Beyond the Nyström Approximation: Speeding up Spectral Clustering using Uniform Sampling and Weighted Kernel $k$-means

Mahesh Mohan, Claire Monteleoni
George Washington University, Washington D.C.
mahesh_mohan@gwu.edu, cmontel@gwu.edu

Abstract

In this paper we present a framework for spectral clustering based on the following simple scheme: sample a subset of the input points, compute the clusters for the sampled subset using weighted kernel $k$-means (Dhillon et al. 2004) and use the resulting centers to compute a clustering for the remaining data points. For the case where the points are sampled uniformly at random without replacement, we show that the number of samples required depends mainly on the number of clusters and the diameter of the set of points in the kernel space. Experiments show that the proposed framework outperforms the approaches based on the Nyström approximation both in terms of accuracy and computational time.

1 Introduction

Clustering is one of the fundamental problems in machine learning. Spectral clustering techniques [Von Luxburg, 2007] have been shown to outperform other clustering methods, such as $k$-means or single-linkage algorithms. However, two significant obstacles to scaling up spectral clustering to large datasets include building an affinity matrix between pairs of data points and computing the eigenvectors of the resulting normalized graph Laplacian, both of which become computationally prohibitive for large data-sets.

Approaches to scaling up spectral clustering typically focus on improving the efficiency of computing the eigenvectors using various matrix approximation schemes. A popular technique for matrix approximation is the Nyström method [Williams and Seeger, 2001] which constructs a low rank approximation by sampling a subset of the columns of the affinity matrix. Various sampling schemes have been proposed in the literature to sample this subset of columns [Kumar et al., 2012; Gittens and Mahoney, 2013], offering a tradeoff between the complexity of the sampling distribution and quality of the approximation. However other than uniform sampling, the practicality of these approaches for massive datasets may be limited [Yan et al., 2009]. The major disadvantage of these methods is that if $l$ columns are sampled, they involve the computation of eigenvectors of a sub-matrix of size $l \times l$, which can itself be computationally expensive when $l$ is large.

Also, the subsequent $k$-means step used for rounding the cluster assignment has to be performed for all the input points.

The connection between spectral clustering and weighted kernel $k$-means was first introduced in [Dhillon et al., 2004]. In this paper, we use this connection to propose a simple framework for spectral clustering that samples a subset $S$ of the input points, computes the clusters for $S$ using weighted kernel $k$-means and uses the resulting centers to compute a clustering for rest of the data points. The proposed framework does not require computation of the entire affinity matrix and does not require the computation of eigenvectors of the selected sub-matrix. Sampling algorithms developed for the Nyström approximation can be leveraged to obtain a tradeoff between accuracy and computational efficiency. In fact, our experiments show that the use of sampling followed by weighted kernel $k$-means outperforms sampling followed by the Nyström approximation. Since the proposed method uses weighted kernel $k$-means to compute the clusters, it is easy to parallelize.

The contributions of this paper are:

- Prior work on the Nyström approximation such as [Gittens and Mahoney, 2013; Talwalkar and Rostamizadeh, 2010] provide approximation guarantees with respect to various matrix norms, but not with respect to the spectral clustering objective itself. We show that matrix norm errors are not always indicative of the quality of spectral clustering.
- We propose a framework for spectral clustering that applies sampling followed by the weighted kernel $k$-means algorithm instead of approximating the affinity matrix or the Laplacian.
- Without making any assumptions about the data, we show that when points are sampled uniformly at random without replacement, the following theorem holds.

**Theorem 1.** Let $0 < \delta < 1$, $\alpha \geq 1$, $0 < \beta < 1$ and $\epsilon > 0$ be approximation parameters. Let $A$ be an $\alpha$-approximation algorithm for the weighted kernel $k$-means problem. Given a set of $n$ points, $V$, suppose we sample a subset $S \subset V$ of size $s$ uniformly at random without replacement such that,

$$s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) / \left( \frac{2\beta^2\epsilon^2}{\Delta^2\alpha^2} + \frac{1}{n} \ln \left( \frac{1}{\delta} \right) \right)$$

where $\Delta = \max_{i,j} w_i w_j \| \phi(a_i) - \phi(a_j) \|^2$. If we run algorithm $A$ with input $S$, then for the solution $C^*$ obtained, with
probability at least $1 - \delta$.

$$NCut(G, C^*) \leq 4(\alpha + \beta)NCut(G, C_{opt}) + \epsilon$$

In other words, by choosing sufficient number of samples $s$, we can obtain centers that provide a good approximation to the optimal cluster centers. As the input size $n$ increases, the term $(1/n)$ vanishes, resulting in a bound for $s$ that is independent of $n$. To the best of our knowledge, the relation between the number of samples and the spectral clustering objective studied in the proposed work has not been explored in the literature.

