J. Nonlinear Var. Anal. 5 (2021), No. 1, pp. 59-77
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.5.2021.1.05

ROBUST FEATURE SELECTION VIA NONCONVEX SPARSITY-BASED METHODS

NGUYEN THAI AN[1], PHAM DINH DONG[2], XIAOLONG QIN[3,*]

[1]*Department of Mathematics, College of Education, Hue University, 34 Le Loi, Hue City, Vietnam*
[2]*Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309, USA*
[3]*Department of Mathematics, Zhejiang Normal University, Zhejiang, China*

Abstract. In this paper, we propose a new model for supervised multiclass feature selection which has the $\ell_{2,1}$-norm in both the fidelity loss and the regularization terms with an additional $\ell_{2,0}$-constraint. This problem is challenging for applying available optimization methods because of the discontinuous and nonconvex nature of the $\ell_{2,0}$-norm. We first convert the constraint defined by the $\ell_{2,0}$-norm into a new constraint defined by a difference of two matrix norms. Then we reformulate the problem as an unconstrained problem using the exact penalty method. Based on a derived formula for the proximal mapping of this difference of matrix norms and Nesterov's smoothing techniques, the nonmonotonic accelerated proximal gradient method is applied to solve the unconstrained problem. Numerical experiments are conducted on many benchmark data sets to show the effectiveness of our proposed method in comparison with existing methods.

Keywords. Feature selection; $\ell_{2,0}$-norm constraint; Exact penalty; Proximal operator; Smoothing technique.

1. INTRODUCTION

Many machine learning applications require dealing with high dimensional data. For instance, in problems of face recognition [1, 2], bioinformatics [3] or video semantic recognition [4], each instance often has several hundreds or thousands of features. The high dimensional nature of the data requires much time and a lot of space to process as they usually contain noise and redundant features. In addition, an enormous amount of training data is required to ensure that learning algorithms work properly as in low dimensional settings. To overcome this difficulty, feature selection algorithms are used to preprocess data in order to identify a small subset of relevant and important features that can give a compact and accurate representation for the original data; see [3, 5, 6] and the references therein. Once the best features are selected, subsequent analysis such as visualization, regression or classification can be favorably applied. Feature selection is also useful for reducing storage requirements and training time, facilitating data visualization and data understanding, eliminating noise and improving the prediction performance. We refer the reader to [7, 8, 9] for more details.

*Corresponding author.
E-mail addresses: nguyenthaian@hueuni.edu.vn (N.T. An), ddpham@oakland.edu (P.D. Dong), qxlxajh@163.com (X. Qin).
Received July 24, 2020; Accepted October 5, 2020.

Machine learning methods for feature selection can be roughly divided into three classes: wrapper, filter, and embedded methods. Wrapper methods exploit the predictive performance of a predefined learning algorithm to evaluate the quality of selected features; see, e.g., [6, 10]. Filter-type methods usually rank each feature individually based on certain statistical measures, where no learning algorithm is involved. Representative filter-type methods include the **reliefF** [11, 12, 13], **T-test** [14, 15] and **Fisher** score [16, 17]. Embedded-type methods embed the feature selection procedure into a learning model and only a single optimization problem is involved; see, e.g., [18, 19]. Our main goal in this paper is to study an embedded-type method for multiclass feature selection using sparsity-based feature selection models.

Many sparsity-based feature selection models have been proposed and have shown promising performance in practical applications; see [7, 9] and the references therein. In these models, the feature selection problem is reformulated as an optimization problem in which a sparsity-inducing regularization term is added to the fitting error. Sparsity-inducing regularization terms defined by the ℓ_1-norm or the ℓ_0-norm are often used in binary classification, while those defined by the $\ell_{2,1}$-norm or the $\ell_{2,0}$-norm are often used in multiclassification. The presence of sparse regularization terms shrinks irrelevant feature coefficients towards zero and then the corresponding features should simply be eliminated. From the sparsity perspective, the ℓ_0-norm or the $\ell_{2,0}$-norm is more desirable to select the features because it can induce the sparsest solution, i.e., each feature should be associated with either the zero coefficient or a large coefficient. However, the resulting optimization problem is of nonconvex nature and is very difficult to solve; see [20]. Therefore, it is often relaxed using the ℓ_1- or the $\ell_{2,1}$-regularization. The ℓ_1-norm based feature selection method, also known as Lasso [21], is commonly used for binary-class data sets. In multi-task learning, $\ell_{2,1}$-norm regularization has been successfully employed to couple feature selection across tasks. Many $\ell_{2,1}$-norm based feature selection methods for multiclass data have been proposed over the last decade; see [22, 23] and the references therein.

In this paper, we propose a new model for multiclass feature selection which can be seen as the combination of two popular sparsity-inducing methods proposed by Nie et al. [22] and Cai et al. [24].

Paper's Contributions:

- Our feature selection model, which has the $\ell_{2,1}$-norm in both the fidelity loss and the regularization terms with an additional $\ell_{2,0}$-constraint, inherits the merits from existing $\ell_{2,1}$- and $\ell_{2,0}$- feature selection approaches. The proposed model is robust to outliers and it can select features across all instances in the data with joint sparsity. In addition, with the use of an additional $\ell_{2,0}$-constraint, the model becomes very efficient for selecting essential features in the data.
- We provide a new approach to deal with the discontinuous and nonconvex $\ell_{2,0}$-constraint of the form $\|W\|_{2,0} \leq K$ by converting it into a continuous penalized term represented as the difference of two matrix norms.
- We provide a new explicit formula for computing the proximal mapping for the difference of two matrix norms and therefore open up the possibility of applying the proximal gradient methods and its variants to deal with the $\ell_{2,0}$-constraint.

Notation: All vectors are in column format. For a matrix $W = (w_{ij}) \in \mathbb{R}^{n\times c}$, we denote its ith row by w^i and its jth column by w_j, i.e.,

$$w^i = \begin{pmatrix} w_{i1} \\ \vdots \\ w_{ic} \end{pmatrix} \text{ and } w_j = \begin{pmatrix} w_{1j} \\ \vdots \\ w_{nj} \end{pmatrix}.$$

Given $W = (w_{ij}) \in \mathbb{R}^{n\times c}$ and $Z = (z_{ij}) \in \mathbb{R}^{n\times c}$, define the inner product:

$$\langle W, Z\rangle := \sum_{i,j} w_{ij} z_{ij} = \text{tr}(W^T Z).$$

This inner product induces the *Frobenius norm* of W:

$$\|W\|_F := \sqrt{\langle W, W\rangle} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}^2} = \sqrt{\sum_{i=1}^{n}\|w^i\|_2^2}.$$

Note that the squared Frobenious norm function defined on the Euclidean space $E := \mathbb{R}^{n\times c}$ is differentiable with gradient given by $\nabla\|\cdot\|_F^2(W) = 2W$. The $\ell_{2,1}$-norm is defined by

$$\|W\|_{2,1} := \sum_{i=1}^{n}\|w^i\|_2 = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{c} w_{ij}^2}.$$

Throughout this paper, we use $X = [x_1,\ldots,x_m] \in \mathbb{R}^{n\times m}$ to denote the training data set of m instances in $\mathbb{R}^n$ in which each instance x_i belongs to exactly one class from $\{1,\ldots,c\}$. We use $Y = [y_1,\ldots,y_m]^\top \in \{0,1\}^{m\times c}$ to denote the associated label matrix in which the ij-entry of Y is 1 if x_i is associated with the jth class and is 0 otherwise.

Paper's organization: The rest of this paper is organized as follows. In Section 2, we present our model as a nonconvex discontinuous optimization problems. Section 3 is devoted to reformulating of the problem under consideration as an unconstrained continuous problem based on the exact penalty method. Nesterov's smoothing techniques for functions with matrix variables are presented in Section 4. In Section 5, accelerated proximal gradient methods based on a new proximal mapping are employed to solve the problem. Finally, numerical experiments on many benchmark data sets are presented in Section 6.

