Multi-view Subspace Clustering on Topological Manifold

Shudong Huang[1#], Hongjie Wu[1#], Yazhou Ren[2],
Ivor W. Tsang[3], Zenglin Xu[4], Wentao Feng[1*] and Jiancheng Lv[1*]

[1]College of Computer Science, Sichuan University, China
[2]School of Computer Science and Engineering, UESTC, China
[3]Centre for Frontier AI Research, A*STAR, Singapore
[4]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, China

{huangsd, Wtfeng2021, lvjiancheng}@scu.edu.cn, wuhongjie0818@gmail.com,
yazhou.ren@uestc.edu.cn, ivor.tsang@gmail.com, xuzenglin@hit.edu.cn

Outline

- Multi-view Subspace Clustering Revisit
- Multi-view Subspace Clustering on Topological Manifold
- Optimization Algorithm
- Experimental Results
- Conclusions

Multi-view Subspace Clustering Revisit

Subspace Clustering: given a dataset $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in R^{k \times n}$, with n data points and k features, the self-expression based subspace clustering problem can be defined as

$$\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 + \alpha \|\mathbf{Z}\|_F^2$$
$$\text{s.t. } \mathbf{Z} \geq 0, diag(\mathbf{Z}) = 0,$$

where $\boldsymbol{Z} \in R^{n \times n}$ is the similarity graph of data points, and α is a trade-off parameter.

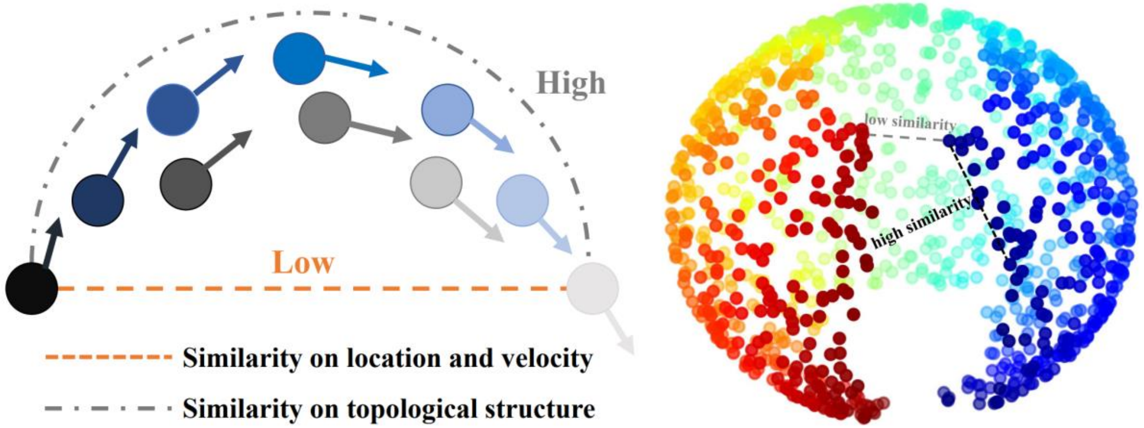
Multi-view Subspace Clustering: when data is presented in multiple view $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(m)}\}$, we can easily extend the above formula to a multi-view version:

$$\min_{\mathbf{Z}^{(v)}} \sum_{v=1}^{m} \left\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{Z}^{(v)}\right\|_F^2 + \alpha \left\|\mathbf{Z}^{(v)}\right\|_F^2$$
$$\text{s.t. } \mathbf{Z}^{(v)} \geq 0, diag\left(\mathbf{Z}^{(v)}\right) = 0,$$

Multi-view Subspace Clustering Revisit

Drawbacks of existing methods:

- Real-world datasets are usually sampled from a nonlinear low-dimensional manifold. But existing clustering methods **do not** consider the manifold topological structure.



- Existing clustering methods usually adopt predefined similarity graphs as input. The graph learning and subsequent multi-view clustering are separated. Thus the constructed graph may **not be suitable** for the subsequent clustering.

Multi-view Subspace Clustering on Topological Manifold

Our contributions:

- We argue to explore the implied data manifold by learning the topological relationship, and propose to integrate multiple affinity graphs into a consensus one with the topological relevance considered.
- Our method is a unified framework which combining affinity graph constructing, topological relevance learning, and label partitioning. And each subtask can be enhanced in a mutual reinforcement manner.
- An effective alternating iterative algorithm is carefully designed to solve the optimization problem of the proposed model. Experimental results on several benchmark datasets demonstrate the effectiveness of our method.

