Hubness-based Sampling Method for Nyström Spectral Clustering

Hongmin Li, Xiucai Ye, Akira Imakura and Tetsuya Sakurai
Department of Computer Science
University of Tsukuba
Tsukuba, Japan
li.hongmin.xa@alumni.tsukuba.ac.jp, {yexiucai, imakura, sakurai}@cs.tsukuba.ac.jp

***Abstract*—Nyström method is widely used for spectral clustering to obtain low-rank approximations of a large matrix. Sampling is crucial to Nyström method, since selecting the representative sample points that can reflect the data structure is important for obtaining good approximation results. To improve the performance of Nyström based spectral clustering, in this paper, we propose a new sampling method by considering the hubness score of sample points. The data points with the high hubness scores, i.e., appearing frequently in the nearest neighbor lists of other data points, have high probabilities to be selected as the sample points. Taking advantage of the topological property of hubs (i.e., data points with high hubness score), the selected sampling points have close relationships with other data points, thus the proposed method is able to achieve scalable and accurate clustering results. We further design fast computation methods, i.e., local hubness approximated methods, to speed up the sampling process. Experimental results on both synthetic and real-world data sets show that the proposed method not only achieves good performance, but also outperforms other sampling methods for Nyström based spectral clustering.**

***Index Terms*—Spectral clustering, Sampling methods, Nyström method, Hubness score**

I. Introduction

Clustering is one of the fundamental problems in machine learning fields [1]. The objective of clustering is to divide unlabeled data into some groups such that the data in the same group are more similar than those in other groups [2]. Spectral clustering is a powerful clustering method, which has superior performance compared to the traditional clustering methods such as k-means [3], [4]. However, the applicability of spectral clustering is limited when the number of data points becomes large [5]. For large-scale data sets, spectral clustering has two bottlenecks, i.e., constructing a large similarity matrix and calculating the eigen-decomposition of the corresponding Laplacian matrix. The computational complexities in the above two steps are $O(n^2)$ and $O(n^3)$, respectively, with n being the number of data points.

Applying spectral clustering to large-scale data has attracted increasing attention in recent years [6]. Several accelerated spectral clustering methods have been proposed. The K-means based approximate spectral clustering (KASP) method has been proposed to reduce the data size beforehand to construct the similarity matrix [7]. The sparse coding technique has been applied to construct a sparse similarity matrix [8]. Nyström method has been used to obtain low-rank approximations of large similarity matrix [9].

Nyström method is efficient and commonly used for spectral clustering to overcome the scalability problem by generating low-rank matrix approximation [9]. Nyström method selects l $(l \ll n)$ landmark points (i.e., sample points) from the n data points and uses a small $l \times l$ matrix to approximate the $n \times n$ matrix for eigen-decomposition. As a result, the computational complexity of eigen-decomposition is reduced to $O(l^3 + nq)$. An important issue of Nyström method is sampling, i.e., how to select the landmark points to construct the small matrix for a good approximation. Uniform sampling is widely used for Nyström method to select the landmark points. However, uniform sampling does not provide a deterministic guarantee on the clustering performance.

Many sampling methods have been proposed to improve the Nyström method. Zhang et al. [10] propose a k-means based sampling method that uses the centers obtained from k-means as the landmark points. Kumar et al. [11] compare the uniform and non-uniform sampling methods and provide a performance bound for the Nyström method with uniform sampling without replacement. Musco et al. [12] propose a recursive sampling method based on ridge leverage scores without assumption on coherence or regularity. These methods aim to reduce the matrix approximation error. However, for spectral clustering, better matrix approximation does not always lead to a better clustering result. The matrix approximation error is not a good criterion to guide the selection of landmark points for Nyström based spectral clustering.

Instead of reducing the matrix approximation error, some sampling methods select the landmark points based on the criterion of clustering performance. Jia et al. [13] propose a probability incremental sampling method, in which more meaningful landmark points are selected by literately updating the sampling probabilities of data points. Similarly, Zhang et al. [14] design an incremental sampling framework for Nyström based spectral clustering, where the landmark points are selected one by one, and each next point with minimum variance is selected as the landmark point. Compared with the methods reducing the matrix approximation error, these methods are more effective for Nyström based spectral clustering. However, these methods mainly consider the relationships among landmark points, which may lose information on the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

data structure and incur incorrectly clustering results.

In this paper, to improve the performance of Nyström based spectral clustering, we propose a new sampling method to select the landmark points by considering the hubness score of data points. In the proposed method, the data points with high hubness scores, i.e., appearing frequently in the nearest neighbor lists of other data points, are selected as the landmark points. The hubness score is an important topological property of data, especially for high-dimensional data, which can better reflect the structure of data points [15]. The data points with high hubness scores have a close relationship with other data points, since they are the nearest neighbors of most of the data points. Hubness based clustering methods have been reported to be effective for clustering accuracy improvement [16]. The proposed sampling method takes advantage of the topological property of hubs to find the representative landmark points, which is able to achieve scalable and accurate clustering results. We further design fast computation methods, i.e., local hubness approximated methods, to speed up the sampling process. Experimental results show that the proposed method makes a balance between efficiency and effectiveness, and outperforms other sampling methods for Nyström based spectral clustering.