2 Preliminaries

2.1 Weighted Kernel $k$-means

Given a set of $n$ points $V = \{v_1, v_2, ..., v_n\}$ with associated weights $\{w_1, w_2, ..., w_n\}$ and a kernel matrix $K$, the weighted kernel $k$-means objective for clusters $V_1, V_2, ..., V_k$ with centers $C = \{C_1, C_2, ..., C_k\}$ is defined as:

$$\mathcal{W}(V, C) = \sum_{i=1}^{k} \sum_{v_j \in V_i} w_j \| \phi(v_j) - C_i \|^2$$

where $C_i = \frac{\sum_{v_j \in V_i} w_j \phi(v_j)}{\sum_{v_j \in V_i} w_j}$. Here $\phi(v)$ is the kernel function that maps the point $v$ to a higher dimensional feature space. It was shown in [Dhillon et al., 2004],

$$\| \phi(v_i) - C_j \|^2 = K_{ii} - \frac{2\tau_1(i,j)}{\text{deg}(V_j)} + \frac{\tau_2(V_j)}{(\text{deg}(V_j))^2}$$

where $\text{deg}(V_i) = \sum_{v_j \in V_i} w_j$, $\tau_1(i,j) = \sum_{v_i \in V_j} w_i K_{ii}$ and $\tau_2(V_j) = \sum_{v_i, v_m \in V_j} w_i w_m K_{im}$.

**Definition 1.** Let $A$ be an approximation algorithm for the weighted kernel $k$-means objective. Let $\alpha \geq 1$. $A$ is an $\alpha$-approximation algorithm if the set of centers, $C$, returned by $A$ satisfies,

$$\mathcal{W}(V, C) \leq \alpha \mathcal{W}(V, C_{opt})$$

where $C_{opt}$ is the set of optimal centers.

2.2 Spectral Clustering Using Norm-Cuts

In spectral clustering, we are given a graph $G = (V, A)$, which is made up of a set of $n$ vertices $V$. The affinity matrix $A$ is $n \times n$ whose entries represent the similarity between vertices. If $V_1, V_2$ are subsets of $V$, let $\text{links}(V_1, V_2) = \sum_{i \in V_1, j \in V_2} A_{ij}$.

Furthermore, let $\text{degree}(V_i) = \text{links}(V_i, V)$. The graph partitioning problem seeks to partition the graph into $k$ disjoint clusters $V_1, ..., V_k$. A number of different graph partitioning objectives have been proposed and studied. In this paper, we will focus on the normalized cut objective. The goal is to minimize the following objective over all possible clusterings $\{V_1, ..., V_k\}$,

$$NCut(G, \{V_1, ..., V_k\}) = \sum_{i=1}^{k} \frac{\text{links}(V_i, V \setminus V_i)}{\text{degree}(V_i)}$$

2.3 Relation Between Weighted Kernel $k$-means And Spectral Clustering

To convert a spectral clustering problem to a weighted kernel $k$-means problem, it was shown in [Dhillon et al., 2004] that we can set $W = D$ and $K = \sigma D^{-1} + D^{-1}AD^{-1}$, where $\sigma = k/(n-k)$. Here the term $\sigma D^{-1}$ is added to ensure $K$ is positive definite and does not change the optimal clustering.

We note that a set of centers $C$ for the weighted kernel $k$-means problem induces a clustering $V_1, ..., V_k$ for the spectral clustering problem. Since both objectives are equivalent, we use $\mathcal{W}(V, C)$, $NCut(G, \{V_1, ..., V_k\})$ and $NCut(C, G)$ interchangeably throughout the remainder of this paper. We also use $NCut(G, k) = \min C NCut(G, C)$ to represent the optimal partitioning of $G$ into $k$ clusters.

3 Why Matrix Norms Are Insufficient For Norm Cuts-based Spectral Clustering

The Nyström approximation has been extensively studied with respect to various matrix error norms, such as the Frobenius norm, trace norm and the spectral norm [Gittens and Mahoney, 2013; Talwalkar and Rostamizadeh, 2010]. In this section, we show that approximation of these norms is neither necessary nor sufficient for producing a good clustering with respect to the norm cuts objective. For simplicity, we focus on the trace norm error. However the examples can be extended to any matrix norm. Also, for this section, we let $NCut(A, k)$ represent the optimal partitioning of the graph with affinity matrix $A$ into $k$ clusters.

