ROBUST SUBSPACE CLUSTERING BY LEARNING AN OPTIMAL STRUCTURED BIPARTITE GRAPH VIA LOW-RANK REPRESENTATION

Wei Chang, Feiping Nie, Rong Wang 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

This paper addresses the subspace clustering problem based on low-rank representation. Combining with the idea of co-clustering, we proposed to learn an optimal structural bipartite graph. It's different with other classical subspace clustering methods which need spectral clustering as post-processing on the constructed graph to get the final result, our method can directly learn a structural graph with k connected components so that the different clusters are obtained easily. Furthermore, we introduce a regularization term of error matrix to our model which helps the proposed algorithm to be more effective to learn an optimal graph under the circumstances of various noise. Experimental results both on synthetic and benchmark datasets are presented to show the effectiveness and robustness of our model.

Index Terms— Subspace Clustering, Bipartite Graph, Low-Rank Representation, Laplacian Rank Constraint.

1. INTRODUCTION

Subspace segmentation plays a very important role in clustering problem as it applies in many research areas including machine learning [1], image compression [2, 3], computer vision, e.g. image/motion/video segmentation [4, 5, 6, 7], and system identification. At present, the graph based methods have a good development in solving the problem of subspace clustering. Sparse Subspace Clustering (SSC) [8] and Robust Subspace Clustering by Low-Rank Representation (LRR) [9] are two typical models which utilize the sparse representation and low-rank representation respectively to solve this problem and the results are very satisfactory. These graph based methods divide the subspace clustering task in two steps' processes: constructing graph and spectral clustering, which means the graph based methods must need spectral clustering as postprocessing to get the final results.

In this paper, according to the idea of co-clustering [10], we propose a novel representation based method to learn an optimal structured bipartite graph with k connect components. By Theorem 1, we constrain the graph with the rank of its Laplacian matrix. Unlike the traditional methods which need spectral clustering to get the final clustering results, our method can obtain the segmentation result directly through the learned bipartite graph without postprocessing. The proposed model is based on the low-rank representation so that it's better at capturing the global structure of the original data than sparse representation. Besides, we introduce the regularization term of error matrix in the proposed method which makes our model more robust to noise and easy to construct the graph. The experimental results verify the validity and high performance of the proposed model.

The rest of this paper is shown as follows: in section 2, we revisit the low-rank representation based method (LRR) and co-clustering. Section 3 proposes the novel low-rank based model with learning an optimal bipartite graph and an alternating iterative algorithm is introduced to optimize it. Section 4 gives the experimental results which demonstrate the effectiveness of proposed model. Finally, we conclude this paper in section 5.

2. LOW-RANK REPRESENTATION AND CO-CLUSTERING REVISITED

The low-rank representation based method (LRR) is proposed by Liu et al. [9]. In order to get the optimal adjacent matrix Z capturing the global structure of original data, LRR utilizes the nuclear norm [11] to solve the low-rank representation problem. The optimization objective function can be described as:

$$\min_{Z} \|Z\|_*, \ s.t. \ X = XZ. \tag{1}$$

LRR has a unique closed form solution $Z = VV^T$, in which V is the right singular matrix drawn from the SVD of X. Besides, it has been proved that the obtained solution Z satisfies the block diagonal property when subspaces are independent.

To have a better clustering performance, co-clustering proposes to utilize the duality information between features and samples to construct the bipartite graph. For a similarity graph S, the bipartite graph G is denoted as follows:

$$G = \begin{bmatrix} 0 & S \\ S^T & 0 \end{bmatrix}. \tag{2}$$

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

Based on bipartite spectral graph partitioning [12], the normalized cut on the graph G is equivalent to the trace norm minimization problem as follows:

$$\min_{F^TF=I} Tr(F^T\tilde{L}F), \tag{3}$$

where $\tilde{L}$ is the normalized Laplacian matrix and $\tilde{L} = I - D^{-\frac{1}{2}}GD^{-\frac{1}{2}}$. D is the diagonal degree matrix and the i-th diagonal element is denoted as $d_i = \sum_i s_{ij}$.

