AMSE JOURNALS-2016-Series: Advances A; Vol. 53; N°2; pp 78-92
Submitted July 2016; Revised Oct. 15, 2016, Accepted Dec. 10, 2016

A Novel Nonnegative Subspace Learning Approach for Unsupervised Feature Selection

* Wei Zheng

* School of Computer Engineering, Jinling Institute of Technology
Nanjing 211169, China, (vividzheng@163.com)

Abstract

Sparse subspace learning has been proven to be effective in data mining and machine learning. In this paper, we propose a novel scheme which performs robust feature selection with non-negative constraint and sparse subspace learning simultaneously. This work emphasizes joint l2, 1-norm and 1l minimization, where the former characterizes the weight matrix and the latter handles residual matrix to improve robustness. The Inexact Augmented Lagrange Multiplier framework has been adopted to solve our object function efficiently and extensive experimental results on original datasets with and without malicious pollutions have demonstrated the superiority of our new method of feature selection.

Key words

Data mining, unsupervised feature selection, sparse subspace learning, clustering

1. Introduction

With the advent of "Big Data" era, it is more convenient for us to obtain various data. However, we will inevitably encounter massive high dimensional data problems when dealing with machine learning tasks and data mining researches, such as pattern recognition, text mining, computer vision, and bio-informatics analysis. Input data without any processing is often correlated, redundant and sometimes noisy, which not only leads to cost of expensive time and memory, but also critically degrades the performance of the learning algorithms [30]. Therefore, seeking the optimal salient features is necessary for us to discover and understand the intrinsic structure of the data. In order to filter out the non-essential features, two typical methods are often used: feature extraction [1, 2] and feature selection [3]. This paper focuses on feature selection,

which is an effective way to reduce the dimensions of the original data. Instead of transforming data values which is utilized by feature extraction, feature selection algorithm selects a subset of the most relevant features based on certain evaluation criteria [31].

According to label availability, the algorithms of feature selection can be grouped into supervised methods [5-10] and unsupervised methods [11-17]. For supervised way, as class labels are given, it is adequate to keep only the features that are related to or lead to these classes. But in some real applications, labels of input data are difficult to obtain or expensive to be demarcated manually, so unsupervised feature selection draws more and more attention in recent years. Traditional unsupervised feature selection methods have addressed this issue by ranking top features based on some evaluation criterion computed independently for each feature (variance [11], trace ratio [8], fisher score [5], information gain [7], and so on). The evaluation criterions in these algorithms usually reflect the power of each feature in different clusters, but they utilize the statistical essentials only which neglect to explore the learning mechanism [18]. For this issue, some researchers put more emphasis upon manifold structure and a hand of feature learning approaches (UDFS [15], NDFS [16], EUFS [20], and MFFS [21]) have been proposed these days. Specifically, UDFS exploits local discriminative information and feature correlations where the manifold structure is considered. However, the strategy used by UDFS is inappropriate since the weight matrix is imposed by an orthogonal constraint [19]. NDFS proposes a non-negative spectral clustering method to learn the cluster labels of the samples, during which feature selection is performed simultaneously. These two methods (UDFS and NDFS) are sensitive to noise or outliers and their computation complexity is the cubic to the dimensions of the features. In addition, EUFS directly embeds feature selection into a clustering algorithm via sparse learning without the transformation used in NDFS. From the view of subspace learning, MFFS treats the feature selection process as a matrix factorization problem. The feature subset obtained by MFFS can approximately represent all features based on a potential assumption that all the features lie in a linear manifold in a real space. Nevertheless, there are three vulnerabilities in this process. First, the number of selected features needs to be specified in advance which is unrealistic in practical applications as the number of useful features is hard to be estimated before the task. The second is that the elements in the indication matrix cannot guarantee to be zeros or ones in the experiment. Last, L1-norm based loss function is sensitive to outlies and noises, which inevitably obstructs the following performance of feature selection.