2. Problem Formulations

Given a *training matrix* $X \in \mathbb{R}^{n\times m}$ and the associated label matrix $Y \in \{0,1\}^{m\times c}$, standard *multivariate regression problems* solve the following optimization problem to find the projection matrix $W = [w_1,\ldots,w_c] \in \mathbb{R}^{n\times c}$ and the bias $b \in \mathbb{R}^c$:

$$\min_{W,b}\sum_{i=1}^{m}\|W^\top x_i + b - y_i\|_2^2 = \sum_{i=1}^{m}\sum_{j=1}^{c}(w_j^\top x_i + b_j - y_{ij})^2.$$

For simplicity, the bias b can be absorbed into W when the constant value 1 is added as an additional dimension for each sample x_i for $i = 1,\ldots,m$. The problem then can be rewritten in a matrix form as

$$\min_{W} \quad \|X^\top W - Y\|_F^2.$$

To select a subset of relevant features, sparsity-inducing feature selection methods [7] aim at solving the following problem with an additional regularization term:

$$\min_{W} \quad \|X^\top W - Y\|_F^2 + \gamma\mathscr{R}(W),$$

where $\gamma > 0$ is a trade-off parameter to balance the overfitting and interpretability of the model. Once a solution W is obtained, the features can be ranked according to the ℓ_2-norm of its rows: the higher the value, the more relevant the feature is. One of the key drawbacks of the squared error loss lies in the fact that it is prone to outliers. Some outliers with large squared errors of the form $\|W^\top x_i - y_i\|^2$ can easily dominate the objective function. To reduce the effect of outliers, other measurements should be used instead of the squared error.

Among the many effective methods is the *robust feature selection method* (**RFS**) proposed by Nie et al. in [22] which used the $\ell_{2,1}$-norm on both fidelity loss and regularization term. Its model is as follows

$$\min_{W} \quad \|X^\top W - Y\|_{2,1} + \gamma\|W\|_{2,1}. \tag{2.1}$$

This model has showed its efficiency in selecting meaningful features across all instances of the data with joint sparsity and also in reducing the effect of outliers. It has been successfully applied to train biological

data, outperforming other methods in many aspects. Problem (2.1) is convex and a global solution can be obtained; see [22].

The *top-K feature selection method* (**TopK**, for brevity) proposed by Cai et al. [24] with an $\ell_{2,1}$-loss and an explicit $\ell_{2,0}$- equality constraint solves

$$\min_{W} \quad \|X^\top W - Y\|_{2,1} \quad \text{subject to} \quad \|W\|_{2,0} = K, \tag{2.2}$$

where $\|W\|_{2,0}$ stands for the number of nonzero rows of W. Although problem (2.2) is nonconvex, an effective alternating algorithm based on the augmented Lagrangian method in [24] can find good approximation solutions. The main advantage of this model is that it uses the $\ell_{2,0}$-norm constraint instead of incorporating it as a regularization term and therefore it alleviates the burden of adjusting the regularized parameters in practice. This model turns out to be very effective when the number of selected features is quite small (e.g., K is less than 20). However, this is also a drawback of this method because in many cases (especially, for large K) we cannot control the sparsity level of the final solution; see section 6.4.

In this paper, we propose a new model which can be seen as a combination of the above two models:

$$\begin{aligned} \min_{W} \quad & \|X^\top W - Y\|_{2,1} + \gamma \|W\|_{2,1} \\ \text{subject to} \quad & \|W\|_{2,0} \le K. \end{aligned} \tag{2.3}$$

In this model, the $\ell_{2,1}$-regularization plays a significant role in selecting features with joint sparsity, i.e., each feature either has small coefficient for all data points or has large coefficient over all data points. The $\ell_{2,0}$-constraint controls the sparsity level, i.e., it controls number of nonzero rows in W. Because of its nonsmoothness and nonconvexity, this problem is very challenging.

3. Exact Penalty Methods for Feature Selection

In this section, we use the *exact penalty method* to introduce an unconstrained optimization model for dealing with the constrained optimization problem (2.3). The difficulty when solving problem (2.3) comes from the discontinuity and nonconvexity of the $\ell_{2,0}$-norm constraint. Our approach is to use a penalty term defined by the difference of the $\ell_{2,1}$-norm and the $\ell_{K,21}$-norm in order to model an unconstrained optimization which has the same optimal solution set as the constrained one when the penalty parameter is sufficiently large.

Definition 3.1. Given an $n \times c$ matrix W and a positive integer K such that $1 \le K \le n$, we rearrange all rows of W in the descent order of their ℓ_2-norms, i.e.,

$$\|w^{(1)}\|_2 \ge \ldots \ge \|w^{(K)}\|_2 \ge \ldots \ge \|w^{(n)}\|_2,$$

and then define

$$\|W\|_{K,21} := \|w^{(1)}\|_2 + \ldots + \|w^{(K)}\|_2.$$

From this definition, we can see that

$$\|W\|_{K,21} = \max_{v} \left\{ \sum_{i=1}^{n} v_i \|w^i\|_2 \;\Big|\; v_i \in \{0,1\}, \sum_{i=1}^{n} v_i = K \right\}.$$

The following proposition allows us to convert the cardinality constraint in (2.3) into a DC constraint and gives a formula for computing the subdifferential of the $\ell_{K,21}$-norm function.

Given a finite dimensional Euclidean space E and a convex function $f\colon E \to \mathbb{R}$, recall that the *subdifferential in the sense of convex analysis* of f at $\bar{x} \in E$ is defined by

$$\partial f(\bar{x}) := \big\{ w \in E \;\big|\; \langle w, x - \bar{x} \rangle \le f(x) - f(\bar{x}) \;\text{ for all } x \in E \big\}.$$

The reader is referred to [25] for more details on subdifferentials of convex functions.

Proposition 3.1. *For any matrix W of n rows and any positive integer $1 \le K \le n$, we have*
(i) $\|W\|_{2,0} \le K$ *if and only if* $\|W\|_{2,1} - \|W\|_{K,21} = 0$.
(ii) *The function $f(W) := \|W\|_{K,21}$ is convex with the subdifferential given by*

$$\partial f(W) = \operatorname{conv}\{\operatorname{diag}(v)p(W) \mid v \in I(W)\}, \tag{3.1}$$

where $p(W)$ is a set consisting of all matrices the same size as W defined row-wisely by

$$[p(W)]_i := \begin{cases} \mathbb{B}, & \text{if } \|w^i\|_2 = 0, \\ \left\{\frac{w^i}{\|w^i\|_2}\right\}, & \text{if } \|w^i\|_2 \neq 0, \end{cases}$$

and the active index set at W is defined by

$$I(W) := \left\{ v \mid v \in \{0,1\}^n, \sum_{i=1}^n v_i = K, \sum_{i=1}^n v_i \|w^i\|_2 = \|W\|_{K,21} \right\}.$$

Proof. (i) Assume that $\|W\|_{2,0} \le K$. This means that W has at most K rows that are not zero vectors. When we rearrange all rows of W in the descent order of their ℓ_2-norm, i.e.,

$$\|w^{(1)}\|_2 \ge \ldots \ge \|w^{(K)}\|_2 \ge \ldots \ge \|w^{(n)}\|_2,$$

these nonzero rows must be among $\{w^{(1)}, \ldots, w^{(K)}\}$. This means that all the rows $w^{(j)}$ for $j = K+1, \ldots, n$ must be zero vectors, and so

$$\|W\|_{2,1} = \sum_{i=1}^n \|w^i\|_2 = \sum_{i=1}^n \|w^{(i)}\|_2 = \sum_{i=1}^K \|w^{(i)}\|_2 = \|W\|_{K,21}.$$

The converse implication can be proved similarly. For (ii), since $f(W) = \|W\|_{K,21}$ is a pointwise supremum of a finite number of convex functions, the function f is convex. Recall that the subdifferential of the Euclidean norm function is given by

$$\partial(\|\cdot\|_2)(w) = \begin{cases} \mathbb{B}, & \text{if } \|w\|_2 = 0, \\ \left\{\frac{w}{\|w\|_2}\right\}, & \text{if } \|w\|_2 \neq 0. \end{cases}$$

The subdifferential formula (3.1) is now obtained from the well-known rule for max functions in convex analysis; see, e.g., [25]. □

Based on Proposition 3.1, we introduce the following penalized version of problem (2.3):

$$\min_W \quad \mathscr{F}(W) = \|X^\top W - Y\|_{2,1} + \gamma \|W\|_{2,1} + \rho\left(\|W\|_{2,1} - \|W\|_{K,21}\right). \tag{3.2}$$

In the theorem below, we establish an important relationship between problem (3.2) and problem (2.3), which allows us to use a fixed threshold for ρ when optimizing (3.2).

Theorem 3.1. *Let $\|X\|_\infty := \max_{i,j} |x_{ij}|$ be the largest absolute value of all entries of X. If the penalty parameter ρ satisfies $\rho > \rho_{\max}$, where*

$$\rho_{\max} := m\|X\|_\infty - \gamma,$$

then any solution $\overline{W}$ of the penalized problem (3.2) is also a solution of problem (2.3).

