Sampling with Minimum Sum of Squared Similarities for Nyström-Based Large Scale Spectral Clustering

Djallel Bouneffouf and Inanc Birol
Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency
Vancouver, British Columbia, Canada
{dbouneffouf, ibirol}@bcgsc.ca

Abstract

The Nyström sampling provides an efficient approach for large scale clustering problems, by generating a low-rank matrix approximation. However, existing sampling methods are limited by their accuracies and computing times. This paper proposes a scalable Nyström-based clustering algorithm with a new sampling procedure, Minimum Sum of Squared Similarities (MSSS). Here we provide a theoretical analysis of the upper error bound of our algorithm, and demonstrate its performance in comparison to the leading spectral clustering methods that use Nyström sampling.

1 Introduction

The amount of large-scale data around us is increasing in size very quickly. Clustering those data with respect to certain features using computational approaches would help us interpret them in a timely manner. Hence, development of clustering algorithms is an active field of research [Kong *et al.*, 2011].

In recent years, spectral clustering has become increasingly popular, often outperforming alternative approaches on a range of datasets [Chen and Cai, 2011]. However, spectral clustering has limited applicability to large-scale problems, due to its computational time complexity. To address this computational challenge, a common approach is to use a low-rank matrix approximation [Zhang and You, 2011], [Chen and Cai, 2011]. The Nyström method is one such technique used in a number of machine learning problems [Fowlkes *et al.*, 2004], [Williams and Seeger, 2001]. The method works by first sampling a set of m landmark points from n instances, with $m \ll n$, to formulate an approximation for the eigen-decomposition of the full dataset using the sampled data.

The most important step of the Nyström method is sampling, because different sampled landmark points give different approximations of the original matrix. Uniform sampling without replacement is the most used approach for this purpose [Fowlkes *et al.*, 2004], [Cohen *et al.*, 2014], where every point has the same probability of being included in the sample. Alternatively, sampling can be performed using local or global properties of the data distribution. Different versions of non-uniform sampling have recently been suggested. For example, in [Zeng *et al.*, 2014] authors proposed an algorithm that considers the similarity between the sample set and the rest of the data points to select the landmark points. Similarly, [Zhang and You, 2011] proposed an algorithm, where points with the smallest variance between the sampled points and the rest of the data are selected as landmark points.

In this paper, we propose "Minimum Sum of Squared Similarities" (MSSS), an algorithm for incremental sampling in Nyström based-spectral clustering. MSSS considers both variance and similarity in its sampling data, increasing the speed of the clustering on large datasets. The algorithm starts sampling with a fixed number of initial landmark points and selects new landmark points one by one, such that the sum of the squared similarities between the previously selected points and the new point is minimized. We also discuss its upper bound on the Frobenius norm error.

2 Related Work

Derived from spectral graph theory [Von Luxburg, 2007], spectral clustering has a wide range of applicability, such as in community detection [Azam and Viktor, 2013], image segmentation [Fowlkes *et al.*, 2004] and clustering of microarray data [Higham *et al.*, 2007].

To apply spectral clustering to large datasets, recent efforts have been concentrating on solving issues around algorithm scalability [Zeng *et al.*, 2014], such as on reducing the time cost of eigen-decomposition of a Laplacian matrix (see 3.1) through parallel computations [Chen *et al.*, 2011], [Wang *et al.*, 2014]. Another approach is to use dimension reduction by Nyström approximation [Williams and Seeger, 2001], a method originally designed for numerical solution of integral equations [Sloan, 1981].

Recently, due to their demonstrated effectiveness, Nyström approximation-based machine learning algorithms are gaining popularity [Fu, 2014]. However the performance of the approach is highly dependent on proper subsampling of the input data to include some *landmark points*, points that capture the inherent complexity and variability of the full dataset. To address this, different sampling methods have been proposed, some as simple as random sampling (RS) [Fowlkes *et al.*, 2004], [Cohen *et al.*, 2014]. However, although it is straightforward to implement, RS makes an implicit assumption that clusters have an equiprobable distribution, which may be a limiting assumption for certain datasets [Fowlkes *et al.*, 2004].

Two proposed alternatives use diagonal sampling [Drineas and Mahoney, 2005] and column-norm sampling [Drineas *et al.*, 2006] algorithms, and are very effective. However, they are shown to perform poorly in various test cases in comparison to RS [Kumar *et al.*, 2009b].

