ISSN 1995-4239, Numerical Analysis and Applications, 2019, Vol. 12, No. 2, pp. 105–115. © Pleiades Publishing, Ltd., 2019.
Russian Text © The Author(s), 2019, published in Sibirskii Zhurnal Vychislitel'noi Matematiki, 2019, Vol. 22, No. 2, pp. 121–136.

Exact Algorithms of Search for a Cluster of the Largest Size in Two Integer 2-Clustering Problems

A. V. Kel'manov[1,2*], A. V. Panasenko[1,2], and V. I. Khandeev[1,2***]**

[1] *Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences, pr. Akad. Koptyuga 4, Novosibirsk, 630090 Russia*

[2] *Novosibirsk State University, ul. Pirogova 1, Novosibirsk, 630090 Russia*

Received May 15, 2018; in final form, June 26, 2018; accepted January 21, 2019

Abstract—We consider two related discrete optimization problems of searching for a subset in a finite set of points in Euclidean space. Both problems are induced by versions of a fundamental problem in data analysis, namely, that of selecting a subset of similar elements in a set of objects. In each problem, given an input set and a positive real number, it is required to find a cluster (i.e., a subset) of the largest size under constraints on a quadratic clusterization function. The points in the input set, which are outside the sought-for subset, are treated as a second (complementary) cluster. In the first problem, the function under the constraint is the sum over both clusters of the intracluster sums of the squared distances between the elements of the clusters and their centers. The center of the first (i.e., the sought-for) cluster is unknown and determined as a centroid, while the center of the second one is fixed at a given point in Euclidean space (without loss of generality, at the origin of coordinates). In the second problem, the function under the constraint is the sum over both clusters of the weighted intracluster sums of the squared distances between the elements of the clusters and their centers. As in the first problem, the center of the first cluster is unknown and determined as a centroid, while the center of the second one is fixed at the origin of coordinates. In this paper, we show that both problems are strongly NP-hard. Also, we present exact algorithms for the problems in which the input points have integer components. If the space dimension is bounded by some constant, the algorithms are pseudopolynomial.

DOI: 10.1134/S1995423919020010

Keywords: *Euclidean space, 2-clustering, largest subset, NP-hardness, exact algorithm, pseudopolynomial-time solvability.*

INTRODUCTION

The subject of this paper is to study two closely related optimization problems of choosing (searching for) a subset of the largest cardinality (size) in a finite set of points in Euclidean space. These simulate one of the key problems of data analysis: choosing a subset of similar elements in a finite set of objects. The goal of the paper is to analyze the computational complexity of the problems and construct algorithms to efficiently solve these problems with guaranteed estimates of quality (accuracy and time complexity).

This study is stimulated, on the one hand, by the fact that these problems have been poorly studied theoretically. No results on their computational complexity have been published so far, and there are no rigorously justified algorithmic solutions. On the other hand, these problems are very important for some applications (see the next section).

The paper is organized as follows. Section 1 provides formulations of the problems, their interpretations, some closely related problems, their distinctive features, and the available algorithmic results. It also presents the results obtained in the present paper. In Section 2, the computational complexity of the problems is analyzed. Section 3 provides some auxiliary results to justify the properties of the algorithms. Section 4 contains a description of the algorithms and justification of some of their qualities (accuracy and time complexity).

[*] E-mail: kelm@math.nsc.ru

[**] E-mail: a.v.panasenko@math.nsc.ru

[***] E-mail: khandeev@math.nsc.ru

1. PROBLEM FORMULATION AND INTERPRETATION; SIMILAR PROBLEMS; OLD AND NEW RESULTS

Below $\mathbb{R}$ is the set of real numbers, d is the space dimension, $\|\cdot\|$ is the Euclidean norm, and $\langle\cdot,\cdot\rangle$ is the inner product. The *centroid* (geometrical center) of a nonempty finite set (cluster) $\mathcal{Y} \subset \mathbb{R}^d$ is the point from $\mathbb{R}^d$ that is equal to the arithmetic mean of the elements of this set; the *center* of a cluster is an arbitrary fixed point $x \in \mathbb{R}^d$ relative to which the squares of the distances from this cluster are added together.

The statement of the first problem being considered is close (but not equivalent) to that of the *minimum sum-of-squares clustering* (MSSC) problem, which is well known since the last century under a different name of k-means (see [1–6]). The computational complexity of this problem was analyzed in [7–10]. It was proved in [7] that even the simplest (basic) two-cluster version of this problem, 2-MSSC (or 2-means), is strongly NP-hard.

Recall that in the 2-MSSC method it is necessary to find a 2-partition of a finite set $\mathcal{Y} \subset \mathbb{R}^d$ minimizing the sum

$$\sum_{y\in\mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y - \overline{y}(\mathcal{Y}\setminus\mathcal{C})\|^2, \tag{1}$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum\limits_{y\in\mathcal{C}} y$ and $\overline{y}(\mathcal{Y}\setminus\mathcal{C}) = \frac{1}{|\mathcal{Y}\setminus\mathcal{C}|}\sum\limits_{y\in\mathcal{Y}\setminus\mathcal{C}} y$ are the centroids of nonempty nonintersecting subsets $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$, respectively.