Proof. Let $\overline{W}$ be a solution of (3.2). It follows from Proposition 3.1 (i) that if $\overline{W}$ is feasible for (2.3), i.e., $\|\overline{W}\|_{2,1} - \|\overline{W}\|_{K,21} = 0$ or equivalently $\|\overline{W}\|_{2,0} \le K$, then it is a solution of (2.3).

Let us now show that if $\rho > \rho_{\max}$, then $\overline{W}$ is indeed a feasible solution of problem (2.3). Suppose by contradiction that $\overline{W}$ is not feasible for (2.3) or $\|\overline{W}\|_{2,1} - \|\overline{W}\|_{K,21} > 0$. Let us fix an index $\bar{v} \in I(\overline{W})$. Since $\|\overline{W}\|_{2,1} - \|\overline{W}\|_{K,21} > 0$, we have $K < n$ and $\sum_{\bar{v}_t = 0} \|\overline{w}^t\|_2 > 0$. Define an $n \times c$ matrix $\widetilde{W}$ as follows

$$\widetilde{w}^t := \begin{cases} \overline{w}^t, & \text{if } \bar{v}_t = 1 \\ 0_{\mathbb{R}^c}, & \text{if } \bar{v}_t = 0 \end{cases} \text{ for all } t = 1, \ldots, n.$$

It follows that

$$\|\widetilde{W}\|_{2,0} \leq K, \ \ \|\overline{W}\|_{K,21} = \|\widetilde{W}\|_{K,21} \text{ and } \|\overline{W}\|_{2,1} - \|\widetilde{W}\|_{2,1} = \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2. \tag{3.3}$$

By the triangle inequality, for each $i \in \{1,\ldots,m\}$ we have

$$\begin{aligned}
\|\widetilde{W}^\top x_i - y_i\|_2 &= \|x_{1i}\widetilde{w}^1 + \ldots + x_{ni}\widetilde{w}^n - y_i\|_2 \\
&= \left\| \sum_{\bar{v}_t=1} x_{ti}\widetilde{w}^t - y_i \right\|_2 = \left\| \sum_{\bar{v}_t=1} x_{ti}\overline{w}^t - y_i \right\|_2 \\
&= \left\| \sum_{\bar{v}_t=1} x_{ti}\overline{w}^t + \sum_{\bar{v}_t=0} x_{ti}\overline{w}^t - y_i - \sum_{\bar{v}_t=0} x_{ti}\overline{w}^t \right\|_2 \\
&= \left\| \sum_{t=1}^{n} x_{ti}\overline{w}^t - y_i - \sum_{\bar{v}_t=0} x_{ti}\overline{w}^t \right\|_2 \\
&\leq \left\| \sum_{t=1}^{n} x_{ti}\overline{w}^t - y_i \right\|_2 + \left\| \sum_{\bar{v}_t=0} x_{ti}\overline{w}^t \right\|_2 \\
&\leq \|\overline{W}^\top x_i - y_i\|_2 + \sum_{\bar{v}_t=0} |x_{ti}| \|\overline{w}^t\|_2 \\
&\leq \|\overline{W}^\top x_i - y_i\|_2 + \|X\|_\infty \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2,
\end{aligned}$$

This implies that, for each $i \in \{1,\ldots,m\}$, the following holds:

$$\|\overline{W}^\top x_i - y_i\|_2 - \|\widetilde{W}^\top x_i - y_i\|_2 \geq -\|X\|_\infty \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2. \tag{3.4}$$

Combining (3.3) and (3.4), we have

$$\begin{aligned}
&\mathscr{F}(\overline{W}) - \mathscr{F}(\widetilde{W}) \\
&= \left[\sum_{i=1}^{m} \|\overline{W}^\top x_i - y_i\|_2 + \gamma \|\overline{W}\|_{2,1} + \rho \left(\|\overline{W}\|_{2,1} - \|\overline{W}\|_{K,21} \right) \right] \\
&\quad - \left[\sum_{i=1}^{m} \|\widetilde{W}^\top x_i - y_i\|_2 + \gamma \|\widetilde{W}\|_{2,1} + \rho (\|\widetilde{W}\|_{2,1} - \|\widetilde{W}\|_{K,21}) \right] \\
&\geq -m\|X\|_\infty \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2 + \gamma \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2 + \rho \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2 \\
&= \sum_{\bar{v}_t=0} \|\overline{w}^t\|_2 (\rho + \gamma - m\|X\|_\infty) > 0,
\end{aligned}$$

by the facts that $\sum_{\bar{v}_t=0} \|\overline{w}^t\|_2 > 0$ and $\rho > m\|X\|_\infty - \gamma$. This is a contradiction because $\overline{W}$ is a solution of (3.2). The proof is now complete. □

Remark 3.1. Theorem 3.1 ensures that problem (2.3) can be solved by minimizing the objective function $\mathscr{F}$ in the penalized problem (3.2) with a fixed parameter of $\rho > \rho_{\max}$. However, a large value of ρ could lead to an approximation solution that is too sparse and is not good for selecting features, especially when K is quite large. We observe from practice that, in many cases, using a smaller value of ρ (not necessary greater than $\rho_{\max}$) gives better results.

4. NESTEROV'S SMOOTHING TECHNIQUES

Using the exact penalty method, we have been able to convert the constrained optimization problem (2.3) with the constraint defined by a discontinuous function to the unconstrained optimization problem (3.2) with a continuous objective function. However, the nonsmoothness and nonconvexity still exist in this new model. One of the key components in our method for solving problem (3.2) involves approximating nonsmooth functions by smooth functions to reduce the *level of nonsmoothness*. The smoothing technique introduced by Nesterov in his seminal paper [26] plays crucial role in our method.

Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^c$ be two given vectors and let Q be a convex compact subset of $\mathbb{R}^c$. Consider the function $f\colon \mathbb{R}^{n\times c} \longmapsto \mathbb{R}$ of matrix variable given by

$$f(W) := \max\left\{\langle W^\top x - y, u\rangle \mid u \in Q\right\}, W \in \mathbb{R}^{n\times c}.$$

This function is convex with respect to W but nonsmooth in general. Given $\mu > 0$, consider the function f_μ defined by

$$f_\mu(W) := \max\left\{\langle W^\top x - y, u\rangle - \frac{\mu}{2}\|u\|_2^2 \mid u \in Q\right\},$$

$W \in \mathbb{R}^{n\times c}$. The theorem below shows that the smooth approximation f_μ as well as its gradient has closed forms that can be expressed in terms of the Euclidean projection. The proof follows directly from [27, Theorem 2.1] and [26, Theorem 1].

Theorem 4.1. *The function f_μ has the following explicit representation:*

$$f_\mu(W) = \frac{\|W^\top x - y\|_2^2}{2\mu} - \frac{\mu}{2}\left[d(\frac{W^\top x - y}{\mu}; Q)\right]^2$$

and is continuously differentiable on $\mathbb{R}^n$ with gradient given by

$$\nabla f_\mu(W) = x\left[P_Q\left(\frac{W^\top x - y}{\mu}\right)\right]^\top.$$

The gradient ∇f_μ is Lipschitz with constant $L = \frac{\|x\|_2^2}{\mu}$. In addition, the following estimate holds

$$f_\mu(W) \le f(W) \le f_\mu(W) + \frac{\mu}{2}C \tag{4.1}$$

for all $W \in \mathbb{R}^{n\times c}$, where $C = \sup\{\|q\|_2^2 \mid q \in Q\} < +\infty$.

Corollary 4.1. *The function $g(W) := \|W\|_{2,1}$ with variable $W \in \mathbb{R}^{n\times c}$ has a smooth approximation given by*

$$g_\mu(W) = \sum_{i=1}^n \left[\frac{\|w^i\|_2^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{w^i}{\mu}; I\!B\right)\right]^2\right],$$

where w^i is the ith row of W and $I\!B$ stands for the closed unit ball in $\mathbb{R}^c$. This function is smooth with gradient $\nabla g_\mu(W)$ is given row-wisely by

$$[\nabla g_\mu(W)]_i = \left[P_{I\!B}\left(\frac{w^i}{\mu}\right)\right]^\top \quad \text{for all } i = 1,\ldots,n.$$

This gradient $\nabla g_\mu(W)$ is Lipschitz continuous with constant $L = \frac{1}{\mu}$. In addition,

$$g_\mu(W) \le g(W) \le g_\mu(W) + \frac{\mu}{2}n, \text{ for all } W \in \mathbb{R}^{n\times c}.$$

5. Solving the Penalized Problem via the Accelerated Proximal Gradient Method

The *proximal gradient method* (see, e.g., [28]) is one of the most widely used convex optimization algorithms for solving convex optimization problems with nonsmooth objective functions. Recent generalizations to the nonconvex setting in [29, 30] allow ones to deal with nonconvex objective functions. In this section, we provide a new formula for finding the proximal mapping of the penalty term defined by a difference of the $\ell_{2,1}$-norm and the $\ell_{K,21}$-norm, which allows us to solve the optimization problem (3.2) using the proximal gradient method and its variants.

