+\gamma_2\|W^{(v)}\|_{2,1}$ Since $J^v(W^{(v)},\boldsymbol{b}^{(v)})$ is only related to the $v$th view and nonnegative, the problem (2) can be decomposed into solving $V$-subproblems:

$$\min_{W^{(v)},\boldsymbol{b}^{(v)}}J^v(W^{(v)},\boldsymbol{b}^{(v)}) \tag{4}$$

Note that the problem (4) cannot be easily solved by the sophisticated optimization algorithms since its objective function is nonsmooth. We utilize IRLS method [Daubechies, Devore, Fornasier et al. (2008)] to solve it, and change it as follows:

$$\min_{W^{(v)},\boldsymbol{b}^{(v)},\alpha^v}\alpha^v\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F^2+\gamma_1\sum_{p=1}^{P}\|\boldsymbol{w}_p^{(v)}\|_2+\gamma_2\|W^{(v)}\|_{2,1} \tag{5}$$

where $\alpha^v=\frac{1}{2\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F}$.

**1. Fix $\alpha^v$, update $W^{(v)},\boldsymbol{b}^{(v)}$.** For problem (5), we can solve the following problem iteratively with the fixed $\alpha^v$:

$$\min_{W^{(v)},\boldsymbol{b}^{(v)}}\alpha^v\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F^2+\gamma_1\sum_{p=1}^{P}\boldsymbol{w}_p^{(v)^T}\tilde{D}_p^{(v)}\boldsymbol{w}_p^{(v)}$$
$$+\gamma_2\mathbf{tr}(W^{(v)^T}\bar{D}^{(v)}W^{(v)}) \tag{6}$$

where $\tilde{D}_p^{(v)}=\frac{1}{2\|\boldsymbol{w}_p^{(v)}\|_2}I_{d_v}$ ($I_{d_v}$ is an identity matrix), $\mathbf{tr}(\cdot)$ is the trace norm of matrix, and $\bar{D}^{(v)}=diag(\frac{1}{2\|W_{1:}^{(v)}\|_2},\cdots,\frac{1}{2\|W_{d_v:}^{(v)}\|_2})$, where $W_{i:}^{(v)}$ is the $i$th row of $W^{(v)}$. Since

$$\mathbf{tr}(W^{(v)^T}\bar{D}^{(v)}W^{(v)})=\sum_{p=1}^{P}\boldsymbol{w}_p^{(v)^T}\bar{D}^{(v)}\boldsymbol{w}_p^{(v)} \tag{7}$$

$$\alpha^v\|W^{(v)^T}X^{(v)}+\boldsymbol{b}^{(v)}\mathbf{1}_N^T-Y\|_F^2=\alpha^v\sum_{p=1}^{P}\sum_{n=1}^{N}(\boldsymbol{w}_p^{(v)^T}\boldsymbol{x}_n^{(v)}+b_p^{(v)}-y_{pn})^2 \tag{8}$$

the problem (6) becomes

$$\min_{W^{(v)},\boldsymbol{b}^{(v)}} \sum_{p=1}^{P} \left( \alpha^v \sum_{n=1}^{N} (\boldsymbol{w}_p^{(v)^T} \boldsymbol{x}_n^{(v)} + b_p^{(v)} - y_{pn})^2 + \gamma_1 \boldsymbol{w}_p^{(v)^T} \tilde{D}_p^{(v)} \boldsymbol{w}_p^{(v)} + \gamma_2 \boldsymbol{w}_p^{(v)^T} \bar{D}^{(v)} \boldsymbol{w}_p^{(v)} \right) \tag{9}$$

Define $J_p^v(\boldsymbol{w}_p^{(v)}, b_p^{(v)}) = \alpha^v \sum_{n=1}^{N} (\boldsymbol{w}_p^{(v)^T} \boldsymbol{x}_n^{(v)} + b_p^{(v)} - y_{pn})^2 + \gamma_1 \boldsymbol{w}_p^{(v)^T} \tilde{D}_p^{(v)} \boldsymbol{w}_p^{(v)} + \gamma_2 \boldsymbol{w}_p^{(v)^T} \bar{D}^{(v)} \boldsymbol{w}_p^{(v)}$. Since $J_p^v(\boldsymbol{w}_p^{(v)}, b_p^{(v)})$ is only related to the $p$th class and nonnegative, the problem (9) can be decomposed into solving $P$-subproblems:

$$\min_{\boldsymbol{w}_p^{(v)}, b_p^{(v)}} J_p^v(\boldsymbol{w}_p^{(v)}, b_p^{(v)}) \tag{10}$$

Taking the derivative of the (10) with the respect to $b_p^{(v)}$ and $\boldsymbol{w}_p^{(v)}$ and setting them to zeros, we can obtain

$$\alpha^v \sum_{n=1}^{N} (2\boldsymbol{x}_n^{(v)^T} \boldsymbol{w}_p^{(v)} + 2b_p^{(v)} - 2y_{pn}) = 0 \tag{11}$$

$$\alpha^v \sum_{n=1}^{N} (2\boldsymbol{x}_n^{(v)} \boldsymbol{x}_n^{(v)^T} \boldsymbol{w}_p^{(v)} + 2\boldsymbol{x}_n^{(v)} b_p^{(v)} - 2\boldsymbol{x}_n^{(v)} y_{pn}) + 2\gamma_1 \tilde{D}_p^{(v)} \boldsymbol{w}_p^{(v)} + 2\gamma_2 \bar{D}^{(v)} \boldsymbol{w}_p^{(v)} = 0 \tag{12}$$

