UNSUPERVISED FEATURE SELECTION BASED ON RECONSTRUCTION ERROR MINIMIZATION

Sheng Yang, Rui Zhang, Feiping Nie , and 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

In this paper, we propose a novel unsupervised feature selection method, which is to minimize the data reconstruction error between each sample and a linear combination of its neighbors. Different from the conventional reconstruction-based feature selection method, we impose a nonnegative orthogonal constraint on the reconstruction weight matrix, so that an ideal neighbor assignment is adaptively captured. To enhance the robustness of the residual term and select the most valuable features, $\ell_{2,1}$-norm is applied to both reconstruction error term and feature selection matrix. At last, we derive an iterative algorithm to effectively solve the proposed objective function, and perform extensive experiments on four benchmark datasets to validate the effectiveness of the proposed method.

Index Terms— data reconstruction error, nonnegative orthogonal constraint, robustness, feature selection.

1. INTRODUCTION

With the development of technology, we can easily get a mass of data. However, the obtained data are often high-dimensional, contain many noise features, and could not be used directly. To select the most valuable features and accelerate data processing, feature selection has attracted much attention in recent years, and is widely researched in many different domains such as selecting the disease genes in medical research [1], image processing in computer vision [2], and data extraction in machine learning [3]. According to using label information or not, feature selection methods are classified into three different types: namely supervised feature selection [4], semi-supervised feature selection [5] and unsupervised feature selection [6]. Without using any label information, unsupervised feature selection becomes more difficult and challenging, but it is much more useful in practice and can spare a large amount of human-labor.

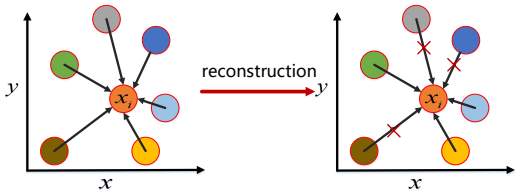


Fig. 1: Sample $\mathbf{x}_i$ is reconstructed by a linear combination of all other samples (left), and sample $\mathbf{x}_i$ can be automatically reconstructed by only a few important neighbors (right)

In recent years, data reconstruction error has become a new criterion for feature selection. For example, zhao et al. [7] proposes a graph regularized feature selection with data reconstruction, where the selected features not only can preserve the local structure of the original data via graph regularization, but also can reconstruct each data point via a linear combination of its neighbors. In [8], it proposes a framework for unsupervised feature selection, which embeds the reconstruction function learning process into feature selection. This method is a greedy search way to select the features, and the finally selected features may not be optimal.

In this paper, we propose a novel unsupervised feature selection method based on *reconstruction error minimization* (REM-FS). Different from other reconstruction-based feature selection method, the proposed method can perform the learning of reconstruction error function (using a few ideal neighbors to reconstruct each sample without any additional parameters) and feature selection simultaneously. To enhance the robustness of reconstruction error term and pick out the discriminative features, we apply the $\ell_{2,1}$-norm to both reconstruction error term and feature selection matrix (i.e. projection matrix). More importantly, a nonnegative orthogonal constraint is imposed on the reconstruction weight matrix, such that each sample is reconstructed by only a few ideal neighbors as the right part of Figure 1, rather than being reconstructed by a linear combination of all the samples as the left part of Figure 1.

Feiping Nie, Sheng Yang and Rui Zhang are with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China (email: feipingnie@gmail.com; 1637789668@qq.com;ruizhang8633@gmail.com).

Xuelong Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (xuelong li@opt.ac.cn).

2. THE PROPOSED ROBUST RECONSTRUCTION MODEL

Given data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, d is the number of features, and n is the number of samples. If each sample $\mathbf{x}_i$ is reconstructed by a linear combination of other samples, we have the following reconstruction model:

$$\min_{\mathbf{V}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} \mathbf{v}_{ij}\mathbf{x}_j \right\|_2^2 \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{n\times n}$ is the reconstruction weight matrix, which measures the contribution of j-th sample to the reconstruction of i-th sample. Besides, to guarantee the probability distribution, each row of matrix $\mathbf{V}$ is constrained with $\sum_j \mathbf{v}_{ij} = 1$.

