On Polynomial Solvability of One Quadratic Euclidean Clustering Problem on a Line

Alexander Kel'manov[1,2] and Vladimir Khandeev[1,2](✉)

[1] Sobolev Institute of Mathematics, 4 Koptyug Ave., 630090 Novosibirsk, Russia
[2] Novosibirsk State University, 2 Pirogova St., 630090 Novosibirsk, Russia
{kelm,khandeev}@math.nsc.ru

Abstract. We consider one problem of partitioning a finite set of points in Euclidean space into clusters so as to minimize the sum over all clusters of the intracluster sums of the squared distances between clusters elements and their centers. The centers of some clusters are given as an input, while the other centers are unknown and defined as centroids (geometrical centers). It is known that the general case of the problem is strongly NP-hard. We show that there exists an exact polynomial algorithm for the one-dimensional case of the problem.

Keywords: Minimum Sum-of-Squares Clustering · Euclidean space · NP-hard problem · One-dimensional case · Polynomial solvability

1 Introduction

The subject of this study is one strongly NP-hard problem of partitioning a finite set of points in Euclidean space into clusters. Our goal is to analyze the computational complexity of the problem in the one-dimensional case. The research is motivated by the openness of the specified mathematical question, as well as by the importance of the problem for some applications, in particular, for Data analysis, Data mining, Pattern recognition, and Data processing.

The paper has the following structure. In Sect. 2, the problem formulation is given. In the same section, a connection is established with a well-known problem that is the closest to we consider one. The next section presents auxiliary statements that reveal the structure of the optimal solution to the problem. These statements allow us to prove the main result. In Sect. 4, our main result of the polynomial solvability of the problem in the 1D case is presented.

2 Problem Formulation, Its Sources and Related Problems

In the well-known clustering *K-Means* problem, an N-element set $\mathcal{Y}$ of points in d-dimension Euclidean space and a positive integer K are given. It is required to

N. F. Matsatsinis et al. (Eds.): LION 13 2019, LNCS 11968, pp. 46–52, 2020.
https://doi.org/10.1007/978-3-030-38629-0_4

find a partition of the input set $\mathcal{Y}$ into non-empty clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$ minimizing the sum

$$\sum_{k=1}^{K} \sum_{y \in \mathcal{C}_k} \|y - \overline{y}(\mathcal{C}_k)\|^2,$$

where $\overline{y}(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{y \in \mathcal{C}_k} y$ is the centroid of the k-th cluster.

Another common name of *K-Means* problem is *MSSC* (*Minimum Sum-of-Squares Clustering*). In statistics, this problem is known from the last century and is associated with Fisher (see, for example, [1,2]). In practice (in a wide variety of applications), this problem arises when there is the following hypothesis on a structure of some given numerical data. Namely, one has assumption that the set $\mathcal{Y}$ of sample (input) data contains K homogeneous clusters (subsets) $\mathcal{C}_1, \dots, \mathcal{C}_K$, and in all clusters, the points are scattered around the corresponding unknown mean values $\overline{y}(\mathcal{C}_1), \dots, \overline{y}(\mathcal{C}_K)$. However, the correspondence between points and clusters is unknown. Obviously, in this situation, for the correct application of classical statistical methods (hypothesis testing or parameter estimating) to the processing of sample data, at first it is necessary to divide the data into homogeneous groups (clusters). This situation is typical, in particular, for the above-mentioned (see Sect. 1) applications.

The *K-Means* strong NP-hardness was proved relatively recently [3]. The polynomial solvability of this problem on a line was proved in [4] in the last century. The cited paper presents an algorithm with $\mathcal{O}(KN^2)$ running time that implements a dynamic programming scheme. This well-known algorithm relies on an exact polynomial algorithm for solving the well-known *Nearest neighbor search* problem [5]. Note that the polynomial solvability in $\mathcal{O}(KN \log N)$-time of the 1D case of the *K-Means* problem follows directly from earlier (than [4]) results obtained in [6–9]. In the cited papers, the authors have proved the faster polynomial-time algorithms for some special cases of the *Nearest neighbor search* problem. Nevertheless, in recent years, for the one-dimensional case of the *K-Means* problem, some new exact algorithms with $\mathcal{O}(KN \log N)$ running time have been constructed. An overview of these algorithms and their properties can be found in [10,11].

The object of our research is the following problem that is close in its formulation to *K-Means* and is poorly studied.

Problem 1 (K-Means and Given J-Centers). *Given* an N-element set $\mathcal{Y}$ of points in d-dimension Euclidean space, a positive integer K, and a tuple $\{c_1, \dots, c_J\}$ of points. *Find* a partition of $\mathcal{Y}$ into non-empty clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$, $\mathcal{D}_1, \dots, \mathcal{D}_J$ such that

$$F = \sum_{k=1}^{K} \sum_{y \in \mathcal{C}_k} \|y - \overline{y}(\mathcal{C}_k)\|^2 + \sum_{j=1}^{J} \sum_{y \in \mathcal{D}_j} \|y - c_j\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_k)$ is the centroid of the k-st cluster.