3. ROBUST SUBSPACE CLUSTERING WITH LEARNED STRUCTURED GRAPH

Problem (1) presents the low-rank representation based model. In general, after obtaining the optimal solution Z, the graph is constructed by $(|Z^T| + |Z|)/2$ and spectral clustering is utilized on this graph. In this work, combining with the idea of co-clustering, we want to learn an optimal bipartite graph with k connected components [10] which can avoid the postprocessing. So based on the coefficient matrix Z obtained by low-rank representation, we can learn a graph S from the matrix Z to construct the bipartite graph G defined in Eq.(2) which has k connected components. Hence, for a given data matrix $X \in R^{d\times n}$, the proposed model can be described as follows:

$$\begin{aligned} &\min_{Z,E,G\in\Omega} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 \\ &s.t. \quad X = XZ + E, S \geq 0, S'\mathbf{1} = \mathbf{1}. \end{aligned} \tag{4}$$

Here, $\mathbf{1} = (1, 1, ..., 1)^T$ and Ω represents the set of the graph G defined in Eq.(2), which has exact k connected components.

Based on the constraint conditions in problem (4), the bipartite graph is nonnegative. And the reference [13] gives an important property about the normalized Laplacian matrix $\tilde{L}_G = I - D_G^{\frac{1}{2}}GD_G^{\frac{1}{2}}$ associated with the graph G which can be seen in Theorem 1.

Theorem 1 *The multiplicity k of the eigenvalue 0 of the normalized Laplacian matrix $\tilde{L}_G$ is equal to the number of connected components in the graph associated with G.*

According to the Theorem 1, we know that if $rank(\tilde{L}_G) = N - k$, the constraint $G \in \Omega$ will be satisfied. Here, $N = 2n$. So the problem (4) can be rewritten as:

$$\begin{aligned} &\min_{Z,S} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 \\ &s.t. \quad X = XZ + E, S \geq 0, S'\mathbf{1} = \mathbf{1}, rank(\tilde{L}_G) = N - k. \end{aligned} \tag{5}$$

Assuming that $\sigma_i(\tilde{L}_G)$ is the i-th smallest eigenvalue of $\tilde{L}_G$. Because the Laplacian matrix $\tilde{L}_G$ is positive semi-definite, we have $\sigma_i(\tilde{L}_G) \geq 0$. Therefore, we can convert problem (5) to the following problem:

$$\begin{aligned} &\min_{Z,E,S} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 + \lambda_3 \sum_{i=1}^{k} \sigma_i \left(\tilde{L}_G\right) \\ &s.t. \quad X = XZ + E, S \geq 0, S'\mathbf{1} = \mathbf{1}. \end{aligned} \tag{6}$$

Cause $\sigma_i(\tilde{L}_G) \geq 0$ for each i, when λ_3 is large enough, the objective function (6) will let the last term $\sum_{i=1}^{k} \sigma_i \left(\tilde{L}_G\right)$ to be zero, which is equivalent to problem (5).

Based on the Ky Fan's Theorem [14], we have:

$$\min \sum_{i=1}^{k} \sigma_i \left(\tilde{L}_G\right) = \min_{F\in R^{N\times k}, F^TF=I} Tr\left(F^T\tilde{L}_GF\right). \tag{7}$$

Therefore, combining with problem (6), the final optimal problem can be described as follows:

$$\begin{aligned} &\min_{Z,E,S,F} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 + \lambda_3 tr(F^T\tilde{L}_GF) \\ &s.t. \quad X = XZ + E, S \geq 0, S'\mathbf{1} = \mathbf{1}, F^TF = I, F \in R^{N\times k}. \end{aligned} \tag{8}$$

In the next section, an alternating iteration based algorithm is proposed to address this model.

3.1. Optimization

For the objective function (8), there are four variables needed to be updated. When fixing the variables S and F, problem (8) can be further transformed into the following problem:

$$\begin{aligned} &\min_{Z,E,J} \|J\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 \\ &s.t. \quad X = XZ + E, Z = J. \end{aligned} \tag{9}$$

So we can get the Augmented Lagrange Multiplier problem of objective function (9) as follows:

$$\begin{aligned} \min_{Z,E,J,Y_1,Y_2} &\|J\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|S - Z\|_F^2 \\ &+ tr[Y_1^T(X - XZ - E)] + tr[Y_2^T(Z - J)] \\ &+ \frac{\mu}{2}(\|X - XZ - E\|_F^2 + \|Z - J\|), \end{aligned} \tag{10}$$

here, Y_1 and Y_2 are Lagrange multipliers and $\mu \geq 0$ is a penalty parameter. The above problem can be solved by exact ALM algorithm [15]. We will show the specific process next.

