A GENERALIZED UNCORRELATED RIDGE REGRESSION WITH NONNEGATIVE LABELS FOR UNSUPERVISED FEATURE SELECTION

Han Zhang, Rui Zhang, Feiping Nie, Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shaanxi, P.R. China, 710072.

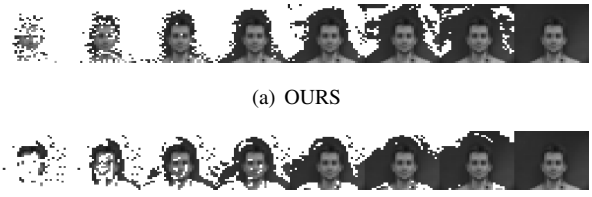
ABSTRACT

The ridge regression has been widely applied in multiple domains and gains the promising performance. However, due to the unavailability of labels, the ridge regression easily incurs the trivial solution towards unsupervised learning. In this paper, we investigate unsupervised feature selection by virtue of an uncorrelated and nonnegative ridge regression model (UNRFS). To be specific, a generalized uncorrelated constraint on the projection matrix, and a nonnegative orthogonal constraint on the indicator matrix are imposed upon the proposed regression model. With the proposed method, the most uncorrelated features on the embedded Stiefel manifold is exploited for feature selection and trivial solutions of projection matrix are avoided as well. Besides, equipped with a generalized scatter matrix, the proposed uncorrelated constraint is superior to conventional uncorrelated constraint, since the closed form solution can be achieved directly. In addition, owing to the nonnegative of real labels, the nonnegative orthogonal constraint is employed to suppress the indicator matrix such that the learned labels confront to reality further.

***Index Terms*—** Feature selection, ridge regression, generalized uncorrelated constraint, nonnegative labels

1. INTRODUCTION

As technologies develop rapidly, a large number of data are generated with high dimensionality. However, most of these features are inessential to their topics, but aggravate the burden of both computation and memory vainly. Therefore, the researches on feature selection are of great significance, which dedicate to selecting the most valuable features from original ones. Since the labels of instances are laborious and expensive to be acquired in most occasions, unsupervised learning becomes more practical for feature selection, which is also under hot investigation in the domains of machine learning, data mining and pattern recognition, etc. Lots of efforts have been made for unsupervised feature selection, which are generally divided into three categories, including filter methods [1, 2, 3], wrapper methods [4] and embedded methods [5, 6, 7, 8, 9]. Essentially speaking, filter methods utilize statistical properties of data to rank the ability of each feature in preserving the intrinsic structure of data. As for wrapper methods, which incorporate feature selection into learning algorithms, they usually depend on specific classifiers tightly. Embedded methods are considered more rational than aforementioned ones, because intrinsic structure of original data tends to existing in the manifold rather than the ambient Euclidean space. Consequently, we propose a novel embedded based approach in this paper.



(a) OURS

(b) SOGFS

Fig. 1. An instance from Imm40 with number of selected features varying in $\{128, 256, 384, 512, 640, 768, 896, 1024\}$ from left to right.

Although plenty of methods concerning unsupervised feature selection have been put forward recently, their performance can still be improved based on three points: (i). most of current approaches learn the labels via spectral analysis, which requires constructing a similarity matrix among data with a high cost of computation; (ii). the unavailability of real labels makes the ridge regression model towards unsupervised learning easily encounter a trivial solution, i.e., null projection matrix; (iii). the common orthogonal or uncorrelated constraint makes the regression models with regularization terms difficult to tackle. Therefore, we propose a novel unsupervised feature selection via evolving the conventional ridge regression model, and main contributions of this paper are highlighted as follows.

- A generalized uncorrelated constraint is imposed to the ridge regression model, which makes the proposed method equip with the closed form solution, and exploits the uncorrelated features, i.e., features with little

*Corresponding author

interrelated information, on the embedded Stiefel manifold meanwhile.

- A nonnegative orthogonal indicator matrix is under investigation to acquire the pseudo labels for unsupervised feature selection, and the $\ell_{2,1}$-norm is employed to ensure the row-sparsity of projection matrix.
- An efficient algorithm is designed to achieve the proposed feature selection method, and extensive experiments are conducted to verify the effectiveness and superiority of the proposed method.

