ICIC Express Letters
Part B: Applications
Volume **8**, Number **11**, November **2017**

 ISSN 2185-2766
pp. 1481–1488

ROBUST CLUSTERING BASED ON K-MEANS WITH LOW-RANK REGULARIZATION

Xiaodong Wang[1,2], Rung-Ching Chen[2,*], Fei Yan[1] and Chen-Yen Peng[2]

[1]College of Computer and Information Engineering
Xiamen University of Technology
No. 600, Ligong Road, Jimei District, Xiamen 361024, P. R. China
{ xdwangjsj; fyan }@xmut.edu.cn

[2]Department of Information Management
Chaoyang University of Technology
No. 168, Jifeng East Rd., Wufeng District, Taichung 41349, Taiwan
*Corresponding author: rungching@gmail.com; davud8407@gmail.com

Received May 2017; accepted August 2017

Abstract. *Clustering is a widely applied technology in many fields. Due to the redundant features contained in the original data, traditional K-means suffers from the unstable performance. Although recent works try to perform dimension reduction and K-means together, they are still sensitive to the outliers with the l_2-norm based loss function. In this paper, we propose a robust K-means type clustering that jointly performs clustering and sparse learning. Different from the previous works, we use $l_{2,1}$-norm based loss function to improve the robustness of clustering and impose low-rank regularization to preserve the most representative features. Experimental results on six benchmark datasets demonstrate the effectiveness of the proposed algorithm.*
Keywords: Clustering, Sparse learning, K-means, $l_{2,1}$-norm, Low-rank regularization

1. **Introduction.** Clustering is a key technology in many fields, such as machine learning, and image processing. The main task of clustering is to categorize data samples into several groups, where samples in the same group are similar and those in different groups are dissimilar. K-means is one of the most popular clustering algorithms, which has been widely used in various applications for its sufficiency and simplicity [1]. As presented in previous works [2], traditional K-means clustering tends to fail when dealing with high-dimensional data, which usually contains redundant features and noises. To cope with this problem, researchers proposed to project the data onto a low-dimensional data space through dimension reduction such as Principal Component Analysis (PCA) and Locality Preserving Projections (LPP), and perform clustering subsequently. However, such projection might not be helpful for improving the clustering performance, due to the separation between subspace learning and clustering.

To overcome this problem, a simple and intuitive way is to integrate dimension reduction and clustering into a joint framework. Over the past two decades, many effective subspace clustering algorithms have been proposed [1,3]. For instance, Ye et al. proposed a discriminative K-means clustering which performs the Linear Discriminant Analysis (LDA) and K-means simultaneously [1]. Ding and Li coherently combined LDA and K-means clustering into a single framework, where the class labels are generated adaptively by K-means in a low-dimensional data space [3]. Nevertheless, LDA suffers from "small size sample" problem, making these LDA based subspace clustering algorithms hard to be applied to real-world applications, such as multimedia understanding, web page classification, and gene expression profiling. Li et al. noticed this problem and tried to address it with a new Maximum Margin Criterion (MMC) which works well when the

number of training samples is smaller than that of features [4]. Nie et al. transformed the objective function of LDA to a least square regression formulation and achieved better results on various kinds of applications [5]. Hou et al. proposed a discriminative embedded clustering framework by jointly combining PCA and K-means into a unified objective function. They also proved that the proposed framework has close relationships with several traditional subspace clustering algorithms [6]. However, the main drawback of the abovementioned methods is that they are sensitive to the outliers for the least-square based distance measurements.

In this paper, we propose a robust clustering by imposing a low-rank constraint on the objective function of K-means. The sparse learning and clustering can be simultaneously performed by an efficient iterative optimization approach. The major contributions of this paper are as follows. (1) An $l_{2,1}$-norm based K-means clustering with low-rank regularization is proposed. (2) The algorithm is able to find the most representative features by performing sparse learning and clustering simultaneously and effectively.

