Robust Multi-class Feature Selection via $l_{2,0}$-Norm Regularization Minimization

Zhenzhen Sun, Yuanlong Yu*

College of Mathematics and Computer Science
Fuzhou University
Fuzhou, Fujian, 350116, China
yu.yuanlong@fzu.edu.cn

Abstract—Feature selection is an important data preprocessing in data mining and machine learning, which can reduce feature size without deteriorating model’s performance. Recently, sparse regression based feature selection methods have received considerable attention due to their good performance. However, these methods generally cannot determine the number of selected features automatically without using a predefined threshold. In order to get a satisfactory result, it often costs significant time and effort to tune the number of selected features carefully. To this end, this paper proposed a novel framework to solve the $l_{2,0}$-norm regularization least square problem directly for multi-class feature selection, which can produce exact row-sparsity solution for the weights matrix, features corresponding to non-zero rows will be selected thus the number of selected features can be determined automatically. An efficient homotopy iterative hard threshold (HIHT) algorithm is derived to solve the above optimization problem and find out the stable local solution. Besides, in order to reduce the computational time of HIHT, an acceleration version of HIHT (AHIHT) is derived. Extensive experiments on eight biological datasets show that the proposed method can achieve higher classification accuracy with fewest number of selected features comparing with the approximate convex counterparts and state-of-the-art feature selection methods. The robustness of classification accuracy to the regularization parameter is also exhibited.

Index Terms—Feature selection, $l_{2,0}$-norm regularization, iterative hard threshold, embedded method.

I. INTRODUCTION

Feature selection, the process of selecting a subset of features which are the most relevant and informative, has been widely researched for many years [1]–[5]. Feature selection has become an essential component in data mining and machine learning because it can reduce the feature size, enhance data understanding, alleviate the effect of the curse of dimensionality, speed up the learning process and improve model’s performance. Therefore, it has been widely used in many real-world applications, e.g., text mining [6], [7], pattern recognition [3], and bioinformatics [3], [9].

In general, feature selection methods can be divided into three categories depending on how they combine the feature selection search with model learning algorithms: filter methods, wrapper methods, and embedded methods. In filter methods, features are selected according to the intrinsic properties of the data before running learning algorithm. Therefore, filter methods are independent of the learning algorithms and can be characterized by utilizing the statistical information. Typical filter methods include Relief [10], Chi-square and information gain [11], and mRMR [12], etc. The wrapper methods use learning algorithm as a black box to score subsets of features, such as correlation-based feature selection (CFS) [13] and support vector machine recursive feature elimination (SVM-RFE) [14]. Embedded methods incorporate the feature selection and model learning into a single optimization problem, such that higher computational efficiency and classification performance can be gained than the filter methods and wrapped methods. Thus, the embedded methods have attracted large attention these years. One of the typical embedded methods is the decision tree algorithm, such as C4.5 [15].

Recently, with the development of sparsity research, sparsity regularization has been widely applied into embedded feature selection methods. The concern behind this is that selecting a minority of features is naturally a problem with sparsity. For binary classification task, the feature selection can be tackled by $l_0$-norm minimization [16] directly in which features corresponding to non-zero weights are selected. However, the non-convexity and non-smooth of $l_0$-norm make it very hard to solve. Most methods relax the $l_0$-norm by $l_1$-norm to make the minimization problem be convex and easy to solve, which is called LASSO [17]. Although some strategies such as one-versus-one or one-versus-all can be used to expand LASSO for multi-class feature selection problem, structural sparsity models are more desirable so that we can obtain the shared pattern of sparsity. Inspired by that, lots of methods have been proposed based on structural sparsity for multi-class feature selection [18]. In [19], Nie et al. proposed a robust feature selection (RFS) method with emphasizing joint $l_{2,1}$-norm minimization on both loss function and regularization. Due to the efficiency of the $l_{2,1}$-norm regularization, it has been widely investigated in these years, such as UDFS [20], L-FS [21], RLSR [22], and URAFS [23], etc.

*This work is supported by National Natural Science Foundation of China (NSFC) under grant #61873067.
Though satisfactory results can be achieved by using $l_{2,1}$-norm regularization for multi-class feature selection, there are still some limitations. First, $l_{2,1}$-norm is just an approximation to $l_{2,0}$-norm and the solutions are essentially different from the original optimum value. Second, $l_{2,1}$-norm over-penalizes large weights, which leads to an unfair competition between different features. Further more, it is hard to tune the regularization parameter of $l_{2,1}$-norm to get exact row-sparsity solution, even a large regularization factor (e.g. $10^5$) cannot produce strong row-sparsity. Thus the number of selected features cannot be determined automatically without using a predefined parameter (e.g., the number of selected features or the threshold of important score). Consequently, it is significant to find a method to solve the original $l_{2,0}$-norm regularization problem.