**Lemma 1.** Preserving trace norm is not sufficient for an approximation to be good with respect to the norm cuts objective.

We construct two matrices that are extremely similar in terms of their trace norms, but have a significantly different norm cuts objective. Here, the trace norm of a matrix $A$ is denoted as $\|A\|_s = \sum_{i=1}^{n} |\lambda_i(A)|$, where $\lambda_i(A)$ is the $i$th eigenvalue of $A$.

In other words, we show that for any value of $\epsilon > 0$, there exists $n, k, A$ and $B$ such that $A$ and $B$ represent graphs $G_A, G_B$ with $n$ vertices, and $\|A - B\|_s \leq (1 + \epsilon)\|A\|_s$ and $NCut(A, k) = O(k) \cdot NCut(B, k)$. Note, that the maximum value of $NCut(A, k)$ is $k$.

Let $A$ and $B$ be the affinity matrices for undirected, unweighted graphs that have $n$ vertices and $k$ and $k-1$ equal sized components, respectively. Further, we assume that each component is completely connected with no self loops. Since each component in $A$ is a complete graph on $n/k$ vertices, its eigenvalues are $-1$ with multiplicity $(n/k) - 1$ and $(n/k) - 1$ with multiplicity 1. Since each component in $A$ is regular, the eigen values of the normalized symmetric Laplacian of $A$ (denoted by $L(A)$) are given as, $\lambda_i(L(A)) = 1 - (1/\Delta)\lambda_i(A)$ where $\Delta = (n/k) - 1$. In other words, $L(A)$ has eigenvalues $n/(n-k)$ with multiplicity $n-k$ and 0 with multiplicity $k$. Similarly, $L(B)$ has eigenvalues $n/(n-k+1)$ with multiplicity $n-k+1$ and 0 with multiplicity $k-1$. A more general statement in this context can be found in [Butler, 2007].

Now we compare the trace norm of the difference of the
Algorithm 1 Proposed Framework for Spectral Clustering

Input: Affinity Matrix $A$, number of clusters $k$, number of samples $s$
Output: Matrix $\hat{Y}$ ($\hat{Y}_{ij} \neq 0$ only if point $i$ belongs to cluster $j$)
Procedure:
1. $sub \leftarrow$ sample($s$)
2. $A_{sub} \leftarrow A(sub, sub)$
3. $Y \leftarrow$ weighted_kernel_kmeans($A_{sub}, k$)
4. $\hat{Y} \leftarrow$ diffuse($Y, A, sub$)

Algorithm 2 Diffuse

Input: Matrix $\hat{Y}$, Affinity Matrix $A$, set of sampled indices $S$
Output: Matrix $Y$ ($Y_{ij} \neq 0$ only if point $i$ belongs to cluster $j$)
Procedure:
for each $i \in S$ do
  for $j = 1: k$ do
    $Y_{ij} \leftarrow Y_{ij}$
  end for
end for
for each $i \notin S$ do
  for $c = 1 : k$ do
    $d(a_i, m_c) \leftarrow K_{ci} - \frac{2}{n} \sum_{a_j \in \pi_c} w_j K_{ij} + \frac{\sum_{a_j \in \pi_c} w_j w_j K_{ij}}{\sum_{a_j \in \pi_c} w_j}$
  end for
  $j \leftarrow \arg\min_{c=1}^{k} d(a_i, m_c)$
  $Y_{ij} \leftarrow 1$
end for

4 Proposed Framework

The proposed framework for spectral clustering is described in Algorithm 1. The overall procedure consists of three stages. In the first stage, a subset of the input points is sampled according to some distribution. This subset is indicated as $sub$. In the second stage, the clusters are computed using weighted kernel $k$-means. However, instead of using the entire affinity matrix, as in [Dhillon et al., 2004], we only use the principal sub-matrix induced by indices in $sub$. The final step involves using the cluster centers returned by the weighted kernel $k$-means procedure to assign clusters to the remaining data points. The procedure for this step is described in Algorithm 2.

This is scalable to larger datasets because it does not require explicit computation of the eigenvectors. Also it only requires the computation of a sub-matrix of size $n \times k$, which can result in significant savings in memory. Weighted kernel $k$-means is extremely easy to parallelize.

