Spectral Clustering Algorithm Based on OptiSim Selection

Xuejuan LIU, Junguo WANG, and Xiangying YUAN

***Abstract*—The spectral clustering (SC) method has a good clustering effect on arbitrary structure datasets because of its solid theoretical basis. However, the required time complexity is high, thus limiting the application of SC in big datasets. To reduce time complexity, we propose an SC algorithm based on OptiSim Selection (SCOSS) in this study. This new algorithm starts from selecting a representative subset by using an optimizable k-dissimilarity selection algorithm (OptiSim) and then uses the Nyström method to approximate the eigenvectors of the similarity matrix. Theoretical deductions and experiment results show that the proposed algorithm can use less clustering time to achieve a good clustering result.**

***Index Terms*—spectral clustering, Nyström method, OptiSim selection, eigen-decomposition.**

I. Introduction

A. Background

As an essential issue in machine learning, clustering analysis is widely used in image processing, text mining, and social networking[1−3]. Clustering analysis divides a dataset into clusters such that intra-cluster similarity is maximized and inter-cluster similarity is minimized without priori knowledge[4]. Commonly used clustering algorithms, including k-means[5], FCM[6], and DBSCAN[7], have some disadvantages, such as quickly obtaining an optimal local solution, being sensitive to the initial setting, or relying heavily on data distribution[8]. The spectral clustering (SC) method has recently been a trending topic because it can obtain an optimal global solution for nonconvex spatial data[9].

Based on spectral graph theory, the SC method has a robust theoretical basis[10]. Regarding each data point as a vertex of an undirected graph, SC starts by calculating each pair similarity between vertices to obtain a similarity matrix, which is then converted into a Laplacian matrix. It solves the eigenvectors of the Laplacian matrix for clustering. Thus, an optimal global solution of the spectral partition criterion in the relaxed continuous domain is obtained. However, the spatial complexity of SC in storing the similarity matrix is $O(N^2)$, and the time complexity in decomposing the Laplacian matrix is $O(N^3)$. In the current big data era, such high computational complexity limits the application of the SC method.

Manuscript received July 13, 2020; revised January 23, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61571226 and in part by Nanjing University of Finance & Economics under Grant LXJXW19001.

Xuejuan LIU is a Lecturer of School of Accounting, Nanjing University of Finance & Economics, Nanjing 210023, China(e-mail: liu_juanjuan80@126.com).

Junguo WANG is an Engineer of Information Management Center, Nanjing Forest Police College, Nanjing 210023, China (e-mail: wangjg@nfpc.edu.cn).

Xiangying YUAN is a Senior Engineer of Information Management Center, Nanjing Forest Police College, Nanjing 210023, China (e-mail: 35000653@qq.com).

B. Related work

Some progress has already been made to solve the efficiency problem of SC. Song et al. proposed a parallel SC approach, in which the similarity matrix was transformed into a sparse one, and the speed of clustering was improved by using computer clusters[11]. Yan et al. used a k-means or RP tree method as a data preprocessor to obtain representative samples, utilized the SC method to cluster these representatives, and then finally assigned each data point with a label according to its distance to each representative[12]. Lin and his collaborators found that the similarity matrix eigenvectors were convergent. They picked the first eigenvector obtained by multiple iterative calculations to cluster the overall dataset[13]. In [14-16], an incomplete Cholesky decomposition method was applied in constructing an approximate Laplacian matrix to reduce computation complexity. In [17], a sampling method based on a neural network was used to accelerate the speed of the SC method. In [18], random Fourier features were used to represent data explicitly in kernel space and thus reduce the computational complexity of SC.

The Nyström extension method is a well-known acceleration means for SC, which only uses some samples to approximate the clustering. Fowlkes et al. first proposed that the Nyström method could be applied in SC to improve the efficiency of clustering[19]. In their study, a small randomly selected subset was utilized to approach the eigenspace of the Laplacian matrix, and it was used in image segmentation with a good performance. For the quality of representatives playing an essential role in the approximation of Nyström, Zhang et al. used the k-means clustering method to preprocess the dataset and then extracted samples[20]. In [21], a sampling method based on the farthest and nearest strategy was developed to get good samples. In this method, each data object was assigned a sampling probability, and whether it could be selected as a sample depended on probabilities. Kumar and his collaborators established an integrated Nyström scheme to minimize the clustering time of SC but at the expense of clustering quality[22]. In [23], the projections on the leading eigenvectors learned from training datasets were used to replace the affinity vector for Nyström extension. Other works that applied the Nyström method in SC include sampling methods based on alternate[24] and maximum diversity[25]. In short, the key to SC acceleration schemes based on the Nyström extension is the selection of better samples because they have an important effect on clustering quality.

C. Contribution

We develop an SC algorithm based on OptiSim Selection (SCOSS) in this study. The novelty of SCOSS is introducing

an optimizable k-dissimilarity selection algorithm (OptiSim) and the utilization of the Nyström method. The OptiSim selection algorithm is a balanced sampling method by which the extracted samples can evenly cover the entire data space. In SCOSS, we use the OptiSim method to pick a representative subset and utilize Nyström method to approximate the resolution of SC.

We organize the rest of the paper as follows: In Section II, we introduce SC method based on graph theory. In Section III, we state the principle of the Nyström extension and OptiSim selection, and present SCOSS algorithm in detail. In Section IV, we conduct experiments to verify the effectiveness of the proposed algorithm. We conclude the study in Section V.

II. SC

For a given dataset $X = \{x_1, \cdots, x_N\}$, where N is the size of X, SC algorithm treats clustering thing as a partition of an undirected graph $G = (X, S)$[10]. In SC, each data point x_i is considered as a vertex in an undirected graph G. The similarity s_{ij} between x_i and x_j is the weight of the edge connecting the two vertices which forms similarity matrix $S = \{s_{ij} | i = 1, \cdots, N, j = 1, \cdots, N\}$. Then, an optimal graph cut criterion is used to divide the graph G into k disjoint subsets $C_1, \ldots, C_k$ so that data points in the same subset are as similar as possible and in different subsets are as dissimilar as possible.