The remainder of this paper is organized as follows. Section 2 shows the related works and Section 3 prespents our proposed method and experiments are given on Section 4. Finally, the conclusion and future works are listed in Section 5.

2. Notations and Related Works.

In this section, we briefly review the related works. This paper is closely related to K-means and $l_{2,1}$-norm based sparse learning.

2.1. Notations.

Suppose that matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in R^{d \times n}$ denotes the input dataset, where d is the number of features and n is the number of data samples. $\|\cdot\|_F$ denotes the Frobenius norm. $\|\cdot\|_*$ means the nuclear norm. Let $F = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n]^T \in R^{n \times c}$ be the cluster indicator matrix and c is the number of clusters, $\boldsymbol{f}_i \in R^{c \times 1}$ is cluster indicator vector of sample i, if data x_i belongs to class j, $F_{ij} = 1$; otherwise $F_{ij} = 0$. Throughout this paper, all the matrices are written as uppercase and vectors are written as bold lowercase.

2.2. Traditional K-means.

Traditional K-means aims to assign the data into different groups with several random initialized centroids, and iteratively update the centroids using data in new clusters. Such assignment and update step are repeated until further refinement can no longer improve the model. Concretely, we can formulate the K-means objective function as follows:

$$\min_F \sum_{k=1}^{c} \sum_{i \in \mathcal{C}_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)^T (\boldsymbol{x}_i - \boldsymbol{m}_k) \tag{1}$$

where $\boldsymbol{m}_k$ is the centroid of cluster $\mathcal{C}_k$.

Denote $G = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_c] \in R^{d \times c}$ as the centroid matrix and recall the definition of F. Problem (1) can be reformulated as:

$$\min_F \sum_{i=1}^{n} (\boldsymbol{x}_i - G\boldsymbol{f}_i)^T (\boldsymbol{x}_i - G\boldsymbol{f}_i) = \min_F \left\| X^T - FG^T \right\|_F^2 \tag{2}$$

2.3. $l_{2,1}$-norm based sparse learning.

As previous works demonstrated, l_2-norm based distance measurement or loss function are sensitive to outliers. To fix this problem, Nie et al. imposed an $l_{2,1}$-norm based regularization technique into their feature learning framework [7]. Other $l_{2,1}$-norm based feature learning algorithms can be found in [8,9]. Chang et al. transformed the standard PCA formulation into a low-rank based linear regression optimization problem and proposed a robust PCA by using the $l_{2,1}$-norm based loss function instead of the l_2-norm based one [10]. Recently, some researchers also incorporated the $l_{2,1}$-norm regularization into clustering [11-13]. For example, Chang et al. proposed a spectral clustering algorithm, where a shrinking strategy based on $l_{2,1}$-norm was utilized

[11]. Cai et al. [12] and Du et al. [13] developed a multi-view and a multi-modal K-means clustering respectively, both of which imposed the $l_{2,1}$-norm on their objective function.

Given an arbitrary matrix $M \in R^{p\times q}$, its $l_{2,1}$-norm can be defined as:

$$\|M\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{q} M_{ij}^2} \tag{3}$$

From Equation (3) we can find that the $l_{2,1}$-norm penalizes the square root of square sums of each row in M, resulting in a group sparsity on M and an increasing robustness to the outliers.

3. The Proposed Method.

3.1. The problem formulation.

As we mentioned before, traditional K-means suffers from 1) it is sensitive to outliers as the l_2-norm based loss function, and 2) it often performs on the original data with high dimensionality, which degrades its clustering performance. Although dimension reduction algorithms can be applied beforehand, the separation between dimension reduction and clustering might be not helpful for improving the clustering performance. In this section, we will deal with these problems step by step. First, to handle the outlier, following the previous works [7], an $l_{2,1}$-norm based loss function is imposed to the objective function of K-means, which is formulated as follows:

$$\min_{W,F,G} \left\|X^T - FG^T\right\|_{2,1} \tag{4}$$

Considering the difficulty of dealing with high-dimensional data in Equation (4), we impose a low-rank regularization term into Equation (4). Concretely, we aim to find a projection matrix $M \in R^{d\times d}$, which projects the data into a new feature space with sparsity. Thereafter, we propose the following objective function:

$$\min_{W,F,G} \left\|X^T W - FG^T\right\|_{2,1} + \gamma\|W\|_* \tag{5}$$

where γ is the regularization parameter.

