Flat Clustering Algorithms for Big Data

Yuanhang Ren
201721220117
ryuanhang@gmail.com

School of Information and Software Engineering
University of Electronic Science and Technology of China

May 2, 2020

Table of Contents

Introduction

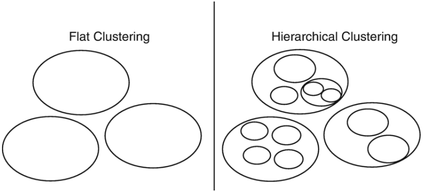
k-means
- Backgrounds
- Provably Good Algorithms
- Efficient Algorithms
- Experiments

Spectral Clustering
- Backgrounds
- Provably Good and Efficient Algorithms
- Experiments

Summary of Contributions

Clustering

- What is clustering?
- Flat clustering vs Hierarchical clustering



Research Problems

Traditional clustering algorithms might be inefficient when datasets are huge and they are also not theoretically guaranteed.
Hence, we focus on clustering algorithms that are:

- provably good
- efficient

We are going to design and analyze such algorithms on the *k-means* and *spectral clustering* problems.

Table of Contents

Introduction

k-means
Backgrounds
Provably Good Algorithms
Efficient Algorithms
Experiments

Spectral Clustering
Backgrounds
Provably Good and Efficient Algorithms
Experiments

Summary of Contributions

Table of Contents

k-means

Backgrounds

Provably Good Algorithms

Efficient Algorithms

Experiments

k-means problem

- It is well-known due to the Lloyd algorithm [Lloyd, 1982] (a.k.a k-means algorithm)
- The k-means problem is NP-hard

Definition (k-means problem)

Given n data points $\mathcal{X} \subseteq \mathbb{R}^d$ and a set of k points $C \subseteq \mathbb{R}^d$, where d is the dimension of the data point. An objective function is defined as follows,

$$\phi_C(\mathcal{X}) = \sum_{x \in \mathcal{X}} d^2(x, C) \tag{1}$$

where $d(x, C) = \min_{c \in C} \|x - c\|$ is the distance of a point to a set. The k-means problem is to find the optimal C such that the $\phi_C(\mathcal{X})$ is minimized given $\mathcal{X}$.

The Solution Quality

Definition (Solution Quality 1)

Let $\alpha \geq 1$. A set C of k centers is an α approximation solution of k-means if

$$\phi_C(\mathcal{X}) \leq \alpha \phi_{\mathsf{OPT}}(\mathcal{X}) \tag{2}$$

$\phi_{\mathsf{OPT}}(\mathcal{X})$ is the minimal objective.

Definition (Solution Quality 2)

Let $\alpha \geq 1$ and $\beta > 0$. A set C of k centers is a β-bad α-approximation solution of k-means if

$$\phi_C(\mathcal{X}) > (\alpha + \beta)\phi_{\mathsf{OPT}}(\mathcal{X}) \tag{3}$$

Otherwise, C is said to be a β-good α-approximation.

Table of Contents

k-means

- Backgrounds
- Provably Good Algorithms
- Efficient Algorithms
- Experiments

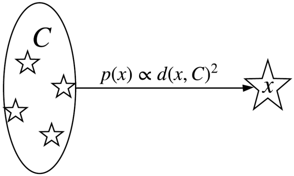
k-MEANS++

k-means++ [Arthur and Vassilvitskii, 2007] employs the d^2 *weighting* to achieve a $O(\log k)$ guarantee.