The statement of the second problem is close (but not equivalent) to that of the well-known *quadratic min-sum all-pairs 2-clustering* problem (see [11–18]), in which it is necessary to find a 2-partition of a finite set $\mathcal{Y} \subset \mathbb{R}^d$ minimizing the sum

$$\sum_{y\in\mathcal{C}}\sum_{x\in\mathcal{C}} \|y - x\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\sum_{x\in\mathcal{Y}\setminus\mathcal{C}} \|y - x\|^2.$$

The *quadratic min-sum all-pairs 2-clustering* problem is equivalent to the *cardinality-weighted minimum sum-of-squares 2-clustering* problem or the *cardinality-weighted* 2-MSSC problem, in which it is necessary to find a 2-partition of a finite set $\mathcal{Y} \subset \mathbb{R}^d$ minimizing the sum

$$|\mathcal{C}|\sum_{y\in\mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y}\setminus\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y - \overline{y}(\mathcal{Y}\setminus\mathcal{C})\|^2. \tag{2}$$

Here, as in the 2-MSSC problem, $\overline{y}(\mathcal{C})$ and $\overline{y}(\mathcal{Y}\setminus\mathcal{C})$ are the centroids of nonempty nonintersecting subsets $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$, respectively. The equivalence of these problems follows from the fact that the well-known equality $2|\mathcal{Y}|\sum\limits_{y\in\mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2 = \sum\limits_{y\in\mathcal{Y}}\sum\limits_{x\in\mathcal{Y}} \|y - x\|^2$ is valid for any nonempty finite set $\mathcal{Y} \subset \mathbb{R}^d$. This equality relates the cardinality-weighted total quadratic scatter of points from the set $\mathcal{Y}$ with respect to its centroid $\overline{y}(\mathcal{Y})$ to the sum of squares of the pairwise distances between the elements of this set. A hypothesis [12] that such problems are hard to solve was proved in [19, 20], in which a proof of strong NP-hardness of the quadratic Euclidean max-cut problem was also given.

Both intracluster sums in the objective functions (1) and (2) are the total quadratic scatters of points of the clusters with respect to their centroids. That is, in these problems both centers are unknown and defined as centroids. In each of the problems formulated below, one of the intracluster sums includes the total quadratic scatter of points of the cluster with respect to a fixed (given) point $x \in \mathbb{R}^d$. Without the loss of generality, this point is considered to be the origin of coordinates. The center of the other cluster is assumed to be equal to its centroid. In other words, only one center is unknown in these problems, which distinguishes these problems from those mentioned above.

The 2-partition problems with one unknown (and one given) center can be formulated as follows:

Problem 1 (2-MSSC problem *with a given center*). *Given*: an N-element set of $\mathcal{Y}$ points in Euclidean space of dimension d. *Find*: a 2-partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min .$$

In this problem, the sum of the intracluster sums of the squared distances between the elements of the clusters and their centers is minimized over both clusters. The center of the cluster $\mathcal{C}$ is unknown and defined as a centroid, and the center of the cluster $\mathcal{Y} \setminus \mathcal{C}$ is fixed at the origin of coordinates. In the 2-MSSC problem, in contrast to Problem 1, both centers are unknown and defined as centroids. Strong NP-hardness of Problem 1 was proved in [21, 22].

Problem 2 (*cardinality-weighted* 2-*MSSC problem with a given center*). *Given*: an N-element set of $\mathcal{Y}$ points in Euclidean space of dimension d. *Find*: a 2-partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$G(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min .$$

In this problem, the sum of the cardinality-weighted intracluster sums of the squared distances between the elements of the clusters and their centers is minimized over both clusters. As in Problem 1, the center of the cluster $\mathcal{C}$ is unknown and defined as a centroid, and the center of the cluster $\mathcal{Y} \setminus \mathcal{C}$ is fixed at the origin of coordinates. Strong NP-hardness of Problem 2 was established in [23, 24].

In the present paper, we investigate the following two problems closely related to Problems 1 and 2:

Problem 3. *Given*: an N-element set of $\mathcal{Y}$ points in Euclidean space of dimension d and a number $\alpha \in (0, 1)$. *Find*: a subset $\mathcal{C} \subset \mathcal{Y}$ of the largest size such that

$$F(\mathcal{C}) \le \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{3}$$

In this problem, find a cluster $\mathcal{C}$ of the largest size under a constraint on the objective function $F(\mathcal{C})$ of Problem 1. This constraint is defined by the right-hand side of inequality (3), that is, by a fraction of the total quadratic scatter of the points of the input set $\mathcal{Y}$ about its centroid $\overline{y}(\mathcal{Y})$.

Problem 4. *Given*: an N-element set of $\mathcal{Y}$ points in Euclidean space of dimension d and a number $\alpha \in (0, 1)$. *Find*: a subset $\mathcal{C} \subset \mathcal{Y}$ of the largest size such that

$$G(\mathcal{C}) \le \alpha N \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{4}$$

In this problem, find a cluster $\mathcal{C}$ of the largest size under a constraint on the objective function $G(\mathcal{C})$ of Problem 2. This constraint is defined by the right-hand side of inequality (4), that is, by a fraction of the total cardinality-weighted scatter of the points of the input set $\mathcal{Y}$ about its centroid $\overline{y}(\mathcal{Y})$.