Let E be a finite dimensional Euclidean space. The proximal gradient method has been widely used for solving optimization problems of the form

$$\min_{W} \quad \mathscr{L}(W) := \Theta(W) + \rho\, \Psi(W),$$

where $\Theta\colon E \to \mathbb{R}$ and $\Psi\colon E \to (-\infty, \infty]$ are two convex functions. Given a constant $\alpha > 0$, define

$$\mathbf{prox}_{\alpha\Psi}(U) := \operatorname*{argmin}_{W} \left\{ \frac{1}{2}\|W - U\|_F^2 + \alpha\Psi(W) \right\} \tag{5.1}$$

for $U \in E$. Note that $\mathbf{prox}_{\alpha\Psi}(U)$ is always a singleton when Ψ is proper, convex, and lower semicontinuous. Assuming further that Θ is Fréchet differentiable with L - Lipschitz continuous gradient, the classical proximal gradient algorithm defines a sequence $\{W^k\}$ by choosing a starting point $W^0 \in E$ and set

$$W^{k+1} = \mathbf{prox}_{\frac{\rho}{L}\Psi}\left(W^k - \frac{1}{L}\nabla\Theta(W^k) \right) \text{ for } k \in \mathbb{N}.$$

It is well-known that this algorithm has the convergence rate of $O(\frac{1}{k})$. By incorporating an extrapolation step, the accelerated version [31, 32] defined by

$$\begin{aligned} W^{k+1} &= \mathbf{prox}_{\frac{\rho}{L}\Psi}\left(U^k - \frac{1}{L}\nabla\Theta(U^k) \right), \\ U^{k+1} &= W^{k+1} + \frac{t_k - 1}{t_{k+1}}(W^{k+1} - W^k) \text{ for } k \in \mathbb{N}, \end{aligned}$$

where $U^0 = W^0 \in E$, $t_0 = 1$ and $t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2}$; improves the convergence rate to $O(\frac{1}{k^2})$. The proximal gradient algorithm and its accelerated versions have recently been extended to the nonconvex setting with the same requirement on Θ but Ψ could be nonconvex; see [29, 30]. The state-of-art proximal gradient algorithm for nonconvex programming is the nonmonotonic APG proposed by Li and Lin in [30]. This **nmAPG** extends Beck and Teboulle's monotone APG [33] and has a rigorous convergence guarantee which ensures that every accumulation point of the iterative sequence is a critical point of the problem; see [30, Theorem 1].

Proposition 5.1. (See [34]) *If Ψ is proper, lower-semicontinuous and $\inf\Psi > -\infty$, then the solution set of (5.1) is nonempty and compact.*

From Theorem 3.1, to solve feature selection problem (2.3) we solve its penalized problem (3.2) with a fixed value of penalty parameter $\rho > \rho_{\max}$. This problem can be written as

$$\min_{W} \quad \mathscr{F}(W) = \mathscr{G}(W) + \rho\, \mathscr{H}_K(W),$$

where

$$\begin{aligned} \mathscr{G}(W) &= \|X^\top W - Y\|_{2,1} + \gamma\|W\|_{2,1}, \\ \mathscr{H}_K(W) &= \|W\|_{2,1} - \|W\|_{K,21}. \end{aligned} \tag{5.2}$$

In what follows, we will derive an explicit formula of the proximal operator for $\mathscr{H}_K$, which plays a crucial role in applying the proximal gradient method.

Proposition 5.2. *Let $\widetilde{W} \in \mathbf{prox}_{\alpha\mathscr{H}_K}(U)$. Then the ith row of $\widetilde{W}$ is zero if and only if the ith row of U is zero.*

Proof. Given $U \in \mathbb{R}^{n\times c}$ and $\alpha \geq 0$, define

$$\Phi(W) := \frac{1}{2}\|W-U\|_F^2 + \alpha(\|W\|_{2,1} - \|W\|_{K,21}). \tag{5.3}$$

It follows from (5.1) that

$$\widetilde{W} \in \mathbf{prox}_{\alpha\mathscr{H}_K}(U) \Longleftrightarrow \widetilde{W} \in \operatorname*{argmin}_{W\in\mathbb{R}^{n\times c}} \Phi(W).$$

Note that $\|W\|_{2,1} - \|W\|_{K,21} \geq 0$ for all $W \in \mathbb{R}^{n\times c}$. By Proposition 5.1, the set $\mathbf{prox}_{\alpha\mathscr{H}_K}(U)$ is nonempty and compact. Consider the case where U has some zero row, say $u^i = 0$. Define the matrix $\widehat{W}$ by replacing $\widetilde{w}^i$ by the zero vector. Since $\widetilde{W}$ is a minimizer of Φ,

$$\frac{1}{2}\|w^i\|^2 \leq \Phi(\widetilde{W}) - \Phi(\widehat{W}) \leq 0.$$

This implies $\widetilde{w}^i = 0$.

Let us now prove that if $\widetilde{w}^i = 0$, then $u^i = 0$. Assume by contradiction that $u^i \neq 0$. We define a new $n \times c$ matrix $\widehat{W}$ such that the ith row is zero and the other rows are the same as those of W. Since we replace a row in $\widetilde{W}$ with a row with zero norm to obtain $\widehat{W}$, we have

$$\|\widetilde{W}\|_{21} - \|\widetilde{W}\|_{K,21} \geq \|\widehat{W}\|_{21} - \|\widehat{W}\|_{K,21}.$$

Taking into account that $\|u^i\|_2 > 0$, we thus have

$$\begin{aligned}\Phi(W) &= \frac{1}{2}\sum_{j\neq i}\|w^j - u^j\|_2^2 + \frac{1}{2}\|u^i\|_2^2 + \alpha(\|W\|_{21} - \|W\|_{K,21}) \\ &> \frac{1}{2}\sum_{j\neq i}\|w^j - u^j\|_2^2 + \alpha(\|\widehat{W}\|_{21} - \|\widehat{W}\|_{K,21}) \\ &= \Phi(\widehat{W}).\end{aligned}$$

This contradicts to the fact that $\widetilde{W}$ is a solution. Thus, we have justified $u^i = 0$ iff the corresponding row $\widetilde{w}^i = 0$. □

Theorem 5.1. *Let $U \in \mathbb{R}^{n\times c}$ and let $\widetilde{v} \in I(U)$. An element $\widetilde{W} \in \mathbf{prox}_{\alpha\mathscr{H}_K}(U)$ can be chosen by*

$$\widetilde{w}^i := \begin{cases} u^i, & \text{if } \widetilde{v}_i = 1 \text{ or } u^i = 0, \\ \left(1 - \frac{\alpha}{\|u^i\|}\right)u^i, & \text{if } \widetilde{v}_i = 0 \text{ and } u^i \neq 0. \end{cases}$$

Proof. By Proposition 5.2, we only need to consider the case where all rows of U are nonzero (and thus all rows of W are nonzero). Note that the function Φ defined in (5.3) can be represented as a difference of convex functions $\Phi = \Phi_1 - \Phi_2$, where

$$\Phi_1(W) := \frac{1}{2}\|W-U\|_F^2 + \alpha\|W\|_{2,1} \text{ and } \Phi_2(W) := \alpha\|W\|_{K,21}.$$

By the optimality condition for differences of convex functions (see, e.g., [35]), if $W \in \mathbf{prox}_{\alpha\mathscr{H}_K}(U)$ then

$$\partial\Phi_2(W) \subset \partial\Phi_1(W).$$

It follows that for any $v \in I(W)$, we have

$$(w^i - u^i) + \alpha(1-v_i)\frac{w^i}{\|w^i\|_2} = 0 \text{ for } i = 1,\ldots,n.$$

This implies that

$$u^i = \begin{cases} w^i, & \text{if } v_i = 1, \\ w^i + \alpha \frac{w^i}{\|w^i\|_2}, & \text{if } v_i = 0. \end{cases} \tag{5.4}$$