In [Belabbas and Wolfe, 2009] the authors developed a weighted sampling (WS) approach using the determinant of the kernel matrix to select landmark points, where the probability of choosing a new landmark point was in proportion to the determinant of the similarity matrix between landmark points. They analyzed the Nyström reconstruction error using the Schur complement [Gowda and Sznajder, 2010], concluding that the larger the determinant, the smaller the error. Although the work provides a solid theoretical basis for measuring the error levels in Nyström approximation, a main drawback of the algorithm provided is in its time complexity.

Assuming that the potential clusters are convex, [Zhang *et al.*, 2008] introduced k-means based sampling (KS) algorithm, as a means to select points near k-means centroids as landmark points. Similarly, [Shinnou and Sasaki, 2008] also pre-processed the data using k-means clustering, to select a *committee* of data points near centroids. Although the latter method does not explicitly state the convexity assumption, both methods perform poorly for non-convex clusters and when clusters are grouped in certain configurations, such as some of the test datasets shown in Figure 1.

In [Zhang and You, 2011], the authors proposed an incremental sampling (IS) algorithm that first randomly samples two points from a dataset, to compute a similarity matrix between the sampled points and the remaining points. The algorithm picks the point with the smallest variance, and then iteratively repeats the process until a desired number of landmarks is reached. While promising, [Zeng *et al.*, 2014] showed that IS performs poorly on high-dimensional data, as the variance of the Euclidean distance tends to zero. In such cases IS may pick inappropriate landmark points for dimension reduction, hence for successful clustering.

The authors in [Zeng *et al.*, 2014] studied how the similarity between the sample set and non-sample set influences the approximation error, and proposed minimum similarity sampling (SS) for high-dimensional space clustering. However, their result depends on the dimensionality of the dataset: SS outperforms IS on high-dimensional data, but not on low dimensional data.

To address the problems raised in [Zeng *et al.*, 2014] and [Zhang and You, 2011], we propose a new sampling algorithm, MSSS, which approximately maximizes the determinant of the reduced similarity matrix that represents the mutual similarities between sampled data points. We demonstrate the performance of MSSS in comparison to leading sampling methods using synthetic benchmark datasets, as well as a University of California, Irvine (UCI) Machine Learning Repository dataset.

3 Key Notion

3.1 Spectral clustering

Spectral clustering algorithms employ the first k eigenvectors of a Laplacian matrix to guide clustering. Loosely following the notation in [Von Luxburg, 2007], this can be outlined as follows.

Algorithm 1 Spectral Clustering Algorithm

1: **Input:** Affinity matrix $S \in R^{n \times n}$, number of clusters to construct, k
2: **Output:** Clusters $c_1, ..., c_k$
3: Compute the Laplacian matrix $P = D - S$; where D is an $n \times n$ diagonal matrix defined by $D_{ii} = \sum_{j=1}^{n} S_{ij}$
4: Compute k eigenvectors $u_1, ..., u_k$ corresponding to the largest k eigenvalues of the generalized eigenproblem $Pu = \lambda Du$; and let $Z \in R^{n \times k}$ be the matrix containing the vectors $u_1, ..., u_k$.
5: Cluster $y_1, ..., y_n$ by k-means algorithm into clusters $c_1, ..., c_k$; with y_i corresponding to the i-th row of Z.

Here, S is the symmetric affinity matrix constructed using the cosine similarity between each pairs of data points.

By analyzing the spectrum of the Laplacian matrix constructed over all data entries, the original data can be compressed into a smaller number of representative points using the Nyström approximation described below.

3.2 Nyström Sampling

The Nyström sampling method was originally designed to approximate the solutions of integral equations of the form:

$$\int_0^1 sim(x, y)\phi(y)dy = \lambda\phi(x) \tag{1}$$

where $x, y \in R$, $\phi(x)$ represents the eigenfunction, and $sim(x, y)$ denotes the similarity between x and y.

If we consider m landmark data points $L = l_1, l_2, ..., l_m$ from a given dataset $X = x_1, x_2, ..., x_n$ with $x_i \in R^n$ and $m \ll n$, then for any given point x in X, Nyström method formulates

$$\frac{1}{m}\sum_{i=1}^{m} sim(x, l_i)\hat{\phi}(l_i) = \hat{\lambda}\hat{\phi}(x) \tag{2}$$

where $\hat{\phi}(x)$ is an approximation to the exact eigenfunction, and $\hat{\lambda}$ is the corresponding approximate eigenvalue. Note that, Eq.2 cannot be solved directly, as both $\hat{\phi}(x)$ and $\hat{\lambda}$ are unknown.