Algorithm 1: k-means++ seeding

Input: dataset $\mathcal{X}$, number of centers k
Output: k centers C

```
c_1 ← Sample a point uniformly at random from X
C ← {c_1}
for i = 2, 3, ... k do
    for x ∈ X do
        p(x) ← d(x, C)^2 / Σ_{x'∈X} d(x', C)^2
    end
    x ← Sample a point from X using p(x)
    C ← C ∪ {x}
end
return C
```



The k-means|| [Bahmani et al., 2012] accelerates the k-means++

Algorithm 2: k-means|| seeding

Input: dataset $\mathcal{X}$, oversampling factor l, number of centers k, number of rounds t

Output: k centers C

$S \leftarrow$ Sample a point uniformly at random from $\mathcal{X}$

for $i = 1,2,\ldots,t$ **do**

  $C' \leftarrow \emptyset$

  for $x \in \mathcal{X}$ **do**

    Add x to C' with probability $\min(1, \frac{ld^2(x,S)}{\phi(\mathcal{X},S)})$

  end

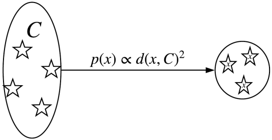
  $S \leftarrow S \cup C'$

end

For $s \in S$, set w_s to be the number of points in $\mathcal{X}$ closer to s than any other point in S

$C \leftarrow$ Let w_s be the weights of s and run an α approximation algorithm on the weighted S

return C



The First Contribution

The classic k-means++ algorithm has been extended to weighted k-means problem and proofs on the clustering quality are given.

The Weighted k-means Problem & Algorithm

Definition (weighted k-means problem)

Given n data points $\mathcal{X} \in \mathbb{R}^d$ and associated weights w. Find a set of k points $C \subseteq \mathbb{R}^d$, such that the following objective function is minimized.

$$\psi_C(\mathcal{X}) = \sum_{x_i \in \mathcal{X}} w_i d^2(x_i, C) \tag{4}$$

Algorithm 3: weighted k-means++ seeding

Input: dataset $\mathcal{X}$, data weights w, number of centers k
Output: k centers C

```
c_1 ← Sample a point from 𝒳 with probability w_x / ∑_{i∈𝒳} w_i
C ← {c_1}
for i = 2, 3, ... k do
    for x ∈ 𝒳 do
        p(x) ← w_x d(x, C)^2 / ∑_{x'∈𝒳} w'_x d(x', C)^2
    end
    x ← Sample a point from 𝒳 with p(x)
    C ← C ∪ {x}
end
return C
```

Guarantee of Weighted k-means++

Theorem (quality of weighted k-means++)
Given data points $\mathcal{X}$ and associated weights w. Let C be the results returned by the weighted k-means++. We have

$$\mathbb{E}[\psi_C(\mathcal{X})] \leq 8(\ln k + 2)\psi_{OPT}(\mathcal{X}) \quad (5)$$

The big picture of the proof:

- Consider the quality of the uniform sampling
- Consider the quality of the d^2 weighting
- Show relationships between intermediate results by induction

Picture of The Proof

Lemma (quality of uniform sampling)

Given an arbitrary optimal cluster A. Denote Z be the chosen point with probability $\frac{w_z}{\sum_{i \in A} w_i}$. Then, $\mathbb{E}[\psi_Z(A)] = 2\psi_{OPT}(A)$

Lemma (quality of d^2 weighting)

Let C be the arbitrary intermediate results, $1 \leq |C| \leq k-1$. Denote A be an arbitrary optimal cluster and let $Z \in A$ be the chosen point using d^2 weighting given C. Then, we have $\mathbb{E}[\psi_{C'}(A) | C, Z \in A] \leq 8\psi_{OPT}(A)$, where $C' = C \cup \{Z\}$.

Picture of The Proof

Lemma (relationships of intermediate results)

Let C be the arbitrary intermediate results, $1 \le |C| \le k-1$. Choose $u > 0$ "uncovered" optimal clusters, and let $\mathcal{X}_u$ denote the set of points in these clusters. Also let $\mathcal{X}_c = \mathcal{X} - \mathcal{X}_u$. Suppose we add $0 \le t \le u$ points with d^2 weighting given C. Denote C' as the resulting points. Then,

$$\mathbb{E}[\psi_{C'}(\mathcal{X})|C] \le [\psi_C(\mathcal{X}_c) + 8\psi_{OPT}(\mathcal{X}_u)](1 + H_t) + \frac{u-t}{u}\psi_C(\mathcal{X}_u) \qquad (6)$$

where $H_t = 1 + \frac{1}{2} + ... + \frac{1}{t}$ is the harmonic sum.