When updating the matrix J, problem (10) becomes:

$$\arg\min_J \frac{1}{\mu} \|J\|_* + \frac{1}{2}\|J - (Z + \frac{1}{\mu}Y_2)\|_F^2. \tag{11}$$

The reference [15] has proved that problem (11) has an analytical solution.

When updating the coefficient matrix Z, problem (10) can be further transformed into convex optimization problem. Hence, by taking the derivative of this convex problem and making it zero, we can get the update form of Z as follows:

$$\begin{aligned} Z =&[(1-\frac{2\lambda_2}{\mu})I + X^TX]^{-1}[X^TX - \frac{2\lambda_2}{\mu}S - X^TE \\ &+ J + \frac{1}{\mu}(X^TY_1 - Y_2)]. \end{aligned} \tag{12}$$

When updating the error matrix E, problem (10) can be rewritten as:

$$\arg\min_E \frac{\lambda_1}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (X - XZ + \frac{1}{\mu}Y_1)\|_F^2. \tag{13}$$

Lin et al. [9] have given the closed-form solution for this problem in Lemma 3.2.

After obtaining the two variables Z and E, we need to update the other matrices S and F. So the problem (8) is equivalent to the following problem:

$$\begin{aligned} &\min_{S,F} \|S - Z\|_F^2 + \lambda tr(F^T\tilde{L}_G F) \\ &s.t. \quad S \geq 0, S'\mathbf{1} = \mathbf{1}, F^TF = I, F \in R^{N\times k}, \end{aligned} \tag{14}$$

here, $\lambda = \lambda_3/\lambda_2$.

When fixing S, cause $\tilde{L}_G = I - D_G^{\frac{1}{2}}GD_G^{\frac{1}{2}}$, the problem for solving matrix F can be rewritten as the following form:

$$\max_{F^TF=I, F\in R^{N\times k}} tr(F^TD_G^{\frac{1}{2}}GD_G^{\frac{1}{2}}F) \tag{15}$$

The matrix F and degree matrix D_G can be rewritten as the following block forms:

$$F = \begin{bmatrix} U \\ V \end{bmatrix}, D_G = \begin{bmatrix} D_{G_u} & \\ & D_{G_v} \end{bmatrix}. \tag{16}$$

Due to the structure of bipartite graph G defined in Eq.(2), the problem (15) can be converted as:

$$\max_{U^TU+V^TV=I} Tr(U^TD_{G_u}^{-\frac{1}{2}}ZD_{G_v}^{-\frac{1}{2}}V). \tag{17}$$

Nie et al. [10] give the solution for this problem in Lemma 1, where the solution is $U = \frac{\sqrt{2}}{2}U_1, V = \frac{\sqrt{2}}{2}V_1$. Here, U_1, V_1 are the leading k left and right singular vectors of matrix $D_{G_u}^{-\frac{1}{2}}ZD_{G_v}^{-\frac{1}{2}}$ respectively.

When F is fixed, the problem (14) becomes:

$$\min_{S\geq 0, S'\mathbf{1}=\mathbf{1}} \|S - Z\|_F^2 + \lambda tr(F^T\tilde{L}_G F) \tag{18}$$

Based on the property of Laplacian matrix and the structure of G define in Eq.(2), we have the following equation:

$$tr(F^T\tilde{L}_G F) = \sum_{i=1}^{n}\sum_{j=1}^{m}\|\frac{f_i}{\sqrt{d_i}} - \frac{f_{j+n}}{\sqrt{d_{j+n}}}\|_2^2 s_{ij}. \tag{19}$$

Based on Eq.(19), we can transform the problem (18) into the following problem:

$$\min_{S\geq 0, S'\mathbf{1}=\mathbf{1}} \sum_{i=1}^{n}\sum_{j=1}^{m}(s_{ij}-z_{ij})^2 + \lambda\|\frac{f_i}{\sqrt{d_i}} - \frac{f_{j+n}}{\sqrt{d_{j+n}}}\|_2^2 s_{ij} \tag{20}$$

It's easy to find that problem (20) is independent between different i. Therefore, this problem is equivalent to optimize the following problem individually for each i. Denote the j-th element of column vector h_i as $h_{ij} = \|\frac{f_i}{\sqrt{d_i}} - \frac{f_{j+n}}{\sqrt{d_{j+n}}}\|_2^2$ and give the same definition for s_i and z_i. So for each i, the problem (20) can be rewritten as the vector form:

$$\min_{s_i\geq 0, s_i'\mathbf{1}=1} \|(s_i - (z_i - \frac{\lambda}{2}h_i)\|_2^2. \tag{21}$$