In practice, the obtained data usually contain many noises. However, the model in problem (1) is sensitive to the large data outliers, because of the square of reconstruction error. If there is a large deviated value, it will be dominant and severely decrease the performance of the model. To enhance the robustness of the model in (1), it is rewritten as

$$\min_{\mathbf{V}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} \mathbf{v}_{ij}\mathbf{x}_j \right\|_2 \tag{2}$$

where the square of reconstruction error is removed, and the model in (2) becomes robust to the noises. Furthermore, there is an implicit weight defined in problem (2), and we can rewrite it as

$$\min_{\mathbf{V}} \sum_{i=1}^{n} d_i \left\| \mathbf{x}_i - \sum_{j=1}^{n} \mathbf{v}_{ij}\mathbf{x}_j \right\|_2^2 \tag{3}$$

where $d_i = 1 \Big/ 2\left\| \mathbf{x}_i - \sum_{j=1}^{n} \mathbf{v}_{ij}\mathbf{x}_j \right\|_2$ is an adaptive weight to measure the importance of data reconstruction. That is to say, if the data reconstruction error is large, d_i will be small, and if the data reconstruction error is small, d_i will be large. Next, for brevity, we rewrite problem (2) into the matrix form as

$$\min_{\mathbf{V}^T\mathbf{1}_n=\mathbf{1}_n, \mathbf{V}\geq 0} \|\mathbf{X} - \mathbf{X}\mathbf{V}\|_{2,1} \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{n\times n}$ is constrained with $\mathbf{V}^T\mathbf{1}_n = \mathbf{1}_n$ (the sum of each column is one), and $\mathbf{V} \geq 0$ (every element is guaranteed to be nonnegative). This constraint for matrix $\mathbf{V}$ is the same as the constraint in problem (1).

Usually, a sample $\mathbf{x}_i$ is not necessary to be reconstructed by a linear combination of all other samples as the left part in Figure 1, and we really expect that a sample is reconstructed by only a few important neighbors as the right part in Figure 1. An intuitive idea is that we can select the k nearest neighbors to reconstruct each sample. If do in this way, the model will involve another variable k, which needs to tune manually. Here, we adopt an easy and elegant way to achieve this motivation without any additional parameter, and rewrite problem (4) as

$$\min_{\mathbf{V}^T\mathbf{V}=\mathbf{I}_n, \mathbf{V}\geq 0} \|\mathbf{X} - \mathbf{X}\mathbf{V}\|_{2,1} \tag{5}$$

where we use the nonnegative orthogonal constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ on the reconstruction weight matrix $\mathbf{V}$, rather than using the original constraint $\mathbf{V}^T\mathbf{1}_n = \mathbf{1}_n$. This nonnegative orthogonal constraint has the following merits: (1) the constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ is not involved with any other parameter. (2) the constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ indicates that the square sum of each column is added up to one, namely we have $\mathbf{v}_1^T\mathbf{v}_1 = 1, \mathbf{v}_2^T\mathbf{v}_2 = 1, ..., \mathbf{v}_n^T\mathbf{v}_n = 1$. In this way, the small values of matrix $\mathbf{V}$ will be forced to closely be zero, and the large values of matrix $\mathbf{V}$ (having major contribution to reconstruct the samples) will be focused.

3. RECONSTRUCTION-BASED MODEL FOR FEATURE SELECTION

For the purpose of performing feature selection, according to the reconstruction model in problem (5), we propose the following objective function for the proposed REM-FS method.