So from (11), we have

$$b_p^{(v)} = \frac{\sum_{n=1}^{N} (y_{pn} - \boldsymbol{x}_n^{(v)^T} \boldsymbol{w}_p^{(v)})}{N} = \frac{Y_p \mathbf{1}_N - \boldsymbol{w}_p^{(v)^T} X^{(v)} \mathbf{1}_N}{N} \tag{13}$$

where $Y_p = (y_{p1}, \cdots, y_{pN}) \in \mathcal{R}^N$. Substituting (13) into (12), we have

$$\boldsymbol{w}_p^{(v)} = (\alpha^v X^{(v)} H X^{(v)^T} + \gamma_1 \tilde{D}_p^{(v)} + \gamma_2 \bar{D}^{(v)})^{-1} \alpha^v X^{(v)} H Y_p^T \tag{14}$$

where $H = I_N - \frac{\mathbf{1}_N \mathbf{1}_N^T}{N}$ with $I_N$ being an identity matrix.

**2. Fix $W^{(v)}, \boldsymbol{b}^{(v)}$, update $\alpha^v$.** The non-negative weight $\alpha^v$ for each view can be updated as

$$\alpha^v = \frac{1}{2\|W^{(v)^T} X^{(v)} + \boldsymbol{b}^{(v)} \mathbf{1}_N^T - Y\|_F}, v = 1, 2, \cdots, V \tag{15}$$

The proposed multi-view feature selection algorithm (RRMVFS) is summarized in Algorithm 1.

### *2.3 Convergence analysis*

**Theorem 1.** *The values of the objective function of (2) monotonically decrease in each iteration in Algorithm 1, and the algorithm converges.*

**Proof:** From Eq. (3), in order to prove $J(W_{t+1}, B_{t+1}) \leq J(W_t, B_t)$, we only need to prove that $J^v(W^{(v)}, \boldsymbol{b}^{(v)})$ monotonically decreases in each iteration.

**Algorithm 1: Robust Re-weighted Multi-view Feature Selection (RRMVFS)**

**Input:** data matrix of each view $X^{(v)}, v = 1, 2, \cdots, V$, label matrix $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N]$.

**Output:** transformation matrix $W^{(v)}, v = 1, 2, \cdots, V$.

**Initialization**

1. Set $t = 0$, threshold $\epsilon = 10^{-5}$, the largest iterative number $T = 20$; Set $(W^{(v)})_0 \in \mathcal{R}^{d_v \times P} (v = 1, \cdots, V)$ all elements 1, and $(b_p^{(v)})_0$ can be calculated by $(b_p^{(v)})_0 = \frac{Y_p 1_N - (\boldsymbol{w}_p^{(v)})_0{}^T X^{(v)} 1_N}{N}$ $(1 \le p \le P, 1 \le v \le V)$ as well as $(\alpha^v)_0 = \frac{1}{2\|(W^{(v)})_0{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_0 \mathbf{1}_N^T - Y\|_F}$ $(1 \le v \le V)$

**while** not converged **do**

2. Compute the diagonal matrices $(\bar{D}^{(v)})_t$ $(1 \le v \le V)$ with the $i$th diagonal element $\frac{1}{2\|(W_{i:}^{(v)})_{t-1}\|_2}$; Compute the diagonal matrices $(\tilde{D}_p^{(v)})_t = \frac{1}{2\|(\boldsymbol{w}_p^{(v)})_{t-1}\|_2} I_{d_v}$ $(1 \le v \le V)$ with $I_{d_v}$ being an identity matrix.

3. For each $\boldsymbol{w}_p^{(v)}$ and $b_p^{(v)}$ $(1 \le p \le P, 1 \le v \le V)$, compute $(\boldsymbol{w}_p^{(v)})_{t+1} = ((\alpha^v)_t X^{(v)} H X^{(v)^T} + \gamma_1 (\tilde{D}_p^{(v)})_t + \gamma_2 (\bar{D}^{(v)})_t)^{-1} (\alpha^v)_t X^{(v)} H Y_p^T$, and $(b_p^{(v)})_{t+1} = \frac{Y_p \mathbf{1}_N - (\boldsymbol{w}_p^{(v)})_{t+1}{}^T X^{(v)} \mathbf{1}_N}{N}$

4. Calculate $(\alpha^v)_{t+1} = \frac{1}{2\|(W^{(v)})_{t+1}{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_{t+1} \mathbf{1}_N^T - Y\|_F}$ $(1 \le v \le V)$;

5. Check the convergence condition $J(W_t, B_t) - J(W_{t+1}, B_{t+1}) < \epsilon$ ($J(\cdot, \cdot)$ is the objective function of (2) ) or $t > T$;

6. t=t+1;

