Yugoslav Journal of Operations Research
33 (2023), Number 1, 59–69
DOI: https://doi.org/10.2298/YJOR211018008P

1-MEAN AND 1-MEDOID 2-CLUSTERING PROBLEM WITH ARBITRARY CLUSTER SIZES: COMPLEXITY AND APPROXIMATION

Artem V. PYATKIN
Sobolev Institute of Mathematics, Russia, 630090, Novosibirsk, Koptyug Ave., 4
Novosibirsk State University, Russia, 630090, Novosibirsk, Pirogova Str., 2
artem@math.nsc.ru

Received: October 2021 / Accepted: April 2022

Abstract: We consider the following 2-clustering problem. Given N points in Euclidean space, partition it into two subsets (clusters) so that the sum of squared distances between the elements of the clusters and their centers would be minimum. The center of the first cluster coincides with its centroid (mean) while the center of the second cluster should be chosen from the set of the initial points (medoid). It is known that this problem is NP-hard if the cardinalities of the clusters are given as a part of the input. In this paper we prove that the problem remains NP-hard in the case of arbitrary clusters sizes and suggest a 2-approximation polynomial-time algorithm for this problem.

Keywords: Euclidean space, mean, medoid, 2-clustering, 2-approximation algorithm, strong NP-hardness.

MSC: 68W25, 90C27.

1. INTRODUCTION

The object of study in this paper is 2-clustering, i. e. partition a set of points in Euclidean space into two non-empty clusters according to some similarity criteria. The aim of the paper is to prove NP-hardness of one particular 2-clustering problem in the case of non-fixed cardinalities of the clusters and to suggest a polynomial-time algorithm with a guaranteed performance for this problem. The research is motivated by the fact that the computaional complexity of the problem and issues of its approximation remained unknown up to date. Note that this paper is a full version of the author's conference paper [1] where only the complexity issues of the problem were studied.

Clustering (partitioning a set of some objects into non-empty subsets containing similar objects) is one of the most actual problems in data analysis, data mining, computational geometry, mathematical statistics and discrete optimization [2, 3, 4]. It includes a wide class of problems differing by the number of clusters, similarity criteria, cluster cardinalities constraints, etc.

One of the most usual similarity criteria is the minimum of the squared distances from the elements of the cluster to some point called a *center* of the cluster. There could be following constraints on the choice of the center:

- An arbitrary point (no restrictions)
- A point from the given set
- A given (fixed) point of the space

This choice of centers types can be motivated by the following problem. Asssume that several sensors should be placed for monitoring the situation in towns of a given area. Some sensors are autonomous and could be placed anywhere; others need regular service and so they should be put in towns. There could be also sensors that had been put earlier and cannot be moved (fixed). Since the energy consumption is proportional to the square of the distances, the clustering problem with different centers types can be interpreted as location problem for the sensors of two types taking into account existing sensors and minimizing the total energy consumption.

If all points of a cluster $\mathcal{C}$ are known and the center can be chosen arbitrarily then it is easy to prove by taking partial derivations that the optimal center coincides with the so-called *centroid* defined as

$$\overline{y}(\mathcal{C}) = \frac{\sum_{y \in \mathcal{C}} y}{|\mathcal{C}|}.$$

If this constraint must hold for all cluster centers then one gets a classical problem MSSC (minimum sum-squared clustering) [5, 6] also known as k-*means* where k is the number of clusters. So, we call such centers *means*. If the center of the cluster must coincide with one of the points from the initial set then we call it *medoid*[1]. Finally, the third type of centers (a fixed point) is called *given*.

