Sampling for Nyström Extension-Based Spectral Clustering: Incremental Perspective and Novel Analysis

XIANCHAO ZHANG and LINLIN ZONG, Dalian University of Technology
QUANZENG YOU, University of Rochester
XING YONG, Mi Inc.

Sampling is the key aspect for Nyström extension based spectral clustering. Traditional sampling schemes select the set of landmark points on a whole and focus on how to lower the matrix approximation error. However, the matrix approximation error does not have direct impact on the clustering performance. In this article, we propose a sampling framework from an incremental perspective, i.e., the landmark points are selected one by one, and each next point to be sampled is determined by previously selected landmark points. Incremental sampling builds explicit relationships among landmark points; thus, they work together well and provide a theoretical guarantee on the clustering performance. We provide two novel analysis methods and propose two schemes for selecting-the-next-one of the framework. The first scheme is based on clusterability analysis, which provides a better guarantee on clustering performance than schemes based on matrix approximation error analysis. The second scheme is based on loss analysis, which provides maximized predictive ability of the landmark points on the (implicit) labels of the unsampled points. Experimental results on a wide range of benchmark datasets demonstrate the superiorities of our proposed incremental sampling schemes over existing sampling schemes.

Categories and Subject Descriptors: I.5.3 [Clustering]: Algorithms

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Spectral clustering, Nyström extension, incremental sampling, clusterability analysis, loss analysis

ACM Reference Format:

Xianchao Zhang, Linlin Zong, Quanzeng You, and Xing Yong. 2016. Sampling for Nyström extension-based spectral clustering: Incremental perspective and novel analysis. *ACM Trans. Knowl. Discov. Data* 11, 1, Article 7 (July 2016), 25 pages.
DOI: <http://dx.doi.org/10.1145/2934693>

1. INTRODUCTION

In recent years, spectral clustering has become one of the most popular modern clustering algorithms, since it often outperforms traditional clustering algorithms such as

This work is supported by the National Natural Science Foundation of China, under grant 61272374 and grant 61300190.

A preliminary version of this article was presented in the proceedings of the 11th International Conference on Data Mining (ICDM'11) [Zhang and You 2011].

Authors' addresses: X. Zhang and L. Zong, School of Software, Dalian University of Technology, Dalian 116024, China; emails: xczhang@dlut.edu.cn, zonglinlin.love@163.com; Q. You, Department of Computer Science, 611 Computer Studies Building, University of Rochester, Rochester, NY 14627; email: ultimateyou@gmail.com; X. Yong, Qinghezhong Road No. 68, Haidian District, Beijing 100085, China; email: xyong.ac@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

2016 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 1556-4681/2016/07-ART7 \$15.00

DOI: <http://dx.doi.org/10.1145/2934693>

k -means in many clustering tasks. Nevertheless, it has limited applicability to large-scale data due to its computational complexity of $O(n^3)$ and space complexity of $O(n^2)$, with n being the number of data points. Applying spectral clustering to large-scale data has been a hot topic in recent years [Liu et al. 2013] and an important way to alleviate the memory and computation burdens of spectral clustering is generating some kind of low-rank matrix approximation [Fowlkes et al. 2004].

Nyström extension is an efficient technique to generate low-rank matrix approximation, which has been widely used in many supervised/unsupervised kernel-based learning tasks [Farahat et al. 2011; Kumar et al. 2012; Wang and Zhang 2014]. When using Nyström extension, first, m ($m \ll n$) landmark points are selected from the n data points, and then the eigendecomposition of the $n \times n$ matrix is approximated by using a much smaller matrix of $m \times m$. Since different sampling schemes produce different approximation results, the key aspect of Nyström extension is sampling. Random sampling (RS) is the most commonly used sampling scheme for Nyström extension. However, RS does not provide a deterministic guarantee on the clustering performance.

Many sampling schemes for Nyström extension-based kernel learning have been proposed [Kumar et al. 2012]. Here, we focus on sampling schemes applied to spectral clustering. Drineas and Mahoney [2005] proposed a column-sampling method and revealed a probabilistic error bound of the sampling scheme. Zhang et al. [2008] provided a more concrete analysis of the Nyström low-rank approximation and proposed a k -means-based sampling scheme. A concrete comparison of uniform versus non-uniform sampling was presented by Kumar et al. [2012]; they also gave a performance bound for the Nyström method with uniform sampling without replacement. Belabbas and Wolfe [2009] proposed two alternative sampling strategies for Nyström methods, one employed a randomized approach and the other one provided a much deterministic way to select landmark points. These schemes aim to lower the matrix approximation error. However, for spectral clustering, though better matrix approximation often leads to a better clustering result, this is not always the case. First, two sampled sets with the same matrix approximation error may result in very different clustering performances; second, in some extreme cases, a sampled set with a lower matrix approximation error may perform worse than another sampled set with a higher matrix approximation error. This indicates that the matrix approximation error is too rough a criterion to guide the selection of landmark points for Nyström extension-based spectral clustering.

In this article, we propose a sampling framework from an incremental perspective, i.e., the landmark points are selected one by one, and each next point to be sampled is determined by previously selected landmark points. Incremental sampling builds explicit relationships among landmark points; thus, they work together well and provide a theoretical guarantee on the clustering performance. We provide two novel analysis methods and propose two schemes for selecting-the-next-one of the framework.

The first scheme is based on clusterability analysis. Note that in spectral clustering, the clusters are calculated with the top k eigenvectors V_k of the similarity matrix. While using Nyström extension, the V_k is approximated by the top k eigenvectors $\hat{V}_k$ calculated with the sampled matrix. This kind of eigenvector permutation causes loss of clustering accuracy and it is expected that the clusterability [Hunter and Strohmer 2010] of V_k and $\hat{V}_k$ should be close to achieve accurate clustering results. We first analyze the difference of clusterabilities between V_k and $\hat{V}_k$, which is characterized by the canonical angle between V_k and $\hat{V}_k$. We give a formula to calculate the canonical angle between V_k and $\hat{V}_k$, which shows that the difference of clusterabilities between V_k and $\hat{V}_k$ is bounded by the Schur complement (SC). We then exploit the properties of the SC. The analysis result shows that when adding a new landmark point to p fixed

landmark points, the SC of the $p + 1$ landmark points could be calculated with the SC of the p landmark points. This immediately suggests an incremental sampling scheme for Nyström extension-based spectral clustering, which starts with a fixed number of initial landmark points and selects new landmark points one by one such that the estimated SC is minimized, until a desired number of landmark points is reached.

The second scheme is based on loss analysis. Instead of analyzing the matrix approximation error or top eigenvectors approximation error, we directly focus on the clustering performance of the sampled set. Since sampling-based clustering is substantially using the landmark points to predict the underlying categories of the unsampled points, we introduce a concept called predictive ability to analyze the clustering performance of the sampled set. With loss analysis, we show that a sampled set with the maximum predictive ability can predict the categories of the unsampled points with the minimum expected loss, thus leading to the optimal clustering result. Our analysis suggests a simple but effective greedy incremental sampling scheme for Nyström extension-based spectral clustering. When adding a new landmark point to p fixed landmark points, the proposed scheme selects the unsampled point, which is the toughest to be correctly predicted by the current sampled set; thus, the predictive ability of the whole $p + 1$ landmark points could be maximized.

The proposed clusterability analysis-based incremental sampling (IS) scheme and loss analysis-based incremental sampling (LS) scheme preponderate over existing sampling schemes in the following aspects:

- Instead of trying to optimize the matrix approximation error, which does not have direct impact on the clustering performance, IS and LS try to optimize criteria (clusterability and predictive ability) that are closely related to the clustering performance; thus, they provide a guarantee on the clustering performance.
- IS and LS are incremental sampling schemes. In each step, only local information related to the landmarks points is needed. Thus, the schemes could be implemented online and could benefit very large datasets.
- IS and LS exploit only the similarities between landmark points and the unsampled points, without considering representations of data attributes; thus, it could easily be applied to applications such as image segmentation where some other schemes like k -means-based sampling cannot be applied.
- Both IS and LS are simple to be implemented. Moreover, they provide flexibility on the number of landmark points.
- Most importantly, IS and LS result in higher clustering quality than existing sampling schemes, which is verified by various experiments on clustering tasks of synthetic datasets, UCI datasets, some large-scale datasets, and image segmentation tasks from the Berkeley Segmentation Dataset and Benchmark.
- Experiments also show that the efficiencies of IS and LS are higher than that of the baselines.

A preliminary version of this article was presented in the proceedings of the 11th International Conference on Data Mining (ICDM'11) [Zhang and You 2011]. In this article, we extend the preliminary work in the following aspects: (1) We propose the general incremental sampling framework and state that this is a novel perspective on sampling for Nyström extension-based spectral clustering. (2) We propose the loss analysis-based incremental sampling scheme, which is a novel one not presented in the preliminary version. (3) We analyze the complexities of the proposed incremental sampling schemes. (4) We experimentally investigate the performances of the proposed schemes on much more and much larger datasets.

