UNSUPERVISED FEATURE SELECTION BY MANIFOLD REGULARIZED SELF-REPRESENTATION

Siqi Liang[⋆†] *Qian Xu*[⋆] *Pengfei Zhu*[⋆] *Qinghua Hu*[⋆] *Changqing Zhang*[⋆]

[⋆] School of Computer Science and Technology, Tianjin Univerity, Tianjin 300350, China
[†]School of Mathematical Sciences Nankai University, Tianjin 300071, China

ABSTRACT

Unsupervised feature selection has been proven to be an efficient technique in mitigating the curse of dimensionality. It helps to understand and analyze the prevalent high-dimensional unlabeled data. Recently, the self-similarity property of objects, which assumes that a feature can be represented by the linear combination of its relevant features, has been successfully used in unsupervised feature selection. However, it does not take the geometry structure of the sample space into consideration. In this paper, we propose a novel algorithm termed manifold regularized self-representation(MRSR). To preserve the local spatial structure, we incorporate an effective manifold regularization into the objective function. An iterative reweighted least square (IRLS) algorithm is developed to solve the optimization problem and the convergence is proved. Extensive experimental results on several benchmark datasets validate the effectiveness of the proposed method.

Index Terms— Unsupervised feature selection, Self-representation, Manifold regularization, Group Sparsity

1. INTRODUCTION

In the past few years, high dimensional data emerge in many domains like computer vision, pattern recognition and biological study [1]. High dimensional data have brought many challenges to the traditional clustering and classification algorithms, such as high time complexity, great storage burden and over-fitting of the learning machine [2]. Feature selection becomes an effective method to alleviate the problem of high-dimensionality, in that it aims to select the most discriminant and representative features and remove the redundant and irrelevant features.

Much attention has been paid to feature selection these years. From the perspective of label availability, feature selection methods can be classified into unsupervised and supervised issues [3]. Supervised feature selection such as robust regression [4] and Fisher score [5] are usually able to effectively select important features with the help of label information. However, the goal for unsupervised feature selection is to find the feature subset that best discovers natural groupings of data. Hence, unsupervised feature selection tends to be much more challenging compared with supervised issue. According to the way of selecting features, feature selection methods can also be classified into filter, wrapper, and embedded methods. Filter methods perform feature selection by leveraging statistical properties of data [5]. Various metrics can be defined to evaluate the importance of feature. For wrapper algorithms, feature selection is wrapped in a learning algorithm and the classification performance of features is taken as the evaluation criterion [6]. Compared with filter and wrapper methods, embedded approaches perform feature selection during the process of model construction [7]. In this paper, we adopt the embedded feature selection algorithm because it avoids a searching process and therefore is more efficient.

Self-similarity property assumes that one part of an object is similar with other parts. For example, a given part of the sky can be similar with other parts in color, shape or texture. Hence, these similar parts can be utilized to reconstruct the given part. This assumption generally holds in nature and has been applied to many real-world applications. For instance, Buades et al. utilized the similarity of image patches for image denoising [8]. Motivated by the self-similarity property, self-representation based algorithms are proposed to solve many intractable problems such as subspace clustering [9] and active learning [10]. Afterwards, self-representation has been applied into feature selection and shown significant potential. It assumes that a feature can be represented by the linear combination of its relevant features. Zhu et al. proposed to address feature selection via self-representation and achieved comparable performance [11]. Liu et al. considered the representativeness and diversity properties of features simultaneously [12].

Recent studies indicate that group sparsity can be effectively applied in various machine learning and computer vision algorithms, such as feature selection and multi-task learning [13]. Tradition L_1-norm termed Lasso requires element sparsity in a vector, while Group Lasso can be seen as an extension of L_1-norm by requiring group sparsity in a matrix. Early usage of group sparsity comes from the field of signal processing. For the recovery of original signals, it

restricts the coefficients in each block are either all zero or all nonzero [14]. Particularly, by formulating feature selection as a linear regression [15] or data reconstruction model [7], group sparsity constraint is imposed on the representation coefficient matrix, thus significant features can be selected via feature ranking.

Manifold learning methods have become increasingly usable in machine learning. The central assumption is that the higher-dimensional data reside on or near a low-dimensional sub-manifold [16]. Inspired by the development of manifold learning, a more simpler and effective technique termed manifold regularization is widely used. It requires that similar samples in the original space should also stay close in the transformed space. By maintaining the similarity, manifold regularization contributes much to model learning [17] .