In this paper we propose a robust unsupervised feature selection method from the viewpoint of non-negative matrix factorization (NSSLFS). Concretely, the global structure of the input data

can be captured in the process of the sparse subspace learning, where the weight matrix can be learned by virtue of the non-negative constraint. By doing this, we perform the robust feature selection and sparse subspace learning simultaneously, where each sub task can be iteratively refined by using the result of the other one. Its appealing characteristics are summarized as follows:

- Motivated by previous research [10], a $l_{2,1}$-norm regularization is performed to select features across all data points with joint sparsity.
- Instead of using l_2-norm based loss function that is vulnerable to outliers, a l_1-norm based loss function is adopted in our work to enhance the robustness to outlier.
- Our method performs the unsupervised feature selection task and sparse subspace learning in a unified framework where a non-negative constraint is imposed into the objective function, which is logically acceptable in real application.
- We propose to optimize our model by combining the IALM and $l_{2,1}$-norm minimization with non-negative constraint.

Extensive experiments have been conducted on various original datasets with or without malicious pollution, whose results validate the effectiveness of the proposed method in terms of clustering.

2. Matrix factorization criterion of feature selection from subspace learning

In machine learning and data mining, subspace learning techniques have been well studied and utilized in many applications. These methods usually learn a low-dimensional representation of high-dimensional space in order to exploit the global structure of data [22, 23]. Clearly, an effective strategy will more capable to reflect the global information. In this section, a matrix factorization criterion for feature selection is used from a subspace learning perspective [21]. For simplicity, we suppose that these features lie in a linear manifold of the real space.

Let $x_i \in R^{d\times 1}$ be the i-th data sample, and $X = [x_1, x_2, \cdots x_n] \in R^{d\times n}$ be a data matrix composed of all samples. H is the coefficient matrix of initial feature space in the selected space, k indicates the number of selected features, E_k is a k-by-k unit matrix, W is an indicator matrix in the following form:

$$W_{ij} = \begin{cases} 1 & \text{if the } i\text{-th feature is selected} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

From the viewpoint of matrix factorization, feature selection problem is expressed as follows:

$$\min_{W,H} \|X - XWH\|_F^2 \quad (2)$$
$$s.t. \quad W \geq 0, W^T W = E_k$$

This object function provides a unified framework for feature selection and feature extraction. However, in the practical application, the number of selected features k is difficult to determine beforehand, leading to the limited feasibility of this model. Experimental results show that the elements of W cannot satisfy the predefined condition (1).

3. Matrix factorization criterion of feature selection from subspace learning

Inspired by the recent advanced researches of feature subspace learning, we propose a novel framework to obtain the most meaningful features and to reject the noisy or irrelevant ones. In this section, we first illustrate the formulation of our method, and then a detailed inexact augmented Lagrange Multiplier approach is used to solve the objective problem.

3.1 The proposed model

In order to achieve feature selection, the feature weight matrix is constrained to be sparse in row, which is formulated as -norm minimization regularization term. Meanwhile, norm of the error matrix is applied to handle the sparse noise on the feature learning procedure. By combining these two terms, the foundation to realize the task of feature selection can be formulated as follow:

$$\min_{W,H,E} \|W\|_{2,1} + \lambda \|E\|_1$$
$$s.t. \quad X = XWH + E, W \in R_+^{d \times k}, H \in R_+^{k \times n}$$

(3)

where λ, k are parameters. We call the above model as robust unsupervised feature selection by nonnegative sparse subspace learning (NSSLFS).

3.2 Optimization algorithm

As can be seen in the previous formulation, our problem involves the $l_{2,1}$-norm which is non-smooth and cannot be solved in a closed form. As a result, we develop a solution based on the inexact augmented Lagrange Multiplier framework and alternating minimization technique [25]. The augment Lagrange function $J(W, H, E, Y, \mu)$ of the above problem can be addressed as

$$J(W,H,E,Y,\mu)=\|W\|_{2,1}+\lambda\|E\|_1+\langle Y,X-XWH-E\rangle+\frac{\mu}{2}\|X-XWH-E\|_F^2$$

(4)

where Y is Lagrange multiplier matrix, μ is a positive parameter and $\|\cdot\|_F$ denotes the Frobenius norm. IALM obtains the optimum solution by iteratively minimizing the augmented Lagrange function:

$$(W_{k+1},H_{k+1},E_{k+1})=\arg\min_{W,H,E} J_k(W,H,E,Y,\mu)$$

(5)

This problem cannot be solved directly as W, H and E are dependent on each other. Therefore, the variable W, H and E are alternatively updated in the solving process. Specifically, only one variable is updated one time with the others fixed.

