SAMPLING-BASED ALGORITHMS FOR CLUSTERING PROBLEMS

ANUP KUMAR BHATTACHARYA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2018

SAMPLING-BASED ALGORITHMS FOR CLUSTERING PROBLEMS

by

ANUP KUMAR BHATTACHARYA

Department of Computer Science and Engineering

Submitted

in Fulfillment of the Requirements of the Degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi

JULY 2018

Certificate

This is to certify that the thesis titled **"Sampling-based Algorithms for Clustering Problems"** being submitted by **Anup Kumar Bhattacharya** to the **Indian Institute of Technology Delhi**, for the award of the degree of **Doctor of Philosophy**, is a record of bonafide research carried out by him under my supervision. The results obtained in this thesis have not been submitted to any other university or institute for the award of any other degree or diploma.

RAGESH JAISWAL
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
Delhi 110016

Acknowledgements

I want to express my sincere thanks to my supervisor Prof. Ragesh Jaiswal. The journey of more than five years has not been without glitches, specially at the beginning. I am most thankful to him for being patient with me in all those moments of uncertainty. What started out as a student-advisor relationship, gradually became friendly and easy-going. This of course goes without saying that I learned most of what I know of theoretical computer science, and which seems to be the direction where I am headed in my career, by working with him.

I want to thank Prof. Amit Kumar. I learned a lot while working with him on different projects. I want to express gratitude to Prof. Sandeep Sen, Prof. Amit Kumar, Prof. Naveen Garg and Dr. Yogish Sabharwal for being committee members in my doctoral research committee, and for mentoring me throughout my doctoral studies.

This journey would have been rather dull without moments I shared with friends at IIT Delhi. I am refraining from naming anyone since I am afraid that I would miss somebody and they would be hurt. Rather let me just acknowledge that you played a crucial part in this journey and your presence is duly noted. Thanks to all those who took part in organizing Frisbee/cricket matches, and movie sessions. Special thanks to those with whom I shared many delightful early morning bicycle rides. I discovered the attraction of trekking during

my doctoral studies. Many thanks to those who fueled this desire. Thanks for moments I shared with the members of Dr group. A special thanks to the Nescafe shop and the adjoining area where innumerable memories were built over cups of coffee. I also want to thank my friends from college. And a special thanks to my master's thesis advisor Prof. Abhijit Das.

I want to thank Prof. Nir Ailon for hosting me during the summer of 2015 at Technion, Israel and for his generous hospitality. My sincere thanks go to Dr. Arijit Ghosh and Prof. Arijit Bishnu for hosting me at ISI Calcutta. A major portion of this thesis was written during my time at ISI Calcutta.

Lastly, I want to thank my family for supporting me through all these years. Literally, this thesis would not have seen the light of the day without you. I dedicate this thesis to my wife, my sister and my mother.

Anup Bhattacharya

Abstract

We study sampling based algorithms for clustering problems, in particular the k-means clustering problem. Given dataset $X \subseteq \mathbb{R}^d$ and an integer k as input, the objective in the k-means problem is to output a set $C \subseteq \mathbb{R}^d$ of k centers such that the k-means cost function, $\phi(C, X) = \sum_{x \in X} \min_{c \in C} ||c - x||^2$, is minimized. This gives a natural clustering of the points where all the points having the same nearest center in C belong to the same cluster. k-means clustering is one of the most widely used clustering problems in practice. The k-means problem is also known to be NP-hard. Approximation algorithms for k-means problem giving constant approximation guarantee in polynomial time are known. However, Lloyd's heuristic (also known as the k-means algorithm) is usually used to solve the k-means problem in practice. This is a local search technique described as follows: Starting with an arbitrary set of k centers, make local improvements until no further improvement is possible. Even though this heuristic performs well in practice, it does not give any theoretical guarantees. The method of choosing the initial k centers for Lloyd's heuristic is known as *seeding*. Note that seeding is done arbitrarily in Lloyd's heuristic. Arthur and Vassilvitskii [17] showed that a simple adaptive seeding technique, known as k-means++ seeding, gives $O(\log k)$-approximation for k-means in expectation.

The k-means++ seeding uses D^2-sampling to sample one center in each of the k iterations. D^2-sampling with respect to any center set C chooses a point with probability proportional to the squared Euclidean distance from the point to its nearest center in C. D^2-sampling is known to be very useful in designing algorithms for clustering problems. Our work in this context may be seen as an attempt to understand the scope and limitations of D^2-sampling based algorithms for k-means problem. Next, we describe the main results from this study.