In 2-clustering, if both clusters have the same requirement on the center (mean, medoid or given) then we denote the problem, respectively, 2-mean, 2-medoid or 2-given. Note that 2-mean 2-clustering problem is the same as the classical 2-means. However, it is possible that different clusters have different constraints on their

[1] The term medoid was introduced in [7] as a representative object of a cluster within a data set whose average dissimilarity to all the objects in the cluster is minimal. Although by dissimilarity usually the distance is meant, applying it for the square of the distance does not contradict the initial definition.

centers. In this case, the corresponding problems are denoted, respectively, 1-mean and 1-medoid, 1-mean and 1-given, or 1-medoid and 1-given. Also, we distinguish whether the cluster cardinalities are fixed (given as a part of input) or can be chosen arbitrarily. Clearly, if a problem with fixed clusters cardinalities is polynomially solvable then the problem with arbitrary clusters sizes is polynomially solvable as well — just consider all $N-1$ possible sizes of the first cluster (where N is the number of points) and choose the best solution. And vice versa, the NP-hardness of a probem with arbitrary clusters cardinalities implies the NP-hardness of the one with fixed sizes of the clusters.

It is easy to see that 2-given 2-clustering problem is polynomially solvable. If the cardinalities of the clusters are not fixed then, clearly, we put each point to the cluster whose center is closer to this point (in fact, we use this procedure in Algorithm A below). In case of fixed cardinalities, if the size of the first cluster is M then just consider for each point the difference between the squared distances from the first and the second center, and choose among them M minimal ones. This implies the polynomial solvability (both for fixed or arbitrary clusters cardinalities) of 1-medoid and 1-given 2-clustering problem that can be reduced to N instances of 2-given 2-clustering problems, and of 2-medoid 2-clustering problem that can be reduced to N^2 instances of 2-given 2-clustering problems.

If at least one cluster center must be a mean then the problem becomes much harder. It is known [8] that the problem k-means in the case of the arbitrary cluster sizes is NP-hard even for $k = 2$. Then by the remark above, the fixed cardinalities version of 2-mean 2-clustering problem is also NP-hard. If both the space dimension d and the number of clusters k are fixed then k-means is polynomially solvable [9]; however, if k is a part of the input, then it remains NP-hard even in the planar case [10], i. e. for $d = 2$. Note also that a PTAS is known [11] for 2-means finding an $(1+\varepsilon)$-approximate solution in time $\mathcal{O}(dN2^{(1/\varepsilon)^{\mathcal{O}(1)}})$.

The 2-clustering problem 1-mean and 1-given was proved to be NP-hard both for the cases of fixed [12, 13] and arbitrary [14, 15] cardinalities. Note that both these variants admit polynomial 2-approximation algorithms of complexity $\mathcal{O}(dN^2)$; for the fixed cardinalities such algorithm can be found in [16], while for the arbitrary arbitrary cardinalities — in [17]. For any fixed space dimension d this problem is polynomially solvable in case of fixed cluster cardinalities. The first algorithm of complexity $\mathcal{O}(dN^{2d+2})$ was suggested in [18]; the best known algorithm of complexity $\mathcal{O}(dN^{d+1})$ can be found in [19].

Finally, the fixed cardinalities vertion of 1-mean and 1-medoid problem was studied in [20] where its NP-hardness was proved (medoid was erraneously called median there).

So, the only case of unknown computation complexity up to date was the 1-mean and 1-medoid 2-clustering problem in the case of arbitrary cluster cardinalities. The conference paper [1] closed this final open case by showing that it was NP-hard. Note that the reduction used in the proof is similar to the one used in [21] for a subset choice problem. The current paper is a journal version of the conference paper [1]; it contains also a 2-approximation polynomial-time algorithm for 1-mean and 1-medoid 2-clustering problem in the case of arbitrary cluster cardinalities.

For the convenience, the review of all cases is given in Table 1, where the contribution of the current paper is shown in **bold**.

Table 1: Complexity of various 2-clustering problems

Centers types	Clusters cardinalities	
	Fixed	Arbitrary
2-given	Polynomially solvable	Polynomially solvable
2-medoid	Polynomially solvable	Polynomially solvable
2-mean	NP-hard [8]	NP-hard [8]
1-medoid and 1-given	Polynomially solvable	Polynomially solvable
1-mean and 1-given	NP-hard [12, 13]	NP-hard [14, 15]
1-mean and 1-medoid	NP-hard [20]	**NP-hard ([1] and this paper)**

The paper is organized as follows. In the next section the strict formulation of the considered problem is given and some preliminary results are proved. In section 3 the main complexity result from [1] is presented. Section 4 contains the main contribution of a journal version, namely, a 2-approximation polynomial-time algorithm. The last section gives some concluding remarks.

