Discriminative Feature Selection via A Structured Sparse Subspace Learning Module

Zheng Wang[1] , **Feiping Nie**[1*] , **Lai Tian**[1] , **Rong Wang**[1,2] and **Xuelong Li**[1]
[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, 710072, P. R. China
[2]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, 710072, P. R. China
{zhengwangml, feipingnie, tianlai.cs}@gmail.com, wangrong07@tsinghua.org.cn, li@nwpu.edu.cn

Abstract

In this paper, we first propose a novel Structured Sparse Subspace Learning (S^3L) module to address the long-standing subspace sparsity issue. Elicited by proposed module, we design a new discriminative feature selection method, named Subspace Sparsity Discriminant Feature Selection (S^2DFS) which enables the following new functionalities: 1) Proposed S^2DFS method directly joints trace ratio objective and structured sparse subspace constraint via $\ell_{2,0}$-norm to learn a row-sparsity subspace, which improves the discriminability of model and overcomes the parameter-tuning trouble with comparison to the methods used $\ell_{2,1}$-norm regularization; 2) An alternative iterative optimization algorithm based on the proposed S^3L module is presented to explicitly solve the proposed problem with a closed-form solution and strict convergence proof. To our best knowledge, such objective function and solver are first proposed in this paper, which provides a new though for the development of feature selection methods. Extensive experiments conducted on several high-dimensional datasets demonstrate the discriminability of selected features via S^2DFS with comparison to several related SOTA feature selection methods. *Source matlab code: https://github.com/StevenWangNPU/L20-FS.*

1 Introduction

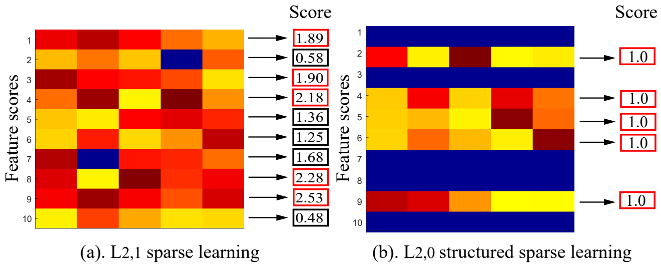
Data in many areas are represented by high-dimensional features, and a natural question emerges out: *what are the most discriminative features in the high-dimensional data?* Feature selection plays a crucial role in machine learning which endeavors to select a subset of features from the high-dimensional data for improving data compactness and reducing noisy features so that the over-fitting, high computational consumption and low performance issues can be alleviated in many real-world applications.

Feature selection approaches can be roughly divided into three categories, i.e., filter methods [Gu *et al.*, 2012], wrapper methods [Maldonado and Weber, 2009] and embedded methods [Xiang *et al.*, 2012; Ming and Ding, 2019; Tian *et al.*, 2019]. Wherein, filter methods focus on evaluating the correlation of features with respect to class label of data, which results in that the correlated features are redundant. Wrapper methods measure the importance of features according to classification performance, so as to the computational cost of wapper model is very high and the representability of learned features may be poor in other tasks. In contrast, embedded methods gain more attentions since it incorporates feature selection and classification model into a unified optimization framework by learning sparse structural projections. Besides, some feature selection approaches based on neural network [Zhang *et al.*, 2020] and auto-encoder model [Han *et al.*, 2018; Abid *et al.*, 2019] pay overmuch attentions on minimizing the reconstruct errors, which possible not benefit for classification task and tuning parameters may affect its efficiency of practical applications.

Most of embedded methods commonly employ $\ell_{2,1}$-norm regularization to improve the row-sparsity in learned subspace. Concretely, Nie et al. propose robust feature selection model (RFS) [Nie *et al.*, 2010] which simultaneously incorporates $\ell_{2,1}$-norm into loss function and regularization term to overcome outliers issues and make feature selection come true. Similar, [He *et al.*, 2012] proposes to use Correntropy Robust loss function in Feature Selection (CRFS) to enhance model's robustness. Additionally, [Xiang *et al.*, 2012] develops a feature selection effort joints discriminative least square regression (DLSRFS) model and sparse $\ell_{2,1}$-norm regularization into an unified framework. Recently, RJFWL [Yan *et al.*, 2016] aims to learn the ranking of all features by imposing $\ell_{2,1}$-norm constraint and non-negative constraint on learned feature weights matrix. Regrettably, all aforementioned approaches are always unable to escape the dilemma of tuning the trade-off parameter between loss term and sparse regularization term. Moreover, the compactness of sparse row is sensitive to the trade-off parameter, namely, the uniqueness of optimum is difficult to guarantee.