2. PROBLEM DEFINITION

2.1. Nyström Extension

The Nyström extension method is originally used to solve the eigenfunction of integral equations in the form

$$\int K(x, y)\phi(y)p(y)dy = \lambda\phi(x), \quad (1)$$

where $p(y)$ indicates the underlying probability density function, $\phi(x)$ represents the eigenfunction, and $K(x, y)$ denotes the similarity between x and y . It is applied to speed up kernel machines in Williams and Seeger [2000]. To start with, one needs to choose m landmark points $Z = \{z_1, z_2, \dots, z_m\}$ from the given datasets $X = \{x_1, x_2, \dots, x_n\}$ with $x_i \in \mathbf{R}^d$ ($m \ll n$). For any given point x in X , using the Nyström method, we have

$$\frac{1}{m} \sum_{i=1}^m K(x, z_i) \hat{\phi}(z_i) = \lambda \hat{\phi}(x), \quad (2)$$

where $\hat{\phi}(x)$ is an approximation to the true $\phi(x)$. Equation (2) cannot be solved directly, as $\hat{\phi}(x)$ and λ are both unknown. Denote the similarity matrix between the landmark points by S with $S_{ij} = K(z_i, z_j)$. Substituting x with z_i in Equation (2) and writing it in matrix form, we have

$$S\hat{\Phi} = m\hat{\Phi}\Lambda, \quad (3)$$

where $\hat{\Phi} = [\hat{\phi}_1 \ \hat{\phi}_2 \ \dots \ \hat{\phi}_m]$ are the eigenvectors of S and $\Lambda = \text{diag}\{\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_m\}$ is a diagonal matrix. For an unsampled point x , the j th eigenfunction at x can be approximated as

$$\hat{\phi}_j(x) = \frac{1}{m\hat{\lambda}_j} \sum_{i=1}^m K(x, z_i) \hat{\phi}_j(z_i). \quad (4)$$

With the above equation, the eigenvector for any given point x can be approximated by the eigenvectors of the landmark points Z .

2.2. Applied to Spectral Clustering

In spectral clustering, the eigenvectors are required to be orthogonal to each other. Therefore, how to orthogonalize the approximated eigenvectors is key to apply Nyström extension to spectral clustering.

Let $S = U_S \Lambda_S U_S^T$ be the eigendecomposition of S and E denote the similarity matrix between landmark points and the remaining points, with $E \in \mathbf{R}^{m \times (n-m)}$. Then, Equation (4) can be written in matrix form as $E^T U_S \Lambda_S^{-1}$, which signifies that the affinities between the unsampled points no longer need to be computed. Denote $W \in \mathbf{R}^{n \times n}$ as the similarity matrix between all the data points and, without loss of generality, rearrange the points such that the landmark points come first. In this way, W can be written in a block matrix form as

$$W = \begin{bmatrix} S & E \\ E^T & A \end{bmatrix}, \quad (5)$$

where $A \in \mathbf{R}^{(n-m) \times (n-m)}$ is the similarity matrix between the unsampled points.

Using the approximated eigenvectors $\hat{U} = [U_S; E^T U_S \Lambda_S^{-1}]$, it is shown in Fowlkes et al. [2004] that W can be estimated as

$$\hat{W} = \hat{U} \Lambda_S \hat{U}^T = \begin{bmatrix} S & E \\ E^T & E^T S^{-1} E \end{bmatrix}. \quad (6)$$

Recall that to calculate the NCut, the similarity matrix is required to be normalized. We need to know the row sums of W to acquire D . Without knowing A , D can be estimated through the row sums of $\hat{W}$. Depending on the definiteness of S , two different approaches are proposed [Fowlkes et al. 2004].

2.2.1. When S is Positive Definite. All the eigenvalues of matrix S are positive and $S^{-1/2}$ is defined. The normalized approximated eigenvectors are obtained by

$$\hat{V} = \begin{bmatrix} S \\ E^T \end{bmatrix} S^{-1/2} U_B \Lambda_B^{-1/2}, \quad (7)$$

where $B = S + S^{-1/2} E E^T S^{-1/2}$ with eigendecomposition $U_B \Lambda_B U_B^T$.

2.2.2. When S is a Non-Positive Matrix. Let $Z = \hat{U} \Lambda^{1/2}$ so that $\hat{W} = Z Z^T$. Let $F \Sigma F^T$ denote the diagonalization of $Z^T Z$. Then, defining $V = Z F \Sigma^{-1/2}$, it is easy to verify that $\hat{W} = V \Sigma V^T$ and $V^T V = I$.

3. THE PROPOSED INCREMENTAL SAMPLING SCHEMES

3.1. Framework

We aim to design an incremental sampling framework for Nyström extension based spectral clustering, which has the following properties: (1) The landmark points are selected one by one, and each next one is determined by existing landmark points. (2) The selecting-next-one procedure only needs local information related to landmark points; thus, it can be implemented online. (3) The sampled set has a theoretical guarantee on the clustering performance, and the number of the landmark points can be approximately estimated. (4) The clustering performance increases with the number of landmark points, and we can improve the performance of a given sampled set by adding more landmark points.

The main steps of the proposed framework are summarized in Algorithm 1.

ALGORITHM 1: Incremental Sampling Framework

Input: $X = \{x_1, x_2, \dots, x_n\}$: datasets
 m : number of landmark points

Output: $I = \{i_1, i_2, \dots, i_m\}$ the indices of sampled points

- 1: Randomly choose some (e.g., p) points from X , and add the corresponding indices into I
 - 2: **while** $|I| < m$ **do**
 - 3: Using a heuristic strategy $\mathfrak{S}$ to select the next landmark point x_{p+1} , and add its index $p + 1$ into I
 - 4: Update related variables used in the heuristic method $\mathfrak{S}$
 - 5: **end while**
-

The key point of incremental sampling is the analysis on the clustering performance of the sampled set. We propose two analysis methods, which lead to two incremental sampling schemes.

3.2. Clusterability Analysis-Based Incremental Sampling

3.2.1. Motivation. The following motivating example could be used to explain our sampling scheme. Assume that we are given a dataset $X = \{x_1, x_2, \dots, x_{10}\}$, and all the points belong to four different clusters, i.e., $\{x_1, x_2, x_3, x_4\} \in C_1$, $\{x_5, x_6\} \in C_2$, $\{x_7, x_8\} \in C_3$, and $\{x_9, x_{10}\} \in C_4$. Let I^i be an $i \times i$ matrix with all entries equal to 1. Then, in the ideal

situation, where the intra-cluster similarities are 1 and the inter-cluster similarities are 0, the similarity matrix of X is

$$W = \begin{bmatrix} I^4 & 0 & 0 & 0 \\ 0 & I^2 & 0 & 0 \\ 0 & 0 & I^2 & 0 \\ 0 & 0 & 0 & I^2 \end{bmatrix}.$$

Suppose that four landmark points $\{s_1, s_2, s_3, s_4\}$ are sampled from the four clusters, respectively. The affinity matrix formed by the sampled points would be a 4×4 identity matrix, of which the i th eigenvector would be a vector e_i with $e_{ii} = 1$ and $e_{ij} = 0$ for $j \neq i$. As discussed in Section 2, the Nyström method can successfully embed the points belonging to the same cluster in X into the same landmark point, whereas points of different clusters will be projected to different landmark points. Hence, it is trivial for k -means (used in the final step in spectral clustering) to achieve the correct clustering results.

However, if the sampled four landmark points are not from the four clusters, but only one cluster, for instance, $s_1, s_2, s_3, s_4 \in C_1$, spectral clustering using the Nyström method will fail to clusters C_2, C_3 , and C_4 since the points in $C_2 \cup C_3 \cup C_4$ will be mapped to the same landmark point. In other words, the points in $C_2 \cup C_3 \cup C_4$ are not well represented by the landmark points. The above phenomenon suggests that the landmark points should be sampled in such a way that all the points are represented by the landmark points appropriately.

Continuing with the above toy example, if we have sampled $S = \{s_1, s_2, s_3\}$ from C_1, C_2, C_3 , respectively, then in order to obtain the correct clustering result, it is desirable that we sample the next point $s_4 \in C_4$. Denote w_u as a similarity vector whose elements are the similarities between one unsampled point u and all the sampled landmark points. Note that for any point $u \in C_i$ ($i = 1, 2, 3$), the similarity matrix between u and S would be $w_u = e_i$, whereas for $u \in C_4$, the similarity vector is $w_u = [0 \ 0 \ 0]^T$. If we view w_u as the observed values of some variable y , for $u \in C_4$, the variance of y is $\text{var}(y) = 0$, and for $u \notin C_4$, $\text{var}(y) = \frac{2}{9}$. Hence, if we want to choose s_4 such that $s_4 \in C_4$, we can calculate the variance as above; the point with the minimum variance should be our best choice.

3.2.2. Clusterability Analysis of Eigenvectors. As discussed in previous sections, to spectral clustering, though better matrix approximation implies a better clustering result, the impact of the matrix approximation error on clustering performance is not clear. It was noted in Hunter and Strohmer [2010] that the perturbation in the similarity matrix differs greatly from the perturbation of eigenvectors. In Hunter and Strohmer [2010], the term clusterability is used to measure the clustering performance of a set of given eigenvectors. If two sets of eigenvectors have a small clusterability difference, their clustering results are similar. Let V_k and $\hat{V}_k$ represent the top- k eigenvectors of W and $\hat{W}$, respectively, the clusterability difference between V_k and $\hat{V}_k$ cannot be simply characterized by the norm $\|V_k - \hat{V}_k\|_F$. Actually, as shown in Hunter and Strohmer [2010], the clusterability difference is closely related to the *canonical angles*, which is defined as below.

Definition 3.1. Let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ be the singular values of $V_k \hat{V}_k^T$. Then, $\theta_i = \arccos \lambda_i$ are the canonical angles between V_k and $\hat{V}_k$. Let $\Theta = \text{diag}(\theta_1, \theta_2, \dots, \theta_k)$ be the diagonal matrix of canonical angles between V_k and $\hat{V}_k$.

Given the above definition, the clusterability difference between V_k and $\hat{V}_k$ is determined by the canonical angle, i.e., a small value of canonical angle represents a small

clusterability difference between V_k and $\hat{V}_k$. $\sin \Theta$ is used to represent the clusterability difference between V_k and $\hat{V}_k$. The following lemma (which is also known as the $\sin \Theta$ theorem) gives the relationship between the perturbed eigenvectors and the canonical angles.