Notations: $\mathbf{Tr}(\mathbf{M})$ denotes the trace of matrix $\mathbf{M}$. $\mathbf{M}^T$ denotes the transpose of $\mathbf{M}$. $\|\mathbf{M}\|_F$ denotes the F-norm of $\mathbf{M}$. $\mathbf{I}_n$ is an $n \times n$ identity matrix. $\mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbb{R}^{n\times 1}$. $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is defined as a centering matrix. $\mathbf{M}^i$ denotes the i-th row of matrix $\mathbf{M}$, while $\mathbf{M}_j$ denotes the j-th column of $\mathbf{M}$, and $\mathbf{M}_{ij}$ denotes the entry in i-th row, j-th column of $\mathbf{M}$. For any matrix $\mathbf{M} \in \mathbb{R}^{d\times c}$, its $\ell_{2,1}$-norm is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{c}\mathbf{M}_{ij}^2} = \sum_{i=1}^{d}\|\mathbf{M}^i\|_2$.

2. METHODOLOGY

Suppose that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ represents an original dataset, which contains n d-dimensional instances and belongs to c clusters. $\mathbf{Y} \in \mathbb{R}^{n\times c}$ preserves binary real labels, where $\mathbf{Y}_{ij}$ is set to 1 if the $\mathbf{x}_i$ belongs to j-th cluster, and 0 otherwise. The classical ridge regression model is represented as:

$$\min_{\mathbf{W},\mathbf{m}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{m}^T - \mathbf{Y}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the embedded subspace; α is a regularization parameter, and $\mathbf{m} \in \mathbb{R}^{c\times 1}$ is the bias. With this model, the intrinsic structure of data could be explored on the manifold. However, the label matrix $\mathbf{Y}$ of data in unsupervised learning is unavailable, which makes problem (1) easily incur a trivial solution, i.e., W is null, provided that $\mathbf{1}\mathbf{m}^T = \mathbf{Y}$. In order to address this problem, we present a novel ridge regression model specifically for unsupervised learning, which is formulated as follows:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{m},\mathbf{F}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{m}^T - \mathbf{F}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 \\ &s.t.\quad \mathbf{W}^T\mathbf{S}_t^{(\mathcal{G})}\mathbf{W} = \mathbf{I}_c, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0\end{aligned} \tag{2}$$

where $\mathbf{W}$ is imposed by a generalized uncorrelated constraint with a generalized scatter matrix $\mathbf{S}_t^{(\mathcal{G})} = \mathbf{S}_t + \mathcal{G}$, where $\mathbf{S}_t = \mathbf{X}\mathbf{H}\mathbf{X}^T$ is the common scatter matrix, and $\mathcal{G}$ depends on the regularization terms of the final model. With this constraint, (i). the trivial null solution of projection matrix is avoided; (ii). the uncorrelated features are exploited, since the covariance matrix of projected dimensions are orthogonal to some extent. (iii). the optimization of model (2) is simplified, since the ingredient of regularization is considered as well, which makes several terms integrally be a constant.

In supervised learnings, $\mathbf{F} \in \mathbb{R}^{n\times c}$ serves as the indicator matrix which preserves the cluster information, and defined as $\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$. This operation indicates that any label matrix could be of orthogonality, i.e., $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$. Based on this, we spontaneously impose an orthogonal constraint to $\mathbf{F}$ in unsupervised learning. Meanwhile, since real labels are nonnegative, $\mathbf{F}$ should be nonnegative so as to approximate the ground truth as much as possible.

Since $\|\mathbf{W}^i\|_2$ is utilized to measure the value of i-th feature, a row-sparse $\mathbf{W}$ is required usually. Considering the effectiveness of $\ell_{2,1}$-norm serving as sparsity regularization has been verified in [10] and employed in plenty of approaches [8, 7, 11, 12, 13], $\ell_{2,1}$-norm is also employed in the proposed method. Consequently, the final UNRFS is formulated as:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{m},\mathbf{F}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{m}^T - \mathbf{F}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 + \beta\|\mathbf{W}\|_{2,1} \\ &s.t.\quad \mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0\end{aligned} \tag{3}$$

where β is another regularization parameter. The generalized scatter matrix $\mathbf{S}_t^{(\mathcal{G})}$ in problem (2) is specifically defined as $\mathbf{S}_t^{(\widetilde{D})}$, where $\mathcal{G} = \alpha\mathbf{I}_d + \beta\mathbf{D}$ is the overall re-weighted matrix of regularization terms in model (3), and $\mathbf{D}$ is a $d\times d$ diagonal matrix with $\mathbf{D}_{ii} = \frac{1}{2\sqrt{\|\mathbf{W}^i\|_2^2+\varepsilon}}(\varepsilon \to 0, i = 1, 2, \cdots, d)$.