Although the self-representation property of features has been successfully applied to unsupervised feature selection, the manifold structure of the sample space cannot be ignored as well. The sample similarity in the raw feature space should be well kept in the reconstructed space. In this paper, we propose an algorithm that consider both the self-representation property and the manifold structure. The main contribution of this paper is summarized as:

- We propose a manifold regularized self-representation (MRSR) algorithm for unsupervised feature selection.
- An iterative reweighted least square (IRLS) algorithm is developed and the convergence can be theoretically guaranteed.
- Experiments on benchmark datasets validate the effectiveness of MRSR over the state-of-the-art algorithms.

2. MANIFOLD REGULARIZED SELF-REPRESENTATION

2.1. Notations

We denote $\mathbf{X} \in \mathbb{R}^{n\times m}$ as a data matrix where n is the total number of samples and m is the number of features. $\mathbf{X} = \{\mathbf{x}_1; \cdots; \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^m (1 \leq i \leq n)$ is the i^{th} sample. We take $\mathbf{f}_1, \ldots, \mathbf{f}_m$ as the m features where $\mathbf{f}_i \in \mathbb{R}^n (1 \leq i \leq m)$. The $L_{2,1}$-norm of a matrix was introduced in [18] as rotational invariant of L_1-norm. It is one of the most popular techniques to achieve group sparsity. For a matrix $\mathbf{M} = (m_{ij})$ whose i^{th} row, j^{th} column are denoted by $\mathbf{m}_i, \mathbf{m}^j$, respectively, the $L_{2,1}$-norm of it is defined as $||\mathbf{M}||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} ||\mathbf{m}_i||_2$

2.2. The proposed model: MRSR

To the best of our knowledge, RSR first introduced self-representation into the task of feature selection and achieved comparable performance [11]. In RSR, data matrix $\mathbf{X}$ is used as the response matrix, and each feature can be represented by all the features with different representation coefficients. However, RSR did not take the structure information of unlabeled data into consideration.

Motivated by the manifold learning, we further incorporate a manifold regularization term to preserve data similarity. According to the discussion above, now we have the following minimization problem:

$$\begin{aligned}\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} &||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1} + \lambda_0 tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{W}) \\ &+ \lambda_1 ||\mathbf{W}||_{2,1}\end{aligned} \tag{1}$$

Here, we use data matrix $\mathbf{X}$ as the response matrix. $\mathbf{W} = \{\mathbf{w}_1; \ldots; \mathbf{w}_m\}$ is the representation coefficients matrix and $\mathbf{w}_i \in \mathbb{R}^m (1 \leq i \leq m)$ is i^{th} row of $\mathbf{W}$. Here, $||\mathbf{w}_i||_2$ can be used as the feature weight since it reflects the importance of the i^{th} feature in representation. For instance, if a feature plays an important role in the representation of all features, then $||\mathbf{w}_i||_2$ must be with big value, and vice versa. The $\mathbf{L}_{2,1}$-norm constraint is imposed on the representation coefficient matrix and the most important features can be selected via a feature ranking with corresponding $\mathbf{w}_i$. $tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{W})$ is the manifold regularization term where $\mathbf{L}$ is the Laplacian matrix. In this paper, we employ the LRGA method to learn a robust Laplacian matrix [19]. The effective manifold regularization term maintains the similarity between the original features and the reconstructed features, thus guarantees the minimum representation residual. We call the above model Manifold Regularized Self-Representation (MRSR) for unsupervised feature selection.

2.3. Optimization and algorithms

The model discussed above is actually convex, but both the loss function and the regularization terms are non-smooth. In this section, we solve the optimization of MRSR using Iterative Reweighted Least-Squares (IRLS) algorithm.

Given the current estimation $\mathbf{W}^t$, we define the diagonal weighting matrice $\mathbf{G}_L^t, \mathbf{G}_R^t$ by $g_{L,i}^t = 1/2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2$, $g_{R,j}^t = 1/2||\mathbf{w}_j^t||_2$, and then $\mathbf{W}^{t+1}$ is updated by solving the following weighted least squares problem:

$$\begin{aligned}\mathbf{W}^{t+1} &= \arg\min_{\mathbf{W}} Q(\mathbf{W}|\mathbf{W}^t) \\ &= \arg\min_{\mathbf{W}} (tr(\mathbf{X} - \mathbf{X}\mathbf{W})^T\mathbf{G}_L^t(\mathbf{X} - \mathbf{X}\mathbf{W}) \\ &+ \lambda_0 tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{W}) + \lambda_1 tr(\mathbf{W}^T\mathbf{G}_R^t\mathbf{W}))\end{aligned} \tag{2}$$