3.2. Optimization.

Since the $l_{2,1}$-norm in problem (5) is non-smooth, making it hard to be directly solved. In this section, we propose an alternative approach to optimize it. Concretely, we optimize some of the variables by keeping the others constant. We first reformulate problem (5) as:

$$\min_{W,F,G} Tr\left(\left(X^T W - FG^T\right)^T D_e \left(X^T W - FG^T\right)\right) + \gamma Tr\left(W^T D_w W\right) \tag{6}$$

Let $E = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n] = X^T W - FG^T$ and define D_e and D_w as:

$$D_e = \begin{bmatrix} \frac{1}{2\sqrt{\boldsymbol{e}_1'\boldsymbol{e}_1}} & & \\ & \ddots & \\ & & \frac{1}{2\sqrt{\boldsymbol{e}_n'\boldsymbol{e}_n}} \end{bmatrix},\ D_w = \frac{1}{2}\left(WW^T\right)^{-\frac{1}{2}} \tag{7}$$

Step 1: Fixing W, G, D_e, D_w and optimizing F

When we fix W and G, and D_e, D_w, the objective function in Equation (6) becomes:

$$F_{i,j} = \begin{cases} 1, & j = \arg\min_k D_e^{ii} \left\|W^T \boldsymbol{x}_i - \boldsymbol{g}_k\right\|_2^2 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Certainly, the optimal solution of F is equal to the cluster indicator matrix assigned by the traditional K-means on projected data $X^T W$ with given cluster center.

Step 2: Fixing F, W, D_e, D_w and optimizing G

By fixing F and W, D_e and D_w, the objective function in Equation (6) equals:

$$\mathcal{L}(G) = Tr\left(\left(X^TW - FG^T\right)^T D_e \left(X^TW - FG^T\right)\right) \tag{9}$$

Taking the deviation of $\mathcal{L}(G)$ *w.r.t.* G, we have:

$$\begin{aligned}\frac{\partial \mathcal{L}(G)}{\partial G} &= \frac{\partial \left(Tr\left(W^TXD_eX^TW - 2W^TXD_eFG^T + GF^TD_eFG^T\right)\right)}{\partial G} \\ &= \frac{\partial \left(Tr\left(GF^TD_eFG^T - 2W^TXD_eFG^T\right)\right)}{\partial G} \\ &= GF^TD_eF - W^TXD_eF\end{aligned} \tag{10}$$

Let the above equations equal zero, and we arrive at:

$$G = W^TXD_eF\left(F^TD_eF\right)^{-1} \tag{11}$$

Step 3: Fixing F, G, D_e, D_w and optimizing W

By fixing F and G, D_e and D_w, the objective function in Equation (6) can be reformulated as:

$$\min_{W^TW=I} Tr\left(W^TXD_eX^TW - 2W^TXD_eFG^T\right) + \gamma Tr\left(W^TD_wW\right) \tag{12}$$

By setting the derivation of Equation (11) *w.r.t.* W to zero, we arrive at:

$$W = \left(XD_eX^T + \gamma D_w\right)^{-1} XD_eFG^T \tag{13}$$

Step 4: Fixing F, G, W and updating D_e, D_w with Equation (7)

According to the abovementioned optimization procedure, we propose an iterative approach to solving the problem in Equation (5). The iterative approach is illustrated in Algorithm 1.