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_m, \mathbf{V}^T\mathbf{V}=\mathbf{I}_n, \mathbf{V}\geq 0} \left\|\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{V}\right\|_{2,1} + \lambda\|\mathbf{W}\|_{2,1} \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{d\times m}$ is the projection matrix to project high-dimensional data to a low subspace (real-world data are often high-dimensional, and we need to map them into the low dimensions), and $\mathbf{V} \in \mathbb{R}^{n\times n}$ is a nonnegative orthogonal matrix to measure the contribution of ideal neighbors to reconstruct each sample. The first term of problem (6) denotes the reconstruction error between each sample and a linear combination of its neighbors after projection. The second regularization term is to force the projection matrix $\mathbf{W}$ to be row sparse for feature selection (selecting the genuinely useful features as in [9]). $\lambda > 0$ is a regularization parameter to balance the first term and the second term.

4. OPTIMIZATION ALGORITHM

To effectively solve problem (6), motivated by reweighted method in [9], it is converted into the following problem

$$\begin{aligned} &\min \left\| \left(\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{V}\right)\mathbf{D}^{\frac{1}{2}} \right\|_F^2 + \lambda Tr\left(\mathbf{W}^T\mathbf{Q}\mathbf{W}\right) \\ &s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_m, \mathbf{V}^T\mathbf{V} = \mathbf{I}_n, \mathbf{V} \geq 0 \end{aligned} \tag{7}$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with i-th element as $\mathbf{D}_{ii} = \frac{1}{2\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}\mathbf{v}_i\|_2}$, and $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is also a diagonal matrix with i-th element as $\mathbf{Q}_{ii} = \frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2 + \varepsilon}}$, where ε is a very small constant to avoid the denominator being zero.

Solving problem (7) is still challenging, because it contains two variables $\mathbf{W}$ and $\mathbf{V}$ simultaneously. We will solve this problem by alternatively optimizing variables $\mathbf{W}$ and $\mathbf{V}$, respectively.

Fix V, update W: With $\mathbf{V}$ fixed, problem (7) becomes

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_m} Tr\left(\mathbf{W}^T\mathbf{A}\mathbf{W}\right)+\lambda Tr\left(\mathbf{W}^T\mathbf{Q}\mathbf{W}\right) \tag{8}$$

where $\mathbf{A}=\mathbf{X}\left(\mathbf{I}_n-\mathbf{V}\right)\mathbf{D}(\mathbf{I}_n-\mathbf{V})^T\mathbf{X}^T$ and by using the property of trace, it is written as

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_m} Tr\left(\mathbf{W}^T\left(\mathbf{A}+\lambda\mathbf{Q}\right)\mathbf{W}\right) \tag{9}$$

Problem (9) has already been solved in spectral clustering [10], and we can know that the optimal solution of $\mathbf{W}$ is formed by the m eigenvectors of $(\mathbf{A}+\lambda\mathbf{Q})$ corresponding to m smallest eigenvalues.

Fix W, update V: When $\mathbf{W}$ is fixed, problem (7) becomes

$$\begin{aligned}&\min Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{V}\mathbf{D}\mathbf{V}^T\right)-2Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{V}^T\right)\\&s.t.\ \ \mathbf{V}^\mathbf{T}\mathbf{V}=\mathbf{I}_n,\mathbf{V}\geq 0\end{aligned} \tag{10}$$

However, this problem is very difficult to solve directly, because it involves the nonnegative orthogonal constraint. To tackle this problem, it is relaxed into the following form

$$\begin{aligned}\min_{\mathbf{V}\geq 0}\ &Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{V}\mathbf{D}\mathbf{V}^T\right)-2Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{V}^T\right)\\&+\tfrac{\gamma}{2}\left\|\mathbf{V}^T\mathbf{V}-\mathbf{I}_n\right\|_F^2\end{aligned} \tag{11}$$

where γ is a parameter to control the orthogonality. When $\gamma\to\infty$, the orthogonality will be satisfied. The Lagrangian function of problem (11) is as follows

$$\begin{aligned}\min_{\mathbf{V}}\ &Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{V}\mathbf{D}\mathbf{V}^T\right)-2Tr\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{V}^T\right)\\&+\tfrac{\gamma}{2}\left\|\mathbf{V}^T\mathbf{V}-\mathbf{I}_n\right\|_F^2+Tr\left(\mathbf{\Lambda}\mathbf{V}^T\right)\end{aligned} \tag{12}$$