LEMMA 3.2 [DAVIS AND KAHAN 1970]. *Denote $V_k = [v_1, v_2, \dots, v_k]$ and $\hat{V}_k = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_k]$, with $\{\lambda_i, v_i | \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n, i = 1, \dots, n\}$ and $\{\hat{\lambda}_i, \hat{v}_i | \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n, i = 1, \dots, n\}$ be the eigensystems of W and $\hat{W}$, respectively. If $|\hat{\lambda}_k - \lambda_{k+1}| \geq \alpha$ and $\hat{\lambda}_k \geq \alpha$ for some $\alpha > 0$, then*

$$\|\sin \Theta\|_F \leq \frac{1}{\alpha} \|W\hat{V}_k - \hat{V}_k\hat{\Sigma}_k\|_F. \quad (8)$$

Applying Lemma 3.2 to Nyström extension-based spectral clustering, we can give an upper bound of $\|\sin \Theta\|$ in the following theorem.

THEOREM 3.3. *Given the definition above, if $|\hat{\lambda}_k - \lambda_{k+1}| \geq \alpha$ and $\hat{\lambda}_k \geq \alpha$ for some $\alpha > 0$, then*

$$\|\sin \Theta\|_F \leq \frac{\sqrt{k}}{\alpha} \|A - E^T S^{-1} E\|_F. \quad (9)$$

PROOF. As the column of $\hat{V}_k$ and the diagonal elements of $\hat{\Sigma}_k$ are the top- k eigenvectors corresponding to the eigenvalues of $\hat{W}$, we have $\hat{W}\hat{V}_k = \hat{V}_k\hat{\Sigma}_k$; therefore,

$$\begin{aligned} \|W\hat{V}_k - \hat{V}_k\hat{\Sigma}_k\|_F &= \|W\hat{V}_k - \hat{W}\hat{V}_k\|_F \\ &\leq \|W - \hat{W}\|_F \|\hat{V}_k\|_F \\ &= \|A - \hat{A}\|_F \|\hat{V}_k\|_F \\ &= \|A - E^T S^{-1} E\|_F \|\hat{V}_k\|_F. \end{aligned}$$

According to the orthogonality of $\hat{V}_k$, it is straightforward to have

$$\|\hat{V}_k\|_F = \sqrt{\text{Tr}(\hat{V}_k^T \hat{V}_k)} = \sqrt{k}, \quad (10)$$

where $\text{Tr}(\cdot)$ denote the trace of a matrix. Substituting the above results into Equation (8) will conclude the proof. $\square$

In Theorem 3.3, the conclusion holds on the condition that columns of $\hat{V}_k$ are orthogonal to each other; this is satisfied for spectral clustering.¹ The term $A - E^T S^{-1} E$ is referred to as the SC. In this way, the performance of the Nyström method-based spectral clustering is related to the SC. Let $W_{I \times J}$ (use W_I as the abbreviation of $W_{I \times I}$) be an $s \times t$ matrix whose (i, j) th entry is given by $(W_{I \times J})_{ij} = W_{I_i J_j}$, where $I = \{I_1, I_2, \dots, I_s\}$ and $J = \{J_1, J_2, \dots, J_t\}$ are the indices of W . Then, $S = W_I$, and the (i, j) th entry of SC can be given by the following lemma.

LEMMA 3.4 [BELABBAS AND WOLFE 2009]. *Given the Schur complement $\text{SC}(W_I) = A - E^T S^{-1} E$, the entry-wise of Schur complement can be given by*

$$\text{SC}_{ij}(W_I) = \frac{\det(W_{I \cup \{i\} \times I \cup \{j\}})}{\det(W_I)}. \quad (11)$$

¹In other low-rank approximations, this condition does not necessarily hold.

According to Lemma 3.4, a sampling method based on the determinant of S is proposed in Belabbas and Wolfe [2009]. In the following, we will analyze the SC from a different viewpoint, which leads to another sampling strategy. Our sampling method behaves in an incremental way. Assuming that we have picked p landmark points from the whole dataset, we want to choose the $(p + 1)$ th landmark point such that the SC will be minimized. Then, we can repeat the strategy, until we reach a total of m landmark points. Denote the set of indices of the sampled landmark points as I , and suppose that the index of the next chosen point is k , then we have

$$(\text{SC}_{ij}(W_{I \cup \{k\}})) = \frac{\det(W_{I \cup \{k, i\} \times I \cup \{k, j\}})}{\det(W_{I \cup \{k\}})}. \quad (12)$$

The following theorem shows that when adding a new landmark point to p fixed landmark points, the SC of the $p + 1$ landmark points could be calculated with the SC of the p landmark points.

THEOREM 3.5. *Given the SC and the matrix W , if $\det(W_{I \cup \{i\} \times I \cup \{j\}}) \neq 0$ and $\det(W_I) \neq 0$, then $\exists \alpha, \beta \in R^{(p+1)}$ and $\gamma \in R^p$, such that*

$$\text{SC}_{ij}(W_{I \cup \{k\}}) = \text{SC}_{ij}(W_I) \frac{\alpha^T W_{I \cup \{i\} \times I \cup \{j\}}^T \beta + 1}{\gamma^T W_I \gamma + 1}. \quad (13)$$

PROOF. Let Δ_n represent the numerator of Equation (12) and Δ_d be the denominator of Equation (12). Using some algebra techniques, let k be the index of the $(p + 1)$ th landmark point, Δ_n can be written as (in the following section we denote $p = |I|$)

$$\begin{aligned} \Delta_n &= \sum_{s=1}^{p+2} (-1)^{(s+p+1)} W_{sk} \det(W_{I \cup \{k, i\} \setminus s \times I \cup \{j\}}) \\ &= \sum_{s=1}^{p+1} (-1)^{(s+p+1)} W_{sk} \sum_{t=1}^{p+1} (-1)^{(t+p+1)} W_{tk} \\ &\quad \det(W_{I \cup \{i\} \setminus s \times I \cup \{j\} \setminus t}) + W_{kk} \det(W_{I \cup \{i\} \times I \cup \{j\}}) \\ &= \sum_{s=1}^{p+1} \sum_{t=1}^{p+1} (-1)^{s+t} W_{sk} W_{tk} \det(W_{I \cup \{i\} \setminus s \times I \cup \{j\} \setminus t}) \\ &\quad + \det(W_{I \cup \{i\} \times I \cup \{j\}}) \\ &= w_i^T \bar{W}^{ij} w_j + \det(W_{I \cup \{i\} \times I \cup \{j\}}), \end{aligned}$$

where $\bar{W}^{ij}$ is the cofactor matrix of $W_{I \cup \{i\} \times I \cup \{j\}}$, $w_i, w_j \in R^{p+1}$ with $w_m = [W_{I_1 k}, W_{I_2 k}, \dots, W_{I_p k}, W_{mk}]$, $m = i, j$. If $\det(W_{I \cup \{i\} \times I \cup \{j\}}) \neq 0$, then columns of $W_{I \cup \{i\} \times I \cup \{j\}}$ are linearly independent vectors (the same is true for the rows). Therefore, w_j is in the range of $W_{I \cup \{i\} \times I \cup \{j\}}$, which is equal to

$$\exists \beta \in R^{p+1} \text{ s.t. } w_j = W_{I \cup \{i\} \times I \cup \{j\}} \beta. \quad (14)$$

Then, one can easily verify that

$$\begin{aligned} w_i^T \bar{W}^{ij} w_j &= \det(W_{I \cup \{i\} \times I \cup \{j\}}) w_i^T \beta \\ &= \det(W_{I \cup \{i\} \times I \cup \{j\}}) \alpha^T W_{I \cup \{i\} \times I \cup \{j\}} \beta, \end{aligned}$$

where, like in the above deduction, we assume w_i is in the range of $W_{I \cup \{i\} \times I \cup \{j\}}^T$ and $w_i = W_{I \cup \{i\} \times I \cup \{j\}}^T \alpha$. In the above equation, we use the fact that $\bar{W}^{ij} W_{I \cup \{i\} \times I \cup \{j\}} = \det(W_{I \cup \{i\} \times I \cup \{j\}}) I(p)$ [here $I(p)$ is an $R^{p \times p}$ identity matrix]. Similarly, the denominator can be reformulated as

$$\begin{aligned} \Delta_d &= \det(W_{I \cup \{k\} \times I \cup \{k\}}) \\ &= \sum_{s=1}^{p+1} (-1)^{s+k} W_{sk} \det(W_{I \cup \{k\} \setminus s \times I}) \\ &= \sum_{s=1}^p \sum_{t=1}^p (-1)^{s+t} W_{sk} W_{tk} \det(W_{I \setminus s \times I \setminus t}) \\ &\quad + \det(W_I) \\ &= w_I^T \bar{W}^I w_I + \det(W_I), \end{aligned}$$

where $w = [W_{I_1k}, W_{I_2k} \dots W_{I_pk}]$ and $\bar{W}^I \in R^{p \times p}$ is the cofactor matrix of W_I . Similarly, according to the assumption $\det(W_I) \neq 0$, we have w_I in the range of W_I with $w_I = W_I \gamma$, and then we reach the following equation:

$$\begin{aligned} \Delta_d &= w_I^T \bar{W}^I w_I + \det(W_I) \\ &= \det(W_I) \gamma^T W_I \gamma + \det(W_I). \end{aligned}$$

From the above results, we conclude that Equation (13) holds. $\square$

3.2.3. The Sampling Scheme. Let $C = \gamma^T W_I \gamma$ (note that W_I is semi-definite since W is semi-definite in spectral clustering; thus, $C \geq 0$), if we choose the next landmark point z_{p+1} such that C is large, then the overall SC error will be reduced.

