Theoretical Computer Science 629 (2016) 80–95

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Exact algorithms for size constrained 2-clustering in the plane

Jianyi Lin *, Alberto Bertoni, Massimiliano Goldwurm

Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39/41, 20135 Milano, Italy

ARTICLE INFO

Article history:
Received 30 January 2015
Received in revised form 2 September 2015
Accepted 1 October 2015
Available online 19 October 2015

Keywords:
Algorithms for clustering
Cluster size constraints
Convex hull
k-Set
Euclidean norm
Manhattan norm

ABSTRACT

We study the problem of determining an optimal bipartition $\{A, B\}$ of a set X of n points in $\mathbb{R}^2$, under the size constraints $|A| = k$ and $|B| = n - k$, that minimizes the dispersion of points around their centroid in A and B, both in the cases of Euclidean and Manhattan norms. Under the Euclidean norm, we show that the problem can be solved in $O(n\sqrt[3]{k}\log^2 n)$ time by using known properties on k-sets and convex hulls; moreover, the solutions for all $k = 1, 2, \ldots, \lfloor n/2 \rfloor$ can be computed in $O(n^2 \log n)$ time. In the case of Manhattan norm, we present an algorithm working in $O(n^2 \log n)$ time, which uses an extended version of red–black trees to maintain a bipartition of a planar point set. Also in this case we provide a full version of the algorithm yielding the solutions for all size constraints k. All these procedures work in $O(n)$ space and rely on separation results of the clusters of optimal solutions.

1. Introduction

The Clustering Problem we consider in this work consists in finding a partition of a set X of n points into m subsets (called clusters) that minimizes the total dispersion of points around the centroid in every subset. This is a fundamental problem in many research areas like data mining, image analysis, pattern recognition and bioinformatics [24]. Clustering is a classical method in unsupervised machine learning, frequently used in statistical data analysis [7,15].

The computational complexity of the problem depends on a variety of parameters: the dimension d of the point space (usually $\mathbb{R}^d$), the distance or semi-distance used to measure the dispersion of points, the number m of clusters (which may be arbitrary, as part of the instance, or fixed in advance), the size of the clusters and possibly other constraints [3,27,28]. In most cases the problem is difficult. For instance, assuming the squared Euclidean semi-distance, when the dimension d is arbitrary the general Clustering Problem is NP-hard even if the number m of clusters is fixed to 2 [1,10]. The same occurs if m is arbitrary and the dimension is $d = 2$ [20]. The problem is solvable in polynomial time when fixing both m and d [16]. Moreover there exists a well-known, usually fast, heuristic for finding an approximate solution, called k-Means [19], which however requires exponential time in the worst case [25].

Thus, a natural goal is the analysis of the problem under input hypotheses that allow us to design polynomial time algorithms. In this work we consider the Clustering Problem in $\mathbb{R}^2$ when the number of clusters is $m = 2$ and their size is given by the instance as useful background information [3], both assuming the Euclidean and Manhattan norms. We call this problem *Size Constrained 2-Clustering* in $\mathbb{R}^2$ (2-SCC-2 for short).

* Corresponding author.
E-mail address: jianyi.lin@unimi.it (J. Lin).

http://dx.doi.org/10.1016/j.tcs.2015.10.005
0304-3975/

The relevance of the 2-clustering problems is due to the wide spread of hierarchical clustering techniques that repeatedly apply the 2-clustering as the key step. The 2-clustering problem with cluster size constraints has been already studied in [18,6], where it is shown that in dimension 1 the problem is solvable in polynomial time for every norm ℓ_p with integer $p \geq 1$, while there is some evidence that the same result does not hold for non-integer p. It is also known that for arbitrary dimension d the same problem is NP-hard even assuming equal sizes of the two clusters [6].

Under the Euclidean norm, an instance of 2-SCC-2 is given by a set $X \subset \mathbb{R}^2$ of n points in general position together with an integer k such that $1 \leq k \leq n/2$, while the solution is a bipartition $\{A, B\}$ of X such that $|A| = k$ minimizing the total weight $W(A) + W(B)$. Here the weight $W(A)$ (respectively, $W(B)$) is the sum of the squares of the ℓ_2-distances of all points $a \in A$ (resp. $b \in B$) from the centroid of A (resp. B). Recall that the unconstrained version of the same problem, with an arbitrary number of clusters, is NP-hard [20].

