Spectral Clustering by Subspace Randomization and Graph Fusion for High-Dimensional Data

Xiaosha Cai[1,2], Dong Huang[1,2⋆], Chang-Dong Wang[3], and Chee-Keong Kwoh[4]

[1] College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China
[2] Guangzhou Key Laboratory of Smart Agriculture, Guangzhou, China
[3] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore
xiaoshacai@hotmail.com, huangdonghere@gmail.com
changdongwang@hotmail.com, asckkwoh@ntu.edu.sg

Abstract. Subspace clustering has been gaining increasing attention in recent years due to its promising ability in dealing with high-dimensional data. However, most of the existing subspace clustering methods tend to only exploit the subspace information to construct a single affinity graph (typically for spectral clustering), which often lack the ability to go beyond a single graph to explore multiple graphs built in various subspaces in high-dimensional space. To address this, this paper presents a new **s**pectral **c**lustering approach based on **s**ubspace **r**andomization and **g**raph **f**usion (SC-SRGF) for high-dimensional data. In particular, a set of random subspaces are first generated by performing random sampling on the original feature space. Then, multiple K-nearest neighbor (K-NN) affinity graphs are constructed to capture the local structures in the generated subspaces. To fuse the multiple affinity graphs from multiple subspaces, an iterative similarity network fusion scheme is utilized to achieve a unified graph for the final spectral clustering. Experiments on twelve real-world high-dimensional datasets demonstrate the superiority of the proposed approach. The MATLAB source code is available at https://www.researchgate.net/publication/338864134.

Keywords: Data clustering; Spectral clustering; Subspace clustering; High-dimensional data; Random subspaces; Graph fusion

1 Introduction

Data clustering is a fundamental yet still very challenging problem in data mining and knowledge discovery [13]. A huge number of clustering techniques have been developed in the past few decades [2–7, 9–12, 14–18, 21, 22], out of which the spectral clustering has been a very important category with its effectiveness and robustness in dealing with complex data [3, 7, 14, 18, 22]. In this paper, we focus on the spectral clustering technique, especially for high-dimensional scenarios.

⋆ Corresponding author

In high-dimensional data, it is often recognized that the cluster structures of data may lie in some low-dimensional subspaces [3]. Starting from this assumption, many efforts have been made to enable the spectral clustering for high-dimensional data by exploiting the subspace information from different technical perspectives [1, 3, 4, 15, 17, 21, 23]. Typically, a new affinity matrix is often learned with the subspace structure taken into consideration, upon which the spectral clustering process is then performed to obtain the final clustering. For example, Liu et al. [17] proposed a low-rank representation (LRR) approach to learn an affinity matrix, whose goal is to segment the data points into their respective subspaces. Chen et al. [1] exploited K-nearest neighbor (K-NN) based sparse representation coefficient vectors to build an affinity matrix for high-dimensional data. He et al. [4] used information theoretic objective functions to combine structured LRRs, where the global structure of data is incorporated. Li et al. [15] presented a subspace clustering approach based on Cauchy loss function (CLF) to alleviate the potential noise in high-dimensional data. Elhamifar and Vidal [3] proposed the sparse subspace clustering (SSC) approach by incorporating the low-dimensional neighborhood information, where each data point is represented by a combination of other points in its own subspace and a new similarity matrix is then constructed. You et al. [23] extended the SSC approach by introducing orthogonal matching pursuit (OMP) to learn a subspace-preserving representation. Wang et al. [21] combined SSC and LRR into a novel low-rank sparse subspace clustering (LRSSC) approach.

Although these methods [1, 3, 4, 15, 17, 21, 23] have made significant progress in exploiting subspace information for enhancing spectral clustering of high-dimensional data, most of them tend to utilize a single affinity graph (associated with a single affinity matrix) by subspace learning, but lack the ability to go beyond a single affinity graph to jointly explore a variety of graph structures in various subspaces in the high-dimensional space. To overcome this limitation, this paper presents a new **s**pectral **c**lustering by **s**ubspace **r**andomization and **g**raph **f**usion (SC-SRGF) approach. Specifically, multiple random subspaces are first produced, based on which we construct multiple K-NN affinity graphs to capture the locality information in various subspaces. Then, the multiple affinity graphs (associated with multiple affinity matrices) are integrated into a unified affinity graph by using an iterative similarity network fusion scheme. With the unified graph obtained, the final spectral clustering result can be obtained by partitioning the this new affinity graph. We conduct experiments on twelve high-dimensional datasets, which have shown the superiority of our approach.