Thus, for any $W \in \mathbf{prox}_{\alpha \mathscr{H}_K}(U)$ and $v \in I(W)$, we have

$$\|u^i\|_2 = \begin{cases} \|w^i\|_2, & \text{if } v_i = 1, \\ \|w^i\|_2 + \alpha, & \text{if } v_i = 0. \end{cases} \tag{5.5}$$

Fix $\widetilde{v} \in I(U)$ and define $\widetilde{W}$ as follows

$$\widetilde{w}^i := \begin{cases} u^i, & \text{if } \widetilde{v}_i = 1, \\ \left(1 - \frac{\alpha}{\|u^i\|}\right) u^i, & \text{if } \widetilde{v}_i = 0. \end{cases}$$

We will show that $\widetilde{W} \in \mathbf{prox}_{\alpha \mathscr{H}_K}(U)$. It suffices to show that $\Phi(\widetilde{W}) \le \Phi(W)$ for all W satisfying the necessary optimality condition (5.4). By the definition of $\widetilde{W}$, we have

$$\|u^i\|_2 = \begin{cases} \|\widetilde{w}^i\|_2, & \text{if } \widetilde{v}_i = 1, \\ \|\widetilde{w}^i\|_2 + \alpha, & \text{if } \widetilde{v}_i = 0. \end{cases} \tag{5.6}$$

Since $\alpha \ge 0$, we have from (5.6) that $\|\widetilde{w}^i\|_2 \le \|u^i\|_2$ for all $i = 1, \ldots, n$. We first claim that $\widetilde{v} \in I(\widetilde{W})$. Indeed, taking any $\widehat{v} \in \{0,1\}^n$ with $\sum_{i=1}^n \widehat{v}_i = K$, we have

$$\sum_{i=1}^n \widetilde{v}_i \|\widetilde{w}^i\|_2 = \sum_{\widetilde{v}_i = 1} \|\widetilde{w}^i\|_2 = \sum_{\widetilde{v}_i = 1} \|u^i\|_2 = \sum_{i=1}^n \widetilde{v}_i \|u^i\|_2 \ge \sum_{i=1}^n \widehat{v}_i \|u^i\|_2 \ge \sum_{i=1}^n \widehat{v}_i \|\widetilde{w}^i\|_2,$$

where the first inequality is due to $\widetilde{v} \in I(U)$ and the second inequality is due to $\|\widetilde{w}^i\|_2 \le \|u^i\|_2$ for all $i = 1, \ldots, n$. Thus, we have clarified that $\widetilde{v} \in I(\widetilde{W})$ and hence

$$\sum_{i=1}^n \widetilde{v}_i \|\widetilde{w}^i\|_2 = \|\widetilde{W}\|_{K,21}. \tag{5.7}$$

From the definition of $\widetilde{W}$ and (5.7), we have

$$\begin{aligned} \Phi(\widetilde{W}) &= \frac{1}{2} \sum_{i=1}^n \|\widetilde{w}^i - u^i\|_2^2 + \alpha(\|\widetilde{W}\|_{2,1} - \|\widetilde{W}\|_{K,21}) \\ &= \frac{1}{2} \sum_{\widetilde{v}_i = 0} \alpha^2 + \alpha \sum_{i=1}^n (1 - \widetilde{v}_i) \|\widetilde{w}^i\|_2 \\ &= \frac{1}{2}(n-K)\alpha^2 + \alpha \sum_{\widetilde{v}_i = 0} (\|u^i\|_2 - \alpha). \end{aligned}$$

Observe that $v \in I(W)$. Using (5.4) and (5.5), we have

$$\begin{aligned} \Phi(W) &= \frac{1}{2} \sum_{i=1}^n \|w^i - u^i\|_2^2 + \alpha(\|W\|_{2,1} - \|W\|_{K,21}) \\ &= \frac{1}{2} \sum_{v_i = 0} \alpha^2 + \alpha \left(\sum_{v_i = 0} \|w^i\|_2 \right) \\ &= \frac{1}{2}(n-K)\alpha^2 + \alpha \left(\sum_{v_i = 0} (\|u^i\|_2 - \alpha) \right). \end{aligned}$$

Since $\widetilde{v} \in I(U)$, we have $\sum_{i=1}^{n} \widetilde{v}_i \|u^i\|_2 \geq \sum_{i=1}^{n} v_i \|u^i\|_2$. It follows that

$$\begin{aligned}\sum_{\widetilde{v}_i=0} \|u^i\|_2 &= \sum_{i=1}^{n} (1-\widetilde{v}_i)\|u^i\|_2 \\ &\leq \sum_{i=1}^{n} (1-v_i)\|u^i\|_2 \\ &= \sum_{v_i=0} \|u^i\|_2.\end{aligned}$$

Thus, we can conclude that $\Phi(\widetilde{W}) \leq \Phi(W)$ for all W satisfying (5.4). The proof is now complete. □

From Theorem 3.1, to solve feature selection problem (2.3) we can solve its penalized problem (3.2) with a fixed value of penalty parameter ρ satisfying condition $\rho > m\|X\|_\infty - \gamma$. However, it is not ready to apply the proximal gradient method because the function $\mathscr{G}$ in (5.2) is nonsmooth. Employing Theorem 4.1 and Corollary 4.1, we can approximate this function by

$$\begin{aligned}\mathscr{G}_\mu(W) = \frac{1}{2\mu}\sum_{i=1}^{m} \|W^\top x_i - y_i\|_2^2 - \frac{\mu}{2}\sum_{i=1}^{m}\left[d\left(\frac{W^\top x_i - y_i}{\mu}; I\!B\right)\right]^2 \\ + \frac{\gamma}{2\mu}\sum_{i=1}^{n}\|w^i\|_2^2 - \frac{\gamma\mu}{2}\sum_{i=1}^{n}\left[d\left(\frac{w^i}{\mu}; I\!B\right)\right]^2.\end{aligned}$$

Recall that the Euclidean projection from a vector $u \in \mathbb{R}^c$ onto the closed unit ball $I\!B$ of $\mathbb{R}^c$ can be computed easily by

$$P_{I\!B}(u) = \begin{cases} \{u\}, & \text{if } \|u\|_2 \leq 1, \\ \left\{\frac{u}{\|u\|_2}\right\}, & \text{if } \|u\|_2 > 1. \end{cases}$$

To proceed further, let us denote $\mathscr{B}_m = I\!B \times I\!B \times \ldots \times I\!B$ as a subset of $\mathbb{R}^{m\times c}$. For an $m \times c$ matrix U, the projection from U to $\mathscr{B}_m$ is defined by

$$\Pi(U, \mathscr{B}_m) := \left[P_{I\!B}(u^1), \ldots, P_{I\!B}(u^m)\right]^\top.$$

It follows that

$$\begin{aligned}[d(U;\mathscr{B}_m)]^2 &= \|U - \Pi(U;\mathscr{B}_m)\|_F^2 \\ &= \sum_{i=1}^{m}\|u^i - P_{I\!B}(u^i)\|_2^2 = \sum_{i=1}^{m}[d(u^i; I\!B)]^2.\end{aligned}$$

By rewriting $\mathscr{G}_\mu$ in this notation and using the differentiability of squared distance functions, the result below is a direct consequence of Theorem 4.1.

Proposition 5.3. *The function $\mathscr{F}$ in (3.2) is approximated by the function*

$$\mathscr{F}_\mu(W) := \mathscr{G}_\mu(W) + \rho\, \mathscr{H}_K(W), \tag{5.8}$$

where

$$\begin{aligned}\mathscr{G}_\mu(W) :=& \frac{1}{2\mu}\|X^\top W - Y\|_F^2 - \frac{\mu}{2}\left[d\left(\frac{X^\top W - Y}{\mu}; \mathscr{B}_m\right)\right]^2 \\ &+ \frac{\gamma}{2\mu}\|W\|_F^2 - \frac{\gamma\mu}{2}\left[d\left(\frac{W}{\mu}; \mathscr{B}_n\right)\right]^2, \\ \mathscr{H}_K(W) :=& \|W\|_{2,1} - \|W\|_{K,21}.\end{aligned}$$

We have the following estimate

$$\mathscr{F}_\mu(W) \le \mathscr{F}(W) \le \mathscr{F}_\mu(W) + \frac{\mu}{2}(m+\gamma n). \tag{5.9}$$

The function $\mathscr{G}_\mu(W)$ is differentiable with gradient

$$\nabla \mathscr{G}_\mu(W) = X\Pi\left(\frac{X^\top W - Y}{\mu}; \mathscr{B}_m\right) + \gamma\Pi\left(\frac{W}{\mu}; \mathscr{B}_n\right).$$

The gradient $\nabla\mathscr{G}_\mu$ is Lipschitz continuous with constant

$$L_\mu = \frac{\|X\|_F^2 + n\gamma}{\mu}.$$

Thus, to find an approximate solution for (3.2), we minimize $\mathscr{F}_\mu$ in (5.8) with a sufficiently small value of smoothing parameter μ. The pseudo-code of **nmAPG** is outlined below.