Denote $\{v_i, \lambda_i | \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p, i = 1, \dots, p\}$ to be the eigendecomposition of W_I , according to Theorem 2, if $\det(W_I) \neq 0$, then $V_p = \{v_1, \dots, v_p\}$ will be the orthogonal basis of vector space R^p . Hence, we can rewrite γ as a linear combination of V_p :

$$\gamma = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_p v_p, \quad (15)$$

where $\alpha_i \in R$ is the correlation coefficient. Under this condition, we have

$$C = \gamma^T W_I \gamma = \alpha_1^2 \lambda_1 + \alpha_2^2 \lambda_2 + \dots + \alpha_p^2 \lambda_p. \quad (16)$$

If we fix $\|\gamma\|^2 = \sum_i \alpha_i^2$, then a larger α_p^2 leads to a larger C . In other words, γ should be aligned to the biggest eigenvector of W_I . Note that, as shown in Section 2, the similarity matrix W_I in spectral clustering should be normalized. The largest eigenvector of the normalized similarity matrix $\bar{W}$ is $D^{1/2} \mathbf{1}$. Let $D_I \in R^{p \times p}$ be a diagonal matrix with $(D_I)_{ii} = \sum_{j \in I} W_{ij}$ (row sums of the landmark points) and $E \in R^{p \times p}$ be the diagonal matrix with $E_{ii} = \sum_j W_{ij}$ (row sums of the whole data points). The largest eigenvector of $D_I^{-1/2} W_I D_I^{-1/2}$ is $D_I^{-1/2} \mathbf{1}$. However, in Nyström extension, W_I is normalized with $\bar{W}_I = F^{-1/2} W_I F^{-1/2}$, and we have

$$\begin{aligned} F^{-1/2} W_I F^{-1/2} F^{1/2} \mathbf{1} &= F^{-1/2} D_I D_I^{-1} W_I F^{-1/2} F^{1/2} \mathbf{1} \\ &= F^{-1/2} D_I D_I^{-1} W_I \mathbf{1} \\ &= F^{-1/2} D_I \mathbf{1}, \end{aligned} \quad (17)$$

where $F = D_I$ is a subset of D containing only the selected indices. And, if we further assume that $(D_I)_{ii}$ are almost equal for all i and the same is true for F_{ii} , the largest

eigenvector of $F^{-1/2}W_I F^{-1/2}$ [according to Equation (17)] is well aligned to $\mathbf{1}$. Therefore, to attain a smaller SC error, we require that $\gamma \approx c\mathbf{1}$ (c is a constant scalar), in which case the coefficient vector γ is well aligned to $\mathbf{1}$. Moreover, according to the assumptions above, $w_I = \bar{W}_I \gamma$ is also required to be almost constant everywhere. This can be characterized by the variance of variable W_{ij} . Thus, we can choose z_{p+1} in such a way that the variance of $\{W_{I_1, I_{p+1}}, W_{I_2, I_{p+1}}, \dots, W_{I_p, I_{p+1}}\}$ is the smallest. We call this method IS, which is summarized in Algorithm 2.

ALGORITHM 2: *Clusterability Analysis-Based Incremental Sampling*

Input: $X = \{x_1, x_2, \dots, x_n\}$: datasets
 m : number of landmark points
Output: $I = \{i_1, i_2, \dots, i_m\}$ the indices of sampled points
 $S \in \mathbf{R}^{m \times m}$: similarity matrix between landmark points
 $E \in \mathbf{R}^{m \times (n-m)}$: similarity matrix between landmark points and remaining points

- 1: Randomly chooses 2 points from the X , and add the indices into I
- 2: Calculate E the similarity matrix between the chosen points and remaining points
- 3: Calculate S the similarity matrix of the chosen points
- 4: Calculate $s \in \mathbf{R}^{(n-p)}$, the column variance of E
- 5: **while** $|I| < m$ **do**
- 6: Find $s_{\min} = \min s$, and add its index p into I
- 7: Calculate similarity between x_p and the remaining points, update E such that
 $E \in \mathbf{R}^{(|I|) \times (n-|I|)}$
- 8: Update S to add the similarity between x_p and other samples
- 9: Calculate $s \in \mathbf{R}^{(n-p)}$, the column variance of E
- 10: **end while**

First, the algorithm randomly chooses two points² from the dataset X . Then, it computes the similarity matrix E between the landmark points and the unsampled points. The point with the smallest variance will be picked. Repeat the procedure, until a total of m landmark points are sampled. In terms of computing similarity matrix, during each iteration, only the similarities between the newly sampled point z_{p+1} and the remaining points need to be computed. And, the results can be stored for future use (in the next iteration and in Nyström extension).

3.3. Loss Analysis-Based Incremental Sampling

3.3.1. Motivation. We analyze the example S in Section 3.2.1 from another perspective. Suppose we have sampled $S = \{s_1, s_2, s_3\}$ from C_1, C_2, C_3 , respectively, and we want to sample the next landmark point $s_4 \in C_4$. Denote w_u as a similarity vector defined as above, and suppose four unsampled points $\{u_1, u_2, u_3, u_4\}$ come from the four clusters, respectively. Note that for point u_i ($i = 1, 2, 3$), the similarity matrix between u_i and S would be $w_u^i = e_i$, whereas for u_4 , the similarity vector is $w_u^4 = [0 \ 0 \ 0]^T$. We assume that the similarity can be used to measure the probability of two points belonging to the same cluster, and then one unsampled point will most likely be represented by the landmark point with the largest similarity to it. Hence, for an unsampled point, its largest similarity to the sampled landmark points can be used to measure the certainty of whether it can be represented by landmark points. We can observe that the certainty of being represented by u_1, u_2, u_3, u_4 is 1, 1, 1, 0, respectively, so u_4 is the point with the

²Here, the initialization of the first two points may influence the performance of the algorithm. However, the results in our experiments are superior to other algorithms without using any heuristic method for initialization.

weakest certainty to be represented by the landmark points. In order to represent all the points of X , we should choose u_4 as the next point to be sampled. Therefore, if we want to choose s_4 such that $s_4 \in C_4$, we can first use their largest similarity to the sampled set as their certainty of being represented, and then the point with the minimum certainty of being represented is the best choice.

3.3.2. Loss Analysis of Predicting Labels. We take the sampled set as the classifier to predict the underlying labels of unsampled points and then achieve a heuristic sampling strategy by using predicted loss analysis.

Note that if a point u can be represented by a landmark point s , it means that the category label of u can be estimated by the label of s . In other words, the landmark points can be used to predict the category of the unsampled points. In order to achieve good clustering results, we want to sample landmark points such that all the unsampled points can be well represented. Hence, a good sampling strategy should maximize the predictive ability of the sampled set. Given the dataset $X = \{x_1, x_2, \dots, x_n\}$, sampled set $S = \{s_1, s_2, \dots, s_p\}$, and the unsampled set $S' = \{s'_1, s'_2, \dots, s'_{n-p}\}$. Assume the dataset X can be partitioned into k clusters and the points in the k clusters can be labeled with the set $Y = \{y_1, y_2, \dots, y_k\}$. Given a sampled set, for each unsampled point, we first find the landmark point that can best predict its label and then define the predictive ability of the whole sampled set.

We first introduce a predictive function, which is used to predict the label information of unsampled points. We then evaluate the predictive ability of a sampled set by using the expectation of its predictive loss. Let $l(\cdot)$ be a function of achieving the underlying label of one point x_i . For an unsampled point $s' \in S'$, let $f(s'|s)$ ($s \in S$) be a predictive function, which predicts the label of s' according to the sampled point s . The definition of the predictive function is as follows.

Definition 3.6. Given the sampled set S and an unsampled point $s' \in S'$, the predictive function f , which predicts the label of s' according to s , is defined as

$$f(s'|s) = \hat{y} \quad (\hat{y} = l(s), \forall s \in S). \quad (18)$$

Note that for a given unsampled point s' , we can select any point $s \in S$ to predict the label of s' , and the prediction can be correct or incorrect. If the function f makes a correct prediction, $f(s'|s) = l(s')$; otherwise $f(s'|s) \neq l(s')$. To evaluate the predictive function, we use $L(l(s'), f(s'|s))$, the loss function of f representing the cost associated with a prediction being “wrong,” where $l(s')$ is the real value and $f(s'|s)$ indicates the predictive value. There are two situations for the result of one prediction: $f(s'|s) = l(s')$, where f makes a correct predication and $f(s'|s) \neq l(s')$ otherwise. Thus, we can describe the loss function L with 0–1 loss, which is frequently used in statistics and decision theory. The 0–1 loss assigns zero cost to a correct decision, and unit cost to a wrong decision. Hence, the loss function $L(l(s'), f(s'|s))$ of predictive function $f(s'|s)$ can be defined as follows.

Definition 3.7.

$$L(l(s'), f(s'|s)) = \begin{cases} 1, & f(s'|s) \neq l(s') \\ 0, & f(s'|s) = l(s'). \end{cases} \quad (19)$$

Intuitively, the smaller the expected loss, the stronger the predictive ability of the sampled set, indicating that the sampled set has a higher representativeness for the data. Given an unsampled point s' , to predict its label, we should minimize the expectation of the loss and find the most appropriate landmark point for prediction; thus, the objective of finding the best predictive point for s' is

$$\arg \min_{s \in S} E[L(l(s'), f(s'|s))]. \quad (20)$$

Mathematically, the 0–1 loss can be converted into the indicator function. Suppose $I_{(\cdot)}$ denotes an indicator function of its argument as follows:

$$I_A = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise,} \end{cases}$$

where A denotes an event. Therefore, Equation (19) can be rewritten as

$$L(l(s'), f(s'|s)) = I_{l(s') \neq f(s'|s)}. \quad (21)$$

LEMMA 3.8 (PROPERTY OF $I_{(\cdot)}$). *In probability theory, if I_A denotes the indicator random variable for an event A , then the expectation of I_A is equal to the probability of A , that is,*

$$E[I_A] = \Pr\{A\}. \quad (22)$$

According to Lemma 3.8 and Equation (21), we can get the following theorem.