Arthur and Vassilvitskii gave examples of instances on which k-means++ seeding yields $\Omega(\log k)$-approximation in expectation (along with a matching upper bound). Brunsch and Röglin [40] showed that there are instances on which k-means++ seeding gives $\Omega(\log k)$-approximation with probability close to one. However, these lower bound instances were high dimensional. It was conjectured that k-means++ seeding performs better on low dimensional datasets. We refute this conjecture by constructing low dimensional instances on which k-means++ seeding yields $\Omega(\log k)$-approximation with high probability [36].

In an attempt to better understand the k-means cost function, we study the relationship between the number of centers k and the optimal k-means cost for any dataset $X \subseteq \mathbb{R}^d$. For any $\varepsilon > 0$, we define $\mathcal{L}_X^{k,\varepsilon}$ to be the smallest number such that optimal $\mathcal{L}_X^{k,\varepsilon}$-means cost is at most ε times the optimal k-means cost on dataset X. We compute upper and lower bounds on this number, $\mathcal{L}_X^{k,\varepsilon}$ [35]. Our upper bound is better than the bounds known earlier where as no such lower bound result was previously known. This also helped us in understanding pseudo-approximation behaviour of the D^2-sampling based k-means++ seeding algorithm.

Ding and Xu [48] introduced a unifying framework for considering variants of the k-means problem where some constraints need to be satisfied by the optimal clustering in addition to optimizing the k-means cost function. One example of such a problem is the

r-gather clustering problem where the additional constraint is that the size of every cluster should be at least r. The authors gave a $(1+\varepsilon)$-approximation algorithm for a very general version of the constrained k-means problem. We design a much faster D^2-sampling based $(1+\varepsilon)$-approximation algorithm for the same problem [37]. We also give justification as to why the currently known algorithms have running time that is exponential in k and $\frac{1}{\varepsilon}$.

In recent works semi-supervised learning frameworks have been studied that incorporates domain knowledge to design polynomial time algorithms for NP-hard optimization problems. Ashtiani et al. [19] designed one such semi-supervised framework for active learning of clustering (SSAC) that has access to a *same-cluster* query oracle. The oracle gives answers to queries enquiring whether two objects belong to the same optimal cluster or not. They gave a polynomial time algorithm for k-means in this framework assuming some *separation* condition to hold for the dataset. We give a D^2-sampling based $(1+\varepsilon)$-approximation algorithm for k-means clustering in SSAC framework without any separation assumption on the dataset, and study upper and lower bounds on the query complexity for $(1+\varepsilon)$-approximation of k-means [10]. Our lower bound result is conditioned on the Exponential time hypothesis (ETH). We also get similar results for *correlation clustering* [9].

Clustering problems are routinely solved in practice. Since sampling algorithms play such a crucial role in the algorithms we design for clustering problems, we study how well these sampling algorithms can be implemented, particularly in space restricted settings. We focus on two such settings, the streaming setting and the query model. In streaming setting, we study uniform and non-uniform sampling algorithms. We design algorithms that optimize the number of required random bits and space [34].