An illustrative example is depicted in Figure 1, from which we can figure out that $\ell_{2,1}$-norm calculates the score of features integrally, then ranks them resorting to the learned scores. By doing so, the feature ranking results may be changed dramatically when the feature score vary subtly. In contrast, $\ell_{2,0}$-norm obtains the score matrix only depend

[*] Corresponding Author



(a). L2,1 sparse learning

(b). L2,0 structured sparse learning

Figure 1: Visualization of learned sparse feature score matric, (a). $\ell_{2,1}$ sparse matrix and (b). $\ell_{2,0}$ structured sparse matrix, respectively. Navy blue background denotes the scores that close to zero.

upon the k features, and no further ranking operation is required. Based on these, Cai et al. [Cai *et al.*, 2013] questioning that whether the method based on convex problem is always better than that based on non-convex problem. They exploit top-k features selection method (RPMFS) based on least square regression model with $\ell_{2,0}$-norm equality constraint. Pang et al. [Pang *et al.*, 2018] develop an Efficient Sparse Feature Selection (ESFS) via $\ell_{2,0}$-norm constraint based on least square regression model as well. In [Du *et al.*, 2018], Unsupervised Group Feature Selection (UGFS) algorithm selects a group of features initially then update the selection until a better group appears by using $\ell_{2,0}$-norm constraint. Nevertheless, both of above methods optimize the $\ell_{2,0}$-norm constraint problem by using Augmented Lagrangian Multiplier (ALM) and gradient descent optimization algorithms, which may result in that the solution is sensitive to the initialization and easy to stuck in local optimum. Ultimately, the subsequent performance of classification is fluctuant.

Facing these obstacles that hinder the development of feature selection, we entail solving $\ell_{2,0}$-norm constraint problem along with rigorous theoretical guarantee firstly in this paper. Then, we propose a new discriminative feature selection model via orthogonal $\ell_{2,0}$-norm constraint. Finally, we analyze the performance of proposed method from two aspects, i.e., discriminability and efficiency. **Our contributions can be summarized as follows:**

- We propose a Structured Sparse Subspace Learning (S^{3}L) module to solve the subspace sparsity issue with theoretical guarantee, which is first presented in our paper according to our best knowledge.
- We elaborately design a discriminative feature selection model, named Subspace Sparsity Discriminant Feature Selection (S^{2}DFS) that integrates trace ratio formulated objective and structured sparse subspace constraint for acquiring more informative feature subset.
- Optimizing the proposed S^{2}DFS model is a NP-hard problem. We provide an alternative iterative optimization algorithm to solve trace ratio problem, in which the proposed S^{3}L module is employed to acquire a closed-formed solution rather than an approximate one, so as to ensure the stability of selection results.
- Experimental results show the effectiveness of proposed optimization algorithm in two perspectives, i.e., performance: our method outperforms other related SOTA feature selection methods in terms of classification on several real-world datasets; convergent speed: our algorithm reaches convergence within few iterations.

2 Related Works

In this section, we briefly review constrained Linear Discriminative Analysis (LDA) in trace ratio formulation, then the subspace sparsity issue will be introduced.

2.1 Constrained Trace Ratio LDA

Given the training data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ and the label vector is $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, where $y_i \in \{1, 2, ..., c\}$, and c denotes the total class number. LDA aims to find a transformation matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$ to map original high-dimensional sample $\mathbf{x} \in \mathbb{R}^{d \times 1}$ into low-dimensional subspace with m dimensions where $m << d$. For improving the discriminative power of model, the samples within the same class are pulled together as well as those points between different classes are pushed far away in learned subspace ideally. There are many kinds of objective function of LDA [Bishop, 2006], one of the most discriminative objectives is the following constrained trace ratio one

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{Tr(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{Tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W})}, \tag{1}$$

where $\mathbf{S}_b = \sum_{i=1}^{c} n_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)^T$ and $\mathbf{S}_w = \sum_{i=1}^{c}\sum_{\mathbf{x}_j \in \boldsymbol{\pi}_i}\left(\mathbf{x}_j - \boldsymbol{\mu}_i\right)\left(\mathbf{x}_j - \boldsymbol{\mu}_i\right)^T$ denote between-class and within-class scatter matrix respectively. n_i is the number of samples belong to i-th class, $\boldsymbol{\mu}_i$ denotes the mean of samples in i-th class, $\boldsymbol{\mu}$ is the mean of total samples and $\boldsymbol{\pi}_i$ denotes the set of i-th class points. Orthogonal constraint is used to avoid trivial solution [Nie *et al.*, 2019]. Note that, optimizing such model in Eq.(1) is a non-convex optimization problem that does not have a closed-form solution, and several attempts [Wang *et al.*, 2007; Nie *et al.*, 2019] have tried to solve it. In what follows, we will provide an alternative iterative optimization algorithm to get the solution.

2.2 Subspace Sparsity Problem

Subspace sparsity issue is originally derived from the sparse principal subspace estimation [Vu *et al.*, 2013] which can be generally formulated as

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I_{m\times m}}, \|\mathbf{W}\|_{2,0}=k} Tr(\mathbf{W}^T\mathbf{A}\mathbf{W}), \tag{2}$$

where $m \leq k \leq d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a positive semi-definite matrix. $\|\mathbf{W}\|_{2,0} = \sum_{i=1}^{d}\left\|\sum_{j=1}^{m} w_{ij}^2\right\|_0$ denotes the number of non-zero rows in matrix $\mathbf{W}$, and the constraint $\|\mathbf{W}\|_{2,0} = k$ forces the learned subspace is row sparse, in which the index of non-zero rows equals to the index of selected features. Since both the orthogonal and row sparsity constraints are the non-convex constraint, optimizing problem (2) is challenging. In the following Section 3.1, we will propose the S^{3}L module to solve it with theoretical guarantee.