The rest of the paper is organized as follows. Section 2 introduces the background knowledge of spectral clustering and Nyström spectral clustering. In Section 3, we propose two versions of hubness-based sampling for Nyström spectral clustering and analysis of their complexities. Section 4 compares the proposed methods with other algorithms in the experiments. Finally, we give a summary of this paper in Section 5.

II. Preliminaries

A. Spectral Clustering

Given a set of n data points $x_1, \ldots, x_n$, spectral clustering algorithm first constructs a similarity matrix $S = (s_{ij}) \in \mathbb{R}^{n \times n}$, where $s_{ij} \geq 0$ indicates the relationship between x_i and x_j. The normalized Laplacian matrix is then computed based on S as

$$L = I - D^{-1/2} S D^{-1/2}, \tag{1}$$

where D is a diagonal matrix and its diagonal element d_{ii} is the sum of each rows of similarity matrix, i.e.,

$$d_{ii} = \sum_{j=1}^{n} s_{ij}. \tag{2}$$

The top c eigenvectors of L are computed and the matrix $H \in \mathbb{R}^{n \times c}$ with these eigenvectors as columns is formed. Let each row of H represent a data point in $\mathbb{R}^c$ and cluster these points by k-means. Each original data point x_i is mapped to the data point represented in row i of H and assigned to the same cluster.

Due to the high computational cost in similarity matrix computation and eigen-decomposition, spectral clustering is difficult to directly apply to large-scale clustering tasks. Nyström method is widely used for accelerating spectral clustering by obtaining low-rank approximations of a large similarity matrix.

B. Nyström Spectral Clustering

We show how to apply Nyström method to accelerate spectral clustering. Nyström method finds approximate eigenvectors of a large similarity matrix by conducting eigen-decomposition for a small sub-matrix of the original matrix and employing the Nyström extension to fill in the rest.

Let S again be the similarity matrix of n data points. Denote l ($l << n$) as the number of landmark points and $A \in \mathbb{R}^{l \times l}$ as the similarity matrix constructed from the l landmark points. Denote $B \in \mathbb{R}^{l \times (n-l)}$ as the similarity matrix constructed from the l landmark points and the $(n-l)$ remaining points. Denote C as the similarity matrix constructed from all $(n-l)$ remaining points. Thus, the similarity matrix S can be rearranged as

$$S = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}. \tag{3}$$

The Nyström method uses A and B to approximate S by replacing C with $B^T A^{-1} B$. That is

$$S = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix}. \tag{4}$$

For spectral clustering, to calculate the normalized Laplacian matrix in equation (1), Fowlkes et al. [9] propose an efficient approach without a direct calculation of $B^T A^{-1} B$ in S. In [9], $D^{-1/2} S D^{-1/2}$ is calculated based on $D_A^{-1/2} A D_A^{-1/2}$ and $D_A^{-1/2} B D_B^{-1/2}$, where D_A and D_B are diagonal matrices and the diagonal elements are the sum of rows from A and B, respectively. Let $\hat{A} = D_A^{-1/2} A D_A^{-1/2}$ and $\hat{B} = D_A^{-1/2} B D_B^{-1/2}$, then $D^{-1/2} S D^{-1/2}$ can be represented as

$$D^{-1/2} S D^{-1/2} = \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{B}^T & \hat{B}^T \hat{A}^{-1} \hat{B} \end{bmatrix}. \tag{5}$$

Thus, the top c eigenvectors of L can be approximately obtained based on $\hat{A}$ and $\hat{B}$. Assume the eigen-decomposition of $\hat{A}$ takes the form $\hat{A} = \hat{V}_A \hat{\Sigma}_A \hat{V}_A^T$, where $\hat{\Sigma}_A$ contains the eigenvalues of $\hat{A}$ and $\hat{V}_A$ are the corresponding eigenvectors. The eigenvectors of L can be approximated as

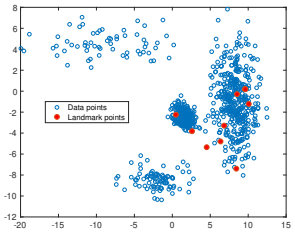
$$V = \sqrt{\frac{l}{n}} \begin{bmatrix} \hat{A} \\ \hat{B}^T \end{bmatrix} \hat{V}_A \hat{\Sigma}_A^{-1}. \tag{6}$$

Finally, k-means is performed on low dimensional space constructed based on the top c eigenvectors obtained from equation (6).

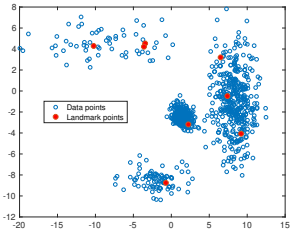
In this paper, our objective is selecting the representative sample points (i.e., landmark points) to construct A and B in equation (4) for obtaining good approximation results.