2. PRELIMINARIES

Call a cluster *trivial* if it contains only one element. Clearly, for a trivial cluster both mean and medoid center coincide with the only cluster element, and the contribution of such cluster into the objective function is zero. Therefore, it looks reasonable to require in 1-mean and 1-medoid 2-clustering problem that the clusters are non-trivial because, otherwise, the specifics of the cluster can be lost. Note also that there are only n possible trivial clusters, so their excludance does not narrow the problem much.

We make use of the following well-known folklore identity (the proof can be found, for instance, in [15]):

$$\sum_{y\in\mathcal{C}} \|y-\overline{y}(\mathcal{C})\|^2 = \frac{\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2}{2|\mathcal{C}|}. \qquad (1)$$

Using (1), we may formulate 1-mean and 1-medoid 2-clustering problem as follows:

Problem 1. *Given a set of points $\mathcal{Y}=\{y_1,\dots,y_N\}$ in Euclidean space $\mathbb{R}^d$, find a subset $\mathcal{C}\subset\mathcal{Y}$ of cardinality $t\in[2,N-2]$ and a point $x\in\mathcal{Y}$ minimizing the objective function*

$$f(\mathcal{C},x)=\frac{1}{2|\mathcal{C}|}\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-x\|^2.$$

The following property of the optimal solution is quite useful.

Proposition 2. *If $t \in [3, N-2]$ then there exists an optimal solution, such that the center x of the second cluster does not lie in $\mathcal{C}$.*

Proof. If $t \in [3, N-2]$ and $x \in \mathcal{C}$ then consider the cluster $\mathcal{C}' = \mathcal{C} \setminus \{x\}$. Clearly, the second addend in $f(\mathcal{C}, x)$ does not change, and we have

$$f(\mathcal{C}, x) - f(\mathcal{C}', x) = \frac{\sum_{y,z\in\mathcal{C}'} \|z-y\|^2 + 2\sum_{y\in\mathcal{C}'} \|y-x\|^2}{2t} - \frac{\sum_{y,z\in\mathcal{C}'} \|z-y\|^2}{2(t-1)}$$

$$= \frac{\sum_{y\in\mathcal{C}'} \|y-x\|^2}{t} - \frac{\sum_{y,z\in\mathcal{C}'} \|z-y\|^2}{2t(t-1)}$$

$$= \frac{\sum_{y\in\mathcal{C}'} \|y-x\|^2 - \sum_{y\in\mathcal{C}'} \|y-\overline{y}(\mathcal{C}')\|^2}{t} \geq 0.$$

The last inequality follows from the well-known fact that the function $g(x) = \sum_{y\in\mathcal{C}'} \|y-x\|^2$ reaches its global minimum at the centroid of the cluster $\mathcal{C}'$, i. e. at $x = \overline{y}(\mathcal{C}')$. □

Remark 3. *Proposition 2 also works for the version of Problem 1 without the restriction on the clusters cardinalities (see Problem 8 below): it may be assumed there that in the optimal solution the center x of the second cluster does not lie in $\mathcal{C}$ unless $\mathcal{C} = \{x\}$ (otherwise, just move x to another cluster).*

We also need the following lemma proved in [16]:

Lemma 4 ([16]). *Let $\overline{y}(\mathcal{C})$ be a centorid or a cluster $\mathcal{C}$. If a point x satisfies the inequality*

$$\|x - \overline{y}(\mathcal{C})\| \leq \|y - \overline{y}(\mathcal{C})\| \tag{2}$$

for all $y \in \mathcal{C}$ then

$$\sum_{y\in\mathcal{C}} \|y-x\|^2 \leq 2\sum_{y\in\mathcal{C}} \|y-\overline{y}(\mathcal{C})\|^2.$$