5 Analysis

In this section, we present the proof of Theorem 1. We extend the analysis in [Czumaj and Sohler, 2010] to the case of sampling without replacement. Specifically, uniform sampling without replacement allows us to use the sharper bound offered by the Serfling inequality stated below. We refer the reader to [Bardenet et al., 2015] for a more thorough discussion.

Lemma 3. [Serfling, 1974; Bardenet et al., 2015] Let $X_1 \ldots X_n$ be random variables sampled uniformly without replacement.
Let \( X = X_1 + \ldots + X_n \) and \( 0 \leq X_i \leq \Delta \). Then for all \( \gamma > 0 \),
\[
P \left[ \sum_{i=1}^{n} \frac{X_i - E[X_i]}{s} \geq \gamma \right] \leq \exp \left( -\frac{2s\gamma^2}{(1 - (s - 1)/n)\Delta^2} \right)
\]

**Definition 2.** Let \( \beta > 0 \) and \( \alpha \geq 1 \). A set of \( k \) centers \( C \) is a \( \beta \)-bad, \( \alpha \)-approximation if
\[
\omega(V, C) > (\alpha + \beta)\omega(V, C_{opt})
\]
If not, \( C \) is said to be a \( \beta \)-good, \( \alpha \)-approximation.

We will need the following lemmas.

**Lemma 4.** Let \( S \) be a subset of \( V \) of size \( s \) such that,
\[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) \left( 2 + \frac{2\beta^2m^2}{\Delta^2a^2} + \frac{\ln(1/\delta)}{n} \right)
\]
If an \( \alpha \)-approximation algorithm for weighted kernel k-means \( \mathcal{A} \) is run on input \( S \), then for the set of centers, \( C_{\mathcal{A}} \), obtained, with probability \( \geq 1 - \delta \),
\[
\omega(S, C_{\mathcal{A}}) \leq (\alpha + \beta)\frac{s}{n}\omega(V, C_{opt})
\]
where \( C_{opt} \) is the set of optimal centers for \( V \) and \( m = \omega(V, C_{opt})/n \).

**Proof.** For each point \( v_i \in S \), let \( X_i = w_i||\phi(v_i) - C_i||^2 \) where \( C_i \) is the nearest center to \( \phi(v_i) \) in \( C_{opt} \). Then
\[
\omega(S, C_{opt}) = \sum_{i=1}^{s} X_i.
\]
We also know that \( E[X_i] = m \). Let \( X_i \) be bounded in the interval \([0, \Delta]\).

\[
Pr[\omega(S, C_{opt}) > (1 + \frac{\beta}{\alpha})s]\omega(V, C_{opt})]
\]
\[
= Pr[\sum_{i=1}^{s} X_i > (1 + \frac{\beta}{\alpha})sm]
\]
\[
\leq e^{-\frac{2s\beta^2m^2}{(1 - (s - 1)/n)\Delta^2}}
\]
The last step is obtained by applying the Serfling inequality. Note, \( e^{-\frac{2s\beta^2m^2}{(1 - (s - 1)/n)\Delta^2}} \leq \delta \) if
\[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) \left( 2 + \frac{2\beta^2m^2}{\Delta^2a^2} + \frac{\ln(1/\delta)}{n} \right)
\]
Let \( C' \) be the set of \( k \) centers in \( S \) obtained by replacing each \( C \in C_{opt} \) by its nearest neighbor in \( S \). From the weakened triangle inequality, it can be shown that \( \omega(S, C') \leq 4\omega(S, C_{opt}) \). This implies with probability at least \( 1 - \delta \), \( S \) contains \( k \) centers with cost at most \( 4(1 + \frac{\beta}{\alpha})\frac{s}{n}\omega(V, C_{opt}) \). In other words,
\[
\omega(S, C') \leq 4(1 + \frac{\beta}{\alpha})\frac{s}{n}\omega(V, C_{opt})
\]
If \( C'' \) is the set of optimum centers for \( \omega(S, C') \), then
\[
\omega(S, C'' \leq \omega(S, C')
\]
Since \( \mathcal{A} \) is an \( \alpha \)-approximation algorithm,
\[
\omega(S, C_{\mathcal{A}}) \leq \alpha\omega(S, C'')
\]
This implies,
\[
\omega(S, C_{\mathcal{A}}) \leq \alpha\omega(S, C')
\]
The lemma now follows.