A popular cut criterion is a normalized cut, which introduces the notion of volume to normalize the correlation between subsets to measure the similarities between inter- and intra-subsets simultaneously[26]. Its constraint function is shown in Equation (1).

$$NCut(C_1, C_2) = \frac{Cut(C_1, C_2)}{Vol(C_1)} + \frac{Cut(C_1, C_2)}{Vol(C_2)} \quad (1)$$

where $Cut(C_1, C_2)$ equals to the sum of weights within C_1 and C_2, and $Vol(C_1) = Cut(C_1, G)$. $Cut(C_1, C_2)$ is computed as follows.

$$Cut(C_1, C_2) = \sum_{x_i \in C_1, x_j \in C_2} S_{ij} \quad (2)$$

The normalized cut criterion can only cut an undirected graph into two clusters, so the multiple normalized cut criterion is a preferred choice when a graph has multiple clusters[27]. Equation (3) defines the constraint function of multiple normalized cuts.

$$MNCut(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{Cut(C_i, G - C_i)}{Vol(C_i)} \quad (3)$$

However, the optimization of the multiple normalized cut is an NP-hard problem, and an alternative way is to obtain its approximate solution in the relaxed continuous real domain. Gu et al. indicated that the relaxed spectral solution of multiple normalized cut is located in the subspace spanned by the eigenvectors of the Laplacian matrix, which corresponds to the top maximum eigenvalues[28]. Therefore, the key process of SC is to resolve the eigenvectors and eigenvalues of the Laplacian matrix, which is shown in Equation (4)[29]:

$$L = D^{-1/2}(D - S)D^{-1/2} = I - D^{-1/2}SD^{-1/2} \quad (4)$$

where D is a diagonal matrix, and each diagonal element d_{ii} in D is the sum of all elements of the ith row in matrix S. A low-dimensional space S_k can be formed with the eigenvectors of matrix L; then, S_k is clustered to obtain the final clustering result by using a classical clustering method.

The time complexity for the decomposition of matrix L is $O(N^3)$, and such high computational complexity limits SC's application in big data. When L is a sparse matrix, the Lanczos method can be utilized to resolve the eigenvectors and thus increase the decomposition speed, whereas its effectiveness has yet to be verified[26]. In 2004, Fowlkes proposed a spectral grouping using the Nyström extension method, which just utilized a small sample dataset to obtain a low-rank approximation of the eigenvectors of the Laplacian matrix to reduce computational complexity[19].

III. SCOSS

A. Nyström extension

The Nyström extension method is an approximation technique that is commonly used to solve eigenfunction problems[30–32]. Supposing a subset has n samples, the remaining dataset's size is $m = N - n$, and the similarity matrix S can be defined as Equation (5) when the Nyström extension is applied to SC.

$$S = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \quad (5)$$

where $A \in R^{n \times n} (A = U \wedge U^T)$ represents the similarity matrix of the sampled subset, $B \in R^{m \times n}$ represents the similarity matrix between the sampled and the remaining, and $C \in R^{m \times n}$ is of the remaining subset.

Under the assumption that $\bar{U}$ approximates the eigenvectors of matrix S, it can be resolved by using the Nyström extension as described in Equation (6).

$$\bar{U} = \begin{bmatrix} A \\ B^T U \wedge^{-1} \end{bmatrix} \quad (6)$$

If $\hat{S}$ is the approximation matrix of S, then $\hat{S}$ can be written as follows:

$$\begin{aligned} \hat{S} = \bar{U} \wedge \bar{U}^T &= \begin{bmatrix} U \\ B^T U \wedge^{-1} \end{bmatrix} \wedge \begin{bmatrix} U^T \wedge^{-1} U^T B \end{bmatrix} \\ &= \begin{bmatrix} U \wedge U^T & B \\ B^T & B^T A^{-1} B \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} \end{aligned} \quad (7)$$

As shown in Equation (7), the Nyström method uses matrix $B^T A^{-1} B$ to replace C and thus reduce computational complexity.

Generally, matrix $\bar{U}$ cannot be used directly because it does not satisfy the orthogonality of eigenvectors. Still, the following step is a satisfactory method to solve the diagonalization of matrix $\hat{S}$. Assume A is a positive matrix and $A^{1/2}$ represents the symmetric square roots of the semi-positive matrices of A, P is defined as $P = A + A^{1/2} B B^T A^{-1/2}$. Furthermore, P can also be written in its diagonal form as $P = U_P \wedge_P U_P^T$, and matrices V and $\hat{S}$ can be written as Equations (8) and (9), respectively. In this case, V is the

eigenvector of $\hat{S}$, and V^T & V are a pair of orthogonal vectors because $V^TV = I$.

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1/2} U_P \wedge_P^{-1/2} \tag{8}$$

$$\begin{aligned} \hat{S} &= \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1} \begin{bmatrix} A & B \end{bmatrix} \\ &= \left\{ \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1/2} U_P \wedge_P^{-1/2} \right\} \wedge_p \\ &\left\{ \wedge_P^{-1/2} U_P^T A^{-1/2} \begin{bmatrix} A & B \end{bmatrix} \right\}^T \\ &= V \wedge_p V \end{aligned} \tag{9}$$

Before clustering in SC, the similarity matrix should be normalized to obtain an ideal result when using the Nyström method. The matrix $\hat{d}$ was proposed in [19] to normalize matrices A and B, and $\hat{d}$ could be calculated as in Equation (10).

$$\begin{aligned} \hat{d} = \hat{S}1 &= \begin{bmatrix} UA1_n + B1_m \\ B^T 1_n + B^T A^{-1} B 1_m \end{bmatrix} \\ &= \begin{bmatrix} a_r + b_r \\ b_c + B^T A^{-1} b_r \end{bmatrix} \end{aligned} \tag{10}$$

where 1 is a vector in which every element is 1, $a_r, b_r \in R^m$ is the sum of the rows in matrices A and B, and $b_c \in R^n$ represents the sum of the columns in matrix B. Hence, matrices A and B can be normalized using Equations (11) and (12), respectively.

$$A_{ij} \leftarrow \frac{A_{ij}}{\sqrt{\hat{d}_i \hat{d}_j}}, i, j = 1, \cdots n \tag{11}$$

$$B_{ij} \leftarrow \frac{B_{ij}}{\sqrt{\hat{d}_i \hat{d}_{j+n}}}, i = 1, \cdots n, j = 1, \cdots m \tag{12}$$

B. OptiSim Selection Method

The OptiSim selection algorithm is a novel sampling method that can flexibly control the balance between the samples representativeness and diversity by defining the size of subset K[33]. The smaller the K value, the better the representativeness of the sample set. On the contrary, the larger the K value, the better the diversity of the sample set.

