Geometric Clusterings*

(extended abstract)

Vasilis Capoyleas †
Günter Rote ‡§¶**
Gerhard Woeginger ¶††

Abstract. A k-clustering of a given set of points in the plane is a partition of the points into k subsets ("clusters"). For any fixed k, we can find a k-clustering which minimizes any monotone function of the diameters or the radii of the clusters in polynomial time. The algorithm is based on the fact that any two clusters in an optimal solution can be separated by a line.

1 Introduction

Problem statement. Let S be a set of n points in the plane. A partitioning of S into k disjoint (possibly empty) sets $C_1, C_2, \ldots, C_k$, is called a *k-clustering*, and the individual sets C_i are called its *clusters*. In cluster analysis, the points represent properties (data) of real-world objects, and the aim is to collect "similar" objects (points which are close to each other) in the same cluster, and to put objects which are very "different" into different clusters.

Let W be some weight function that assigns a real weight to any finite set C of points in the plane, like the diameter of C, the radius of C, or the perimeter or the area of its convex hull. Further, let $\mathcal{F}$ be a k-ary symmetric function, assigning a real value to every k-tuple of reals. Examples for $\mathcal{F}$ are the sum or the maximum.

The *geometric k-clustering problem for W with respect to $\mathcal{F}$* is defined as follows:

INSTANCE: A set S of n points in the plane; a rational number d.
QUESTION: Is there a k-clustering for S into k sets $C_1, C_2, \ldots, C_k$ such that $\mathcal{F}(W(C_1), W(C_2), \ldots, W(C_k)) \leq d$?

Previous work and our result. If k is part of the input, this problem is in general NP-complete. Supowit [16] has shown this result for W being the diameter and for $\mathcal{F}$ being the maximum function; in other words, for k part of the input, minimizing the maximum diameter in a k-clustering is NP-complete. The related problem of minimizing the maximum radius,

*This work was initiated independently by Vasilis Capoyleas and by the other two authors.

†Rutgers University, Computer Science Department, Piscataway, New Jersey 08854, USA
‡University of Waterloo, Dept. of Combinatorics and Optimization, Waterloo, Ontario, Canada N2L 3G1
§On leave from ††

¶The work of Günter Rote and Gerhard Woeginger was done while they were at the Freie Universität Berlin, Fachbereich Mathematik, Institut für Informatik; their work was partially supported by the ESPRIT II Basic Research Actions Program of the EC under contract no. 3075 (project ALCOM).
**Günter Rote also acknowledges the support by the Fonds zur Förderung der wissenschaftlichen Forschung, Projekt S32/01.

††Technische Universität Graz, Institut für Mathematik, Kopernikusgasse 24, A-8010 Graz, Austria

which is also known as the k-center problem in the area of location problems, is also NP-complete (Megiddo and Supowit [14]). For fixed k, minimizing the maximum radius has been shown to be polynomial by Drezner [6]. NP-completeness can also be shown for minimizing the maximum cluster area and for minimizing the sum of all cluster areas, as follows from a result of van Emde Boas [9] that it is NP-complete to decide whether a set of points can be covered by a given number of lines. For more information, the interested reader is referred to Johnson's NP-Completeness Column [13].

In this note we show that for every **fixed** k, the geometric k-clustering problem becomes solvable in polynomial time, if W and $\mathcal{F}$ are as follows:

- W is the diameter or the radius;
- $\mathcal{F}$ is an arbitrary monotone increasing function.

Standard examples for monotone increasing functions $\mathcal{F}$ are the *maximum*, the *sum*, or the *sum of the squares* of k non-negative arguments. The 2-clustering problem for the maximum diameter has been treated by Asano, Bhattacharya, Keil, and Yao [1]. They gave an $O(n \log n)$ algorithm for this problem. Monma and Suri [15] gave an $O(n^2)$ algorithm for finding a 2-clustering with smallest sum of diameters.

