Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the complexity of clustering with relaxed size constraints in fixed dimension ☆

Massimiliano Goldwurm [a], Jianyi Lin [b,*], Francesco Saccà [c]

[a] Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy
[b] Department of Mathematics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[c] Dipartimento di Informatica, Università degli Studi di Milano, Milano, Italy

ARTICLE INFO

Article history:
Received 15 November 2016
Received in revised form 13 April 2017
Accepted 25 April 2017
Available online xxxx

Keywords:
Geometric clustering problems
Cluster size constraints
Computational complexity
Constrained k-Means

ABSTRACT

We study the computational complexity of the problem of computing an optimal clustering $\{A_1, A_2, ..., A_k\}$ of a set of points assuming that every cluster size $|A_i|$ belongs to a given set M of positive integers. We present a polynomial time algorithm for solving the problem in dimension 1, i.e. when the points are simply rational values, for an arbitrary set M of size constraints, which extends to the ℓ_1-norm an analogous procedure known for the Euclidean norm. Moreover, we prove that in dimension 2, assuming Euclidean norm, the problem is (strongly) NP-hard with size constraints $M = \{2, 4\}$. This result is extended also to the size constraints $M = \{2, 3\}$ both in the case of Euclidean and ℓ_1-norm.

1. Introduction

In the area of unsupervised machine learning and statistical data analysis the clustering methods play an important role with applications in pattern recognition, bioinformatics, signal and image processing, medical diagnostics. Clustering consists in grouping a set of objects into subsets, called clusters, that are maximally homogeneous with respect to a suitable criterion for evaluating the similarity of objects [5,8]. Partitional or hard clustering requires the subsets to be disjoint and non-empty, and in the usual geometric setting the similarity between objects is measured by distance between points representing the objects [17].

A classical clustering problem is the so-called Euclidean Minimum-Sum-of-Squares [1], Variance-based [11] or k-Means clustering problem: given a finite point set $X \subset \mathbb{R}^d$, find a k-partition $\{A_1, ..., A_k\}$ of X minimizing the sum of weights $W(A_1, ..., A_k) = \sum_i W(A_i) = \sum_i \sum_{x \in A_i} \|x - \mu(A_i)\|^2$ of all clusters, where $\mu(A_i)$ is the sample mean of A_i and $\|\cdot\|$ is the Euclidean norm (also called ℓ_2-norm). In most cases such a partitional clustering problem is difficult: when d is part of the instance the problem is NP-hard even if the number of clusters is fixed to $k = 2$ [1,7]; the same occurs for arbitrary k with fixed dimension $d = 2$ [18]. Nonetheless, a well-known heuristic for this problem is Lloyd's algorithm [15], also named k-Means Algorithm, which is not guaranteed to converge to the global optimum. This algorithm is usually very fast, but may require exponential time in the worst case [25].

Often one has some a-priori information on the clusters, that can be incorporated into traditional clustering techniques to increase the clustering performance [2]. Problems that include background information are so-called constrained clus-

☆ A preliminary version of this work was presented at the 11th AAIM International Conference [10].

* Corresponding author.
E-mail address: jianyi.lin@kustar.ac.ae (J. Lin).

http://dx.doi.org/10.1016/j.tcs.2017.04.017
0304-3975/

tering and can be divided into two classes based on the constraints: instance-level constraints typically define pairs of elements that must be (must-link) or cannot be (cannot-link) in the same cluster [28], and cluster-level constraints prescribe characteristics of each cluster, such as cluster diameter or cluster size [6,24]. In [29] cluster size constraints are used for improving clustering accuracy, for instance allowing one to avoid extremely small or large clusters in standard cluster analysis. In the *size constrained clustering* (SCC) problem, assuming an ℓ_p-norm with integer $p \geq 1$, typically one is given a finite set $X \subset \mathbb{R}^d$ of n points and k positive integers $m_1, ..., m_k$ such that $\sum_i m_i = n$, and searches for a partition $\{A_1, ..., A_k\}$ of X, with $|A_1| = m_1, ..., |A_k| = m_k$, that minimizes the objective function $W(A_1, ..., A_k) = \sum_{i=1}^{k} \sum_{x \in A_i} \|x - c_i\|_p^p$, where each $c_i = \operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{x \in A_i} \|x - c\|_p^p$ is the ℓ_p-centroid of A_i.

For arbitrary $k \in \mathbb{N}$, the SCC problem is NP-hard also in dimension $d = 1$, for any (fixed) ℓ_p-norm, $p \geq 1$ [3]. The same negative result holds for arbitrary $d \in \mathbb{N}$ when the number of clusters is fixed to $k = 2$, for every ℓ_p-norm with $p > 1$ [3]. On the contrary, in the case $d = 2 = k$ the SCC problem is solvable in $O(n^2 \log n)$ time assuming Manhattan norm (ℓ_1) and in $O(n \sqrt[3]{m} \log^2 n)$ time with Euclidean norm (ℓ_2) [14], where m is the size of one of the two clusters.