where $\mathbf{\Lambda}$ is the Lagrangian multiplier, and problem (12) is now free from any constraint. Taking the derivative of problem (12) with respect to $\mathbf{V}$ and using Karush-Kunh-Tucker (KKT) condition $\mathbf{\Lambda}\circ\mathbf{V}$=0, we get the following updating rule

$$\mathbf{V}_{ij}=\mathbf{V}_{ij}\frac{(\gamma\mathbf{V})_{ij}}{\left(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{V}\mathbf{D}-\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{D}+\gamma\mathbf{V}\mathbf{V}^T\mathbf{V}\right)_{ij}} \tag{13}$$

Based on the above analysis, we summarize the whole procedure for solving problem (6) in Algorithm 1.

Algorithm 1 Algorithm for the proposed REM-FS method.

Input: Data matrix $\mathbf{X}\in\mathbb{R}^{d\times n}$, the parameter λ, a large enough number γ and the selected feature number t.
Initialize: The matrix $\mathbf{W}\in\mathbb{R}^{d\times m}$, $\mathbf{V}\in\mathbb{R}^{n\times n}$, diagonal matrices $\mathbf{D}\in\mathbb{R}^{n\times n}$ and $\mathbf{Q}\in\mathbb{R}^{d\times d}$.

while not converge **do**
1. Update $\mathbf{W}$ by solving problem (9), which is formed by the m eigenvectors of $(\mathbf{A}+\lambda\mathbf{Q})$ corresponding to m smallest eigenvalues.
2. Update $\mathbf{V}$ via Eq. (13).
3. Update diagonal matrices $\mathbf{D}$ and $\mathbf{Q}$, respectively.

end while

Output: Obtain optimal matrix $\mathbf{W}$ and calculate each $\left\|\mathbf{w}^i\right\|_2$, $i=1,2,...d$, then sort in descending order and select top ranking t features.