THEOREM 3.9. *Given the sampled set S , minimizing the expectation of the loss function is equivalent to maximizing the probability of correct prediction, namely,*

$$\arg \min_{s \in S} E[L(l(s'), f(s'|s))] \Leftrightarrow \arg \max_{s \in S} \Pr\{f(s'|s) = l(s')\}. \quad (23)$$

PROOF.

$$\begin{aligned} E[L(l(s'), f(s'|s))] &= E[I_{l(s') \neq f(s'|s)}] \\ &= \Pr\{l(s') \neq f(s'|s)\} \\ &= 1 - \Pr\{f(s'|s) = l(s')\}. \end{aligned}$$

Hence, the objective can be rewritten as

$$\begin{aligned} &\arg \min_{s \in S} \{1 - \Pr\{f(s'|s) = l(s')\}\} \\ &\Leftrightarrow \arg \max_{s \in S} \Pr\{f(s'|s) = l(s')\}, \end{aligned}$$

where $\Pr\{f(s'|s) = l(s')\}$ represents the probability of correct prediction. $\square$

For the predictive function $f(s'|s)$, based on Definition 3.7, we use the underlying label of s as the predictive result for s' . Then, we can find the landmark point that best predicts s' with the following object function:

$$\arg \max_{s \in S} \Pr\{l(s) = l(s')\}. \quad (24)$$

Definition 3.10. Given a sampled set S and unsampled set S' , for each unsampled point $s' \in S'$, denote the landmark point that best predicts s' as $B(s')$, then the predictive ability of the sampled set S is defined as $\sum_{s' \in S'} \Pr\{l(B(s')) = l(s')\}$.

From the above discussion, we have an observation such that with the same number of landmark points, the sampling set with the maximum predictive ability will produce the best final clustering result.

Now the question becomes how to compute $\Pr\{l(s) = l(s')\}$ for choosing one point in S to predict the label of s' . Since the details of function $l(\cdot)$ are unknown, we cannot compute or obtain $l(s)$ and $l(s')$ directly. Note that $\Pr\{l(s) = l(s')\}$ indicates the probability of the landmark point s and the unsampled point s' belonging to the same cluster, and the larger the similarity $w_{ss'}$, the higher the probability that they belong to the same cluster. Thus, instead of calculating $l(\cdot)$, we can estimate $\Pr\{l(s) = l(s')\}$ by their similarity $w_{ss'}$ (after normalization). This means that for an unsampled point s' , with the current sampled set, selecting the landmark point with the largest similarity to

s' to predict its label has the minimum expected loss, though the prediction may be incorrect.

3.3.3. The Sampling Scheme. As discussed in the previous subsection, a good sampled set can well represent all the data and can be used to predict the labels of the unsampled data points. Hence, our goal is to design a sampling scheme such that the predictive ability of the sampled set is maximized.

Based on the loss analysis above, we can predict the category information for every unsampled point using the current sampled set. Nevertheless, for different unsampled points, the accuracy of prediction is different. For an unsampled point s' , denote the sampled point that best predicts s' as $B(s')$. If the probability of $B(s')$ and s' being in the same cluster is low, it means that the current sampled set cannot represent the unsampled point s' well, and it is likely that s' belongs to a new cluster and we should select a new landmark point for predicting s' . Inspired by the idea of the *Adaboost* algorithm, which increases the weight of incorrectly classified example, we have the following proposition for our sampling scheme.

PROPOSITION 3.11. *Suppose we have a set of p landmark points S . If we choose the unsampled point s_{p+1} that has the minimum probability of being correctly predicted by S , then the whole set of the $p + 1$ sampled points $S \cup \{s_{p+1}\}$ will have the maximum predictive ability.*

Based on this proposition, in each round of our algorithm, we select the unsampled point that is most likely to be incorrectly predicted by the current sampled set, formally as follows:

$$s_{p+1} = \arg \min_{s' \in S'} \max_{s \in S} w_{ss'}, \quad (25)$$

where s_{p+1} denotes the next landmark point we will sample.

We call this scheme LS, and the details are described in Algorithm 3.

ALGORITHM 3: *Loss Analysis-Based Incremental Sampling*

Input: $X = \{x_1, x_2, \dots, x_n\}$: datasets
 m : number of landmark points
Output: $I = \{i_1, i_2, \dots, i_m\}$ the indices of sampled points
 $S \in \mathbf{R}^{m \times m}$: affinity matrix between sampled points
 $E \in \mathbf{R}^{m \times (n-m)}$: similarity matrix between sampled points and unsampled points

- 1: Randomly sample one point from the data X , and add index into I
- 2: Calculate E the similarity matrix between the sampled point and remaining points
- 3: **while** $|I| < m$ **do**
- 4: Find s_{p+1} according to E using Equation (25), and add its index into I
- 5: Calculate affinity between s_{p+1} and the remaining points, update E such that $E \in \mathbf{R}^{|I| \times (n-|I|)}$
- 6: **end while**
- 7: Calculate S the similarity matrix of sampled landmark points

First, the algorithm randomly selects one point from the dataset X . And then it calculates the similarity matrix E between the landmark point and the remaining points. When selecting the next landmark point, the one that is most likely incorrectly predicted by the current sampled set will be selected. Repeat the procedure, until a total of m landmark points are sampled.

Fig. 1. Three synthetic datasets. from left to right: three circles, half circle and two points, and one circle and two points.

3.4. Complexity

The total time complexity of Nyström extension is $O(m^3) + O(nm^2)$, where $O(m^3)$ denotes the eigendecomposition of S and $O(nm^2)$ denotes the Nyström extension and orthogonalization of the approximated eigenvectors. The memory usage for Nyström-based spectral clustering is $O(nm)$, mainly used for the storage of the similarity matrix. In Algorithms 2 and 3, the sampling schemes are iterative processes. Compared with RS, the additional expense mainly lies in the calculation of variance of E for IS and searching the extrema of similarity for LS. Both the calculation of variance and the extrema search will give an additional $O(nm^2)$ operation, which is in the same order of magnitude as the RS-based Nyström method.

4. EXPERIMENTS

4.1. Experimental Settings

We compare our proposed incremental sampling schemes (IS and LS) with three Nyström extension-based methods, RS [Fowlkes et al. 2004], the k -means sampling (KS) [Zhang et al. 2008], the weighted sampling (WS) [Belabbas and Wolfe 2009], and two non-Nyström-based approaches, landmark-based spectral clustering (LSC) [Cai and Chen 2015] and fast spectral clustering with k -means (KASP) [Yan et al. 2009]. Also, we report the results of original spectral clustering [Ng et al. 2002]. To make KS more accurate when applying to a spectral clustering algorithm, we also add the orthogonal operation.

The basic parameter configuration for all the algorithms is the same. For WS, as the number of sampling candidates is $\binom{n}{m}$, the *Metropolis* algorithm is employed to reduce the complexity. In order to make a balance between the performance and the time complexity, we adopt the same setting as in Belabbas and Wolfe [2009]. And the number of iterations is set to the order of $50m$ (m is the number of sampled points). Moreover, in the KS scheme, as noted in Zhang et al. [2008], the k -means algorithm does not need to converge. For simplicity, we set the maximum iterations of k -means to be 10, which is the default setting in Zhang et al. [2008].

The normalized mutual information (NMI) [Strehl and Ghosh 2002] is used to evaluate the performance of the schemes for data clustering. The larger the NMI, the better the clustering result. For each fixed number of landmark points, the reported results are averaged over 50 repetitions of running for each sampling scheme.

4.2. Clustering Results on Synthetic Datasets

In this experiment, we compare all the schemes on three synthetic datasets, as shown in Figure 1, that are widely used to evaluate the performance of different clustering algorithms.

The clustering qualities of different schemes are shown in Figure 2, with the number of landmark points increasing. On all the datasets, the performance of Nyström extension-based sampling schemes increases stably and gradually approaches to that

Fig. 2. Clustering accuracy on synthetic datasets.

Fig. 3. CPU time consumption on synthetic datasets.

Table I. Summary of the Four UCI Datasets

Dataset	# of Instances	# of Attributes	# of Classes
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Ecoli	336	7	8

of SC when the number of landmark points increases. LS and IS present significantly better performance than the other Nyström extension-based sampling schemes. The performance of KASP also increases with the increasing number of landmarks, but it is not better than LS and IS. LSC is better than LS and IS sometimes. However, the performance of LSC may decrease with the number of landmark points increasing, which is not desirable for sampling methods. Overall, IS and LS exhibit attractive clustering results, even when the number of landmark points is far more less than the total number of data points.

The time consumptions of the algorithms corresponding to the results in Figure 2 are shown in Figure 3. WS consumes much more time compared with the other schemes. During each iteration, the scheme needs to calculate the determinant of an $m \times m$ matrix, which is the main reason of such an expensive time consumption. The other schemes show comparable results. SC consumes less time since the scales of the datasets are small.

4.3. Clustering Results on UCI Datasets

In this experiment, we test the sampling schemes on four real-world datasets from the UCI Machine Learning Repository,³ namely *Iris*, *Wine*, *Seeds*, and *Ecoli*. A brief summary of the four datasets is given in Table I.

The clustering qualities of different sampling schemes are shown in Figure 4. It can be seen that the clustering accuracies of the Nyström extension-based schemes rise

³<http://archive.ics.uci.edu/ml>.