The rest of the paper is organized as follows. The proposed approach is described in Section 2. The experimental results are reported in Section 3. The paper is concluded in Section 4.

2 Proposed Framework

In this section, we describe the overall process of the proposed SC-SRGF approach. The formulation of the clustering problem is given in Section 2.1. The

construction of multiple K-NN graphs (corresponding to multiple affinity matrices) in a variety of random subspaces is introduced in Section 2.2. Finally, the fusion of the multiple graphs into a unified graph and the spectral clustering process are described in Section 2.3.

2.1 Problem Formulation

Let $X \in \mathbb{R}^{n \times d}$ be the data matrix, where n is the number of data points and d is the number of features. Let $x_i \in \mathbb{R}^d$ denote the i-th data point, corresponding to the i-row in X. Thus the data matrix can be represented as $X = (x_1, x_2, \cdots, x_n)^\top$. Let $f_j \in \mathbb{R}^n$ denote the j-th data feature, corresponding to the j-th column in X. Thus the data matrix can also be represented as $X = (f_1, f_2, \cdots, f_d)$. The purpose of clustering is to group the n data points into a certain number of subsets, each of which is referred to as a cluster.

2.2 Affinity Construction in Random Subspaces

In this work, we aim to enhance the spectral clustering for high-dimensional datasets with the help of the information of various subspaces. Before exploring the subspace information, a set of random subspaces are first generated. Note that each subspace consists of a certain number of features, and thereby corresponds to a certain number of columns in the data matrix X.

Multiple random subspaces are generated by performing random sampling (without replacement) on the data features with a sampling ratio r. Let m denote the number of generated random subspaces. Then the set of random subspaces can be represented as

$$\mathcal{F} = \{F^{(1)}, F^{(2)}, \cdots, F^{(m)}\}, \tag{1}$$

where

$$F^{(i)} = (f_1^{(i)}, f_2^{(i)}, \cdots, f_{d'}^{(i)}) \tag{2}$$

denotes the i-th random subspace, $f_j^{(i)}$ denotes the j-th feature in $F^{(i)}$, and $d' = \lfloor r \cdot d \rfloor$ is the number of features. Each subspace can be viewed as selecting corresponding columns in the original data matrix. Therefore, the data submatrix in a given subspace $F^{(i)}$ can be represented as

$$X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)})^\top \tag{3}$$

where $x_j^{(i)} \in \mathbb{R}^{d'}$ denotes the j-th data point in this subspace.

To explore the locality structures in various subspaces, multiple K-NN graphs are constructed. Specifically, given a subspace $F^{(i)}$, its K-NN graph can be defined as

$$G^{(i)} = \{V, E^{(i)}\}, \tag{4}$$

where $V = \{x_1, x_2, \cdots, x_n\}$ is the node set and $E^{(i)}$ is the edge set. The weights of the edges in the graph are computed as

$$E^{(i)} = \{e_{jk}^{(i)}\}_{n\times n}, \tag{5}$$

$$e_{jk}^{(i)} = \begin{cases} \exp(-\frac{d(x_j^{(i)}, x_k^{(i)})}{2\sigma}), & \text{if } x_j \in KNN^i(x_k) \text{ or } x_k \in KNN^i(x_j), \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $e_{jk}^{(i)}$ is the edge weight between nodes x_j and x_k in $G^{(i)}$, $d(x_j, x_k)$ is the Euclidean distance between $x_j^{(i)}$ and $x_k^{(i)}$, $KNN^i(x_k)$ is the set of K-NNs of x_k in the i-th subspace, and the kernel parameter σ is set to the average distance between all points.

With the m random subspaces, we can construct m affinity graphs (corresponding to m affinity matrices) as follows:

$$\mathcal{G} = \{G^{(1)}, G^{(2)}, \cdots, G^{(m)}\}. \tag{7}$$

Note that these affinity graphs share the same node set (i.e., the set of all data points), but have different edge weights constructed in different subspaces, which enable them to capture a variety of underlying subspace structure information in high-dimensional space for enhanced clustering performance.

2.3 Fusing Affinity Graphs for Spectral Clustering

In this section, we proceed to fuse multiple affinity graphs (corresponding to multiple affinity matrices) into a unified affinity graph for robust spectral clustering of high-dimensional data.

Specifically, we adopt the similarity network fusion (SNF) [20] scheme to fuse the information of multiple graphs. For simplicity, the set of the affinity matrices for the m graphs is represented as $\mathcal{E} = \{E^{(1)}, E^{(2)}, \cdots, E^{(m)}\}$. The goal here is to merge the m affinity matrices in $\mathcal{E}$ into a unified affinity matrix $\tilde{E}$.