In this case we show two results. First, we describe an algorithm that solves 2-SCC-2 in $O(n\sqrt[3]{k}\log^2 n)$ time. This is obtained by using known results of computational geometry concerning dynamic data structures for convex hulls [23,22] and the enumeration of k-sets in $\mathbb{R}^2$ [13,11].[1] Then, we present an algorithm for the full version of the problem, i.e. a procedure yielding a solution for all $k = 1, 2, \ldots, \lfloor n/2 \rfloor$, that works in $O(n^2 \log n)$ time. This procedure is based on a similar approach used in [14] to solve the unconstrained version of the problem.

We also study the 2-SCC-2 problem under the Manhattan distance (ℓ_1 norm). In this case we first present a full algorithm computing a solution for all size constraints in $O(n^3 \log n)$ time. Then, we describe another procedure yielding a solution for a single (arbitrary) size constraint k, that works in $O(n^2 \log n)$ time.

All these algorithms work in $O(n)$ space and are based on separation results of the clusters of optimal solutions of the problem, under ℓ_1 and ℓ_2 norms, which intuitively extend to the bidimensional case the so-called String Property on the real line [21,26].

We note that there exists a wide literature on clustering algorithms based on different optimality criteria to choose the best partition. For instance, in [2] efficient procedures are given that find solutions in the plane that either minimize the cluster diameters or maximize the minimum intercluster distance.

2. Problem definition

In this section we define the 2-clustering problem in the plane and fix our notation. For any point $a \in \mathbb{R}^2$, we denote by a_x and a_y the abscissa and the ordinate of a, respectively. Moreover, for every real $p \geq 1$, let $\|a\|_p$ be the usual ℓ_p norm of point a, i.e. $\|a\|_p = (|a_x|^p + |a_y|^p)^{1/p}$. Clearly, $\|a\|_2$ and $\|a\|_1$ are the Euclidean and the Manhattan norm of a, respectively.

To define our problem formally, let us consider a finite set $X \subset \mathbb{R}^2$: we say that X is in *general position* if it does not contain three collinear points. A *cluster* of $X \subset \mathbb{R}^2$ is a non-empty subset $A \subset X$, while the pair $\{A, \bar{A}\}$ is a *2-clustering* of X, where $\bar{A} = X \setminus A$. Assuming the ℓ_p norm, the *centroid* and the *weight* of A are the values $C_A \in \mathbb{R}^2$ and $W(A) \in \mathbb{R}_+$ defined, respectively, by

$$C_A = \operatorname*{argmin}_{\mu \in \mathbb{R}^2} \sum_{a \in A} \|a - \mu\|_p^p, \tag{1}$$

$$W(A) = \sum_{a \in A} \|a - C_A\|_p^p. \tag{2}$$

Then, the Size Constrained 2-Clustering Problem in $\mathbb{R}^2$ (2-SCC-2, for short) is defined as follows:

2-SCC-2 problem under ℓ_p norm
Given a set $X \subset \mathbb{R}^2$ of cardinality n in general position and an integer k, $1 \leq k \leq n/2$, find a 2-clustering $\{A, \bar{A}\}$ of X, with $|A| = k$, that minimizes the weight $W(A, \bar{A}) = W(A) + W(\bar{A})$.

In the case $p > 1$ the solutions of the problem above satisfy the following property, proved in [6] in a more general dimension.

Theorem 1 *(Separation Result in ℓ_p, for $p > 1$). For any fixed $p > 1$, let $\{A, B\}$ be an optimal solution of the 2-SCC-2 problem under ℓ_p norm for an instance $X \subset \mathbb{R}^2$ with size constraint $|A| = k$. Then, we have $C_A \neq C_B$ and there exists $c \in \mathbb{R}$ such that for every $u \in X$*

$$u \in A \text{ implies } \|u - C_A\|_p^p - \|u - C_B\|_p^p < c$$

$$u \in B \text{ implies } \|u - C_A\|_p^p - \|u - C_B\|_p^p > c.$$

The previous theorem states that (for variable z ranging over $\mathbb{R}^2$) equation

[1] We notice that such a time complexity can be reduced of a logarithmic factor if one uses the dynamic data structure for convex hull proposed in [8], whose journal version however, to our knowledge, has not appeared yet.