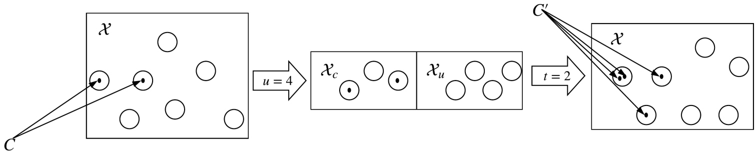


Table of Contents

k-means

Backgrounds

Provably Good Algorithms

Efficient Algorithms

Experiments

Clustering Based on Uniform Sampling

Algorithm 4: clustering based on uniform sampling

Input: dataset $\mathcal{X}$, number of clusters k, number of points to sample s, clustering algorithm $\mathcal{A}_c$

Output: k centers C

$S \leftarrow$ Sample s points uniformly without replacement

$C \leftarrow$ Solve the k-means problem on S with $\mathcal{A}_c$

return k centers C

Theorem (quality of Algorithm 4)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ be approximation parameters. Let C be the set of centers returned by Algorithm 4 and $\mathcal{A}_c$ is an α approximation algorithm. Suppose we sample s points uniformly without replacement such that,

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

we have

$$\phi_C(\mathcal{X}) \leq 4(\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$, where $\Delta = \max_{i,j} \|v_i - v_j\|^2$ is the squared diameter of the data, $m = \phi_{OPT}(\mathcal{X})/n$ is the average of the optimal objective.

The Second Contribution

- A sharper bound for the uniform sampling algorithm is proved, and a further proof indicate that this algorithm runs in polylogarithmic time given mild assumptions on datasets.
- A novel algorithm called Double-K-MC^2 is proposed to approximate weights.
- MATLAB implementations of uniform sampling, K-MC^2, Double-K-MC^2, and their corresponding kernel versions are given. Experiments are carried out to verify the efficiency and effectiveness of these algorithms.

A Sharper Bound

Theorem (a sharper bound of uniform sampling)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ be approximation parameters. Let C be the set of centers returned by Algorithm 4 and $\mathcal{A}_c$ is an α approximation algorithm. Suppose we sample s points uniformly without replacement such that,

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

we have

$$\phi_C(\mathcal{X}) \leq (\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$, where $\Delta = \max_{i,j} \|v_i - v_j\|^2$ is the squared diameter of the data, $m = \phi_{OPT}(\mathcal{X})/n$ is the average of the optimal objective.

The big picture of the proof:

1. Show that C will be a *good* solution for S.
2. Suppose C is a *bad* solution for $\mathcal{X}$, it will probably be a *bad* solution for S.
3. According to 1 and 2, C will be a *good* solution for $\mathcal{X}$

A Poly-Log Time Algorithm

Assume that a dataset is sampled i.i.d. according to a probability distribution F

- F has finite variance and exponential tails, *i.e.* $\exists c, t$ such that $P[\mathrm{d}(x, \mu(F)) > a] \leq ce^{-at}$, where $\mu(F)$ is the mean of F.
- F's minimal and maximal density on a hypersphere with non zero probability mass is bounded by a constant.

Theorem (efficiency of uniform sampling)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ *be approximation parameters. Assume (A1) and (A2) hold, and let* C *be the set of centers returned by Algorithm 4, we have the following*

$$\phi_C(\mathcal{X}) \leq (\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$ *if we sample* $O(\ln(\frac{1}{\delta})\frac{\alpha^2}{\beta^2}k^2 \log^4 n)$ *points*

Table of Contents

k-means

- Backgrounds
- Provably Good Algorithms
- Efficient Algorithms
- Experiments

Baseline Algorithms

- Since the uniform sampling algorithm is efficient and provably good, we design experiments to verify this.
- Baselines are K-MC^2 [Bachem et al., 2016] and Double-K-MC^2 sampling.
- As computing weights in k-means$||$ is time-consuming, we propose a novel algorithm called *Double-K-MC^2 sampling* to approximate weights.