On the one hand, Problem 1 may be considered as some modification of K-Means. On the other hand, the introduced notation allows us to call Problem 1 as *K-Means and Given J-Centers.*

Unlike *K-Means*, Problem 1 models an applied clustering problem in which for a part of clusters (i.e., for $\mathcal{D}_1, \ldots, \mathcal{D}_J$) the quadratic scatter data centers (i.e., $c_1, \ldots, c_J$) are known in advance, i.e., they are given as input instance. This applied problem is also typical for Data analysis, Data mining, Pattern recognition, and Data processing. In particular, the two-cluster Problem 1, i.e., 1-*Mean and Given* 1-*Center*, is related to the solution of the applied signal processing problem. Namely, this two-clusters problem is related with the problem of joint detecting a quasi-periodically repeated pulse of unknown shape in a pulse train and evaluating this shape under Gaussian noise with given zero value (see [12–14]). In this two-cluster Problem 1, the zero mean corresponds to the cluster with the center specified at the origin. Apparently, the first mention has been made in [12] on this two-cluster Problem 1. It should be noted that simpler optimization problems induced by the applied problems of noise-proof detection and discrimination of impulses of specified shapes are typical, in particular, for radar, electronic reconnaissance, hydroacoustics, geophysics, technical and medical diagnostics, and space monitoring (see, for example, [15–17]).

Problem 1 strong NP-hardness was proved in [18–20]. Note that the *K-Means* problem is not equivalent to Problem 1 and is not a special case of it. Therefore, the solvability of Problem 1 in the 1D case requires independent study. This question until now remained open.

The main result of this paper is the proof of Problem 1 polynomial solvability in the one-dimensional case.

3 Some Auxiliary Statements: Properties of the Problem 1 Optimal Solution in the 1D Case

In what follows, we assume that $d = 1$. Below we will call by Problem 1D the one-dimensional case of Problem 1.

Our proof is based on the few given below auxiliary statements, which reveal the structure of Problem 1D optimal solution. For briefness, we present these statements without proofs, limiting ourselves to the presentation of their ideas.

Denote by $\mathcal{C}_1^*, \ldots, \mathcal{C}_K^*$, $\mathcal{D}_1^*, \ldots, \mathcal{D}_J^*$ the optimal clusters in Problem 1D.

Lemma 1. *If in Problem 1D $c_m < c_\ell$, where $1 \le m \le J$, $1 \le \ell \le J$, then for each $x \in \mathcal{D}_m^*$ and $z \in \mathcal{D}_\ell^*$ the inequality $x \le z$ holds.*

Lemma 2. *If in Problem 1D $\overline{y}(\mathcal{C}_m^*) < \overline{y}(\mathcal{C}_\ell^*)$, where $1 \le m \le K$, $1 \le \ell \le K$, then for each $x \in \mathcal{C}_m^*$ and $z \in \mathcal{C}_\ell^*$ the inequality $x \le z$ holds.*

Lemma 3. *For an optimal solution of Problem 1D, the following statements are true:*

(1) *If $\overline{y}(\mathcal{C}_m^*) < c_\ell$, where $1 \le m \le K$, $1 \le \ell \le J$, then for each $x \in \mathcal{C}_m^*$ and $z \in \mathcal{D}_\ell^*$ the inequality $x \le z$ holds.*

(2) *If $\overline{y}(\mathcal{C}_m^*) > c_\ell$, where $1 \le m \le K$, $1 \le \ell \le J$, then for each $x \in \mathcal{C}_m^*$ and $z \in \mathcal{D}_\ell^*$ the inequality $x \ge z$ holds.*

The proof of Lemmas 1–3 is carried out by the contrary method using the following equality

$$(x - c_m)^2 + (z - c_\ell)^2 = 2(x - z)(c_\ell - c_m) + (z - c_m)^2 + (x - c_\ell)^2.$$

The validity of this equality follows from the well-known formula for the sum of squares of the trapezoid diagonals.

Lemma 4. *In Problem 1D, for each $k \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, J\}$ it is true that $\overline{y}(\mathcal{C}_k^*) \neq c_j$.*

Lemma 5. *In Problem 1D, for each $k, j \in \{1, \ldots, K\}$, $k \neq j$, it is true that $\overline{y}(\mathcal{C}_k^*) \neq \overline{y}(\mathcal{C}_j^*)$.*

The proof of Lemmas 4 and 5 is carried out by the contrary method.