5. EXPERIMENT

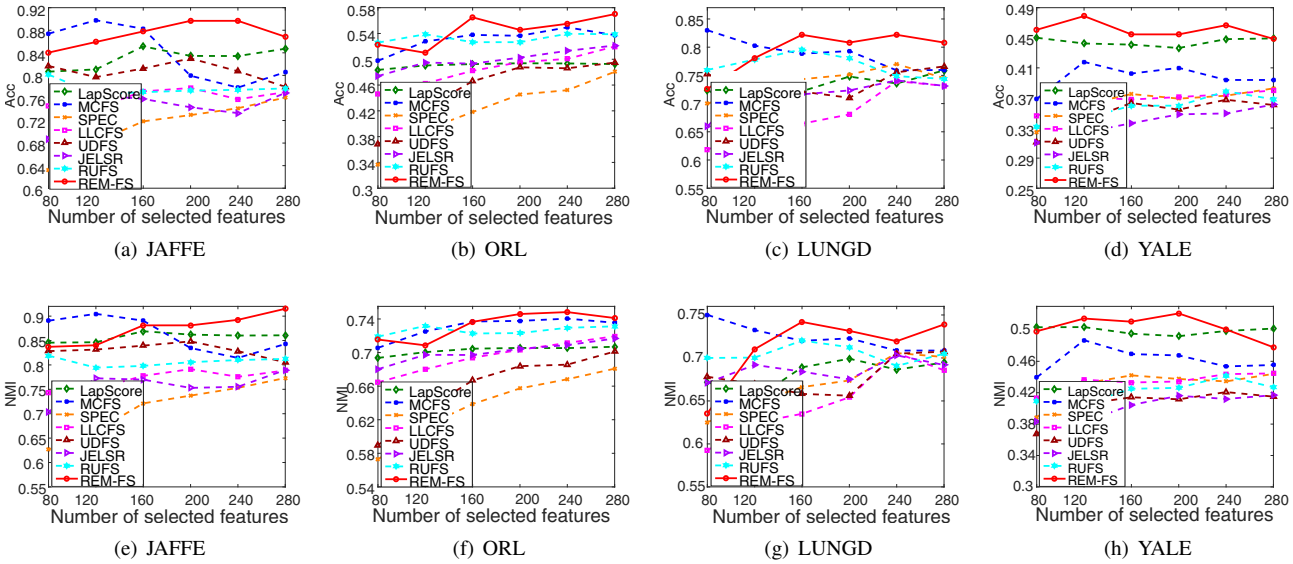
5.1. Benchmark Datasets and Competitors

In this section, to validate the performance of the proposed REM-FS method, we conduct extensive experiments on four benchmark datasets including: the Japanese Female Facial Expression (JAFFE) dataset [11], which contains 213 images of 7 facial expressions posed by 10 Japanese female models; the ORL dataset [12] has 40 different classes, and each class has 10 samples; the Lung_Discrete (LUNGD) dataset [13] is a gene expression microarray dataset, which has 73 samples belonging to 7 different classes; the last one is the Yale dataset [14], which contains 165 images selecting from 15 different persons, and every person has 11 different images.

In addition, seven state-of-the-art unsupervised feature selection methods are compared with the proposed method including: Laplacian Score (LapScore) [15], MCFS [16], SPEC [17] , LLCFS [18] , UDFS [19], JELSR [20] and RUFS [3].

5.2. Performance on Benchmark Datasets

As a convention in [3, 19], we use k-means to perform the clustering task, and record the average results (repeating 20 times). Clustering accuracy (Acc) and Normalized Mutual Information (NMI) [21] are as the basic evaluation metrics to measure the performance of different feature selection methods. As for regularization parameter λ in problem (6), the optimal parameter is selected at the candidate set $\left\{10^{-3},10^{-2},10^{-1},1,10,10^2,10^3\right\}$, and parameter γ is to control the orthogonality, which should be a large enough number and fixed as 10^4 in the experiment. The clustering accuracy and NMI with different number of features are shown in Figure 2, where the red line denotes the proposed REM-FS method. Generally speaking, the performance of proposed REM-FS first increases as the feature number size becomes large, then its performance falls down slightly. This is because more features can provide more information at the beginning, but if the feature number increases excessively, some noise features will be brought into the selected feature subset, and deteriorate the performance of the proposed method. To take a further analysis, we show the clustering accuracy and NMI on top 200 features in Table 1, where



(a) JAFFE (b) ORL (c) LUNGD (d) YALE

(e) JAFFE (f) ORL (g) LUNGD (h) YALE

Fig. 2: Clustering accuracy (top) and NMI (bottom) with different number of selected features

Table 1: Clustering accuracy (left) and NMI (right) on top 200 features.

Methods	JAFFE	ORL	LUNGD	YALE	JAFFE	ORL	LUNGD	YALE
LapScore	0.8345	0.4938	0.7479	0.4364	0.8621	0.7055	0.6991	0.4920
MCFS	0.7993	0.5364	0.7932	0.4100	0.8346	0.7377	0.7228	0.4675
SPEC	0.7296	0.4454	0.7514	0.3691	0.7371	0.6573	0.6738	0.4375
LLCFS	0.7777	0.4955	0.6808	0.3712	0.7914	0.7033	0.6542	0.4343
UDFS	0.8300	0.4875	0.7103	0.3545	0.8482	0.6838	0.6561	0.4116
JELSR	0.7434	0.5026	0.7233	0.3482	0.7529	0.7045	0.6751	0.4160
RUFS	0.7730	0.5265	0.7801	0.3597	0.8056	0.7232	0.7125	0.4262
REM-FS(our)	**0.8967**	**0.5455**	**0.8082**	**0.4545**	**0.8811**	**0.7459**	**0.7315**	**0.5210**

best results are in bold face and the second-best results are underlined. From Table 1, we conclude that the proposed REM-FS obtains the satisfactory performance and outperforms other seven compared methods. The distinct merit of REM-FS compared with other seven compared methods is that (1) a few ideal neighbors can be automatically captured by using nonnegative orthogonal constraint without involving any additional parameter. (2) In addition, the projection matrix $\mathbf{W}$ serves as a bridge to make the sample reconstruction and feature selection perform simultaneously. (3) By using $\ell_{2,1}$-norm, the reconstruction term becomes more robust, and projection matrix becomes row sparse such that the valuable features can be selected.