Four data structures, namely, sample dataset M, subsample dataset M', candidate buffer pool C, and recycling station R, are defined in the OptiSim algorithm. M stores a subset sampled by OptiSim, and its size is n. M' stores a subsampled set whose size is K, and the similarity between each data point in M' and M is lower than a given threshold θ. Candidates that can be selected in M' are placed in C. Data in M' that cannot be sampled in M are moved into R.

OptiSim is an iterative algorithm, except the first sample is randomly selected from the dataset X, all the others in M are iteratively selected from M'. Each iteration begins with the establishment of a subsample set M', which contains K objects chosen from the candidate buffer pool C. The pairwise similarity of data points between datasets M and M' must be less than the threshold value θ. When data points in C are insufficient, data in R are moved into C. After the establishment, only the data point with the minimum similarity to the data in M can be selected as a sample, and the remaining data M' are moved into R. The sampling process of OptiSim is described in Algorithm 1.

***Algorithm 1**. OptiSim selection algorithm*

Input: dataset X, size of sample dataset n, size of subsample dataset K, and threshold θ.

Output: indices of samples.
1. Create four empty data structures: sample dataset M, subsample dataset M', candidate buffer pool C, and recycling station R.
2. Randomly select a data point from X to be placed into M and remove the remaining to C.
3. Randomly select a data point x from C. If the similarity between x and each data point in M is lower than a given threshold θ, then x is selected into M'; otherwise, it is moved into R.
4. Repeat Step 3 until K data points in M' or C is empty.
5. If C is empty and the size of M' is less than K, then remove all the data in R to C, and return to Step 3.
6. If M' is empty, then exit.
7. Find the data point in M' that has the minimum similarity to the data in M and place it into M.
8. Move all the other unselected data points in M' into R.
9. If the size of M equals n, then exit; otherwise, return to Step 3.

As shown in Algorithm 1, the computational complexity in OptiSim is $O(Kn)$. In the sampling process, OptiSim requires that the pairwise similarity among each object in M' and M is less than the threshold θ and that each sample obtained must have the minimum similarity to the data in M. These steps ensure the representativeness of the samples to a certain extent. Meanwhile, the setting of K can further guarantee the representativeness of the sample.

C. SCOSS algorithm

The SCOSS algorithm uses Algorithm 1 to obtain adequate samples, which are used in approximating similarity matrix S to get the approximate matrix $\hat{S}$. A low-dimensional embedding space $Y \in R^{N \times k}$is formed by the eigenvectors of $\hat{S}$ whose eigenvalues are the top k largest. Finally, Y is to be clustered by using the k-means algorithm. The SCOSS algorithm is described in Algorithm 2.

***Algorithm 2**. SCOSS algorithm*

Input: dataset X, size of samples n, and the number of clusters k.

Output: k clusters.
1. Compute similarity matrix S of X: $S \in R^{N \times N}$ and each element $s_{ij} \in S$ represents the similarity between x_i and x_j.
2. Use Algorithm 1 to obtain a sample dataset whose size is n. Then, compute $A \in R^{n \times n}$ and $B \in R^{(N-n) \times N}$, which represents the similarity matrix among samples and between samples and the remaining, respectively.
3. Calculate matrix $\hat{d}$ with Equation (10). Then, normalize matrices A and B by Equations (11) and (12), respectively.

4. Apply normalized matrices A and B to compute matrix P: $P = A + A^{1/2}BB^TA^{-1/2}$.
5. Diagonalize matrix P, and let $P = U_P \wedge_p U_P^T$.
6. Resolve orthogonal eigenvector V by adding $U_P, \wedge_P$ *in Equation (8).*
7. Use the eigenvectors of $\hat{S}$ *with the top k largest eigenvalue to form matrix* $W : W \in R^{N\times k}$.
8. Normalize each row of W to obtain matrix $Y : y_{ij} = w_{ij}/\sqrt{\sum_{j=1}^{k} w_{ij}^2}$, *and* $Y \in R^{N\times k}$.
9. Use k-means algorithm to cluster Y, and the ith row of Y corresponds to data point x_i *in X. If the ith row of Y is clustered in* C_j, *then* x_i *belongs to the jth cluster.*

D. Efficiency analysis of SCOSS algorithm

The time complexity of SCOSS is first analyzed, then the proof that SCOSS can achieve good clustering quality is deduced.

In Algorithm 2, Step 1 computes each pair similarity of N data points and the time complexity is $O(N^2)$. In Step 2, n samples are selected using Algorithm 1, matrices A and B are calculated, and the complexity is $O(Kn^2 + mn)$. The required time for computing $\hat{d}$ and normalizing A and B in Step 3 is $O(mn)$ and $O(n^2) + O(mn)$, respectively. Given $n \ll m$, the complexity of Step3 is $O(mn)$. Steps 4-6 use the Nyström extension method to approach the eigenvectors of $\hat{s}$, and the complexity is $O(n^3)$. Steps 7 and 8 require the calculations of $O(k)$ and $O(kN)$ times, respectively. In Step 9, the time complexity is $O(kNt)$, where t is the time of iteration in the k-means algorithm. Specifically, the entire time complexity of SCOSS is $O(n^3)$. However, the time complexity of SC is $O(N^3)$. By comparing the complexities of these two algorithms, SCOSS can achieve acceleration against the original SC method.