The closed form solution of $\mathbf{W}_{t+1}$ can be obtained by

$$\mathbf{W}^{t+1} = (\mathbf{X}^T\mathbf{G}_L^t\mathbf{X} + \lambda_0\mathbf{X}^T\mathbf{L}\mathbf{X} + \lambda_1\mathbf{G}_R^t)^{-1}\mathbf{X}^T\mathbf{G}_L^t\mathbf{X} \tag{3}$$

In order to avoid the overflow error, a sufficiently small value ϵ is introduced by defining $g_{L,i}^t = 1/\max(2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2, \epsilon)$ and $g_{R,j}^t = 1/\max(2||\mathbf{w}_j^t||_2, \epsilon)$. The formal algorithm is stated in Algorithom 1. After obtaining the values of $||\mathbf{w}_j||_2$, features are ranked and then the top k features can be selected.

Algorithm 1 MRSR based unsupervised feature selection

Input:

Training data $\mathbf{X} \in \mathbb{R}^{n \times m}$

Parameters λ_0 and λ_1

1: Set $t = 0$ and initialize $\mathbf{G}_L^t$ and $\mathbf{G}_R^t$;
2: Compute the Laplacian matrix $\mathbf{L}$;
3: **repeat**
4: $\mathbf{W}^{t+1} = (\mathbf{X}^T\mathbf{G}_L^t\mathbf{X} + \lambda_0\mathbf{X}^T\mathbf{L}\mathbf{X} + \lambda_1\mathbf{G}_R^t)^{-1}\mathbf{X}^T\mathbf{G}_L^t\mathbf{X}$;
5: update $\mathbf{G}_L^{t+1}$ and $\mathbf{G}_R^{t+1}$, $t = t + 1$;
6: **until** Converge.
7: Calculate feature weights $v_i = ||\mathbf{w}_i||_2$

Output:

$v_i, i = 1, 2, \ldots, m$

2.4. Time complexity

In MRSR, we mainly need to update $\mathbf{W}$ during each iteration, whose computational complexity is $O(m^3 + n^2m)$, where n and m are the number of samples and features, respectively. Hence, the complexity of MRSR is $O(T(m^3 + n^2m))$, where T is the total number of iterations.

2.5. Convergence analysis

The objective function in Eq. (1) is convex but non-smooth. In this paper, we develop an iterative reweighted least squares algorithm to solve the optimization problem of MRSR. We verify that the proposed model can converge to the optimal solution. The details can be found in the supplemental materials.

3. EXPERIMENTS

In this section, we conduct experiments to verify the effectiveness of MRSR. We evaluate the performance in terms of clustering results(including ACC and NMI) and classification results on six real-world benchmark datasets. Five benchmark unsupervised feature selection algorithms are selected for comparison.

3.1. Datasets

In this paper, we use six real-world datasets for extensive experiments. There are 4 face image datasets(i.e., warpPIE10P, warpAR10P, pixraw10p, orlraws10P) and 2 microarray datasets(i.e., TOX-171 and CLL-SUB-111). Detailed information about the datasets is summarized in Table 1.

Datasets	Instances	Features	Classes	Domains
warpPIE10P	210	2420	10	Image, Face
warpAR10P	130	2400	10	Image, Face
pixraw10P	100	10000	10	Image, Face
orlraw10P	100	10304	10	Image, Face
TOX-171	171	5748	4	Microarray, Bio
CLL-SUB-111	111	11340	3	Microarray, Bio

Table 1. Dataset Description.

3.2. Comparison methods

MRSR is compared with the following representative unsupervised feature selection algorithms.

Laplacian Score: Laplacian Score [5] computes Laplacian score for each feature to evaluate its locality preserving power.

MCFS: Mutli-Cluster Feature Selection [20] selects features using spectral regression with L_1-norm regularization.

UDFS: Unsupervised Discriminative Feature Selection[21] selects features by exploiting local discriminative information and feature correlations simultaneously. Manifold structure is also considered in UDFS.

SPEC: Spectral Feature Selection[22] selects features using spectral clustering.

RSR: Regularized Self-Representation Model[11] selects features by assuming that a feature can be represented by a linear combination of other features.