1) Computing W: Fixing H and E, the optimization problem for updating W is equivalent to minimize the following model:

$$\min_W \|W\|_{2,1}+\langle Y,X-XWH-E\rangle+\frac{\mu}{2}\|X-XWH-E\|_F^2 \quad (6)$$

$$s.t.\quad W\geq 0$$

It can easily be solved by converting to the problem below [10]

$$\min_W Tr(W^TDW)+\langle Y,X-XWH-E\rangle+\frac{\mu}{2}\|X-XWH-E\|_F^2 \quad (7)$$

$$s.t.\quad W\geq 0$$

Here, D is a matrix with their diagonal elements $D_{ii}=\frac{1}{2}\|w^i\|_2$, w^i is the *i*-th row of W. Following the solution of Non-negative Matrix Factorization (NMF) [27], we introduce multiplicative updating rules. Let $\Phi=[\phi_{ij}]$ be the Lagrange multiplier for the constraint $W\geq 0$, the Lagrange function is

$$J(W)=Tr(W^TDW)+\langle Y,X-XWH-E\rangle+\frac{\mu}{2}\|X-XWH-E\|_F^2+Tr[\Phi W^T] \quad (8)$$

By computing the derivative of (8) with respect to W and setting it as zero, we have

$$2DW+\mu\left[X^TXWH+X^T\left(E-\frac{1}{\mu}Y-X\right)\right]H^T+\Phi=0$$

(9)

Using the KKT condition [29], $\phi_{ij}w_{ij}=0$ we obtain the updating rules:

$$W_{ij} \leftarrow W_{ij} \frac{(X^T Y H^T + \mu X^T X H^T)_{ij}}{(2DW + \mu X^T X W H H^T + \mu X^T E H^T)_{ij}} \tag{10}$$

2) Computing H: Fixing W and E, the optimization problem for H is equivalent to minimize the following function :

$$J(H) = \langle Y, X - XWH - E \rangle + \frac{\mu}{2} \|X - XWH - E\|_F^2 \tag{11}$$

$s.t. \quad H \geq 0$

Which can be solved using the similar technique as in 1) , thus, the non-negative updating rule of H is

$$H_{ij} \leftarrow H_{ij} \frac{(W^T X^T Y + \mu W^T X^T X)_{ij}}{(\mu W^T X^T X W H + \mu W^T X^T E)_{ij}}$$

(12)

3) Computing E: Fixing W and H, the optimization problem for E is equivalent to minimize the following function :

$$J(E) = \lambda \|E\|_1 + \frac{\mu}{2} \left\| X - XWH - E + \frac{1}{\mu} Y \right\|_F^2$$

(13)

Obviously, the minimizer of (16) can be computed using the soft-thresholding operator [4].

$$E_{s+1} = S_{\frac{\lambda}{\mu}} [X - XW_{s+1}H_{s+1} + \frac{1}{\mu_t} Y_s]$$

(14)

Here S is the soft-thresholding operator.

4) Updating Y

$$Y_{s+1} = Y_s + \mu_s \left(X - XW_{s+1}H_{s+1} - E_{s+1} \right)$$

(15)

5) Updating μ

$$\mu_{s+1} = \begin{cases} \rho\mu_s & if \ \ \mu_s \|E_{s+1} - E_s\| < \varepsilon \|X_{s+1}\|_F \\ \mu_s & otherwise \end{cases}$$

(16)

Where, $\rho > 1$ is a constant , and ε is the parameter in the stopping criteria.

Finally, we summarize the H and W update rules of the proposed optimization algorithm in Algorithm 1 and the IALM strategy adopted to solve the NSSLFS is listed in Algorithm 2.

3.3 Time complexity analysis

In this section, we briefly discuss the time complexity of each step within the IALM framework of algorithm 2. Giving the input data matrix size $d \times n$ and the indicator matrix size $d \times k$. Due to the inner loop in updating W, the time complexity in step 2 is $O\left(nd^2 + kd^2\right)$. The computational complexity of matrix multiplications in step 3 is $O\left(knd + kd\right)$. The computation cost for E and Y are both $O\left(knd\right)$ in step 4 and 5. Thus the total time complexity of Algorithm 2 is $O\left(knd + \left(k+n\right)d^2\right)$, in our case $d >> k$, the overall computation cost for NSSLFS is $O\left(t\left(knd + nd^2\right)\right)$, where *t* is the number of iterations.