In SCOSS, the process of OptiSim selection can be viewed as getting samples from the $(n-1)$ subsampled dataset M', and the construction of each subsample dataset M' is equivalent to a traverse of dataset X. The $(n-1)$ samples in M are the results of a traverse of the union dataset $T = M_1' \cup M_2' \cdots M_{n-1}'$. First to prove that with the scale of T increases, that is with the increase of the sampling number, sampling error will be reduced. Theorem 1 was demonstrated in the literature [19].

Theorem 1. For a given matrix $A \in R^{m\times n}$, $Z \subseteq R^n$ is an vector subspace, $\pi_k(A)$ is the best Rank-k approximation of A, and $\pi_k(A) = \pi_{R^n,k}(A)$. M' is a subsample dataset, then have

$$E_s(\|A - \pi_{Z+span(M'),k}(A)\|_F^2) \leq \|A - \pi_k(A)\|_F^2 + \varepsilon\|E\|_F^2 \tag{13}$$

Theorem 1 is right for a subsample dataset M'. Defining $E = A - \pi_{M_1' \cup \cdots \cup M_{n-1}'}(A)$, and for the union T, Formula (14) can also be established.

$$E_T(\|A - \pi_{M_1' \cup \cdots \cup M_{n-1}',k}(A)\|_F^2) \leq \|A - \pi_k(A)\|_F^2 + \varepsilon\|E\|_F^2 \tag{14}$$

Also owing to

$$\|E\|_F^2 \leq \|A - \pi_{M_1' \cup \cdots \cup M_{n-2}',k}(A)\|_F^2 \tag{15}$$

Replace the Formula (15) into the Formula (14)

$$\begin{aligned}
&E_T(\|A - \pi_{M_1' \cup \cdots \cup M_{n-1}',k}(A)\|_F^2) \\
&\leq \|A - \pi_k(A)\|_F^2 + \varepsilon\|A - \pi_{M_1' \cup \cdots \cup M_{n-2}',k}(A)\|_F^2 \\
&\leq \|A - \pi_k(A)\|_F^2 + \varepsilon E_{M_1' \cup \cdots \cup M_{n-2}'} \\
&(\|A - \pi_{M_1' \cup \cdots \cup M_{n-2}',k}(A)\|_F^2) \\
&\leq \|A - \pi_k(A)\|_F^2 + \varepsilon(\frac{1}{1-\varepsilon}\|A - \pi(A)\|_F^2 + \varepsilon^{n-2}\|A\|_F^2) \\
&= \frac{1}{1-\varepsilon}\|A - \pi_k(A)\|_F^2 + \varepsilon^{n-1}\|A\|_F^2
\end{aligned} \tag{16}$$

In Formula (16), the first item is a constant. The second will decrease with the increase of n, that is, the sampling error will become smaller with the rise of the size of samples.

Therefore, the definition of the subsample dataset and selection of the smallest similarity to the selected samples in the OptiSim method ensure that it can get a better sample dataset.

TABLE I
THE UCI DATASETS

Dataset	Instances	Attributes	Clusters
Breast	699	8	2
banknote	1372	4	2
Steel	1941	27	7
Imageseg	2100	19	7
Wilt	4839	4	2
pageblocks	5473	10	5

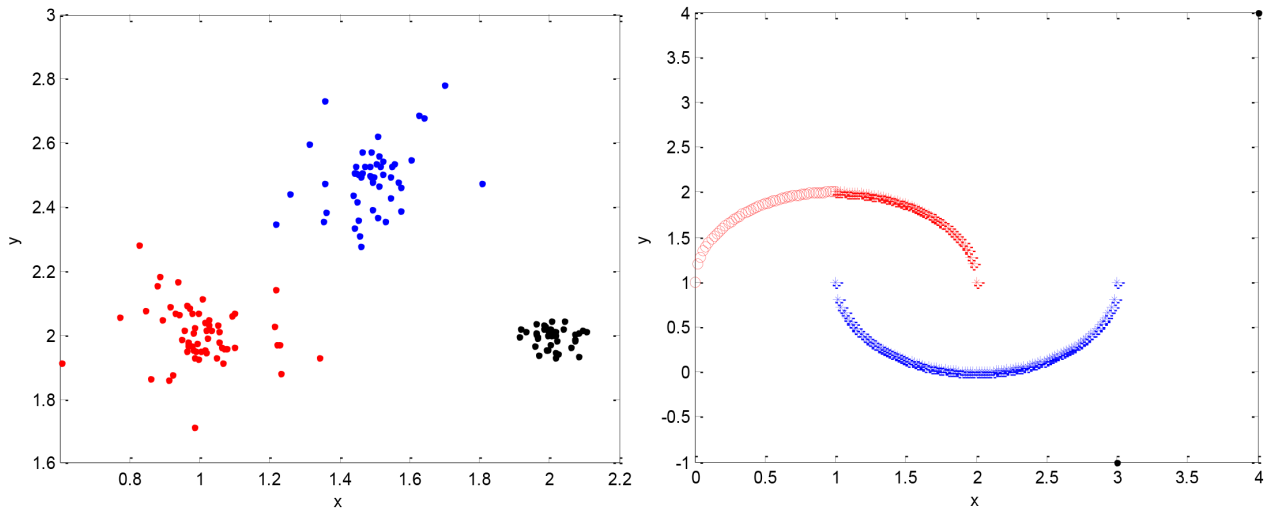
IV. EXPERIMENTS

We condeucted experiments on two synthetic and six UCI datasets to verify whether SCOSS could accelerate the SC algorithms speed. The synthetic datasets include a nonspherical (Test 1) and spherical set (Test 2), as shown in Figure 1. The UCI datasets used in the experiments are shown in Table 1. All the UCI datasets were normalized before running the experiments. In dataset Breast, the column where some values are missing was deleted. The proposed algorithm was compared with three other algorithms, namely, SC, SC based on random sampling (SRS)[19], and SC based on Maximum dissimilarity Sampling (SMDS)[25]. Experiments were conducted with MATLAB 2014a; the processor was Intel (R) Core (TM) i5-3210m, 2.5 GHz, and the memory size was 4 GB.