Algorithm 1: The optimization algorithm for problem (5)

Input:
- The centered training data $X \in R^{d\times n}$;
- The parameters γ

Output:
- Optimal $W \in R^{d\times d}$, $F \in R^{n\times c}$, $G \in R^{d\times c}$;
- Set $t = 0$ and initialize $W_0 \in R^{d\times d}$ as an identity matrix;
- **repeat**
  - Compute F^t according to Equation (8);
  - Compute G^t according to Equation (10);
  - Compute W^t according to Equation (13);
  - Update D_e^t and D_w^t according to Equation (7);
  - $t = t + 1$;
- **until** Convergence
- Return W, F, G

4. **Experimental Results and Discussion.** To evaluate the effectiveness of the proposed algorithm, we applied it to several kinds of open benchmark datasets including three image datasets (AR_ImData, MSRA25, and Coil20) and three UCI datasets (Cars, Vehicle, and Wine). We compared the proposed algorithm with three closely related algorithms. A brief description of these compared algorithms is listed as follows.

KM: The traditional clustering algorithm, which performs clustering with all features preserved. It is used as the baseline method in this paper.

LPPKM: A two-stage subspace learning based on local manifold learning and K-means clustering. It uses locality preserving projections (LPP) to get low-dimensional features, following the cluster labels calculation via a standard K-means clustering subsequently.

LDAKM: A joint subspace clustering that combines LDA and K-means in a coherent way to adaptively select the most discriminative subspace.

4.1. **Experimental setup.** In our experiments, all the parameters (if any) of our algorithm and the compared algorithms are tuned from $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6\}$. The evaluation procedures are repeated 5 times. Then the average results with standard deviation are reported. Following [14], Accuracy (ACC) is used as the evaluation metric. Denote $\boldsymbol{g}_i$ as the ground truth label of $\boldsymbol{x}_i$ and $\boldsymbol{q}_i$ as the clustering results. ACC is defined as:

$$ACC = \frac{\sum_{i=1}^{n} \delta(\boldsymbol{g}_i, map(\boldsymbol{q}_i))}{n} \tag{14}$$

where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$, otherwise. $map(.)$ is the best mapping function that matches the ground truth label and obtained cluster label using Kuhn-Munkres algorithms. A larger ACC indicates a better performance.

Table 1. A brief description of the selected datasets

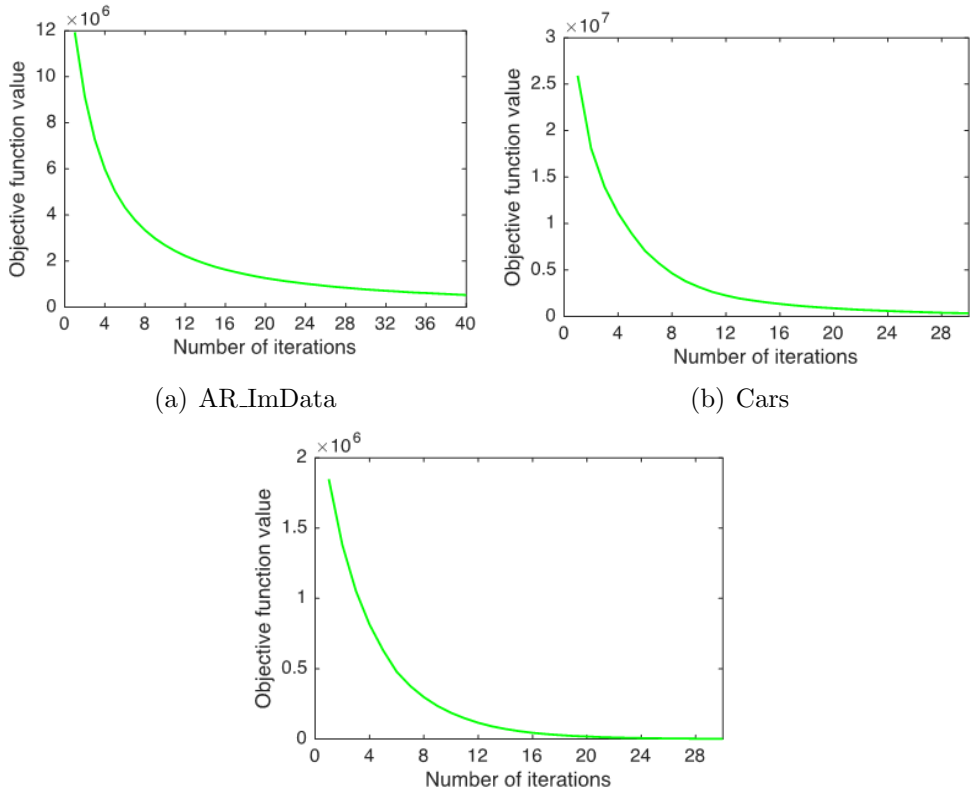
Dataset	Size (n)	Classes	Features
AR_ImData	840	120	768
MSRA25	1799	12	256
Coil20	1440	20	1024
Cars	392	3	8
Vehicle	846	4	18
Wine	178	3	13