Overview of the paper and related results. The key result that we will use is that for any given 2-clustering, there is always a 2-clustering which is at least as good as the given one (as regards the diameters or the radii of *both* clusters in each clustering) and in which the two clusters can be separated by a line. For the case of radii, this is easy to see, whereas the proof for the case of diameters is more elaborate. It is the subject of section 2. This theorem allows us to limit the number of possible candidates for optimal solutions. From this, a polynomial-time algorithm, which essentially tests all these candidates, follows in a quite straightforward way. This algorithm is derived in section 3.

The problem of testing whether a 2-clustering with specified bounds on the two diameters exists has been treated by Hershberger and Suri [10,11]. They gave an $O(n \log n)$-time algorithm which does not use the separability of the two clusters.

A separability result related to ours was known for the problem of minimizing the sum of the variances of the clusters. (The variance of a cluster is the sum of the squares of the distances of all pairs of points in the cluster, divided by the number of points, cf. Bock [4, section 15, pp. 162–176].) With this objective function, two clusters of an optimal clustering are always separated by a line. Similarly, for the problem where the sum of the squares of all distances between points in the same cluster is to be minimized (without division by the cluster sizes), Boros and Hammer [3] showed that two clusters in an optimal solution can always be separated by circle. In both of these cases the separability result is due to the special form of the objective function.

Definitions and Notations. The convex hull of a point set A is denoted by $conv(A)$, the *diameter* (the maximum distance of two points in A) by $diam(A)$. By the *perimeter* of a point set A we mean the perimeter of it convex hull, i. e., the length of the boundary of $conv(A)$. The *radius* $r(A)$ of a finite point set A is the radius of the smallest enclosing circle. We can define the radius, the perimeter, and the diameter of the empty set as 0 or $-\infty$, as we like. Two sets are said to be *linearly separable*, if they can be strictly separated by a straight line. It is well known that two finite sets are linearly separable if and only if their convex hulls are disjoint.

2 Separability of Two Clusters in the Diameter Case

Our results about the case where the function W is the diameter are based on the following theorem which shows how we can separate two intersecting clusters by a line without increasing the diameters.

Theorem 1 *Let A and B be two sets of points in the plane with diameters d_A and d_B. Then there are two linearly separable sets A' and B' with diameters $d_{A'}$ and $d_{B'}$ such that $d_{A'} \leq d_A$, $d_{B'} \leq d_B$ and $A' \cup B' = A \cup B$.*

The proof consists of taking $A \cup B$ and finding a line which separates this set into two new clusters A' and B' which fulfill the claimed property. The details, which are hairier than one might expect, are contained in the full paper [5].

A weaker version of Theorem 1 with an erroneous proof was stated in [1].

Lemma 2 *In the construction in Theorem 1,*

$$\mathit{perimeter}(A) + \mathit{perimeter}(B) \geq \mathit{perimeter}(A') + \mathit{perimeter}(B')$$

holds. If $\mathit{conv}(A) \cap \mathit{conv}(B) \neq \emptyset$, then the inequality is strict.

3 The Polynomial Time Result

In this section, we extend Theorem 1 of the preceding section to more than two clusters, and we also show the corresponding result for the case of radii. Finally, we will apply these separability theorems to obtain a polynomial algorithm.

Theorem 3 *Consider the optimal k-clustering problem for the diameter with a monotone increasing function $\mathcal{F}$. For every point set P in the plane, there is an optimal k-clustering such that each pair of clusters is linearly separable.*

Proof. This is an easy consequence of Theorem 1 and Lemma 2. □

So far, we have only dealt with the diameter as the quality measure of a cluster. For the radius, an analog of Theorem 3 can be shown directly.