3. COMPLEXITY RESULT

Reformulate Problem 1 as a decision problem:

Problem 5. *Given a set of points $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in Eulidean space $\mathbb{R}^d$ and a number $K > 0$, are there a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality $t \in [2, N-2]$ and a point $x \in \mathcal{Y}$ so that $f(\mathcal{C}, x) \leq K$?*

We need the following known NP-complete variant of the exact cover by 3-sets problem [22] where each vertex belongs to at most three subsets:

Problem 6 (X3C3). *Given a* 3*-uniform hypergraph of maximum degree* 3 *on* $n = 3q$ *vertices, is there a subset of* q *edges covering all its vertices? In other words, there is a set of vertices* $V = \{v_1, \dots, v_n\}$ *where* $n = 3q$ *and a collection of edges (subsets)* $E = \{e_1, \dots, e_m\}$ *such that each* $e_i \subseteq V$, $|e_i| = 3$ *and every* v_j *lies in at most three edges; the question is whether there is a subset* $E_0 \subseteq E$ *of cardinality* q *such that* $\cup_{e \in E_0} e = V$*?*

Note that we may assume $m > q+2$ since otherwise the problem X3C3 can be solved by brute force in time $\mathcal{O}(m^2)$.

Theorem 7. *Problem 5 is NP-complete in a strong sense.*

Proof. Consider an arbitrary instance of X3C3 problem and reduce it to an instance of Problem 5 in a following way. Put $d = 3n + 1 = 9q + 1$ and $N = m + 1$. Choose an integer a so that $a^2 > \max\{(m - q - 1)(m - 1)/6, m/18\}$ and let $K = 18a^2(m-1)+m-q$. Each hyperedge e_i corresponds to a point $y_i \in \mathcal{Y}$, $i = 1, \dots, m$ and each vertex v_j corresponds to three coordinates $3j$, $3j - 1$, $3j - 2$ that are referred to as j-th coordinate triple, $j = 1, \dots, n$. Denote by $y_i(k)$ the k-th coordinate of the point y_i. If $v_j \notin e_i$ then put $y_i(3j - 2) = y_i(3j - 1) = y_i(3j) = 0$. Otherwise, one of these three coordinates is $2a$ and other two are $-a$. To determine which one is $2a$, define a parameter s_{ij} as the number of hyperedges with lesser indices than i, containing the vertex v_j, i. e. $s_{ij} = |\{l < i \mid v_j \in e_l\}|$. Note that $s_{ij} \in \{0, 1, 2\}$ since the maximum degree of the hypergraph is 3. Put

$$\begin{aligned} &y_i(3j-2) = 2a,\ y_i(3j-1) = y_i(3j) = -a, \text{ if } s_{ij} = 0;\\ &y_i(3j-1) = 2a,\ y_i(3j-2) = y_i(3j) = -a, \text{ if } s_{ij} = 1;\\ &y_i(3j) = 2a,\ y_i(3j-2) = y_i(3j-1) = -a, \text{ if } s_{ij} = 2. \end{aligned}$$

Also, put $y_i(d) = 1$ for all $i \in \{1, \dots, m\}$ and $y_N(k) = 0$ for all $k \in \{1, \dots, d\}$.

Since the hypergraph is 3-uniform, we have $\|y_i\|^2 = \|y_i - y_N\|^2 = 18a^2 + 1$ and $\|y_i - y_j\|^2 \geq 36a^2$ for all $i, j \in \{1, \dots, N - 1\}$, $i \neq j$. Note also that the equality $\|y_i - y_j\|^2 = 36a^2$ holds if and only if $e_i \cap e_j = \emptyset$.

If a subset E_0 of cardinality q covering all vertices of the hypergraph exists then let $\mathcal{C} = \{y_i \mid e_i \in E_0\}$ and $x = y_N$. Clearly,

$$f(\mathcal{C}, x) = \frac{(q^2 - q)36a^2}{2q} + (m - q)(18a^2 + 1) = 18a^2(m - 1) + m - q = K,$$

as required.