3 Proposed Method

3.1 Structured Sparse Subspace Learning Module

In this section, we propose a novel S^{3}L module to solve problem (2) that has been tentatively solved in [Wang *et al.*, 2014; Yang and Xu, 2015], but none of them provides deterministic guarantee on global convergence. We first develop a straightforward strategy for solving the case that $\text{rank}\left(\mathbf{A}\right) \leq m$. Then, an iterative optimization **Algorithm 1** is derived to solve the general case that $\text{rank}\left(\mathbf{A}\right) > m$ in which a low-rank proxy covariance $\mathbf{P}$ is designed to approximate $\mathbf{A}$. For readability, we define two definitions that will be used in following deducing:

Definition 3.1 *Given an indices $\mathcal{I}$ whose elements span from 1 to d. Then, the row extraction matrix $\mathbf{Q}_d^k \in \mathbb{R}^{d\times k}$ is defined as:*

$$q_{ij} = \begin{cases} 1, & \textit{if} \quad i = \mathcal{I}(j) \\ 0, & \textit{otherwise.} \end{cases}$$

Define $\mathcal{Q}_{d,k}(\mathcal{I}) = \mathbf{Q}_d^k$ as an operator that inputs indices $\mathcal{I}$ and outputs row extraction matrix $\mathbf{Q}_d^k \in \mathbb{R}^{d\times k}$.

Case 1: $\text{rank}\left(\mathbf{A}\right) \leq m$

We first consider the simplest problem, i.e., $k = m$, thus, the problem (2) reduces to

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I_{m\times m}}, \|\mathbf{W}\|_{2,0}=m} Tr\left(\mathbf{W}^T\mathbf{A}\mathbf{W}\right). \tag{3}$$

Due to the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I_{m\times m}}$ and $\|\mathbf{W}\|_{2,0} = m$, $\mathbf{W}$ can be rewritten as $\mathbf{W} = \mathbf{Q}_d^m\mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^{m\times m}$ and satisfies $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{m\times m}$. Additionally, $\mathbf{Q}_d^m = \mathcal{Q}_{d,m}(\mathcal{I}) \in \mathbb{R}^{d\times m}$ is the row extraction matrix that has been defined. Then, problem (3) has been reduced to solve

$$\max_{\mathbf{V}\in\mathbb{R}^{m\times m}, \mathbf{V}^T\mathbf{V}=\mathbf{I}_{m\times m}} Tr\left(\mathbf{V}^T\widetilde{\mathbf{A}}\mathbf{V}\right), \tag{4}$$

where $\widetilde{\mathbf{A}} = \mathbf{Q}_d^{mT}\mathbf{A}\mathbf{Q}_d^m$. So far, the row extraction matrix $\mathbf{Q}_d^m$ is still unknown. As a consequence, considering $\mathbf{V} \in \mathbb{R}^{m\times m}$ is a square matrix and $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, we can infer the fact $Tr(\mathbf{V}^T\widetilde{\mathbf{A}}\mathbf{V}) = Tr(\widetilde{\mathbf{A}}\mathbf{V}\mathbf{V}^T)$. Thus, problem (4) can be reduced to

$$\max Tr\left(\widetilde{\mathbf{A}}\right), \tag{5}$$

which has a globally optimal solution that m largest diagonal elements of $\mathbf{A}$, and $\mathbf{Q}_d^m$ is generated by operator $\mathcal{Q}_{d,m}(\mathcal{I})$ with input $\mathcal{I}$ that equals to the indices of m largest diagonal elements of $\mathbf{A}$.

Next, we consider the case that $k \neq m$, and problem in **Case 1** becomes to

$$\begin{aligned} &\max_{\mathbf{W}\in\mathbb{R}^{d\times m}} Tr\left(\mathbf{W}^T\mathbf{A}\mathbf{W}\right) \\ &s.t. \quad \mathbf{W}^T\mathbf{W} = \mathbf{I_{m\times m}}, \|\mathbf{W}\|_{2,0} = k, \text{rank}(\mathbf{A}) \leq m. \end{aligned} \tag{6}$$

We use similar technique to solve problem (6) as

$$\max Tr\left(\mathbf{Q}_d^{k^T}\mathbf{A}\mathbf{Q}_d^k\right), \tag{7}$$

which can be easily solved globally by selecting k largest diagonal elements of $\mathbf{A}$, and $\mathbf{Q}_d^k = \mathcal{Q}_{d,k}(\mathcal{I})$ where $\mathcal{I}$ denotes the indices of k largest diagonal elements of $\mathbf{A}$. Throughout above analysis, our S^{3}L module is able to efficiently (non-iteratively) obtain **globally optimal solution** of NP-hard problem (2) when $\text{rank}(\mathbf{A}) \leq m$.

Nevertheless, the **Case 1** is not usually occurred in practice, since PCA often used as a data-preprocessing technique to remove null space of data, which cause a general **Case 2**, i.e., $\text{rank}(\mathbf{A}) = d > m$. Next, we develop an iterative S^{3}L module to solve the problem (2) in the following **Case 2**.