III. The Proposed Hubness-based Sampling for Nyström spectral clustering

In this section, we introduce the proposed hubness-based sampling method for Nyström spectral clustering.

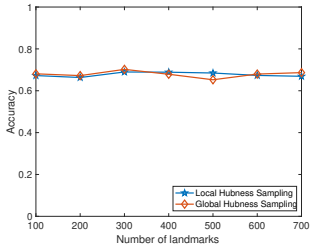


(a) Uniform landmark sampling

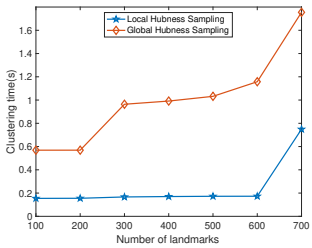


(b) Hubness based landmark sampling

Fig. 1. Uniform sampling is extremely uncertain sampling for data set with uneven distribution. A better sampling method should select points that more equally cover the relevant data.



(a) Clustering accuracy



(b) Clustering time

Fig. 2. Comparision between local hubenss sampling (10 pre-partitions created by random partition) and global hubness sampling on data set USPS: clustering accuracy of different methods are very close while local hubness sampling consume less time.

A. Hubness Scores

Generally, Nyström method uses uniform sampling to select landmark points from data sets, which often fails for many data sets with uneven distribution. Figure 1 shows an example that the uniform sample is more possible oversampling in the bigger clusters for a dataset with different densities of clusters. As a result, the smaller but still important clusters are missed for sampling. Uniform sampling is observed that shows the limited clustering quality for Nyström spectral clustering.

To address this problem, we introduce the hubness score as a measure of point importance to select landmarks for Nyström spectral clustering. The hubness score that indicates the number of k-occurrences of a point in k-nearest-neighbor lists is defined as

Definition 1: Hubness score [15]. For any integer $k > 0$, the hubness score of data points x_i is defined as

$$N_i^k \stackrel{\text{def}}{=} \sum_{j=1}^{n} t_{ij}, \tag{7}$$

where

$$t_{ij} = \begin{cases} 1, & x_i \text{ is among the } k \text{ nearest neighbors of } x_j, \\ 0, & \text{otherwise.} \end{cases}$$

Let T^k be the binary adjacency matrix of k-nearest-neighbors graph, we can also write the matrix form as

$$N^k = T^k \mathbf{1}_n. \tag{8}$$

The data points with largest hubness scores are referred to as hubs.

It has been discussed in [16] that hubness does not depend on scale. In other words, for a point with a high hubness score, its density can be a low value.

B. Local Hubness based Sampling

Computing the hubness scores naively requires a costly matrix inversion, which can be prohibitive for large-scale data sets. It will be $O(n^2)$ computational complexity to directly calculate the hubness score which is very time-consuming for large-scale data sets. To avoid the high computational cost, we consider an alternative scheme which computes local hubness score in the local partition. That is, first dividing the data points into some partitions, then computing hubness score based on the data points in local partitions.

As shown in Figure 2, global hubness-based sampling often achieves better clustering accuracy but consumes much more time. In this paper, we utilize two pre-partition methods: random and k-means partition.

1) Hubness-based Sampling using Random Partition: Random partition randomly divides the data points into some partitions. We suppose that all the data points are divided into m rep-partitions $P_1, P_2, \ldots, P_m$ at random. To better distinguish, we let P indicate rep-partition and p indicates the probability. Thus, we can write the approximated form of equation (8) as

$$\hat{N}^k \approx \begin{bmatrix} T_1^k & & \\ & \ddots & \\ & & T_m^k \end{bmatrix} \mathbf{1}_n, \tag{9}$$

where T_i^k is thebinary adjacency matrix of k-nearest-neighbor in the P_i local partition.

Instead of directly selecting the data points with the highest hubness scores, we select data points based on the following probability

$$p_i = \hat{N}_i^k / \sum_{j=1}^{n} \hat{N}_j^k. \tag{10}$$

That is, the data points with higher hubness scores have higher probabilities to be selected. The hubness-based sampling method using a random partition is summarized in Algorithm 1.

Obversely, there are two conditions for better approximation: (1) maintaining a similar size $|P_i|$ between each part; (2) x_j should have neighbor x_v in the same part P_i as many as possible. The random partition can easily provide a similar size of parts to satisfy the first condition. Although the second condition is hard to satisfy by the random partition, the data points with the largest hubness scores still have higher possibilities to become hubs. To illustrate this idea, consider a simple example in Figure 3 which supposes the pre-partition is in a bad situation where data points in different clusters are randomly mix up into pre-partitions, but that approximated hubs still have reasonable distribution.