Suppose now that there is a cluster $\mathcal{C} \subset \mathcal{Y}$ of cardinality $t \in [2, N - 2]$ and a point x such that $f(\mathcal{C}, x) \leq K$. Consider two cases.

Case 1. Assume $y_N \in \mathcal{Y} \setminus \mathcal{C}$. In this case, clearly, $x = y_N$. So, the second addend in $f(\mathcal{C}, x)$ is

$$(m - t)(18a^2 + 1). \tag{3}$$

To calculate the first addend, consider several types of coordinate triples. Note that each coordinate triple may be non-zero in $0, 1, 2$, or 3 points from $\mathcal{C}$. So, denote by a_i the number of coordinate triples that are non-zero in exactly i points from $\mathcal{C}$, where $i = 0, 1, 2, 3$. Since the total number of coordinate triples is n and each point from $\mathcal{C}$ has exactly three non-zero coordinate triples, we have $a_0 + a_1 + a_2 + a_3 = n = 3q$ and $a_1 + 2a_2 + 3a_3 = 3t$. The contribution of the a_0 zero coordinate triples into the first addend of the objective function is 0. Each of the a_1 coordinate triples that is non-zero in one point from $\mathcal{C}$ contirbutes $(t-1)6a^2/t$; so, their total contribution is

$$\frac{6(t-1)a^2a_1}{t}. \tag{4}$$

If a coordinate triple is non-zero in two points from $\mathcal{C}$, it contributes $(18a^2 + 2(t-2)6a^2)/t = 6(2t-1)a^2/t$. The total contribution of such triples is

$$\frac{6(2t-1)a^2a_2}{t}. \tag{5}$$

Finally, the total contribution of the triples that are non-zero in three points from $\mathcal{C}$ equals

$$\frac{(3 \cdot 18a^2 + 3(t-3)6a^2)a_3}{t} = 18a^2a_3. \tag{6}$$

Summing (3)–(6) we get

$$f(\mathcal{C}, x) = \frac{6(t-1)a^2a_1 + 6(2t-1)a^2a_2}{t} + 18a^2a_3 + (m-t)(18a^2+1)$$

$$= 6(a_1 + 2a_2 + 3a_3)a^2 + 18(m-t)a^2 - \frac{6(a_1+a_2)a^2}{t} + m - t$$

$$= 18ma^2 - \frac{6(a_1+a_2)a^2}{t} + m - t = K + 18a^2 - \frac{6(a_1+a_2)a^2}{t} + q - t.$$

Note that $a_1 + a_2 = 3t - a_2 - 3a_3 \le 3t$; hence $f(\mathcal{C}, x) > K$ if $q > t$. Therefore, $q \le t$. If $a_1 + a_2 \le 3t - 1$ then using $t \le m - 1$ we have

$$f(\mathcal{C}, x) \ge K + 18a^2 - \frac{6(3t-1)a^2}{t} + q - t \ge K + \frac{6a^2}{m-1} + q - m + 1 > K$$

by the choice of a. So, $a_1 + a_2 = 3t = a_1 + 2a_2 + 3a_3$, i. e., $a_2 = a_3 = 0$ and $a_1 = 3t$. On the other hand, $a_0 + a_1 = 3q \le 3t$, giving $a_0 = 0$ and $q = t$. But then the set $E_0 = \{e_i \mid y_i \in \mathcal{C}\}$ induces a cover of cardinality q in the hypergraph.

Case 2. Assume $y_N \in \mathcal{C}$. If $x = y_N$ then $\sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|x - y\|^2 = (18a^2 + 1)(N - t)$ while if $x = y_i$ for some $y_i \in \mathcal{Y} \setminus \mathcal{C}$ then $\sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|x - y\|^2 \ge 36a^2(N - t - 1)$. Clearly, $(18a^2+1)(N-t) < 36a^2(N-t-1)$ whenever $N - t \ge 3$. So, two subcases are available: either $x \ne y_N$ and $t = N - 2$ or, by Proposition 2, $x = y_N$ and $t = 2$. Consider them separately.