Algorithm 1: H and W update algorithm

Input: $W_s \in R^{d\times k}, H_s \in R^{k\times d}, E_s \in R^{n\times d}, D_s \in R^{d\times d}, \mu_s, Y_s$

Initialize: $t = 0, l = 0, c = 0, W_t^{'} = W_s, H_l^{'} = H_s$

1. Repeat
2. While not converged do
3. Update $\left(W_{t+1}^{'}\right)_{ij} = \left(W_t^{'}\right)_{ij} \cdot \dfrac{\left(X^T Y_s H_l^{'T} + \mu X^T X H_l^{'T}\right)_{ij}}{(2D_t W_t^{'} + \mu X^T X W_t^{'} H_l^{'} H_l^{'T} + \mu X^T E_s H_l^{'T})_{ij}}$
4. $t = t + 1$
5. end while
6. Update the diagonal matrix D_{c+1} as

$$D_{c+1} = diag\left(\frac{1}{2\left\|(w')_t^1\right\|_2}, ..., \frac{1}{2\left\|(w')_t^d\right\|_2}\right)$$

7. $c = c + 1$
8. **until** Converges
9. **repeat**
10. Update $\left(H_{l+1}^{'}\right)_{ij} = \left(H_l^{'}\right)_{ij} \cdot \dfrac{\left(W_t^{'T} X^T Y_s + \mu_s W_t^{'T} X^T X\right)_{ij}}{\left(\mu_s W_t^{'T} X^T E_s + \mu_s W_t^{'T} X^T X W_t^{'} H_l^{'}\right)_{ij}}$
11. $l = l + 1$
12. **until** Converges

Output: $W_{s+1} = W_{t+1}^{'}, H_{s+1} = H_{l+1}^{'}$

Algorithm 2: NSSLFS for optimizing problem (3) via IALM

Input: Data matrix $X \in R^{d \times n}$, regularization parameter λ, the dimensions of the subspace k, and selected feature number p

Initialize: $X, E_0 = 0, \mu_0 = 1000, \rho = 1.6, s = 0, W_s, H_s, D_0 \in R^{d \times d}$ as an identity matrix.

1. While not converge, do
2. Update W_{s+1} by Algorithm 1;
3. Update H_{s+1} by Algorithm 1;
4. Update Y_s using (10);
5. Update μ_s using (12);
6. $s = s + 1$
7. If convergence criterion is satisfied then end while else continue

Output: Sort all d features according to $\left\| w_t^i \right\|_2, i = 1 \cdots d,$ in descending order and select the top p ranked features.

4 Experiments

In this section, experiments on several datasets are performed to show the effectiveness of our proposed formulation for the task of unsupervised feature selection.

4.1 Dataset

We conduct our tests on six available benchmark datasets from different fields serve as a good test bed for an overall merit. Details of these datasets are described in Table 1.

Table 1 Description of the datasets in the experiments

Data Sets	#samples	#features	#class	Data type
warpAR10P	130	2400	10	Face image
USPS_test	2007	256	10	Handwritten digit dataset
scene-classification	2407	294	6	Image library
COIL20	1440	1024	20	Object image dataset

Brain_Tumor1	90	5920	5	Biomedical dataset
AMLALL	72	7129	2	Biomedical dataset

4.2 Experiment setup and compared scheme

On each dataset, we compare NSSLFS with some existing well-known feature selection algorithms:

1. Baseline: All original features are selected.
2. LS: Feature selection method that can best preserve the local manifold structure of data.
3. MCFS: Features are selected based on spectral regression with l_1-norm regularization.
4. NDFS: Features are selected by a joint framework of nonnegative spectral analysis and $l_{2,1}$-norm regularized regression.
5. EUFS: Feature selection is embedded directly into a clustering algorithm via sparse learning without transformation.
6. MFFS: A new unsupervised feature selection criterion is developed from the viewpoint of subspace learning which is treated as a matrix factorization problem.