The similarity of Problems 1, 2 to Problems 3, 4, respectively, and the difference between them are reflected in the following remark. In Problems 3 and 4 $F(\mathcal{C})$ and $G(\mathcal{C})$ are not objective functions. They only define some constraints on these problems. If $\mathcal{C}^*$ is an optimal solution to Problem 1 (or 2), $F(\mathcal{C}^*) \le F(\mathcal{C})$ in Problem 1 (or $G(\mathcal{C}^*) \le G(\mathcal{C})$ in Problem 2) for any $\mathcal{C} \subset \mathcal{Y}$. Therefore, in Problems 3 and 4, inequalities (3) and (4) define some subsets of the admissible solutions to Problems 1 and 2. This means that in each of Problems 3 and 4 it is necessary to find a cluster of the largest size in the subset of admissible solutions to Problems 1 and 2, respectively.

All of the above-formulated extremal problems can be treated as problems of approximation, combinatorial geometry, graph theory, and statistics. They can be applied to problems in data science, data mining, pattern recognition, and machine learning. In these applications and research areas clusterization algorithms are key tools for solving problems of computer assisted data analysis (see,

for instance, [2–6, 25–31] and the references therein). The following single argument is sufficient to demonstrate the importance of the problems being considered: It is well known that problems of data mining and classical problems of mathematical statistics are closely related: Their major goal is to determine the structural properties of some sets (data or samples). In classical statistics the samples being analyzed are homogeneous, whereas in data mining the sample (experimental) data are inhomogeneous.

A well-known statistical problem is verifying the hypothesis that the mean of a sample coincides (or does not coincide) with a given value. There exist several classical criteria to solve this problem.

How to handle a situation when the sample is inhomogeneous, consists of elements of two distributions, and no relation between the sample elements and the distribution is known? This situation (with inhomogeneous sample data) is typical for data mining and, in particular, for big data problems. It is clear that using the classical statistical methods requires solving the problem of partitioning the data into homogeneous sets (samples). The above extremal problems simulate only few poorly studied problems of this kind. The experience of studying extremal problems induced by problems similar to the above ones has shown that most of them are difficult to solve. The questions of hardness of the induced problems and the possible algorithms to approximate them are important mathematical problems.

The hardness of Problems 3 and 4 has not yet been determined, and no algorithmic results are available by now. Below we present some results for closely related problems, namely, for Problems 1 and 2.

Recall some of the results published for Problem 1. These are results obtained for a version of Problem 1 with given cardinalities of the clusters. This version is called the 2-*MSSC problem with a given center and cluster cardinalities* (Problem 5 in Section 4). Strong NP-hardness of this version of the problem follows directly from the strong NP-hardness of Problem 1 (without constraints on the cardinalities of the clusters) proved in [21, 22]. However, the fact of NP-hardness was first established in [33–35] before the results obtained in [21, 22].

In what follows, when describing the existing algorithmic solutions, M is the cardinality of the cluster with an unknown center given as the input of the problem, and D is the maximum absolute value of the coordinates of the input set points.

It follows from [36] that the problem can be solved in time $\mathcal{O}(d^2N^{2d})$, which is polynomial if the dimension d of the space is fixed (bounded from above by a constant). A fast exact algorithm of hardness $\mathcal{O}(dN^{d+1})$ was proposed in [37]. An exact algorithm for the problem with integer inputs was proposed in [38]. The time complexity of this algorithm is $\mathcal{O}(dN(2MD+1)^d)$. If the space dimension is fixed, this algorithm is pseudopolynomial.

A 2-approximate polynomial algorithm with hardness $\mathcal{O}(dN^2)$ was justified in [39].

A polynomial-time approximation scheme (PTAS) with hardness $\mathcal{O}(dN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, where ε is relative error, was proposed in [40].

It was established in [41] that if P $\neq$ NP, no fully polynomial-time approximation scheme (FPTAS) exists for this problem. Paper [41] presents an algorithm for finding a $(1+\varepsilon)$-approximate solution in time $\mathcal{O}(dN^2(\sqrt{2q/\varepsilon}+2)^d)$ for a given $\varepsilon \in (0,1)$. If the dimension d of the space is fixed, the hardness of the algorithm is $\mathcal{O}(N^2(1/\varepsilon)^{q/2})$ and it implements a FPTAS scheme. In paper [42], a faster algorithm with time $\mathcal{O}\big(\sqrt{d}N^2\big(\frac{\pi e}{2}\big)^{d/2}\big(\sqrt{2/\varepsilon}+2\big)^d\big)$ was justified. This algorithm implements an FPTAS scheme with hardness $\mathcal{O}(N^2(1/\varepsilon)^{d/2})$, if the dimension d of the space is fixed, and it remains polynomial if $d=\mathcal{O}(\log N)$, that is, if the dimension of the space is a slowly increasing function of the input set cardinality. In this case it implements a PTAS scheme with hardness $\mathcal{O}\Big(N^{C\left(1.05+\log\left(2+\sqrt{2/\varepsilon}\right)\right)}\Big)$, where C is a positive constant.