Under the assumption that $C = \{c_1, \cdots, c_k\}$ and $S = \{s_1, \cdots, s_t\}$ are the experimental and actual clustering results, k and t are the number of clusters in C and S, respectively. $N_{i,j}$is defined as the number of objects jointly contained in C_i and S_j, N_i^c and N_j^s count the number of objects in c_i and s_j, respectively. Indicators, namely, Clustering Accuracy (CA) and Normalized Mutual Information (NMI), were used to evaluate the clustering quality. CA compares the real label of each data point with its experimental label and is defined as Equation (17):

$$CA = \frac{1}{N}\sum_{i=1}^{k} max_j^t N_{i,j} \tag{17}$$

NMI utilizes external information to evaluate the clustering effect, and it can be calculated as Equation (18):



(a) Test1 (b)Test2

Fig. 1. The synthetic datasets. Test1 is generated by the points whose cluster centers are (1,2),(1.6,2.5) and (2,2). Test 2 is made up of two spherical sudsets.

TABLE II
THE COMPARISONS OF CA VALUES(%)

Sampling ration(%)	Algorithm	Datasets							
		Test1	Test2	Breast	Banknote	Steel	Imageseg	Wilt	Pageblocks
100	**Spectral**	**99.8**	**95.3**	**95.1**	**62.6**	**34.5**	**65.7**	**51.1**	**38.1**
1	SRS	55.1	61.5	81.2	48.4	31.4	53.4	48.5	23.1
	SMDS	54.3	65.0	72.4	52.1	32.0	54.6	**49.3**	24.5
	SCOSS	**62.7**	**68.2**	**81.9**	**57.2**	**32.2**	**59.3**	48.7	**26.2**
3	SRS	58.2	63.1	80.1	50.4	30.5	54.7	48.8	24.1
	SMDS	64.7	66.7	75.5	52.2	32.2	55.2	50.3	24.9
	SCOSS	**77.2**	**72.6**	**84.6**	**59.6**	**33.4**	**61.5**	**51.3**	**27.5**
5	SRS	75.0	68.7	80.7	53.6	32.5	55.8	49.2	24.6
	SMDS	72.1	71.2	80.3	55.7	32.3	56.2	50.6	25.7
	SCOSS	**89.5**	**84.1**	**86.2**	**61.5**	**34.2**	**62.7**	**52.0**	**33.6**
10	SRS	79.7	73.4	85.7	54.6	32.9	56.3	50.2	26.3
	SMDS	81.6	78.8	82.4	56.9	33.5	56.5	51.5	26.7
	SCOSS	**91.3**	**88.9**	**92.3**	**61.9**	**34.6**	**63.3**	**52.1**	**35.7**
15	SRS	84.5	82.9	85.6	54.9	33.8	57.9	50.3	28.3
	SMDS	85.4	85.1	83.2	57.8	33.6	58.2	51.9	28.9
	SCOSS	**96.7**	**92.4**	**93.0**	**62.3**	**34.9**	**64.6**	**52.3**	**36.9**

NOTES:Table2 provides the comparisons of Clustering Accuracy(CA) values among Spectral Clustering(SC), Sc based on Random Sampling(SRS), Sc based on Maximum Dissimilarity Sampling(SMDS) and the proposed algorithm (SCOSS) methods in synthetic datasets and UCI datasets. The values of SC method are first provided, then followed with the other three sampling schemes . The largest CA value is shown in bold. As shown in Table2, SCOSS obtains the highest CA values in almost all datasets for each given sampling ratio.

TABLE III
THE COMPARISONS OF NMI VALUES(%)

Sampling ratio(%)	algorithm	Datasets							
		Test1	Test2	Breast	Banknote	Steel	Imageseg	Wilt	Pageblocks
100	**Spectral**	**99.3**	**94.7**	**96.9**	**75.7**	**19.4**	**62.8**	**55.6**	**27.1**
1	SRS	45.1	46.5	72.1	53.8	9.1	47.5	46.3	20.6
	SMDS	54.3	58.4	71.2	56.5	10.3	50.6	49.2	21.8
	SCOSS	**68.4**	**64.2**	**76.4**	**62.9**	**12.6**	**53.2**	**52.0**	**23.5**
3	SRS	54.2	62.9	72.9	54.9	11.1	48.4	46.9	21.4
	SMDS	55.3	63.2	73.5	58.9	11.7	51.2	50.2	21.9
	SCOSS	**72.7**	**69.9**	**81.8**	**64.1**	**15.9**	**54.7**	**52.3**	**23.7**
5	SRS	65.6	66.5	73.5	58.4	12.2	50.2	47.1	22.7
	SMDS	64.1	64.3	76.9	60.4	13.5	53.1	51.7	22.1
	SCOSS	**84.6**	**81.1**	**86.3**	**66.8**	**17.8**	**56.2**	**52.8**	**24.2**
10	SRS	82.1	74.3	73.6	60.7	13.4	52.8	51.3	23.2
	SMDS	83.0	76.6	81.4	62.8	15.7	54.9	52.1	23.5
	SCOSS	**91.3**	**85.8**	**90.2**	**72.9**	**17.5**	**58.9**	**53.8**	**25.8**
15	SRS	86.7	81.5	81.9	61.2	14.0	54.6	52.0	23.3
	SMDS	88.9	85.9	84.7	64.8	15.9	55.2	52.9	23.7
	SCOSS	**95.1**	**90.4**	**92.1**	**73.6**	**17.9**	**60.1**	**54.2**	**26.1**

NOTES:Table3 provides the comparisons of Normalized Mutual Information (NMI) values among Spectral Clustering(SC), Sc based on Random Sampling(SRS), Sc based on Maximum Dissimilarity Sampling(SMDS) and the proposed algorithm (SCOSS) methods in synthetic datasets and UCI datasets. The values of SC method are also first provided, then followed with the other three sampling schemes. The largest NMI is shown in bold too. As shown in Table3, SCOSS performs better than SRS and SMDS.