**End While**

By Step 3 in Algorithm 1, we have

$$\begin{aligned}\Big((W^{(v)})_{t+1}, (\boldsymbol{b}^{(v)})_{t+1}\Big) &= \arg\min_{W^{(v)}, \boldsymbol{b}^{(v)}} (\alpha^v)_t \|W^{(v)^T} X^{(v)} + \boldsymbol{b}^{(v)} \mathbf{1}_N^T - Y\|_F^2 \\ &+ \gamma_1 \sum_{p=1}^{P} \boldsymbol{w}_p^{(v)^T} (\tilde{D}_p^{(v)})_t \boldsymbol{w}_p^{(v)} + \gamma_2 \mathbf{tr}\Big((W^{(v)})^T (\bar{D}_v)_t (W^{(v)})\Big) \end{aligned} \tag{16}$$

Noticing that $(\alpha^v)_t = \frac{1}{2\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t \mathbf{1}_N^T - Y\|_F}$ and substituting $(\tilde{D}_p^{(v)})_t$ and $(\bar{D}_v)_t$ by

definitions, we can get

$$\frac{\|(W^{(v)})_{t+1}{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_{t+1}\mathbf{1}_N^T - Y\|_F^2}{2\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F} + \gamma_1 \sum_{p=1}^{P} \frac{\|(\boldsymbol{w}_p^{(v)})_{t+1}\|_2^2}{2\|(\boldsymbol{w}_p^{(v)})_t\|_2} + \gamma_2 \sum_{i=1}^{d_v} \frac{\|(W_{i:}^{(v)})_{t+1}\|_2^2}{2\|(W_{i:}^{(v)})_t\|_2}$$
$$\leq \frac{\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F^2}{2\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F} + \gamma_1 \sum_{p=1}^{P} \frac{\|(\boldsymbol{w}_p^{(v)})_t\|_2^2}{2\|(\boldsymbol{w}_p^{(v)})_t\|_2} + \gamma_2 \sum_{i=1}^{d_v} \frac{\|(W_{i:}^{(v)})_t\|_2^2}{2\|(W_{i:}^{(v)})_t\|_2} \tag{17}$$

Since for $f(z) = z - \frac{z^2}{2\beta}$, given any $z \neq \beta \in \mathcal{R}^n$, $f(z) \leq f(\beta)$ holds, we get

$$\|(W^{(v)})_{t+1}{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_{t+1}\mathbf{1}_N^T - Y\|_F - \frac{\|(W^{(v)})_{t+1}{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_{t+1}\mathbf{1}_N^T - Y\|_F^2}{2\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F}$$
$$\leq \|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F - \frac{\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F^2}{2\|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F} \tag{18}$$

$$\gamma_1 \sum_{p=1}^{P} \|(\boldsymbol{w}_p^{(v)})_{t+1}\|_2 - \gamma_1 \sum_{p=1}^{P} \frac{\|(\boldsymbol{w}_p^{(v)})_{t+1}\|_2^2}{2\|(\boldsymbol{w}_p^{(v)})_t\|_2} \leq \gamma_1 \sum_{p=1}^{P} \|(\boldsymbol{w}_p^{(v)})_t\|_2 - \gamma_1 \sum_{p=1}^{P} \frac{\|(\boldsymbol{w}_p^{(v)})_t\|_2^2}{2\|(\boldsymbol{w}_p^{(v)})_t\|_2} \tag{19}$$

$$\gamma_2 \sum_{i=1}^{d_v} \|(W_{i:}^{(v)})_{t+1}\|_2 - \gamma_2 \sum_{i=1}^{d} \frac{\|(W_{i:}^{(v)})_{t+1}\|_2^2}{2\|(W_{i:}^{(v)})_t\|_2} \leq \gamma_2 \sum_{i=1}^{d_v} \|(W_{i:}^{(v)})_t\|_2 - \gamma_2 \sum_{i=1}^{d} \frac{\|(W_{i:}^{(v)})_t\|_2^2}{2\|(W_{i:}^{(v)})_t\|_2} \tag{20}$$

From (17)-(20), we obtain

$$\|(W^{(v)})_{t+1}{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_{t+1}\mathbf{1}_N^T - Y\|_F + \gamma_1 \sum_{p=1}^{P} \|(\boldsymbol{w}_p^{(v)})_{t+1}\|_2 + \gamma_2 \sum_{i=1}^{d_v} \|(W_{i:}^{(v)})_{t+1}\|_2$$
$$\leq \|(W^{(v)})_t{}^T X^{(v)} + (\boldsymbol{b}^{(v)})_t\mathbf{1}_N^T - Y\|_F + \gamma_1 \sum_{p=1}^{P} \|(\boldsymbol{w}_p^{(v)})_t\|_2 + \gamma_2 \sum_{i=1}^{d_v} \|(W_{i:}^{(v)})_t\|_2 \tag{21}$$

That is, $J^v(W_{t+1}^{(v)}, \boldsymbol{b}_{t+1}^{(v)}) \leq J^v(W_t^{(v)}, \boldsymbol{b}_t^{(v)})$. Thus,

$$\sum_{v=1}^{V} J^v(W_{t+1}^{(v)}, \boldsymbol{b}_{t+1}^{(v)}) \leq \sum_{v=1}^{V} J^v(W_t^{(v)}, \boldsymbol{b}_t^{(v)})$$

So, we get $J(W_{t+1}, B_{t+1}) \leq J(W_t, B_t)$, that is, Algorithm 1 decreases the objective values in each iteration. Since $J(W, B) \geq 0$, Algorithm 1 converges.

## 3 Experiments

In this section, we evaluate the effectiveness of the proposed method by comparing it with several related feature selection methods.