---

**Lemma 5.** Let \( S \) be a subset of \( V \) of size \( s \), sampled uniformly at random such that,
\[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) \left( \frac{\beta^2m^2}{\Delta^2} + \frac{1}{n}\ln \left( \frac{1}{\delta} \right) \right)
\]
Let \( \chi \) be the set of 12\( \beta \)-bad 4\( \alpha \)-approximations of \( \omega(V, C_{opt}) \) then,
\[
Pr[C_b \subset S \text{ and } C_b \in \chi \text{ and } \omega(S, C_b) \leq 4(\alpha + \beta)m] \leq \delta
\]
**Proof.** Let \( s \geq \frac{4\alpha + 5\beta}{\beta}k \). Let \( C_b \in \chi \) be a 12\( \beta \)-bad 4\( \alpha \)-approximation of \( \omega(V, C_{opt}) \). If \( C_b \subset S \), let \( S' \) be the subset of \( S \) with \( C_b \) removed, such that \( |S'| = s - k \). Thus,
\[
Pr[C_b \subset S \text{ and } \omega(S, C_b) \leq 4(\alpha + \beta)m]
\]
\[
= Pr \left[ \frac{\omega(S, C_b)}{s} \leq \frac{4(\alpha + \beta)m}{s} \right] \cdot Pr[C_b \subset S]
\]
\[
\leq Pr \left[ \frac{\omega(S', C_b)}{s - k} \leq \frac{4}{s - k}(\alpha + \beta)m \right] \cdot Pr[C_b \subset S]
\]
\[
\leq Pr \left[ \frac{\omega(S', C_b)}{s - k} \leq (4\alpha + 5\beta)m \right] \cdot Pr[C_b \subset S]
\]
(1)
The second step holds because the mean remains unchanged at each step when the elements are chosen uniformly at random without replacement and \( \omega(S, C_b) = \omega(S', C_b) \) when \( C_b \subset S \). The third step holds since \( s \geq \frac{4\alpha + 5\beta}{\beta}k \).

Let \( Y_i \) denote the distance of point \( v_i \in S \) from its closest center in \( C_b \). Since \( C_b \in \chi \), \( E[Y_i] \geq (4\alpha + 12\beta)m \). Thus,
\[
Pr \left[ \frac{\omega(S', C_b)}{s - k} \leq (4\alpha + 5\beta)m \right]
\]
\[
= Pr \left[ \sum_{i=1}^{s-k} Y_i \leq (4\alpha + 5\beta)m \right]
\]
\[
\leq Pr \left[ \sum_{i=1}^{s-k} Y_i \leq \frac{4\alpha + 5\beta}{4\alpha + 12\beta} E[\sum_{i=1}^{s-k} Y_i] \right]
\]
\[
\leq Pr \left[ \sum_{i=1}^{s-k} Y_i - E[\sum_{i=1}^{s-k} Y_i] \right] \leq \left( \frac{7\beta}{4\alpha + 12\beta} \right) E[\sum_{i=1}^{s-k} Y_i]
\]
\[
\leq \exp \left( -2(s - k) \left( \frac{7\beta E[\sum_{i=1}^{s-k} Y_i]}{4\alpha + 12\beta} \right) \right)
\]
(2)
Similarly to the previous lemma, we have used the Serfling bound to obtain (2). Thus it follows that,
\[
Pr \left[ \frac{\omega(S', C_b)}{s - k} \leq (4\alpha + 5\beta)m \right] \leq e^{-\frac{4\beta^2m^2}{(1 - (s - 1)/n)\Delta^2}}
\]
(3)
We note, \( Pr[C_b \subset S] \leq (s/n)^k \) and \( |\chi| \leq n^k \). Using this and (3) in (1), we have,
\[
Pr[C_b \subset S \text{ and } C_b \in \chi \text{ and } \frac{\mathcal{W}(S, C_b)}{s} \leq 4(\alpha + \beta)m]
\leq \sum_{C_b \in \chi} Pr[C_b \subset S \text{ and } \frac{\mathcal{W}(S, C_b)}{s} \leq 4(\alpha + \beta)m] \leq \sum_{C_b \in \chi} Pr \left[ \frac{\mathcal{W}(S^*, C_b)}{s-k} \leq (4\alpha + 5\beta)m \right] \cdot (s/n)^k 
\leq s^n \cdot e^{-\frac{\sigma^2 m^2}{(1+(\gamma-1)/n)^2 \Delta^2}} \cdot (s/n)^k 
\leq s^k \cdot e^{-\frac{\sigma^2 m^2}{(1+(\gamma-1)/n)^2 \Delta^2}} 
\] 
which is upper bounded by \( \delta \) when,
\[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) / \left( \frac{\beta^2 m^2}{\Delta^2} + \frac{1}{n} \ln \left( \frac{1}{\delta} \right) \right)
\]

Next, we use these lemmas to prove Theorem 1.