This generalized uncorrelated constraint contains two terms, i.e., $\mathbf{W}^T\mathbf{S}_t\mathbf{W}$ and $\mathbf{W}^T(\alpha\mathbf{I}_d + \beta\mathbf{D})\mathbf{W}$. When α and β are small values, the first term dominates $\mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W}$, which makes the projected dimensions uncorrelated to each other to a great extent, since the first term, i.e., the covariance matrix of projected dimensions, extremely approaches to orthogonal. Besides, this constraint equips problem (3) with the closed form solution, and it will be intuitively reflected in the optimization procedure.

3. OPTIMIZATION

3.1. A Counterpart of Problem (3)

First, we optimize the bias $\mathbf{m}$ to simplify the problem (3). Denote the Lagrangian function of Eq. (3) w.r.t. $\mathbf{m}$ as $\mathcal{L}(\mathbf{m})$. According to the extreme value condition, the derivative of $\mathcal{L}(\mathbf{m})$ is zero at the optimal point, i.e., $\frac{\partial\mathcal{L}(\mathbf{m})}{\partial\mathbf{m}} = 0$, thus the optimal $\mathbf{m}$ could be achieved. This optimal bias essentially centers the estimated errors. Besides, according to Theorem 1 in [10], $\|\mathbf{W}\|_{2,1}$ in problem (3) could be replaced by its re-weighted counterpart, i.e., $\mathbf{Tr}(\mathbf{W}^T\mathbf{D}\mathbf{W})$. Thus, problem (3) is addressed by solving the following counterpart:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{F}} \|\mathbf{H}(\mathbf{X}^T\mathbf{W} - \mathbf{F})\|_F^2 + \alpha\mathbf{Tr}(\mathbf{W}^T\mathbf{W}) + \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{D}\mathbf{W}) \\ &s.t.\quad \mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0\end{aligned} \tag{4}$$

3.2. Optimize F with Fixed W

With $\mathbf{W}$ being fixed, problem (4) can be transformed into:

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c,\mathbf{F}\geq 0} \mathbf{Tr}(\mathbf{F}^T\mathbf{H}\mathbf{F} - 2\mathbf{F}^T\mathbf{H}\mathbf{X}^T\mathbf{W}) \tag{5}$$

Due to the non-convexity of $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$, the problem above is difficult to solve directly. Consequently, we could relax this problem into another one:

$$\min_{\mathbf{F}\geq 0} \mathbf{Tr}(\mathbf{F}^T\mathbf{H}\mathbf{F} - 2\mathbf{F}^T\mathbf{H}\mathbf{X}^T\mathbf{W}) + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 \tag{6}$$

For problem (6), F would satisfy the orthogonality provided that $\gamma \to \infty$. Considering the Lagrangian function of Eq. (6), it is described as follows:

$$\min_{\mathbf{F}\geq 0} \mathbf{Tr}(\mathbf{F}^T\mathbf{H}\mathbf{F} - 2\mathbf{F}^T\mathbf{H}\mathbf{X}^T\mathbf{W}) + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 - \mathbf{Tr}(\mathbf{\Lambda}\mathbf{F}^T) \tag{7}$$

where $\mathbf{\Lambda} \geq 0$ is the Lagrangian multiplier matrix, and $\mathbf{\Lambda} \in \mathbb{R}^{n\times c}$. In order to achieve the optimal solution of $\mathbf{F}$, the Karush-Kuhn-Tuckre (KKT) condition of Eq. (6) is necessary to be satisfied, i.e.,

$$\begin{cases} 2\mathbf{H}\mathbf{F} - 2\mathbf{H}\mathbf{X}^T\mathbf{W} + 2\gamma(\mathbf{F}\mathbf{F}^T\mathbf{F} - \mathbf{F}) - \mathbf{\Lambda} = 0 \\ (\forall i, j)\ \ \mathbf{F}_{ij} \geq 0,\ \ \mathbf{\Lambda}_{ij} \geq 0,\ \ \mathbf{\Lambda}_{ij}\mathbf{F}_{ij} = 0 \end{cases} \tag{8}$$

So, depending on the equations and inequations above, the indicator matrix $\mathbf{F}$ could be updated by

$$\mathbf{F}_{ij} = \mathbf{F}_{ij}\frac{(\gamma\mathbf{F})_{ij}}{(\mathbf{H}\mathbf{F} - \mathbf{H}\mathbf{X}^T\mathbf{W} + \gamma\mathbf{F}\mathbf{F}^T\mathbf{F})_{ij}} \tag{9}$$

Noting that the achieved $\mathbf{F}$ is an approximate optimal solution to problem (5), and it needs to be normalized so as to satisfy $(\mathbf{F}^T\mathbf{F})_{ii} = 1(i = 1, 2, \cdots, c)$ in practice.