4.2. **Experimental results for multimedia understanding.** Table 2 illustrates the results for clustering on six datasets. We can conclude the following. 1) LPPKM fails KM on most of the selected datasets. The main reason might be the separation of dimension reduction and clustering. However, as the dimension is considerably reduced, the speed of clustering will be improved. 2) LDA outperforms KM on these datasets except for Coil20, which demonstrates that jointly performing dimension reduction is beneficial to the subspace clustering. 3) Our algorithm achieves the best results compared with the other algorithms by imposing the robust $l_{2,1}$-norm based loss function and low-rank regularization into a joint objective function.

Table 2. Performance comparison in terms of ACC (±Standard Deviation (%))

	AR_ImData	Cars	Coil20	MSRA25	Vehicle	Wine
KM	30.90 ± 0.30	44.90 ± 0.00	47.47 ± 1.70	51.28 ± 0.96	44.15 ± 0.37	64.48 ± 0.77
LPPKM	24.31 ± 1.19	45.31 ± 0.14	47.27 ± 1.54	48.40 ± 5.81	44.00 ± 0.58	60.00 ± 6.03
LDAKM	24.88 ± 0.54	44.90 ± 0.00	43.27 ± 3.59	54.26 ± 3.35	44.30 ± 0.84	67.75 ± 5.52
OURS	$\mathbf{31.21 \pm 0.31}$	$\mathbf{46.68 \pm 0.00}$	$\mathbf{50.54 \pm 1.25}$	$\mathbf{59.54 \pm 1.71}$	$\mathbf{45.27 \pm 0.00}$	$\mathbf{70.79 \pm 0.40}$

4.3. **Convergence analysis.** Following previous works [12,15], it can be easily verified that the iterative optimization method in Algorithm 1 will converge to a local minimum. In this section, we conduct an experiment to study the convergence speed of our algorithm. The parameter γ of our algorithm is set to 1, which is the median of the tuned range. Figure 1 shows the results on three datasets, i.e., AR_ImData, Cars, and Wine. From the