Subcase 2a. Let $t = N - 2 = m - 1$ and $x \ne y_N$. Then the second addend in the objective function is at least $36a^2$. Introduce a_0, a_1, a_2, a_3 in the same way as in the Case 1. Note, however, that now $a_1 + 2a_2 + 3a_3 = 3t - 3 = 3m - 6$ since $y_N \in \mathcal{C}$. Note also that the last coordinate contributes $(m-2)/(m-1)$ into the first addend of the objective function. Using (3)–(5) with $t = m - 1$ and $a_1 + a_2 \leq 3q$, we have

$$f(\mathcal{C}, x) \geq \frac{6(m-2)a^2a_1 + 6(2m-3)a^2a_2}{m-1} + 18a^2a_3 + \frac{m-2}{m-1} + 36a^2$$

$$= 6(a_1 + 2a_2 + 3a_3)a^2 + 36a^2 + 1 - \frac{6(a_1+a_2)a^2 + 1}{m-1}$$

$$\geq 18ma^2 + 1 - \frac{18qa^2+1}{m-1} = K + 18a^2 - m + q + 1 - \frac{18qa^2+1}{m-1}$$

$$= K + q + 1 + \frac{18a^2(m-q-1) - m^2 + m - 1}{m-1}$$

$$> K + q + 1 - \frac{mq+1}{m-1} = K + \frac{m-q-2}{m-1} > K$$

because $a^2 > m/18$ and $m > q + 2$. A contradiction.

Subcase 2b. Let $t = 2$ and $x = y_N$. Then $\mathcal{C} = \{y_i, y_N\}$ for some i. So,

$$f(\mathcal{C}, x) = (18a^2+1)/2 + (m-1)(18a^2+1) = 9a^2 + 1/2 + K + q - 1 > K.$$

Since in both subcases we have a contradiction, Case 2 is impossible.

Note that K and all coordinates of points from $\mathcal{Y}$ are bounded by a polynomial of m and q. Hence, Problem 5 is NP-complete in a strong sense. □

4. 2-APPROXIMATION ALGORITHM

In this section we provide an approximate solution for a non-restricted version of Problem 1 (without the restriction on the clusters cardinalities):

Problem 8. *Given a set of points $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in Eulidean space $\mathbb{R}^d$, find a non-empty subset $\mathcal{C} \subset \mathcal{Y}$ and a point $x \in \mathcal{Y}$ minimizing the objective function $f(\mathcal{C}, x)$.*

The idea of the algorithm is as follows: we first solve a 2-medoid 2-clustering problem and then substitute the center of one of the found clusters by its mean. Here is the formal description of the algorithm.

Algorithm A: Finding an Approximate Solution to Problem 8

Step 0: For all $i, j = 1, \ldots, N$ calculate $a_{ij} = \|y_i - y_j\|^2$.

Step 1: Consider all possible pairs y_i, y_j where $i = 1, \ldots, N-1$ and $j = i + 1, \ldots, N$ and fulfil Steps 2–4 for each of these pairs.

Step 2: For a fixed pair y_i, y_j put $\mathcal{C}_1 = \{y_i\}$, $\mathcal{C}_2 = \{y_j\}$ and do Steps 3 and 4.

Step 3: For each $y_k \in \mathcal{Y} \setminus \{y_i, y_j\}$, if $a_{ik} \le a_{jk}$ let $\mathcal{C}_1 := \mathcal{C}_1 \cup \{y_k\}$ else let $\mathcal{C}_2 := \mathcal{C}_2 \cup \{y_k\}$

Step 4: Calculate $f_{ij} = \min\{f(\mathcal{C}_1, y_j), f(\mathcal{C}_2, y_i)\}$.

Step 5: Output $f_A = \min\{f_{ij} \mid i = 1, \dots, N-1,\ j = i+1, \dots, N\}$.