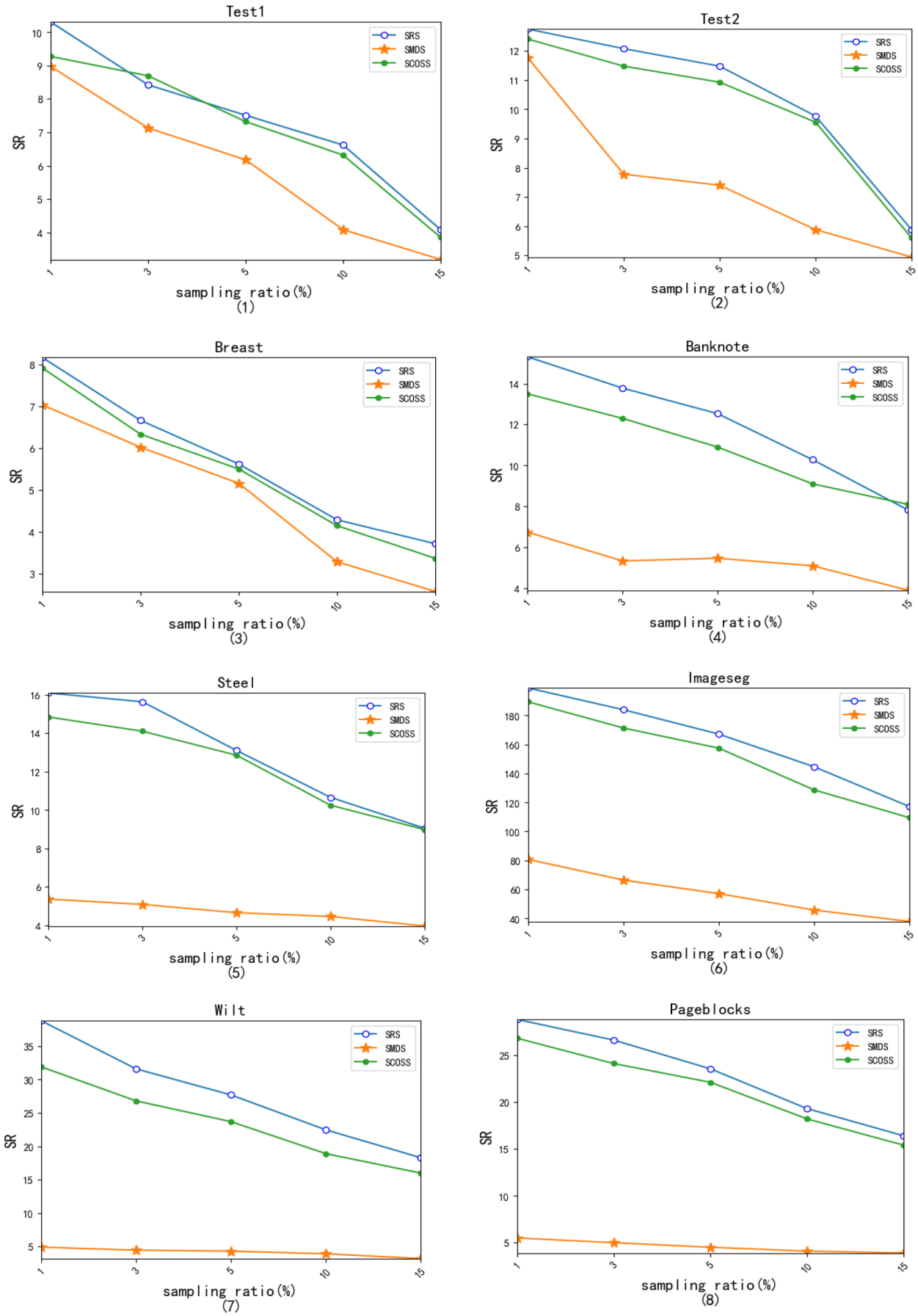


Fig. 2. The comparisons of SR values. NOTES: Figure 2 presents the comparisons of Speedup Ratio (SR) among Sc based on Random Sampling(SRS), Sc based on Maximum Dissimilarity Sampling(SMDS) and the proposed algorithm (SCOSS) methods in synthetic datasets and UCI datasets. The x-axis stands for sampling ratio and the y-axis stands for SR value. Each subgraph reveals a comparison on a certain dataset. The blue line shows the trend of SR with the increase of sampling ration for SRS, the green for SCOSS, and the orange for SMDS.

TABLE IV
THE COMPARISONS OF CT VALUES(S)

Sampling ration(%)	Algorithm	Datasets							
		Test1	Test2	Breast	Banknote	Steel	Imageseg	Wilt	Pageblocks
100	**Spectral**	27.8	45.9	25.3	68.9	325	5131	692	1516
1	SRS	**2.7**	**3.6**	**3.1**	**4.5**	**20.2**	**25.8**	**17.8**	**52.6**
	SMDS	3.1	3.9	3.6	10.2	60.5	63.7	142	278
	SCOSS	3.0	3.7	3.2	5.1	21.9	27.1	21.7	56.5
3	SRS	**3.3**	**3.8**	**3.8**	**5.0**	**20.8**	**27.9**	**21.9**	**57.8**
	SMDS	3.9	5.9	4.2	12.9	63.9	77.6	156	305
	SCOSS	3.2	4.0	4.0	5.6	23.1	30.0	25.8	62.9
5	SRS	**3.7**	**4.0**	**4.5**	**5.5**	**24.8**	**30.7**	**25.0**	**65.2**
	SMDS	4.5	6.2	4.9	12.6	69.8	88.9	160	336
	SCOSS	3.8	4.2	4.6	6.3	25.3	32.6	29.2	68.7
10	SRS	**4.2**	**4.7**	**5.9**	**6.7**	**30.5**	**35.5**	**30.8**	**78.7**
	SMDS	6.8	7.8	7.7	13.5	72.8	112.3	178	375
	SCOSS	4.4	4.8	6.1	7.6	31.7	39.9	36.7	83.4
15	SRS	**6.8**	**7.8**	**6.8**	**8.8**	**35.9**	**43.8**	**37.9**	**92.4**
	SMDS	8.7	9.3	9.8	17.6	81.9	135.7	215	392
	SCOSS	7.2	8.2	7.5	8.5	36.2	46.9	43.3	98.2

NOTES:Table4 provides the comparisons of Cluster Time (CT) among Spectral Clustering(SC), Sc based on Random Sampling(SRS), Sc based on Maximum Dissimilarity Sampling(SMDS) and the proposed algorithm (SCOSS) methods in synthetic datasets and UCI datasets. The values of SC method are also first provided , then followed with the other three sampling schemes'. The least CT is shown in bold. It can be get that the three sampling methods only uses less time than SC, and SRS method has the minimum CT in every case. The CTs of SCOSS are almost the same as SRS's.