Following similar settings in previous work, we list the parameters and the numbers of selected features. In the compared test, we fix the neighborhood size to be 5 for all methods in all datasets. For MFFS we fix the penalty term to be 10^8 as used in [21]. All the parameters are tuned for each method by a "grid-search" strategy from the following set $\{10^{-6},10^{-4},\cdots,10^{4},10^{6}\}$ to compare different unsupervised feature selection algorithms fairly [26, 28]. The number of selected features p is taken as $\{20,40,60,\cdots,200\}$ for all datasets. The dimension of the subspace k is tuned in the range of $\{20,40,60,\cdots,200\}$. We report the best result of every method among various parameters.

For simplicity, the method based on k-means clustering is adopted to evaluate the performance of selected features by two widely used metrics: Accuracy (ACC) and Normalized Mutual Information (NMI). The bigger value is, the better effect is.

However, unstable results were often obtained while using k-means arithmetic owing to random selection of initial centers. Following the common way, clustering experiments are repeated 20 times to reduce the statistical variation and the average values with standard deviation are recorded.

4.3 Convergence results

As discussed in Section 3, we employ the IALM method to solve the optimization problem given in (3). Now, we conduct empirical study on convergence of NSSLFS and analysis the convergence behavior. The convergence curves on 8 test datasets are list in Fig.1, it can be observed that: 1) Objective function value decreases sharply until convergences to a fixed value. 2) Our model converges within 10 iterations, demonstrating that the proposed algorithm is efficiency.

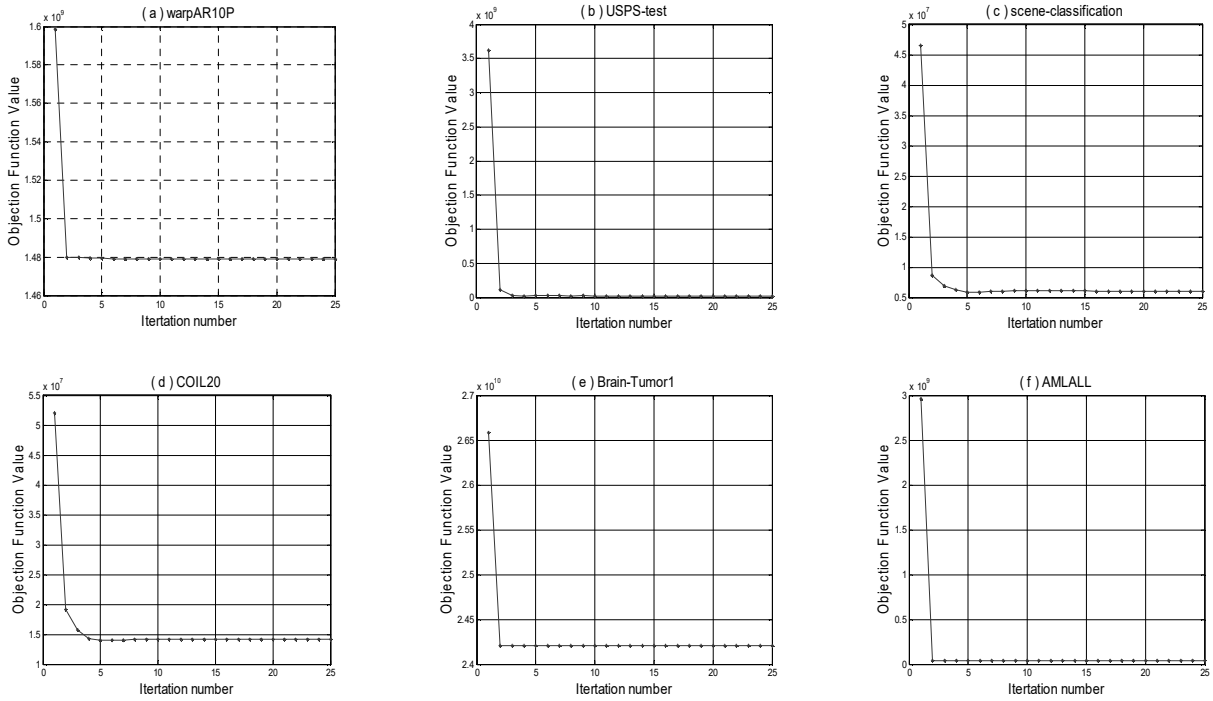

Fig. 1. Convergence curves of the objective function value in (3).