Fig. 4. Clustering accuracy on UCI datasets.

stably with increasing number of landmark points. The performances of LS and IS are better than that of KS, WS, and RS. With enough landmark points, LS and IS will approach SC on all the datasets. KASP and LSC perform better than other schemes sometimes. However, their performances do not increase stably with increasing number of landmark points, e.g., see the curve in *Iris* after six sampled points. This is not desirable since guaranteed performance could not be achieved, i.e., we do not know whether we can get better performance with increasing number of landmark points. Overall, LS and IS perform better than other sampling schemes.

The time consumptions of the sampling schemes corresponding to the results in Figure 4 are shown in Figure 5. The behaviors of the schemes are similar to their behaviors in synthetic datasets. WS consumes much more time compared with other sampling schemes. The other schemes present similar time consumption.

4.4. Clustering Results on Large-Scale Datasets

We also conduct experiments on four real-world large-scale datasets from the LibSVM datasets page⁴ (*Shuttle*, *Letter*) and from the UCI Machine Learning Repository (*PenDigits*, *a7a*). (We report the result of *PenDigits* and *a7a* in this subsection since it is many times larger than the UCI datasets in the above subsection.) A brief description of the datasets is given in Table II.

The results are shown in Figure 6. We can see that the performance of LS and IS also rises stably with increasing number of landmark points. With enough landmark points, LS and IS will approach SC, and they are the best two in terms of performance, as shown in Figure 6(a) and (d). The clustering accuracy of other schemes may decrease sharply on adding landmark points to existing ones, for example, KASP in Figure 6(a), and WS, RS, and LSC in Figure 6(d). In Figure 6(b) and (c), KASP performs better than

⁴<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

Fig. 5. CPU time consumption on UCI data sets.

Table II. Summary of the Four Large-Scale Datasets

Dataset	# of Instances	# of Attributes	# of Classes
<i>Shuttle</i>	14500	9	7
<i>Letter</i>	20000	16	26
<i>PenDigits</i>	10992	16	10
<i>a7a</i>	16100	122	2

LS and IS, but it does not perform well on other datasets. From the results, we can conclude that IS and LS are effective for large datasets.

The time consumptions of different sampling schemes are shown in Figure 7. SC consumes the most time among all the schemes on all the datasets. There is a little difference between the performances of the other schemes on the datasets *Letter*, *PenDigits*, and *a7a*. On the dataset *Shuttle*, WS consumes the most time, and KASP consumes the second most time.

4.5. Image Segmentation Results

In this subsection, we compare the performances of the five sampling schemes LS, IS, WS, RS, and LSC when applied to image segmentation. We use 60 images selected from the Berkeley Segmentation Dataset [Martin et al. 2001]. For the construction of a similarity matrix between pixels, we apply the χ^2 -distance described in Fowlkes et al. [2004].

We use the F -measure to evaluate the image segmentation results [Martin et al. 2004]. There exist several human segmentations for each image in the Berkeley Segmentation Dataset. We use the human segmentations as ground truth and compute the F -measure with regard to the human segmentations. The averaged F -measure for each scheme is reported. Note that KS and KASP are not appropriate for application to image segmentation. For each pixel, it contains only the color values, but in

Fig. 6. Clustering accuracy on large-scale datasets.

Fig. 7. CPU time consumption on large-scale datasets.

constructing the similarity matrix of an image, not only the color values are taken into consideration. As a result, using the k -means does not give any informative clue for the selection of landmark pixels. Also, SC is not appropriate due to the large memory space. Here, we compare the performances of the proposed schemes IS and LS with WS, RS, and LSC. The number of samples is set as 50.

The left half of Table III shows the F -measure values of IS, LS, WS, RS, and LSC on the images. The best result is highlighted in boldface, and the worst one is highlighted in italic and underline. In the last three rows of Table III, we report the average result on the images, the number of images having the best performance, and the number of images having the worst performance. The results show that LS and IS are the best two in terms of performance on most of the images and they are the worst two on a very small number of images. The average results of LS and IS are relatively higher than that of other schemes. In terms of Accuracy, WS and RS are the best two on several images but they also are the worst two on a greater number of images. RS gets the lowest average result among the schemes. LSC performs the best on 12 images but also performs the worst on 16 images. Overall, we can conclude that IS and LS are more stable than other schemes and they are the best two sampling schemes for image segmentation tasks.

The right half of Table III shows the time consumption of the five schemes. We find that LSC is faster than RS, which is due to the fact that LSC does not use the standard routine of SVD. IS and LS take more time than RS and LSC, but they are much faster than WS.

4.6. Protein Complexes Detection Results

In this subsection, we compare the performances of LS, IS, WS, RS, KS, KASP, and LSC when applying to protein complexes detection in protein–protein interaction (PPI) networks. A PPI network is a biomolecule relationship network, which consists of a set of nodes representing the interacting proteins and a set of edges describing interactions between those proteins. Studies of protein complexes in a PPI network contribute greatly to the understanding of the biological mechanism. Spectral clustering has been widely applied to detect protein complexes from PPI networks [Christel and Kim 2005; Sen et al. 2006; Kentaro et al. 2010; Qin and Gao 2010].

We experiment on two PPI network datasets for yeast. The first dataset was constructed by Collins et al. [2007] with data from Gavin et al. [2006]. It consists of 68,438 PPIs among 2,574 yeast proteins. The second dataset was constructed by Collins et al. [2007] with data from Krogan et al. [2006]. It consists of 454,506 PPIs among 5,359 yeast proteins. To evaluate the complexes mined by clustering, the resulting complexes are matched to known protein complexes from Gavin et al. [2006] and Krogan et al. [2006], 478 and 547 in number, respectively. We use the method in Qin and Gao [2010] to determine the number of clusters. The number of samples is set as 800.

To evaluate the clustering results, we use sensitivity, positive predictive value (PPV), and accuracy [Brohe and Helden 2006]. Sensitivity describes how many of the reference complex sets are also contained in the candidate complex set. PPV measures how much of the prediction in the candidate set is correct. Accuracy is the geometric mean of sensitivity and PPV.

Tables IV and V show the protein complexes detection results and the time consumption of the schemes. On the Gavin PPI network, the sensitivity, PPV, and accuracy of LS and IS are higher than those of the other sampling schemes. On the Krogan PPI network, the sensitivity, PPV, and accuracy of LS and IS are higher than those of the other sampling schemes, besides KASP. The sensitivity of LS and IS is lower than that of KASP, whereas their PPV is higher than that of KASP. In an integrated manner, the accuracy of LS and IS is higher than that of KASP. Overall, we can conclude that the

Table III. F -1 and Time for Image Segmentation

Image ID	F -Measure					CPU Time (s)				
	LS	IS	WS	RS	LSC	LS	IS	WS	RS	LSC
8023	0.317	0.307	0.282	<u>0.242</u>	0.291	18.783	18.725	69.007	2.494	1.375
14037	0.571	0.523	0.523	<u>0.496</u>	0.577	18.199	18.456	48.353	1.911	1.357
19021	0.583	0.569	0.593	<u>0.518</u>	0.528	18.322	18.154	47.887	2.628	1.407
24077	<u>0.522</u>	0.554	0.525	0.542	0.524	18.637	18.295	56.600	2.114	1.310
33039	0.414	0.402	<u>0.388</u>	0.473	0.463	19.679	18.348	52.990	2.163	1.493
37073	0.615	0.584	0.518	0.502	<u>0.399</u>	18.736	18.073	46.301	2.213	1.431
38092	0.576	0.553	<u>0.552</u>	0.560	0.559	18.405	18.287	62.916	1.748	1.318
41069	0.408	0.376	0.411	<u>0.278</u>	0.417	18.719	18.967	50.180	1.981	1.350
42012	0.513	<u>0.457</u>	0.496	0.500	0.483	18.328	18.022	66.271	2.037	1.279
42049	0.702	0.771	<u>0.559</u>	0.687	0.638	17.955	17.782	45.778	1.786	1.352
43074	0.448	<u>0.360</u>	0.441	0.404	0.397	18.708	18.761	55.748	2.162	1.360
55073	0.487	0.495	0.482	0.515	<u>0.458</u>	18.759	19.014	50.378	2.182	1.348
58060	0.633	<u>0.487</u>	0.500	0.498	0.573	18.633	17.945	46.235	1.925	1.355
62096	0.504	0.503	0.545	0.676	<u>0.488</u>	19.062	18.249	46.159	1.724	1.286
65033	<u>0.518</u>	0.560	0.545	0.556	0.540	18.480	18.370	50.980	2.040	1.338
66053	0.523	0.521	<u>0.487</u>	0.610	0.548	18.016	18.439	49.103	1.870	1.320
69015	0.736	0.743	0.737	<u>0.690</u>	0.696	19.080	17.699	45.461	2.364	1.365
69040	0.380	<u>0.367</u>	0.410	0.403	0.403	18.341	18.374	77.834	2.269	1.354
76053	0.540	0.541	0.539	0.564	<u>0.526</u>	18.676	18.331	51.842	2.782	1.363
78004	0.683	0.683	0.658	<u>0.584</u>	0.690	18.587	18.847	54.252	2.364	1.376
85048	0.576	0.525	0.595	<u>0.520</u>	0.538	19.397	18.224	56.612	2.572	1.449
86000	0.550	0.550	0.580	<u>0.544</u>	0.583	19.136	18.365	52.088	2.211	1.359
86068	0.382	0.391	0.383	0.321	<u>0.310</u>	18.787	19.096	50.508	1.983	1.327
87046	0.511	0.510	0.462	<u>0.457</u>	0.496	18.814	18.841	61.049	1.935	1.394
89072	0.465	0.471	0.471	0.489	<u>0.445</u>	18.718	18.731	48.294	2.504	1.352
97033	0.633	0.599	<u>0.489</u>	0.587	0.589	18.738	18.092	54.310	1.801	1.336
101085	<u>0.587</u>	0.532	<u>0.455</u>	0.547	0.609	19.046	18.551	49.349	3.454	1.433
101087	<u>0.542</u>	0.621	0.647	0.578	0.591	18.368	17.782	44.377	1.997	1.284
102061	0.504	0.502	0.502	<u>0.435</u>	0.484	18.731	18.943	51.116	3.088	1.352
103070	0.533	0.535	0.544	<u>0.491</u>	0.587	18.419	18.586	51.483	2.677	1.352
105025	0.459	0.474	0.453	0.456	<u>0.445</u>	19.097	18.421	48.343	1.924	1.345
109053	0.383	0.405	0.398	0.399	<u>0.362</u>	19.489	18.438	47.137	2.081	1.377
119082	0.699	0.698	0.701	0.682	<u>0.671</u>	18.333	18.151	51.487	1.899	1.275
126007	0.575	0.578	0.576	<u>0.553</u>	0.567	18.259	18.106	45.147	1.788	1.391
130026	0.359	0.360	0.356	0.332	<u>0.322</u>	18.565	17.648	46.509	2.089	1.359
134035	0.560	0.639	<u>0.550</u>	0.588	0.574	19.072	18.113	50.137	2.133	1.294
143090	0.542	0.547	0.558	0.538	<u>0.500</u>	19.136	18.537	48.784	2.266	1.377
145086	0.654	0.652	0.653	<u>0.485</u>	0.646	18.665	18.409	48.863	2.150	1.743
147091	0.622	0.593	0.579	0.561	<u>0.543</u>	18.318	17.725	46.610	2.425	1.667
157055	0.562	<u>0.529</u>	0.558	0.532	0.540	18.652	18.294	78.509	1.808	1.317
159008	0.515	0.520	<u>0.514</u>	0.529	0.533	18.705	19.342	49.964	2.013	1.304
160068	0.479	0.621	0.479	<u>0.361</u>	0.493	18.854	18.417	46.239	1.869	1.267
163085	0.444	<u>0.406</u>	0.435	0.434	0.421	18.775	18.266	52.857	1.830	1.397
167062	<u>0.231</u>	0.248	0.247	0.379	0.254	18.647	18.224	45.905	2.218	1.270
167083	0.589	0.588	0.565	<u>0.513</u>	0.526	19.659	18.545	60.909	2.025	1.279
175043	0.633	0.573	<u>0.531</u>	0.603	0.610	18.736	18.738	51.056	2.169	1.535
216081	0.679	0.641	0.623	0.665	<u>0.580</u>	18.406	18.396	58.810	1.921	1.491
219090	0.520	0.519	<u>0.517</u>	0.537	0.554	18.699	18.222	52.664	1.671	1.404