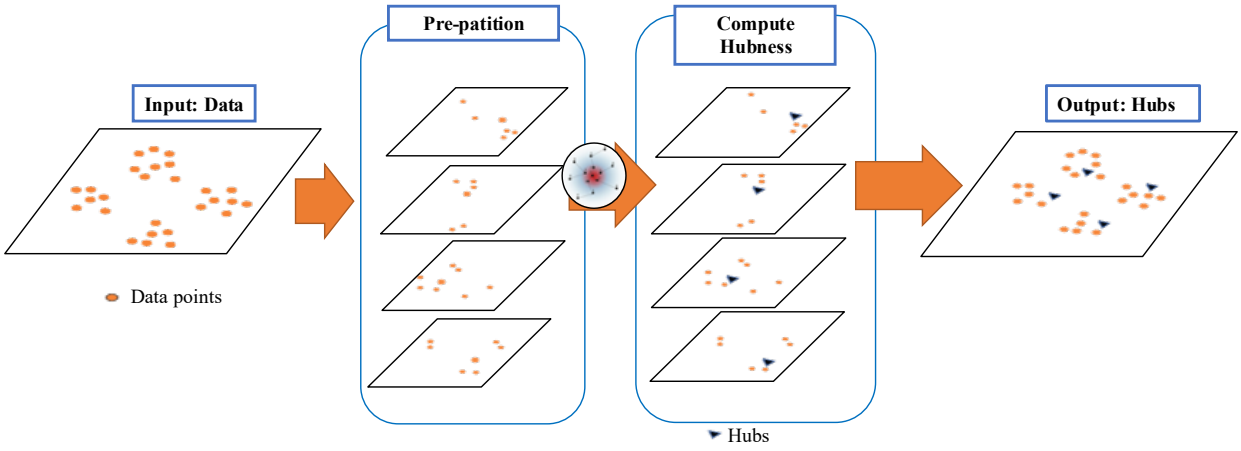


Fig. 3. Illustrative example of computation of local hubness: In worst case, different clusters are mixed in the pre-partition while we still have a reasonable set of hubs computed by N^1.

Algorithm 1 Local Hubness Sampling using random partition

Input:
$X = x_1, x_2, \ldots, x_n$: datasets
l: number of landmark points
k: hubness parameter
m: number of pre-partitions;

Output:
$I = i_1, i_2, \ldots, i_l$: the indices of the sampled points

Obtain initial m pre-partitions $C_i, \ldots, C_m$ via random partition;
for all pre-partition P_i **do**
  Uniformly obtain m pre-partitions from X;
end for
Compute approximate hubness scores $\hat{N}^k$ by equation (9);
Obtain I by selecting data points with highest hubness score as landmark;
return I.

2) Hubness-based Sampling using k-means partition: Besides random partition, we also apply the k-means algorithm to obtain partitions and then compute the local hubness score to select landmark points.

Different from random partition, in k-means partition, the data points in the same part are more similar than those in others and the number of groups varies greatly. The value of the hubness score in different parts can be varied. Thus, we decide the number of selected landmark points n_i in a local partition P_i based on $|P_i|$ (the number of data points in partition P_i) as

$$n_i = \frac{|P_i|}{n} l. \tag{11}$$

It maybe necessary to modify n_i to satisfy the equation $\sum_{i=1}^{m} n_i = l$ where n_i is an integer. In each local partition P_i, a data point is selected as the landmark with the probability according to equation (10).

The hubness-based sampling method using k-means partition is summarized as Algorithm 2.

Algorithm 2 Local Hubness Sampling using k-means partition

Input:
$X = x_1, x_2, \ldots, x_n$: datasets
l: number of landmark points
k: hubness parameter
m: number of pre-partition;

Output:
$I = i_1, i_2, \ldots, i_l$: the indices of the sampled points

Obtain initial m pre-partitions $C_i, \ldots C_m$ via k-means partition;
Set $n_i = \frac{|P_i|}{n} l$, and modify n_i to satisfy $\sum_{i=1}^{m} n_i = l$ where n_i is an integer;
for all pre-partition P_i **do**
  Construct the local adjacency matrix T_i^k of k-nearest-neighbors graph of P_i;
  Set $N_i^k = T_i^k \mathbf{1}$ and $p_{i_j} = N_{i_j}^k / \sum_{j=1}^{|P_i|} N_{i_j}^k$;
  Sample n_i indices of C_i according to the probability vector p_i, and add them into I;
end for
return I.

C. Applying Landmark Points to Nyström Spectral Clustering

The Nyström spectral clustering using the proposed sampling method can be summarized as follows.

1) Obtain landmark points by Algorithm 1 or 2.
2) Compute similarity matrix A between landmark points and similarity matrix B between landmark points and remaining data points. A common measure of similarity S_{ij} is Gaussian kernel function:

$$S_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{12}$$

where σ is a scaling kernel parameter that determine how fast the similarity decrease with the Euclidean distance between data points x_i and x_j.