Lemmas 1–5 establish the relative position of the optimal clusters $\mathcal{D}_1^*, \ldots, \mathcal{D}_J^*$ and $\mathcal{C}_1^*, \ldots, \mathcal{C}_K^*$ on a line. These lemmas are the base of the following statement.

Theorem 1. *Let in Problem 1D points $y_1, \ldots, y_N$ of $\mathcal{Y}$, and points $c_1, \ldots, c_J$ be ordered so that*

$$y_1 < \ldots < y_N,$$
$$c_1 < \ldots < c_J.$$

Then optimal partition of $\mathcal{Y}$ into clusters $\mathcal{C}_1^*, \ldots, \mathcal{C}_K^*, \mathcal{D}_1^*, \ldots, \mathcal{D}_J^*$ corresponds to a partition of the positive integer sequence $1, \ldots, N$ into disjoint segments.

4 Polynomial Solvability of the Problem in the 1D Case

The following theorem is the main result of the paper.

Theorem 2. *There exists an algorithm that finds an optimal solution of Problem 1D in polynomial time.*

Our proof of Theorem 1 is constructive. Namely, we justify an algorithm that implements a dynamic programming scheme and allows one to find an exact solution of Problem 1D in $\mathcal{O}(KJN^2)$ time.

The idea of the proof is as follows. Without loss of generality, we assume that the points $y_1, \ldots, y_N$ of $\mathcal{Y}$, as well as the points $c_1, \ldots, c_J$ are ordered as in Theorem 1.

Let $\mathcal{Y}_{s,t} = \{y_s, \ldots, y_t\}$, where $1 \le s \le t \le N$, be a subset of $t - s + 1$ points of $\mathcal{Y}$ with numbers from s to t.

Let

$$f_{s,t}^j = \sum_{i=s}^{t} (y_i - c_j)^2, \quad j = 1, \ldots, J,$$

$$f_{s,t} = \sum_{i=s}^{t} (y_i - \overline{y}(\mathcal{Y}_{s,t}))^2,$$

where $\overline{y}(\mathcal{Y}_{s,t})$ is the centroid of the subset $\mathcal{Y}_{s,t}$.

We prove that the optimal value of the Problem 1 objective function is found by the following formula

$$F^* = F_{K,J}(N),$$

and the values

$$F_{k,j}(n),\;\; k = -1, 0, 1, \dots, K;\;\; j = -1, 0, 1, \dots, J;\;\; n = 0, \dots, N,$$

are calculated by the recurrent formulas. The formula

$$F_{k,j}(n) = \begin{cases} 0, & \text{if } n = k = j = 0; \\ +\infty, & \text{if } n = 0;\; k = 0, \dots, K;\; j = 0, \dots, J;\; k + j \neq 0; \\ +\infty, & \text{if } k = -1;\; j = -1, \dots, J;\; n = 0, \dots, N; \\ +\infty, & \text{if } j = -1;\; k = -1, \dots, K;\; n = 0, \dots, N; \end{cases} \tag{1}$$

sets the initial and boundary conditions for subsequent calculations. Formula (1) follows from the properties of the optimal solution. The basic formula

$$F_{k,j}(n) = \min\Big\{\min_{i=1}^{n}\Big\{F_{k-1,j}(i-1) + f_{i,n}\Big\},\; \min_{i=1}^{n}\Big\{F_{k,j-1}(i-1) + f_{i,n}^{j}\Big\}\Big\},$$
$$k = 0, \dots, K;\; j = 0, \dots, J;\; n = 1, \dots, N, \tag{2}$$

defines recursion. In general, the formulas (1), (2) implement the forward algorithm.

Further, we have proved that the optimal clusters $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ may be found using the following recurrent rule, that implements the backward algorithm.

The step-by-step rule looks as follows:

Step 0. $k := K$, $j := J$, $n := N$.

Step 1. If

$$\min_{i=1}^{n}\Big(F_{k-1,j}(i-1) + f_{i,n}\Big) \leq \min_{i=1}^{n}\Big(F_{k,j-1}(i-1) + f_{i,n}^{j}\Big),$$

then

$$\mathcal{C}_k^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\},$$

where

$$i^* = \arg\min_{i=1}^{n}\Big(F_{k-1,j}(i-1) + f_{i,n}\Big);$$

$k := k - 1$; $n := i^* - 1$.

If, however,

$$\min_{i=1}^{n}\Big(F_{k-1,j}(i-1) + f_{i,n}\Big) > \min_{i=1}^{n}\Big(F_{k,j-1}(i-1) + f_{i,n}^{j}\Big),$$

then

$$\mathcal{D}_j^* = \{y_{i^*}, y_{i^*+1}, \ldots, y_n\},$$

where

$$i^* = \arg\min_{i=1}^{n}\Big(F_{k,j-1}(i-1) + f_{i,n}^{j}\Big);$$

$j := j-1$; $n := i^* - 1$.