### *3.1 Experimental setup*

We compare the performance of our method RRMVFS with several related methods: Single, CAT, RFS [Nie, Huang, Cai et al. (2011)], SSMVFS [Wang, Nie and Huang (2013)], SMML [Wang, Nie, Huang et al. (2013)], DSML-FS [Gui, Rao, Sun et al. (2014)], and WMCFS [Xu, Wang and Lai (2016)]. Single refers to using single-view features to find best performance for classification. CAT refers to using concatenated vectors for classification without feature selection. RFS is an efficient robust feature selection method, but not designed for multi-view feature selection. The other feature selection methods are designed for multi-view feature selection. The parameters $\gamma_1$ and $\gamma_2$ in feature selection methods are tuned from the set $\{10^i | i = -5, -4, \cdots, 5\}$. The exponential parameter $\rho$ in WMCFS method is set to be 5 according to Xu et al. [Xu, Wang and Lai (2016)].

The public available data sets, including images data set NUS-WIDE-OBJECT (NUS), handwritten numerals data set mfeat, and Internet pages data set Ads are employed in the experiments. For NUS data set, we choose all 12 animal classes including 8182 images. For Ads, there exist some incomplete data. We first discard the incomplete data and then randomly choose some non-AD samples so that the number of non-AD data is the same as that of AD data. The total number of the samples employed in Ads is 918. For mfeat data set, all of data are employed. In each data set, the samples are randomly and averagely divided into two parts. One part is used for training and the other part is used for testing. In the training set, we randomly choose 6 (9, or 12) samples from each class to learn transformation matrix of these compared methods. In the test set, we employ 20% of data for validation, and the parameters that achieve the best performance on the validation set are employed for testing. We arrange features in descending order based on the values of $\|W_{i:}\|_2, i = 1, 2, \cdots, d$, and select a certain number of top-ranked features. The numbers of selected features are $\{10\%, \cdots, 90\%\}$ of the total amount of features, respectively. Then the selected features are taken as the inputs of the subsequent 1-nearest neighbour (1NN) classifier. We conduct two kinds of experiments. First, we conduct experiments on all views to evaluate these methods. Then, we conduct experiments on the subsets formed by two views and four views to make the evaluation of views. For all experiments, ten independent and random choices of training samples are employed, and the averaged accuracies (AC) and F1 scores (macroF1) are reported.

### *3.2 Evaluation of feature selection*

Figs. 1, 2, and 3 show the performance of the compared methods w.r.t. the different percentages of selected features on NUS, mfeat and Ads data sets, respectively. From these figures, we can see that, the performance of these methods under AC are consistent with the one under macroF1 scores. CAT has the better performance than Single in all cases, which means that it is essential to combine different views to select features. The feature selection methods, including ours, achieve the comparable or even better performance than CAT. Specifically, from Fig. 1, we can see that three feature selection methods including RFS, SSMVFS, and DSML-FS show the comparable performance on NUS data set. SSMVFS and DSML-FS are the feature selection methods designed for multi-view