By normalizing the rows in the affinity matrix $E^{(i)}$, we have $\bar{E}^{(i)} = \{\bar{e}_{jk}^{(i)}\}_{n\times n} = (D^{(i)})^{-1}E^{(i)}$, where $D^{(i)}$ is the degree matrix of $E^{(i)}$. Then the initial status matrix $P_{t=0}^{(i)}$ can be defined as

$$P_{t=0}^{(i)} = \frac{\bar{E}^{(i)} + (\bar{E}^{(i)})^\top}{2}, \tag{8}$$

$$\tag{9}$$

And the kernel matrix $S^{(i)} = \{s_{jk}^{(i)}\}_{n\times n}$ can be defined as

$$s_{jk}^{(i)} = \begin{cases} \frac{\bar{e}_{jk}^{(i)}}{\sum_{x_l \in KNN(x_j)} \bar{e}_{jl}^{(i)}}, & \text{if } x_k \in KNN(x_j), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

With the above two types of matrices defined, we can iteratively update the status matrices by exploiting the information of multiple affinity matrices. Particularly, in each iteration, the i-th status matrix is updated as follows [20]:

$$P_{t+1}^{(i)} = S^{(i)} \times (\frac{\sum_{j \neq i} P_t^{(j)}}{m-1}) \times (S^{(i)})^\top, \quad i = 1, 2, \cdots, m. \tag{11}$$

After each iteration, $P_{t+1}^{(i)}$ will be normalized by $P_{t+1}^{(i)} = (D_{t+1}^{(i)})^{-1} P_{t+1}^{(i)}$ with $D_{t+1}^{(i)}$ being the degree matrix of $P_{t+1}^{(i)}$.

When the status matrices converge or the maximum number of iterations is reached, the iteration process stops and the fused affinity matrix will be computed as

$$\tilde{E} = \frac{1}{m} \sum_{i=1}^{m} P^{(i)}. \tag{12}$$

Then the unified matrix $\tilde{E}$ will be symmetrized by $\tilde{E} = (\tilde{E} + \tilde{E}^\top)/2$. With the unified affinity matrix $\tilde{E}$ obtained by fusing information of multiple affinity matrices from multiple subspaces, we can proceed to perform spectral clustering on this unified matrix to build the clustering result with a certain number of, say, k', clusters.

Let $\tilde{D}$ be the degree matrix of $\tilde{E}$. Its graph Laplacian can be computed as

$$\tilde{L} = \tilde{D} - \tilde{E}. \tag{13}$$

After that, eigen-decomposition is performed on the graph Laplacian $\tilde{L}$ to obtain the k' eigenvectors that correspond to its first k' eigenvalues. Then the k' eigenvectors are stacked to form a new matrix $\tilde{U} \in \mathbb{R}^{n \times k'}$, where the i-th column corresponds to the i-th eigenvector. Then, by treating each row as a new feature vector for the data point, some discretization techniques like k-means [18] can be performed on the matrix $\tilde{U}$ to achieve the final spectral clustering result.

3 Experiments

In this section, we conduct experiments on a variety of high-dimensional datasets to compare our approach against several other spectral clustering approaches.

3.1 Datasets and Evaluation Measures

In our experiments, twelve real-world high-dimensional datasets are used, namely, *Armstrong-2002-v1* [19], *Chowdary-2006* [19], *Golub-1999-v2* [19], *Alizadeh-2000-v2* [19], *Alizadeh-2000-v3* [19], *Bittner-2000* [19], *Bredel-2005* [19], *Garber-2001* [19], *Khan-2001* [19], *Binary-Alpha* (*BA*) [14], *Coil20* [14], and *Multiple Features* (*MF*) [6]. To simplify the description, the twelve benchmark datasets are abbreviated as *DS-1* to *DS-12*, respectively (as shown in Table 1).

Table 1. Dataset Description

Dataset	Abbr.	#Instance	Dimension	#Class
Armstrong-2002-v1	*DS-1*	72	1081	2
Chowdary-2006	*DS-2*	104	182	2
Golub-1999-v2	*DS-3*	72	1868	3
Alizadeh-2000-v2	*DS-4*	62	2093	3
Alizadeh-2000-v3	*DS-5*	62	2093	4
Bittner-2000	*DS-6*	38	2201	2
Bredel-2005	*DS-7*	50	739	3
Garber-2001	*DS-8*	66	4553	4
Khan-2001	*DS-9*	83	1069	4
Binary Alpha	*DS-10*	1404	320	36
Coil20	*DS-11*	1440	1024	20
Multiple Features	*DS-12*	2000	649	10

Table 2. Average NMI over 20 runs by different methods on the benchmark datasets. The best score in each row is in bold.