4.4 Clustering results of datasets

We evaluate the performance using raw data first; Table 2 and Table 3 demonstrate ACC and NMI comparison results of six algorithms and all features as baseline over eight datasets. The plots of clustering performance in terms of ACC and NMI versus the number of selected feature on datasets individually are given in Fig.2 and 3 respectively.

Table 2 ACC (%±std) of various unsupervised feature selection methods on different datasets. The best results are highlighted in bold.

Dataset	AllFea	LapScore	MCFS	NDFS	EUFS	MFFS	NSSLFS
warpAR10P	23.58±3.94	30.88±3.91	28.6±3.24	39.50±3.07	33.54±2.42	33.73±4.06	**44.23±3.85**

USPS_test	59.0±3.50	56.93±3.37	60.39±3.55	62.57±3.97	60.31±3.87	60.46±3.12	**63.64±1.85**
scene-classification	57.74±4.12	58.10±4.18	52.37±3.87	59.94±4.34	56.94±4.36	56.16±3.93	**61.19±4.44**
COIL20	59.17±3.98	57.66±4.14	59.75±3.69	61.03±4.46	59.58±5.09	53.39±3.49	**62.76±4.42**
Brain_Tumor1	49.44±3.79	50.61±6.07	49.06±5.76	67.72±6.75	46.44±3.88	60.33±8.51	**77.44±4.05**
AMLALL	66.6±5.29	75.0±0.00	65.4±6.70	75.5±3.62	74.2±0.71	67.7±5.66	**79.6±1.22**

Table 3 NMI (%±std) of various unsupervised feature selection methods on different datasets. The best results are highlighted in bold.

Dataset	AllFea	LapScore	MCFS	NDFS	EUFS	MFFS	NSSLFS
warpAR10P	20.28±5.33	33.7±3.69	27.05±3.33	41.2±3.92	35.2±3.26	33.2±3.84	50.20±4.40
USPS_test	56.24±1.68	55.9±1.60	57.46±1.75	58.2±1.68	57.7±1.43	57.6±1.59	58.37±1.44
scene-classification	39.03±1.49	39.4±1.31	36.89±1.36	40.2±0.93	38.7±0.80	36.3±3.10	41.43±1.94
COIL20	75.58±1.89	72.0±1.70	73.64±1.95	75.9±2.15	73.6±2.81	69.2±1.35	76.98±2.19
Brain_Tumor1	29.2±4.70	30.5±4.98	33.99±6.17	42.1±5.49	25.9±3.56	39.±4.51	46.±12.27
AMLALL	8.07±4.42	17.13±0.00	9.45±2.88	18.6±4.55	16.41±1.83	9.61±4.02	23.56±2.80

4.5 Datasets with Sparse Noise

The second experiment investigates the robustness of the proposed method to different levels of sparse noise. We take the warpAR10P dataset for instance, sparse noise within a ratio varies from 10% to 30% is added onto the original dataset and the clustering results over the malicious polluted data are evaluated. Similar to the experiment configuration in the first one, we record ACC and NMI for the methods over the datasets with different ratio of noise in Table 4 and 5.

Table 4 ACC (%±std) of warpAR10P dataset with different ratio of sparse noise. The best results are highlighted in bold.

ACC	AllFea	LapScore	MCFS	NDFS	EUFS	MFFS	NSSLFS
ratio = 0.1	24.04±5.22	25.81±2.25	36.31±3.23	35.35±4.10	32.96±3.57	36.38±4.12	**43.04±4.50**
ratio = 0.2	25.42±5.49	28.81±2.41	30.46±2.59	34.00±3.06	29.58±2.68	33.58±2.78	**36.65±2.82**
ratio = 0.3	25.77±4.91	31.42±2.44	29.12±3.13	33.19±2.80	29.12±3.01	29.50±4.30	**33.58±3.16**

Table 5 NMI (%±std) of warpAR10P dataset with different ratio of sparse noise. The best results are highlighted in bold.