3.3. Evaluation metrics

Following the existed evaluation practice for unsupervised feature selection, we evaluate the unsupervised feature selection algorithms from two perspectives: clustering performance and classification performance. Clustering accuracy (ACC) and Normalized Mutual Information (NMI) are used to measure the clustering result. The classification performance is evaluated by comparing the groundtruth with that learned by using classification algorithms.

3.4. Parameter setting

Following previous work, for Laplacian Score, MCFS, UDFS, SPEC, RSR and MRSR, we fix $k = 5$ for all the datasets to specify the neighborhood size. In MCFS, UDFS, RSR and MRSR, the regularization parameter needs to be chosen, while in Laplacian score and SPEC, the bandwidth parameter for Gaussian kernel needs to be chosen. To make the comparison fair, we tune the bandwith and two regularization parameters from $\{10^{-6}, 10^{-5}, 10^{-4}, \ldots, 10^5, 10^6\}$ and record the best result. For feature dimension, we set the number of features as $\{10, 20, 30, \ldots, 150\}$ and report the average results over different dimensions. The K-means clustering algorithm is performed on the selected features by different algorithms. The experiment is run for 20 times with different random initializations since the K-means clustering result varies with initialization. The average results are reported for all the comparing algorithms.

3.5. Results and discussion

The experiment results of different methods on six real-world datasets in terms of clustering accuracy, NMI and classification accuracy are listed in Table 2, 3 and 4, respectively. The best feature selection results are highlighted with boldface. MCFS and RSR select the features using a regression model while Laplacian Score and SPEC select features one after another. It turns out that MCFS and RSR outperform Laplacian Score and SPEC, which indicates the superiority of transforming the unsupervised feature selection problem to a regression problem. Moreover, UDFS selects features in batch and takes the correlation between different features into account. As a result, it achieves a better performance than Laplacian Score and SPEC which analyze features separately. This observation validates the conclusion that features should be analyzed jointly for feature selection.

According to the experiment results, MRSR achieves the best performance in terms of clustering accuracy, NMI and classification accuracy among all the competing methods. MRSR converts the feature selection problem into a regression problem, selects data features in batch and incorporates the local structure information of unlabeled examples. Therefore, MRSR incorporates the strengths of the baseline methods introduced above and has a good performance in the experiment. Compared with RSR, the performance is significantly improved, which verifies that it is necessary to preserve the locality of the sample space.

Dataset	Laplacian	MCFS	UDFS	SPEC	RSR	MRSR
warpPIE10P	20.6	42.1	**47.9**	36.1	43.4	41.4
warpAR10P	21.1	22.7	44.2	45.6	32.8	**44.8**
pixraw10P	50.7	**87.6**	69.2	48.1	63.9	85.1
orlraw10P	40.1	**78.4**	72.3	37.81	60.1	76.7
TOX-171	40.4	40.3	40.3	38.8	42.3	**50.0**
CLL-SUB-111	37.4	50.9	50.3	50.9	50.6	**54.4**
Average	35.1	53.7	54.0	42.9	48.9	**58.7**

Table 2. Clustering results (ACC) of different unsupervised feature selection methods.

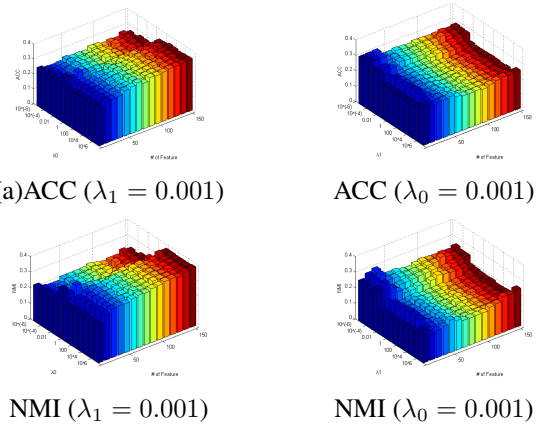
Dataset	Laplacian	MCFS	UDFS	SPEC	RSR	MRSR
warpPIE10P	20.6	**54.6**	52.3	39.7	50.3	51.6
warpAR10P	20.2	20.0	**48.6**	48.0	34.7	47.9
pixraw10P	67.1	91.4	77.4	54.7	72.4	**92.3**
orlraw10P	49.4	84.4	78.1	44.5	67.4	**84.7**
TOX-171	11.9	11.8	11.4	9.8	14.8	**27.9**
CLL-SUB-111	2.9	19.7	14.9	19.9	19.4	**22.9**
Average	28.7	47.0	47.1	36.1	43.2	**54.6**

Table 3. Clustering results (NMI) of different unsupervised feature selection methods.