Theorem 4 *Consider the optimal k-clustering problem for the radius with a monotone increasing function $\mathcal{F}$. For every point set P in the plane, there is an optimal k-clustering such that each pair of clusters is linearly separable.*

Proof. The power diagram of the cluster circles (cf. Aurenhammer [2]; Imai, Iri, and Murota [12]; or Edelsbrunner [7], section 13.6) is a polygonal dissection of the plane, which can easily be used to show the theorem. □

For the diameter, such a decomposition into convex regions like the power diagram need not necessarily exist. However, by a result of Edelsbrunner, Robison, and Shen [8, Lemmas 1 and 2], we can enlarge the non-intersecting convex hulls of the clusters until they touch while still keeping them convex. This will yield a set of convex polygonal regions which does not necessarily cover the whole plane but which is otherwise completely sufficient for our purposes.

Theorem 5 *For any fixed k, the geometric k-clustering problem for the diameter or for the radius with respect to some monotone increasing function $\mathcal{F}$ is solvable in $O(n^{6k})$ time.*

Proof. The dual of the dissection mentioned above is a plane graph with at most $3k-6$ edges. For each of these edges, we have to specify one of $O(n^2)$ possible dividing lines. We have to do this for all planar graphs on k vertices, and in this way, we generate $O(n^{6k-12})$ possible candidates for the optimal solution. We simple have to compare them and take the best. □

References

[1] Tetsuo Asano, B. Bhattacharya, M. Keil, and F. Yao, Clustering algorithms based on minimum and maximum spanning trees, *Proc. Fourth Ann. Symp. Computational Geometry*, Urbana-Champaign, 1988, pp. 252–257, Association for Computing Machinery 1988.

[2] F. Aurenhammer, Power diagrams: properties, algorithms, and applications, *SIAM J. Computing* **16** (1987), 78–96.

[3] E. Boros and P. L. Hammer, On clustering problems with connected optima in Euclidean spaces, *Discrete Mathematics* **75** (1989), 81–88.

[4] H. H. Bock, *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Göttingen 1974.

[5] V. Capoyleas, G. Rote, and G. Woeginger, Geometric clusterings, submitted to *J. Algorithms* (1990).

[6] Z. Drezner, The p-centre problem — heuristic and optimal algorithms, *J. Operational Research Society* **35** (1984), 741–748.

[7] H. Edelsbrunner, *Algorithms in Combinatorial Geometry*, Springer-Verlag, 1987.

[8] H. Edelsbrunner, A. D. Robison, and X. Shen, Covering convex sets with non-overlapping polygons, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Report UIUCDCS-R-87-1364, 1987; to appear in *Geometriae Dedicata* (1990).

[9] P. van Emde Boas, Another NP-complete covering problem, unpublished manuscript, 1982.

[10] J. Hershberger and S. Suri, Finding tailored partitions, *Proc. Fifth Ann. Symp. Computational Geometry*, Saarbrücken, West Germany, June 1989, pp. 255–265, Association for Computing Machinery 1989.

[11] J. Hershberger and S. Suri, Finding tailored partitions, to appear in *J. Algorithms* (1991).

[12] H. Imai, M. Iri, and K. Murota, Voronoi diagrams in the Laguerre metric and its applications, *SIAM J. Computing* **14** (1985), 93–105.

[13] D. S. Johnson, The NP-completeness column: an ongoing guide, *Journal of Algorithms* **3** (1982), 182–195.

[14] N. Megiddo and K. J. Supowit, On the complexity of some common geometric location problems, *SIAM J. Computing* **13** (1984), 182–196.

[15] C. Monma and S. Suri, Partitioning points and graphs to minimize the maximum or the sum of diameters, to appear in: *The Theory and Applications of Graphs*, Proc. Sixth Int. Conf. Theory and Appl. of Graphs, Kalamazoo, Michigan, May 1988, ed. G. Chartrand et al.; Wiley, 1990 or 1991.

[16] K. J. Supowit, Topics in Computational Geometry, Ph. D. Thesis, 1981, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Report UIUCDCS-R-81-1062.