Dataset	SC	KASP	SSC	SSC-OMP	SC-SRGF
DS-1	$0.366_{\pm 0.000}$	$0.263_{\pm 0.104}$	$0.366_{\pm 0.000}$	$0.351_{\pm 0.000}$	$\mathbf{0.546}_{\pm 0.117}$
DS-2	$0.081_{\pm 0.000}$	$0.171_{\pm 0.295}$	$0.764_{\pm 0.000}$	$\mathbf{0.860}_{\pm 0.000}$	$0.849_{\pm 0.022}$
DS-3	$0.596_{\pm 0.000}$	$0.404_{\pm 0.245}$	$0.690_{\pm 0.000}$	$0.700_{\pm 0.000}$	$\mathbf{0.801}_{\pm 0.049}$
DS-4	$0.605_{\pm 0.000}$	$0.851_{\pm 0.164}$	$0.734_{\pm 0.000}$	$0.620_{\pm 0.000}$	$\mathbf{0.913}_{\pm 0.000}$
DS-5	$0.560_{\pm 0.000}$	$0.614_{\pm 0.061}$	$0.442_{\pm 0.001}$	$0.441_{\pm 0.007}$	$\mathbf{0.626}_{\pm 0.002}$
DS-6	$0.032_{\pm 0.000}$	$0.032_{\pm 0.027}$	$0.035_{\pm 0.000}$	$0.035_{\pm 0.000}$	$\mathbf{0.053}_{\pm 0.003}$
DS-7	$0.249_{\pm 0.000}$	$\mathbf{0.367}_{\pm 0.089}$	$0.102_{\pm 0.000}$	$0.115_{\pm 0.000}$	$0.311_{\pm 0.075}$
DS-8	$0.082_{\pm 0.005}$	$0.139_{\pm 0.055}$	$0.086_{\pm 0.004}$	$\mathbf{0.172}_{\pm 0.011}$	$0.161_{\pm 0.024}$
DS-9	$0.604_{\pm 0.000}$	$0.328_{\pm 0.073}$	$0.835_{\pm 0.000}$	$0.533_{\pm 0.009}$	$\mathbf{0.881}_{\pm 0.014}$
DS-10	$0.503_{\pm 0.005}$	$0.591_{\pm 0.009}$	$0.580_{\pm 0.006}$	$0.260_{\pm 0.006}$	$\mathbf{0.613}_{\pm 0.007}$
DS-11	$0.780_{\pm 0.000}$	$0.860_{\pm 0.022}$	$0.864_{\pm 0.005}$	$0.517_{\pm 0.201}$	$\mathbf{0.888}_{\pm 0.001}$
DS-12	$0.655_{\pm 0.000}$	$0.866_{\pm 0.018}$	$0.824_{\pm 0.001}$	$0.556_{\pm 0.002}$	$\mathbf{0.871}_{\pm 0.030}$
Avg. score	0.425	0.457	0.527	0.430	**0.626**
Avg. rank	3.83	3.17	3.00	3.50	**1.25**

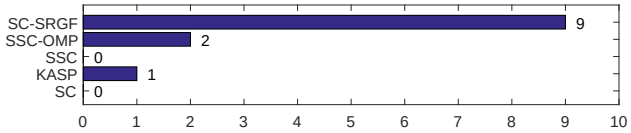

Fig. 1. Number of times being ranked in the first position in Table 2.

To quantitatively evaluate the clustering results of different algorithms, two widely-used evaluation measures are used, namely, normalized mutual information (NMI) [8] and adjusted Rand index (ARI) [8]. Note that larger values of NMI and ARI indicate better clustering results.

Table 3. Average ARI over 20 runs by different methods on the benchmark datasets. The best score in each row is in bold.