Multi-view Subspace Clustering on Topological Manifold

Topological Manifold Learning:

We propose to learn a more suitable manifold topological structure, such that the intrinsic similarities can be explicitly uncovered.

Given a predefined similarity graph $\boldsymbol{Z} \in R^{n\times n}$, we investigated the topological structure of data by solving

$$\min_{\mathbf{S}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{Z}_{jk} \left(\mathbf{S}_{ij} - \mathbf{S}_{ik}\right)^2 + \beta \left\|\mathbf{S} - \mathbf{I}\right\|_F^2 ,$$

where β is a balance parameter, i, j, and k are data point indexes.

$\boldsymbol{S}$ represents the target topological relationship matrix, and $\boldsymbol{S}_{ij}$ denotes the data point j's topological relevance to i.

Multi-view Subspace Clustering on Topological Manifold

With the multi-view subspace representation, connectivity constraint, and normalize term, the final objective function is:

$$\min_{\mathbf{Z}^{(v)},\mathbf{S}} \sum_{v=1}^{m} \left\| \mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{Z}^{(v)} \right\|_F^2 + \alpha \left\| \mathbf{Z}^{(v)} \right\|_F^2 +$$

$$\frac{1}{2}\sum_{v=1}^{m} w_v \sum_{i,j,k=1}^{n} \mathbf{Z}_{jk}^{(v)} \left(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}^{(v)}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}^{(v)}}} \right)^2 + \beta \left\| \mathbf{S} - \mathbf{I} \right\|_F^2$$

$$\text{s.t. } \mathbf{Z}^{(v)} \geq 0, diag\left(\mathbf{Z}^{(v)}\right) = 0, \mathbf{s}_i^T \mathbf{1} = 1, s_{ij} \geq 0, rank\left(\mathbf{L}_S\right) = n - c,$$

where w_v is the weight of v-th view, $\mathbf{D}^{(v)}$ is the degree matrix of $\mathbf{S}^{(v)}$, $\mathbf{L}_S$ is the Laplacian matrix of $\mathbf{S}$, and rank $(\mathbf{L}_S) = n - c$ is a rank constraint that manipulates the target graph $\mathbf{S}$ containing exactly c connected components.

Optimization Algorithm

Since the corresponding optimization problem is not jointly convex in all variables, we choose to solve it by updating one variable while fixing other variables.

Algorithm 2: The Algorithm for Eq. (5)

Input: Multi-view data $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}\}$ with m views, cluster number c, parameters α and β.

Initialize the weight of each view $w_v = \frac{1}{m}$.

Initialize the affinity graph $\mathbf{Z}^{(v)}$ according to Eq. (2).

Initialize the consensus graph. $\mathbf{S} = \sum_{v=1}^{m} w_v \mathbf{Z}^{(v)}$.

Output: The indicator matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ with exactly c connected components.

1: **repeat**
2: Update $\mathbf{Z}^{(v)}$ according to Eq. (9).
3: Update $\mathbf{S}$ by Algorithm 1.
4: Update $\mathbf{F}$ according to Eq. (13).
5: Update w_v according to Eq. (6).
6: **until** converge

Experimental Results

We evaluate the proposed method on seven benchmark datasets, compared with ten state-of-the-art methods.