ACC	AllFea	LapScore	MCFS	NDFS	EUFS	MFFS	NSSLFS
ratio = 0.1	21.1±6.99	24.0±2.79	35.5±3.69	37.3±3.97	33.5±2.36	37.4±3.31	**44.7±4.07**
ratio = 0.2	21.8±7.06	27.6±3.82	27.1±2.80	34.9±2.95	29.8±2.21	31.2±3.50	**38.4±2.23**
ratio = 0.3	22.2±6.53	31.6±3.49	26.8±3.94	31.8±3.83	26.7±3.12	27.2±4.05	**33.1±4.48**

4.6 Discussion

From the results of the Table 2 and Table 3, we derive the following conclusions:

1. Compared with the baseline (all features), feature selection methods can use very few features to achieve better clustering performance so we obtain higher accuracy and computational efficiency simultaneously.
2. The nonlinear local geometric structure of data distribution introduced by LS, MCFS and NDFS plays an important role in clustering. And the non-negative constraint contained in NDFS and EUFS makes the scaled cluster indication matrix more accurate and reasonable. DNFS usually yields relatively better results by containing the two superiorities mentioned above. Although NSSLFS does not use the local structure explicitly, it still achieves higher ACC and NMI than NDFS by exploiting the underlying structure instead of learning pseudo labels. Both MFFS and NSSLFS leverage the low dimensional subspace learning scheme, while NSSLFS is superior to MFFS as the minimization of the regression model and $l_{2,1}$-norm regularization term enables W to have massive zero rows which is particularly suitable for feature selection.

We also have the following two observations from Table 4 and Table 5:

1. Compared to other approaches, better performance is obtained by the proposed robust method against corrupted dataset with sparse noise.
2. When the ratio of corruption varies from 0.1 to 0.3, our algorithm offers quite outstanding results which only use very few selected features.

Conclusion

The indicator matrix plays an important role in feature selection. In this paper, we introduce a novel approach to obtain the indicator matrix W with $l_{2,1}$-normal constraint based on the low-

dimensional sparse subspace learning, where the data may be compact and discriminative instead of original space. An efficient strategy is designed to optimize the $l_{2,1}$ -norm regularized minimization problem with non-negative constraint, which integrates alternate iteration into IALM. Extensive experiments demonstrate that NSSLFS is more efficient than some other existing algorithms in clustering problem with sparse noise.

Acknowledgements

This work is supported by the Natural Science Fund for Colleges and Universities in Jiangsu Province, China (Grant No 16KJB520012).

References

1. C.Y. Xu, Y. Zhang, Cell contour irregularity feature extraction methods based on linear geometric heat flow curve evolution, 2014, International Journal of Heat and Technology, vol. 6, no. 2-4, pp. 96-110.
2. H. Liu, H. Motoda, R. Setiono, and Z. Zhao, Feature Selection: An Ever Evolving Frontier in Data Mining, Journal of Machine Learning Research, 2010, The Fourth Workshop on Feature Selection in Data Mining 2010, vol. 10, pp. 4-13.
3. Isabelle Guyon， Steve Gunn， Masoud Nikravesh， Lofti A. Zadeh, 2008, Feature extraction: foundations and application, Springer.
4. L. Luo, J. Yang, J.J. Qian, Y. Tai, Nuclear-L1 Norm Joint Regression for Face Reconstruction and Recognition, 2015, Pattern Recognition vol. 48, no. 12, pp. 3811-3824.
5. R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2000, Wiley-Interscience, Hoboken, NJ, 2nd edition.
6. Robnik-Šikonja M, Kononenko I, Theoretical and empirical analysis of ReliefF and RReliefF, 2003, Machine learning, vol. 53, no. 1-2. pp. 23-69.
7. T. M. Cover and J. A. Thomas, Elements of Information Theory, 2006, Wiley-Interscience, 2nd edition.
8. F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, Trace ratio criterion for feature selection, 2008, Proceedings of the 23rd national Conference on Artificial Intelligence (AAAI), Chicago, Illinois, USA, July , 2008, vol. 2, pp. 671-676.