In recent years, researchers have been working on solving the $l_{2,0}$-norm problem directly. In [24], Cai et al. proposed a robust and pragmatic multi-class feature selection (RPMFS) method to solve the original $l_{2,0}$-norm constrained problem. RPMFS sets the objective function as a $l_{2,1}$-norm loss term with a $l_{2,0}$-norm equality constraint, and use the augmented Lagrangian method to solve this equality constraint problem. In [25], Pang et al. also proposed an efficient sparse feature selection method (ESFS) based on $l_{2,0}$-norm constrained problem, then they transform the model into the same structure as LDA to calculate the ratio of inter-class scatter to intra-class scatter of features. However, these methods also need to predefine the number of selected features to construct the equality constraint, which cannot be determined automatically. To obtain satisfactory classification results, the number of selected features need be tuned carefully which costs a large amount of time and effort thus not suitable for practical application. From literatures of sparsity research, it has been proved that the regularized problem is more effective than the equality constraint problem to find a sparse solution.

In this paper, we propose a novel and simple framework for multi-class feature selection which can solve the original $l_{2,0}$-norm regularization least square problem (denoted as $l_{2,0}$-FS) directly. This method can produce exact row-sparsity solution thus select features in group and automatically determine the number of selected features. In order to effectively solve the proposed objective function, the homotopy iterative hard threshold (HIHT) algorithm is designed to solve the proposed optimization problem, in which each row of the weights matrix is treated as a whole and updated. Thus it can achieve exact row-sparsity solution. Besides, an acceleration version of HIHT (AHIHT) is derived to reduce the computational time which is more practical for feature selection task.

2) Experiment results on eight benchmark biological datasets show that our approach outperforms ESFS, RPMFS, the relaxed or approximate counterparts, and state-of-the-art feature selection methods evaluated in terms of classification accuracy using two popular classifiers and the number of selected features. The robustness of classification accuracy to the regularization parameter is also exhibited.

The rest of this paper is organized as follows. Section II presents the notations and definitions used in this paper, and related work on $l_{2,1}$-norm and $l_{2,0}$-norm constrained multi-class feature selection methods are introduced. In Section III, the new method $l_{2,0}$-FS is proposed and the convergence of HIHT algorithm is analyzed. The experimental results are presented in Section IV. Conclusions and future work are given in Section V.

II. RELATED WORK

A. Notations and Definition

The notations and the definition used in this paper are shown in this subsection. Vectors are written as boldface lowercase letters and matrices are written as boldface uppercase letters. For a vector $x \in \mathbb{R}^n$, $x_i$ denotes the $i$-th element of $x$. For a matrix $X = \{x_{ij}\} \in \mathbb{R}^{n \times m}$, $x_i$ and $x_j$ denote its $i$-th row and $j$-th column, respectively.

For $p \neq 0$, the $p$-norm of the vector $x$ is defined as:

$$||x||_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}. $$

The $l_0$-norm of the vector $x$ is defined as:

$$||x||_0 = \sum_{i=1}^{n} |x_i|^0, $$

which count the number of non-zero elements in $x$.

The Frobenius norm of $X$ is defined as:

$$||X||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2} = \sqrt{tr(X^T X)}. $$

The $l_{2,1}$-norm of $X$ is defined as:

$$||X||_{2,1} = \sum_{i=1}^{n} ||x_i||_2 = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} x_{ij}^2}. $$

The $l_{2,0}$-norm of matrix $X$ is defined as:

$$||X||_{2,0} = \sum_{i=1}^{n} 1_{||x_i||_2 \neq 0}, $$

where $1_{a}$ denotes the indicator function.
where \( I_A \) stands for the indicator function. For a scalar \( x \), if \( x \neq 0 \), \( I_x = 1 \), otherwise \( I_x = 0 \). Thus the \( l_{2,0} \)-norm of matrix \( X \) is defined as the number of non-zero rows in \( X \). If a matrix has a large number of zero rows (the row vector is zero vector), we define it has the character of row-sparsity.

### B. Multi-class Feature Selection based on \( l_{2,1} \)-Norm

In general, the most multi-class feature selection algorithms based on \( l_{2,1} \)-norm regularization can be formulated as follows:

\[
\min_{W,b} ||W^T X + b^T 1 - Y||_F^2 + \lambda ||W||_{2,1}, \tag{1}
\]

where \( X = \{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{d \times N} \) is the training data. \( Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^{C \times N} \) is the binary label matrix with \( y_{ij} = 1 \) if \( x_i \) has label \( y_i = j \); otherwise \( y_{ij} = 0 \). \( W \in \mathbb{R}^{d \times C} \) denotes the model weights and \( b \in \mathbb{R}^{1 \times C} \) denotes the learned biased vector. \( 1 \in \mathbb{R}^{1 \times N} \) is a column vector with all its entries being 1. \( N \) is the sample number, \( d \) is the feature dimension, \( C \) denotes the class number, and \( \lambda \) is the regularization parameter. After optimizing, the features of \( X \) can be selected according to the norm of each \( w^i \).