If we denote the similarity matrix between the landmark points by $\widetilde{S}$ with $\widetilde{s}_{ij} = sim(l_i, l_j)$, and substitute x with l_i in Eq.2, we can write it in matrix form,

$$\widetilde{S}\hat{\Phi} = m\hat{\Phi}\hat{\Lambda} \tag{3}$$

where $\hat{\Phi} = [\hat{\phi}_1\hat{\phi}_2...\hat{\phi}_m]$ are the eigenvectors of $\widetilde{S}$ and $\hat{\Lambda} = diag\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m\}$ is a diagonal matrix of the corresponding approximate eigenvalues.

For an unsampled point x, the j-th eigenfunction at x can be approximated by

$$\hat{\phi}_j(x) \simeq \frac{1}{m\hat{\lambda}_j}\sum_{i=1}^{m} sim(x, l_i)\hat{\phi}_j(l_i) \tag{4}$$

With the equation above, the eigenvector for any given point x can be approximated through the eigenvectors of the landmark points L [Belabbas and Wolfe, 2009].

The same idea can be applied to extend the solution of a reduced matrix eigenvalue problem, to approximate the eigenvectors of a similarity matrix. Specifically, one may approximate k eigenvectors of S by decomposing and then extending a $k \times k$ principal sub-matrix of S.

First, let S be partitioned as

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

with $A \in R^{k \times k}$.

Now, define spectral decompositions $S = U \Lambda U^T$ and $A = U_A \Lambda_A U_A^T$; the Nyström extension then provides an approximation for k eigenvectors in

$$\widetilde{U} = \begin{bmatrix} U_A \\ B U_A \Lambda_A^{-1} \end{bmatrix} \quad (6)$$

where the approximations of $\widetilde{U} \approx U$ and $\widetilde{\Lambda} \approx \Lambda$ may then be composed, yielding an approximation $\widetilde{S} \approx S$ according to

$$\widetilde{S} = \widetilde{U} \Lambda_A \widetilde{U}^\top = \begin{bmatrix} A & B^\top \\ B & B A^{-1} B^\top \end{bmatrix} \quad (7)$$

We call $\widetilde{S}$ the Nyström approximation to S.

We note from Eq.6 that the main computational load to calculate this approximation is centred around the principal sub-matrix A of dimension $k < n$, and hence the Nyström extension provides a practical scalability for the spectral decomposition, hence for the spectral clustering problem.

4 Proposition

Before we present our approach, let us consider two toy examples to motivate the proposed algorithm.

Example 1. *Assume that in a given dataset $X = \{x_1, x_2, ..., x_n\}$ the data points belong to four clusters. In the ideal case of zero inter-cluster similarities, when three landmark points are selected, the similarity matrix would be $\widetilde{S} = I_3$ where I_a is the $a \times a$ identity matrix with $a = 3$. Similarly, for a fourth landmark point x_i, we would ideally have $\widetilde{S}_{\cup\{x_i\}} = I_4$, implying that the new landmark point is again not at all similar to the precedent landmarks points.*

Suppose, however, that we have to choose between two non-ideal landmark points, x_1 and x_2, such that

$$\widetilde{S}_{\cup\{x_1\}} = \begin{bmatrix} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0.2 \\ 0 & 0 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 \end{bmatrix}$$

and

$$\widetilde{S}_{\cup\{x_2\}} = \begin{bmatrix} 1 & 0 & 0 & 0.1 \\ 0 & 1 & 0 & 0.3 \\ 0 & 0 & 1 & 0.2 \\ 0.1 & 0.3 & 0.2 & 1 \end{bmatrix}$$

It was shown in [Zhang and You, 2011] that the larger the determinant of the reduced dimension similarity matrix, the smaller the error of the Nyström approximation. Hence, in this case, one should pick x_1 over x_2, because $det(\widetilde{S}_{\cup\{x_1\}})$ is larger than $det(\widetilde{S}_{\cup\{x_2\}})$.

However, the search for an optimum landmark point involves repeated computation of determinants, which has a time complexity of $O(n^3)$. Our motivation is to reduce this to $O(mn)$ by considering the sum of squared similarities between a landmark candidate and the preceding landmark data points.

In effect, by this approach we consider both the variance and the sum of the similarities between the new and preceding landmark data points, as proposed by [Zeng et al.*, 2014] and [Zhang and You, 2011], respectively, since by definition*

$$\frac{1}{m-1} \sum_{i<m} b_i^2 = var(b) + \left\{ \frac{1}{m-1} \sum_{i<m} b_i \right\}^2$$

where b is the similarity vector between a new (m^{th}) landmark point and the preceding $(m-1)$ landmarks, and $var(b)$ is the variance of b.

In this example, this will allow us to pick x_1 over x_2, while the sum of similarities would not be able to distinguish between the two choices. The variance method would similarly pick x_1 over x_2.