A randomized algorithm was proposed in [43]. If $M \geq \beta N$, where $\beta \in (0,1)$ is a constant, at given $\varepsilon > 0$ and $\gamma \in (0,1)$ the algorithm finds a $(1+\varepsilon)$-approximate solution to the problem with a probability not less than $1-\gamma$ in time $\mathcal{O}(dN)$. In the same paper, conditions are established at which the algorithm finds a $(1+\varepsilon_N)$-approximate solution to the problem in time $\mathcal{O}(dN^2)$ with a probability not less than $1-\gamma_N$, where $\varepsilon_N \to 0$ and $\gamma_N \to 0$ as $N \to \infty$. Hence, these are conditions at which the algorithm is asymptotically exact. To date this algorithm has record quality, since it can find, with a probabilistic

guarantee, an approximate solution in a time that is linear in N and d and obtain an asymptotically exact solution in a time that is quadratic in N and linear in d.

The above results can be used to solve the main problem with unknown cardinalities. In fact, by exhaustive search of not more than N possible combinations of the cardinalities of two clusters, one can construct a family of N solutions of the problem version with given cardinalities and choose in this family the best solution in terms of the objective function. Of great interest are some approximate algorithms for solving Problem 1 without exhaustive search of the cardinalities, since these algorithms are $\mathcal{O}(N)$ times faster. Such a polynomial approximate algorithm was proposed in [32]. It finds a 2-approximate solution in $\mathcal{O}(dN^2)$ operations. For comparison, an algorithm from paper [39] with an exhaustive search of admissible combinations of the cardinalities finds it in $\mathcal{O}(dN^3)$ operations.

Recall the results available for Problem 2. The problem has been only recently formulated (in [23, 24]). Therefore, there are much less algorithmic solutions for this problem than for Problem 1, which has been studied for a much longer time. In the above-cited papers, in addition to presenting a proof of the solution hardness, it was shown that no FPTAS scheme is available for Problem 2 if P $\neq$ NP.

Most available algorithmic solutions have been obtained by using efficient techniques of constructing algorithms for Problem 1. Some of the algorithms have been obtained for the version of Problem 2 with given cardinalities of the clusters: the *cardinality-weighted 2-MSSC problem with a given center and cluster cardinalities* (Problem 6 in Section 4). The available estimates of quality of these algorithms are essentially the same as those presented above for Problem 1. Therefore, they are only mentioned here but not considered separately.

A 2-approximate algorithm was constructed in [44]. An exact algorithm for the integer version of the problem was proposed in [45]. A $(1+\varepsilon)$-approximate algorithm implementing an FPTAS scheme for a fixed space dimension was considered in [46]. A modification of this algorithm providing faster performance was proposed in [42]. This modification implements a PTAS scheme if the space dimension is a slowly increasing function of the input set cardinality ($d = \mathcal{O}(\log N)$). A randomized algorithm was justified in [47].

In conclusion of the review it may be said that the construction of algorithms for Problems 1 and 2 without exhaustive search through N admissible cardinalities of the sought-for clusters remains a topical issue.

The present paper gives first results on new data clusterization Problems 3 and 4. Specifically, it is found that both problems are strongly NP-hard. Some exact algorithms are justified for these problems if the points of the input set have integer coordinates. If the space dimension is fixed, both algorithms are pseudopolynomial.

2. COMPUTATIONAL COMPLEXITY ANALYSIS

Before analyzing the computational complexity of Problems 3 and 4, note that the right-hand sides of (3) and (4) do not depend on the sought-for cluster $\mathcal{C}$ and are constants for the given input. We set

$$A = \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2, \quad B = \alpha N \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{5}$$

Let us formulate Problems 3 and 4 as verification procedures of their properties.

Problem 3A. *Given*: An N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d, a natural number M, and a real number $A > 0$. *Question*: Does there exist in $\mathcal{Y}$ a subset $\mathcal{C}$ with a cardinality not less than M such that

$$F(\mathcal{C}) \leq A? \tag{6}$$

Problem 4A. *Given*: An N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d, a natural number M, and a real number $B > 0$. *Question*: Does there exist in $\mathcal{Y}$ a subset $\mathcal{C}$ with a cardinality not less than M such that

$$G(\mathcal{C}) \le B? \tag{7}$$

A complexity analysis of these problems is performed in the following

Statement. *Problems* 3A *and* 4A *are strongly NP-complete.*

Proof. It is easy to see from (3) and (4) that both problems, 3A and 4A, belong to the NP-class.

Let us formulate optimization Problems 1 and 2 as verification procedures of their properties.

Problem 1A. *Given*: an N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a real number $A > 0$. *Question*: does there exist in $\mathcal{Y}$ a subset $\mathcal{C}$ such that inequality (6) is valid?

Problem 2A. *Given*: an N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a real number $B > 0$. *Question*: does there exist in $\mathcal{Y}$ a subset $\mathcal{C}$ such that the inequality (7) is valid?

It is easy to see that at $M = 1$ the answer is positive in Problems 3A and 4A if and only if it is positive in Problems 1A and 2A. Therefore, the theorem is valid, since the strongly NP-complete Problems 1A and 2A are special cases (at $M = 1$) of Problems 3A and 4A, respectively. □

It follows from this theorem that optimization Problems 3 and 4 are strongly NP-hard.

3. FUNDAMENTALS OF THE ALGORITHMS

To construct the algorithms for solving Problems 1 and 2 and to analyze their quality, we will need some auxiliary problems, statements, sets, and algorithms.