(Continued)

Table III. Continued

227092	0.551	<u>0.524</u>	0.533	0.615	0.530	18.315	18.300	49.239	1.987	1.355
241004	0.525	0.509	0.518	<u>0.468</u>	0.593	18.764	17.885	46.004	2.407	1.337
241048	<u>0.532</u>	0.576	0.543	0.633	0.681	19.087	18.251	49.211	2.057	1.325
271035	<u>0.554</u>	0.593	0.590	0.609	0.612	18.844	18.339	52.564	1.931	1.473
285079	0.578	0.647	0.596	0.554	<u>0.521</u>	18.571	17.933	49.806	1.931	1.434
296007	0.614	0.652	0.448	<u>0.232</u>	0.451	18.725	17.794	54.189	2.208	1.372
296059	<u>0.548</u>	0.604	0.600	0.552	0.570	19.274	18.601	50.496	1.941	1.392
300091	0.448	0.490	0.483	0.386	<u>0.287</u>	19.133	18.321	45.428	2.056	1.438
302008	0.635	0.633	0.635	<u>0.592</u>	0.648	17.961	17.777	46.476	1.950	1.342
306005	0.559	0.557	0.560	0.578	<u>0.466</u>	18.402	18.798	45.524	2.974	1.296
361010	0.554	0.506	0.546	<u>0.465</u>	0.534	18.309	18.974	51.608	2.798	1.341
376043	0.474	0.472	0.470	<u>0.429</u>	0.432	18.645	19.018	61.892	2.562	1.282
AVG	0.534	0.532	0.519	0.509	0.515	18.702	18.369	52.069	2.163	1.371
#best	18	15	6	9	12	—	—	—	—	—
#worst	7	7	10	20	16	—	—	—	—	—

Table IV. Result on the Gavin PPI Network

Algorithm	Sensitivity	PPV	Accuracy	CPU Time (s)
RS	0.542	0.471	0.505	9.486
WS	0.552	0.473	0.511	247.996
KS	0.585	0.484	0.532	15.997
KASP	0.575	0.478	0.524	33.414
LSC	0.580	0.480	0.527	12.459
IS	0.625	0.511	0.565	20.915
LS	0.632	0.511	0.568	15.907

Table V. Result on the Krogan PPI Network

Algorithm	Sensitivity	PPV	Accuracy	CPU Time (s)
RS	0.323	0.234	0.275	10.792
WS	0.334	0.283	0.308	342.962
KS	0.357	0.284	0.319	38.918
KASP	0.416	0.280	0.341	85.636
LSC	0.366	0.277	0.319	13.721
IS	0.381	0.317	0.348	36.286
LS	0.377	0.312	0.343	24.979

performances of IS and LS are better than those of other sampling schemes for protein complexes detection tasks.

In terms of running time, WS is the most time consuming method. Such high time consumption is due to the fact that WS calculates the determinant of an $m \times m$ matrix in each iteration. The running time of the other sampling schemes is significantly less than WS. IS and LS take more time than RS and LSC, but they are faster than KASP and comparable with KS.

5. RELATED WORK

Recently, applying spectral clustering to large-scale problems has become a hot topic, and Nyström extension is an important method for large-scale spectral clustering. The Nyström method was introduced to the machine learning community to construct a low-rank approximation by Williams and Seeger [2000]. Kernel-based learning with Nyström extension has been extensively studied in recent years [Li et al. 2010; Farahat et al. 2011; Kumar et al. 2012; Wang and Zhang 2014]. Since landmark point sampling

is crucial for the performance of Nyström extension-based algorithms, many works have been devoted to sampling methods in the past decade. Here, we mainly review sampling schemes that have been applied to spectral clustering; one can refer to Kumar et al. [2012] for a comprehensive survey of sampling methods for kernel-based learning.

Fowlkes et al. [2004] were the first to address the normalization and orthogonality problems and successfully apply Nyström extension to spectral clustering. They used a random method to sample the landmark points. Drineas and Mahoney [2005] proposed the column-norm sampling strategy. In this algorithm, the probability of the i th row being sampled depends on the L_2 norm of the column, which needs to compute the norms of all the rows/columns of the kernel matrix and hence is quite expensive. Li et al. [2011] propose a time- and space-efficient spectral clustering algorithm, which can scale to very large datasets. Furthermore, Kumar et al. [2012] showed that the method is even inferior to RS. They also provided the performance bounds for different sampling methods, but their analysis mainly involved the problem of low-rank approximation. Zhang et al. [2008] introduced a k -means-based sampling method. They analyzed that the Nyström approximation error is mostly affected by the quantization error (error of quantizing each data point with the closest landmark point). As k -means is known to be capable of giving a local minimum of the quantization error, the centers obtained from k -means can be designated as the landmark points. However, their analysis is based on an assumption that the kernel K satisfies $K_{ij} - K_{st} \leq C(\|x_i - x_j\|^2 + \|x_s - x_t\|^2)$, limiting the applicability of the algorithm, e.g., not being applicable to image segmentation. Furthermore, k -means-based sampling adds supplementary data points into the dataset, which may lead to unsteady results. Belabbas and Wolfe [2009] proposed two other approaches, which use the determinant of the kernel matrix to sample landmark points. The first approach was a randomized one. Let I denote the set of indices of the landmark points, then the probability of choosing I was in proportion to the determinant of the similarity matrix between landmark points. They used the SC to analyze the Nyström reconstruction error and concluded that the bigger the determinant, the smaller the error. Let n be the number of total data samples and m be the number of landmark points, then there will be a total of $\binom{n}{m}$ candidates for I . To circumvent this heavy computational burden, they employed a simulation algorithm to choose the desired I . The second method chose I containing the indices of the k largest diagonal elements of the approximated matrix, which also needs to access all the columns of the matrix and has a huge complexity. Actually, our first incremental sampling scheme is also based on the SC, but we use the SC in a different way. Ding et al. [2014] designed an adaptive Nyström sampling method. The sampling probability of every data point would be updated after each sampling pass, and a proof was given that the sampling error will decrease exponentially with increase in sample times. Li et al. [2015] proposed an accurate and scalable Nyström scheme that first sampled a large column subset from the input matrix but then only performed an approximate SVD on the inner sub-matrix using the recent randomized low-rank matrix approximation algorithms. Lin et al. [2015] proposed a novel criterion named the Hilbert Space Embedding Nyström criterion, which directly minimized the eigenvector approximation error. The proposed criterion was closely related to the kernel density estimation, in which the kernel width also needs to be adaptively chosen based on the sampled data. Kang et al. [2015] designed a spectral clustering for large-scale social networks via the pre-coarsening sampling-based Nyström method. Wang et al. [2014] modified the Nyström method by using spectral shifting to solve the following problem. That is, when the spectrum of the symmetric positive semi-definite matrix decays slowly, the Nyström approximation is of low accuracy.