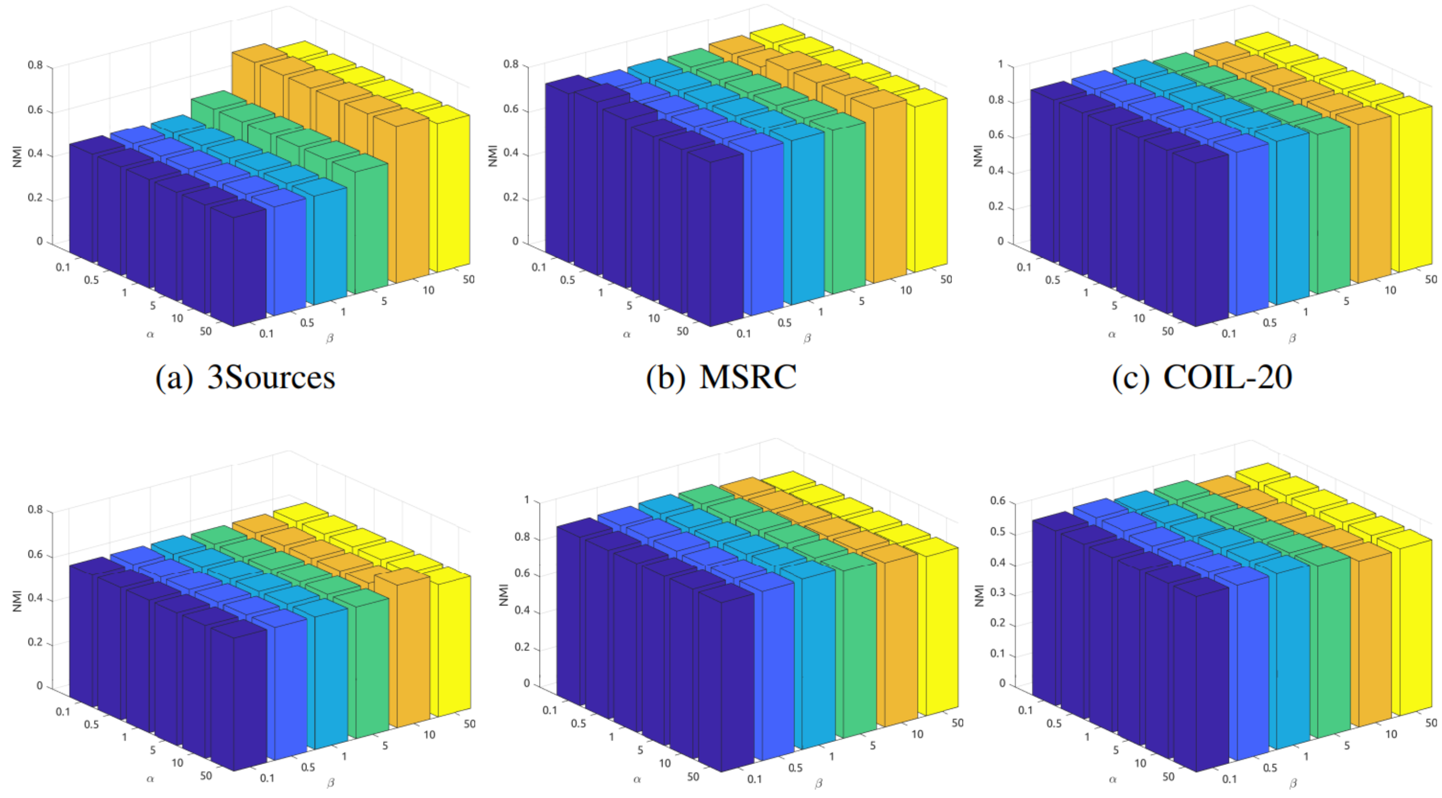
Our proposed method achieves the best clustering results in the majority of cases, and the improvement is remarkable.

Table 2: Clustering results of all methods on different datasets (%). The best performance is **bolded**, and the second best performance is underlined.