Example 2. *Expanding on the above example, let us now consider a third candidate for the new landmark, x_3, such that*

$$\widetilde{S}_{\cup\{x_3\}} = \begin{bmatrix} 1 & 0 & 0 & 0.3 \\ 0 & 1 & 0 & 0.3 \\ 0 & 0 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}$$

In this case, again x_1 is preferable over x_3, but this time the variance method will not be able to pick between x_1 and x_3, while sum of similarities and sum of squared similarities would both prefer x_1, as they should.

4.1 Theoretical analysis

Next, following the formalism of [Zhang and You, 2011] we show that, under certain assumptions, selecting landmarks incrementally in a way that minimizes the sum of the squared similarities between the new and the precedent landmark points would maximize the determinant of the landmark similarity matrix.

Theorem 1. *For a given dataset $X = \{x_1, x_2, \ldots, x_n\}$, let S be a real $n \times n$, symmetric positive definite matrix, where its entries $s_{\alpha\beta} = sim(x_\alpha, x_\beta)$ describe the similarity (in $[0, 1]$) between x_α and x_β, with $s_{\alpha\alpha} = 1$. Let $\widetilde{S}$ be an optimum $(m-1)^{st}$ order Nyström approximation of S, such that $\widetilde{S} \simeq I$. Denote the landmark data points that constitute this Nyström approximation with $\widetilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{m-1}\}$. Then, for two data points x_p and x_q in X but not in $\widetilde{X}$, if*

$$\sum_{i<m} sim^2(x_p, \tilde{x}_i) \leq \sum_{i<m} sim^2(x_q, \tilde{x}_i)$$

then

$$det(\widetilde{S}_{\cup\{x_p\}}) \gtrapprox det(\widetilde{S}_{\cup\{x_q\}})$$

Therefore, the optimum new landmark data point $\tilde{x}_m$ that minimizes $\sum_{i<m} sim^2(\tilde{x}_m, \tilde{x}_i)$ approximately maximizes $det(\widetilde{S}_{\cup\{\tilde{x}_m\}})$.

Proof. First, consider the augmented Nyström approximation, with

$$\widetilde{S}_{\cup\{\tilde{x}_m\}} = \begin{bmatrix} & \widetilde{S} & \cdot & b \\ \cdot & \cdot & \cdot & \cdot \\ & b^\top & \cdot & 1 \end{bmatrix} \tag{8}$$

where $b = [b_1, ..., b_{m-1}]^\top$ is the vector of similarities between new landmark $\tilde{x}_m$ and the preceding landmarks $\tilde{x}_i$, i.e., $b_i = sim(\tilde{x}_i, \tilde{x}_m)$ for $i = 1, \ldots, m-1$.

To take the determinant of this augmented matrix, one can consider two levels of co-factor expansions over column m then by row $m-1$, to give

$$det(\widetilde{S}_{\cup\{\tilde{x}_m\}}) = det(\widetilde{S}) - \sum_{i=1}^{m-1} b_i^2 det(\widetilde{S}_{/i}) \tag{9}$$
$$-2 \sum_{i,j=1 \wedge i \neq j}^{m-1} (-1)^{i+j} b_i b_j det(M_{<im>,<(m-1)j>})$$

where $M_{<im>,<(m-1)j>}$ is the minor $< (m-1)j >$ of $M_{<im>}$, which in turn is the minor $< im >$ of $\widetilde{S}_{\cup\{\tilde{x}_m\}}$, and $\widetilde{S}_{/i}$ is a matrix obtained by deleting the rows and columns $i \in 1, ..., m-1$ of the matrix $\widetilde{S}_{\cup\{\tilde{x}_m\}}$.

Note that, because $\widetilde{S} \simeq I$, all determinants of the second order cofactors will be negligible, i.e.,$det(M_{<im>,<(m-1)j>}) \simeq 0$, and the double summation in the second line of Eq.9 will drop.

Next, due to the constructive nature of the theorem, we will have $\widetilde{S}_{/i} \simeq I$ for all $i \in 1, ..., m-1$. Thus, we can write

$$\max_i \left| det(\widetilde{S}_{/i}) - det(I) \right| < \epsilon$$

for some $\epsilon > 0$, which puts the second term on the right hand side of Eq.9 in some $\delta > 0$ neighbourhood of $\sum_{i=1}^{m-1} b_i^2$, where $\delta < \epsilon \sum_{i=1}^{m-1} b_i^2$.