(a) AR_ImData (b) Cars

(c) Wine

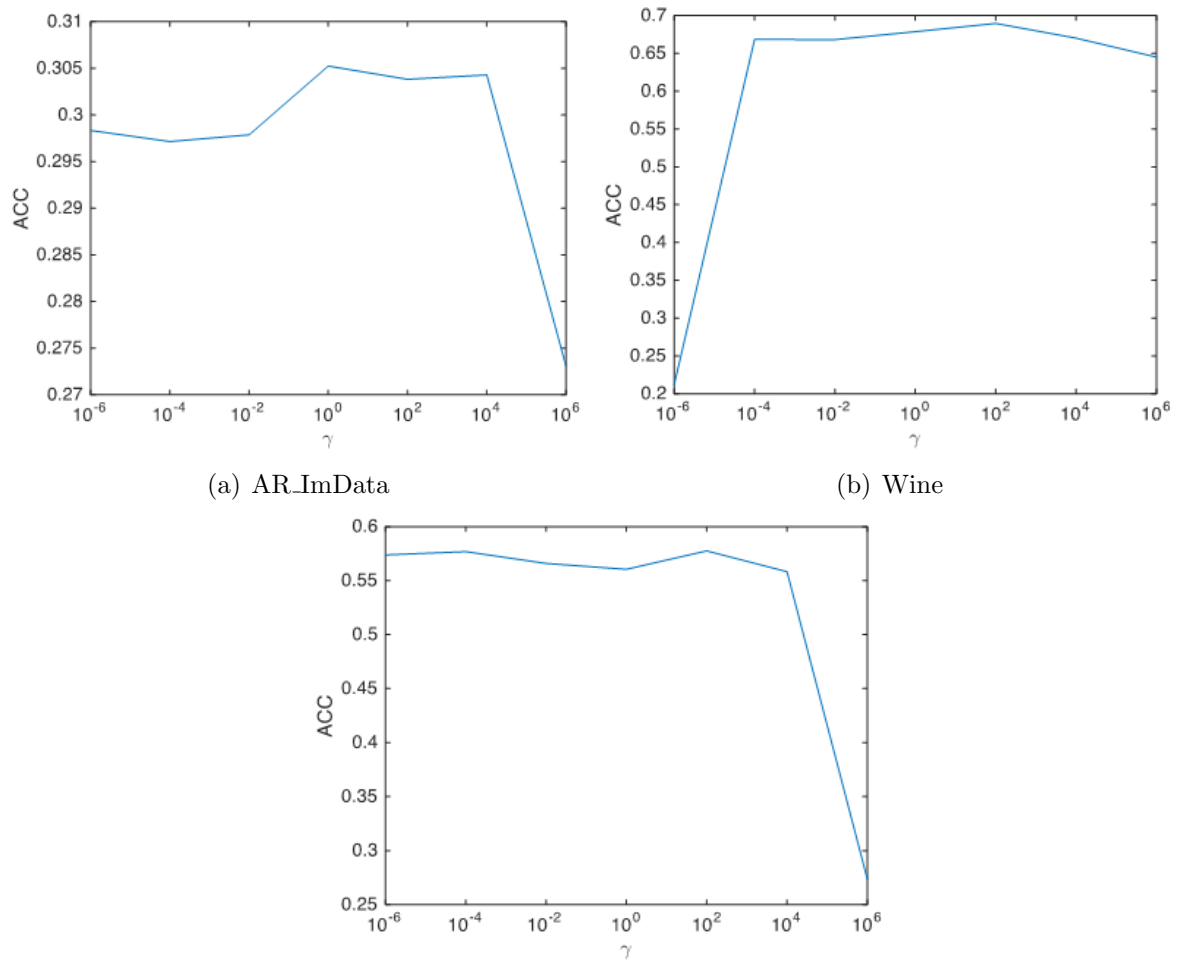
FIGURE 1. Convergence analysis of the proposed algorithm

results, we can observe that our algorithm converges fast, concretely, within 40 iterations for these three datasets.

4.4. **Parameter sensitivity analysis.** In this section, we study the parameter sensitivity of the proposed algorithm. Our algorithm has only one parameter γ. In Figure 2, we show the clustering performance influence of this parameter on different applications using AR_ImData, Wine, and MSRA25 datasets. For each dataset, we independently run our algorithm 5 times with the tuned parameter γ and report the average results. Note that γ is used to control the sparsity of transform matrix W. The larger γ is, the sparser W will be. From the results, we can observe the following. 1) A sparse transform matrix W is helpful for improving clustering performance. For example, the clustering accuracy increases when the parameter γ arises and is smaller than 1 for AR_ImData and Wine datasets. 2) Higher sparsity of W does not always reflect better results. Concretely, when $\gamma > 10^4$ for AR_ImData and MSRA25 datasets, the clustering accuracy significantly decreases.

