






Randomized Algorithms for Some Sequence Clustering Problems

Sergey Khamidullin¹ , Vladimir Khandeev^{1,2} , and Anna Panasenکو^{1,2} 

¹ Sobolev Institute of Mathematics, 4 Koptyug Avenue, 630090 Novosibirsk, Russia
{kham,khandeev,a.v.panasenko}@math.nsc.ru

² Novosibirsk State University, 2 Pirogova Street, 630090 Novosibirsk, Russia

Abstract. We consider two problems of clustering a finite sequence of points in Euclidean space. In the first problem, we need to find a cluster minimizing intracluster sum of squared distances from cluster elements to its centroid. In the second problem, we need to partition a sequence into two clusters minimizing cardinality-weighted intracluster sums of squared distances from clusters elements to their centers; the center of the first cluster is its centroid, while the center of the second one is the origin. Moreover, in the first problem, the difference between any two subsequent indices of cluster elements is bounded above and below by some constants. In the second problem, the same constraint is imposed on the cluster with unknown centroid. We present randomized algorithms for both problems and find the conditions under which these algorithms are polynomial and asymptotically exact.

Keywords: Clustering · Euclidean space · Minimum sum-of-squares · NP-hard problem · Randomized algorithm · Asymptotic accuracy

1 Introduction

The subject of this study are two strongly NP-hard problems of clustering a finite sequence of points in Euclidean space. Our goal is to construct a randomized algorithm for the problems. The research is motivated by the fact that the considered problems are related to mathematical time series analysis problems, approximation and discrete optimization problems, and also by their importance for applications such as signals analysis and recognition, remote object monitoring, etc. (see the next section and the papers therein).

The paper has the following structure. In Sect. 2, formulation of the problems is given. In the same Section, the known results are listed. The next Section contains the auxiliary problem and the algorithm for solving it, which are needed to construct our proposed algorithms. In Sect. 4, the randomized algorithms for the considered problems are presented.

2 Problems Formulation, Related Problems, and Known Results

We consider the following two problems.

Problem 1. Given a sequence $\mathcal{Y} = (y_1, \dots, y_N)$ of points in \mathbb{R}^d and positive integers T_{\min} , T_{\max} and $M > 1$. Find a subset $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N} = \{1, \dots, N\}$ of the index set of \mathcal{Y} such that

$$F_1(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \bar{y}(\mathcal{M})\|^2 \longrightarrow \min ,$$

where $\bar{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$ is the centroid of $\{y_j \mid j \in \mathcal{M}\}$, under the constraints

$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \quad m = 2, \dots, M, \quad (1)$$

on the elements of the set (n_1, \dots, n_M) .

Problem 2. Given a sequence $\mathcal{Y} = (y_1, \dots, y_N)$ of points in \mathbb{R}^d and positive integers T_{\min} , T_{\max} , and $M > 1$. Find a subset $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N} = \{1, \dots, N\}$ of the index set of \mathcal{Y} such that

$$F_2(\mathcal{M}) = |\mathcal{M}| \sum_{j \in \mathcal{M}} \|y_j - \bar{y}(\mathcal{M})\|^2 + |\mathcal{N} \setminus \mathcal{M}| \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min ,$$

where $\bar{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$ is the centroid of $\{y_j \mid j \in \mathcal{M}\}$, under the constraints (1) on the elements of the set (n_1, \dots, n_M) .

Problem 1 is induced by the following applied problem. Given a sequence \mathcal{Y} of N time-ordered measurements of d numerical characteristics of some object. M of these measurements correspond to a repeating (identical) state of the object. There is an error in each given measurement result. The correspondence of the measurement results to the states of the object is unknown. However, it is known that the time interval between two consecutive identical states is bound from above and below by the specified constants T_{\min} and T_{\max} . It is required to find a subsequence of numbers corresponding to the measurements of the repeated state of the object.

In the special case when $T_{\min} = 1$ and $T_{\max} = N$, Problem 1 is equivalent to the well-known M -variance problem (see, e.g., [1]). A list of known results for M -variance problem can be found in [2].

When T_{\min} and T_{\max} are parameters, Problem 1 is strongly NP-hard for any $T_{\min} < T_{\max}$ [3]. When $T_{\min} = T_{\max}$, it is solvable in polynomial time.

In [4], a 2-approximation algorithm with $\mathcal{O}(N^2(MN + d))$ running time is proposed.

