Computational Geometry ••• (••••) •••–•••

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

A streaming algorithm for 2-center with outliers in high dimensions

Behnam Hatami, Hamid Zarrabi-Zadeh *

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 31 October 2015
Accepted 1 July 2016
Available online xxxx

Keywords:
k-Center
Outlier
High dimensions
Data stream

ABSTRACT

We study the 2-center problem with outliers in high-dimensional data streams. Given a stream of points in arbitrary d dimensions, the goal is to find two congruent balls of minimum radius covering all but at most z points. We present a $(1.8+\varepsilon)$-approximation streaming algorithm, improving over the previous $(4+\varepsilon)$-approximation algorithm available for the problem. The space complexity and update time of our algorithm are $\text{poly}(d, z, 1/\varepsilon)$, independent of the size of the stream.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The k-center problem—covering a set of points using k congruent balls of minimum radius—is a fundamental problem, arising in many applications such as data mining, machine learning, statistics, and image processing. In real-world applications where input data is often noisy, it is very important to consider outliers, as even a small number of outliers can greatly affect the quality of the solution. In particular, the k-center problem is very sensitive to outliers, and even a constant number of outliers can increase the radius of the k-center unboundedly. Therefore, it is natural to consider the following generalization of the k-center problem: given a set P of n points in arbitrary d dimensions and a bound z on the number of outliers, find k congruent balls of minimum radius to cover at least $n-z$ points of P. See Fig. 1 for an example. In this paper, we focus on the *data stream* model of computation where only a single pass over the input is allowed, and we have only a limited amount of working space available. This model is in particular useful for processing massive data sets, as it does not require the entire data set to be stored in memory.

The Euclidean k-center problem has been extensively studied in the literature. If k is part of the input, the problem in known to be NP-hard in two and more dimensions [10], and is even hard to approximate to within a factor better than 1.82, unless P = NP [9]. Factor-2 approximation algorithms are available for the problem in any dimension [9,11]. For small k and d, better solutions are available. The 1-center problem in fixed dimensions is known to be LP-type and can be solved in $O(n)$ time [7]. For 2-center in the plane, the current best algorithm runs in $O(n\log^2 n\log^2\log n)$ time [5].

For k-center with outliers, Charikar *et al.* [6] gave the first algorithm with an approximation factor of 3, which works in any dimension. Better results are known for small k in the plane. The 1-center problem with z outliers in the plane can be solved in $O(n\log n + z^3 n^{\varepsilon})$ time, for any $\varepsilon > 0$, using Matoušek's framework [14]. Agarwal [1] gave a randomized $O(nz^7\log^3 z)$-time algorithm for 2-center with z outliers in the plane.

* Corresponding author.
E-mail addresses: bhatami@ce.sharif.edu (B. Hatami), zarrabi@sharif.edu (H. Zarrabi-Zadeh).

http://dx.doi.org/10.1016/j.comgeo.2016.07.002
0925-7721/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: B. Hatami, H. Zarrabi-Zadeh, A streaming algorithm for 2-center with outliers in high dimensions, Comput. Geom. (2016), http://dx.doi.org/10.1016/j.comgeo.2016.07.002

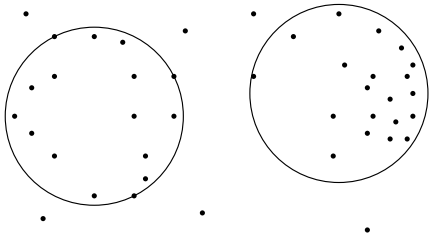

Fig. 1. An example of 2-center with 6 outliers.

Table 1
Summary of the streaming algorithms for k-center with and without outliers in high dimensions.