This problem has been widely studied and a lot of variants have been proposed. Nie et al. \[19\] first combine the \( l_{2,1} \)-norm regularization with a \( l_{2,0} \)-norm loss term instead of the Frobenius norm loss term and demonstrate the proposed model is more robust for outliers than the original model. Yang et al. \[20\] incorporated discriminative analysis into the \( l_{2,1} \)-norm minimization. By doing that, an unsupervised feature selection joint model was yielded. In \[21\], the authors combine the \( l_{2,1} \)-norm regularization with the fisher criterion to select more discriminative features. Recently, Yan et al. \[26\] imposed both nonnegative and \( l_{2,1} \)-norm constraints on the feature weights matrix. The nonnegative property ensures the row-sparsity of learned feature weights combining with the \( l_{2,1} \)-norm minimization, which makes it clearer for which feature should be selected.

Although the \( l_{2,1} \)-norm based model can achieve satisfactory results, one of the biggest problem is that we don’t know how many features need to be selected after the regularization parameter is tuned. From the sparsity perspective, \( l_{2,0} \)-norm is more desirable.

#### C. Multi-Class Feature Selection based on \( l_{2,0} \)-Norm Constrained Problem

The multi-class feature selection based on \( l_{2,0} \)-norm always construct an equality constraint to determine the number of non-zero rows of weights matrix. In \[24\], Cai et al. construct the objective function as a \( l_{2,1} \)-norm loss term with a \( l_{2,0} \)-norm equality constraint, which can be written as follows:

\[
\min_{W,b} ||W^T X + b^T 1 - Y||_{2,1}
\]

\[
\text{s.t.} ||W||_{2,0} = k
\]

then they used the Augmented Lagrangian Multiplier (ALM) method to solve this problem.

In \[25\], the authors form a similar model which is written as follows:

\[
\min_{W,b} ||W^T X + b^T 1 - Y||_{2,1}
\]

\[
\text{s.t.} ||W||_{2,0} = k
\]

where \( Q \in \mathbb{R}^{C \times C} \) can be any reversible matrix which is used to code labels. By using this label coding method, they transform model (3) into the same structure as LDA which can calculate the ratio of inter-class scatter to intra-class scatter of features.

Though the above mentioned feature selection methods are based on \( l_{2,0} \)-norm, they also need to predefined the number of selected features to construct the equality constraint. How many features need be selected is unknown, so it will cost a large amount of time and effort to tune the number of selected features to get a satisfactory result. From literatures of sparsity research, it has been proved that the regularized problem is more effective than the equality constraint problem to find a sparse solution.

### III. THE PROPOSED METHOD

#### A. Problem Formulation

In this work, we propose a novel multi-class feature selection method by using the least square regression combined with a \( l_{2,0} \)-norm regularization, the optimization function is formulated as follows:

\[
\varphi_{2,0}(W) = \min_W \frac{1}{2} ||W^T X - Y||_F^2 + \lambda ||W||_{2,0}
\]

(4)

where the notations are defined as in \[II.B\] and the bias \( b \) is neglected.

The \( l_{2,0} \)-norm regularized least square problem (4) can not only select the discriminative features through least square regression but also indispensable features through the \( l_{2,0} \)-norm. Since \( l_{2,0} \)-norm can promote row-sparsity, the selected features are useful for all outputs. Iterative hard threshold (IHT) algorithm \[27\]–[30] is usually used to solve \( l_0 \)-norm regularization minimization for vector sparsity problem, inspired by it, we extended it to solve the matrix sparsity problem (4).

#### B. Optimization Algorithm

The core idea of the IHT algorithm is using the proximal point technique to iteratively update the solution. Firstly we define \( f(W) \) as

\[
f(W) = \frac{1}{2} ||W^T X - Y||_F^2
\]

(5)

Since \( f(W) \) is a differentiable convex function whose gradient is Lipschitz continuous (denote its Lipschitz constant
as $L_f$), it can be approximatively iterative updated by the projected gradient method:

$$W^{t+1} = \arg \min_W f(W^t) + tr(\nabla f(W^t)^T (W - W^t)) + \frac{\lambda}{2} ||W - W^t||_2^2, \tag{6}$$

where $L \geq 0$ is a constant, which should essentially be an upper bound on the Lipschitz constant of $\nabla f(W)$, i.e., $L \geq L_f$.

By adding $\lambda ||W||_{2,0}$ into the right side of (9), the solution of (4) can be obtained by iteratively solving the subproblem:

$$W^{t+1} = \arg \min_W f(W^t) + tr(\nabla f(W^t)^T (W - W^t)) + \frac{\lambda}{2} ||W - W^t||_2^2 + \lambda ||W||_{2,0}. \tag{7}$$

By removing the item $f(W^t)$ and adding an item $\frac{1}{L} ||\nabla f(W^t)||_2^2$, both of which are independent on $W$ and can be considered as constant items, the right hand of (7) can be rewritten as:

$$W^{t+1} = \arg \min_W L \frac{1}{2} \{||W - (W^t - \frac{1}{L} \nabla f(W^t))||_2^2 + \frac{2\lambda}{L} ||W||_{2,0} \}, \tag{8}$$

