An Efficient k-Center Clustering Algorithm for Geometric Objects

Guang Xu and Jinhui Xu
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260, USA
{ *guangxu, jinhui*}*@cse.buffalo.edu*

1 Overview

Clustering is a fundamental problem in computational geometry and finds numerous applications in many different fields such as data mining, image processing, and pattern classification and recognition. Extensive research has been done on investigating theoretically and/or practically efficient approaches for solving various variants of this problem, and a number of important results have been obtained in recent years. Among many of its variants, the Euclidean k-center clustering problem (or more generally, the k-center clustering problem in a metric space) of a set of points has received considerable attention from the fields of computational geometry and approximation algorithms, and an $O(n \log k)$-time, 2-approximation algorithm has been obtained [4, 5] by using a farthest point method. The success of this constant approximation algorithm relies on a key fact that the distances among the set of points satisfy the metric property. This algorithm, although simple, is actually quite close to the limit of approximation. It has been shown that approximating the Euclidean k-center problem within any ratio smaller than 1.822 is NP-hard [3].

In this paper, we consider an interesting generalization of the Euclidean k-center clustering problem, called *k-center clustering problem of point sets (or KCS for short)* in d-dimensional Euclidean space R^d. In the KCS problem, we are given a sequence of finite point sets in R^d space, denoted by $S_1, S_2, \ldots, S_n$, respectively, and an integer $k \geq 1$, and are required to find k congruent spheres $B_1, B_2, \ldots, B_k$ of minimum radius so that for each $S_i, 1 \leq i \leq n$, there is at least one ball $B_j, 1 \leq j \leq k$, containing all points in S_i. The KCS problem models the k-center clustering problem of a set of bounded geometric objects represented by polyhedra. This is because for any polyhedron, it is sufficient to consider the vertices of its convex hull. Different from the ordinary k-center clustering problem, in our KCS problem, the geometric objects represented by S_i and S_j, for any $i, j \leq n$, could overlap with each other.

The KCS problem we study is motivated by applications in biology where a set of replication/transcription sites, each represented by a 3-D polyhedron, needs to be clustered into clusters to form replication or transcription zones for studying the behavior of chromatic domains in cell nucleus. It also appears in many applications of the Euclidean k-center clustering problem where each point in the k-center clustering problem is actually the representative point (e.g., the center) of some geometric object. When the sizes of the geometric objects are small enough, simplifying each object to a point does not affect the quality of clustering too much, since the distance function of the set of objects can be well approximated by that of the representative points. However, when the size of each object is large, the distance between two objects could be significantly different from that of their representative points. Thus the quality of the obtained clustering through representative points could be rather poor.

In general, the distances among a set of geometric objects do not satisfy the triangle inequality, thus making almost all previously known k-center clustering algorithms fail. To overcome the difficulty caused by the loss of the triangle inequality property, we use the ideas of shape simplification and core set. More specifically, we first simplify each object (or point set) to some simple geometry shape, and then determine several representative points from each simplified shape. The representative points of each object may not belong to its original object. In order to achieve a quality solution, the simplified shapes and representative points should have the following features:

1. The size and shape of the original objects are somewhat preserved.

2. The position of the original objects are well reflected.

Based on several interesting observations, we will show that each geometric object can be represented by a small set of representative points, and an efficient 3-approximation algorithm can be obtained by clustering the set of representative points. In summary, we prove the following theorems.

Theorem 1 *There is an $O(m+n\log k)$ time 3-approximation algorithm for the k-center clustering problem of point sets in R^d space, where $m=\sum_{i=1}^{i=n}|S_i|$.*

The constant hidden in the O-notation of the running time has exponential dependence on the dimension d. If the dimension d is very high, we have the following result.

Theorem 2 *There is an $O(\frac{dm}{\epsilon^2}+dnk+\frac{1}{\epsilon^{O(1)}})$ time $3+\epsilon$-approximation algorithm for the k-center clustering problem of point sets in R^d space.*

References

[1] M. Bădoiu, S. Har-Peled, P. Indyk. "Approximate Clustering via Core-Sets," *Proceedings of the 34th ACM Symposium on Theory of Computing*, pp. 250-257, 2002.

[2] M. Bern and D. Eppstein. "Approximation algorithms for geometric problems," In *Approximation algorithms for NP-Hard problems*, D.S. Hochbaum (editors), PWS Publishing Company, 1997.

[3] T. Feder and D. H. Greene. "Optimal algorithm for approximation clustering," *Proc. 20th ACM Symp. Theory of Computing,* pages 434-444, 1988.

[4] T. Gonzalez. "Clustering to minimize the maximum intercluster distance," *Theoritical Computer Science*, Vol. 38, pages 293-306, 1985.

[5] P. Vaidya. "An $O(n\log n)$ algorithm for the all-nearest-neighbors problem," *Disc. and Comp. Geometry*, Vol. 4, pages 101-115, 1989.

[6] V. V. Vazirani. "Approximation Algorithms," Springer, 2001.