Finally, this gives us,

$$det(\widetilde{S}_{\cup\{\tilde{x}_m\}}) \simeq det(\widetilde{S}) - \sum_{i=1}^{m-1} b_i^2.$$

Hence, minimizing $\sum_{i=1}^{m-1} b_i^2$ approximately maximizes $det(\widetilde{S}_{\cup\{\tilde{x}_m\}})$. □

4.2 Minimum Sum of Squared Similarities

Our algorithm is directly deduced from **Theorem 1**, where rather than finding a new landmark point that maximizes the determinant, we find a point that minimizes the sum of squared similarities (MSSS). The MSSS algorithm initially randomly chooses two points from the dataset X. It then computes the sum of similarities between the sampled points and a subset, T, selected randomly from the remaining data points. The point with the smallest sum of squared similarities is then picked as the next landmark data point. The procedure is repeated until a total of m landmark points are picked.

Algorithm 2 The Minimum Sum of Squared Similarities Algorithm

Input: $X = \{x_1, x_2, ..., x_n\}$: dataset
m: number of landmark data points
γ: size of the subsampled set from the remaining data, in percentage
Output: $\widetilde{S} \in R^{m \times m}$: similarity matrix between landmark points
Initialize $\widetilde{S} = I_0$
For (i=0 to i<2) do
    $\widetilde{x}_i = Random(X)$
    $\widetilde{S} := \widetilde{S}_{\cup x_i}$
    $\widetilde{X} := \widetilde{X} \cup \{\widetilde{x}_i\}$
End For
While $i < m$ **do**
    $T = Random(X \backslash \{\widetilde{X}\}, \gamma)$
    Find $\widetilde{x}_i = argmin_{x \in T} \sum_{j<i-1} sim^2(x, \widetilde{x}_j)$
    $\widetilde{S} := \widetilde{S}_{\cup \widetilde{x}_i}$
    $\widetilde{X} := \widetilde{X} \cup \{\widetilde{x}_i\}$
End While

Theorem 2. *For a dataset $X = \{x_1, x_2, ..., x_n\}$, define the following positive definite similarity matrices:*

- *S: the $n \times n$ similarity matrix for the overall dataset with a maximum diagonal entry $S_{\max}$;*
- *$\widetilde{S}_l$: a similarity matrix with l landmark point selected randomly from X;*
- *$\widetilde{S}_m$: a similarity matrix with m landmark point selected using MSSS, with $m \leq l \leq n$; and*
- *S_m: the best rank-m approximation of S.*

Then with some probability $1-p$ or more, we can write

$$||S - \widetilde{S}_m|| \leq (m+1) \sum_{i=m+1}^{n} \lambda_i + ||S - S_m|| \tag{10}$$
$$+ \; nS_{\max} \sqrt[4]{\frac{64m}{l}} \left(1 + \sqrt{\frac{wd_S^*}{S_{\max}}}\right)^{\frac{1}{2}}$$

where

$$d_S^* = \max_{ij} (S_{ii} + S_{jj} 2S_{ij})$$

and

$$w = -\frac{n-1}{2n-1} \frac{2}{\beta(l,n)} \log p$$

with

$$\beta(l,n) = 1 - \frac{1}{2\max\{l, n-l\}}$$

Proof. Using the above notation, let us introduce some facts.

Fact 1 [Belabbas and Wolfe, 2009] Let $\lambda_1 \leq ... \leq \lambda_n$ be the eigenvalues of the similarity matrix S. Then

$$||S - \widetilde{S}_l|| \leq (l+1) \sum_{i=l+1}^{n} \lambda_i \tag{11}$$

Fact 2 [Kumar *et al.*, 2009a] With some probability $1-p$ or higher, we can write the following inequality:

$$||\widetilde{S}_l - \widetilde{S}_m|| \leq ||S - S_m|| + nS_{\max} \sqrt[4]{\frac{64m}{l}} \left(1 + \sqrt{\frac{wd_S^*}{S_{\max}}}\right)^{\frac{1}{2}} \tag{12}$$

Then, adding both sides of Eq.11 and Eq.12, noting that $\sum_{i=m+1}^{n} \geq \sum_{i=l+1}^{n}$, and using the Triangle inequality

$$||S - \widetilde{S}_m|| \leq ||S - \widetilde{S}_l|| + ||\widetilde{S}_l - \widetilde{S}_m|| \tag{13}$$

we prove **Theorem 2**. □

5 Evaluation

We tested MSSS with two subsampling percentages, $\gamma = 0$ and 10, (MSSS and MSSS-10, respectively), and compared its performance to the results of three state-of-the-art approaches described in Section 2: random sampling (RS), k-means sampling (KS) [Zhang *et al.*, 2008] and minimum similarity sampling (SS) [Zeng *et al.*, 2014]. Note that, we did not compare our algorithm to WS and IS, since it was shown earlier that SS performs better than WS and IS [Zeng *et al.*, 2014].