Dataset	SC	KASP	SSC	SSC-OMP	SC-SRGF
DS-1	$0.268_{\pm 0.000}$	$0.152_{\pm 0.058}$	$0.268_{\pm 0.000}$	$0.238_{\pm 0.000}$	$\mathbf{0.578}_{\pm 0.181}$
DS-2	$0.066_{\pm 0.000}$	$0.168_{\pm 0.340}$	$0.851_{\pm 0.000}$	$\mathbf{0.924}_{\pm 0.000}$	$0.916_{\pm 0.015}$
DS-3	$0.656_{\pm 0.000}$	$0.378_{\pm 0.270}$	$0.707_{\pm 0.000}$	$0.729_{\pm 0.000}$	$\mathbf{0.844}_{\pm 0.047}$
DS-4	$0.506_{\pm 0.000}$	$0.897_{\pm 0.148}$	$0.796_{\pm 0.000}$	$0.627_{\pm 0.000}$	$\mathbf{0.947}_{\pm 0.000}$
DS-5	$0.360_{\pm 0.003}$	$\mathbf{0.479}_{\pm 0.057}$	$0.261_{\pm 0.006}$	$0.289_{\pm 0.005}$	$0.427_{\pm 0.005}$
DS-6	$0.018_{\pm 0.000}$	$0.009_{\pm 0.029}$	$0.020_{\pm 0.000}$	$0.020_{\pm 0.000}$	$\mathbf{0.047}_{\pm 0.036}$
DS-7	$0.277_{\pm 0.000}$	$0.387_{\pm 0.169}$	$0.105_{\pm 0.000}$	$0.112_{\pm 0.000}$	$\mathbf{0.404}_{\pm 0.122}$
DS-8	$0.068_{\pm 0.010}$	$0.059_{\pm 0.067}$	$0.0004_{\pm 0.003}$	$0.103_{\pm 0.020}$	$\mathbf{0.128}_{\pm 0.024}$
DS-9	$0.466_{\pm 0.000}$	$0.206_{\pm 0.056}$	$0.826_{\pm 0.000}$	$0.433_{\pm 0.009}$	$\mathbf{0.860}_{\pm 0.011}$
DS-10	$0.210_{\pm 0.005}$	$0.291_{\pm 0.011}$	$0.300_{\pm 0.008}$	$0.051_{\pm 0.004}$	$\mathbf{0.327}_{\pm 0.008}$
DS-11	$0.638_{\pm 0.000}$	$0.682_{\pm 0.055}$	$0.701_{\pm 0.017}$	$0.260_{\pm 0.019}$	$\mathbf{0.744}_{\pm 0.002}$
DS-12	$0.559_{\pm 0.000}$	$0.818_{\pm 0.029}$	$0.754_{\pm 0.000}$	$0.445_{\pm 0.006}$	$\mathbf{0.826}_{\pm 0.056}$
Avg. score	0.341	0.377	0.466	0.353	**0.587**
Avg. rank	3.67	3.42	3.08	3.50	**1.17**

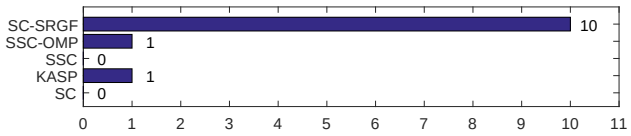

Fig. 2. Number of times being ranked in the first position in Table 3.

In terms of the experimental setting, we use $m = 20$, $K = 5$, and $r = 0.5$ on all the datasets in the experiments. In the following, the robustness of our approach with varying values of the parameters will also be evaluated in Section 3.3.

3.2 Comparison Against the Baseline Approaches

In this section, we compare the proposed SC-SRGF method against four baseline spectral clustering methods, namely, original spectral clustering (SC) [18], k-means-based approximate spectral clustering (KASP) [22], sparse subspace clustering (SSC) [3], and sparse subspace clustering by orthogonal matching pursuit (SSC-OMP) [23]. The detailed comparison results are reported in Tables 2, 3, and 4, and Figures 1 and 2.

In terms of NMI, as shown in Table 2, the proposed SC-SRGF method obtains the best scores on the *DS-1*, *DS-3*, *DS-4*, *DS-5*, *DS-6*, *DS-9*, *DS-10*, *DS-11*, and *DS-12* datasets. The average NMI score (across the twelve datasets) of our method is 0.626, which is much higher than the second highest average score of 0.527 (obtained by SSC). The average rank of our method is 1.25, whereas the second best method only achieves an average rank of 3.00. As shown in Figure 1, our SC-SRGF method yields the best NMI scores on nine out of the

Table 4. Average time costs (s) by different methods on the benchmark datasets.