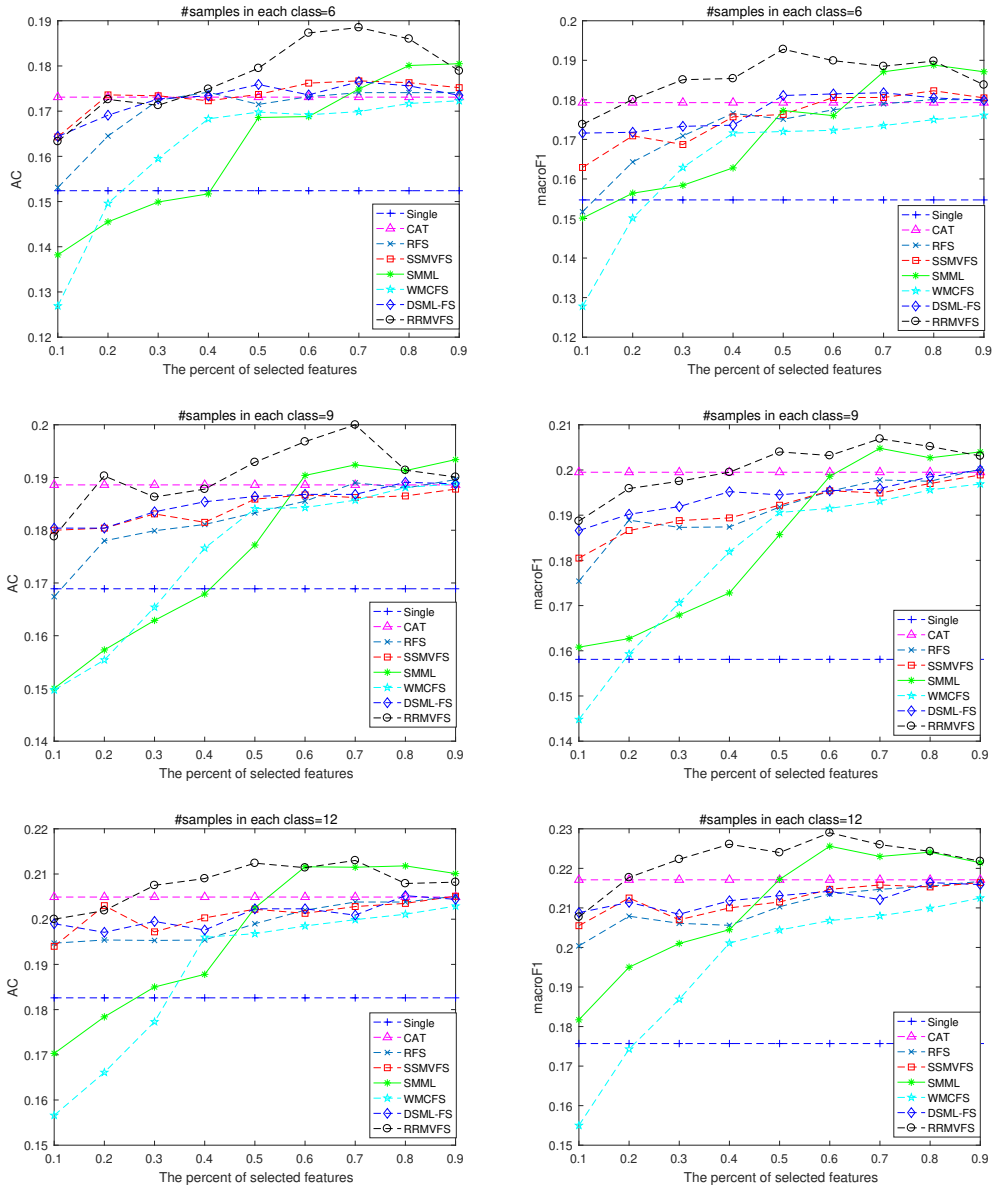


**Figure 1:** ACs and macroF1 scores of compared methods *vs.* percents of selected features on NUS data set

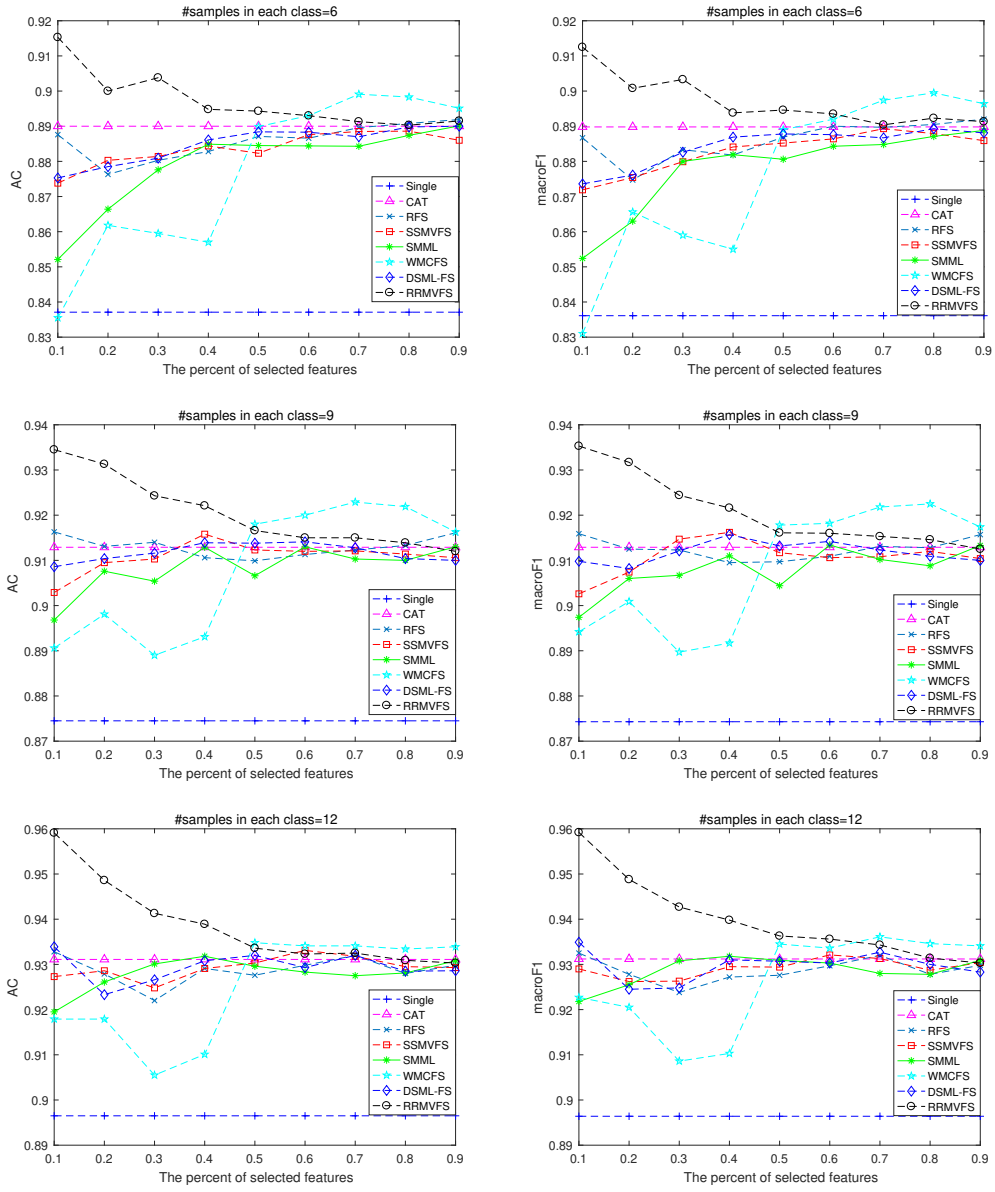


**Figure 2:** ACs and macroF1 scores of compared methods *vs.* percents of selected features on mfeat data set

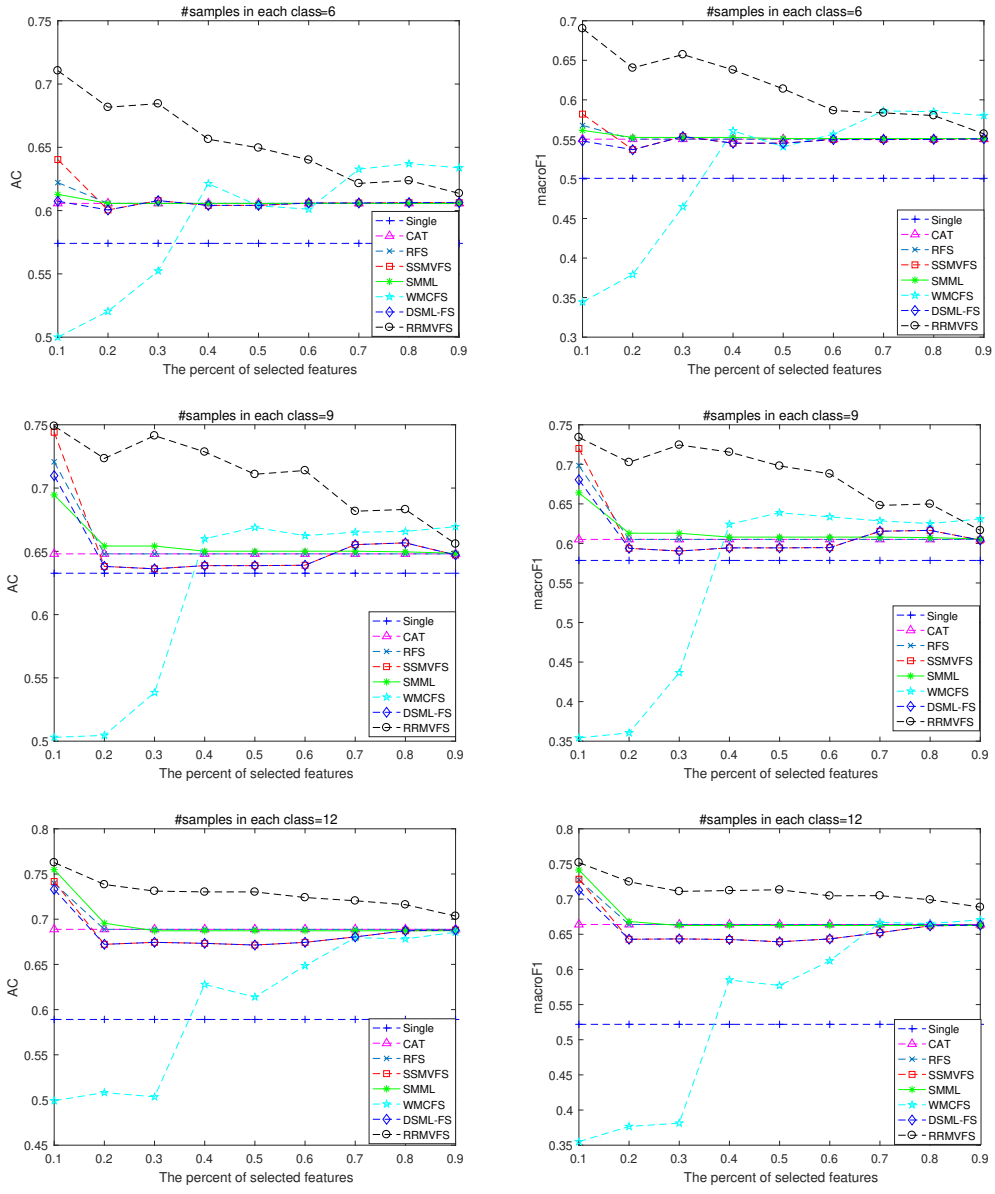


**Figure 3:** ACs and macroF1 scores of compared methods *vs.* percents of selected features on Ads data set

learning, while RFS is a robust feature selection method that is not specially designed for multi-view learning. This means that it is necessary to build a robust method since data may be corrupted with noise. Along with the number of selected features increasing, the ACs and macroF1 scores of the multi-view feature selection methods WMCFS, SMML, and the proposed RRMVFS are greatly improved. Moreover, RRMVFS achieves the best results in most cases. This phenomenon might be attributed to the robust penalty which may help RRMVFS select more representative features.

From Fig. 2, we can see that RFS, SSMVFS, and DSML-FS show the comparable performance on mfeat data set. Along with the number of selected features increasing, the ACs and macroF1 scores of SMML and WMCFS are increased except at a few percentages, especially for WMCFS. The proposed RRMVFS can achieve the best performance when only a small number of features (the percent of selected features is 10) are selected.

