Approximation Algorithms for Problems of Partitioning a Sequence Into Clusters

A. Kel'manov, L. Mikhailova, S. Khamidullin, V. Khandeev

Sobolev Institute of Mathematics,
Novosibirsk State University,
Novosibirsk, Russia

Intelligent Data Processing: Theory and Applications (IDP-2016)

Barcelona, Spain,
October 10–14, 2016

Introduction

Subjects of the study

are problems of partitioning a finite sequence of points in Euclidean space into subsequences.

The goal of the study

is to find out the computational complexity of the problems and to provide polynomial-time factor-2 approximation algorithms.

Applications:

problems of approximation, clustering, sequence (time series) analysis.

Problem formulation

Problem 1

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L and M.

Find nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$, i.e. subsets of indices of the elements from the sequence $\mathcal{Y}$, such that

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min,$$

where $\mathcal{M} = \cup_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum_{j \in \mathcal{M}_l} y_j$ is the centroid of subset $\{y_j | j \in \mathcal{M}_l\}$, under the following constraints: (i) the cardinality of $\mathcal{M}$ is equal to M, (ii) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order, (iii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied:

$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \quad m = 2, \ldots, M.$$

Problem formulation

Problem 2

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, and L.

Find nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$, i.e. subsets of indices of the elements from the sequence $\mathcal{Y}$, such that

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min,$$

where $\mathcal{M} = \cup_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum_{j \in \mathcal{M}_l} y_j$ is the centroid of subset $\{y_j | j \in \mathcal{M}_l\}$, under the following constraints: (i) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order, (ii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied:

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

where the cardinality M of $\mathcal{M}$ is not known in advance.

Interpretation

Interpretation

There is a table with the results of the chronologically ordered measurements of a tuple of numerical characteristics of some object. The object can be in either a passive state or in one of some active states.

Interpretation

It is assumed that:

1) in the passive state all the numerical characteristics in the tuple equal zero, while, in each active state the value of at least one characteristic is nonzero;

2) the data contains some measurement errors;

3) the correspondence of the sequence element to some state of the object is not known in advance;

4) all the active states of the object are accompanied by a switching into the passive state for some unknown time interval which is bounded from above and below.

Interpretation

Interpretation

It is required

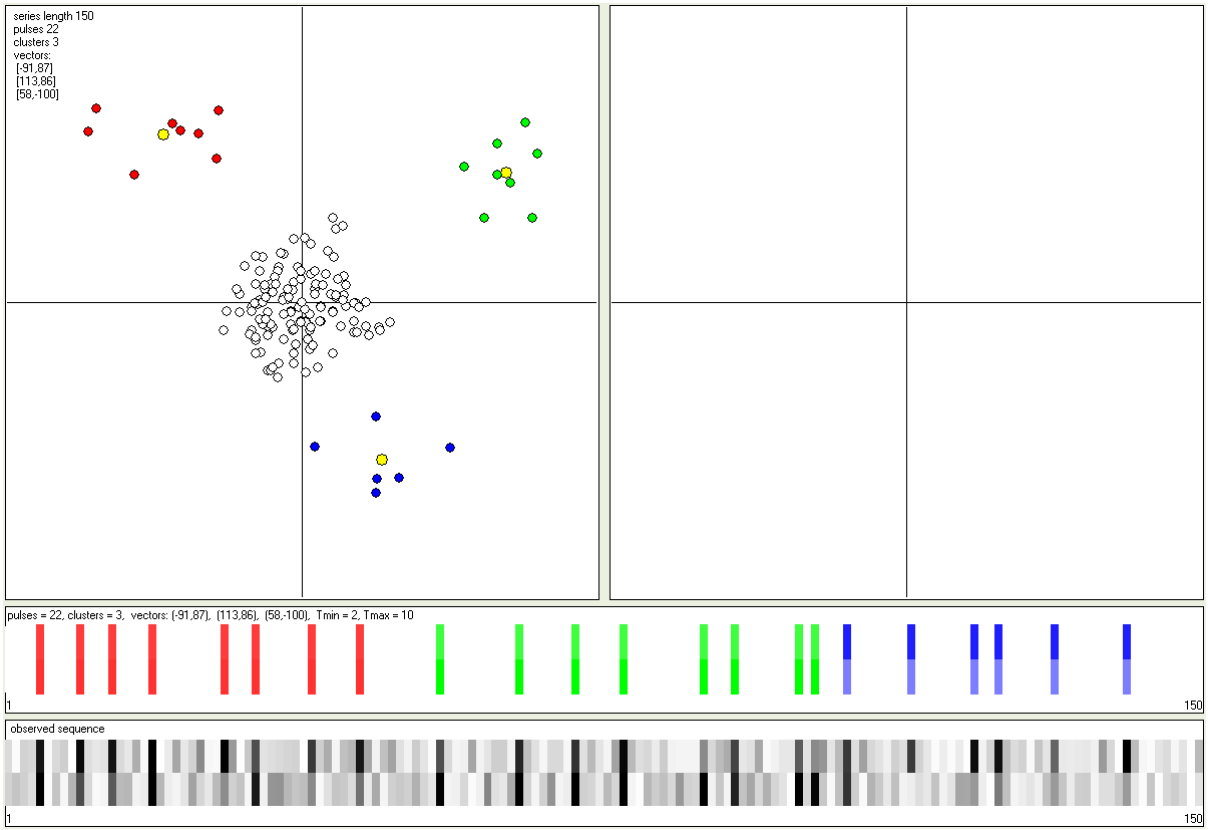
1) to find the sequence of active states of the object;

2) to estimate the characteristics of the object in each of the active states.