First, we need the following problems (mentioned in Section 2):

Problem 5 (2-*MSSC with a given center and cluster cardinalities*). *Given*: an N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a natural number M. *Find*: a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M minimizing the function $F(\mathcal{C})$.

Problem 6 (*cardinality-weighted 2-MSSC with a given center and cluster cardinalities*). *Given*: an N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a natural number M. *Find*: a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M minimizing the function $G(\mathcal{C})$.

The computational basis of the algorithms proposed in the present paper are some algorithms for solving these problems. The algorithms for solving Problems 5 and 6 are based on the following two Lemmas, 1 and 2, whose proofs can be found in [39] and [45], respectively.

For an arbitrary point $x \in \mathbb{R}^d$, we set

$$r^x(y) = \langle y, x \rangle, \quad y \in \mathcal{Y}, \tag{8}$$

$$h^x(y) = (2M - N)\,\|y\|^2 - 2M\,\langle y, x \rangle, \quad y \in \mathcal{Y}, \tag{9}$$

and

$$f^x(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \quad \mathcal{C} \subseteq \mathcal{Y}, \tag{10}$$

$$g^x(\mathcal{C}) = M \sum_{y \in \mathcal{C}} \|y - x\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \quad \mathcal{C} \subseteq \mathcal{Y}. \tag{11}$$

Lemma 1. *The minimum of the function* (10) *over all subsets* $\mathcal{C} \subseteq \mathcal{Y}$ *of cardinality* M *is reached on a subset consisting of* M *vectors of the set* $\mathcal{Y}$ *with the greatest values of the function* (8).

Lemma 2. *The minimum of the function* (11) *over all subsets* $\mathcal{C} \subseteq \mathcal{Y}$ *of cardinality* M *is reached on a subset consisting of* M *vectors of the set* $\mathcal{Y}$ *with the smallest values of the function* (9).

In what follows, it is assumed that the points of the set $\mathcal{Y}$ have integer coordinates. That is, we consider special (integer) cases of the problems. Let an auxiliary set be defined as the totality of nodes (points) of a uniform grid with a rational step:

$$\mathcal{D} = \left\{ x \mid (x)^j = \frac{1}{M}(v)^j, \quad (v)^j \in \mathbb{Z}, \, |(v)^j| \leq MD, \quad j = 1, \ldots, d \right\}, \tag{12}$$

where

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1,\ldots,d\}} |(y)^j|, \tag{13}$$

and let $(*)^j$ be the jth coordinate of the point $*$.

Note that

$$|\mathcal{D}| = (2MD + 1)^d \, .$$

Let us present an algorithm for the integer case of Problem 5.

Algorithm $\mathcal{A}_1$.

Input: the set $\mathcal{Y}$ and the natural number M.

Step 1. Find D and construct the nodes of the lattice $\mathcal{D}$ by formulas (13) and (12).

Step 2. For every $x \in \mathcal{D}$, construct a set $\mathcal{C}(x)$ consisting of M points $y \in \mathcal{Y}$ with the largest values of the function (8). Calculate $f^x(\mathcal{C}(x))$ by formula (10).

Step 3. Find the point $x_A = \arg\min_{x \in \mathcal{D}} f^x(\mathcal{C}(x))$ and the corresponding subset $\mathcal{C}(x_A)$. As a solution to the problem, take $\mathcal{C}^M_{A_1} = \mathcal{C}(x_A)$. If there are several solutions, take any of them.

Output: the set $\mathcal{C}^M_{A_1}$.

Remark 1. It was proved in [38] using Lemma 1 that if the coordinates of all points of the input set $\mathcal{Y}$ are integer and lie in the interval $[-D, D]$, algorithm $\mathcal{A}_1$ finds the optimal solution to Problem 5 in time $\mathcal{O}(dN(2MD + 1)^d)$.

Finally, we need an algorithm for finding a solution to the integer Problem 6. The following algorithm differs from algorithm $\mathcal{A}_1$ only in step 2.

Algorithm $\mathcal{A}_2$.

Input: the set $\mathcal{Y}$ and the positive integer number M.

Step 1. Find D and construct the nodes of the lattice $\mathcal{D}$ by formulas (13) and (12).

Step 2. For every $x \in \mathcal{D}$, construct a set $\mathcal{C}(x)$ consisting of M points $y \in \mathcal{Y}$ with the smallest values of the function (98). Calculate $g^x(\mathcal{C}(x))$ by formula (11).

Step 3. Find the point $x_A = \arg\min_{x \in \mathcal{D}} g^x(\mathcal{C}(x))$ and the corresponding subset $\mathcal{C}(x_A)$. As a solution to the problem, take $\mathcal{C}^M_{A_2} = \mathcal{C}(x_A)$. If there are several solutions, take any of them.

Output: the set $\mathcal{C}^M_{A_2}$.

Remark 2. It was proved in [45] using Lemma 1 that if the coordinates of all points of the input set $\mathcal{Y}$ are integer and lie in the interval $[-D, D]$, algorithm $\mathcal{A}_2$ finds the optimal solution to Problem 6 in time $\mathcal{O}(dN(2MD + 1)^d)$.