Algorithm 1. nmAPG with fixed μ.

INPUT: training matrix $X \in \mathbb{R}^{n\times m}$, label matrix $Y \in \mathbb{R}^{m\times c}$,
integer $K > 0, \gamma > 0$, $\rho > 0$, $\mu > 0$, initial guess W^0.
Set $Z^1 = W^1 = W^0$, $t_0 = 0, t_1 = 1$, $\eta \in [0,1)$, $c_1 = \mathscr{F}_\mu(W^1)$,
$q_1 = 1$, $\tau > L_\mu = \frac{\|X\|_F^2 + n\gamma}{\mu}$ and $\delta \in (0, \tau - L_\mu)$.
for $k = 1, \ldots, T$ **do**
$U^k = W^k + \frac{t_{k-1}}{t_k}(Z^k - W^k) + \frac{t_{k-1}-1}{t_k}(W^k - W^{k-1})$
$Z^{k+1} \in \mathbf{prox}_{\frac{\rho}{\tau}\mathscr{H}_K}\left(U^k - \frac{1}{\tau}\nabla\mathscr{G}_\mu(U^k)\right)$
if $\mathscr{F}_\mu(Z^{k+1}) \le c_k - \frac{\delta}{2}\|Z^{k+1} - U^k\|_F^2$, **then**
$W^{k+1} = Z^{k+1}$
else
$V^{k+1} \in \mathbf{prox}_{\frac{\rho}{\tau}\mathscr{H}_K}\left(W^k - \frac{1}{\tau}\nabla\mathscr{G}_\mu(W^k)\right)$
$W^{k+1} = \begin{cases} Z^{k+1}, & \mathscr{F}_\mu(Z^{k+1}) \le \mathscr{F}_\mu(V^{k+1}), \\ V^{k+1}, & \text{otherwise.} \end{cases}$
end if
$t_{k+1} = \dfrac{\sqrt{4t_k^2 + 1} + 1}{2}$
$q_{k+1} = \eta q_k + 1$
$c_{k+1} = \dfrac{\eta q_k c_k + \mathscr{F}_\mu(W^{k+1})}{q_{k+1}}$
end for
OUTPUT: W^{T+1}.

Remark 5.1. It often happens that using a small value of the smoothing parameter μ is better because it reduces the approximation gap between $\mathscr{F}$ and $\mathscr{F}_\mu$. However, a too small value of μ implies a very large value of the Lipschitz constant L_μ. This makes the step-size $\frac{1}{\tau}$ in **nmAPG** too small which leads to a slow convergence rate. Thus, the time cost of the algorithm is expensive if we fix μ ahead of time. In practice, we often decrease μ gradually until a preferred threshold μ_* is attained. From the estimate (5.9), we choose

$$\mu_* = \frac{2\varepsilon}{m+\gamma n},$$

which ensures that the gap $\mathscr{F} - \mathscr{F}_\mu$ is less than ε. The final optimization scheme is outlined as follows.

Algorithm 2. nmAPG with gradually decreasing μ.

INPUT: training matrix $X \in \mathbb{R}^{n\times m}$, label matrix $Y \in \mathbb{R}^{m\times c}$,
$\gamma > 0$, $\rho > 0$, integer $K > 0$,
$W^0 \in \mathbb{R}^{n\times c}$, $\sigma \in (0,1)$, $\mu_0 > 0$, $\varepsilon > 0$

Set $k \longleftarrow 0$.

repeat the following

Compute $W^{k+1} \leftarrow$ **Algorithm1**$(X, Y, \gamma, \rho, K, \mu_k, W^k)$

Update $\mu_{k+1} \longleftarrow \sigma\mu_k$

Set $k \longleftarrow k+1$

until $\mu < \mu_* = \frac{2\varepsilon}{m+\gamma n}$

OUTPUT: W

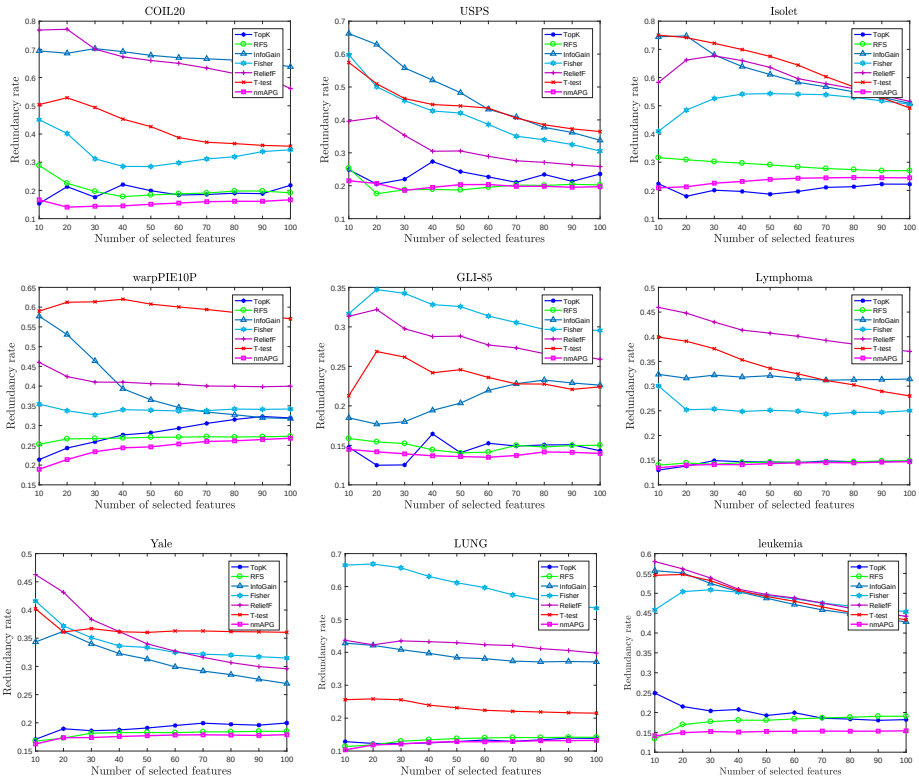

FIGURE 1. Redundancy rate comparison of 7 feature selection methods on 9 data sets.

6. NUMERICAL EXPERIMENTS

In order to evaluate the performance of our method (**nmAPG**, for brevity), we will compare it with the following popular multiclass supervised feature selection methods: **RFS** [22], **TopK** [24], **ReliefF** [11, 12], **Fisher** [16, 17], **InfoGain** [36, 37] and **T-test** [14, 15]. The efficiency of each method is estimated by the average of the redundancy rate and the classification accuracy.

For **TopK**, we use the following parameters: $\mu = 0.1, \rho = 1.02, \text{maxiter} = 1000$, and for **RFS** we vary its regularization in the set $\gamma = \{0.001, 0.01, 0.1, 0.2, 0.8, 1\}$ and then chose the best result to report. The others four methods are parameter free. We set the parameters for **nmAPG** as

$$\tau = 2L_\mu, \delta = 0.1(\tau - L_\mu), \eta = 0.9,$$

and set the starting guess W^0 to be the zero matrix.