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{t} N_{i,j} log\frac{N \cdot N_{i,j}}{N_i^c \cdot N_j^s}}{\sqrt{\sum_{i=1}^{k} N_i^c \cdot log\frac{N_i^c}{N} \cdot \sum_{j=1}^{t} N_j^s \cdot log\frac{N_j^s}{N}}} \tag{18}$$

The larger the value of CA and NMI, the better the clustering quality; the maximum value of both indicators is 1. Indicators, namely, Cluster Time (CT) and Speedup Ratio (SR), were used to evaluate the clustering speed. CT is the time used for clustering. SR is the clustering time ratio of SC to the sampling-based SC clustering methods, and it is defined as Equation (19):

$$SR = \frac{T_{SC}}{T_{SSC}} \tag{19}$$

where T_{SC} is the time used for the spectral method and T_{SSC} the time used for the sampling-based SC method.

In all the other methods except for SC methods, the sampling ratios were set as R=1%, 3%, 5%,10%and 15%.The parameters of minimum similarity threshold in SCOSS and SMDS were all assigned as $\theta = 0.7{*}maxdistance$, and the K value in SCOSS was set as $K = 0.05{*}N$. Every clustering method was run ten times to obtain the average clustering effect.

A. Analysis of the CA indicator

The comparisons of CA values among the four clustering methods are shown in Table 2. The results of SC method are first provided for the convenience of comparison in Table 2, then followed with the values of three sampling schemes (The same is true in the following comparisons in NMI and CT indicators). The largest CA value is shown in bold.

The results in Table 2 show that the quality of SC method is the best, and it is very intelligible because SC utilizes the whole dataset while the other three ways only use some samples for clustering. But the superiority of SCOSS is evident among these three sampling methods. It obtains the highest CA values in almost all datasets for each given sampling ratio. Meanwhile, the quality of SCOSS method is getting better with the increase of sampling proportion, and the CA value is also closer to the original SC algorithm.

B. Analysis of the NMI indicator

The comparisons of NMI values are presented in Table 3. The largest CA value is also shown in bold. SC method also achieves the maximum value in terms of the NMI indicator, which is the same as in CA. Similarly, the proposed SCOSS algorithm also obtains better clustering results in all datasets regardless of any sampling ratio than SRS and SMDS methods. Moreover, the NMI value of SCOSS is also getting higher as the number of samples increases and closer to that of the SC method.

SCOSS outperforms because it utilizes OptiSim selection to sample data object, which not only takes advantage of the dissimilarity between data objects but also avoids too many outliers to be selected in samples. Furthermore, the clustering effect of SRS is inferior and unstable because it only randomly extracts some samples from the dataset. SMDS performs better than SRS but worse than SCOSS just because it makes the best of the max dissimilarity during sampling but easily extracts too many outliers meanwhile.

C. Analysis of the CT indicator

The comparisons of CT values are shown in Table 4, where the least CT is shown in bold. Table 4 reveals that no matter in which dataset SC method always takes the longest time, and this is because its complexity is $O(N^3)$. In the meantime, it can also be get from Table 4 that the three sampling methods only uses less time, that is they all could achieve the goal of speedup.

Furthermore, SRS method has the minimum CT in each sampling ration case of any dataset among the three algorithms, and the CT values of SCOSS are almost the same as SRS's. This finding can be attributed to the required time of SRS for sampling being $O(n)$ and the time complexities of SCOSS and SMDS for sampling respectively being $O(Kn)$ and $O(N^2)$.

D. Analysis of the SR indicator

The level of speedup of sampling methods against SC can be reflected by SR indicator, and Figure 2 displays these comparisons in 2D coordinate maps, where the x-axis stands for sampling ratio and the y-axis stands for SR value. There are eight subgraphs in Figure 2, and each subgraph reveals the SR values of the three sampling methods on a certain dataset. The blue line in each subgraph shows the trend of SR with the increase of sampling ration for SRS, the green for SCOSS, and the orange for SMDS.

As shown in Figure 2, the proposed SCOSS has almost the same SR values as the SRS's in whatever dataset, which means they have nearly the same acceleration effect. Whereas the SMDS has the worst speedup effect, and it is due to taking more time in the process of sampling. On the other hand, all the SR values decrease with the increase of sampling proportion in each dataset, and it is mainly owing to the increase of sample size. Furthermore, it can also get that the larger the scale of the dataset, the better the acceleration effect regardless of any sampling ration and any sampling method, and this also implies that the sampling method is more available for large datasets.

V. CONCLUSION

The Nyström extension method can improve the SC algorithm's speed by selecting some samples to make an approximate calculation; however, the quality of clustering depends heavily on the representativeness of the selected samples. In the proposed SCOSS algorithm, an OptiSim sampling method is utilized to extract some good representatives, and the Nyström method is used to approximate the clustering. Theoretical analysis and experiments show that the proposed algorithm has a novel acceleration effect and can achieve a better clustering quality than the other two sampling methods.

REFERENCES

[1] D.Gracia and S.Sudha, "Adaptive Clustering of Embedded Multiple Web Objects for Efficient Group Prefetching", *Arabian Journal for Science and Engineering*, vol.42, no.2, pp715-724, 2017.