4. THE ALGORITHMS

The idea of the algorithms being proposed is simple: the grid approach is used to approximate the unknown centroid of the largest sought-for cluster by one of the nodes of a uniform grid with a rational step. For every grid node, on the basis of Lemma 1 (in solving Problem 3) or Lemma 2 (in solving Problem 4) algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ are used to construct a family of admissible solutions—subsets. In the thus constructed family of admissible subsets, a subset of the largest size satisfying the constraint (3) (for Problem 3) or the constraint (4) (for Problem 4) is chosen.

The following algorithm for solving Problem 3 is proposed:

Algorithm $\mathcal{A}_3$.

Input: the set $\mathcal{Y}$ and the number α.

Step 1. Calculate A by formula (5).

Step 2. For every $M = 1, \ldots, N$, using algorithm $\mathcal{A}_1$, find an exact solution $\mathcal{C}_{A_1}^M$ of Problem 5, and calculate the objective function $F(\mathcal{C}_{A_1}^M)$ for this solution.

Step 3. In the family $\{\mathcal{C}_{A_1}^M, M = 1, \ldots, N\}$ of sets obtained in step 2, find a set $\mathcal{C}_{A_1}$ of the largest cardinality for which $F(\mathcal{C}_{A_1}) \le A$.

Output: the set $\mathcal{C}_{A_1}$.

The algorithm for solving Problem 4 is similar; its major difference from algorithm $\mathcal{A}_3$ is in constructing the admissible solution to the problem in step 2.

Algorithm $\mathcal{A}_4$.

Input: the set $\mathcal{Y}$ and the number α.

Step 1. Calculate B by formula (5).

Step 2. For every $M = 1, \ldots, N$, using algorithm $\mathcal{A}_2$, find an exact solution $\mathcal{C}_{A_2}^M$ of Problem 6, and calculate the objective function $G(\mathcal{C}_{A_2}^M)$ for this solution.

Step 3 In the family $\{\mathcal{C}_{A_2}^M, M = 1, \ldots, N\}$ of sets obtained in step 2, find a set $\mathcal{C}_{A_2}$ of the largest cardinality for which $G(\mathcal{C}_{A_2}) \le B$.

Output: the set $\mathcal{C}_{A_2}$.

We have the following

Theorem. *Let the points of the input set $\mathcal{Y}$ have integer coordinates lying in the interval $[-D, D]$. Then algorithms $\mathcal{A}_3$ and $\mathcal{A}_4$ find exact solutions to Problems 3 and 4 in time $\mathcal{O}(dN^2(2ND+1)^d)$.*

Proof. Let us prove that the solution obtained by algorithm $\mathcal{A}_3$ is optimal. Let $\mathcal{C}_1^*$ be the optimal solution to Problem 3, $M_1^* = |\mathcal{C}_1^*|$. Note that algorithm $\mathcal{A}_3$ finds an admissible solution of Problem 5 at $M = M_1^*$. Since $\mathcal{C}_{A_1}^{M^*}$ is the optimal solution to this problem,

$$F(\mathcal{C}_{A_1}^{M^*}) \le F(\mathcal{C}_1^*) \le A.$$

Thus, the set $\mathcal{C}_{A_1}^{M^*}$ was considered in step 3 of the algorithm, and the set $\{\mathcal{C}_{A_1}^M, M = 1, \ldots, N \mid F(\mathcal{C}_{A_1}) \le A\}$ is not empty. From the definition of step 3 we have

$$|\mathcal{C}_{A_1}| \ge |\mathcal{C}_{A_1}^{M^*}| = M_1^* = |\mathcal{C}_1^*|.$$

On the other hand, since $\mathcal{C}_{A_1}$ is an admissible solution of Problem 1, we have $|\mathcal{C}_{A_1}| \le |\mathcal{C}_1^*|$; hence, $|\mathcal{C}_{A_1}| = |\mathcal{C}_1^*|$.

The proof of optimality of the solution obtained by algorithm $\mathcal{A}_4$ is constructed similarly to the proof of optimality of the solution obtained by algorithm $\mathcal{A}_3$.

Let us estimate the time complexity of algorithm $\mathcal{A}_3$. Step 1 is made in $\mathcal{O}(dN)$ operations. The hardest step 2 requires $\mathcal{O}(dN^2(2ND+1)^d)$ operations, since for every $M = 1, \ldots, N$ algorithm $\mathcal{A}_1$ is executed in $\mathcal{O}(dN(2MD+1)^d)$ operations. Finally, step 3 is executed in $\mathcal{O}(N)$ operations. By adding together the computational costs in all steps, we obtain the estimate of the time complexity of algorithm $\mathcal{A}_3$ presented in the theorem. The time complexity of algorithm $\mathcal{A}_4$ is estimated in the same way. □

Remark 3. If the space dimension is fixed, algorithms $\mathcal{A}_3$ and $\mathcal{A}_4$ are pseudopolynomial, since in this case the time of their execution can be estimated as $\mathcal{O}(N^2(ND)^d)$.

CONCLUSIONS

In the present paper, strong NP-hardness has been proved for two optimization problems of clusterization that are closely related and important for some applications. For the first time, algorithms constructed in a similar way have been proposed to solve these problems. These algorithms make it possible to find exact solutions if the coordinates of the input points are integer. If the space dimension is bounded from above by a constant, the algorithms are pseudopolynomial. In other words, pseudopolynomial solvability of integer versions of the problems has been demonstrated if the space dimension is fixed (is not part of the input).