We required each algorithm to sample 1%, 2%,..,10% of the data as landmark points, which are used by Nyström-based spectral clustering methods to cluster the datasets.

Because sampling algorithms are sensitive to the datasets used, and clustering algorithms contain a degree of randomness, we used various benchmark datasets, and repeated our evaluations 1000 times. We measured the clustering quality of each algorithm using their average accuracy across these tests, also recording their standard deviations.

5.1 Clustering Synthetic Data

We evaluated algorithms on eight commonly used synthetic datasets[1] with different shapes (Figure 1), instances and number of clusters (Table 1). We note that, on these datasets, all tested algorithms have better performance than the baseline of random sampling.

Table 2 reports the average accuracy of each algorithm, along with their standard deviations across 1000 tests. As expected, the accuracies depend on the dataset. For example, the accuracy of all algorithms in A.K Jain's Toy problem and PathBased1 datasets stay in the range of $50\% - 60\%$, while going as high as over 90% for the R15 and D31 datasets. We can say from this observation that the A.K Jain's Toy problem and PathBased1 datasets present difficulties to Nyström method-based spectral clustering.

In the eight datasets, we see that overall MSSS algorithms perform better than the baseline random sampling. MSSS gave the best results in three datasets and MSSS-10 gave the best results in two datasets, indicating that the MSSS method offers a flexible approach, and that tuning the γ parameter for specific datasets might help for its optimum performance.

The results of the SS algorithm show overall better performance compared to KS and RS sampling. SS also gave the best performance on the R15 dataset, which has a uniform density distribution. When the density distribution is not uniform though MSSS gave higher performance than SS.

Figure 1: Eight synthetic datasets used for benchmarking

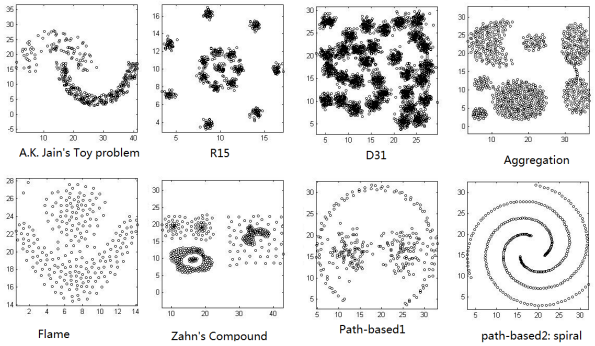

Table 1: Datasets used for benchmarking

Dataset	Instances	Attributes	Classes
Synthetic Datasets			
A.K Jain's Toy problem	373	2	2
R15	600	2	15
D31	3100	2	31
Aggregation	788	2	7
Flame	240	2	2
Zahn's Compound	399	2	6
Pathbased1	300	2	3
Pathbased2:Spiral	312	2	3
UCI Datasets			
Iris	150	4	3
Wine	178	13	3
Glass	214	9	7
Breast	699	9	2
Wdbc	569	32	2
Haberman	306	3	2
DUMD	259	5	4

5.2 Clustering Real Data

We also compared the performance of the four sampling methods using data from University of California, Irvine (UCI) Machine Learning Repository[2]. We randomly chose seven datasets: Iris, Wine, Glass, Breast, Wdbc, Haberman and Data User Modeling Dataset (DUMD). A brief summary of the datasets is listed in Table 1.

The clustering accuracy results on the UCI datasets are shown in Table 2. The results show that MSSS provided better clustering than the other algorithms on five datasets. Ranking the algorithms with respect to their mean accuracies, we note that the top three performing algorithms were MSSS, MSSS-10 and SS, in that order. Results on the UCI dataset