5. **Conclusions.** In this paper, to overcome the shortcomings of traditional K-means clustering, we proposed an $l_{2,1}$-norm based K-means clustering with low-rank regularization. Compared with several K-means based subspace clustering algorithms, our algorithm is efficient and is able to find the most representative features by performing sparse learning and clustering simultaneously. As proposed algorithm is non-smooth, we solved it using an iterative approach. Empirical results on six benchmark datasets showed the effectiveness of the proposed algorithm. In the future work, we will combine our method with fuzzy system to different application domains.

Acknowledgements. This paper was supported by National Natural Science Foundation of China (Grant No. 61502405), National Natural Science Foundation of Fujian



(a) AR_ImData (b) Wine

(c) MSRA25

FIGURE 2. Performance variance $w.r.t.\ \gamma$ on three datasets

Province, China (No. 2016J01324, No. 2017J01511), Scientific Research Fund of Fujian Provincial Education Department (No. JA15385, No. JAS160368), and Ministry of Science and Technology, Taiwan, (Grant Nos. MOST-104-2221-E-324-019-MY2, MOST-103-2632-E-324-001-MY3).

REFERENCES

[1] J. Ye, Z. Zhao and M. Wu, Discriminative K-means for clustering, *Advances in Neural Information Processing Systems*, pp.1649-1656, 2008.

[2] C. Ding and X. He, K-means clustering via principal component analysis, *Proc. of the 21st International Conference on Machine Learning*, pp.225-232, 2004.

[3] C. Ding and T. Li, Adaptive dimension reduction using discriminant analysis and K-means clustering, *Proc. of the 24th International Conference on Machine Learning*, vol.10, no.1, pp.521-528, 2007.

[4] H. L. H. Li, T. J. T. Jiang and K. Z. K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks*, vol.17, no.1, pp.157-165, 2006.

[5] F. Nie, S. Xiang, Y. Liu, C. Hou and C. Zhang, Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction, *Pattern Recognition Letters*, vol.33, no.5, pp.485-491, 2012.

[6] C. Hou, F. Nie, D. Yi and D. Tao, Discriminative embedded clustering: A framework for grouping high-dimensional data, *IEEE Trans. Neural Networks and Learning Systems*, vol.26, no.6, pp.1287-1299, 2015.

[7] F. Nie, H. Huang, X. Cai and C. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, *Advances in Neural Information Processing Systems*, vol.23, pp.1813-1821, 2010.

[8] X. Wang, X. Zhang, Z. Zeng, Q. Wu and J. Zhang, Unsupervised spectral feature selection with l_1-norm graph, *Neurocomputing*, vol.200, pp.47-54, 2016.

[9] X. Wang, R.-C. Chen, F. Yan and Z. Zeng, Semi-supervised feature selection with exploiting shared information among multiple tasks, *Journal of Visual Communication & Image Representation*, vol.41, pp.272-280, 2016.

[10] X. Chang, F. Nie, Y. I. Yang, C. Zhang and H. Huang, Convex sparse PCA for unsupervised feature learning, *ACM Trans. Knowledge Discovery from Data*, vol.11, no.1, pp.1-16, 2016.

[11] X. Chang, F. Nie, Z. Ma, Y. Yang and X. Zhou, A convex formulation for spectral shrunk clustering, *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pp.2532-2538, 2015.

[12] X. Cai, F. Nie and H. Huang, Multi-view K-means clustering on big data, *The 23rd International Joint Conference on Artificial Intelligence*, pp.2598-2604, 2013.

[13] L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, W. Wang and Y. D. Shen, Robust multiple kernel K-means using $l_{2,1}$-norm, *IJCAI International Conference on Artificial Intelligence*, pp.3476-3482, 2015.

[14] X. Chang, F. Nie, Y. Yang and H. Huang, A convex formulation for semi-supervised multi-label feature selection, *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pp.1171-1177, 2014.

[15] F. Nie, X. Wang and H. Huang, Clustering and projected clustering with adaptive neighbors, *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.977-986, 2014.