Case 2: $\text{rank}\left(\mathbf{A}\right) > m$

We solve the problem (2) in **case 2** based on the Minimization-Majorization framework (MM) [Sun *et al.*, 2016] whose key insight is to maximize a lowerbound function of target problem. Therefore, how to design a suitable lowerbound function is the key to the success of MM. Fortunately, through a lot of attempts, we elaborately design an ideal lowerbound function of problem (2) as follow:

$$\begin{aligned} g\left(\mathbf{W}|\mathbf{W}_t\right) &= Tr(\mathbf{W}^T\mathbf{\Gamma}_t\mathbf{W}) \\ s.t. \quad &\mathbf{W} \in \mathcal{K}, \end{aligned} \tag{8}$$

where $\mathcal{K} = \{\mathbf{W} \in \mathbb{R}^{d\times m} | \mathbf{W}^T\mathbf{W} = \mathbf{I}, \|\mathbf{W}\|_{2,0} = k\}$ is the structured sparse subspace constraint, the surrogate variable $\mathbf{\Gamma}_t = \mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^\dagger\mathbf{W}_t^T\mathbf{A}$ and $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse. As a lowerbound function, a deserved property should be invariably held:

$$Tr(\mathbf{W}^T\mathbf{\Gamma}_t\mathbf{W}) \leq Tr(\mathbf{W}^T\mathbf{A}\mathbf{W}) \quad \forall \mathbf{W} \in \mathcal{K}, \tag{9}$$

which will be proved in Section 3.2.

According to MM framework, solving problem (2) in **Case 2** becomes to solve

$$\max_{\mathbf{W}\in\mathcal{K}} Tr\left(\mathbf{W}^T\mathbf{\Gamma}_t\mathbf{W}\right), \tag{10}$$

which is still a difficult problem that can not be optimized directly. Now, we discuss some properties of designed surrogate variable $\mathbf{\Gamma}_t$ through the following **Theorem 1**.

Theorem 1 *In the t-th iteration $(t \geq 1)$, the conditions rank $\left(\mathbf{\Gamma}_t\right) \leq m$ and $\mathbf{\Gamma}_t \succeq \mathbf{0}$ will always hold.*

Proof 1 *According to the property of matrix rank $rank(\mathbf{AB}) \leq \min\{rank(\mathbf{A}), rank(\mathbf{B})\}$, we have $rank(\mathbf{\Gamma}_t) \leq rank(\mathbf{W}_t) = m$, then the first condition has been proved. Subsequently, we suppose that $\mathbf{\Omega} = \mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^\dagger\mathbf{W}_t^T\mathbf{A}^{\frac{1}{2}}$, then according to the fact that $\mathbf{A} \succeq \mathbf{0}$, $\mathbf{A}$ is a symmetric matrix and $\mathbf{B}^\dagger = \mathbf{B}^\dagger\mathbf{B}\mathbf{B}^\dagger$, we can obtain*

$$\begin{aligned} \mathbf{\Omega}\mathbf{\Omega}^T &= \mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^\dagger\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^\dagger\mathbf{W}_t^T\mathbf{A} \\ &= \mathbf{\Gamma}_t \succeq \mathbf{0}, \end{aligned} \tag{11}$$

*which completes the proof of **Theorem 1**.*

According to **Theorem 1**, we wondrously discover an interesting phenomenon that the designed surrogate variable $\mathbf{\Gamma}_t$ satisfies all conditions of $\mathbf{A}$ in **case 1**, i.e, $\text{rank}(\mathbf{A}) \leq m$ and $\mathbf{\Gamma}_t \succeq \mathbf{0}$. Those facts enlighten a thought that problem (10) can be solved by using S^{3}L module iteratively, which is summarized in the following **Algorithm 1**.

Algorithm 1 Solve problem (2), when rank$(\mathbf{A}) > m$

```
Input: $\mathbf{A}$, $k$, $m$, $d$ and $\mathbf{W}_0$;
Initialization: $\mathbf{W}_0 = \text{rand}(d, m)$, $t = 0$;
Repeat:
  1. $\mathbf{\Gamma}_t \leftarrow \mathbf{A}\mathbf{W}_t \left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^{\dagger} \mathbf{W}_t^T\mathbf{A}$;
  2. $\mathcal{I}_t \leftarrow \textsf{sort}$ ( diag($\mathbf{\Gamma}_t$), 'descend', $k$ );
  3. $\mathbf{Q}_{td}^{k} \leftarrow \mathcal{Q}_{d,k}(\mathcal{I}_t)$;
  4. $\mathbf{V}_t \leftarrow \textsf{eigs}\left(\left(\mathbf{Q}_{td}^{k}\right)^T \mathbf{A}\mathbf{Q}_{td}^{k}, \text{'descend'}, m\right)$;
  5. $\mathbf{W}_{t+1} \leftarrow \mathbf{Q}_{td}^{k}\mathbf{V}_t$;
  6. $t = t + 1$;
Until convergence
Output: $\mathbf{W}^* \in \mathbb{R}^{d\times m}$.
```

3.2 Algorithm Analysis

We prove the convergence of the proposed **Algorithm 1** through the following **Theorem 2**.

Theorem 2 *Algorithm 1 increases the objective function value of Eq.(2) in each iteration, when rank*$(\mathbf{A}) > m$.