Algorithm 5: Double-K-MC^2 sampling

Input: dataset $\mathcal{X}$, $\#$ of points to sample s, chain length u
Output: k centers C
$S_1 \leftarrow$ Sample s points from V via K-MC^2
$V' \leftarrow$ Remove S_1 from V
$S_2 \leftarrow$ Sample s points from V' via K-MC^2
For point $s_i \in S_1$, let w_i be the number of points in S_2 closer to s_i than to any other points in S_1
Let $w_i + 1$ be the weight of s_i
$C \leftarrow$ Solve the weighted k-means problem on S_1 with an α approximation algorithm
return k centers C

Traditional Clustering

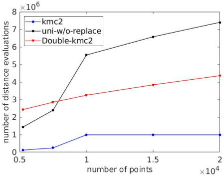
Table 1: data size n, number of clusters k, dimension d

datasets	n	k	d
a2	5250	35	2
a3	7500	50	2
b2-random-10	10000	100	2
b2-random-15	15000	100	2
b2-random-20	20000	100	2
KDD	145751	200	74
RNA	488565	200	8
Poker Hand	1000000	200	10

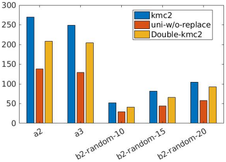
- chain length: $u = 200$
- sampling size: $1.5 \log^2 n$ and $0.7 \log^4 n$ for Double-K-MC^2 and uniform sampling
- α approximation algorithm: (weighted) k-means++ with Lloyd
- evaluation metrics: number of distance evaluations and k-means objective
- algorithms are run 40 times repeatedly with different initial random seeds

Results

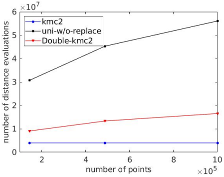
1. The time cost of uniform sampling is about 10 times higher than that of K-MC^2 and it increases slowly with respect to the data size. The k-means objective of uniform sampling is roughly 60% of the objective of K-MC^2.
2. Double-K-MC^2 achieves a better clustering quality compared with K-MC^2 and a lower time cost compared with uniform sampling.
3. Double-K-MC^2 could be the first choice if you prefer a good clustering quality with reasonable time costs. For the best quality, uniform sampling is recommended.



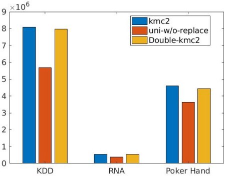
(A) the number of distance evaluations on synthetic data



(B) k-means objective on synthetic data



(C) the number of distance evaluations on real data



(D) k-means objective on real data

Figure 1: k-means objective and time cost versus the number of points

Image Segmentation

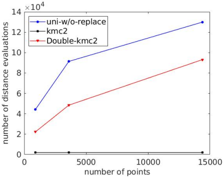
Table 2: data size n, number of clusters k

datasets	n	k
baby	900(30 * 30)	5
kitten	3600(60 * 60)	5
bear	14400(120 * 120)	5

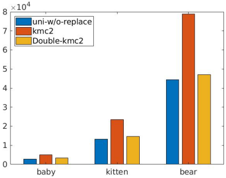
- The kernel versions of uniform sampling, Double-K-MC^2, and K-MC^2.
- Construct an affinity matrix A via the approach in Stella and Shi [2003] and find the nearest positive definite matrix K as the kernel.
- chain length: $u = 200$
- sampling size: $0.25 \log^2 n$ and $0.4 \log^4 n$ for Double-K-MC^2 and uniform sampling
- α approximation algorithm: (weighted) kernel k-means++ with kernel Lloyd
- evaluation metric: number of distance evaluations and kernel k-means objective
- algorithms are run 30 times repeatedly with different initial random seeds

Results

1. The kernel uniform sampling has the best clustering quality while the growth of the time cost is not too rapid.
2. The kernel Double-K-MC^2 has a similar clustering quality with much lower time cost compared with the kernel uniform sampling.
3. Thus, we recommend using kernel Double-K-MC^2 if the quality is your major concern. For a more efficient result, the kernel K-MC^2 is a better choice.