There are many other approaches to speed up spectral clustering. Hunter and Strohmer [2010] provided a general theory for approximate spectral clustering under

perturbations to the eigenvectors. Sakai and Imiya [2009], Khoa and Chawla [2012], and Gittens and Mahoney [2013] proposed algorithms that approximate the eigenvectors of the Laplacian matrix using random projections. Yan et al. [2009] developed a general framework for fast approximate spectral clustering in which a distortion-minimizing local transformation is first applied to the data. Li et al. [2013] proposed an algorithm based on the importance of rows. Cai and Chen [2015], Shamir and Tishby [2011], and Liu et al. [2013] proposed data transformation-based algorithms to compress or sparsify the data. Si et al. [2014] considered both low-rank and clustering structures of the kernel matrix. Semertzidis et al. [2015] and Li et al. [2015] proposed to utilize constraints to speed up spectral clustering. Couillet and Benaych-Georges [2015] analyzed the exact behavior of the eigenpairs of kernel graph Laplacians used for clustering by means of random matrix theory. However, these are beyond the scope of this article, since we aim to dig the capacity of Nyström extension-based spectral clustering.

6. CONCLUSIONS

In this article, we have investigated the problem of landmark points sampling for Nyström extension-based spectral clustering from an incremental perspective. We proposed an incremental sampling framework in which each landmark point is carefully selected by previous landmark points such that the landmark points work together well and provide a theoretical guarantee on the clustering performance. Instead of matrix approximation error analysis used in previous works, we provided novel analysis methods, clusterability analysis, and loss analysis for the incremental sampling framework. Based on these, we proposed two incremental sampling schemes for Nyström extension-based spectral clustering. Experiments have demonstrated the superiorities of the proposed incremental sampling schemes. In our current schemes, the initial landmark points are randomly selected. In future work, we will find smart initializations to improve the performances of the proposed schemes. We will also exploit our analysis methods and sampling scheme to other Nyström extension-based learning tasks.

REFERENCES

- Gavin Anne-Claude, Aloy Patrick, Grandi Paola, Krause Roland, Boesche Markus, Marzioch Martina, Rau Christina, Jensen Lars Juhl, Bastuck Sonja, and Dmpelfeld Birgit. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 7084 (Mar. 2006), 631–636.
- Mohamed-Ali Belabbas and Patrick J. Wolfe. 2009. Spectral methods in machine learning and new strategies for very large datasets. *Proc. Nat. Acad. Sci. USA* 51, 6 (Jan. 2009), 369–374.
- Sylvain Brohe and Jacques Van Helden. 2006. Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinform.* 7, 1602 (2006), 2791–2797.
- Deng Cai and Xinlei Chen. 2015. Large scale spectral clustering via landmark-based sparse representation. *IEEE Trans. Cybern.* 45, 8 (Aug. 2015), 1669–1680.
- Kamp Christel and Christensen Kim. 2005. Spectral analysis of protein–protein interactions in *Drosophila melanogaster*. *Phys. Rev. E* 71, 4 (Apr. 2005), 100–119.
- Sean R. Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F. Greenblatt, Forrest Spencer, Frank C. P. Holstge, Jonathan S. Weissman, and Nevan J. Krogan. 2007. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteom.* 6, 3 (Jan. 2007), 439–450.
- Romain Couillet and Florent Benaych-Georges. 2015. Understanding big data spectral clustering. In *Proceedings of the IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*. IEEE, Cancun, Mexico, 29–32.
- Chandler Davis and W. M. Kahan. 1970. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* 7, 1 (1970), 1–46.
- Shi Fei Ding, Hong Jie Jia, and Zhong Zhi Shi. 2014. Spectral clustering algorithm based on adaptive Nyström sampling for big data analysis. *J. Softw.* 25, 9 (2014), 2037–2049.
- Petros Drineas and Michael W. Mahoney. 2005. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* 6 (Dec. 2005), 2153–2175.

- Ahmed K. Farahat, Ali Ghodsi, and Mohamed Kamel. 2011. A novel greedy algorithm for Nyström approximation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR, Fort Lauderdale, USA, 269–277.
- Charless Fowlkes, Serge Belongie, Fan R. K. Chung, and Jitendra Malik. 2004. Spectral grouping using Nyström methods. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 2 (Feb. 2004), 214–225.
- Alex Gittens and Michael Mahoney. 2013. Revisiting the Nyström for improved large-scale machine learning. In *Proceedings of the 30th International Conference on Machine Learning*. IMLS, Atlanta, GA, USA, 567–575.
- Blake Hunter and Thomas Strohmer. 2010. Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements. *Comput. Res. Repository* (2010), arXiv: abs/1011.0997.
- Ying Kang, Bo Yu, Weiping Wang, and Dan Meng. 2015. Spectral clustering for large-scale social networks via a pre-coarsening sampling based Nyström method. In *Advances in Knowledge Discovery and Data Mining*. Springer, Ho Chi Minh City, Vietnam, 106–118.
- Inoue Kentaro, Li Weijiang, and Kurata Hiroyuki. 2010. Diffusion model based spectral clustering for protein-protein interaction networks. *PLoS One* 5, 9 (Sep. 2010), 2771–2794.
- Nguyen Lu Dang Khoa and Sanjay Chawla. 2012. Large scale spectral clustering using resistance distance and Spielman–Teng solvers. In *Proceedings of 15th International Conference on Discovery Science*. Springer, Lyon, France, 7–21.
- Nevan J. Krogan, Cagney Gerard, Yu Haiyuan, Zhong Gouqing, Guo Xinghua, Ignatchenko Alexandr, Li Joyce, Pu Shuye, Datta Nira, and Aaron P. Tikuisis. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 7084 (Mar. 2006), 637–643.
- Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. 2012. Sampling methods for the Nyström method. *J. Mach. Learn. Res.* 13 (Apr. 2012), 981–1006.
- Jianyuan Li, Yingjie Xia, Zhenyu Shan, and Yuncai Liu. 2015. Scalable constrained spectral clustering. *IEEE Trans. Knowl. Data Eng.* 27, 2 (Feb. 2015), 589–593.
- Mu Li, Wei Bi, James T. Kwok, and Bao-Liang. Lu. 2015. Large-scale Nyström kernel matrix approximation using randomized SVD. *IEEE Trans. Neural Netw. Learn. Syst.* 26, 1 (Jan. 2015), 152–164.
- Mu Li, James T. Kwok, and B.-L. Lu. 2010. Making large-scale Nyström approximation possible. In *Proceedings of the International Conference on Machine Learning*. ACM, Haifa, Israel, 631–638.
- Mu Li, Xiao-Chen Lian, James T. Kwok, and Bao-Liang Lu. 2011. Time and space efficient spectral clustering via column sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Colorado Springs, CO, USA, 2297–2304.
- Mu Li, Gary L. Miller, and Rongkun Peng. 2013. Iterative row sampling. In *Proceedings of the IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Berkeley, CA, USA, 127–136.
- Ming Lin, Fei Wang, and Changshui Zhang. 2015. Large-scale eigenvector approximation via Hilbert space embedding Nyström. *Pattern Recognit.* 48, 5 (May 2015), 1904–1912.
- Jialu Liu, Chi Wang, Marina Danilevsky, and Jiawei Han. 2013. Large-scale spectral clustering on graphs. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, Beijing, China, 1486–1492.
- David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Barcelona, Spain, 416–425.
- David R. Martin, Charless C. Fowlkes, and Jitendra Malik. 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 5 (May 2004), 530–549.
- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*. MIT Press, Vancouver, British Columbia, Canada, 849–856.
- Guimin Qin and Lin Gao. 2010. Spectral clustering for detecting protein complexes in protein–protein interaction (PPI) networks. *Math. Comput. Model.* 52, 11C12 (Dec. 2010), 2066–2074.
- Tomoya Sakai and Atsushi Imiya. 2009. Fast spectral clustering with random projection and sampling. In *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, Leipzig, Germany, 372–384.
- T. Semertzidis, D. Rafailidis, M. G. Strintzis, and P. Daras. 2015. Large-scale spectral clustering based on pairwise constraints. *Inf. Process. Manag.* 51, 5 (Sep. 2015), 616–624.
- Taner Z. Sen, Andrzej Kloczkowski, and Robert L. Jernigan. 2006. Functional clustering of yeast proteins from the protein–protein interaction network. *BMC Bioinform.* 7, 1 (Jul. 2006), 355.

- Ohad Shamir and Naftali Tishby. 2011. Spectral clustering on a budget. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR, Fort Lauderdale, USA, 661–669.
- Si Si, Cho-Jui Hsieh, and Inderjit S. Dhillon. 2014. Memory efficient kernel approximation. In *Proceedings of the 31th International Conference on Machine Learning*. ACM, Beijing, China, 701–709.
- Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3 (Dec. 2002), 583–617.
- Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang. 2014. Improving the modified Nyström method using spectral shifting. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 611–620.
- Shusen Wang and Zhihua Zhang. 2014. Efficient algorithms and error analysis for the modified Nyström method. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR, Reykjavik, Iceland, 996–1004.
- Christopher K. I. Williams and Matthias Seeger. 2000. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*. MIT Press, Vancouver, British Columbia, Canada, 682–688.
- Donghui Yan, Ling Huang, and Michael I. Jordan. 2009. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Paris, France, 907–916.
- Kai Zhang, Ivor W. Tsang, and James T. Kwok. 2008. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the International Conference on Machine Learning*. ACM, Helsinki, Finland, 1232–1239.
- Xianchao Zhang and Quanzeng You. 2011. Clusterability analysis and incremental sampling for Nyström extension based spectral clustering. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, New York, NY, USA, 942–951.

Received August 2014; revised May 2016; accepted May 2016