Proof 2 *Note the fact that* $\mathbf{B} = \mathbf{B}\mathbf{B}^{\dagger}\mathbf{B}$, *we can infer that*

$$\begin{aligned} Tr\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right) &= Tr\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^{\dagger}\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right) \\ &= Tr\left(\mathbf{W}_t^T\mathbf{\Gamma}_t\mathbf{W}_t\right), \end{aligned} \tag{12}$$

where rank $(\mathbf{\Gamma}_t) \leq m$. *Then, we can use the* S^{3}L *module to maximize the above Eq.(12) as*

$$\begin{aligned} & Tr\left(\mathbf{W}_t^T\mathbf{\Gamma}_t\mathbf{W}_t\right) \\ =& Tr\left(\widehat{\mathbf{W}}_{t+1}^T\mathbf{A}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^{\dagger}\mathbf{W}_t^T\mathbf{A}\widehat{\mathbf{W}}_{t+1}\right), \end{aligned} \tag{13}$$

where $\widehat{\mathbf{W}}_{t+1}^T$ *is generated according to Eq.(4) with input* $\mathbf{A} = \mathbf{\Gamma}_t$. *Thanks to the commutative law in algebraic operation of matrix trace and* $\mathbf{A}$ *is a symmetric matrix, Eq.(13) can be decomposed as*

$$Tr\left(\mathbf{\Phi}\mathbf{\Theta}\right), \tag{14}$$

where two symmetric matrices are respectively defined as $\mathbf{\Phi} = \mathbf{A}^{\frac{1}{2}}\mathbf{W}_t\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right)^{\dagger}\mathbf{W}_t^T\mathbf{A}^{\frac{1}{2}} \in \mathbb{R}^{d\times d}$, $\mathbf{\Theta} = \mathbf{A}^{\frac{1}{2}}\widehat{\mathbf{W}}_{t+1}\widehat{\mathbf{W}}_{t+1}^T\mathbf{A}^{\frac{1}{2}} \in \mathbb{R}^{d\times d}$. *Considering **Theorem 4.3.53** in [Horn and Johnson, 2012], we have*

$$Tr\left(\mathbf{\Phi}\mathbf{\Theta}\right) \leq \sum_{i=1}^{d}\lambda_i\left(\mathbf{\Phi}\right)\lambda_i\left(\mathbf{\Theta}\right) \leq \sum_{i=1}^{m}\lambda_i\left(\mathbf{\Theta}\right), \tag{15}$$

which provides an evidence of correctness for Eq.(9). Moreover, note that rank$(\mathbf{\Theta}) \leq$ *rank*$(\widehat{\mathbf{W}}_{t+1}) = m$, *then we have* $\sum_{i=1}^{m}\lambda_i\left(\mathbf{\Theta}\right) = Tr\left(\mathbf{\Theta}\right)$. *Combining Eq.(12), Eq.(13) and Eq.(15), we can conclude that*

$$Tr\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right) \leq Tr\left(\widehat{\mathbf{W}}_{t+1}^T\mathbf{A}\widehat{\mathbf{W}}_{t+1}\right). \tag{16}$$

According to the previous definitions, i.e., $\widehat{\mathbf{W}}_{t+1} = \mathbf{Q}_{td}^{k}\widehat{\mathbf{V}}_t$ *and* $\mathbf{W}_{t+1} = \mathbf{Q}_{td}^{k}\mathbf{V}_t$, *we can finally obtain that*

$$\begin{aligned} Tr\left(\mathbf{W}_t^T\mathbf{A}\mathbf{W}_t\right) &= Tr\left(\left(\widehat{\mathbf{V}}_t\right)^T\left(\mathbf{Q}_{td}^{k}\right)^T\mathbf{A}\mathbf{Q}_{td}^{k}\widehat{\mathbf{V}}_t\right) \\ &\leq Tr\left(\widehat{\mathbf{W}}_{t+1}^T\mathbf{A}\widehat{\mathbf{W}}_{t+1}\right), \end{aligned} \tag{17}$$

then the proof of Theorem 2 has been completed.

3.3 Subspace Sparsity Discriminant Feature Selection

In this part, we propose to introduce a novel feature selection method named S^{2}DFS, and optimize it by using S^{3}L module.

Model Formulation. In order to improve model's discriminative power, S^{2}DFS leverages trace ratio LDA model in Eq.(1) for pulling samples within same class together and pushing those samples between different classes far away in learned subspace. Besides, we use the structured sparse subspace constraint to directly facilitate the exploration and inference of important features of data. In light of these points, the objective function of S^{2}DFS model can be formulated as

$$\begin{aligned} &\max_{\mathbf{W}\in\mathbb{R}^{d\times m}} \frac{Tr\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{Tr\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)} \\ &s.t. \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}_{m\times m}, \|\mathbf{W}\|_{2,0} = k, \end{aligned} \tag{18}$$

where $m \leq k < d$, m is the number of selected features and k means the number of non-zero rows in $\mathbf{W}$. Due to the nonlinear constraints and trace ratio objective, optimizing such a maximization ratio problem in Eq.(18) is challenging.

Optimization Algorithm Before solving problem (18), we first consider to solve a general maximization ratio problem as follow:

$$\max_{\mathbf{x}\in\mathcal{C}} \frac{f(\mathbf{x})}{g(\mathbf{x})}, \tag{19}$$

where $\mathbf{x} \in \mathcal{C}$ is arbitrary constraint on $\mathbf{x}$ and as the denominator, $g(\mathbf{x}) > 0$. In **Algorithm 2**, we summarize an alternative optimization algorithm to solve the general problem (19), and a series of theorems and proofs are presented to guarantee its convergence.