From Fig. 3, we can see that four feature selection methods including RFS, SSMVFS, SMML, and DSML-FS obtain comparable performance on Ads data set. They achieve their best performance at 10%, and when the percentages of selected features change from 20% to 90%, their ACs and macroF1 scores are substantially unchanged and comparable to those of CAT. Just like the performance on the other two data sets, the proposed RRMVFS still achieves the best performance on Ads data set.

### *3.3 Evaluation of views*

In order to evaluate the effect of views, we test ACs and macroF1 scores of the compared methods in terms of views on NUS, mfeat, and Ads data sets, respectively. The experiments are conducted on the subsets that contains 12 samples of each class. For each data sets, we conduct two kinds of experiments. Firstly, we randomly choose two views for experiments. Secondly, we randomly choose two views from the remaining views, and combine them with the previous two views to form the subsets for experiments.

**Table 1:** ACs and macroF1 scores with standard deviation of the compared methods in terms of views on NUS data set

| | AC | | | macroF1 | | |
|---|---|---|---|---|---|---|
| method | V3&V5 | V1&V2 &V3&V5 | all views | V3&V5 | V1&V2 &V3&V5 | all views |
| CAT | 0.1932±0.0130 | 0.2003±0.0140 | 0.2049±0.0101 | 0.1956±0.0107 | 0.2065±0.0132 | 0.2171±0.0096 |
| RFS | 0.1949±0.0123 | 0.2012±0.0138 | 0.2051±0.0104 | 0.1979±0.0130 | 0.2079±0.0164 | 0.2167±0.0097 |
| SSMVFS | 0.1997±0.0148 | 0.2031±0.0136 | 0.2051±0.0106 | 0.2002±0.0137 | 0.2119±0.0129 | 0.2164±0.0098 |
| SMML | **0.2083±0.0107** | 0.2092±0.0118 | 0.2118±0.0125 | 0.2104±0.0095 | 0.2143±0.0114 | 0.2256±0.0240 |
| DSML-FS | 0.2019±0.0120 | 0.2009±0.0143 | 0.2052±0.0118 | 0.2045±0.0101 | 0.2116±0.0155 | 0.2164±0.0116 |
| WMCFS | 0.1917±0.0121 | 0.2019±0.0141 | 0.2029±0.0103 | 0.1962±0.0090 | 0.2079±0.0128 | 0.2125±0.0094 |
| RRMVFS | 0.2072±0.0135 | **0.2102±0.0132** | **0.2130±0.0137** | **0.2076±0.0102** | **0.2150±0.0103** | **0.2290±0.0156** |

The experimental results on these three data sets are shown in Tabs. 1-3, respectively. It can be seen that with the increase of views, generally speaking, the performance of all compared methods gets better. On the subsets which consist of two views, our method RRMVFS does not show the best performance, but on the subsets formed by four views and the whole sets, our method is superior to others significantly. This phenomenon might be attributed to the learning of view weights which may help RRMVFS select more relevant views.

**Table 2:** ACs and macroF1 scores with standard deviation of the compared methods in terms of views on mfeat data set

| | AC | | | macroF1 | | |
|---|---|---|---|---|---|---|
| method | V1&V6 | V1&V2 &V3&V6 | all views | V1&V6 | V1&V2 &V3&V6 | all views |
| CAT | 0.8915±0.0111 | 0.9226±0.0109 | 0.9311±0.0107 | 0.8913±0.0111 | 0.9226±0.0106 | 0.9312±0.0105 |
| RFS | 0.8939±0.0118 | 0.9243±0.0133 | 0.9328±0.0108 | 0.8948±0.0112 | 0.9258±0.0090 | 0.9325±0.0109 |
| SSMVFS | 0.8950±0.0117 | 0.9243±0.0155 | 0.9331±0.0072 | 0.8956±0.0123 | 0.9245±0.0100 | 0.9321±0.0076 |
| SMML | 0.8899±0.0104 | 0.9210±0.0130 | 0.9318±0.0088 | 0.8904±0.0107 | 0.9209±0.0127 | 0.9318±0.0088 |
| DSML-FS | 0.8955±0.0114 | 0.9251±0.0130 | 0.9339±0.0075 | 0.8941±0.0175 | 0.9251±0.0128 | 0.9349±0.0094 |
| WMCFS | **0.9094±0.0112** | 0.9310±0.0106 | 0.9341±0.0106 | **0.9101±0.0107** | 0.9314±0.0103 | 0.9361±0.0130 |
| RRMVFS | 0.8964±0.0162 | **0.9397±0.0092** | **0.9591±0.0089** | 0.8970±0.0144 | **0.9400±0.0129** | **0.9592±0.0089** |