An exact algorithm for the case of integer inputs was substantiated in [5]. When the space dimension is fixed, the algorithm is pseudopolynomial and runs in $\mathcal{O}(N^3(MD)^d)$ time.

In [6], an FPTAS was presented for the case of Problem 1 when the space dimension is fixed. Given relative error ε , this algorithm finds a $(1 + \varepsilon)$ -approximate solution to the problem in $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ time.

Problem 2 simulates the following applied problem. As in Problem 1, we have a sequence \mathcal{Y} of N time-ordered measurement results for d characteristics of some object. This object can be in two different states (active and passive, for example). Each measurement has an error and the correspondence between the elements of the input sequence and the states is unknown. One knows that the object was in the active state exactly M times (or the probability of the active state is $\frac{M}{N}$) and the time interval between every two consecutive active states is bounded from below and above by some constants T_{\min} and T_{\max} . It is required to find a 2-partition of the input sequence and evaluate the object characteristics.

If $T_{\min} = 1$ and $T_{\max} = N$, Problem 2 is equivalent to *Cardinality-weighted variance-based 2-clustering with given center* problem. One can easily find a list of known results for this special case in [8].

Cardinality-weighted variance-based 2-clustering with given center problem is related but not equivalent to the well-known *Min-sum all-pairs 2-clustering* problem (see, e.g., [9, 10]). Many algorithmic results are known for this closely related problem, but they are not directly applicable to *Cardinality-weighted variance-based 2-clustering with given center* problem.

Problem 2 is strongly NP-hard [11]. Only two algorithmic results have been proposed for this problem until now.

An exact pseudopolynomial algorithm was proposed in [11] for the case of integer instances and the fixed space dimension d . The running time of this algorithm is $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + d)(2MD + 1)^d)$, where D is the maximum absolute value of coordinates of the input points.

In [12], a 2-approximation algorithm was presented. The running time of the algorithm is $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + d))$.

The main results of this paper are randomized algorithms for Problems 1 and 2. These algorithms find a $(1 + \varepsilon)$ -approximate solution with probability not less than $1 - \gamma$ in $\mathcal{O}(dMN^2)$ time, for the given $\varepsilon > 0$, $\gamma \in (0, 1)$ and under assumption $M \geq \beta N$ for $\beta \in (0, 1)$. The conditions are found under which these algorithms are asymptotically exact (i.e. the algorithms find a $(1 + \varepsilon_N)$ -approximate solutions with probability $1 - \gamma_N$, where $\varepsilon_N, \gamma_N \rightarrow 0$) and find the solutions in $\mathcal{O}(dMN^3)$ time.

3 Auxiliary Problem

To construct the algorithms for Problems 1 and 2, we need the following auxiliary problem.

Problem 3. Given a sequence $g(n)$, $n = 1, \dots, N$, of real values, positive integers T_{\min} , T_{\max} and $M > 1$. Find a subset $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$ of indices of sequence elements such that

$$G(\mathcal{M}) = \sum_{i \in \mathcal{M}} g(i) \rightarrow \min ,$$

under constraints (1) on the elements of the tuple (n_1, \dots, n_M) .

The following algorithm finds the solution of Problem 3.

Algorithm A.

Input: a sequence $g(n)$, $n = 1, \dots, N$, numbers T_{\min} , T_{\max} and $M > 1$.

Step 1. Compute

$$G_m(n) = \begin{cases} g(n), & \text{if } n \in \omega_1, m = 1; \\ g(n) + \max_{j \in \gamma_{m-1}^-(n)} G_{m-1}(j), & \text{if } n \in \omega_m, m = 2, \dots, M, \end{cases}$$

where

$$\omega_m = \{n \mid 1 + (m-1)T_{\min} \leq n \leq N - (M-m)T_{\min}\}, m = 1, \dots, M,$$

$$\gamma_{m-1}^-(n) = \{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \leq j \leq n - T_{\min}\}, \\ n \in \omega_m, m = 2, \dots, M.$$

Step 2. Compute

$$G_{\max}^x = \max_{n \in \omega_M} G_M^x(n)$$

and find the tuple $\mathcal{M} = (n_1, \dots, n_M)$ by the formulae

$$n_M^x = \arg \max_{n \in \omega_M} G_M^x(n),$$

$$n_{m-1}^x = \arg \max_{n \in \gamma_m^-(n_m^x)} G_m^x(n), \quad m = M, M-1, \dots, 2.$$

Output: the tuple $\mathcal{M} = (n_1, \dots, n_M)$.