Dataset	SC	KASP	SSC	SSC-OMP	SC-SRGF
DS-1	0.197	0.215	0.309	0.229	0.296
DS-2	0.193	0.225	0.316	0.219	0.734
DS-3	0.204	0.217	0.376	0.222	0.770
DS-4	0.203	0.226	0.291	0.221	0.298
DS-5	0.205	0.224	0.295	0.223	0.296
DS-6	0.194	0.216	0.299	0.214	0.290
DS-7	0.200	0.214	0.299	0.215	0.287
DS-8	0.213	0.230	0.727	0.257	0.295
DS-9	0.201	0.217	0.337	0.218	0.295
DS-10	0.681	0.330	7.740	0.640	19.592
DS-11	0.871	0.440	21.571	0.769	20.915
DS-12	1.060	0.489	25.134	0.787	27.666

twelve datasets in Table 2, whereas the second and third best methods only achieves the best scores on two and one benchmark datasets, respectively.

In terms of ARI, as shown in Table 3, our SC-SRGF method also yields overall better performance than the baseline methods. Specifically, our method achieves an average ARI score (across twelve datasets) of 0.587, whereas the second best score is only 0.466. Our method obtains an average rank of 1.17, whereas the second best average rank is only 3.08. Further, as can be seen in Figure 2, our method achieves the best ARI score on ten out of the twelve datasets, which also significantly outperforms the other spectral clustering methods.

In terms of time cost, as shown in Table 4, it takes our SC-SRGF method less than 1 second to process the first nine smaller datasets and less than 30 seconds to process the other three larger datasets, which is comparable to the time costs of the SSC method. Therefore, with the experimental results in Tables 2, 3, and 4 taken into account, it can be observed that our method is able to achieve significantly better clustering results for high-dimensional datasets (as shown in Tables 2 and 3) while exhibiting comparable efficiency with the important baseline of SSC (as shown in Table 4).

All experiments were conducted in MATLAB R2016a on a PC with i5-8400 CPU and 64GB of RAM.

3.3 Parameter Analysis

In this section, we evaluate the performance of our SC-SRGF approach with three different parameters, i.e., the number of affinity matrices (or random subspaces) m, the number of nearest neighbors K, and the sampling ratio r.

Influence of the Number of Affinity Matrices m The parameter m controls the number of random subspaces to be generated, which is also the number of affinity matrices to be fused in the affinity fusion process. Figure 3 illustrates

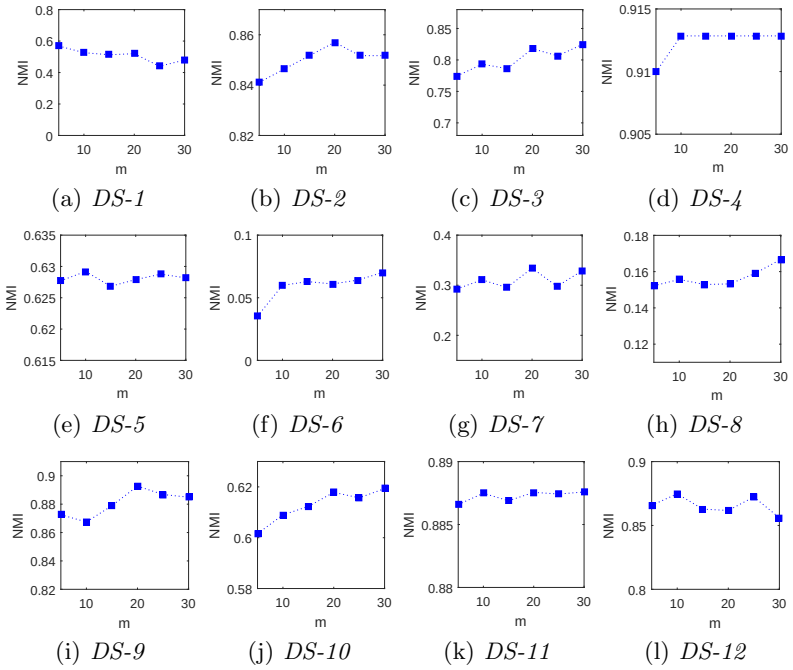

Fig. 3. Average NMI over 20 runs by SC-SRGF with varying number of affinity matrices m.

the performance (w.r.t. NMI) of our SC-SRGF approach as the number of affinity matrices goes from 5 to 30 with an interval of 5. As shown in Fig. 3, the performance of SC-SRGF is stable with different values of m. Empirically, a moderate value of m, say, in the interval of $[10, 30]$, is preferred. In the experiments, we use $m = 20$ on all of the datasets.

Influence of the Number of Nearest Neighbors K The parameter K controls the number of nearest neighbors when constructing the K-NN graphs for the multiple random subspaces. As can be seen in Figure 4, a smaller value of K can be beneficial to the performance, probably due to the fact that the K-NN graph with a smaller K may better reflect the locality characteristics in a given subspace. In the experiments, we use $K = 5$ on all of the datasets.

