Theoretical Computer Science 634 (2016) 55–66

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Tight lower bound instances for *k*-means++ in two dimensions

Anup Bhattacharya [a], Ragesh Jaiswal [a,*,1], Nir Ailon [b,2]

[a] *IIT Delhi, India*
[b] *Technion, Haifa, Israel*

ARTICLE INFO

Article history:
Received 24 April 2015
Received in revised form 8 April 2016
Accepted 8 April 2016
Available online 13 April 2016
Communicated by V.Th. Paschos

Keywords:
k-means++
Lower bounds

ABSTRACT

The *k*-means++ seeding algorithm is one of the most popular algorithms that is used for finding the initial k centers when using the Lloyd's algorithm for the *k*-means problem. It was conjectured by Brunsch and Röglin [9] that *k*-means++ behaves well for datasets with small dimension. More specifically, they conjectured that the *k*-means++ seeding algorithm gives $O(\log d)$ approximation with high probability for any d-dimensional dataset. In this work, we refute this conjecture by giving two dimensional datasets on which the *k*-means++ seeding algorithm achieves an $O(\log k)$ approximation ratio with probability exponentially small in k. This solves open problems posed by Mahajan et al. [12] and by Brunsch and Röglin [9].

1. Introduction

The *k*-means clustering problem is one of the most important problems in Data Mining and Machine Learning that has been widely studied. The problem is defined as follows:

(*k*-means problem): Given a set of n points $X = \{x_1, \ldots, x_n\}$ in a d-dimensional space, find a set of k points $C = \{c_1, \ldots, c_k\}$ (these are called *centers*) such that the cost function $\Phi_C(X) = \sum_{x \in X} \min_{c \in C} D(x, c)$ is minimized. Here $D(x, c)$ denotes the square of the Euclidean distance between points x and c.

The problem is known to be NP-hard even for small values of the parameters such as when $k = 2$ [10] and when $d = 2$ [13,12]. There are various approximation algorithms for the problem. However, in practice, a heuristic known as the *k*-means algorithm (also known as Lloyd's algorithm) is used because of its excellent performance on real datasets even though it does not give any performance guarantees. This algorithm is simple and can be described as follows:

(*k*-means Algorithm): (i) Arbitrarily, pick k points C as centers. (ii) Cluster the given points based on the nearest distance to centers in C. (iii) For all clusters, find the mean of all points within a cluster and replace the corresponding member of C with this mean. Repeat steps (ii) and (iii) until convergence.

* Corresponding author.
E-mail addresses: anupb@cse.iitd.ernet.in (A. Bhattacharya), rjaiswal@cse.iitd.ac.in, rjaiswal@cse.iitd.ernet.in (R. Jaiswal), nailon@cs.technion.ac.il (N. Ailon).

[1] Ragesh Jaiswal acknowledges the support of the ISF–UGC India–Israel joint research grant 2014.
[2] Nir Ailon acknowledges the support of a Marie Curie International Reintegration Grant PIRG07-GA-2010-268403, as well as the support of The Israel Science Foundation (ISF) no. 1271/13.

http://dx.doi.org/10.1016/j.tcs.2016.04.012
0304-3975/

Even though the above algorithm performs very well on real datasets, it guarantees only convergence to local minima. This means that this *local search* algorithm may either converge to a local optimum solution or may take a large amount of time to converge [5,6]. Poor choice of the initial k centers (step (i)) is one of the main reasons for its bad performance with respect to approximation factor. A number of *seeding* heuristics have been suggested for choosing the initial centers. One such seeding algorithm that has become popular is the k-means++ seeding algorithm. The algorithm is extremely simple and runs very fast in practice. Moreover, this simple randomized algorithm also gives an approximation factor of $O(\log k)$ in expectation [7]. In practice, this seeding technique is used for finding the initial k centers to be used with the k-means algorithm and this ensures a theoretical approximation guarantee. The simplicity of the algorithm can be seen by its simple description below:

(k-means++ seeding): Pick the first center randomly from the given points. After picking $(i-1)$ centers, pick the ith center to be a point p with probability proportional to the square of the Euclidean distance of p to the closest previously $(i-1)$ chosen centers.

A number of recent work has been done in understanding the power of this simple sampling based approach for clustering. We discuss these in the following paragraph.

1.1. Related work

There is a *discrete* version of the classical k-means problem. In the discrete version, the centers are constrained to be a subset of the given points. It can be shown that any optimal solution to the discrete version of the k-means problem is a 2-factor approximation solution for the k-means problem. All the results discussed below are for the discrete version of the k-means problem. However, since the approximation guarantees are constant (or worse), similar approximation guarantees also hold for the k-means problem. Arthur and Vassilvitskii [7] showed that the sampling algorithm gives an approximation guarantee of $O(\log k)$ in expectation. They also gave an example dataset on which this approximation guarantee is best possible. Ailon et al. [3] and Aggarwal et al. [2] showed that sampling more than k centers in the manner described above gives a constant *pseudo-approximation*.[3] Ackermann and Blömer [1] showed that the results of Arthur and Vassilvitskii [7] may be extended to a large class of other distance measures. Jaiswal and Garg [11] and Agarwal et al. [4] showed that if the dataset satisfies certain separation conditions, then the seeding algorithm gives constant approximation with probability $\Omega(1/k)$. Bahmani et al. [8] showed that the seeding algorithm performs well even when fewer than k sampling iterations are executed provided that more than one center is chosen in a sampling iteration. We now discuss our main results.

1.2. Main results

The lower-bound examples of Arthur and Vassilvitskii [7] and Aggarwal et al. [2] have the following two properties: (a) the examples are high dimensional, and (b) the examples lower-bound the *expected* approximation factor. That is, they gave high dimensional datasets on which k-means++ seeding gives an approximation factor of $\Omega(\log k)$ in expectation. Note that such results do not rule out the possibility that k-means++ seeding gives a constant approximation with probability that is not too small (say $1/poly(k)$). Brunsch and Röglin [9] gave a high dimensional dataset where this is not true and showed that an $O(\log k)$ approximation is achieved with probability exponentially small in k. More specifically, they showed that the k-means++ seeding gives an approximation ratio of at most $(2/3-\epsilon)\cdot\log k$ only with probability that is exponentially small in k.

An important open problem mentioned in their work was to understand the behavior of the seeding algorithm on low-dimensional datasets. This problem was also mentioned as an open problem by Mahajan et al. [12] who showed that the *planar* (dimension $= 2$) k-means problem is NP-hard. In this work, we construct two dimensional datasets on which the k-means++ seeding algorithm achieves an approximation ratio $O(\log k)$ with probability exponentially small in k. More formally, here is the main theorem that we prove in this work.

Theorem 1 *(Main theorem). Let $r(k) = \delta\cdot\log k$ for a fixed real $\delta \in \left(0, \frac{1}{120}\right)$. There exists a family of instances for which k-means++ achieves an $r(k)$-approximation with probability at most $2^{-k} + e^{\left(-(k-1)^{1-120\delta-o(1)}\right)}$.*

Note that the theorem refutes the conjecture by Brunsch and Röglin [9]. They conjectured that the k-means++ seeding algorithm gives an $O(\log d)$-approximation for any d-dimensional instance.

[3] Here pseudo-approximation means that the algorithm is allowed to output more than k centers but the approximation factor is computed by comparing with the optimal solution with k centers.