Dataset	SC_{best}	DiMSC	AMGL	MVGL	WMSC	CSMSC	GMC	LMVSC	SMVSC	CoMSC	**Ours**
					ACC						
3Sources	53.67	76.33	44.14	42.54	57.75	78.34	69.23	50.18	43.14	64.26	**81.07**
MSRC	58.95	72.38	70.67	70.48	69.00	80.48	74.76	74.71	81.43	80.86	**85.24**
COIL-20	72.75	76.15	79.30	75.21	76.58	75.06	79.10	75.56	61.07	71.90	**80.42**
Caltech-7	48.58	41.51	64.66	56.38	38.95	62.08	69.20	60.91	57.22	64.65	**77.61**
100Leaves	69.62	47.87	79.09	54.12	78.23	76.78	82.38	67.32	38.03	78.75	**83.56**
Caltech-20	41.74	28.45	49.69	57.29	33.98	47.47	45.64	47.10	61.36	53.32	**68.99**
MNIST	52.74	51.79	85.10	30.55	51.91	50.64	84.37	71.45	77.16	69.65	**87.44**
					NMI						
3Sources	49.99	63.77	18.35	27.11	49.33	70.75	54.80	30.51	24.21	59.32	**70.81**
MSRC	46.81	60.08	66.80	58.18	59.53	71.43	74.21	65.55	70.18	74.08	**77.35**
COIL-20	81.91	83.02	91.43	83.80	84.16	84.17	91.79	83.24	73.06	81.42	**91.90**
Caltech-7	28.99	32.10	52.76	51.63	28.08	51.82	60.56	44.33	44.96	55.96	**64.51**
100Leaves	86.17	70.98	90.48	63.96	90.44	89.05	90.25	84.64	64.92	90.42	**92.48**
Caltech-20	45.47	27.59	54.47	58.59	41.81	57.83	38.46	49.21	57.56	**59.38**	56.53
MNIST	47.13	34.08	76.08	24.04	47.31	46.13	76.39	63.46	62.40	64.80	**77.49**
					Purity						
3Sources	71.18	80.47	49.94	48.46	71.48	83.67	74.56	75.74	53.08	72.01	**84.62**
MSRC	60.00	72.38	74.14	70.48	71.38	80.48	79.05	75.33	81.43	81.76	**85.24**
COIL-20	75.25	78.94	84.37	77.78	78.19	77.56	84.79	79.08	61.72	78.94	**85.00**
Caltech-7	79.61	76.11	84.83	86.84	79.58	86.95	88.47	70.98	85.80	72.73	**88.60**
100Leaves	72.94	50.47	83.42	57.44	80.55	79.44	85.06	77.39	39.49	85.44	**86.01**
Caltech-20	70.54	54.83	68.33	74.85	67.29	**77.91**	55.49	52.48	71.32	61.12	75.02
MNIST	56.27	52.37	85.43	30.55	55.94	54.22	84.37	77.00	77.16	76.38	**87.44**
					F-score						
3Sources	48.49	70.68	38.18	44.75	50.79	73.17	60.47	41.87	38.46	60.49	**75.25**
MSRC	43.88	58.61	62.22	54.56	57.52	70.13	69.68	64.71	69.36	71.35	**75.29**
COIL-20	69.09	72.27	75.95	71.43	73.40	70.75	79.42	70.23	53.68	66.33	**82.29**
Caltech-7	40.01	42.26	61.41	59.77	37.78	61.74	72.17	56.37	55.46	64.92	**79.77**
100Leaves	61.94	33.12	59.14	8.58	72.63	69.64	50.42	58.20	23.20	**73.19**	69.29
Caltech-20	33.21	20.10	39.78	47.05	30.54	42.30	34.03	39.78	**66.27**	47.72	53.13
MNIST	41.53	32.80	74.99	24.46	41.08	41.41	74.43	59.42	62.39	61.04	**77.67**

Experimental Results

Parameter Analysis



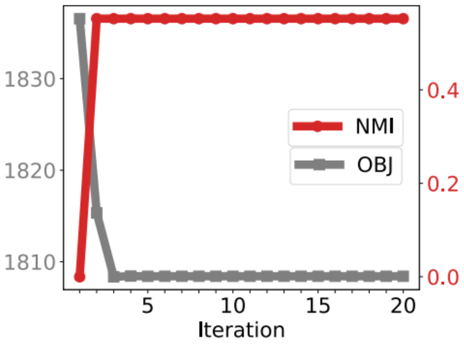
(a) 3Sources (b) MSRC (c) COIL-20

(d) Caltech-7 (e) 100Leaves (f) Caltech-20

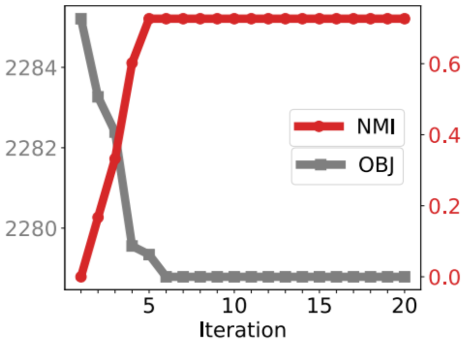
Figure 5: NMI w.r.t. α and β on different datasets.