Problem	Approximation factor	
	Without outliers	With outliers
1-center	1.22 [4]	1.73 [17]
2-center	$1.8+\varepsilon$ [13]	$1.8+\varepsilon$ [Here]
k-center	$2+\varepsilon$ [12,15]	$4+\varepsilon$ [15]

In the streaming model, where only a single pass over the input is allowed, McCutchen and Khuller [15] and independently Guha [12] presented algorithms to maintain a $(2+\varepsilon)$-approximation to k-center in any dimension using $O((kd/\varepsilon)\log(1/\varepsilon))$ space. For $k=1$, Zarrabi-Zadeh and Chan [16] presented a simple algorithm achieving an approximation factor of $3/2$ using only $O(d)$ space. Agarwal and Sharathkumar [2] improved the approximation factor to $(1+\sqrt{3})/2+\varepsilon \approx 1.37$ using $O((d/\varepsilon^3)\log(1/\varepsilon))$ space. The approximation factor of their algorithm was later improved to 1.22 by Chan and Pathak [4]. For $k=2$, Kim and Ahn [13] have recently obtained a $(1.8+\varepsilon)$-approximation using $O(d/\varepsilon)$ space. Their algorithm extends to any fixed k, with the same approximation factor.

For k-center with z outliers in the streaming model, McCutchen and Khuller [15] gave a $(4+\varepsilon)$-approximation algorithm using $O(\frac{zk}{\varepsilon})$ space. When dimension is fixed, a $(1+\varepsilon)$-approximation to 1-center with outliers can be maintained in $O(z/\varepsilon^{((d-1)/2)})$ space using the notion of robust ε-kernels [3,18]. For 1-center with outliers in high dimensions, Zarrabi-Zadeh and Mukhopadhyay [17] gave a $(\sqrt{2}\alpha)$-approximation, where α is the approximation factor of the underlying algorithm for maintaining 1-center. Combined with the 1.22-approximation algorithm of Chan and Pathak [4], it yields an approximation factor of $(\sqrt{2}\times 1.22)\approx 1.73$ using $O(d^3 z)$ space.

Our result In this paper, we study the 2-center problem with outliers in high dimensional data streams. We present a streaming algorithm that achieves an approximation factor of $1.8+\varepsilon$, for any $\varepsilon>0$, using $\text{poly}(d, z, \frac{1}{\varepsilon})$ space and update time. This improves over the previous $(4+\varepsilon)$-approximation streaming algorithm available for the problem presented by McCutchen and Khuller [15]. The approximation factor of our algorithm matches that of the best streaming algorithm for the 2-center problem with no outliers. This is somewhat surprising, considering that the current best approximation factors for streaming k-center with and without outliers differ by a multiplicative factor of $\sqrt{2}$ for $k=1$ [4,17], and by a factor of 2 for general k [12,15]. See Table 1 for a comparison.

To obtain our result, we have used a combination of several ideas including parallelization, far/close ball separation, centerpoint theorem, and keeping lower/upper bounds on the radius and distance of the optimal balls. We have also employed ideas of [13] for the 2-center problem with no outliers. However, our problem is much harder here, as we not only need to find balls of minimum radius, but we also need to decide which subset of points to cluster. This is in particular more challenging in the streaming model, where we only have a single pass over the input, and we must decide on the fly which point is an outlier, and which one can be safely ignored as a non-outlier point, to comply with the working space restriction enforced by the model.

2. Preliminaries

Let $B(c,r)$ denote a ball of radius r centered at c. We use $r(B)$ to denote the radius of a ball B. For two points p and q, the distance between p and q is denoted by $\|pq\|$. Given two balls $B(c,r)$ and $B'(c',r')$, we define $\delta(B,B') \equiv \max\{0, \|cc'\|-r-r'\}$ to be the *distance* between B and B'. Two balls B_1 and B_2 are said to be α-*separated*, if $\delta(B_1,B_2) > \alpha\cdot\max\{r(B_1), r(B_2)\}$.

Given an n-point set P in d-dimensions, a point $c\in\mathbb{R}^d$ is called a *centerpoint* of P, if any halfspace containing c contains at least $\lceil n/(d+1)\rceil$ points of P. It is well-known that any finite set of points in d dimensions has a centerpoint [8]. The following observation is a corollary of this fact.

Please cite this article in press as: B. Hatami, H. Zarrabi-Zadeh, A streaming algorithm for 2-center with outliers in high dimensions, Comput. Geom. (2016), http://dx.doi.org/10.1016/j.comgeo.2016.07.002

Download English Version:

https://daneshyari.com/en/article/4949140

Download Persian Version:

https://daneshyari.com/article/4949140

Daneshyari.com