Since the Frobenius norm and $l_{2,0}$-norm are all separable function, each row of $W$ can be updated individually:

$$(W^{t+1})^i = \arg \min_{w^i} \frac{L}{2} \{||w^i - (w^i - \frac{1}{L} \nabla f((w^i)^T))||_2^2 + \frac{2\lambda}{L} ||w^i||_{2,0} \}, \tag{9}$$

this subproblem has a closed-form solution as:

$$(W^{t+1})^i = \begin{cases} (w^i)^T - \frac{1}{L} \nabla f((w^i)^T), & \text{if } ||(w^i)^T - \frac{1}{L} \nabla f((w^i)^T)||_2^2 > \frac{2\lambda}{L} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

The solution of problem (4) can be obtained by using (10) to iteratively update $(W^j)^i$ in $W^j$ until convergence. In (10), it can be seen that there are a parameter need to be tuned: $L$. The upper bound on $L_f$ is unknown or may not be easily calculated, thus we use line search method to search $L$ as suggested in (29) until the objective value descend.

**Homotopy Strategy:** many works [31–33] have verified that the sparse coding approaches benefit from a good starting point. Thus we can use the solution of (4), for a given value of $\lambda$, to initialize HIHT in a nearby value of $\lambda$. The second solve will typically take fewer iterations than the first one. Using this warm-start technique, we can efficiently solve for a sequence of values of $\lambda$, which is called homotopy strategy. An outline of the proposed HIHT algorithm for solving (4) is described as Alg. 1.

**Algorithm 1 Homotopy iterative hard threshold method to solve problem (4)**

**Input:** Training data $X \in \mathbb{R}^{d \times N}$, training labels $Y \in \mathbb{R}^{C \times N}$; parameters $L_0, \lambda_0, L_{min}, L_{max}$; $/ L_0 \in [L_{min}, L_{max}]$;

**Output:** $W^*$;
1: initialize $k \leftarrow 0, \rho \in (0, 1), \gamma > 1, \eta > 0, \epsilon > 0, W^0 = 0$;
2: repeat
3: $i \leftarrow 0$;
4: $W^{k,0} = W^k$;
5: $L_{k,0} \leftarrow L_k$;
6: repeat
7: update $W^{k,i+1}$ by Eq. (10);
8: while $\varphi_k(W^{k,i}) - \varphi_k(W^{k,i+1}) < \frac{\eta}{2} ||W^{k,i} - W^{k,i+1}||_F^2$ do
9: $L_{k,i} \leftarrow \min \{\gamma L_{k,i}, L_{max}\}$;
10: update $W^{k,i+1}$ by Eq. (10);
11: end while
12: $L_{k,i+1} \leftarrow L_{k,i}$;
13: $i \leftarrow i + 1$;
14: until $||W^{k,i} - W^{k,i+1}||_F^2 \leq \epsilon$
15: $W^{k+1} \leftarrow W^{k,i}$;
16: $L_{k+1} \leftarrow L_{k,i}$;
17: $\lambda_{k+1} = \rho \lambda_k$;
18: $k \leftarrow k + 1$;
19: until $\lambda_{k+1}$ is small enough
20: $W^* \leftarrow W^k$.

**C. Convergence Analysis**

For a fixed $\lambda$, since $\nabla f$ is Lipschitz continuous with constant $L_f$, we have:

$$f(W^{k+1}) \leq f(W^k) + tr(\nabla f(W^k)^T (W^{k+1} - W^k)) + \frac{L_f}{2} ||W^{k+1} - W^k||_2^2. \tag{11}$$

Using this inequality, the fact that $L > L_f$, and (7), we obtain that

$$\varphi_k(W^{k+1}) = f(W^{k+1}) + \lambda_k ||W^{k+1}||_0 \leq f(W^k) + tr(\nabla f(W^k)^T (W^{k+1} - W^k)) + \frac{L_f}{2} ||W^{k+1} - W^k||_2^2 + \lambda_k ||W^{k+1}||_0 \leq f(W^k) + tr(\nabla f(W^k)^T (W^{k+1} - W^k)) + \frac{L_f}{2} ||W^{k+1} - W^k||_2^2 + \lambda_k ||W^{k+1}||_0 \leq f(W^k) + \lambda_k ||W^{k+1}||_0 = \varphi_k(W^k), \tag{12}$$

where the last inequality follows from (7). The above inequality implies that for a fixed $\lambda$, $\varphi_k(W^k)$ is non-increasing and moreover,

$$\varphi_k(W^k) - \varphi_k(W^{k+1}) \geq \frac{L_f}{2} ||W^{k+1} - W^k||_2^2 \tag{13}$$

for a fixed $\lambda$, $\varphi_k(W^k)$ is non-increasing and moreover,

$$\varphi_k(W^k) - \varphi_k(W^{k+1}) \geq \frac{L_f}{2} ||W^{k+1} - W^k||_2^2 \tag{13}$$
Since $f(\mathbf{W})$ is bounded below, it then follows that $\varphi_{\lambda_k}(\mathbf{W}^k)$ is bounded below. Hence, $\varphi_{\lambda_k}(\mathbf{W}^k)$ converges to a finite value as $k \to \infty$ and a local optimal solution $\mathbf{W}_{\lambda_k}$ can be achieved.

