Randomized Sketches for Clustering: Fast and Optimal Kernel k-Means

Rong Yin, Yong Liu, Weiping Wang, and Dan Meng

{yinrong, wangweiping, mengdan}@iie.ac.cn, liuyonggsai@ruc.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS

Introduction

In this paper, we investigate the efficiency of kernel k-means combined with randomized sketches in terms of both statistical analysis and computational requirements. More precisely, we propose a unified randomized sketches framework to kernel k-means and investigate its excess risk bounds, obtaining the state-of-the-art risk bound with only a fraction of computations. Indeed, we prove that it suffices to choose the sketch dimension $\Omega(\sqrt{n})$ to obtain the same accuracy of exact kernel k-means with greatly reducing the computational costs, for sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nyström kernel k-means, where n is the number of samples. To the best of our knowledge, this is the first result of this kind for unsupervised learning. Finally, the numerical experiments on simulated data and real-world datasets validate our theoretical analysis.

Motivation

Kernel k-means is one of the fundamental approaches in unsupervised learning. The Voronoi cell associated with a centroid $\mathbf{c}_j$ is defined as

$$\mathcal{C}_j := \{i : j = \arg\min_{s=[k]} \|\Phi_i - \mathbf{c}_s\|^2\}. \quad (1)$$

The expected squared norm criterion is defined as

$$W(\mathbf{C}, \mu) := \mathbb{E}_{\Phi\sim\mu}[\min_{j=[k]} \|\Phi - \mathbf{c}_j\|^2]. \quad (2)$$

The excess clustering risk $\mathcal{E}(\mathbf{C}_n)$ of the empirical risk minimizer is defined as:

$$\mathcal{E}(\mathbf{C}_n) := \mathbb{E}_{\mathcal{S}\sim\mu}[W(\mathbf{C}_n, \mu)] - W^*(\mu), \quad (3)$$

where $W^*(\mu) := \inf_{\mathbf{C}\in\mathcal{H}^k} W(\mathbf{C}, \mu)$ is the optimal clustering risk.
The statistical properties of kernel k-means have been studied for decades, but they may not appear to be sufficient. And due to the high time and space requirements, it has no capability to large scale scenarios.

The Proposed Framework

We propose a framework of randomized sketches kernel k-means by reducing the original column $\mathbf{k}_i \in \mathbb{R}^n$ to an m-dimensional subspace of $\mathbb{R}^n$, where $m \ll n$ is the sketch dimension.
The proposed randomized sketches method:

$$\tilde{\mathbf{K}} = \mathbf{RK} = \mathbf{SQK} \in \mathbb{R}^{m\times n}. \quad (4)$$

The unified randomized sketches kernel k-means:

$$\bar{\mathbf{C}}_{n,m} = \arg\min_{\bar{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \left\|\tilde{\mathbf{k}}_i - \bar{\mathbf{c}}_j\right\|^2$$
$$= \frac{1}{n}\min_{\nu} \sum_{j=1}^{k}\sum_{i\in\mathcal{C}_j} \left\|\tilde{\mathbf{k}}_i - \frac{1}{|\mathcal{C}_j|}\sum_{s\in\mathcal{C}_j}\tilde{\mathbf{k}}_s\right\|^2. \quad (5)$$

Define the clustering centers by

$$\tilde{\mathbf{c}}_j = \frac{\sum_{i=1}^{n}\mathbf{k}_i\mathbb{I}_{\{\tilde{\mathbf{k}}_i\in\tilde{\mathcal{C}}_j\}}}{\sum_{i=1}^{n}\mathbb{I}_{\{\tilde{\mathbf{k}}_i\in\tilde{\mathcal{C}}_j\}}}, \quad j = 1,\dots,k. \quad (6)$$

Example 1

Sub-Gaussian Sketches Kernel k-Means:
The matrix $\mathbf{S} \in \mathbb{R}^{m\times m}$ in Eq.(4) is described by a hash function.

$$\mathbf{S}_{i,j} = \begin{cases} \sigma(i)/\sqrt{m}, & \textbf{with the pro } \frac{1}{\sqrt{n}}, \\ 0 & \textbf{with the pro } 1 - \frac{1}{\sqrt{n}}. \end{cases}$$

- Time complexity: $\mathcal{O}(\sqrt{n}m^2 + nmkt)$.
- Space complexity: $\mathcal{O}(nm)$.

Example 2

ROS Sketches Kernel k-Means:
$\mathbf{S} \in \mathbb{R}^{m\times m}$ in Eq.(4) can be defined as below:

$$\mathbf{S} = \mathbf{DA}. \quad (7)$$

$\mathbf{D} \in \mathbb{R}^{m\times m}$ is a random diagonal matrix whose entries are i.i.d. Rademacher variables. $\mathbf{A} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix with uniformly bounded entries, for example the Hadamard matrix and the discrete Fourier transform matrix. We use the Hadamard matrix in this paper.
The Hadamard matrix:

$$\mathbf{A}_m = \begin{bmatrix} \mathbf{A}_{m/2} & \mathbf{A}_{m/2} \\ \mathbf{A}_{m/2} & -\mathbf{A}_{m/2} \end{bmatrix}$$

with $\mathbf{A}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\mathbf{A} = \frac{1}{\sqrt{m}}\mathbf{A}_m$.

- Time complexity: $\mathcal{O}(nm\log m + nmkt)$.
- Space complexity: $\mathcal{O}(nm)$.