Theorem 9. *Algorithm A finds a 2-approximate solution to Problem 8 in $\mathcal{O}(dN^2 + N^3)$ time.*

Proof. Clearly, on Steps 2–4 Algorithm A finds an optimum solution to 2-fixed 2-clustering problem with the clusters centers y_i, y_j since moving any point from one cluster to another could only increase the objective function. Hence,

$$\sum_{y\in\mathcal{C}} \|y - y_i\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y - y_j\|^2 \ge \sum_{y\in\mathcal{C}_1} \|y - y_i\|^2 + \sum_{y\in\mathcal{C}_2} \|y - y_j\|^2 \tag{7}$$

for every non-empty subset $\mathcal{C} \subset \mathcal{Y}$. Since the function $g(x) = \sum_{y\in\mathcal{C}} \|y - x\|^2$ reaches its global minimum at the centroid of the cluster $\mathcal{C}$, we also have

$$\sum_{y\in\mathcal{C}_2} \|y - y_i\|^2 + \sum_{y\in\mathcal{C}_1} \|y - y_j\|^2 \ge f_{ij} \tag{8}$$

Let f^* be an optimum solution to Problem 8. Then $f^* = f(\mathcal{C}^*, x^*)$ for some $\mathcal{C}^* \subset \mathcal{Y}$ and $x^* \in \mathcal{Y}$. Note that if $|\mathcal{C}^*| = 1$ then Algorithm A finds an optimum solution to Problem 8 because the contribution of the first addend into the objective function is zero. So, assume $|\mathcal{C}^*| \ge 2$. By Remark 3, we may assume that $x^* \notin \mathcal{C}^*$.

Denote by $y^* = \overline{y}(\mathcal{C}^*)$ the centroid of the cluster $\mathcal{C}^*$ and let x_1 and x_2 be the closest and second closest to y^* points of $\mathcal{Y}$, respectively. Put $y_i = x^*$. If $x_1 \ne x^*$ then put $y_j = x_1$; otherwise, put $y_j = x_2$. Since $x^* \notin \mathcal{C}^*$, the point y_j satisfies (2) anyway. Applying Lemma 4 and (1), we have

$$2f^* = 2\sum_{y\in\mathcal{C}^*} \|y - y^*\|^2 + 2\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y - x^*\|^2 \ge \sum_{y\in\mathcal{C}^*} \|y - y_j\|^2 + 2\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y - y_i\|^2$$

$$\ge \sum_{y\in\mathcal{C}_1} \|y - y_j\|^2 + \sum_{y\in\mathcal{C}_2} \|y - y_i\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y - y_i\|^2 \ge f_{ij} \ge f_A$$

by (7) and (8). So, Algorithm A finds a 2-approximate solution to Problem 8.

Step 0 requires calculations of $\mathcal{O}(N^2)$ norms, each of which taking $\mathcal{O}(d)$ operations. Steps 2–4 are fulfiled $\mathcal{O}(N^2)$ times (once for each of $\mathcal{O}(N^2)$ pairs chosen at Step 1). Among them Step 3 is the most time-consuming: it requires $\mathcal{O}(N)$ comparisons. So, the total time complexity of Algorithm A is $\mathcal{O}(dN^2 + N^3)$. □

5. CONCLUSIONS

In this paper we have proved that 1-mean and 1-medoid 2-clustering problem remains NP-hard in the case of arbitrary clusters cardinalities. This finishes the classification of the complexity of 2-clustering problems where the centers of the clusters can be either means, or medoids or given points. We have also presented a 2-approximation polynomial-time algorithm for this problem.

The question of existence of an approximation scheme for 1-mean and 1-medoid 2-clustering problem remains open.

Anknowledgements. The author is grateful to the unknown referees for their comments improving the paper.

Funding. The research was supported by the program of fundamental scientific researches of the SB RAS, project 0314-2019-0014 and by the Russian Foundation for Basic Research, project 19-01-00308

REFERENCES

[1] A.V. Pyatkin, "NP-hardness of 1-Mean and 1-Medoid 2-Clustering Problem with Arbitrary Clusters Sizes", in *MOTOR 2021. Communications in Computer and Information Science*, vol. 1476, pp. 248–256, 2021.