It is clear that the algorithms proposed can be used to solve practical problems of only small dimension. Nevertheless, these algorithms may serve as a starting point for obtaining improved algorithmic solutions.

For further studies it is important to justify the efficient approximate algorithms with guaranteed estimates of quality for some problems that are difficult to solve.

FUNDING

This work was supported by the Russian Science Foundation (project no. 16-11-10041, Problem 1), by the Russian Foundation for Basic Research (projects no. 19-01-00308 and 18-31-00398-mol-a, Problem 2), by the RAS Fundamental Research Program (project no. 0314-2016-0015), and by RF Ministry of Education and Science (under Program 5-10).

REFERENCES

1. MacQueen, J.B., Some Methods for Classification and Analysis of Multivariate Observations, *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: Univ. of California Press, 1967, vol. 1, pp. 281–297.
2. Rao, M., Cluster Analysis and Mathematical Programming, *J. Am. Stat. Assoc.*, 1971, vol. 66, pp. 622–626.
3. Hansen, P., Jaumard, B., and Mladenovich, N., Minimum Sum of Squares Clustering in a Low Dimensional Space, *J. Classific.*, 1998, vol. 15, pp. 37–55.
4. Hansen, P. and Jaumard, B., Cluster Analysis and Mathematical Programming, *Math. Programming*, 1997, vol. 79, pp. 191–215.
5. Fisher, R.A., *Statistical Methods and Scientific Inference*, New York: Hafner, 1956.
6. Jain, A.K., Data Clustering: 50 Years beyond k-Means, *Patt. Recog. Lett.*, 2010, vol. 31, iss. 8, pp. 651–666.
7. Aloise, D., Deshpande, A., Hansen, P., and Popat, P., NP-Hardness of Euclidean Sum-of-Squares Clustering, *Mach. Learning*, 2009, vol. 75, iss. 2, pp. 245–248.
8. Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V., Clustering Large Graphs via the Singular Value Decomposition, *Mach. Learning*, 2004, vol. 56, pp. 9–33.
9. Dolgushev, A.V. and Kel'manov, A.V., On the Algorithmic Complexity of a Problem in Cluster Analysis, *J. Appl. Ind. Math.*, 2011, vol. 5, no. 2, pp. 191–194.
10. Mahajan, M., Nimbhorkar, P., and Varadarajan, K., The Planar k-Means Problem Is NP-Hard, *Theor. Comp. Sci.*, 2012, vol. 442, pp. 3–21.