Since the $\lambda$ is monotone decreased, and $\mathbf{W}_{\lambda_k}$ is set as the initial solution for HIHT in $\lambda_{k+1}$, we obtain that:

$$\varphi_{\lambda_k}(\mathbf{W}_{\lambda_k}^*) > \varphi_{\lambda_{k+1}}(\mathbf{W}_{\lambda_{k+1}}^*) = \varphi_{\lambda_{k+1}}(\mathbf{W}_{\lambda_{k+1}}^0) \geq \varphi_{\lambda_{k+1}}(\mathbf{W}_{\lambda_{k+1}}^{\lambda_{k+1}})$$

(14)

it implies that the objective value is monotone decreasing and a local optimal solution can be achieved by the proposed algorithm. We show an example of the objective value at each iteration in Fig. 1, it can be seen that the objective function values monotonically decrease at each iteration until convergence, which verifies the convergence of Alg. 1 experimentally.

**D. Acceleration of HIHT**

Inspired by [32], we derived an acceleration version for HIHT to reduce the computational time. For each fixed value of $\lambda$, Alg. 1 iterates to get a solution of problem (4) between steps 6-14. In the following, for acceleration of Alg. 1 we replace steps 6-14 by just calling one outer loop. The outline of the AHIHT method is described as Alg. 2.

**IV. EXPERIMENT**

The proposed $l_2, \theta$-norm regularization multi-class feature selection method is evaluated by several experiments. The experiments are divided into three parts: (1) We evaluate the proposed method in terms of the number of selected features and classification accuracy by KNN and softmax classifiers compared with baseline and other six state-of-art feature selection algorithms. (2) We evaluate the stability of our algorithm in terms of classification accuracy when the regularization parameter $\lambda$ and the number of nearest neighbour $k$ of KNN changed. We also evaluate the influence of initialization of AHIHT for feature selection. (3) We compare the convergence speed and computational time of some sparsity-based methods.

**A. DATA SETS DESCRIPTION**

We use eight biological benchmark datasets to validate the performance of our method in the experiments: Brain [34], Breast3 [35], Leukemia [36], Lung [37], Lymphoma [38], NCI [39], Prostate [40], and Srbct [41]. The 8 benchmark data sets were selected from Feiping Nie’s homepage table. 1 shows some characteristics of these datasets.

**TABLE I \DATASETS DESCRIPTION**

| Datasets  | #samples | #Features | #Classes |
|-----------|----------|-----------|----------|
| Brain     | 42       | 5597      | 5        |
| Breast3   | 95       | 4869      | 3        |
| Leukemia  | 38       | 3051      | 2        |
| Lung      | 203      | 3312      | 5        |
| Lymphoma  | 62       | 4026      | 3        |
| NCI       | 61       | 5244      | 8        |
| Prostate  | 102      | 6033      | 2        |
| Srbct     | 63       | 2308      | 4        |

**B. EXPERIMENT SETUP**

In the experiment, the feature selection performance is evaluated by classification accuracy on two popular classifiers, i.e. $K$ nearest neighbor (KNN) and softmax, we set up KNN with $k = 5$. For each data set, $\frac{3}{4}$ of samples per class are randomly selected for training and the rest samples are

\[\text{http://www.escience.cn/system/file?fileId=82035}\]
responsible for testing, ten repeated trials are carried out and average results are recorded for comparison. We compare our feature selection method with baseline (without feature selection) and six state-of-art feature selection algorithms: (1) two basic filter methods: Relief [10], and mRMR [12]; (2) two $l_{2,1}$-norm based methods: RLSR [22], and URAFS [23]; (3) two $l_{2,0}$-norm constrained methods: RPMFS [24], and EFSF [25]. The codes of Relief and mRMR are provided by the FEAST package [42], and others can be downloaded from the authors’ homepages.

For our method, the parameter $\lambda$ is first tuned in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^0\}$, and then fine-tuned in a small range, and for RLSR it is tuned from $\{10^{-5}, 10^{-3}, \ldots, 10^3\}$. For our method, the features correspond to the non-zero rows of $W$ are selected. For other methods, the number of selected features is tuned from $\{20, 40, \ldots, 500\}$. All other parameters take the default values as suggested by the authors. For RLSR, all training data are used as labeled data thus it can be seen as a supervised method here. The number of selected features with highest classification accuracy are recorded.