Interpretation

Example

150 results of the measurements of a tuple of numerical characteristics of some object.
22 measurements correspond to one of three active states; 128 measurements correspond to a passive state.



Known results

Except for the special case with $L = 1$, no algorithms with guaranteed approximation factor are known at the moment for Problems 1 and 2.

Problem 1: known results (for $L = 1$)

1. The variant of Problem 1 in which $T_{\min}$ and $T_{\max}$ are the parameters: (Kel'manov, Pyatkin, 2013):
(1) the problem is strongly NP-hard for any $T_{\min} < T_{\max}$;
(2) the problem is solvable in polynomial time when $T_{\min} = T_{\max}$.

2. A 2-approximation polynomial-time algorithm running in $\mathcal{O}(N^2(MN + q))$ time was presented (Kel'manov, Khamidullin, 2013).

Known results

Problem 1: known results (for $L = 1$)

3. For the case of integer inputs and fixed space dimension q an exact pseudopolynomial algorithm was constructed. The time complexity of this algorithm is $\mathcal{O}(MN^2(MD)^q)$, where D is the maximum absolute in any coordinate of the input points. (Kel'manov, Khamidullin, Khandeev, 2015).

4. For the case of fixed space dimension a fully polynomial-time approximation scheme was proposed which, given a relative error ε, finds a $(1+\varepsilon)$-approximate solution of the problem in $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ time (Kel'manov, Khamidullin, Khandeev, 2016).

Known results

Problem 2: known results (for $L = 1$)

1. The variant of Problem 1 in which $T_{\min}$ and $T_{\max}$ are the parameters: (Kel'manov, Pyatkin, 2013):
(1) the problem is strongly NP-hard for any $T_{\min} < T_{\max}$;
(2) the problem is solvable in polynomial time when $T_{\min} = T_{\max}$.

2. A 2-approximation polynomial-time algorithm running in $\mathcal{O}(N^2(N + q))$ time was presented (Kel'manov, Khamidullin, 2015).

Obtained results

Main results of this paper

1. An algorithm is proposed that allows to find a 2-approximate solution of Problem 1 in $\mathcal{O}(LN^{L+1}(MN+q))$ time, which is polynomial if L is fixed.

2. An algorithm is proposed that allows to find a 2-approximate solution of Problem 2 in $\mathcal{O}(LN^{L+1}(N+q))$ time, which is polynomial if L is fixed.

The complexity status

The next statement establishes the complexity status of Problems 1 and 2.

Proposition

Problems 1 and 2 are strongly NP-hard.

This proposition follows from the fact that the special cases of Problems 1 and 2 with $L = 1$ are strongly NP-hard (Kel'manov, Pyatkin, 2013).

Problem 1. The approach

The approach to Problem 1

1. For each ordered set (tuple) containing L elements of the sequence $\mathcal{Y}$, we find an exact solution of the auxiliary problem, i.e. a family containing disjoint subsets of indices of the input sequence, which is a feasible solution of the original Problem 1. The found family of subsets we declare a solution candidate for Problem 1 and include this family in the set of solution candidates.

2. From the obtained set as the final solution we choose a family of subsets which yields the largest value for the objective function of the auxiliary problem.

Auxiliary problem

From now on we use $f^x(y)$ to denote a function $f(x, y)$ for which x is fixed. Moreover, let

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - x_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2,$$

$$G(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l \rangle - \|x_l\|^2),$$

where $x_1, \ldots, x_L$ are points from $\mathbb{R}^q$, and elements of the sets $\mathcal{M}_l, \ldots, \mathcal{M}_L$, and $\mathcal{M}$ satisfy restrictions of Problem 1.

Lemma 1

1. For any nonempty fixed subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ the minimum of function S over $x_1, \ldots, x_L$ is reached at the points $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$, and is equal to $F(\mathcal{M}_1, \ldots, \mathcal{M}_L)$.

Auxiliary problem

From now on we use $f^x(y)$ to denote a function $f(x, y)$ for which x is fixed. Moreover, let

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - x_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2,$$

$$G(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l \rangle - \|x_l\|^2),$$

where $x_1, \ldots, x_L$ are points from $\mathbb{R}^q$, and elements of the sets $\mathcal{M}_l, \ldots, \mathcal{M}_L$, and $\mathcal{M}$ satisfy restrictions of Problem 1.

Lemma 1 (continued)

2. For any tuple $x = (x_1, \ldots, x_L)$ of fixed points from $\mathbb{R}^q$ the minimum of function $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ over $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is reached at the subsets $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ that maximize function $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$.