Example 3

Nyström Kernel k-Means:
$\mathbf{S} \in \mathbb{R}^{m\times m}$ in Eq.(4) can be defined as below:

$$\mathbf{S} = \mathbf{I},$$

where $\mathbf{I}$ is an identify matrix.
Therefore, the proposed Nyström kernel k-means can be converted into:

$$\tilde{\mathbf{C}}_{n,m} = \arg\min_{\tilde{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \left\|\tilde{\mathbf{k}}_i - \tilde{\mathbf{c}}_j\right\|^2$$
$$= \arg\min_{\tilde{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \left\|\tilde{\Phi}_i - \tilde{\mathbf{c}}_j\right\|^2, \quad (8)$$

where $\tilde{\Phi}_i = \mathbf{\Phi}_m^T\Phi_i$, $\tilde{\mathbf{c}}_j = \mathbf{\Phi}_m^T\mathbf{c}_j$, $\mathbf{\Phi}_m = [\Phi_{\pi(1)},\dots,\Phi_{\pi(m)}]$, $\pi(i) \in [1,n]$, and the dictionary (i.e., subset) $\{\Phi_{\pi(i)}\}_{i=1}^m$ is m points Φ_j sampled from $\{\Phi_j\}_{j=1}^n$ through $\mathbf{Q}$.

- Time complexity: $\mathcal{O}(nmkt)$.
- Space complexity: $\mathcal{O}(nm)$.

Simulation Experiment

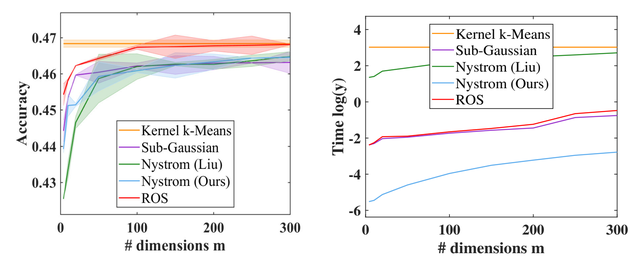

Figure 1: Test accuracy and training time (in seconds) with different dimensions m.

Theoretical Analysis

Theorem 1. *If $\|\Phi_{\mathbf{x}}\| \le 1$ for any $\mathbf{x} \in \mathcal{X}$, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and, in either one of the three cases of sub-Gaussian, ROS, and Nyström, the sketch dimension is $m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right)$, then, with probability at least $1-\delta$, we have $\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m},\mu)] - W^*(\mu) = \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right) + \mathcal{O}\left(\frac{\varepsilon}{1-\varepsilon}\right)$.*

We adopt the improved kernel k-means++ sampling for the proposed randomized sketches kernel k-means.

Theorem 2. *Let $C^+_{n,m}$ be obtained by the improved k-means++ algorithm with a local search strategy. If $\|\Phi_{\mathbf{x}}\| \le 1$ for any $\mathbf{x} \in \mathcal{X}$, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and, in either one of the three cases of sub-Gaussian, ROS, and Nyström, the sketch dimension is $m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right)$, then, with probability at least $1-\delta$, we have $\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m},\mu)\right]\right] = \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + W^*(\mu)\right) + \mathcal{O}\left(\frac{\varepsilon}{1-\varepsilon}\right)$, where $\mathcal{J}$ is the randomness derived from the k-means++ initialization.*

Experiments on Real-World Scenarios

Table 1: The datasets used in this paper. Test accuracy and training time (in seconds) on real datasets.

Dataset	Instance	Class	Kernel k-Means		Gaussian		Nyström (Liu)	
			Time	Accuracy	Time	Accuracy	Time	Accuracy
dna	2000	3	0.16	0.50±0.01	0.12	0.49± 0.02	0.09	0.50±0.02
segment	2310	7	0.13	0.50±0.02	0.09	0.45±0.03	0.05	0.43±0.01
mushrooms	8124	2	0.56	0.64±0.01	0.32	0.63±0.02	0.11	0.61±0.01
pendigits	10992	10	0.61	0.11 ±0.01	0.34	0.11±0.01	0.21	0.10± 0.02
protein	17766	3	5.07	0.46±0.01	3.16	0.44±0.03	1.09	0.45±0.02
a8a	32561	2	6.47	0.75±0.01	3.21	0.73±0.03	1.12	0.73±0.02
w7a	49749	2	29.7	0.97±0.02	15.3	0.95±0.02	1.36	0.96± 0.01
connect-4	67557	3	0.28	0.61±0.01	0.22	0.60±0.03	0.11	0.59±0.02
covtype	581012	7	/	/	/	/	/	/
Dataset	Instance	Class	Sub-Gaussian (Ours)		ROS (Ours)		Nyström (Ours)	
			Time	Accuracy	Time	Accuracy	Time	Accuracy
dna	2000	3	0.06	0.49± 0.01	0.07	0.50±0.01	0.04	0.50±0.01
segment	2310	7	0.03	0.47±0.03	0.03	0.49±0.01	0.02	0.42±0.01
mushrooms	8124	2	0.04	0.63±0.01	0.04	0.62±0.02	0.03	0.60±0.01
pendigits	10992	10	0.14	0.11±0.01	0.16	0.11± 0.01	0.03	0.11±0.02
protein	17766	3	0.16	0.45±0.01	0.21	0.46±0.01	0.03	0.44±0.02
a8a	32561	2	0.11	0.74±0.01	0.12	0.74±0.02	0.03	0.73±0.02
w7a	49749	2	0.30	0.94±0.02	0.36	0.95± 0.01	0.03	0.97±0.01
connect-4	67557	3	0.05	0.59±0.01	0.06	0.60±0.02	0.03	0.58±0.02
covtype	581012	7	1.02	0.32±0.02	1.36	0.33±0.04	0.66	0.32±0.03