3) Calculate D by Eq. (2) and construct $R = \hat{A} + \hat{A}^{-1/2}\hat{B}\hat{B}^T\hat{A}^{-1/2}$.
4) Calculate D_A and D_B by Eq. (2), Let $\hat{A} = D_A^{-1/2}AD_A^{-1/2}$ and $\hat{B} = D_A^{-1/2}BD_B^{-1/2}$.
5) Compute the approximate eigenvectors $\hat{V}$ by equation (6).
6) Normalize the top c eigenvectors of $\hat{V}$ as U.
7) Apply k-means clustering to n rows of U into c clusters.

D. Complexity Analysis

The total time complexity of Nyström method is $O(l^3) + O(nl^2)$, where $O(l^3)$ presents the eigen-decomposition and $O(nl^2)$ presents the orthogonalizing of eigenvectors. In Algorithm 1 and 2, the sampling schemes are based on random partition and k-means partition, respectively. In the case of k-means local partition, the time complexity is $O(tmn)$, where t indicates the number of iterations. While, if we use the random local partition, this time complexity can be ignored. The computation of hubness score is happened in each local partition. Thus, the time complexity of computing hubness score is $O(s^2 + ks)$, where $s = max|P_i|$, $i = 1, 2, \ldots, m$. Because of the partition approximation, we can simply adjust the computational complexity by setting a reasonable value of m.

IV. Experiments

In this section, we conduct several experiments based on both synthetic and real-world data sets to evaluate the performance of two proposed methods.

A. Experimental Settings

There are four parameters in our method: m, i.e., the number of pre-partition; k, i.e., the number of nearest landmarks; l, i.e., the number of landmarks; σ, i.e., the scale parameter of Gassain kernel equation in equation (12). For each data set, we employ grid search to search for the possibly best parameters and execute experiments. m is searched in the set of $\{10, 20, \ldots, 100\}$. k is searched in the set of $\{5, 10, \ldots, 30\}$. σ is set by the mean distance value between x_i and x_j in the equation (12). We also range l in $\{100, 200, \ldots, 700\}$ in the experiments. We repeat each method 10 times and report the average value as the final result.

All experiments were run in MATLAB 9.4.0 (R2018a) on Ubuntu 16.04.5 LTS with Intel CPU E5-1650 v4 @ 3.60GHz (6Core/12Thread, Broadwell) and 16GB x4 = 64GB ram.

B. Evaluation Metric

To evaluate the clustering performance, we adopt one widely used evaluation metric, i.e., Accuracy (ACC), to evaluate the clustering results. Let $X = [x_1, x_2, \ldots, x_n]$ be the data matrix. For each data point x_i, denote t_i and c_i as the cluster label of ground truth and obtained cluster label from clustering algorithms, respectively. The ACC is defined as:

$$\text{ACC} = \frac{\sum_{i=1}^{n} \delta(t_i, \text{map}(c_i))}{n}, \tag{13}$$

where n is the number of data and $\delta(t, c)$ is a function to check t and c are equal or not, returning 1 if equals otherwise returning 0. The map(c) is a best mapping function that maps each predicted label to the most possibly true cluster label by permuting operations. The map function can be found by Kuhn-Munkres algorithm [17]. As the name implies, a better clustering result will provide a more significant value of ACC in the range of $[0, 1]$.

C. Compared methods

We compare the proposed sampling methods with other sampling methods for Nyström spectral clustering. There are several versions of Nyström spectral clustering, we use a Matlab version according to [8] in our experiments. The compared methods are

1) US: Short for uniform sampling. As we mentioned in section 2.2, Nyström often use random sampling in practice. We compare the uniform sampling scheme as baseline landmark selection methods.
2) KS: Short for k-means sampling proposed by [10].
3) RS: Short for recursive sampling for the Nyström method by [12]. The Authors have provided their Matlab code on their GitHub [1].

For our algorithm, there are two versions:

1) HRS: Short for hubness-based sampling using the random partition.
2) HKS: Short for hubness-based sampling using k-means partition.

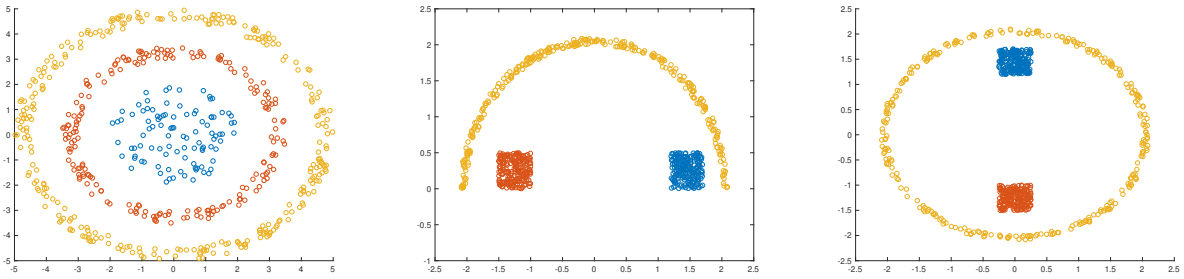
D. Clustering Results on Synthetic Data sets

In this experiment, we evaluate the clustering performance of compared methods on three synthetic data sets, as shown in Figure 4.