Auxiliary problem

Let

$$g_l^x(n) = 2\langle y_n, x_l\rangle - \|x_l\|^2, \quad n \in \mathcal{N}, \quad l = 1, \ldots, L,$$

where x_l is a point from tuple x, and y_n is an element of sequence $\mathcal{Y}$. In accordance with this definition, we have

$$G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{n \in \mathcal{M}_l} g_l^x(n).$$

Problem 3

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ and a tuple $x = (x_1, \ldots, x_L)$ of points from $\mathbb{R}^q$, and some positive $T_{\min}$, $T_{\max}$ and M.
Find: nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ that maximize the objective function $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$, under the same constraints on the optimized variables as in Problem 1.

For solving this problem the following dynamic programming scheme is justified.

Auxiliary problem. The algorithm

Lemma 2

For any positive integers L and M such that $(M-1)T_{\min} < N$ and $L \le M$, the optimal value $G^x_{\max}$ of the objective function of Problem 3 is given by the formula

$$G^x_{\max} = \max_{n \in \{1+(M-1)T_{\min}, \ldots, N\}} G^x_{L,M}(n);$$

here, the values of $G^x_{L,M}(n)$ are calculated using the recurrence formula

$$G^x_{l,m}(n) = g^x_l(n) + \begin{cases} 0, & \text{if } l = 1,\ m = 1, \\ \max\limits_{j \in \gamma_{m-1}(n)} G^x_{1,m-1}(j), & \\ & \text{if } l = 1,\ m = 2, \ldots, M-(L-1), \\ \max\limits_{j \in \gamma_{m-1}(n)} G^x_{l-1,m-1}(j), & \\ & \text{if } l = 2, \ldots, L,\ m = l, \\ \max\{\max\limits_{j \in \gamma_{m-1}(n)} G^x_{l,m-1}(j), \max\limits_{j \in \gamma_{m-1}(n)} G^x_{l-1,m-1}(j)\}, & \\ & \text{if } l = 2, \ldots, L, \\ & \quad m = l+1, \ldots, M-(L-l), \end{cases}$$

Auxiliary problem. The algorithm

Lemma 2 (continued)

where

$$\gamma_{m-1}(n) = \left\{ j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \le j \le n - T_{\min} \right\}, \quad m = 2, \ldots, M,$$

for every $n = 1 + (m-1)T_{\min}, \ldots, N - (M-m)T_{\min}$.

Auxiliary problem. The algorithm

Algorithm $\mathcal{A}_1$

Input: sequence $\mathcal{Y}$, tuple $(x_1, \ldots, x_L)$ of points, numbers $T_{\min}$, $T_{\max}$, and M.

Step 1. Compute the values $g_l^x(n)$ for $l = 1, \ldots, L$, $n = 1 + (l-1)T_{\min}, \ldots, N - (L-l)T_{\min}$.

Step 2. Using the recurrence formulae, compute the values $G_{l,m}^x(n)$ for each $l = 1, \ldots, L$, $m = l, \ldots, M - (L-l)$, $n = 1 + (m-1)T_{\min}, \ldots, N - (M-m)T_{\min}$.

Step 3. Find the maximum $G_{\max}^x$ of the objective function G^x, and the optimal subsets $\mathcal{M}_l^x$.

Output: the family $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of subsets.

Theorem 1

Algorithm $\mathcal{A}_1$ finds the optimal solution of Problem 3 in $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$ time.

Problem 1. The algorithm

Algorithm $\mathcal{A}$

Input: sequence $\mathcal{Y}$, numbers $T_{\min}$, $T_{\max}$, M, and L.

Step 1. For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using Algorithm $\mathcal{A}_1$, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of Problem 3.

Step 2. Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$. If the optimum is taken by several tuples, we choose any of them.

Output: the family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ of subsets.

Problem 1. The algorithm

Lemma 3

Let $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ be the optimal solution of Problem 1, and $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ be the solution found by Algorithm $\mathcal{A}$. Then

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*).$$

Theorem 2

Algorithm $\mathcal{A}$ finds a 2-approximate solution of Problem 1 in $\mathcal{O}(LN^{L+1}(M(T_{\max} - T_{\min} + 1) + q))$ time. The performance guarantee 2 of the algorithm is tight.

Remark

In the expression of the time complexity of Algorithm $\mathcal{A}$, the value of $(T_{\max} - T_{\min} + 1)$ is at most N. Therefore, the running time of the algorithm is $\mathcal{O}(LN^{L+1}(MN + q))$, which is polynomial if L is fixed.

Conclusion

In this paper we have shown the strong NP-hardness of two problems of partitioning a finite sequence of points of Euclidean space into clusters. We also have shown approximation algorithms for these problems. The proposed algorithms allow to find 2-approximate solutions of the problems in a polynomial time if the number of clusters is fixed.

In our opinion, the presented algorithms would be useful as tools for solving problems in applications related to data mining, and analysis and recognition of time series (signals).

Of considerable interest is the development of faster polynomial-time approximation algorithms for the cases when the number of clusters is not fixed. An important direction of study is searching subclasses of these problems for which faster polynomial-time approximation algorithms can be constructed.

Thank you for your attention!