11. Brucker, P., On the Complexity of Clustering Problems, *Lect. Not. Econ. Math. Systems*, 1978, vol. 157, pp. 45–54.
12. Bern, M. and Eppstein, D., Approximation Algorithms for Geometric Problems, in *Approximation Algorithms for NP-Hard Problems*, Boston: PWS, 1997, pp. 296–345.
13. Indyk, P., A Sublinear Time Approximation Scheme for Clustering in Metric Space, *Proc. of the 40th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, 1999, pp. 154–159.
14. De la Vega, F. and Kenyon, C., A Randomized Approximation Scheme for Metric Max-Cut, *J. Comput. Syst. Sci.*, 2001, vol. 63, pp. 531–541.
15. De la Vega, F., Karpinski, M., Kenyon, C., and Rabani, Y., Polynomial Time Approximation Schemes for Metric Min-Sum Clustering, *Electronic Colloquium on Computational Complexity (ECCC)*, 2002, rep. no. 25.
16. Hasegawa, S., Imai, H., Inaba, M., Katoh, N., and Nakano, J., Efficient Algorithms for Variance-Based k-Clustering, *Proc. of the 1st Pacific Conference on Computer Graphics and Applications (Pacific Graphics'93, Seoul, Korea)*, River Edge, NJ: World Scientific, 1993, vol. 1, pp. 75–89.
17. Inaba, M., Katoh, N., and Imai, H., Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering, *SCG'94 Proc. of the Tenth Annual Symposium on Computational Geometry*, Stony Brook, NY, USA, 1994, pp. 332–339.
18. Sahni, S. and Gonzalez, T., P-Complete Approximation Problems, *J. ACM*, 1976, vol. 23, pp. 555–566.
19. Ageev, A.A., Kel'manov, A.V., and Pyatkin, A.V., NP-Hardness of the Euclidean Max-Cut Problem, *Dokl. Math.*, 2014, vol. 89, no. 3, pp. 343–345.
20. Ageev, A.A., Kel'manov, A.V., and Pyatkin, A.V., Complexity of the Weighted Max-Cut in Euclidean Space, *J. Appl. Ind. Math.*, 2014, vol. 8, no. 4, pp. 453–457.
21. Kel'manov, A.V. and Pyatkin, A.V., On the Complexity of a Search for a Subset of "Similar" Vectors, *Dokl. Math.*, 2008, vol. 78, no. 1, pp. 574/575.
22. Kel'manov, A.V. and Pyatkin, A.V., On a Version of the Problem of Choosing a Vector Subset, *J. Appl. Ind. Math.*, 2009, vol. 3, no. 4, pp. 447–455.
23. Kel'manov, A.V. and Pyatkin, A.V., NP-Hardness of Some Quadratic Euclidean 2-Clustering Problems, *Dokl. Math.*, 2015, vol. 92, no. 2, pp. 634–637.
24. Kel'manov, A.V. and Pyatkin, A.V., On the Complexity of Some Quadratic Euclidean 2-Clustering Problems, *Comput. Math. Math. Phys.*, 2016, vol. 56, no. 3, pp. 491–497.
25. Bishop, C.M., *Pattern Recognition and Machine Learning*, New York: Springer, 2006.
26. James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning*, New York: Springer, 2013.
27. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
28. Aggarwal, C.C., *Data Mining: The Textbook*, Springer, 2015.
29. Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning (Adaptive Computation and Machine Learning Series)*, MIT Press, 2017.
30. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., and Herawan, T., Big Data Clustering: A Review, *LNCS*, 2014, vol. 8583, pp. 707–720.
31. Pach, J. and Agarwal, P.K., *Combinatorial Geometry*, New York: Wiley, 1995.
32. Kel'manov, A.V. and Khandeev, V.I., A 2-Approximation Polynomial Algorithm for a Clustering Problem, *J. Appl. Ind. Math.*, 2013, vol. 7, no. 4, pp. 515–521.
33. Gimadi, E.Kh., Kel'manov, A.V., Kel'manova, M.A., and Khamidullin, S.A., A Posteriori Detection of a Quasiperiodic Fragment in a Numerical Sequence with a Given Number of Recurrences, *Sib. Zh. Ind. Mat.*, 2006, vol. 9, no. 1, pp. 55–74.
34. Gimadi, E.Kh., Kel'manov, A.V., Kel'manova, M.A., and Khamidullin, S.A., A Posteriori Detecting a Quasiperiodic Fragment in a Numerical Sequence, *Pattern Recogn. Image An.*, 2008, vol. 18, no. 1, pp. 30–42.
35. Baburin, A.E., Gimadi, E.Kh., Glebov, N.I., and Pyatkin, A.V., The Problem of Finding a Subset of Vectors with the Maximum Total Weight, *J. Appl. Ind. Math.*, 2008, vol. 2, no. 1, pp. 32–38.
36. Gimadi, E.Kh., Pyatkin, A.V., and Rykov, I.A., On Polynomial Solvability of Some Problems of a Vector Subset Choice in a Euclidean Space of Fixed Dimension, *J. Appl. Ind. Math.*, 2010, vol. 4, no. 1, pp. 48–53.
37. Shenmaier, V.V., Solving Some Vector Subset Problems by Voronoi Diagrams, *J. Appl. Ind. Math.*, 2016, vol. 10, no. 4, pp. 560–566.
38. Kel'manov, A.V. and Khandeev, V.I., An Exact Pseudopolynomial Algorithm for a Problem of the Two-Cluster Partitioning of a Set of Vectors, *J. Appl. Ind. Math.*, 2015, vol. 9, no. 4, pp. 497–502.
39. Dolgushev, A.V. and Kel'manov, A.V., An Approximation Algorithm for Solving a Problem of Cluster Analysis, *J. Appl. Ind. Math.*, 2011, vol. 5, no. 4, pp. 551–558.

40. Dolgushev, A.V., Kel'manov, A.V., and Shenmaier, V.V., Polynomial-Time Approximation Scheme for a Problem of Partitioning a Finite Set into Two Clusters, *Proc. Steklov Inst. Math.*, 2016, vol. 295, suppl. 1, pp. 47–56.
41. Kel'manov, A.V. and Khandeev, V.I., Fully Polynomial-Time Approximation Scheme for a Special Case of a Quadratic Euclidean 2-Clustering Problem, *J. Appl. Ind. Math.*, 2016, vol. 56, no. 2, pp. 334–341.
42. Kel'manov, A.V., Motkova, A.V., and Shenmaier, V.V., An Approximation Scheme for a Weighted Two-Cluster Partition Problem, *LNCS*, 2018, vol. 10716, pp. 323–333.
43. Kel'manov, A.V. and Khandeev, V.I., A Randomized Algorithm for Two-Cluster Partition of a Set of Vectors, *Comput. Math. Math. Phys.*, 2015, vol. 55, no. 2, pp. 330–339.
44. Kel'manov, A.V. and Motkova, A.V., Polynomial-Time Approximation Algorithm for the Problem of Cardinality-Weighted Variance-Based 2-Clustering with a Given Center, *Comput. Math. Math. Phys.*, 2018, vol. 58, no. 1, pp. 130–136.
45. Kel'manov, A.V. and Motkova, A.V., Exact Pseudopolynomial Algorithms for a Balanced 2-Clustering Problem, *J. Appl. Ind. Math.*, 2016, vol. 10, no. 3, pp. 349–355.
46. Kel'manov, A.V. and Motkova, A.V., A Fully Polynomial-Time Approximation Scheme for a Special Case of a Balanced 2-Clustering Problem, *LNCS*, 2016, vol. 9869, pp. 182–192.
47. Kel'manov, A.V., Khandeev, V.I., and Panasenko, A.V., Randomized Algorithms for Some Clustering Problems, *Comm. Comput. Inform. Sci.*, 2018, vol. CCIS-871, pp. 109–119.