Algorithm 2 Algorithm to solve the general maximization ratio problem (19).

```
1. Initialization: $x \in \mathcal{C}$, $t = 1$.
While not converge do
  2. Calculate $\lambda_t = \frac{f(\mathbf{x}_t)}{g(\mathbf{x}_t)}$.
  3. Calculate $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) - \lambda_t g(\mathbf{x})$.
  4. $t = t + 1$.
end while
```

Theorem 3 *The global solution of the general maximization ratio problem (19) is equivalent to the root of following function:*

$$h(\lambda) = \arg\max_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) - \lambda_t g(\mathbf{x}). \tag{20}$$

Proof 3 *Supposing* $\mathbf{x}^*$ *is the global solution of problem (19) and its corresponding maximal objective function value is* λ^*, *then the following equation holds:* $\frac{f(\mathbf{x}^*)}{g(\mathbf{x}^*)} = \lambda^*$. *As a result,* $\forall\ \mathbf{x} \in \mathcal{C}$, *we always have* $\frac{f(\mathbf{x})}{g(\mathbf{x})} \leq \lambda^*$. *As the condition*

$g(\mathbf{x}) > 0$, *we can infer that* $f(\mathbf{x}) - \lambda^* g(\mathbf{x}) \leq 0$ *which means:*

$$\max_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) - \lambda^* g(\mathbf{x}) = 0 \quad \Rightarrow \quad h(\lambda^*) = 0. \tag{21}$$

Consequently, the global maximal objective function value λ^* *of problem (19) is the root of function* $h(\lambda)$. *Thus, the proof of* ***Theorem 3*** *is completed.*

□

Theorem 4 ***Algorithm 2*** *increases the objective function value of problem (19) in each iteration until it reaches convergence.*

Proof 4 *According to the step 2 in* ***Algorithm 2****, we have* $f(\mathbf{x}_t) - \lambda_t g(\mathbf{x}_t) = 0$*, and from step 3, we can infer that* $f(\mathbf{x}_{t+1}) - \lambda_t g(\mathbf{x}_{t+1}) \geq f(\mathbf{x}_t) - \lambda_t g(\mathbf{x}_t)$*. Combining above two inequalities, we can obtain* $f(\mathbf{x}_{t+1}) - \lambda_t g(\mathbf{x}_{t+1}) \geq 0$ *that equals to* $\frac{f(\mathbf{x}_{t+1})}{g(\mathbf{x}_{t+1})} \geq \lambda_t = \frac{f(\mathbf{x}_t)}{g(\mathbf{x}_t)}$*. That is,* ***Algorithm 2*** *increases the objective function value in Eq.(19) in each iteration and the proof of* ***Theorem 4*** *has been finished.*

□

Systematically, it is worth noting that **Theorem 3–4** provide a complete framework to optimize the general maximization ratio problem in Eq.(19) with rigorously convergent proof.

According to the **Algorithm 2**, we can infer that the key of solving proposed maximization ratio problem in Eq.(18) is to solve the following problem:

$$\begin{aligned} &\max_{\mathbf{W}\in\mathbb{R}^{d\times m}} Tr\left(\mathbf{W}^T\mathbf{B}\mathbf{W}\right) \\ &s.t. \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}, \|\mathbf{W}\|_{2,0} = k, \end{aligned} \tag{22}$$

where $\mathbf{B} = \mathbf{S}_b - \lambda\mathbf{S}_w + \eta\mathbf{I}$ is a positive semi-definite matrix if η is large enough. In theory, the relaxation parameter η is easily to be calculated as the largest eigenvalue of $\mathbf{S}_b - \lambda\mathbf{S}_w$. The first way that can be thought to acquire η is eigenvalue decomposition, which however is time-consuming in dealing with high-dimensional data. Here, we employ a power iteration method [Nie *et al.*, 2017] to solve above issue, which can obtain the relaxation parameter η faster than eigenvalue decomposition and improves the applicability of our method.

Note that problem (22) is a special case of problem (2), which can be directly addressed by using proposed S^{3}L module and obtain a closed-form solution.

4 Experiments

In this section, for verifying the effectiveness of proposed method, we first present an visualization experiment on Mnist dataset. Then, we evaluate its classification performance on nine publicly available high-dimensional datasets. Finally, we exhibit the convergent speed of our method.

4.1 Preliminary

Datasets. We evaluate the performance of proposed method on several high-dimensional real-world datasets, and more details about them are shown in Table **??**. For the color image datasets, i.e., Pubfig [Xu *et al.*, 2018], we firstly downsample each image into a suitable scale and extract 100 LOMO features [Liao *et al.*, 2015] for data representation.

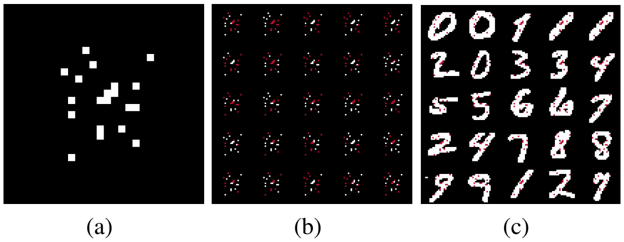
Datasets	Dim.	Class	Num.	Type
20News[1]	8,014	4	3,970	Text
WebKB[2]	4,165	7	1,166	Web
Binalpha[3]	320	36	1,404	Letter
Pixraw10P[3]	10,000	10	100	Face image
Text1[3]	7,511	2	1,946	Text
Pubfig	65,536	8	772	Face image

Table 1: Descriptions of datasets.