### C. Classification Performance

We evaluate the classification performance of our algorithm in this part. Tab. I shows the average results in terms of the number of selected feature and classification accuracy. From this table, it can be seen that for most datasets, the proposed $l_{2,0}$-FS can achieve highest accuracy with fewest selected features or comparable results to the best ones, since our method can find a more sparse solution by $l_{2,0}$-norm regularization instead of the $l_{2,1}$-norm regularization problem. RPMFS, ESFS and our algorithm both solve the original $l_{2,0}$-norm feature selection problem, but RPMFS and ESFS try to solve a equality constraint problem so that the number of selected features need to be tuned carefully to obtain a satisfactory classification result. From the numerical comparison, we can see that our method outperforms RPMFS and ESFS most of the time, which means that our method is more efficient than RPMFS and ESFS. The classification results demonstrate that our method can remove more redundant features while maintaining the discriminative performance.

Fig. 2 shows the classification accuracy V.S. the number of selected feature using softmax classifier. From it we can see that when the number of selected feature is small (less than 100), the classification results of our method can beat other compared methods consistently. And when the number of selected features is more than 100, our method still achieved best or comparable classification results. It also can be seen that the performance of RPMFS is more sensitive to the number of selected features than other methods. The classification results demonstrate that our method can select more informative features than other methods.

### D. Stability Evaluation

In this part, we evaluate the effect of $\lambda$ and the number of nearest neighbour $k$ of KNN in classification accuracy. The datasets of Breast3, Lung, NCI and Srbct are used for testing, and the experimental results are shown in Fig. 3. It can be seen from this figure that the performance is not sensitive to the number of nearest neighbour $k$ of KNN, and is not very sensitive to the $\lambda$ as long as it is in the range of $[10^{-4}, 10^{-2}]$, thus it is no need to speed much time to tune the value of $\lambda$. From this figure we can see the performance get worst when $\lambda = 10^{-1}$, the reason is that in this case the proposed method will obtain a very sparse solution and the number of selected features tend to zero.

$L_{2,0}$-norm is a non-convex problem, which may make the solution sensitive to initialization. We use three different kinds of initialization to explore the influence of initialization: zero initialization, random Gaussian distribution initialization, and random uniform distribution initialization. For Gaussian and uniform distribution, ten repeated trials are carried out and average results are recorded. Fig. 4 shows the results, in which datasets Breast3, Lung, and Srbct are used. It can be seen that these three different initialization methods can get the same results as the parameter $\lambda$ are set the same. The results show that our algorithm is not sensitive to initialization when apply to feature selection.

### E. Comparison of Convergence Speed and Time Consumption

Fig. 5 plots the objective function value for each iteration of the three $l_{2,0}$-norm based methods. As can be observed, our method can decrease the objective value quickly in early iterations, which indicates that our method converges fast than RPMFS and ESFS. Fig. 6 shows the average time consumption of each sparsity regularization method on the eight datasets, which also compares AHIHT with HIHT. It can be seen that the computational time of RPMFS and AHIHT are much less than the other four methods, the reason is that the computational complexity of RPMFS and AHIHT is in proportion to $d^2$ while for other methods it is in proportion to $d^3$. Comparing AHIHT with HIHT, it can be found that AHIHT can reduce the computational complexity of HIHT effectively, thus is more suitable for practical applications.

### F. Comparison of AHIHT and HIHT

In this part, we compare the performance of HIHT and its accelerated version in terms of accuracy and the number of selected features with different value of $\lambda$. Fig. 7 show the results, in which datasets Breast3, Lung, and NCI are used. From this figure it can be seen that, when $\lambda$ is less than 0.1, HIHT can get a more sparse solution than AHIHT, while AHIHT can get higher classification accuracy than HIHT, it means that the features selected by AHIHT is more useful for
**TABLE II**
The number of selected features and classification accuracy using selected features (red is the best result and blue is the second one).

| Dataset | KNN | Softmax |
|---------|-----|---------|
|         | Baseline | Relief | miRMR | RLSK | URAFS | RPMFS | ESFS | $l_2,0$-FS |
|         | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc | No.fea | acc |
| Brain   | 71.67 | 480 | 86.67 | 200 | 81.67 | 360 | 78.33 | 140 | 76.67 | 420 | 83.33 | 200 | 81.67 | 360 | 78.33 | 140 | 76.67 | 420 | 83.33 | 200 | 81.67 |
| Breast3 | 50.05 | 160 | 54.58 | 380 | 58.39 | 340 | 55.16 | 460 | 52.90 | 460 | 51.94 | 380 | 58.39 | 100 | 59.33 |
| Leukemia | 96.67 | 150 | 98.33 | 240 | 99.17 | 440 | 100.00 | 80 | 97.50 | 420 | 98.33 | 240 | 99.17 | 101 | 100.00 |
| Lung    | 94.39 | 380 | 94.39 | 200 | 94.09 | 480 | 94.39 | 380 | 94.39 | 480 | 94.39 | 380 | 94.39 | 380 | 94.39 | 380 | 94.39 |
| Lymphoma | 97.50 | 440 | 99.50 | 60 | 100.00 | 280 | 98.50 | 240 | 99.50 | 240 | 99.50 | 240 | 99.50 | 240 | 99.50 |
| NCI     | 73.89 | 440 | 74.44 | 240 | 73.89 | 320 | 74.44 | 240 | 71.67 | 420 | 73.89 | 320 | 73.89 | 320 | 73.89 |
| Prostate | 81.82 | 40 | 91.82 | 60 | 90.30 | 40 | 89.09 | 480 | 81.82 | 40 | 83.03 | 20 | 87.88 |
| Srbct   | 93.16 | 240 | 98.95 | 260 | 98.95 | 100 | 97.89 | 160 | 96.32 | 300 | 94.21 | 420 | 99.47 |

**Fig. 2**
The classification accuracy using selected features by Softmax. (A) Brain. (B) Breast3. (C) Lung. (D) NCI. (E) Prostate. (F) Srbct.