[2] P. Berkhin, "A Survey of Clustering Data Mining Techniques", in *Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data.* Springer, Berlin, Heidelberg, 2006.

[3] R. C. Dubes, and A. K. Jain, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey, 07632, 1988.

[4] S. Ghoreyshi, and J. Hosseinkhani, "Developing a Clustering Model based on K-Means Algorithm in order to Creating Different Policies for Policyholders", *International Journal of Advanced Computer Science and Information Technology*, vol. 4, no. 2, pp. 46–53, 2015.

[5] W. D. Fisher, "On Grouping for Maximum Homogeneity", *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 789–798, 1958.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the 5-th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5.1, pp. 281–297, 1967.

[7] L. Kaufman, and P. J. Rousseeuw, "Clustering by means of medoids", in *Y. Dodge (ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods.* North-Holland, Amsterdam, 1987, pp. 405–416 .

[8] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering", *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[9] M. Inaba, N. Katoh, and H. Imai "Applications of weighted Voronoi diagrams and randomization to variance-based clustering", in *Proceedings of the tenth annual symposium on Computational geometry*, pp. 332–339, 1994.

[10] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard", *Theoretical Computer Science*, vol. 442, pp. 13–21, 2012.

[11] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time $(1+\varepsilon)$-approximation algorithm for geometric k-means clustering in any dimensions", in *Proceedings - Annual Symposium on Foundations of Computer Science*, pp. 454–462, 2004.

[12] A. E. Baburin, E. Kh. Gimadi, N. I. Glebov, and A. V. Pyatkin, "The problem of finding a subset of vectors with the maximum total weight", *Journal of Applied and Industrial Mathematics*, vol. 2, no. 1, pp. 32–38, 2008.

[13] E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova, and S. A. Khamidullin, "A posteriori detection of a quasi periodic fragment in numerical sequences with given number of recurrences", *Sibirskii Zhurnal Industrial'noi Matematiki*, vol. 9, no. 1, pp. 55–74, 2006. (in Russian).

[14] A. V. Kelmanov, and A. V. Pyatkin, "On the complexity of a search for a subset of '"similar" vectors", *Doklady Mathematics*, vol. 78, no. 1, pp. 574–575, 2008.

[15] A. V. Kel'manov, and A. V. Pyatkin, "On a Version of the Problem of Choosing a Vector Subset", *Journal of Applied and Industrial Mathematics*, vol. 3, no. 4, pp. 447–455, 2009.

[16] A. V. Dolgushev, and A. V. Kel'manov, "An approximation algorithm for solving a problem of cluster analysis", *Journal of Applied and Industrial Mathematics*, vol. 5, no. 4, pp. 551–558, 2011.

[17] A. V. Kel'manov, and V. I. Khandeev, "A 2-approximation polynomial algorithm for a clustering problem", *Journal of Applied and Industrial Mathematics*, vol. 7, no. 4, pp. 515–521, 2013.

[18] E. Kh. Gimadi, A. V. Pyatkin, and I. A. Rykov, "On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension", *Journal of Applied and Industrial Mathematics*, vol. 4, no. 1, pp. 48–53, 2010.

[19] V.V. Shenmaier, "Solving some vector subset problems by Voronoi diagrams", *Journal of Applied and Industrial Mathematics*, vol. 10, no. 4, pp. 560–566, 2016.

[20] A. V. Kel'manov, A. V. Pyatkin, and V. I., Khandeev, "NP-Hardness of Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with Constraints on the Cluster Sizes", *Doklady Mathematics*, vol. 100, no. 3, pp. 545–548, 2019.

[21] A.V. Pyatkin, "Easy NP-hardness Proofs of Some Subset Choice Problems", in *Kochetov Y., Bykadorov I., Gruzdeva T. (eds) Mathematical Optimization Theory and Operations Research. MOTOR 2020. Communications in Computer and Information Science*, vol. 1275, pp. 70–79, 2020.

[22] M. J. Garey, and D. S. Johnson, *Computers and Intractability. The Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, San Francisco, 1979.