[1] http://cs.joensuu.fi/sipu/datasets/

[2] https://archive.ics.uci.edu/ml/datasets.html

Table 2: Accuracy on Datasets

	MSSS	MSSS-10	SS	RS	KS
Synthetic Datasets					
A.K Jain's Toy problem	**61.06 ± 0.63**	**61.06 ± 0.63**	**61.06 ± 0.63**	**61.06 ± 0.63**	53.79 ± 0.68
R15	91.53 ± 0.85	90.46 ± 0.80	**91.73 ± 0.78**	89.72 ± 1.20	89.71 ± 1.27
D31	94.79 ± 0.14	94.81 ± 0.13	94.80 ± 0.14	95.07 ± 0.30	**95.07 ± 0.29**
Aggregation	78.88 ± 0.98	**79.03 ± 0.98**	78.90 ± 0.92	74.73 ± 1.18	74.51 ± 1.30
Flame	73.69 ± 8.17	**73.70 ± 8.17**	73.69 ± 8.17	73.69 ± 8.17	70.91 ± 6.40
Zahn's Compound	**71.52 ± 1.98**	70.35 ± 2.08	71.51 ± 1.85	65.73 ± 3.39	65.69 ± 3.67
PathBased1	**62.71 ± 4.89**	58.49 ± 7.15	61.86 ± 5.50	57.32 ± 4.77	56.21 ± 4.25
Pathbased2:Spiral	**72.40 ± 4.34**	72.27 ± 4.46	72.38 ± 4.06	71.50 ± 4.84	66.94 ± 5.20
UCI Datasets					
Iris	69.06 ± 4.30	68.74 ± 4.46	68.76 ± 4.37	75.16 ± 6.70	**75.86 ± 7.66**
Wine	59.71 ± 4.71	**60.37 ± 4.81**	60.36 ± 4.67	59.59 ± 4.83	60.17 ± 5.19
Glass	**72.71 ± 4.01**	72.41 ± 4.04	72.10 ± 4.01	72.51 ± 4.16	71.36 ± 2.46
Breast	62.46 ± 4.09	62.50 ± 4.00	**62.71 ± 4.01**	58.64 ± 3.48	58.75 ± 3.50
Wdbc	**50.38 ± 0.27**	50.37 ± 0.27	50.37 ± 0.27	50.37 ± 0.27	50.09 ± 0.27
Haberman	**55.33 ± 1.46**	55.22 ± 1.44	55.30 ± 1.47	55.10 ± 1.39	55.19 ± 1.37
DUMD	58.97 ± 2.57	**59.08 ± 2.40**	59.07 ± 2.81	57.87 ± 3.01	57.56 ± 3.22

recapitulate our observation on synthetic data that γ in our algorithm can be tuned according to the dataset to yield better performance.

5.3 Comparison on Computation Time

To compare the computational time of the tested sampling algorithms, we measured the average run times for each algorithm over 1000 replicates (Table 3). Overall, we observed that all the tested algorithms have a higher computational time than RS, and that KS algorithm consumed much more time compared to other algorithms. We speculate that the poor time performance of the KS algorithm is due to the fact that it calculates k-means clusters – a computationally expensive process. The other algorithms had similar computational times. Hence we conclude that the accuracy improvement in MSSS is achieved with comparable computational time with respect to the state-of-the-art sampling algorithms.

6 Conclusion

In this paper, we introduced a new sampling algorithm, MSSS, for Nyström method-based spectral clustering. The new algorithm is scalable; it can handle large-scale data with computational times comparable to the state-of-the-art. Through theoretical analyses, we provided an upper bound to the matrix approximation error for the algorithm. In benchmarking experiments we demonstrated the competitive performance of MSSS over the current state-of-the-art.

Intriguing as these results are though, we note that the second best algorithm on every dataset, be it synthetic or real, was always within one standard deviation away from the top performing algorithm. This implies that achieving performance improvement in this problem domain is becoming increasingly difficult. Yet, the MSSS algorithm demonstrates a marginal but consistent advantage over the state-of-the art. Further, the new algorithm comes with an important parameter that can be tuned for better performance over various datasets, promising added gains if it can be optimized over any given dataset, hence paving the way for future research.

MSSS is implemented in java, and its source code is available from: www.bcgsc.ca/platform/bioinfo/software/msss

7 Acknowledgments

The authors thank the funding organizations, Genome Canada, British Columbia Cancer Foundation, and Genome British Columbia for their support of this work. We also thank Dr. Victoria Stuart for her critical reading of the draft manuscript.

References

[Azam and Viktor, 2013] Nadia Farhanaz Azam and Herna L Viktor. *Spectral clustering: An explorative study of proximity measures*. Springer, 2013.

[Belabbas and Wolfe, 2009] Mohamed-Ali Belabbas and Patrick J Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009.

[Chen and Cai, 2011] Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[Chen *et al.*, 2011] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y Chang. Parallel spectral clustering in distributed systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):568–586, 2011.

[Cohen *et al.*, 2014] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. *arXiv preprint arXiv:1408.5099*, 2014.