Besides, we also investigate the sensitiveness of the parameters of MRSR. Due to the space limit, we only report the results on warpAR10P in Fig. 1. The experiment result indicates that our method is not very sensitive to the number of the features. Furthermore, the performance of MRSR is also not very sensitive to parameters λ_0 and λ_1 in a wide range in term of ACC. For NMI, the performance is a little sensitive to the parameters and there exists a fault phenomenon when the $\lambda_1 = 1$ and $\lambda_0 = 0.01$.

Dataset	Laplacian	MCFS	UDFS	SPEC	RSR	MRSR
warpPIE10P	87.0	**99.1**	96.2	86.5	94.8	**99.1**
warpAR10P	63.7	74.4	83.2	74.7	57.5	**86.6**
pixraw10P	70.1	97.5	97.2	49.3	87.3	**99.0**
orlraw10P	45.7	91.6	92.9	67.1	71.7	**96.9**
TOX-171	54.7	**65.5**	56.9	54.1	57.2	63.5
CLL-SUB-111	62.3	58.1	**81.9**	59.5	66.3	66.0
Average	63.9	81.0	84.7	65.2	72.5	**85.2**

Table 4. Classification rates(%) of different unsupervised feature selection methods.



(a)ACC ($\lambda_1 = 0.001$) ACC ($\lambda_0 = 0.001$)

NMI ($\lambda_1 = 0.001$) NMI ($\lambda_0 = 0.001$)

Fig. 1. ACC and NMI of MRSR with different λ_0, λ_1 and features on dataset warpAR10P.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a manifold regularized self-representation (MRSR) model for unsupervised feature selection. The $L_{2,1}$-norm is used to measure the self-representation residual to alleviate the impact of the outliers. The representation coefficients are also regularized by the $L_{2,1}$-norm sparsity to select effective features. To maintain the sample similarity of the raw space in the reconstructed space, a manifold regularization is imposed on reconstructed samples. As a result, the most representative features which can reconstruct other features and preserve locality are selected. The experiment results validated the effectiveness of MRSR in terms of both the clustering and classification performances. In the future, we will extend MRSR tasks to multi-view or classification problems.

5. REFERENCES

[1] Jiliang Tang and Huan Liu, "Unsupervised feature selection for linked social media data," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 904–912.

[2] Isabelle Guyon and Andr Elisseeff, "An introduction to variable feature selection," vol. 3, pp. 1157–1182, 2003.

[3] Jennifer G Dy and Carla E Brodley, "Feature selection for unsupervised learning," *Journal of machine learning research*, vol. 5, no. Aug, pp. 845–889, 2004.

[4] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," *Advances in Neural Information Processing Systems*, pp. 1813–1821, 2010.

[5] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," in *NIPS*, 2005, vol. 186, p. 189.

[6] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[7] Suhang Wang, Jiliang Tang, and Huan Liu, "Embedded unsupervised feature selection," in *AAAI*, 2015.

[8] A Buades, B Coll, and J. M Morel, "A non-local algorithm for image denoising," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 60–65.

[9] E Elhamifar and R Vidal, "Sparse subspace clustering: algorithm, theory, and applications.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–81, 2013.

[10] F. Nie, H. Wang, H. Huang, and C. Ding, "Early active learning via robust representation and structured sparsity," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1572–1578.

[11] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon CK Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.

[12] Yanbei Liu, Kaihua Liu, Changqing Zhang, Jing Wang, and Xiao Wang, "unsupervised feature selection via diversity-induced self-representation," *Neurocomputing*, 2016.

[13] Jianhui Chen, Jiayu Zhou, and Jieping Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 42–50.

[14] Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2008.

[15] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou, "L2,1 -norm regularized discriminative feature selection for unsupervised learning," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.

[16] Ayyoob Jafari and Farshad Almasganj, "Using laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy," *Speech Communication*, vol. 52, no. 9, pp. 725–735, 2010.

[17] Xiaofei He, "Laplacian regularized d-optimal design for active learning and its application to image retrieval," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 254–263, 2010.

[18] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha, "R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288.

[19] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 175–184.

[20] 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.

[21] 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.

[22] 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.