6. CONCLUSION

In this paper, we propose a novel unsupervised feature selection method (REM-FS) based on the reconstruction error minimization. We impose a nonnegative orthogonal constraint on the reconstruction weight matrix, such that having major contribution of neighbors will be focused and ignoring little contribution of neighbors (forcing small elements in matrix $\mathbf{V}$ to closely be zero) without involving any additional parameter. To reduce the impacts of large data outliers and select the valuable features, $\ell_{2,1}$-norm is applied for the residual term to enhance the robustness, and the projection matrix to be row sparse for feature selection. At last, we derive an iterative optimization algorithm to solve the objective function, and perform extensive experiments on four benchmark datasets to prove the effectiveness of the proposed method.

7. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China grant under numbers 61772427, 61751202 and 61761130079.

8. REFERENCES

[1] Mohua Banerjee, Sushmita Mitra, and Haider Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Transactions on Systems Man & Cybernetics Part C*, vol. 37, no. 4, pp. 622–632, 2007.

[2] Ming Ming Cheng, Yun Liu, Qibin Hou, Jiawang Bian, Philip Torr, Shi Min Hu, and Zhuowen Tu, *HFS: Hierarchical Feature Selection for Efficient Image Segmentation*, Springer International Publishing, 2016.

[3] Mingjie Qian and Chengxiang Zhai, "Robust unsupervised feature selection," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1621–1627.

[4] Le Song, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo, "Supervised feature selection via dependence estimation," in *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, 2007, pp. 823–830.

[5] Z. Xu, I King, M. R. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–47, 2010.

[6] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 24, no. 3, pp. 301–312, 2002.

[7] Zhou Zhao, Xiaofei He, Deng Cai, Lijun Zhang, Wilfred Ng, and Yueting Zhuang, "Graph regularized feature selection with data reconstruction," *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 3, pp. 689–700, 2016.

[8] Jundong Li, Jiliang Tang, and Huan Liu, "Reconstruction-based unsupervised feature selection: An embedded approach," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 2159–2165.

[9] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *International Conference on Neural Information Processing Systems*, 2010, pp. 1813–1821.

[10] Ulrike von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[11] Michael J. Lyons, Miyuki Kamachi, and Jiro Gyoba, "The japanese female facial expression (jaffe) database," 2017.

[12] Alan J. Chaney, Ian D. Wilson, and Andrew Hopper, "The design and implementation of a raid-3 multimedia file server," in *Proceedings of the 5th International Workshop on Network and Operating System Support for Digital Audio and Video*, London, UK, UK, 1995, NOSSDAV '95, pp. 306–317, Springer-Verlag.

[13] Worrawat Engchuan and Jonathan H. Chan, "Pathway activity transformation for multi-class classification of lung cancer datasets," *Neurocomputing*, vol. 165, pp. 81–89, 2015.

[14] A. S Georghiades, P. N Belhumeur, and D Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[15] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," in *International Conference on Neural Information Processing Systems*, 2005, pp. 507–514.

[16] Deng Cai, Chiyuan Zhang, and Xiaofei He, "Unsupervised feature selection for multi-cluster data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[17] Zheng Zhao and Huan Liu, "Spectral feature selection for supervised and unsupervised learning," in *Machine Learning, Proceedings of the Twenty-Fourth International Conference*, 2007, pp. 1151–1157.

[18] Hong Zeng and Yiu Ming Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 8, pp. 1532–47, 2011.

[19] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.

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

[21] Rui Zhang, Feiping Nie, and Xuelong Li, "Embedded clustering via robust orthogonal least square discriminant analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2332–2336.