Figure 5 shows the clustering accuracies by varying number of landmarks for all methods. In most cases, the accuracy increases as the number of landmark points increases. It can be seen that both the proposed HRS and HKS methods significantly improve the clustering result from baseline (US) and obtain better results than other Nyström based sampling methods.

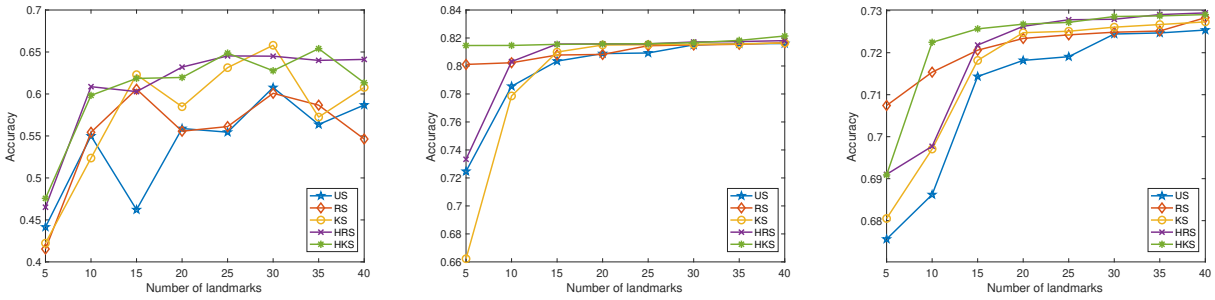
The US method provides the shortest clustering time and the KS method spends the most time. The proposed HRS method consumes the 2nd shortest time while HKS spend more time because the k-means partition leads to imbalance number of elements in local parts. Overall, the two proposed HRS and HKS methods achieve better clustering results on the there synthetic data sets with reasonable clustering time.

[1] https://github.com/cnmusco/recursive-nystrom



(a) Three circles (b) Half circle and two squares (c) one circle and two squares

Fig. 4. Synthetic data sets



(a) Three circles (b) Half circle and two squares (c) half-kernel

Fig. 5. Accuracy by varying number of landmarks on synthetic data sets

TABLE I
PROPERTIES OF DATA SETS

data set	# of samples	# of Features	# of Clusters
PenDigits	10,992	16	10
USPS	9,298	256	10
RCV1_4Class	9,625	29,992	4
TDT2	9,394	36,772	30
Letter	20,000	16	26

E. Clustering Results on Real-world Data Sets

In this experiment, we evaluate the clustering performance of compared methods on five real-world data sets. The properties of the five real-world data sets are summarized in Table I and introduced as follows.

USPS: A data set of handwritten image data within 10 clusters from 0 to 9 digit. Each column represents an image that has 256 dimensions [18].

PenDigits: As the name suggests, it is handwriting image data set [19].

RCV1_4Class: A subset of the test benchmark data collection for machine learning research [20].

TDT2: A subset of the document data set TDT2 which consists of articles from 6 media sets [21].

Letter: A letters recognition data set that consists of 26 English capital letters (from A to Z) [22].

We show the clustering accuracy of different methods on all data set in Figure 6 and report the best accuracy for each method in Table II. To show how spectral clustering is accelerated in large-scale data sets, we also report the results of the Normalized Spectral Clustering (SC) with sparse similarity which is popular in practice. As shown in Figure 6, the accuracy of different methods increases generally when the number of landmark points increases. On two handwriting digits data sets, the result is somewhat counter-intuitive where accuracy does not progressively increase. The proposed HRS and HKS methods significantly improve the clustering accuracy from the baseline US and outperform other compared methods.

For a fair comparison, we report the maximum clustering time of different methods on all data set in Table III. The original spectral clustering method consumes the most time and the US method consumes the least time on all data sets. The proposed HRS and HKS methods have good speed comparing with KS and RS.

Overall, the two proposed HRS and HKS methods achieve better clustering results on four real-world data sets with reasonable clustering time. However, HKS often consumes more runtime than HRS because of using k-means partition. Note that all clustering methods reduce the runtime from the original spectral clustering. For the data set TDT2 with the largest feature size and the data set Letter with the largest sample size, the speedups are more remarkable.