**Table 3:** ACs and macroF1 scores with standard deviation of the compared methods in terms of views on Ads data set

| | AC | | | macroF1 | | |
|---|---|---|---|---|---|---|
| method | V1&V3 | V1&V3 &V4&V5 | all views | V1&V3 | V1&V3 &V4&V5 | all views |
| CAT | 0.6520±0.0767 | 0.6807±0.0651 | 0.6888±0.0545 | 0.6228±0.0941 | 0.6572±0.0730 | 0.6639±0.0585 |
| RFS | 0.6932±0.0926 | 0.7351±0.0738 | 0.7398±0.0630 | 0.6745±0.1089 | 0.7233±0.0811 | 0.7274±0.0658 |
| SSMVFS | **0.7510±0.0803** | 0.7482±0.0941 | 0.7417±0.0913 | **0.7472±0.0867** | 0.7319±0.1256 | 0.7283±0.1280 |
| SMML | 0.7253±0.0830 | 0.7466±0.0657 | 0.7548±0.0394 | 0.7087±0.0964 | 0.7307±0.0823 | 0.7414±0.0447 |
| DSML-FS | 0.7335±0.0917 | 0.7316±0.1046 | 0.7330±0.0860 | 0.7283±0.1046 | 0.7129±0.1300 | 0.7126±0.1237 |
| WMCFS | 0.5809±0.0644 | 0.6496±0.0928 | 0.6856±0.0849 | 0.5140±0.0753 | 0.6119±0.1230 | 0.6705±0.0745 |
| RRMVFS | 0.7144±0.0905 | **0.7485±0.0809** | **0.7627±0.0523** | 0.7043±0.0890 | **0.7434±0.0772** | **0.7519±0.0588** |

## 4 Conclusion

In this paper, we have proposed a robust re-weighted multi-view feature selection method by assigning all views of each sample to the same class while imposing penalty based on latent subspaces of each view through least-absolute criterion, which can take both the complementary property of different views and the specificity of each view into consideration and induce the robustness. The proposed model can be efficiently solved by decomposing it into several small scale optimization subproblems, and the convergence of the proposed iterative algorithm is presented. The comparison experiments with several state-of-the-art feature selection methods verify the effectiveness of the proposed method.

Many real-world applications, such as text categorization, are multi-label problems. The future work is to develop the proposed method for multi-label multi-view feature selection.

**Acknowledgement:** The work was supported by the National Natural Science Foundation of China (Grant Nos. 61872368, U1536121).

## References

**Chen, J.; Zhou, J.; Ye, J.** (2011): Integrating low-rank and group-sparse structures for robust multi-task learning. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 42-50.

**Daubechies, I.; Devore, R.; Fornasier, M.; Gunturk, S.** (2008): Iteratively re-weighted

least squares minimization: proof of faster than linear rate for sparse recovery. *42nd Annual Conference on Information Sciences and Systems*, pp. 26-29.

**Du, S.; Ma, Y.; Li, S.; Ma, Y.** (2017): Robust unsupervised feature selection via matrix factorization. *Neurocomputing*, vol. 241, pp. 115-127.

**Fang, S.; Cai, Z.; Sun, W.; Liu, A.; Liu, F. et al.** (2018): Feature selection method based on class discriminative degree for intelligent medical diagnosis. *Computers, Materials and Continua*, vol. 55, no. 3, pp. 419-433.

**Gui, J.; Sun, Z.; Ji, S.; Tao, D.; Tan, T.** (2016): Feature selection based on structured sparsity: a comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1490-1507.

**Gui, J.; Tao, D.; Sun, Z.; Luo, Y.; You, X. et al.** (2014): Group sparse multiview patch alignment framework with view consistency for image classification. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 23, no. 7, pp. 3126-3137.

**Jebara, T.** (2011): Multi-task sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, vol. 12, pp. 75-110.

**Nie, F.; Huang, H.; Cai, X.; Ding, C.** (2011): Efficient and robust feature selection via joint $L_{2,1}$-norms minimization. *International Conference on Neural Information Processing Systems, Curran Associates Inc.*, pp. 1813-1821.

**Obozinski, G.; Taskar, B.; Jordan, M.** (2006): *Multi-Task Feature Selection*. Department of Statistics University of California, Berkeley.

**Tibshirani, R.** (2011): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, vol. 73, no. 3, pp. 273-282.

**Wang, H.; Nie, F.; Huang, H.; Risacher, S.** (2011): Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. *International Conference on Computer Vision*, pp. 557-562.

**Wang, H.; Nie, F.; Huang, H.** (2013): Multi-view clustering and feature learning via structured sparsity. *International Conference on Machine Learning*, pp. 352-360.

**Wang, H.; Nie, F.; Huang, H.; Ding, C.** (2013): Heterogeneous visual features fusion via sparse multimodal machine. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3097-3102.

**Wang, H.; Nie, F.; Huang, H.; Risacher, S. L.; Saykin, A. J. et al.** (2012): Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, vol. 28, no. 12, pp. 127-136.

**Xiao, L.; Sun, Z.; He, R.; Tan, T.** (2013): Coupled feature selection for cross-sensor iris recognition. *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems*, pp. 1-6.

**Xu, J.; Han, J.; Nie, F.; Li, X.** (2017): Re-weighted discriminatively embedded k-means for multi-view clustering. *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3016-3027.

**Xu, Y.; Wang, C.; Lai, J.** (2016): Weighted multi-view clustering with feature selection.

*Pattern Recognition*, vol. 53, pp. 25-35.

**Yang, Y.; Ma, Z.; Hauptmann, A. G.; Sebe, N.** (2013): Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 661-669.

**Zhang, Q.; Tian, Y.; Yang, Y.; Pan, C.** (2014): Automatic spatial-spectral feature selection for hyperspectral image via discriminative sparse multimodal learning. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 261-279.

**Zhu, P.; Zuo, W.; Zhang, L.; Hu, Q.; Shiu, S. C. K.** (2015): Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, vol. 48, no. 2, pp. 438-446.