(A) the number of distance evaluations on image data



(B) kernel k-means objective on image data

Figure 2: kernel k-means objective and time cost versus the number of points

Table of Contents

Introduction

k-means

- Backgrounds
- Provably Good Algorithms
- Efficient Algorithms
- Experiments

Spectral Clustering

- Backgrounds
- Provably Good and Efficient Algorithms
- Experiments

Summary of Contributions

Table of Contents

Spectral Clustering
Backgrounds
Provably Good and Efficient Algorithms
Experiments

Spectral Clustering Problem

- The problem is introduced from the graph theory.
- This problem is also NP-hard.

Definition (Normalized Cut)

Denote v_i as the node i and C_j as the cluster j. Find a partition matrix $F_{ij} = \begin{cases} 1 & v_i \in C_j \\ 0 & otherwise \end{cases}$, such that the following objective is minimized,

$$\min_F \mathrm{Tr}(\frac{F^T LF}{F^T DF}) \tag{7}$$

where $L = D - A$, A is the affinity matrix of the graph, and D is the degree matrix of A.

Table of Contents

Spectral Clustering

Backgrounds

Provably Good and Efficient Algorithms

Experiments

Transform into Traditional Clustering

Definition (weighted kernel k-means problem)

Given n data points $\mathcal{X} \subseteq \mathbb{R}^d$, associated weights w, and a mapping function $\varphi(.)$. Find a set C of size k such that the following objective is minimized,

$$\Psi_C(\mathcal{X}) = \sum_{i=1}^{k} \sum_{x_j \in \pi_i} w_j \left\| \varphi(x_j) - C_i \right\|^2 \tag{8}$$

where π_i is the ith cluster, $\cup_{i=1}^{k} \pi_i = \mathcal{X}$, $C_i = \frac{\sum_{x_j \in \pi_i} w_j \varphi(x_j)}{\sum_{j \in \pi_i} w_j}$.

Lemma (relationships between two clusterings)

Let W and K be the weight and kernel matrix of the weighted kernel k-means. Choose $W = D$ and $K = D^{-1} A D^{-1}$, where D and A are the degree and affinity matrix of the spectral clustering. Then, we have

$$\Psi_C(\mathcal{X}) = \mathrm{Tr}(D^{-1/2} A D^{-1/2}) - k + Ncut \tag{9}$$

The Uniform Sampling Algorithm

Algorithm 6: uniform sampling and weighted kernel k-means based spectral clustering [Mohan and Monteleoni, 2017]

Input: dataset $\mathcal{X}$, number of clusters k, sample size s, affinity matrix A, degree matrix D

Output: k partitions of $\mathcal{X}$

$S \leftarrow$ Sample s numbers uniformly from $1, 2, ..., n$ without replacement

$W \leftarrow D, K \leftarrow D^{-1}AD^{-1}$

$W_s \leftarrow W(S, S), K_s \leftarrow K(S, S)$

$\pi' \leftarrow$ Run an α approximation algorithm on S given W_s and K_s

/* diffuse */

for $i = 1, ..., n$ **do**

  for $c = 1, ..., k$ **do**

$$d(a_i, m_c) \leftarrow K_{ii} - \frac{2\sum_{a_j \in \pi'_c} w_j K_{ij}}{\sum_{a_j \in \pi'_c} w_j} + \frac{\sum_{a_j, a_l \in \pi'_c} w_j w_l K_{jl}}{\left(\sum_{a_j \in \pi'_c} w_j\right)^2}$$

  end

  $j \leftarrow \operatorname*{argmin}_{c=1,2,...,k} d(a_i, m_c)$

  $Y_{ij} \leftarrow 1$

end

return Y

The Third Contribution

- A sharper bound for the uniform sampling and weighted kernel k-means based spectral clustering algorithm has been proved.
- We give MATLAB implementations of this algorithm and use experiments to validate the efficiency and the quality of it.