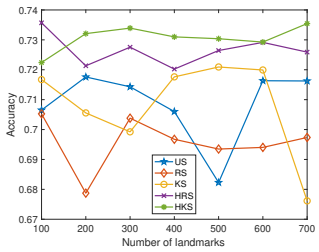
We compare the parameters, i.e., the number of nearest-neighbor k and the number of pre-partition m, regarding clustering accuracy and time on PenDigits data set and report

TABLE II
ACCURACY OF DIFFERENT SAMPLING METHODS FOR NYSTRÖM SPECTRAL CLUSTERING

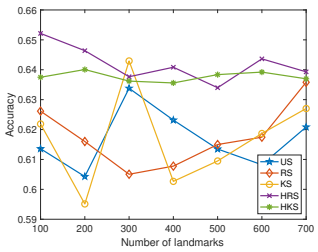
Data sets	SC	US	RS	KS	HRS	HKS
PenDigits	$0.640_{0.000}$	$0.718_{0.042}$	$0.705_{0.034}$	$0.721_{0.033}$	$0.736_{0.035}$	$0.735_{0.040}$
USPS	$0.666_{0.001}$	$0.634_{0.031}$	$0.636_{0.041}$	$0.643_{0.014}$	$0.652_{0.028}$	$0.640_{0.035}$
RCV1_4Class	$0.305_{0.000}$	$0.633_{0.066}$	$0.616_{0.037}$	$0.651_{0.053}$	$0.664_{0.069}$	$0.642_{0.055}$
TDT2	$0.762_{0.012}$	$0.398_{0.039}$	$0.410_{0.023}$	$0.400_{0.027}$	$0.417_{0.040}$	$0.450_{0.040}$
Letter	$0.312_{0.017}$	$0.292_{0.034}$	$0.302_{0.022}$	$0.303_{0.026}$	$0.314_{0.026}$	$0.315_{0.027}$

TABLE III
CLUSTERING TIME OF DIFFERENT SAMPLING METHODS FOR NYSTRÖM SPECTRAL CLUSTERING (SECOND)

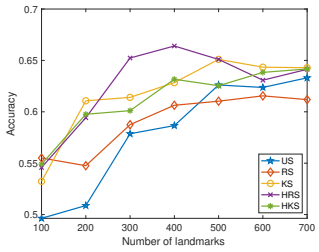
Data sets	SC	US	RS	KS	HRS	HKS
PenDigits	$1.518_{0.291}$	$0.029_{0.001}$	$2.310_{0.021}$	$0.572_{0.021}$	$0.069_{0.037}$	$0.660_{0.006}$
USPS	$2.321_{0.051}$	$0.602_{0.001}$	$2.451_{0.031}$	$0.884_{0.027}$	$0.673_{0.029}$	$0.696_{0.008}$
RCV1_4Class	$20.589_{0.501}$	$0.192_{0.002}$	$3.580_{0.020}$	$0.483_{0.017}$	$0.847_{0.008}$	$1.243_{0.000}$
TDT2	$57.745_{0.149}$	$0.277_{0.002}$	$6.836_{0.011}$	$2.867_{0.030}$	$0.649_{0.033}$	$1.891_{0.017}$
Letter	$15.571_{1.130}$	$0.451_{0.014}$	$1.562_{0.002}$	$3.443_{0.521}$	$0.694_{0.026}$	$0.811_{0.022}$



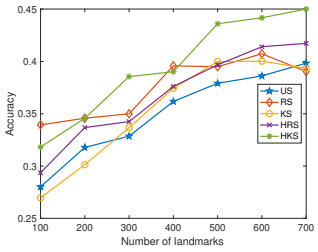
(a) PenDigits



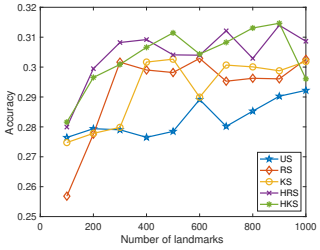
(b) USPS



(c) RCV1_4Class

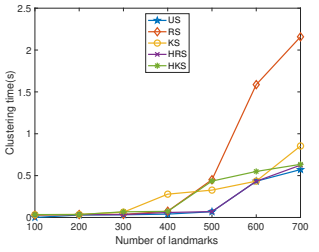


(d) TDT2

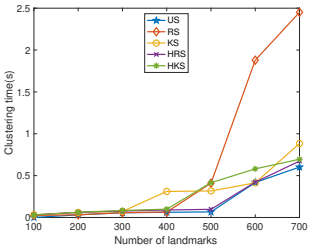


(e) Letter

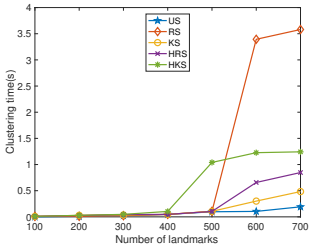
Fig. 6. Accuracy by varying number of landmarks on real-world data sets