Experimental setting. We compare our method to several SOTA feature selection methods including **DLSRFS** [Xiang *et al.*, 2012], **RFS** [Nie *et al.*, 2010], **CRFS** [He *et al.*, 2012], **RJFWL** [Yan *et al.*, 2016], **infFS** [Roffo *et al.*, 2015] and **ESFS** [Pang *et al.*, 2018] in the classification task. We use the k-nearest neighbor algorithm as the classifier, all experiments are repeatedly conducted 10 times and the average recognition rate and standard deviation are recorded as the measurement of performance for all competitors.

4.2 Visualization Experiment

In order to intuitively show the performance of feature selection algorithm, we provide an visualization experiment conducted on Mnist digits dataset. Figure 2b exhibits the two types of selected features, i.e., digital pixels (white dots) and background pixels (red dots) respectively. In the Figure 2c, we show the selected features lied on the Mnist digits, from which we can observe that the shape of digits can be distinctly reconstructed by selected features generated by proposed method. Experimental results verify the discriminability of selected features that learned from our method.



(a) (b) (c)

Figure 2: The results of using proposed S^{2}DFS to select 20 most informative pixels of Mnist digits. (a) The selected 20 features on each MNIST digit that are shown in white pixels. (b) The selected features in training samples which are in red and white dots. (c) The selected features (red points) lied on the Mnist digits.

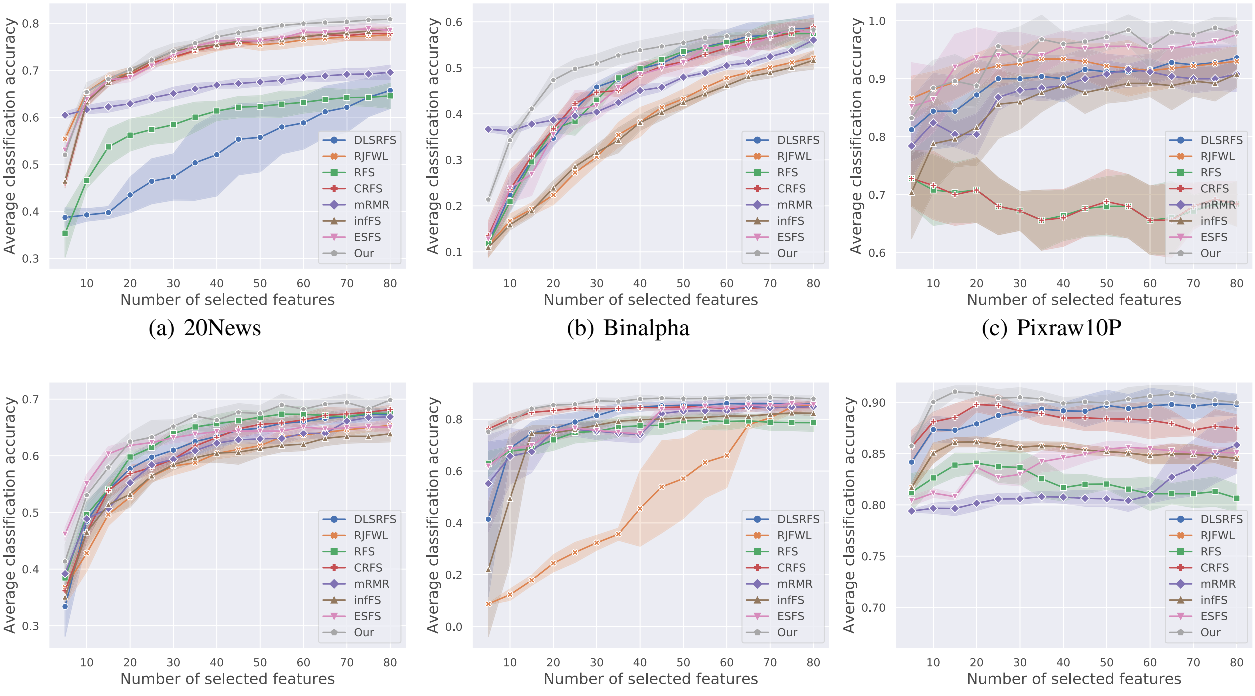
4.3 Classification Experiment

Figure 3 shows the statistical estimation of classification results on six high-dimensional real-world datasets, which demonstrates the discriminability of feature subsets selected by our method. Concretely, in general, comparing to all competitors, our method always achieves comparable or even

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/

[3] http://www.escience.cn/system/file?fileId=82035



(a) 20News (b) Binalpha (c) Pixraw10P

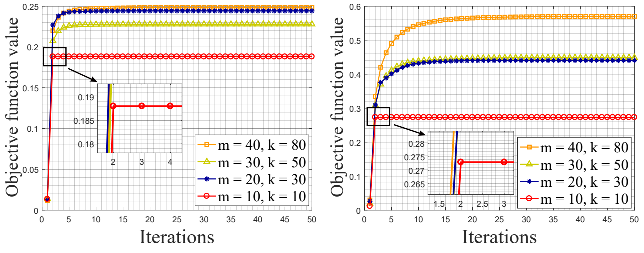
(d) Pubfig (e) WebKB (f) Text1

Figure 3: The error bar figure of classification accuracy with different numbers of selected features on six real-world datasets.

highest mean accuracies on almost all datasets with different number of selected features. Specifically, due to the nonparameter property of our method, it achieves stable performance on all datasets. Furthermore, our method outperforms ESFS in different degrees on various datasets, which results from that ESFS applies approximate gradient descent algorithm to obtain a suboptimal solution, while our method obtains a closed-form solution, which guarantees stability of our method's performance.