9. Z. Zhao, L. Wang, and H. Liu, Efficient spectral feature selection with minimum redundancy, 2010, Proceedings of 24th national Conference on Artificial Intelligence (AAAI), Georgia, USA, July, 2010, pp. 1-6.
10. F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization, 2010, Proceedings of 24th Neural Information Processing Systems (NIPS), Vancouver ,CANADA, December, 2010, pp. 1813-1821, 2010.
11. W. J. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components. Journal of the Royal Statistical Society,1987, Series C (Applied Statistics), vol. 36, no. 1, pp. 22-33.
12. X. He, D. Cai, and P. Niyogi, Laplacian score for feature selection, 2005, In Advances in Neural Information Processing Systems 18 (NIPS), Vancouver, CANADA, 2005, pp. 507-514.
13. Z. Zhao and H. Liu, Spectral feature selection for supervised and unsupervised learning, 2007, Proceedings of the 24th international conference on Machine learning (ICML),Corvallis, OR, USA, June ,2007, pp.1151- 1157.
14. D. Cai, C, Zhang, and X. He. Unsupervised feature selection for multi-cluster data, 2010, Proceedings of the 16th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining Washington DC, DC, USA, July, 2010, pp.333-342.
15. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, l2,1-norm regularized discriminative feature selection for unsupervised learning, 2011, In International Joint Conferences on Artificial Intelligence (IJCAI), New York City, NY, USA, 2011, pp. 1589-1594.
16. Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis, 2012, the 26th Conference on Artificial Intelligence (AAAI), Toronto, Ontario, Canada, 2012 pp.1026-1032.
17. L. Du, Z. Shen, X. Li, P. Zhou, and Y.-D. Shen, Local and global discriminative learning for unsupervised feature selection, 2013, The 13th IEEE International Conference on Data Mining (ICDM), Dallas, Texas, USA, December,2013, pp. 131-140.
18. C. Hou, F. Nie, X. Li, D. Yi, Y. Wu, Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection, 2014, IEEE Transactions on Cybernetics, vol.44, no. 6, pp: 793-804.
19. M. Qian and C. Zhai, Robust unsupervised feature selection, 2013, Proceedings of the 23rd international joint conference on Artificial Intelligence (IJCAI), Beijing, China, August, 2013, pp. 1621-1627.

20. W. Suhang, T. Jiliang, L. Huan, Embedded Unsupervised Feature Selection, 2015, Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI), Austin Texas, USA, January, 2015**,** pp.470-476, 2015.
21. J.R W. Shiping, Witold Pedrycz, Z. Qingxin, Z. William, Subspace learning for unsupervised feature selection via matrix factorization, 2015, Pattern Recognition vol. 48, no. 1, pp. 10-19.
22. X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu. Global and local structure preservation for feature selection, 2014, IEEE Transactions on Neural Networks and Learning Systems (NNLS), vol. 25, no. 6, pp. 1083-1095.
23. Z. Zhao, L. Wang, H. Liu, and J. Ye, On similarity preserving feature selection, 2013, IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 25, no. 3, pp. 619-632.
24. C. Ramirez, V. Kreinovich, M. Argaez. Why l1 is a Good Approximation to l0: A Geometric Explanation, 2013, Journal of Uncertain Systems vol.7, no.3, pp. 203-207.
25. L. Zhouchen, L. Risheng, and S. Zhixun, Linearized Alternating Direction Method with Adaptive Penalty for Low Rank Representation, In neural information processing systems (NIPS), Granada SPAIN, 2011, December, pp. 612–620.
26. R. J. Wang, X. H. Li, and Q. D. Wu, Optimizing process parameters for selective laser sintering based on neural network and genetic algorithm, 2009, International Journal of Heat and Technology, vol. 42, no. 11-12, pp. 1035-1042.
27. DD Lee, HS Seung, Learning the parts of objects by nonnegative matrix factorization, 1999, Nature vol. 401, no. 6755, pp.788–791.
28. Otsuka,Takayuki, Wakasu,Yutaka and Inoue,Tatsuo, A simple identification of transformation plastic behaviour and some data for heat treating materials, 2009, International Journal of Heat and Technology, vol. 24, no. 1-4, pp. 292-305.
29. D. Cai, H. Xiaofei, W. Xiaoyun, H. Jiawei, 2008, Non-negative Matrix Factorization on Manifold, 2008, Eighth IEEE International Conference on Data Mining (ICDM), Pisa, Italy, July, 2008, pp.63-72.
30. L. Huawei, W. Xindong, and Z. Shichao, Feature selection using hierarchical feature clustering, 2011, Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM), Glasgow, United Kingdom, October, pp. 974 - 984.
31. H. Liu and Z. Zhao. Manipulating Data and Dimensionality Reduction Methods: Feature Selection, 2009, In Encyclopedia of Complexity and Systems Science, Robert Meyers (Ed.), Springer.