[Drineas and Mahoney, 2005] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

[Drineas *et al.*, 2006] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for ma-

Table 3: Average Computational Time of Tested Algorithms (ms)

	MSSS	MSSS-10	SS	RS	KS
Synthetic Datasets					
A.K Jain's Toy problem	669 ± 68	699 ± 72	664 ± 57	$\mathbf{664 \pm 58}$	737 ± 73
R15	2324 ± 202	2333 ± 222	2498 ± 242	$\mathbf{2249 \pm 201}$	2778 ± 302
D31	1278 ± 322	1299 ± 321	$\mathbf{1232 \pm 302}$	1327 ± 332	1525 ± 342
Aggregation	2597 ± 291	$\mathbf{2549 \pm 302}$	2579 ± 262	2581 ± 263	2624 ± 274
Flame	101 ± 32	103 ± 32	95 ± 29	$\mathbf{95 \pm 28}$	198 ± 30
Zahn's Compound	2007 ± 135	1950 ± 136	1949 ± 125	$\mathbf{1947 \pm 123}$	2239 ± 170
PathBased1	$\mathbf{155 \pm 51}$	158 ± 53	157 ± 48	157 ± 47	170 ± 49
Pathbased2:Spiral	345 ± 56	344 ± 58	339 ± 49	$\mathbf{338 \pm 50}$	354 ± 56
Real Datasets					
Iris	237 ± 11	231 ± 10	222 ± 10	$\mathbf{221 \pm 10}$	283 ± 18
Wine	96 ± 32	92 ± 33	92 ± 28	$\mathbf{92 \pm 28}$	196 ± 31
Glass	459 ± 73	456 ± 75	452 ± 67	$\mathbf{451 \pm 67}$	494 ± 84
Breast	582 ± 44	501 ± 46	512 ± 36	$\mathbf{386 \pm 28}$	591 ± 28
Wdbc	101 ± 43	102 ± 41	101 ± 42	$\mathbf{98 \pm 41}$	111 ± 42
Haberman	124 ± 41	122 ± 43	123 ± 37	$\mathbf{121 \pm 35}$	133 ± 37
DUMD	353 ± 58	350 ± 59	347 ± 52	$\mathbf{347 \pm 51}$	379 ± 61

trices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.

[Fowlkes *et al.*, 2004] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nystrom method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):214–225, 2004.

[Fu, 2014] Zhouyu Fu. Optimal landmark selection for nystrm approximation. In ChuKiong Loo, KeemSiah Yap, KokWai Wong, Andrew Teoh, and Kaizhu Huang, editors, *Neural Information Processing*, volume 8835 of *Lecture Notes in Computer Science*, pages 311–318. Springer International Publishing, 2014.

[Gowda and Sznajder, 2010] M Seetharama Gowda and Roman Sznajder. Schur complements, schur determinantal and haynsworth inertia formulas in euclidean jordan algebras. *Linear Algebra and Its Applications*, 432(6):1553–1559, 2010.

[Higham *et al.*, 2007] Desmond J Higham, Gabriela Kalna, and Milla Kibble. Spectral clustering and its use in bioinformatics. *Journal of computational and applied mathematics*, 204(1):25–37, 2007.

[Kong *et al.*, 2011] Tengteng Kong, Ye Tian, and Hong Shen. A fast incremental spectral clustering for large data sets. In *Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2011 12th International Conference on*, pages 1–5. IEEE, 2011.

[Kumar *et al.*, 2009a] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble nystrom method. In *Advances in Neural Information Processing Systems*, pages 1060–1068, 2009.

[Kumar *et al.*, 2009b] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 553–560. ACM, 2009.

[Shinnou and Sasaki, 2008] Hiroyuki Shinnou and Minoru Sasaki. Spectral clustering for a large data set by reducing the similarity matrix size. In *LREC*, 2008.

[Sloan, 1981] Ian H Sloan. Quadrature methods for integral equations of the second kind over infinite intervals. *Mathematics of computation*, 36(154):511–523, 1981.

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

[Wang *et al.*, 2014] Minchao Wang, Wu Zhang, Wang Ding, Dongbo Dai, Huiran Zhang, Hao Xie, Luonan Chen, Yike Guo, and Jiang Xie. Parallel clustering algorithm for large-scale biological data sets. *PloS one*, 9(4):e91315, 2014.

[Williams and Seeger, 2001] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, pages 682–688, 2001.

[Zeng *et al.*, 2014] Zhicheng Zeng, Ming Zhu, Hong Yu, and Honglian Ma. Minimum similarity sampling scheme for nyström based spectral clustering on large scale high-dimensional data. In *Modern Advances in Applied Intelligence*, pages 260–269. Springer, 2014.

[Zhang and You, 2011] Xianchao Zhang and Quanzeng You. Clusterability analysis and incremental sampling for nyström extension based spectral clustering. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 942–951. IEEE, 2011.

[Zhang *et al.*, 2008] 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.