Experimental Results

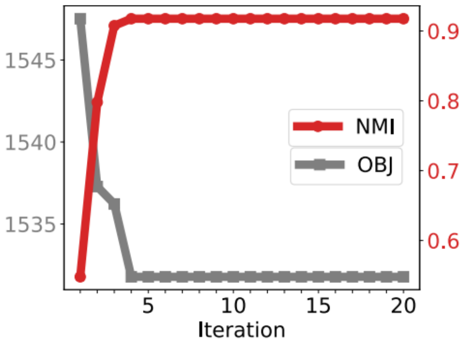
Convergence Study



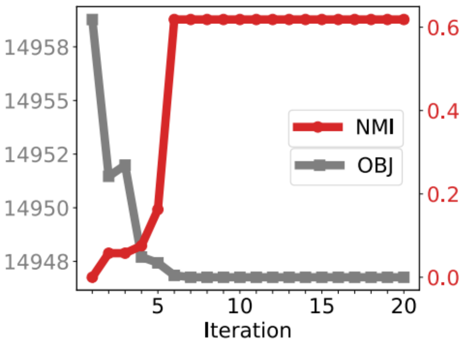
(a) 3Sources



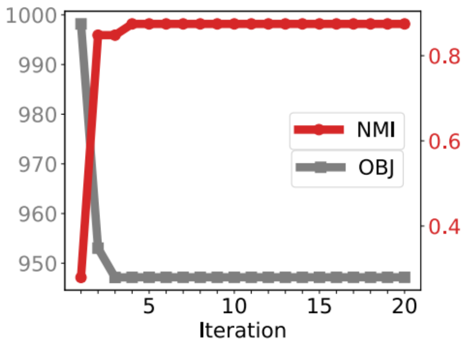
(b) MSRC



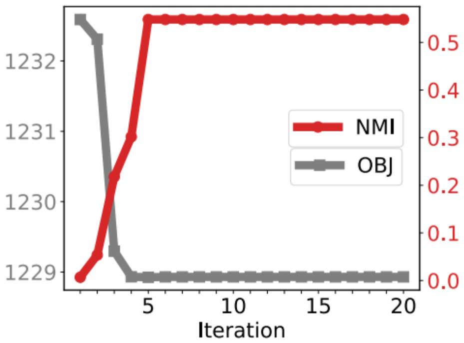
(c) COIL-20



(d) Caltech-7



(e) 100Leaves

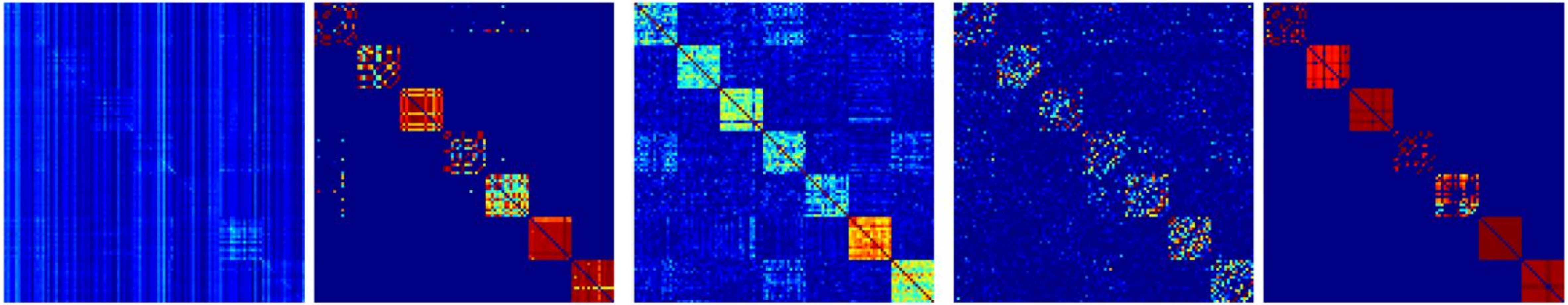


(f) Caltech-20

Figure 6: Convergence analysis of the proposed method, where OBJ denotes the objective value.

Experimental Results

We visualize the target graph learned by different methods, our model almost achieves a pure structured graph with a much clear clustering structure.



(a) DiMSC (b) MVGL (c) CSMSC (d) CoMSC (e) Ours

Conclusions

- In this paper, we propose to explore the implied data manifold by learning the topological relationship between data points.
- To do so, we integrate multiple affinity graphs into a consensus one with the topological relevance considered.
- An alternating iterative algorithm is designed to solve the optimization problem of the proposed model.

The experimental results show that:

- manifold topological structure is suitable and beneficial for multi-view subspace clustering tasks;
- our model is quite robust with respect to different parameter settings;
- the proposed optimization algorithm is very efficient and converges fast.

- In the future, we are interested in extending the proposed model to other machine learning framework, such as semi-supervised learning and deep learning.

Thanks for listening