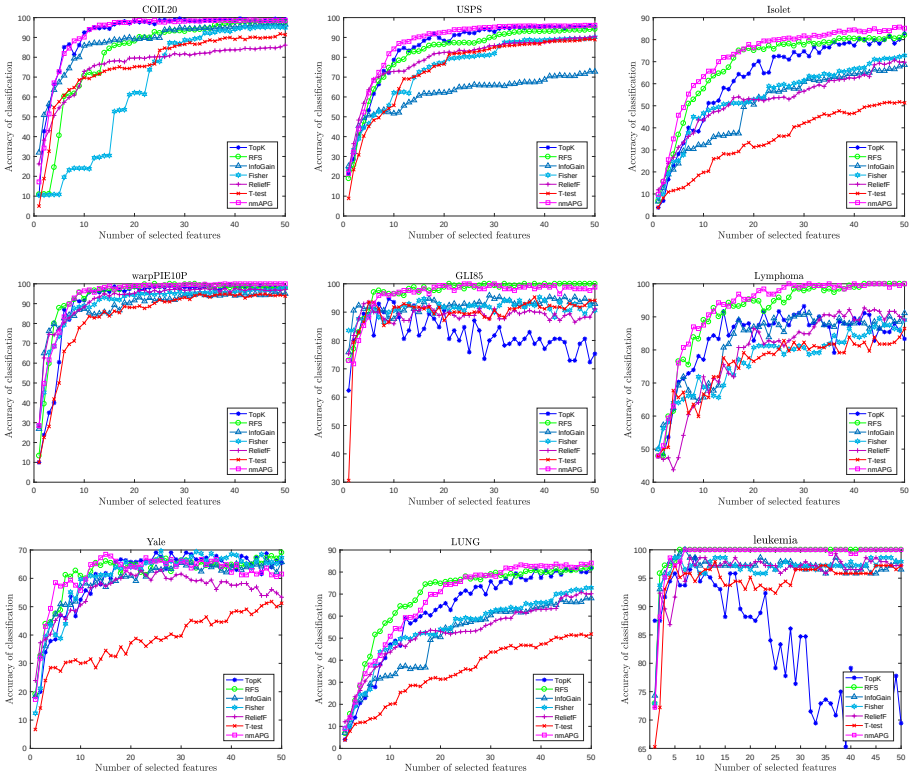

FIGURE 2. 1-NN classification accuracy comparison of 7 feature selection methods on 9 data sets.

6.1. **Datasets Description.** The following 9 benchmark data sets for feature selection are used in our experiments. All of them are available at
http://featureselection.asu.edu/datasets.php.

TABLE 1. Summary of 9 data sets.

Data	**# Instances**	**# Features**	**# Classes**	**Types**
COIL20	1440	1024	20	Image
WarpPIE10P	210	2420	10	Image
USPS	9298	256	10	Image
Yale	165	1024	15	Image
LUNG	203	3312	5	Biology
GLI-85	85	22283	2	Biology
Lymphoma	96	4026	9	Biology
Isolet	1560	617	26	Voice
leukemia	72	7070	2	Biology

6.2. **Redundancy Rate Comparison.** The redundancy rate is a popular measurement to evaluate the quality of selected features. Let F be the set of features being selected by some method, the redundancy rate is defined by

$$\mathrm{RED}(F) := \frac{1}{|F|(|F|-1)} \sum_{f_i, f_j \in F,\, i>j} \mathrm{corr}_{ij},$$

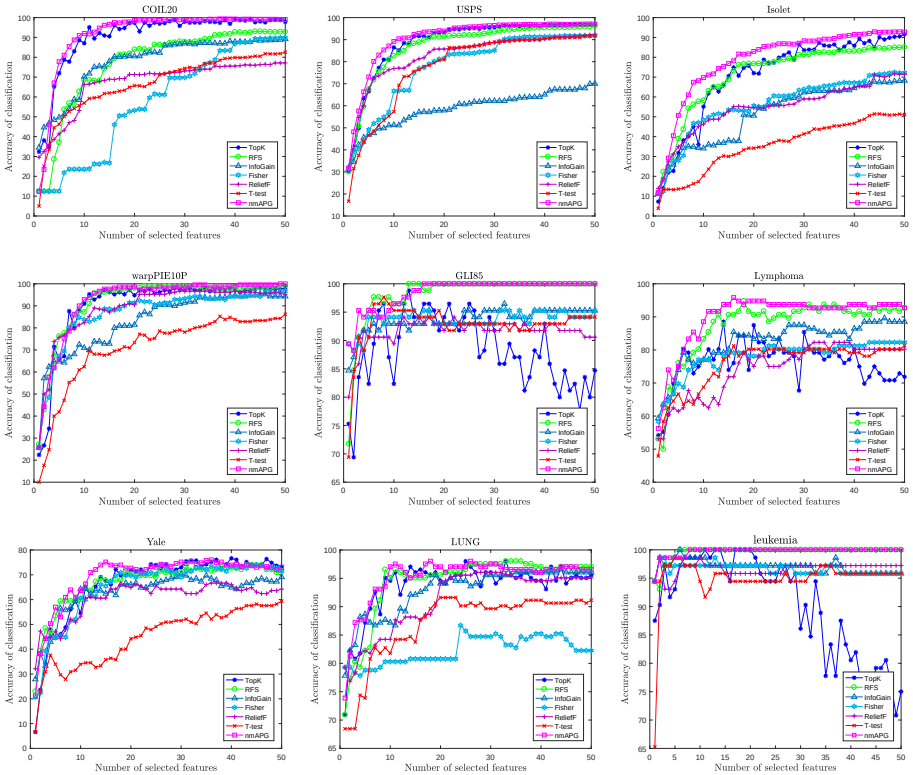

FIGURE 3. SVM classification accuracy comparison of 7 feature selection methods on 9 data sets.

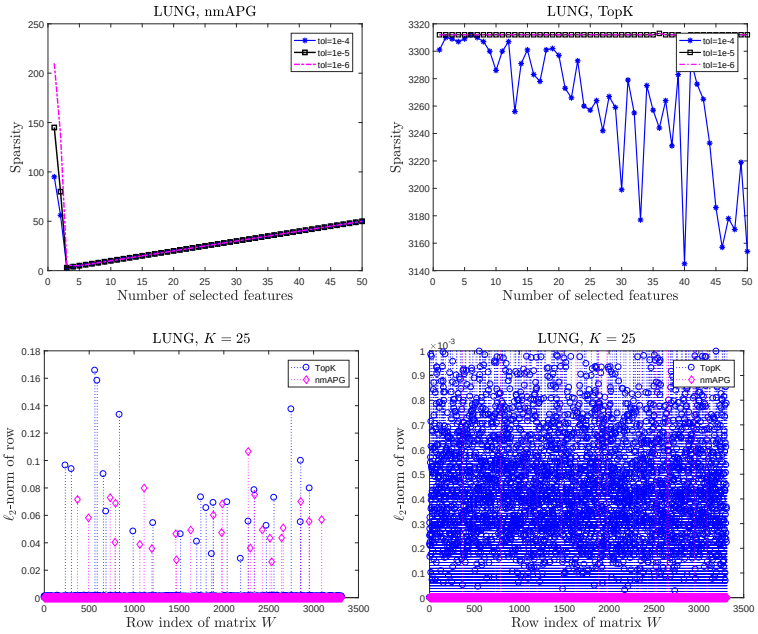

FIGURE 4. Sparsity level comparison between **nmAPG** and **TopK** on LUNG. The bottom-right is a zoom of the bottom-left in the interval $[0, 0.001]$.

where $|F|$ is the cardinality of F and corr_{ij} is the correlation between two features f_i and f_j. Here the correlation is computed on the reduced data represented by features in F. A large value of $\mathrm{RED}(F)$ indicates that many selected features are correlated and there exist some redundancy in F. We apply Algorithm 1 for all runs in this section with the same setting of parameters below:

$$\gamma = 1, \mu = 0.01, \rho = 0.01.$$

The algorithm will terminate whenever the maximum number of iterations is 1000 or the following conditions is satisfied:

$$\frac{\|W^k - W^{k-1}\|_F}{\max(\|W^k\|_F, 1)} \leq 10^{-7}. \tag{6.1}$$

Each data set is randomly divided into training set and testing set with the ratio 70% train - 30% test. For each case, we perform 10 independent runs and report the average RED in Fig. 1. The result show that the set of features chosen by our method has lower redundancy rate in almost all cases than the others and therefore our method is very efficient in eliminating irrelevant features. Note also that the RED is computed directly on the reduced data without employing any learning algorithm. The comparison result also shows that our method does not rely on any specific classifier. This fact will be clarified in what follows.

6.3. **Classification Accuracy Comparison.** In this section, 1-NN and SVM with 5-fold cross-validation are employed to evaluate the quality of selected features. We use the support vector machine (SVM) with linear kernel and set its parameter $C = 1$. The number of selected features K on each data set varies from 1 to 50 with the incremental step 1. The better FS method is expected to have higher classification accuracies. We apply Algorithm 2 for all runs in this section with the setting of parameters as follows

$$\gamma = 0.1, \mu_0 = 1, \sigma = 0.1, \varepsilon = 0.1,$$
$$\rho \in \{0.001\rho_{\max}, 0.01\rho_{\max}\},$$

where $\rho_{\max}$ is defined in Theorem 3.1. We switch to the next smaller value of μ after every 200 iterations or when condition (6.1) is reached. We also use a "warm start" technique by using the solution of the previous problem as an initial guess for the next one.