In this work we study a *relaxed version* of the SCC problem, where the size of each cluster belongs to a given set M of integers, rather than being fixed by the instance of the problem. We show that in dimension $d = 1$, assuming the ℓ_1-norm, for an arbitrary (finite) $M \subset \mathbb{N}$ the solution can be obtained in $O(n(ks + n))$ time, where n is the number of input points, $s = |M|$ and k is the number of clusters. This extends an analogous algorithm proposed for the problem assuming the Euclidean norm [4]. It further emphasizes the difference w.r.t. SCC problem, which is NP-hard in dimension 1 for every ℓ_p-norm, showing that relaxing the size constraints is a key condition to guarantee a solution computable in polynomial time. We recall that clustering problems in dimension 1 have already been studied in the literature, especially in connections with problems of computational biology [4,22]. In particular in [4] an algorithm for solving clustering problem in dimension 1 (with size constraints) is applied for determining promoter regions in genomic sequences, which can be defined intuitively as positions in DNA molecules where the occurrence of certain patterns of nucleotides allows the cell to activate or silence the genes (hence regulating gene expression).

Other results of the present contribution concern the relaxed size constrained clustering problem in dimension $d = 2$. In this case, assuming ℓ_2-norm and fixing $M = \{2, 4\}$, we prove that the problem is strongly NP-hard and it is also easy to see that it does not admit FPTAS unless $P = NP$. Moreover, we prove the same results for the case $M = \{2, 3\}$ both with ℓ_2 and ℓ_1 norm. This also implies that the general relaxed size constrained problem, where M is part of the instance, is strongly NP-hard on the plane both assuming ℓ_2 and ℓ_1-norm.

The introduction of relaxed size constraints is motivated by all applications where one wants to bound the cluster size to certain values, possibly avoiding too large or too small clusters, up to the balanced case where all clusters have almost the same size. Situations of this type are rather common in several contexts [2,16,29].

2. Problem definition

In this section we define the problem and fix our notation. Given two positive integers d and p, for every point $a = (a_1, ..., a_d) \in \mathbb{R}^d$, we denote by $\|a\|_p$ the ℓ_p-norm of a, i.e. $\|a\|_p = (\sum_1^d |a_i|^p)^{1/p}$. Clearly, $\|a\|_2$ and $\|a\|_1$ are the Euclidean and the Manhattan (or Taxicab) norm of a, respectively.

Given a finite set $X \subset \mathbb{R}^d$, a *cluster* of X is a non-empty subset $A \subset X$, while a *clustering* is a partition $\{A_1, ..., A_k\}$ of X in k clusters for some k. Assuming the ℓ_p norm, the *centroid* and the *weight* of a cluster A are the values $C_A \in \mathbb{R}^d$ and $W_p(A) \in \mathbb{R}_+$ defined, respectively, by

$$C_A = \operatorname*{argmin}_{c \in \mathbb{R}^d} \sum_{a \in A} \|a - c\|_p^p, \qquad W_p(A) = \sum_{a \in A} \|a - C_A\|_p^p$$

The *weight* of a clustering $\{A_1, ..., A_k\}$ is $W_p(A_1, ..., A_k) = \sum_1^k W_p(A_i)$. We recall that, in case of ℓ_2-norm, the weight of a cluster A can be computed by relation

$$W_2(A) = \frac{1}{|A|} \sum_{(*)} \|a - b\|_2^2 \tag{1}$$

where the sum is extended to all unordered pairs $\{a, b\}$ of distinct elements in A. Moreover, given a set $\mathcal{M} \subset \mathbb{N}$, any clustering $\{A_1, ..., A_k\}$ such that $|A_i| \in \mathcal{M}$ for every $i = 1, \ldots, k$, is called *$\mathcal{M}$-clustering*.

RSC-d Problem (with ℓ_p-norm): Relaxed Size Constrained Clustering in $\mathbb{R}^d$
Given a set $X \subset \mathbb{Q}^d$ of n points, an integer k such that $1 < k < n$ and a finite set $\mathcal{M}$ of positive integers, find an $\mathcal{M}$-clustering $\{A_1, ..., A_k\}$ of X that minimizes $W_p(A_1, ..., A_k)$.[1]

When $\mathcal{M}$ is not included in the instance, but fixed in advance, we call the problem **$\mathcal{M}$-RSC-d** (with ℓ_p-norm). In this work we study these problems in dimension $d = 1, 2$ assuming ℓ_1 and ℓ_2-norm.

[1] If X does not admit a $\mathcal{M}$-clustering then symbol $\perp$ is returned.

Please cite this article in press as: M. Goldwurm et al., On the complexity of clustering with relaxed size constraints in fixed dimension, Theoret. Comput. Sci. (2017), http://dx.doi.org/10.1016/j.tcs.2017.04.017