Remark 1. It follows from [4, 7] that Algorithm A finds the optimal solution of Problem 3 in $\mathcal{O}(NM(T_{\max} - T_{\min} + 1))$ time.

4 Randomized Algorithms

Below is a randomized algorithm for Problem 1.

Algorithm A₁.

Input: a sequence \mathcal{Y} , positive integers T_{\min} , T_{\max} , M , a positive integer parameter k .

Step 1. Generate a multiset \mathcal{T} of points by randomly and independently choosing k elements from \mathcal{Y} with replacement.

Step 2. For every nonempty subset $\mathcal{H} \subseteq \mathcal{T}$ compute the centroid $\bar{y}(\mathcal{H})$ and find a solution $\mathcal{M} = \mathcal{M}(\mathcal{H})$ of Problem 3 for $g(n) = \|y_n - \bar{y}(\mathcal{H})\|^2$, $n = 1, \dots, N$.

Step 3. From the family of solutions $\{\mathcal{M}(\mathcal{H}) \mid \mathcal{H} \subseteq \mathcal{T}\}$ found at Step 2, choose the set $\mathcal{M}_{A_1} = \mathcal{M}(\mathcal{H})$ for which the value of $F_1(\mathcal{M}(\mathcal{H}))$ is minimal.

Output: the set \mathcal{M}_{A_1} .

The next randomized algorithm allows one to find approximate solution of Problem 2.

Algorithm \mathcal{A}_2 .

Input: a sequence \mathcal{Y} , positive integers T_{\min} , T_{\max} , M , a positive integer parameter k .

Step 1. Generate a multiset \mathcal{T} of points by randomly and independently choosing k elements from \mathcal{Y} with replacement.

Step 2. For every nonempty subset $\mathcal{H} \subseteq \mathcal{T}$ compute the centroid $\bar{y}(\mathcal{H})$ and find a solution $\mathcal{M} = \mathcal{M}(\mathcal{H})$ of Problem 3 for $g(n) = 2M\langle y_n, \bar{y}(\mathcal{H}) \rangle - (2M - N)\|y_n\|^2 - M\|\bar{y}(\mathcal{H})\|^2$, $n = 1, \dots, N$.

Step 3. From the family of solutions $\{\mathcal{M}(\mathcal{H}) \mid \mathcal{H} \subseteq \mathcal{T}\}$ found at Step 2, choose the set $\mathcal{M}_{\mathcal{A}_2} = \mathcal{M}(\mathcal{H})$ for which the value of $F_2(\mathcal{M}(\mathcal{H}))$ is minimal.

Output: the set $\mathcal{M}_{\mathcal{A}_2}$.

The following theorem describes the properties of algorithms \mathcal{A}_1 and \mathcal{A}_2 .

Theorem 1. *Assume that in Problems 1 and 2, $M \geq \beta N$ for $\beta \in (0, 1)$. Then, given $\varepsilon > 0$ and $\gamma \in (0, 1)$, for a fixed parameter*

$$k = \max\left(\left\lceil \frac{2}{\beta} \left\lceil \frac{2}{\gamma\varepsilon} \right\rceil \right\rceil, \left\lceil \frac{8}{\beta} \ln \frac{2}{\gamma} \right\rceil\right)$$

algorithms \mathcal{A}_1 and \mathcal{A}_2 find $(1 + \varepsilon)$ -approximate solutions of Problem 1 and 2 with probability $1 - \gamma$ in $\mathcal{O}(dMN^2)$ time.

Finally, in the next theorem, conditions are established under which algorithms \mathcal{A}_1 and \mathcal{A}_2 are polynomial and asymptotically exact.

Theorem 2. *Assume that in Problems 1 and 2, $M \geq \beta N$ for $\beta \in (0, 1)$. Then, for fixed $k = \lceil \log_2 N \rceil$, algorithms \mathcal{A}_1 and \mathcal{A}_2 find $(1 + \varepsilon_N)$ -approximate solutions of Problem 1 and 2 with probability $1 - \gamma_N$ in $\mathcal{O}(dMN^3)$ time, where $\varepsilon_N, \gamma_N \rightarrow 0$.*

The idea of proving Theorems 1 and 2 is to estimate the probability of events $F_i(\mathcal{M}_{\mathcal{A}_i}) \geq (1 + \frac{1}{\delta t})F_i(\mathcal{M}_i^*)$ in the case when the multiset \mathcal{T} contains at least t elements of the optimal solution \mathcal{M}_i^* , where $\delta \in \mathbb{R}$, $t \in \mathbb{N}$, $i = 1, 2$. To do this, we use the Markov inequality. Then, using Chernov's inequality, we show that it is sufficient to put $\delta = \gamma/2$, $t = \lceil 2/(\gamma\varepsilon) \rceil$ in Theorem 1 and $\delta = (\log_2 N)^{-1/2}$, $t = \lceil kM/(2N) \rceil$ in Theorem 2.