Classification than those of HIHT. What’s more, the results obtained by AHIHT demonstrate AHIHT is more robust to $\lambda$ than HIHT. For HIHT, it should tune a small value of $\lambda$ to get satisfactory result, while a small value of $\lambda$ will spend more computational time, which has been show in Fig. 6. In conclusion, AHIHT is more practical than HIHT for feature.
Fig. 3
Classification accuracy of the proposed method with respect to $\lambda$ as well as the number of neighbors. (a) Breast3. (b) Lung. (c) NCI. (d) SRBCT.

Fig. 4
Results of classification accuracy and the number of selected features with respect to different initialization. (a) Breast3. (b) Lung. (c) SRBCT.
selection.

V. CONCLUSIONS

In this paper, we proposed a novel method to solve the original $l_{2,0}$-norm regularization least square problem for multi-class feature selection, instead of solving its relaxed problem like most of other existing methods. A homotopy iterative hard threshold is proposed to optimize the proposed model which can obtain exact row-sparsity solution. Besides, in order to reduce the computational time of HIHT for feature selection task, an acceleration version of HIHT (AHIHT) is derived. Experiments on eight biological datasets show that we can achieve comparable or better classification performance comparing with other six state-of-the-art feature selection algorithms. In the future, we are interested in combining the $l_{2,1}$-norm loss function with the $l_{2,0}$-norm regularization.

REFERENCES

[1] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, “Clustering-guided sparse structural learning for unsupervised feature selection,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 9, pp. 2138–2150, 2014.
[2] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang, “Graph regularized feature selection with data reconstruction,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 689–700, 2016.
[3] N. K. Suchetha, A. Nihal, and P. Hrudya, “Comparing the wrapper feature selection evaluators on twitter sentiment classification,” 2019 Inter. Conf. Comp. Intell. in Data Sci. (ICCDIS), pp. 1–6, 2019.
[4] M. Zabihimayyan and D. Doran, “Fuzzy rough set feature selection to enhance phishing attack detection,” in 2019 IEEE Int. Conf. on Fuzzy Sys. (FUZZ-IEEE), 2019, pp. 1–6.
[5] O. S. Qasim, M. S. Mahmoud, and F. M. Hasan, “Hybrid binary dragonfly optimization algorithm with statistical dependence for feature selection,” Int. J. Math. Eng. and Manag. Sci., vol. 5, no. 6, pp. 1420–1428, 2020.
[6] B. Tang, S. Kay, and H. He, “Toward optimal feature selection in naive bayes for text categorization,” IEEE Trans. Knowl. Data Eng., vol. 28, pp. 1602–1606, 2016.
[7] C. Wan, Y. Wang, Y. Liu, J. Ji, and G. Feng, “Composite feature extraction and selection for text classification,” IEEE Access, vol. 7, pp. 35 208–35 219, 2019.
[8] Y. Saeyes, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” Bioinf., vol. 23, no. 19, pp. 2507–2517, 2007.
[9] B. Y. Lei, Y. J. Zhao, Z. W. Huang, and et al., “Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis,” Medical Image Anal., vol. 61, p. 101632, 2020.
[10] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in Proc. 9th Int. Workshop Mach. Learn., pp. 249–256, 1992.
[11] L. E. Raileanu and K. Stoffel, “Theoretical comparison between the gini index and information gain criteria,” Ann. Math. Artif. Intell., vol. 41, no. 1, pp. 77–93, 2004.
[12] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.
[13] M. A. Hall and L. A. Smith, “Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper,” in Proc. 12th Int. Florida Artif. Intell. Res. Society Conf., vol. 1999, pp. 235–239, 1999.
[14] I. Guyon, J. Weston, S. Bannhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Mach. Learn., vol. 46(1-3), pp. 389–422, 2002.
[15] J. R. Quinlan, “C4. 5: Programs for machine learning,” Amsterdam, The Netherlands: Elsevier, 2014.
[16] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, “Use of the zero-norm with linear models and kernel methods,” J. Mach. Learn. Res., vol. 3, pp. 1439–1461, 2003.
[17] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Statistical Soc. Series B (Methodological), vol. 58, pp. 267–288, 1996.
[18] J. Gai, Z. Sun, S. Ji, D. Tao, and T. Tan, “Feature selection based on structured sparsity: A comprehensive study,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 7, pp. 1490–1507, 2017.
[19] F. Nie, H. Huang, X. Cai, and C. H. Ding, “Efficient and robust feature selection via joint $l_{2,1}$-norms minimization,” in Proc. Advances Neural Inf. Process. Syst., pp. 1813–1821, 2010.
[20] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, “$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning,” in Proc. Int. Joint Conf. Artif. Intell, pp. 1589–1594, 2011.
[21] J. Zhang, J. Yu, J. Wang, and Z. Q. Zeng, “$l_{2,1}$ norm regularized fisher criterion for optimal feature selection,” Neurocomputing, vol. 166, pp. 455–463, 2015.
[22] X. Chen, F. Nie, G. Yuan, and J. Z. Huang, “Semi-supervised feature selection via rescaled linear regression,” in Proc. 26th Int. Joint Conf. Artif. Intell., pp. 1525–1531, 2017.
Fig. 7
CLASSIFICATION ACCURACY AND THE NUMBER OF SELECTED FEATURES WITH RESPECT TO $\lambda$ OF AHIHT AND HIHT. (A) BREAST3, (B) LUNG, (C) NCI.