सार

हम क्लस्टरिंग समस्याओं के लिए नमूना आधारित एल्गोरिदम का अध्ययन करते हैं, विशेष रूप से k-मीन्स क्लस्टरिंग समस्या। दिया गया डेटासेट X ⊆ R^d और इनपुट के रूप में एक पूर्णांक k, k-मीन्स समस्या का उद्देश्य k केंद्रों के सेट C ⊆ R^d को आउटपुट करना है जैसे कि k-मीन्स लागत फ़ंक्शन, φ(C, X) = x∈X min c∈C ||c − x||^2, कम किया गया है। यह उन बिंदुओं का प्राकृतिक क्लस्टरिंग देता है जहां C में समान निकटतम केंद्र वाले सभी बिंदु एक ही क्लस्टर से संबंधित होते हैं। k-मीन्स क्लस्टरिंग अभ्यास में सबसे व्यापक रूप से इस्तेमाल क्लस्टरिंग समस्याओं में से एक है। k-मीन्स समस्या को एनपी-हार्ड भी कहा जाता है। k-मीन्स समस्या के लिए अनुमानित एल्गोरिदम बहुपद समय में निरंतर अनुमान गारंटी प्रदान करते हैं। हालांकि, लॉयड के ह्युरिस्टिक (जिसे k-मीन्स एल्गोरिदम भी कहा जाता है) आमतौर पर अभ्यास में k-मीन्स समस्या को हल करने के लिए उपयोग किया जाता है। यह निम्नानुसार वर्णित एक स्थानीय खोज तकनीक है: k केंद्रों के मनमानी सेट से शुरू होने तक, स्थानीय सुधार करें जब तक कि कोई और सुधार संभव न हो। भले ही यह ह्युरिस्टिक अभ्यास में अच्छा प्रदर्शन करता है, यह किसी भी सैद्धांतिक गारंटी नहीं देता है। लॉयड के हेरिस्टिक के लिए प्रारंभिक के केंद्रों को चुनने की विधि को सीडिंग के रूप में जाना जाता है। ध्यान दें कि सीडिंग लॉयड के हेरिस्टिक में मनमाने ढंग से किया जाता है। आर्थर और वासिलविट्स्की [17] ने दिखाया कि एक सरल अनुकूली सीडिंग तकनीक, जिसे k-मीन्स++ सीडिंग के नाम से जाना जाता है, उम्मीद में k-मीन्स के लिए ओ O(log k)-एप्रोक्साइमिशन देता है। k-मीन्स++ सीडिंग D^2-सैंपलिंग का उपयोग प्रत्येक के पुनरावृत्तियों में एक केंद्र का नमूना करने के लिए करता है। किसी भी केंद्र सेट C के संबंध में D^2-सैंपलिंग बिंदु से यूक्लिडियन दूरी के लिए आनुपातिक संभावना के साथ एक बिंदु चुनता है, सी। D^2-सैंपलिंग में निकटतम केंद्र में क्लस्टरिंग समस्याओं के लिए एल्गोरिदम डिजाइन करने में बहुत उपयोगी माना जाता है । इस संदर्भ में हमारे काम को के-साधन समस्या के लिए D^2-सैंपलिंग आधारित एल्गोरिदम के दायरे और सीमाओं को समझने के प्रयास के रूप में देखा जा सकता है। इसके बाद, हम इस अध्ययन से मुख्य परिणामों का वर्णन करते हैं।

आर्थर और वासिलविट्स्की ने उदाहरणों के उदाहरण दिए जिन पर k-मीन्स++ सीडिंग पैदावार Ω(log k) - उम्मीद में अनुमान (ऊपरी बाउंड के साथ)। ब्रंसच और रोग्लिन [40] ने दिखाया कि ऐसे उदाहरण हैं जिन पर k-मीन्स++ सीडिंग Ω(log k) देता है - एक के करीब संभावना के साथ अनुमान। हालांकि, ये निचले बाध्य उदाहरण उच्च आयामी थे। यह अनुमान लगाया गया था कि k-मीन्स++ सीडिंग कम आयामी डेटासेट पर बेहतर प्रदर्शन करती है। हम इस अनुमान को कम आयामी उदाहरणों का निर्माण करके अस्वीकार करते हैं जिन पर k-मीन्स++ सीडिंग उपज Ω(log k) - उच्च संभावना वाले प्रतिकृति [36]।

k-मीन्स लागत समारोह को बेहतर ढंग से समझने के प्रयास में, हम किसी भी डेटासेट X ⊆ R^d के लिए केंद्रों की संख्या और इष्टतम के-साधन लागत के बीच संबंधों का अध्ययन करते हैं। किसी भी ε> 0 के लिए, हम L_ {k, ε} (X) को सबसे छोटी संख्या के रूप में परिभाषित करते हैं जैसे कि इष्टतम L_ {k, ε} (X) -मेन्स लागत डेटासेट पर इष्टतम k-मीन्स लागत सबसे अधिक ε बार होती है X. हम इस नंबर पर ऊपरी और निचली सीमाओं की गणना करते हैं, L_ {k, ε} (X) [35]। हमारी ऊपरी सीमा पहले ज्ञात सीमाओं से बेहतर है, जहां पहले इस तरह के निचले बाध्य परिणाम को ज्ञात नहीं किया गया था। इससे हमें D^2-सैंपलिंग आधारित k-मीन्स++ सीडिंग एल्गोरिदम के छद्म-अनुमानित व्यवहार को समझने में भी मदद मिली।

डिंग और जू [48] ने के-के संस्करणों पर विचार करने के लिए एक एकीकृत ढांचा प्रस्तुत किया।
इसका मतलब है कि k-मीन्स लागत समारोह को अनुकूलित करने के अलावा कुछ बाधाओं को इष्टतम क्लस्टरिंग से संतुष्ट होने की आवश्यकता है। ऐसी समस्या का एक उदाहरण आर-इकट्ठा क्लस्टरिंग समस्या है जहां अतिरिक्त बाधा यह है कि प्रत्येक क्लस्टर का आकार कम से कम आर होना चाहिए। लेखकों ने बाध्यकारी के-साधन समस्या के एक बहुत ही सामान्य संस्करण के लिए एक (1+ε)-एप्रोक्साइमिशन एल्गोरिदम दिया। हम एक ही समस्या के लिए एक बहुत तेजी से डी ^ 2-नमूना आधारित (1+ε)-एप्रोक्साइमिशन एल्गोरिदम डिजाइन [37]। हम औचित्य भी देते हैं कि वर्तमान में ज्ञात एल्गोरिदम में समय चल रहा है जो कि के और 1 / ε में घातीय है।