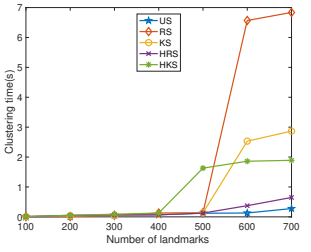
(a) PenDigits



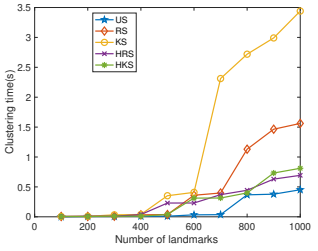
(b) USPS



(c) RCV1_4Class

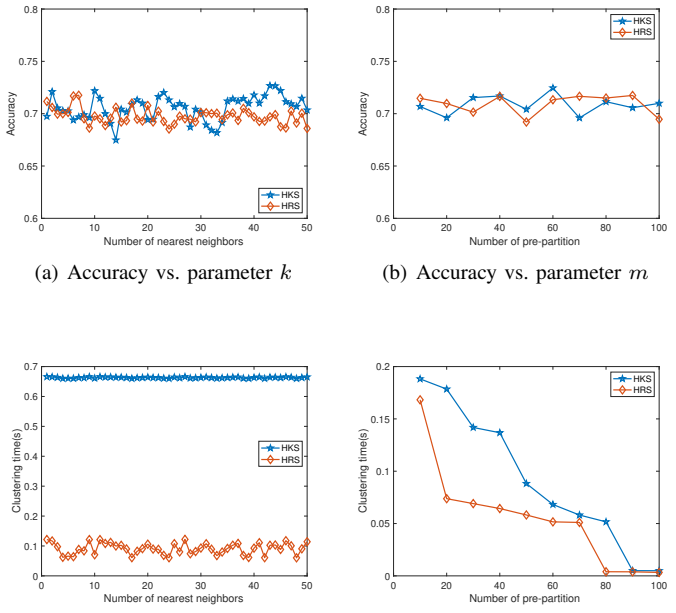


(d) TDT2



(e) Letter

Fig. 7. Clustering time by varying number of landmarks on real-world data sets



(a) Accuracy vs. parameter k (b) Accuracy vs. parameter m

(c) Clustering time vs. parameter k (d) Clustering time vs. parameter m

Fig. 8. Accuracy and clustering time by varying parameters on PenDigits data set

the results in Figure 8. The parameters are set as $m = 50$ and $l = 700$ in Figures 8(a) and 8(c); $k = 15$ and $l = 700$ in Figures 8(b) and 8(d). When the two parameters are changed within a certain range, the performance also changes within a certain range. Both the proposed methods show low dependency on parameters and robust for clustering accuracy. On the other hand, the clustering time highly depends on the number of pre-partition m.

V. Conclusion

In this paper, we propose a new sampling method for Nyström spectral clustering based on the hubness score of data points. The data points with high hubness scores have high probabilities to be selected as the landmark points. We further design fast computation methods, i.e., local hubness approximated methods, to speed up the sampling process. We propose two versions of local hubness-based sampling methods and evaluate the performance of the proposed methods by comparing them with three related methods as well as the original spectral clustering. The experimental results on both synthetic and real-world data sets demonstrate the effectiveness of the proposed methods in comparison to other sampling methods for Nyström spectral clustering.

References

[1] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[2] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[3] Xiucai Ye and Tetsuya Sakurai. Robust similarity measure for spectral clustering based on shared neighbors. *ETRI journal*, 38(3):540–550, 2016.

[4] Xiucai Ye and Tetsuya Sakurai. Spectral clustering with adaptive similarity measure in kernel space. *Intelligent Data Analysis*, 22(4):751–765, 2018.

[5] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[6] Xiucai Ye, Hongmin Li, Tetsuya Sakurai, and Zhi Liu. Large scale spectral clustering using sparse representation based on hubness. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1731–1737. IEEE, 2018.

[7] Donghui Yan, Ling Huang, and Michael I Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM, 2009.

[8] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y Chang. Parallel spectral clustering in distributed systems. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):568–586, 2010.

[9] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nystrom method. *IEEE transactions on pattern analysis and machine intelligence*, 26(2):214–225, 2004.

[10] Kai Zhang, Ivor W Tsang, and James T Kwok. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1232–1239. ACM, 2008.

[11] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(Apr):981–1006, 2012.

[12] Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017.

[13] Hongjie Jia, Shifei Ding, and Mingjing Du. A nyström spectral clustering algorithm based on probability incremental sampling. *Soft Computing*, 21(19):5815–5827, 2017.

[14] Xianchao Zhang, Linlin Zong, Quanzeng You, and Xing Yong. Sampling for nyström extension-based spectral clustering: incremental perspective and novel analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(1):7, 2016.

[15] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010.

[16] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):739–751, 2013.

[17] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[19] Fevzi Alimoğlu and Ethem Alpaydin. Combining multiple representations for pen-based handwritten digit recognition. *Turkish Journal of Electrical Engineering & Computer Sciences*, 9(1):1–12, 2001.

[20] Duy Khuong Nguyen and Tu Bao Ho. Fast parallel randomized algorithm for nonnegative matrix factorization with kl divergence for large sparse datasets. *arXiv preprint arXiv:1604.04026*, 2016.

[21] Deng Cai, Xuanhui Wang, and Xiaofei He. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 105–112, 2009.

[22] Peter W Frey and David J Slate. Letter recognition using holland-style adaptive classifiers. *Machine learning*, 6(2):161–182, 1991.