Step 2. If $k > 0$ or $j > 0$, then go to Step 1; otherwise — the end of calculations.

The validity of this rule we have proved by induction.

Finally, we have proved that the running time of the algorithm is $\mathcal{O}(KJN^2)$, that is, the algorithm is polynomial. The algorithms running time is defined by the complexity of implementation of formula (2). This formula is calculated $\mathcal{O}(KJN)$ times and every calculation of $F_{k,j}(n)$ requires $\mathcal{O}(N)$ operations.

5 Conclusion

In the present paper, we have proved the polynomial solvability of the one-dimensional case of one strongly NP-hard problem of partitioning a finite set of points in Euclidean space. The construction of approximate efficient algorithms with guaranteed accuracy bounds for the general case of Problem 1 and faster polynomial-time exact algorithms for the 1D case of this problem seems to be the directions of future studies.

Acknowledgments. The study presented in Sects. 3 and 4 was supported by the Russian Foundation for Basic Research, projects 19-01-00308 and 18-31-00398. The study presented in the other sections was supported by the Russian Academy of Science (the Program of basic research), project 0314-2019-0015, and by the Russian Ministry of Science and Education under the 5-100 Excellence Programme.

References

1. Fisher, R.A.: Statistical Methods and Scientific Inference. Hafner, New York (1956)
2. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
3. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of euclidean sum-of-squares clustering. Mach. Learn. **75**(2), 245–248 (2009)
4. Rao, M.: Cluster analysis and mathematical programming. J. Am. Stat. Assoc. **66**, 622–626 (1971)
5. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
6. Glebov, N.I.: On the convex sequences. Discrete Anal. **4**, 10–22 (1965). in Russian
7. Gimadutdinov, E.K.: On the properties of solutions of one location problem of points on a segment. Control. Syst. **2**, 77–91 (1969). in Russian
8. Gimadutdinov, E.K.: On one class of nonlinear programming problems. Control. Syst. **3**, 101–113 (1969). in Russian

9. Gimadi (Gimadutdinov) E.Kh.: The choice of optimal scales in one class of location, unification and standardization problems. Control. Syst. **6**, 57–70 (1970). in Russian
10. Wu, X.: Optimal quantization by matrix searching. J. Algorithms **12**(4), 663–673 (1991)
11. Grønlund, A., Larsen, K.G., Mathiasen, A., Nielsen, J.S., Schneider, S., Song, M.: Fast exact k-means, k-medians and Bregman divergence clustering in 1D. CoRR arXiv:1701.07204 (2017)
12. Kel'manov, A.V., Khamidullin, S.A., Kel'manova, M.A.: Joint finding and evaluation of a repeating fragment in noised number sequence with given number of quasiperiodic repetitions. In: Book of Abstract of the Russian Conference "Discrete Analysis and Operations Research" (DAOR-4), p. 185. Sobolev Institute of Mathematics SB RAN, Novosibirsk (2004)
13. Gimadi, E.K., Kel'manov, A.V., Kel'manova, M.A., Khamidullin, S.A.: A posteriori detection of a quasi periodic fragment in numerical sequences with given number of recurrences. Sib. J. Ind. Math. **9**(1(25)), 55–74 (2006). in Russian
14. Gimadi, E.K., Kel'manov, A.V., Kel'manova, M.A., Khamidullin, S.A.: A posteriori detecting a quasiperiodic fragment in a numerical sequence. Pattern Recogn. Image Anal. **18**(1), 30–42 (2008)
15. Kel'manov, A.V., Khamidullin, S.A.: Posterior detection of a given number of identical subsequences in a guasi-periodic sequence. Comput. Math. Math. Phys. **41**(5), 762–774 (2001)
16. Kel'manov, A.V., Jeon, B.: A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train. IEEE Trans. Sig. Process. **52**(3), 645–656 (2004)
17. Carter, J.A., Agol, E., et al.: Kepler-36: a pair of planets with neighboring orbits and dissimilar densities. Science **337**(6094), 556–559 (2012)
18. Kel'manov, A.V., Pyatkin, A.V.: On the complexity of a search for a subset of "similar" vectors. Dokl. Math. **78**(1), 574–575 (2008)
19. Kel'manov, A.V., Pyatkin, A.V.: On a version of the problem of choosing a vector subset. J. Appl. Ind. Math. **3**(4), 447–455 (2009)
20. Kel'manov, A.V., Pyatkin, A.V.: Complexity of certain problems of searching for subsets of vectors and cluster analysis. Comput. Math. Math. Phys. **49**(11), 1966–1971 (2009)