Proof. Let \( \beta^* \) be some constant. Let \( S \subseteq V \) be a subset of \( s \) points chosen uniformly at random such that,
\[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) / \left( \frac{(\beta^*)^2 m^2}{\Delta^2 \alpha^2} + \frac{1}{n} \ln \left( \frac{1}{\delta} \right) \right)
\]

Then from Lemma 3, we have, if \( C \subseteq S \) and \( C \in \chi \), with probability at least \( 1 - \delta \),
\[
\mathcal{W}(S, C) > 4(\alpha + \beta^*)m
\]

On the other hand, if we run algorithm \( A \) for set \( S \), then by Lemma 2, the resulting set of centers satisfies, with probability at least \( 1 - \delta \),
\[
\mathcal{W}(S, C) \leq 4(\alpha + \beta^*)m
\]

This implies with probability at least \( 1 - 2\delta \) the set \( C \) is a \( (12\beta) \)-good \( (4\alpha) \)-approximation. In order to remove the dependence on the number of samples on \( m \),

- If \( m \leq \epsilon \), we simply set \( \beta = (1/\sqrt{2})\epsilon/m \). In this case we will obtain the following bound on \( s \).
  \[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) / \left( \frac{\epsilon^2}{\Delta^2 \alpha^2} + \frac{1}{n} \ln \left( \frac{1}{\delta} \right) \right)
\]

- If \( m > \epsilon \), we set \( \beta = 3\beta^* \) to obtain,
  \[
s \geq \ln \left( \frac{1}{\delta} \right) \left( 1 + \frac{1}{n} \right) / \left( \frac{2\beta^* \epsilon^2}{\Delta^2 \alpha^2} + \frac{1}{n} \ln \left( \frac{1}{\delta} \right) \right)
\]

The theorem now follows. \( \square \)

### Table 1: The synthetic datasets used in our experiments.

<table>
<thead>
<tr>
<th>Name</th>
<th>Instances</th>
<th>Attributes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation (Agg)</td>
<td>788</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Flame (Fla)</td>
<td>240</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A.K’s toy problem (AK)</td>
<td>373</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

![Figure 2](image-url)
approaches based on combining sampling and the weighted kernel \( k \)-means algorithm outperform approaches based on the Nyström approximation both in terms of accuracy and computation time.

6.2 Image Segmentation

One of the most popular applications of spectral clustering is image segmentation. In this section, we describe results obtained on an image segmentation benchmark [Yu and Shi, 2003]. The affinity matrix and the final discretization were computed using the approach of [Yu and Shi, 2003].

Experimental results in Figure 2 show that wkk-unif and wkk-col obtain the same accuracy as the Nyström based approaches, while taking significantly less time. In cases where only 10% of the columns were sampled, wkk-unif and wkk-col achieved a lower norm cuts ratio compared to nys-unif and nys-col. The baseline methods shi and kulis use the entire affinity matrix. Thus they can achieve lower norm cuts ratio, but have a significantly higher computation time.

In order to study how the proposed approach scales to larger graphs, we vary the size of the input images. The results are shown in Figure 3. For shi, the time taken scales cubically with the size of the graph. Hence it is unsuitable for large graphs. In contrast, kulis, wkk-unif and wkk-col scale almost linearly with the size of the graph. Both wkk-unif and wkk-col outperform kulis in terms of execution time. They also outperform the Nyström-based approaches both in terms of accuracy and computation time. These results show that uniform sampling followed by the weighted kernel \( k \)-means algorithm scales well in practice.

7 Conclusion

This paper presented a framework for spectral clustering based on combining sampling and the weighted kernel \( k \)-means algorithm. If the points are sampled uniformly at random without replacement, we show that for large datasets, the number of samples required, is independent of the input size and depends only on the number of clusters and the diameter of the set of points in the kernel space. Experiments show that approaches based on the proposed framework outperform approaches based on the Nyström approximation.
References