Influence of the Sampling Ratio r The parameter r controls the sampling ratio when producing the multiple random subspaces from the high-dimensional space. As shown in Figure 5, a moderate value of r is often preferred on the benchmark datasets. Empirically, it is suggested that the sampling ratio be set in the interval of $[0.2, 0.8]$. In the experiments, we use $r = 0.5$ on all of the datasets.

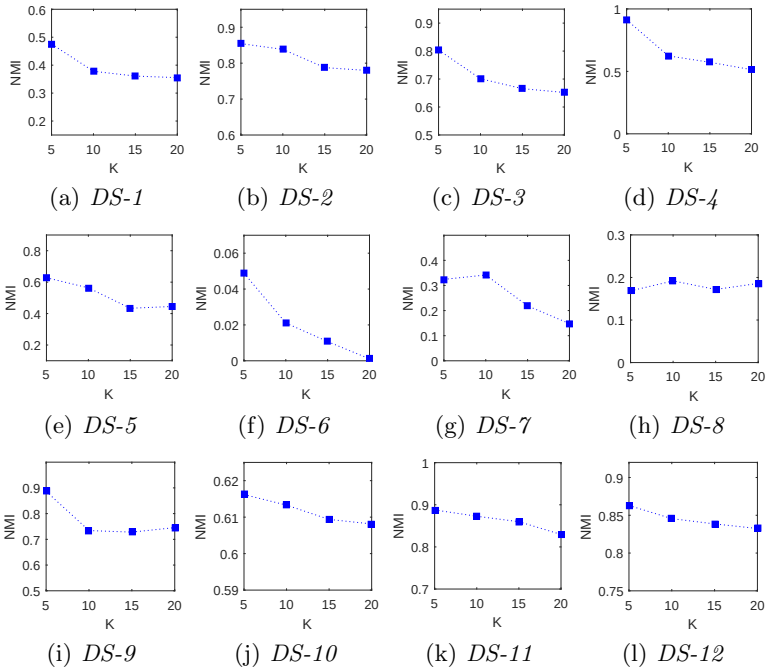

Fig. 4. Average NMI over 20 runs by SC-SRGF with varying number of nearest neighbors K.

Brief Summary From the above experimental results, we can observe that the proposed SC-SRGF approach exhibits quite good consistency and robustness w.r.t. the three parameters. That is, the three parameters in our approach are not sensitive ones. They do not need any sophisticated parameter tuning, and can be safely set to some moderate values across different datasets.

4 Conclusion

In this paper, we propose a new spectral clustering approach termed SC-SRGF for high-dimensional data, which is able to explore diversified subspace information inherent in high-dimensional space by means of subspace randomization and affinity graph fusion. In particular, a set of multiple random subspaces are first generated by performing random sampling on the original feature space repeatedly. After that, multiple K-NN graphs are constructed to capture the locality information of the multiple subspaces. Then, we utilize an iterative graph fusion scheme to combine the multiple affinity graphs (i.e., multiple affinity matrices) into a unified affinity graph, based on which the final spectral clustering result can be achieved. We have conducted extensive experiments on twelve real-world

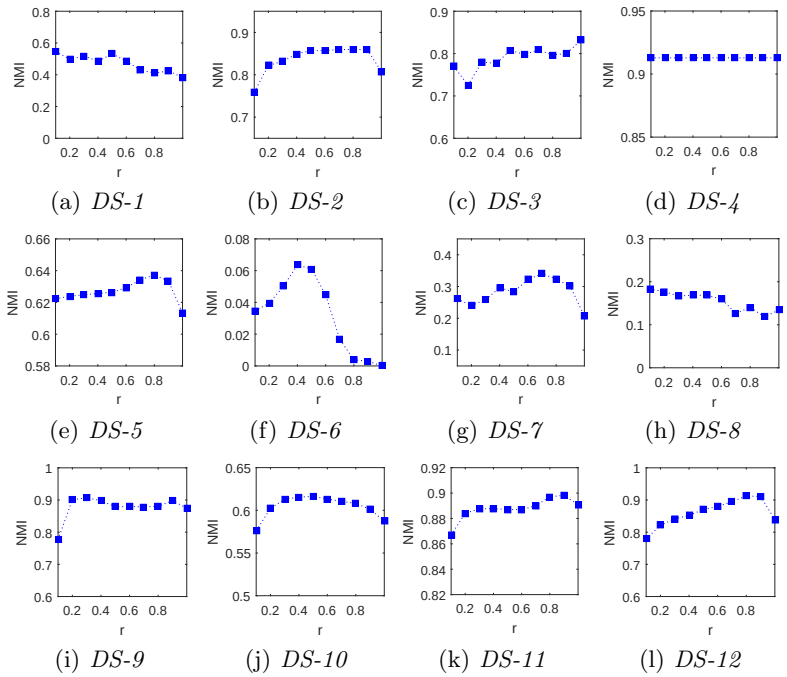

Fig. 5. Average NMI over 20 runs by SC-SRGF with varying sampling ratio r.

high-dimensional datasets, which demonstrate the superiority of our SC-SRGF approach when compared with several baseline spectral clustering approaches.