3.3. Optimize W with Fixed F

Considering the generalized uncorrelated constraint, i.e., $\mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c$, problem (4) with fixed $\mathbf{F}$ holds the following deduction:

$$\begin{aligned} &\min_{\mathbf{W}} \|\mathbf{H}(\mathbf{X}^T\mathbf{W} - \mathbf{F})\|_F^2 + \alpha\mathbf{Tr}(\mathbf{W}^T\mathbf{W}) + \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{D}\mathbf{W}) \\ &s.t.\ \ \mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c \\ &= \min_{\mathbf{W}} \mathbf{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W} - 2\mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{F} + \mathbf{F}^T\mathbf{H}\mathbf{F}) \\ &+\alpha\mathbf{Tr}(\mathbf{W}^T\mathbf{W}) + \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{D}\mathbf{W})\ \ s.t.\ \ \mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c \\ &\Leftrightarrow \max_{\mathbf{W}} \mathbf{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{F})\ \ s.t.\ \ \mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c \end{aligned} \tag{10}$$

Due to the uncorrelated constraint $\mathbf{W}^T\mathbf{S}_t^{(\widetilde{D})}\mathbf{W} = \mathbf{I}_c$, problem (10) is greatly simplified into the last formulation above, and we could address it by solving

$$\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_c} Tr(\mathbf{Q}^T\mathbf{B}) \tag{11}$$

where

$$\mathbf{Q} = (\mathbf{S}_t^{(\widetilde{D})})^{\frac{1}{2}}\mathbf{W},\quad \mathbf{B} = (\mathbf{S}_t^{(\widetilde{D})})^{-\frac{1}{2}}\mathbf{X}\mathbf{H}\mathbf{F} \tag{12}$$

The optimal $\mathbf{Q}$ can be achieved according to Lemma 1, and thus based on the definition of $\mathbf{Q}$ hereinbefore, the optimal $\mathbf{W}$ is achieved by $\mathbf{W} = (\mathbf{S}_t^{(\widetilde{D})})^{-\frac{1}{2}}\mathbf{Q}$.

Lemma 1. *The optimal solution* $\mathbf{Q}$ *to problem (11) is achieved by*

$$\mathbf{Q} = \mathbf{U}\mathbf{V}^T \tag{13}$$

where $\mathbf{U}$ *and* $\mathbf{V}$ *consist of singular vectors of compact SVD decomposition of* $\mathbf{B}$ *defined in Eq. (12) corresponding to its left and right singular values, respectively.*

Algorithm 1 Algorithm to solve the proposed UNRFS (3)

Input: Given data $\mathbf{X} \in \mathbb{R}^{d\times n}$, clustering number c, number of selected features h, coefficients α, β and γ.
Initialize a random and normalized matrix $\mathbf{F} \in \mathbb{R}^{n\times c}$ satisfying $\mathbf{F} \geq 0, \forall i, \|\mathbf{F}_i\|_2 = 1$; $\mathbf{D} = \mathbf{I}_d$;
Repeat:
1: Calculate $\mathbf{S}_t^{(\widetilde{D})} = \mathbf{X}\mathbf{H}\mathbf{X}^T + \alpha\mathbf{I}_d + \beta\mathbf{D}$.
2: Calculate $\mathbf{B}$ with Eq. (12).
3: Calculate $\mathbf{Q}$ by solving problem (11).
4: Update $\mathbf{W}$ by $\mathbf{W} = (\mathbf{S}_t^{(\widetilde{D})})^{-\frac{1}{2}}\mathbf{Q}$.
5: Update $\mathbf{F}$ by Eq. (9).
6: Normalize $\mathbf{F}$ to ensure $(\mathbf{F}^T\mathbf{F})_{ii} = 1, (i = 1, 2, \cdots, c)$.
7: Update $\mathbf{D} = \mathbf{diag}(\frac{1}{2\sqrt{\|\mathbf{W}^1\|_2^2+\varepsilon}}, \frac{1}{2\sqrt{\|\mathbf{W}^2\|_2^2+\varepsilon}}, \cdots, \frac{1}{2\sqrt{\|\mathbf{W}^d\|_2^2+\varepsilon}})$.
until **convergence**
Output: Calculate and sort $\|\mathbf{W}^i\|_2 (i = 1, 2, \cdots, d)$ in the descending order, then select the top h ranked features.