A Sharper Bound

Theorem (a sharper bound of Algorithm 6)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, *and* $\beta > 0$ *be approximation parameters. Suppose we sample* s *points in Algorithm 6 such that,*

$$s \geq \ln(\frac{1}{\delta})(1 + \frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n}) \tag{10}$$

we have

$$Ncut \leq (\alpha + \beta) Ncut^* + (\alpha + \beta - 1)c \tag{11}$$

with probability at least $1 - 2\delta$, *where* $\Delta = \max_{i,j} \|\varphi(x_i) - \varphi(x_j)\|^2$ *is the squared diameter of the data in the mapped space,* $m = \Psi_{OPT}(\mathcal{X})/n$ *is the average of the optimal objective,* c *is a constant irrelevant to partitions,* $Ncut^*$ *is the optimal* $Ncut$.

Table of Contents

Spectral Clustering
Backgrounds
Provably Good and Efficient Algorithms
Experiments

Traditional Clustering

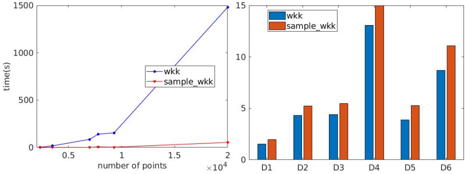
Table 3: data size n, number of clusters k, dimension d

datasets	name	n	d	k
D_1	segment	2310	19	7
D_2	MnistData-05	3495	784	10
D_3	MnistData-10	6996	784	10
D_4	isolet5	7797	617	26
D_5	USPS	9298	256	10
D_6	letter-recognition	20000	16	26

- Experiments will be performed to verify the efficiency and the effectiveness of uniform sampling
- algorithms: weighted kernel k-means(uniform sampling vs no-sampling)
- anchor based graph construction method
- sampling size: 20% of data points
- α approximation algorithm: weighted kernel k-means++
- evaluation metrics: time(s) and Ncut

Results

1. The sampling version is about 20∼50 times faster than the no-sampling one while Ncut is not larger than the no-sampling version by 25%
2. Hence, the uniform sampling method is a good choice for its amazing efficiency and reasonable clustering quality.



(A) time costs(s) (B) Ncut

Figure 3: results of spectral clustering

Table of Contents

Introduction

k-means

Backgrounds

Provably Good Algorithms

Efficient Algorithms

Experiments

Spectral Clustering

Backgrounds

Provably Good and Efficient Algorithms

Experiments

Summary of Contributions

Summary

1. (Theoretical) The classic k-means++ algorithm has been extended to weighted k-means problem and proofs on the clustering quality are given.
2. (Theoretical) A sharper bound for the uniform sampling algorithm on k-means is proved, and a further proof indicates that this algorithm runs in polylogarithmic time given mild assumptions on datasets.
3. (Theoretical) A sharper bound for the spectral clustering is proved.
4. (Empirical) We give MATLAB implementations of uniform sampling, K-MC^2, Double-K-MC^2, and their corresponding kernel versions on k-means. Experiments validate the efficiency and effectiveness of these algorithms.
5. (Empirical) We give MATLAB implementations of the weighted kernel k-means based spectral clustering algorithm and its corresponding sampling versions. Experiments are used to justify these algorithms on the efficiency and solution quality.

Questions?

References

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Approximate k-means++ in sublinear time. 2016.

B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633, 2012.

A. Czumaj and C. Sohler. Sublinear-time approximation for clustering via random sampling. In *International Colloquium on Automata, Languages, and Programming*, pages 396–407. Springer, 2004.

S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

M. Mohan and C. Monteleoni. Beyond the nystrom approximation: Speeding up spectral clustering using uniform sampling and weighted kernel k-means. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2494–2500. AAAI Press, 2017. ISBN 978-0-9992411-0-3. URL `http://dl.acm.org/citation.cfm?id=3172077.3172235`.

X. Y. Stella and J. Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003.