The plots of average classification accuracy versus the number of selected features for 10 independent runs are reported in Fig. 2 for 1-NN and in Fig. 3 for SVM.

Our numerical examples show that method outperforms the compared FS methods on both classifiers. The accuracy curves of **RFS** and **nmAPG** are much more stable than that of **TopK**. The result also shows that, although both **TopK** and **nmAPG** use the $\ell_{2,0}$-norm constraint, our method outperforms **TopK** in the cases where K is quite large, e.g., $K > 20$.

6.4. **Sparsity Level Investigation.** Both **TopK** and our model employ the $\ell_{2,0}$-norm constraint to promote sparsity. We now conduct experiments to examine the effect of the sparsity constraint on the selected features. The row-sparsity of a solution W obtained by some FS method is determined by the number of rows whose ℓ_2-norms are greater than a tolerance tol. With the same setting as section 6.3, we plot the sparsity investigation of both methods on LUNG data set in Fig. 4. The sparsity curve of **TopK** is not linearly increasing with respect to K as expected. The sparsity level of **nmAPG** is nearly the same as the number of features to be selected. This shows that the $\ell_{2,0}$-norm constraint in our model is very effective from the sparsity perspective.

6.5. **Convergence of nmAPG.** We now chose two labeled data sets COIL20 and Isolet to illustrate the convergence of Algorithm 1 and Algorithm 2. Recall that for minimizing the function $\mathscr{F}$ in (3.2), we minimize its approximation $\mathscr{F}_\mu$ in (5.8) by Algorithm 1 with a fixed μ or by Algorithm 2 with decreasing μ. Fig. 5 plots values of $\mathscr{F}$ versus iterations of both algorithms for solving (3.2) in which

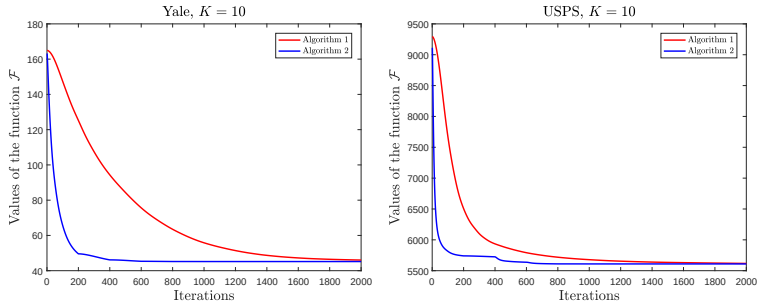

FIGURE 5. The convergence of Algorithm 1 and Algorithm 2 on Yale and USPS.

$\gamma = 0.1, \rho = 0.01$ and $K = 10$. We can see that both algorithms decrease the objective function very fast, and Algorithm 2 converges much faster than Algorithm 1.

7. CONCLUSIONS

In this paper, we introduced a new supervised FS model which has the $\ell_{2,1}$-norm on both loss and regularization terms with an additional $\ell_{2,0}$-norm constraint. Using the exact penalty method, we reformulated the problem as a continuous one by adding to the objective a penalized term of the form $\ell_{2,1} - \ell_{K,21}$. Based on an explicit formula for the proximal mapping of the $\ell_{2,1} - \ell_{K,21}-$norm as well as Nesterov's smoothing techniques, the problem can be effectively solved by the accelerated proximal gradient methods. Our numerical examples show that the proposed method outperforms many state-of-art FS methods on several benchmark data sets. The approach to deal with $\ell_{2,0}$ in this paper can be applied to many other problems in constrained sparse optimization. This is a promising research topic we would like to pursue in the future.

Acknowledgements

Research of the first author was supported by the Vietnam National Foundation for Science and Technology Development under grant 101.01-2017.325. Research of the third author was supported by the National Natural Science Foundation of China under Grant 11401152. This work was also supported by the Strong Research Group Program of Hue University.

REFERENCES

[1] A. Destrero, C. De Mol, F. Odone, A. Verri, A regularized framework for feature selection in face detection and authentication, Int. J. Comput. Vis. 83 (2009), 164-177.

[2] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, Pattern Recognit. 45 (2012), 2884-2893.

[3] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2017), 2507-2517.

[4] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015), 252-264.

[5] C. Freeman, D. Kulic, O. Basir, An evaluation of classifier-specific filter measure performance for feature selection, Pattern Recognit. 48 (2015), 1812-1826.

[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Machine Learn. Res. 3 (2003), 1157-1182.

[7] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, Feature selection based on structured sparsity: A comprehensive study, IEEE Trans. Neural Netw. Learn. Syst. 28 (2017), 1490-1507.

[8] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer, Berlin, 2006.
[9] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, 50 (6), p.94, ACM Computing Surveys (CSUR), 2017.
[10] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997), 273-324.
[11] K. Kira, L.A. Rendell, A Practical Approach to Feature Selection. in Proceedings of the 9th International Workshop on Machine Learning, ML92, pp. 249–256, Morgan Kaufmann Publishers Inc., San Francisco, 1992.
[12] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, In European Conference on Machine Learning, pp. 171-182, 1994.
[13] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Mach. Learn. 53 (2003), 23-69.
[14] J.C. Davis, R.J. Sampson, Statistics and Data Analysis in Geology, 2nd ed. John Wiley & Sons, New York, 1986.
[15] R. Peck, J.L. Devore, Statistics: The exploration & analysis of data, Cengage Learning, 2011.
[16] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, (2nd Edition), Wiley-Interscience, 2000.
[17] Q. Gu, Z. Li, J. Han, Generalized Fisher score for feature selection, in Proc. 27th Conf. Uncertainty Artif. Intell., pp. 266-273, 2011.
[18] A. Argyriou, A. Evgeniou, M. Pontil, Multi-task feature learning, Proc. 19th Ann. Conf. Neural Information Processing Systems, pp. 41-48, 2007.
[19] L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification, In ICML, 2007.
[20] E. Amaldi, V. Kann, On the approximability of minimizing non zero variables or unsatisfied relations in linear systems, Theor. Comput. Sci. 209 (1998), 237-260.
[21] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B. Stat. Methodol. 58 (1994), 267-288.
[22] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, in Proc. Adv. Neural Inf. Process. Syst., pp. 1813–1821, 2010.
[23] S. Xiang, F. Nie, G. Meng, C. Pan, C. Zhang, Discriminative least squares regression for multiclass classification and feature selection, IEEE Trans. Neural Netw. Learn. Syst. 23 (2012), 1738-1754.
[24] X. Cai, F. Nie, H. Huang, Exact top-k feature selection via $\ell_{2,0}$-norm constraint, 23rd International Joint Conference on Artificial Intelligence (IJCAI), pp. 1240-1246, 2013.
[25] B.S. Mordukhovich, N.M. Nam, An Easy Path to Convex Analysis and Applications, Morgan & Claypool, 2014.
[26] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. Ser. A 103 (2005), 127-152.
[27] N.M. Nam, N.T. An, N. R.B. Rector, J. Sun, Nonsmooth algorithms and Nesterov's smoothing technique for generalized Fermat-Torricelli problems, SIAM J. Optim. 24 (2014)), 1815-1839.
[28] N. Parikh, S. Boyd, Proximal Algorithms, Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123-231, 2013.
[29] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (2014), 459-494.
[30] H. Li, Z. Lin, Accelerated proximal gradient methods for nonconvex programming. in Advances in Neural Information Processing Systems, vol. 28, pp. 379–387, Montreal, 2015.
[31] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009), 183-202.
[32] Y. Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140 (2013), 125-161.
[33] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE Trans. Image Process 18 (2009), 2419–2434.
[34] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality, Math. Oper. Res. 35 (2010), 438-457.

[35] P.D. Tao, L.T.H. An, A d.c. optimization algorithm for solving the trust-region subproblem, SIAM J. Optim. 8 (1998), 476-505.

[36] S. Lei, A feature selection method based on information gain and genetic algorithm, 2012 International Conference on Computer Science and Electronics Engineering, pp. 355-358. Hangzhou, 2012.

[37] L.E. Raileanu, K. Stoffel, Theoretical comparison between the gini index and information gain criteria, Univ. Neuchatel, Neuchatel, 2000.