हाल के कार्यों में अर्द्ध पर्यवेक्षित शिक्षण ढांचे का अध्ययन किया गया है जिसमें एनपी-हार्ड ऑप्टिमाइज़ेशन समस्याओं के लिए बहुपद समय एल्गोरिदम डिजाइन करने के लिए डोमेन ज्ञान शामिल है। एसटिआनी एट अल। [19] क्लस्टरिंग (एसएसएसी) की सक्रिय शिक्षा के लिए इस तरह के एक अर्द्ध पर्यवेक्षित ढांचे को डिजाइन किया गया है जिसमें एक ही क्लस्टर क्वेरी ऑरैकल तक पहुंच है। ऑरैकल पूछताछ के जवाब देता है कि क्या दो ऑब्जेक्ट एक ही इष्टतम क्लस्टर से संबंधित हैं या नहीं। उन्होंने इस ढांचे में डेटासेट के लिए कुछ अलगाव स्थिति रखने के लिए के-साधनों के लिए बहुपद समय एल्गोरिदम दिया। हम डेटासेट पर किसी भी पृथक्करण धारणा के बिना एसएसएसी फ्रेमवर्क में k-मीन्स क्लस्टरिंग के लिए D^2-सैंपलिंग आधारित (1+ε)-प्रोक्सिमेशन एल्गोरिदम देते हैं, और क्वेरी जटिलता (1+ε) के लिए ऊपरी और निचली सीमाओं का अध्ययन करते हैं - के-साधनों का अनुमान [10]। हमारे निचले बाध्य परिणाम घातीय समय परिकल्पना (ईटीएच) पर सशर्त है। हमें सहसंबंध क्लस्टरिंग के लिए भी इसी तरह के परिणाम मिलते हैं [9]।

क्लस्टरिंग समस्याओं को नियमित रूप से अभ्यास में हल किया जाता है। नमूना एल्गोरिदम खेलने के बाद से क्लस्टरिंग समस्याओं के लिए डिजाइन किए गए एल्गोरिदम में ऐसी महत्वपूर्ण भूमिका, हम अध्ययन करते हैं कि इन नमूना एल्गोरिदम को कितनी अच्छी तरह से कार्यान्वित किया जा सकता है, खासकर स्थान प्रतिबंधित सेटिंग्स में। हम दो ऐसी सेटिंग्स, स्ट्रीमिंग सेटिंग और क्वेरी मॉडल पर ध्यान केंद्रित करते हैं। स्ट्रीमिंग सेटिंग में, हम समान और गैर समान नमूना एल्गोरिदम का अध्ययन करते हैं। हम एल्गोरिदम डिज़ाइन करते हैं जो आवश्यक यादृच्छिक बिट्स और स्पेस की संख्या को अनुकूलित करते हैं [34]।

Contents

List of Figures

List of Tables

List of Symbols

$\phi(C, X)$	k-means cost of X with respect to set C
$D(x, c)$	Squared Euclidean distance between x and c
$C^\star$	Set of optimal centers in k-means
$\Delta_k(X)$	Optimal k-means cost of dataset X
$\phi(c, X)$	k-means cost of X with respect to center c
$\mu(Y)$	Centroid of points in Y
$\mathcal{L}_X^{k,\varepsilon}$	Smallest number such that optimal $\mathcal{L}_X^{k,\varepsilon}$-means cost is at most ε-times optimal k-means cost
$\mathbb{O}^\star$	Arbitrary partition $\mathbb{O}^\star$ of X into k clusters $O_1^\star, \ldots, O_k^\star$
$\phi(C, \mathbb{O}^\star)$	k-means cost of clustering $\mathbb{O}^\star$ by center set C
$\lambda(x, r)$	The covering number of set X with respect to diameter r
$A^{\mathbb{C}}$	Partition algorithms in Constrained k-means
$\mathtt{opt}_k(\mathbb{O}^\star)$	Optimal k-means cost of clustering $\mathbb{O}^\star$

$\Delta(X)$ Optimal 1-means cost of dataset X