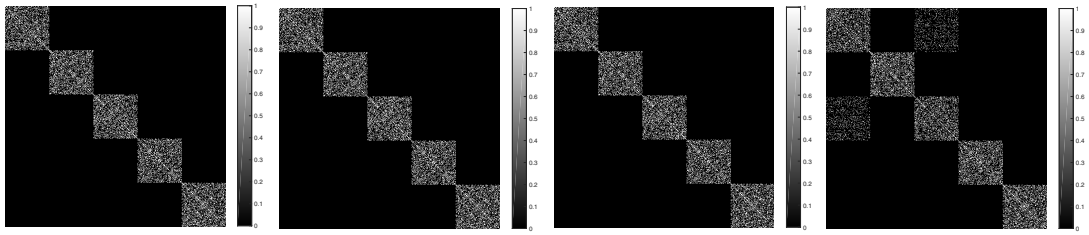
An efficient iterative algorithm is given to solve this problem by reference [16]. Hence, we conclude the algorithm to solve the problem (8) in Algorithm 1. In this algorithm, we can only update the c nearest similarities for each column in S. So we can reduce the complexity of updating S and F significantly. Besides, we only need the SVD on an $n \times n$ matrix in each iteration, which don't need to conduct eigen-decomposition on the $N \times N$ bipartite graph G. So the proposed Algorithm 1 is very efficient to solve the subspace clustering problem.

Algorithm 1 Algorithm to solve problem (8) (LOSBG)

Input: data matrix X, the cluster number k.
Output: the learned bipartite graph G and the cluster label.
Initialize: Randomly initialize the matrix S to satisfy the constraint condition in problem (8).
repeat
1. Fix others, update J by solving problem (11).
2. Fix others, update Z by formula (12).
3. Fix others, update E by solving problem (13),
4. Fix others, update S, for each column vector s_i, which is updated by solving the problem (21).
5. update F which can be solved by problem (17) based on the definition of F in formula (16).
6. update the multipliers Y_1, Y_2 and parameter μ.

until converge

For the proposed model (8), it has three parameters which makes the model seems very complexity. In fact, there is only one parameter need to be regulated. The graph S is determined by Z, so solving for Z cannot be affected by S, which means that the parameter λ_2 must be small enough in problem (9). Besides, when updating S, F in problem (14), we need the parameter $\lambda = \lambda_3/\lambda_2$ large enough. So based on the above analysis, we just let the parameter λ_2 small enough and set λ_3 be a constant, which makes our model to be affected only by one parameter λ_1. Besides, according to the analysis on LRR [9], our model is not sensitive to the parameter λ_1.



(a) noise = 0.6 (b) noise = 0.7 (c) noise = 0.8 (d) noise = 0.9

Fig. 1. The learned graphs after binarization, obtained by LOSBG conducting on synthetic dataset in different settings of noise. And the clustering results are 100%, 100%, 100%, 79.80% respectively.

Table 1. The clustering accuracy (%) and standard error (%) on the Hopkins 155.

Method	two motions				three motions				All			
	mean	median	min	std	mean	median	min	std	mean	median	min	std
RANSAC	94.88	98.63	54.33	9.38	81.51	86.18	52.55	14.26	91.86	97.22	52.55	12.01
LSA	95.29	99.56	51.59	10.91	90.57	95.86	46.59	14.35	94.22	99.33	46.59	11.89
SSC	92.93	99.29	53.46	12.79	80.45	88.27	47.34	15.69	90.11	97.45	47.34	14.43
LRR	97.32	**100.00**	53.53	7.13	88.29	93.50	58.20	14.06	95.28	**100.00**	53.53	9.86
LatLRR	97.89	**100.00**	55.35	7.71	96.26	**100.00**	64.09	8.58	97.52	**100.00**	55.35	7.92
BDR	98.74	**100.00**	51.52	6.46	**98.78**	99.79	**87.59**	**2.50**	**98.74**	**100.00**	51.52	5.80
LOSBG	**99.30**	**100.00**	**79.87**	**2.51**	95.97	99.60	72.19	6.80	98.55	**100.00**	**72.19**	**4.12**

4. EXPERIMENT

In this section, we will verify the validity and effectiveness of the proposed model (named LOSBG) on a synthetic datasets and Hopkins 155 Datasets.