5 Conclusion

In the present paper, we have proposed randomized algorithms for two sequence clustering problems. The algorithms find $(1 + \varepsilon)$ -approximate solutions with probability not less than $1 - \gamma$ in $\mathcal{O}(dMN^2)$ time. Conditions are found under which the algorithms are polynomial and asymptotically exact.

In our opinion, the algorithms presented in this paper can be used to quickly obtain solutions to large-scale applied problems of signal analysis and recognition.

Acknowledgments. The study of Problem 1 was supported by the Russian Foundation for Basic Research, projects 19-01-00308 and 19-07-00397, 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. The study of Problem 2 was supported by the Russian Foundation for Basic Research, project 19-31-90031.

References

1. Aggarwal, H., Imai, N., Katoh, N., Suri, S.: Finding K points with minimum diameter and related problems. *J. Algorithms* **12**(1), 38–56 (1991)
2. Kel'manov, A.V., Panasenko, A.V., Khandeev, V.I.: Randomized algorithms for some hard-to-solve problems of clustering a finite set of points in Euclidean space. *Comput. Math. Math. Phys.* **59**(5), 842–850 (2019). <https://doi.org/10.1134/S0965542519050099>
3. Kel'manov, A.V., Pyatkin, A.V.: On the complexity of some problems of choosing a subsequence of vectors. *Zh. Vych. Mat. Mat. Fiz.* (in Russian) **52**(12), 2284–2291 (2012)
4. Kel'manov, A.V., Romanchenko, S.M., Khamidullin, S.A.: Approximation algorithms for some intractable problems of choosing a vector subsequence. *J. Appl. Indu. Math.* **6**(4), 443–450 (2012)
5. Kel'manov, A.V., Romanchenko, S.M., Khamidullin, S.A.: Exact pseudopolynomial-time algorithms for some intractable problems of finding a subsequence of vectors. *Zh. Vych. Mat. Mat. Fiz.* (in Russian) **53**(1), 143–153 (2013)
6. Kel'manov, A.V., Romanchenko, S.M., Khamidullin, S.A.: An approximation scheme for the problem of finding a subsequence. *Numer. Anal. Appl.* **10**(4), 313–323 (2017). <https://doi.org/10.1134/S1995423917040036>
7. Kel'manov, A.V., Khamidullin, S.A.: Posterior detection of a given number of identical subsequences in a quasi-periodic sequence. *Comput. Math. Math. Phys.* **41**(5), 762–774 (2001)
8. Panasenko, A.: A PTAS for one cardinality-weighted 2-clustering problem. In: Khachay, M., Kochetov, Y., Pardalos, P. (eds.) *MOTOR 2019*. LNCS, vol. 11548, pp. 581–592. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22629-9_41
9. Sahni, S., Gonzalez, T.: P-complete approximation problems. *J. ACM* **23**, 555–566 (1976)
10. de la Vega, F., Karpinski, M., Kenyon, C., Rabani, Y.: Polynomial time approximation schemes for metric min-sum clustering. *Electronic Colloquium on Computational Complexity (ECCC)*. Report No. 25 (2002)
11. Kel'manov, A., Khamidullin, S., Panasenko, A.: Exact algorithm for one cardinality-weighted 2-partitioning problem of a sequence. In: Matsatsinis, N.F., Marinakis, Y., Pardalos, P. (eds.) *LION 2019*. LNCS, vol. 11968, pp. 135–145. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38629-0_11
12. Kel'manov, A., Khamidullin, S., Panasenko, A.: 2-approximation polynomial-time algorithm for a cardinality-weighted 2-partitioning problem of a sequence. In: Sergeev, Y.D., Kvasov, D.E. (eds.) *NUMTA 2019*. LNCS, vol. 11974, pp. 386–393. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-40616-5_34