An alternative algorithm for solving problem (3) is summarized in Algorithm 1, which outputs the most h valuable features selected by the proposed UNRFS.

4. EXPERIMENTS

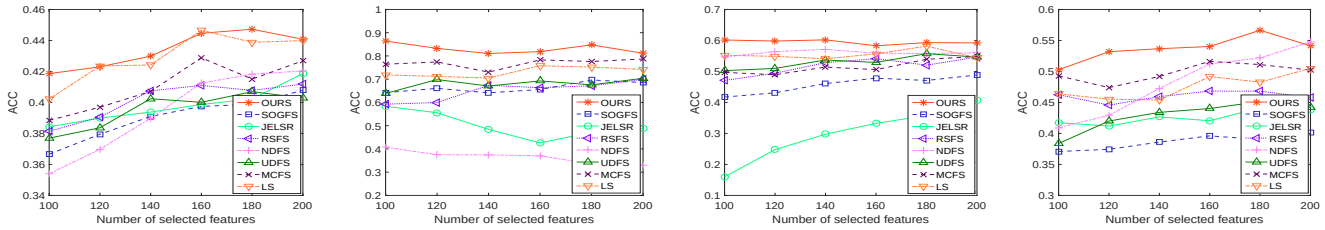
In this section, we conduct experiments to illustrate the superiority of the proposed method. Specific schemes for experiments involving datasets, compared methods and some settings are introduced at first.

Datasets: Columbia University Image Library (COIL20) [14], Imm40 [15], Binary Alphabet (BA) [1] and Pixraw10P [2].

Competitors: Seven state-of-the-art approaches are compared including MCFS [5], JELSR [7], NDFS [8], UDFS [11], SOGFS [12], RSFS [16] and LS [17].

Settings: Two metrics including clustering accuracy (ACC) [18] and normalized mutual information (NMI) [19]

[1] http://www.cs.nyu.edu/~roweis/data.html
[2] http://featureselection.asu.edu/datasets.php



(a) BA (b) Pixraw10P (c) COIL20 (d) Imm40

Fig. 2. k-means clustering accuracy on four benchmarks w.r.t. different numbers of selected features.

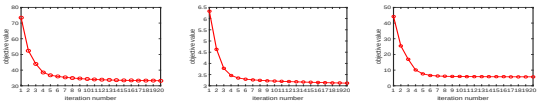
Table 1. Comparisons on ACC±std (%) and NMI±std (%) with 100 selected features.

	Datasets	OURS	SOGFS[12]	JELSR[7]	RSFS[16]	NDFS[8]	UDFS[11]	MCFS[5]	LS[17]
ACC±std (%)	BA	**41.86±1.9**	36.66±1.1	38.43±1.6	38.13±0.9	35.39±1.3	37.68±1.1	38.84±1.1	40.22±1.8
	Pix10P	**86.40±6.9**	64.20±5.4	58.35±5.0	59.25±5.0	40.75±8.5	63.85±5.5	76.45±10.5	71.85±8.3
	COIL20	**60.15±3.9**	41.73±1.9	15.98±1.8	47.09±3.1	54.67±4.2	50.25±2.8	49.69±3.4	55.21±4.5
	Imm40	**50.27±2.6**	37.08±2.0	41.73±2.8	46.25±3.1	40.91±1.3	38.37±1.4	49.29±2.9	46.37±2.6
NMI±std (%)	BA	**57.89±0.7**	53.03±0.7	53.44±0.9	54.41±0.7	50.45±0.5	53.56±0.6	55.41±0.6	56.07±0.7
	Pix10P	**90.56±4.1**	70.54±3.3	65.10±3.1	67.08±4.4	53.02±7.9	72.68±3.4	83.17±5.6	81.46±5.3
	COIL20	**73.55±1.5**	50.62±1.4	19.89±2.4	62.61±1.5	67.67±2.3	60.84±1.3	62.72±1.9	67.07±1.7
	Imm40	**72.65±1.3**	62.88±1.1	67.83±2.1	70.78±1.8	65.39±0.7	64.03±0.8	72.01±1.2	70.73±1.9

are employed to measure the performance among all comparisons. Both of them indicate the better performance with a larger value. Parameter γ in the proposed method is set to 100, which is large enough to ensure the orthogonality of $\mathbf{F}$. To achieve the best performance, α is searched in the range of small values from 0.01 to 0.1 with the interval of 0.01, while β is searched from 0.1 to 1 with the interval of 0.1. In order to avoid the occasionality triggered by random starting-points in k-means, clustering with the same selected features are performed for 20 times, and the average results are recorded.