Synthetic Data. First, we apply LOSBG to a high-dimensional toy dataset as a sanity check, which is drawn from an 200-dimensional Euclidean space. We randomly select five 50-dimensional subspaces from this space and take 100 samples from them respectively to form this toy dataset. Besides, in order to verify the robustness of LOSBG under the situation of noise, we add Gaussian noise to this dataset and set the portion of noise to be $r = 0.6, 0.7, 0.8, 0.9$ respectively.

We conduct LOSBG on this high-dimensional synthetic dataset with noise. Figure 1 shows the learned graph S after binarization, which constructs the bipartite graph G defined in Eq.(2). It can be seen that, when the portion of noise $r = 0.6, 0.7, 0.8$, our model can learn a perfect block diagonal structure. When $r = 0.9$, which means the data has been heavily contaminated, the Figure 1(d) shows that LOSBG also has the good clustering performance.

Motion Segmentation. Hopkins 155 Dataset is a motion dataset and provided by the Vision Lab of Johns Hopkins University [17]. This dataset is made up of 155 sequences and each sequence is one clustering task that need to be segmented into two or three motions. For comparison, we also list the results of RANSAC (Random Sample Consensus) [18], LSA (Local Subspace Analysis) [6], SSC, LRR, LatLRR [19] and BDR [20]. These methods are tested on the same datasets and the parameters are tune to the best.

Table 1 displays the comparison results of LOSBG with other methods. We notice that LOSBG is better than the algorithms which utilize spectral clustering as postprocessing like LRR, LatLRR etc. Besides, LOSBG achieves almost the best results among all comparison methods. So the effectiveness of LOSBG are verified in practical circumstance.

5. CONCLUDE

In this paper, we proposed a novel low-rank representation based method LOSBG which can learn an optimal structured bipartite graph to solve the subspace clustering problem. Unlike the traditional methods which transform the clustering problem in two steps: constructing graph and spectral clustering, LOSBG can get the segmentation result straightly from the learned bipartite graph. Besides, due to the regularization term of error matrix in our model, LOSBG is better at dealing with the situation under various noise. Experiment results on benchmark datasets demonstrates that our model has a better performance than other classical method. And it's easy to see that we can also introduce sparse constraint instead of low-rank representation to our algorithm framework.

6. REFERENCES

[1] Le Lu and René Vidal, "Combined central and subspace clustering for computer vision applications," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 593–600.

[2] Wei Hong, John Wright, Kun Huang, and Yi Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.

[3] Allen Y Yang, John Wright, Yi Ma, and S Shankar Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, 2008.

[4] Ken-ichi Kanatani, "Motion segmentation by subspace separation and model selection," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 586–591.

[5] René Vidal, Roberto Tron, and Richard Hartley, "Multiframe motion segmentation with missing data using powerfactorization and gpca," *International Journal of Computer Vision*, vol. 79, no. 1, pp. 85–105, 2008.

[6] Jingyu Yan and Marc Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *European conference on computer vision*. Springer, 2006, pp. 94–106.

[7] L Zelnik-Manor and M Irani, "Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 2, pp. II–287.

[8] Ehsan Elhamifar and Rene Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[9] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, "Robust recovery of subspace structures by low-rank representation," *arXiv preprint arXiv:1010.2955*, 2010.

[10] Feiping Nie, Xiaoqian Wang, Cheng Deng, and Heng Huang, "Learning a structured optimal bipartite graph for co-clustering," in *Advances in Neural Information Processing Systems*, 2017, pp. 4129–4138.

[11] Emmanuel J Candes and Yaniv Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[12] Inderjit S Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 269–274, 2001.

[13] Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann, "The laplacian spectrum of graphs," *Graph theory, combinatorics, and applications*, vol. 2, no. 871-898, pp. 12, 1991.

[14] Ky Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i," *Proceedings of the National Academy of Sciences*, vol. 35, no. 11, pp. 652–655, 1949.

[15] Zhouchen Lin, Minming Chen, and Yi Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[16] Jin Huang, Feiping Nie, and Heng Huang, "A new simplex sparse learning model to measure data similarity for clustering.," in *IJCAI*, 2015, pp. 3569–3575.

[17] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2790–2797.

[18] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[19] Guangcan Liu and Shuicheng Yan, "Latent low-rank representation for subspace segmentation and feature extraction," IEEE, 2011, pp. 1615–1622.

[20] Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan, "Subspace clustering by block diagonal representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.