Acknowledgments. This work was supported by NSFC (61976097 & 61876193) and A*STAR-NTU-SUTD AI Partnership Grant (No. RGANS1905).

References

1. Chen, F., Wang, S., Fang, J.: Spectral clustering of high-dimensional data via k-nearest neighbor based sparse representation coefficients. In: Proc. of International Joint Conference on Neural Networks (IJCNN). pp. 363–374 (2015)
2. Chen, M.S., Huang, L., Wang, C.D., Huang, D.: Multi-view clustering in latent embedding space. In: Proc. of AAAI Conference on Artificial Intelligence (2020)
3. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(11), 2765–2781 (2013)
4. He, R., Wang, L., Sun, Z., Zhang, Y., Li, B.: Information theoretic subspace clustering. IEEE Transactions on Neural Networks and Learning Systems **27**(12), 2643–2655 (2016)

5. Huang, D., Wang, C., Peng, H., Lai, J., Kwoh, C.: Enhanced ensemble clustering via fast propagation of cluster-wise similarities. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2018). https://doi.org/10.1109/TSMC.2018.2876202
6. Huang, D., Wang, C.D., Lai, J.H.: Locally weighted ensemble clustering. IEEE Transactions on Cybernetics **48**(5), 1460–1473 (2018)
7. Huang, D., Wang, C.D., Wu, J.S., Lai, J.H., Kwoh, C.K.: Ultra-scalable spectral clustering and ensemble clustering. IEEE Transactions on Knowledge and Data Engineering (2019). https://doi.org/10.1109/TKDE.2019.2903410
8. Huang, D., Cai, X., Wang, C.D.: Unsupervised feature selection with multi-subspace randomization and collaboration. Knowledge-Based Systems **182**, 104856 (2019)
9. Huang, D., Lai, J.H., Wang, C.D.: Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. Neurocomputing **170**, 240–250 (2015)
10. Huang, D., Lai, J.H., Wang, C.D.: Robust ensemble clustering using probability trajectories. IEEE Transactions on Knowledge and Data Engineering **28**(5), 1312–1326 (2016)
11. Huang, D., Lai, J.H., Wang, C.D., Yuen, P.C.: Ensembling over-segmentations: From weak evidence to strong segmentation. Neurocomputing **207**, 416–427 (2016)
12. Huang, D., Lai, J., Wang, C.D.: Ensemble clustering using factor graph. Pattern Recognition **50**, 131–142 (2016)
13. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters **31**(8), 651–666 (2010)
14. Kang, Z., Peng, C., Cheng, Q., Xu, Z.: Unified spectral clustering with optimal graph. In: Proc. of AAAI Conference on Artificial Intelligence. pp. 3366–3373 (2018)
15. Li, X., Lu, Q., Dong, Y., Tao, D.: Robust subspace clustering by Cauchy loss function. IEEE Transactions on Neural Networks and Learning Systems **30**(7), 2067–2078 (2019)
16. Liang, Y., Huang, D., Wang, C.D.: Consistency meets inconsistency: A unified graph learning framework for multi-view clustering. In: Proc. of IEEE International Conference on Data Mining (ICDM) (2019)
17. Liu, G., Lin, Z., Yan, S., Sun, J., Ma, Y., Yu, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 171–184 (2013)
18. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)
19. de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B., Schliep, A.: Clustering cancer gene expression data: A comparative study. BMC bioinformatics **9**(1), 497 (2008)
20. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature Methods **11**, 333–337 (2014)
21. Wang, Y., Xu, H., Leng, C.: Provable subspace clustering: When LRR meets SSC. IEEE Transactions on Information Theory **65**(9), 5406–5432 (2019)
22. Yan, D., Huang, L., Jordan, M.I.: Fast approximate spectral clustering. In: Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 907–916 (2009)
23. You, C., Robinson, D.P., Vidal, R.: Scalable sparse subspace clustering by orthogonal matching pursuit. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)