[2] Jie-MingYang, Zhi-Ying Liu and Zhao-Yang Qu, "Clustering of Words Based on Relative Contribution for Text Categorization," *IAENG International Journal of Computer Science*, vol.40, no.3, pp207-219, 2013.

[3] Tao Ma, Heng Liu, and Yu Zhang, "A Method for Establishing Tropospheric Atmospheric Refractivity Profile Model Based on Multiquadric RBF and k-means Clustering," *Engineering Letters*, vol. 28, no.3, pp733-741, 2020.

[4] A. K.Jain and R.C.Dubes , "Algorithms for clustering data," Prentice-Hall, Englewood Cliffs, New Jersey,1988, pp45-46,1988.

[5] X.Wu , V.Kumar, J. R.Quinlan, et al, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol.14, no.1, pp1-37, 2007.

[6] P.Salgado and P.Garrido, "Fuzzy Clustering of Fuzzy Systems," *IEEE International conference on systems man and cybernetics*, pp2368 - 2373, 2004.

[7] M.Ester, H.Kriegel, J.Sander, and X.Xu, , "A density-based algorithm for discovering clusters in large spatial databases with noise," In Proc. 2th Kdd Conf, vol. 96, no.34: pp226-231,1996.

[8] D. Xu and Y.Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol.2, no.2, pp165-193, 2015.

[9] A.Y.Ng, M.I. Jordan and Y.Weiss, "On spectral clustering: analysis and an algorithm," *Advances in Neural Information Processing Systems*, USA, MIT Press, pp849-856, 2002.

[10] F.R.KChung , Spectral Graph Theory, Am. Math, Soc. 1997.

[11] Y.Song, W.Chen, H.Bai, C.Lin, and E. Y.Chang, "Parallel Spectral Clustering," *Lecture Notes in Computer Science*, vol. 5212, pp.374-389, 2008.

[12] D.Yan, L.Huang, and M.I.Jordan, "Fast approximate spectral clustering," In Proc. 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009, pp907-916, 2009.

[13] F. Lin and W.W.Cohen , "Power iteration clustering," In Proc. 27th international conference on machine learning, ICML-10, 2010, pp655-662, 2010.

[14] K.Frederix and M.Van Barel, "Sparse spectral clustering method based on the incomplete Cholesky decomposition," *Journal of Computational and Applied Mathematics*, vol.237, no.1, pp 145-161, 2013.

[15] M.Novak, C.Alzate and R.Langone, "Fast kernel spectral clustering based on incomplete Cholesky factorization for large scale data analysis," *Internal Report 14-119[R], ESAT-SISTA:KU Leuven*, 2014.

[16] R. Langone, M. Van Barel and J. A. K. Suykens, "Entropy-based incomplete cholesky decomposition for a scalable spectral clustering algorithm: Computational studies and sensitivity analysis," *Entropy* ,vol.18, no.5, pp182, 2016.

[17] K.Taşdemir, "Vector quantization based approximate spectral clustering of large datasets," *Pattern Recognition*, vol.45, no.8, pp 3034-3044, 2012.

[18] L.He, R.Nilanjan, Y.Guan , et al"Fast large-scale spectral clustering via explicit feature mapping," *IEEE Transactions on Systems, Man, and Cybernetics*,vol.49, no.3, pp1058-1071, 2019.

[19] C. C.Fowlkes, S.Belongie, F.Chung, and J.Malik, "Spectral grouping using the Nystrom method," *IEEE transactions on pattern analysis and machine intelligence*, vol.26, no.2, pp 214-225, 2004.

[20] K.Zhang and J. TKwok , "Clustered Nyström method for large scale manifold learning and dimension reduction," *IEEE Transactions on Neural Networks*, vol.21, no.10, pp1576-1587, 2010.

[21] L.Wang, J. C.Bezdek, C.Leckie, and R.Kotagiri, "Selective sampling for approximate clustering of very large data sets," *International Journal of Intelligent Systems*, vol.23,no.3, pp 313-331, 2008.

[22] S.Kumar, M.Mohri and A.Talwalkar , "Sampling methods for the Nyström method," *Journal of Machine Learning Research*, vol.13, pp 981-1006, 2012.

[23] L.He, H.Zhu, T.Zhang, et al, "Projected affinity values for Nyström spectral clustering," *Entropy*, vol.20, no.7, pp519-535, 2018.

[24] J.Liu, C.Wang, M.Danilevsky, and J.Han, "Large-scale spectral clustering on graphs," In Proc. 23th on Artificial Intelligence, AAAI Press, 2013, pp1486-1492.

[25] Q.Zhan and Y.Mao, "Improved spectral clustering based on Nyström method," *Multimedia Tools and Applications*, vol.76, pp20149-20165, 2017.

[26] J.Shi and J.Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp888-905, 2000.

[27] M.Meila and M.Xu, "Multiway cuts and spectral clustering". U. Washington Tech Report. 2003.

[28] M.GU, et al. "Spectral relaxation models and structure analysis for k-way graph Clustering and bi-clustering," Penn. State Univ. Tech. Report CSE-01-007, 2001.

[29] M.Filippone, F.Camastra, F.Masulli and S.Rovetta , "A survey of kernel and spectral methods for clustering," *Pattern Recognition*, vol.41, pp 176-190, 2008.

[30] "Nyström E.J. Über die Praktische Auflösung von Linearen Integralgleichungen mit Anwendungen auf Randwertaufgaben der Potentialtheorie," *Commentationes Physico-Mathematicae*. vol .4, no.15, pp1-52,1928.

[31] S. N. Jator, and E. O. Adeyefa, "Direct Integration of Fourth Order Initial and Boundary Value Problems using Nystrom Type Methods," *IAENG International Journal of Applied Mathematics*, vol. 49, no.4, pp638-649, 2019.

[32] W. H.Press, B.Flannery, S.Teukolsky , et al. Numerical Recipies in C, Cambridge, 1995.

[33] R. D.Clark, "OptiSim: an extended dissimilarity selection method for finding diverse representative subsets ," *Journal of Chemical Information and Computer Sciences*, vol.37, no.6, pp1181-1188, 1997.