4.1. Comparison on Performance

Fig. 2 shows the results of clustering accuracy with different numbers of selected features which range from 100 to 200 with 20 intervals, and the red lines with asterisks represent the performance of the proposed UNRFS. It is clear that UNRFS is totally beyond others on the datasets. Besides, we also compare NMI. Table 1 records the results of both ACC and NMI with standard deviations, and the number of selected features is set to 100. The best results are bold, and it is obvious that UNRFS gains considerable improvement than others. In addition, Fig. 1 shows an instance sampled from Imm40 with different numbers of features selected by UNRFS and SOGFS [12]. Distinctly, the proposed method selects more valuable features (e.g., face, eyes, nose, etc) primarily, specifically with less selected features, while SOGFS selects some valueless features (e.g., hair, body, background, etc).



(a) BA (b) Pixraw10P (c) COIL20

Fig. 3. Convergence behavior of UNRFS on different datasets, the x-axis represents the number of iteration, and y-axis records the objective value in Eq. (3).

4.2. Convergence Demonstration

Fig. 3 shows the objective values in Eq. (3) with respect to the number of iteration. All of them illustrate speedy convergence of the proposed UNRFS.

5. CONCLUSION

In this paper, we propose a novel unsupervised feature selection method via exploring the uncorrelated features on Stiefel manifold. The proposed method takes advantages of the classical ridge regression and evolves it for preventing the case of unsupervised learning from the trivial solution and equipping the model with the closed form solution. Besides, since real labels are nonnegative values, a nonnegative constraint is imposed on indicator matrix so as to learn exact labels. To validate the effectiveness and superiority of the proposed UNRFS, extensive experiments are conducted and demonstrate the promising performance of the UNRFS.

6. REFERENCES

[1] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2006, pp. 507–514.

[2] Zheng Zhao and Huan Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1151–1157.

[3] Mingjie Qian and Chengxiang Zhai, "Robust unsupervised feature selection," in *IJCAI*, 2013, pp. 1621–1627.

[4] Sina Tabakhi, Parham Moradi, and Fardin Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Engineering Applications of Artificial Intelligence*, vol. 32, pp. 112–123, 2014.

[5] Deng Cai, Chiyuan Zhang, and Xiaofei He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 333–342.

[6] Zheng Zhao, Lei Wang, Huan Liu, et al., "Efficient spectral feature selection with minimum redundancy," in *AAAI*, 2010, pp. 673–678.

[7] Chenping Hou, Feiping Nie, Xuelong Li, Dongyun Yi, and Yi Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793–804, 2014.

[8] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, Hanqing Lu, et al., "Unsupervised feature selection using nonnegative spectral analysis," in *AAAI*, 2012, vol. 2012, pp. 1026–1032.

[9] Suhang Wang, Jiliang Tang, and Huan Liu, "Embedded unsupervised feature selection," in *AAAI*, 2015, pp. 470–476.

[10] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[11] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou, "l2, 1-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI proceedings-international joint conference on artificial intelligence*, 2011, vol. 22, p. 1589.

[12] Feiping Nie, Wei Zhu, Xuelong Li, et al., "Unsupervised feature selection with structured graph optimization," in *AAAI*, 2016, pp. 1302–1308.

[13] Hongfu Liu, Ming Shao, and Yun Fu, "Consensus guided unsupervised feature selection," in *AAAI*, 2016, pp. 1874–1880.

[14] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al., "Columbia object image library (coil-20)," 1996.

[15] M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann, "The IMM face database - an annotated dataset of 240 face images," Tech. Rep., Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, may 2004.

[16] Lei Shi, Liang Du, and Yi-Dong Shen, "Robust spectral learning for unsupervised feature selection," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 977–982.

[17] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems 18*, Y. Weiss and P. B. Sch Eds.

[18] Christos H Papadimitriou and Kenneth Steiglitz, *Combinatorial optimization: algorithms and complexity*, Courier Corporation, 1998.

[19] Alexander Strehl and Joydeep Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.