4.4 Convergence Analysis

Figure 4 plots the convergence curves of our iterative algorithm with different number of selected features m and model's sparsity k on 20News and WebKB datasets. Overall, our algorithm reaches convergent within 15 iterations in most cases or even less than 5 iterations on 20News dataset. In detail, a subfigure embedded in each figure is used to facilitate observation of the curve convergence when $k = m$. We discover the fact that the curve reaches convergent with one iteration on all datasets, which verifies the high-efficiency of proposed optimization algorithm.



(a) 20News (b) WebKB

Figure 4: Convergence curve of objective function value in Eq.(18) with different number of selected features m and model sparsity k on 20News (a) and WebKB (b) datasets respectively.

5 Conclusion

In this paper, we have proposed explicit discriminative feature selection algorithm via employing structured sparse subspace constraint, which is a NP-hard optimization problem. To our best knowledge, there are no techniques able to obtain closed-form solution and proves convergence. As the major contribution of this paper, we provide an ideal iterative optimization algorithm that can directly solve proposed NP-hard problem and reaches convergent very fast in both theory and practice. Experimental results demonstrate the effectiveness of proposed method. In future works, we intend to develop nonlinear feature selection model via proposed S^{3}L module for future improving discriminability of selected features.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the National Natural Science Foundation of China under Grant 61936014, Grant 61772427 and Grant 61751202, and in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501.

References

[Abid *et al.*, 2019] Abubakar Abid, Muhammad Fatih Balin, and James Zou. Concrete autoencoders for differentiable feature selection and reconstruction. *arXiv preprint arXiv:1901.09346*, 2019.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Exact top-k feature selection via l2, 0-norm constraint. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1240–1246, 2013.

[Du *et al.*, 2018] Xingzhong Du, Feiping Nie, Weiqing Wang, Yi Yang, and Xiaofang Zhou. Exploiting combination effect for unsupervised feature selection by $\ell_{2,0}$-norm. *IEEE transactions on neural networks and learning systems*, 30(1):201–214, 2018.

[Gu *et al.*, 2012] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*, 2012.

[Han *et al.*, 2018] Kai Han, Yunhe Wang, Chao Zhang, Chao Li, and Chao Xu. Autoencoder inspired unsupervised feature selection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2941–2945. IEEE, 2018.

[He *et al.*, 2012] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. $\ell_{2,1}$ regularized correntropy for robust feature selection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2504–2511. IEEE, 2012.

[Horn and Johnson, 2012] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[Maldonado and Weber, 2009] Sebastián Maldonado and Richard Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009.

[Ming and Ding, 2019] Di Ming and Chris Ding. Robust flexible feature selection via exclusive l21 regularization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3158–3164. AAAI Press, 2019.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[Nie *et al.*, 2017] Feiping Nie, Rui Zhang, and Xuelong Li. A generalized power iteration method for solving quadratic problem on the stiefel manifold. *Science China Information Sciences*, 60(11):112101, 2017.

[Nie *et al.*, 2019] Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, and Xuelong Li. Towards robust discriminative projections learning via non-greedy $l_{2,1}$-norm minmax. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, DOI: 10.1109/TPAMI.2019.2961877.

[Pang *et al.*, 2018] Tianji Pang, Feiping Nie, Junwei Han, and Xuelong Li. Efficient feature selection via $\ell_{2,0}$-norm constrained sparse regression. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):880–893, 2018.

[Roffo *et al.*, 2015] Giorgio Roffo, Simone Melzi, and Marco Cristani. Infinite feature selection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4202–4210, 2015.

[Sun *et al.*, 2016] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.

[Tian *et al.*, 2019] Lai Tian, Feiping Nie, and Xuelong Li. Learning feature sparse principal components. *arXiv preprint arXiv:1904.10155*, 2019.

[Vu *et al.*, 2013] Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[Wang *et al.*, 2007] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[Wang *et al.*, 2014] Zhaoran Wang, Huanran Lu, and Han Liu. Tighten after relax: Minimax-optimal sparse pca in polynomial time. In *Advances in neural information processing systems*, pages 3383–3391, 2014.

[Xiang *et al.*, 2012] Shiming Xiang, Feiping Nie, Gaofeng Meng, Chunhong Pan, and Changshui Zhang. Discriminative least squares regression for multiclass classification and feature selection. *IEEE transactions on neural networks and learning systems*, 23(11):1738–1754, 2012.

[Xu *et al.*, 2018] Jie Xu, Lei Luo, Cheng Deng, and Heng Huang. Bilevel distance metric learning for robust image recognition. In *Advances in Neural Information Processing Systems*, pages 4198–4207, 2018.

[Yan *et al.*, 2016] Hui Yan, Jian Yang, and Jingyu Yang. Robust joint feature weights learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 28(5):1327–1339, 2016.

[Yang and Xu, 2015] Wenzhuo Yang and Huan Xu. Streaming sparse principal component analysis. In *International Conference on Machine Learning*, pages 494–503, 2015.

[Zhang *et al.*, 2020] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):659–673, 2020.