[23] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, “Generalized uncorrelated regression with adaptive graph for unsupervised feature selection,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 5, pp. 1587–1595, 2019.

[24] X. Cai, F. Nie, and H. Huang, “Exact top-k feature selection via $l_{2,0}$-norm constraint,” in Proc. 23rd Int. Joint Conf. Artif. Intell., pp. 1240–1246, 2013.

[25] T. Pang, F. Nie, J. Han, and X. Li, “Efficient feature selection via $l_{2,0}$-norm constrained sparse regression,” IEEE Trans. on Knowl. Data Eng., vol. 31, no. 5, pp. 880–893, 2019.

[26] H. Yan, J. Yang, and J. Yang, “Robust joint feature weights learning framework,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 5, pp. 1327–1339, 2016.

[27] T. Blumensath and M. E. Davie, “Iterative thresholding for sparse approximations,” Fourier Anal. Appl., vol. 27, no. 5, pp. 629–654, 2008.

[28] ——, “Iterative hard thresholding for compressed sensing,” Appl. Comp. Harm. Anal., vol. 27, no. 3, pp. 265–274, 2009.

[29] Z. S. Lu, “Iterative hard thresholding methods for $l_0$ regularized convex cone programming,” Math. Prog., vol. 147, pp. 125–154, 2014.

[30] Q. Jiang, R. C. de Lamare, Y. Zakharov, S. Li, and X. He, “Knowledge-aided normalized iterative hard thresholding algorithms for sparse recovery,” 2018, pp. 1965–1969.

[31] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” IEEE Journ. Select. Topics Sign. Proc., vol. 1, no. 4, pp. 586–597, 2007.

[32] Z. Dong and W. Zhu, “Homotopy methods based on $l_0$ -norm for A. Bhattacharjee, W. G. Richards, I. Staunton, and et al., “Classification of human lung carcinomas by mra expression profiling reveals distinct adenocarcinoma subclasses,” in Proc. Nat. Academy Sci. United States America, vol. 98, no. 24, pp. 13 790–13 795, 2001.

[33] J. Ge, X. Li, H. Jiang, H. Liu, T. Zhang, M. Wang, and T. Zhao, “Picasso: A sparse learning library for high dimensional data analysis in R and python,” J. Mach. Learn. Res., vol. 20, pp. 44:1–44:5, 2019.

[34] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, and et al., “Prediction of central nervous system embryonal tumour outcome based on gene expression,” Nature, vol. 415, pp. 436–442, 2002.

[35] L. van’t Veer, H. Dai, M. van de Vijver, and et al., “Gene expression profiling predicts clinical outcome of breast cancer,” Nature, vol. 415, pp. 530–536, 2002.

[36] C. Nutt, D. Mani, R. A. Betensky, and et al., “Gene expression-based classification of malignant gliomas correlates better with survival than histological classification,” Cancer Research, vol. 63, no. 7, p. 1602, 2003.

[37] A. A. Alizadeh, M. B. Eisen, R. E. Davis, and et al., “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” Nature, vol. 403, pp. 503–511, 2000.

[38] S. S. Jeffrey, M. V. de Rijn, M. Waltham, and et al., “Systematic variation in gene expression patterns in human cancer cell lines,” Nature Genetics, vol. 24, no. 3, pp. 227–235, 2002.

[39] D. Singh, P. G. Febbo, K. Ross, and et al., “Gene expression correlates of clinical prostate cancer behavior,” Cancer Cell, vol. 1, pp. 203–209, 2002.

[40] J. Khan, J. S. Wei, M. Ringner, and et al., “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” in Proc. Nat. Academy Sci. United States America, vol. 7, no. 6, pp. 673–679, 2001.

[41] G. Brown, A. Pocock, M. J. Zhao, and M. Lujan, “A unifying framework for information theoretic feature selection,” J. Mach. Learn. Res., vol. 13, pp. 27–66, 2012.