A Randomized Algorithm for Two-Cluster Partition of a Sequence

Alexander Kel'manov[1,2], Sergey Khamidullin[1], 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,kham,khandeev}@math.nsc.ru

In the paper we consider the following strongly NP-hard [1]

Problem. *Given* a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, and some positive integer numbers $T_{\min}$, $T_{\max}$, and M. *Find* a subset $\mathcal{M} = \{n_1, \ldots, n_M\}$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

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

where $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$, under constraints

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

on the elements of $(n_1, \ldots, n_M)$.

This problem is important, for example, in time series analysis, data mining, machine learning, and noise-proof clusterization of signals.

In this work, we present a randomized algorithm for the problem. Under assumption $M \geq \beta N$, where $\beta \in (0, 1)$ is some constant, and given $\varepsilon > 0$ and $\gamma \in (0, 1)$, the algorithm finds a $(1 + \varepsilon)$-approximate solution of the problem with probability not less than $1 - \gamma$ in $\mathcal{O}(qMN^2)$ time. The conditions are found under which the algorithm finds a $(1 + \varepsilon_N)$-approximate solution of the problem with probability not less than $1 - \gamma_N$, where $\varepsilon_N \to 0$ and $\gamma_N \to 0$ as $N \to \infty$, in $\mathcal{O}(qMN^3)$ time.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research, project nos. 15-01-00462, 16-31-00186, and 16-07-00168.

References

1. Kel'manov, A.V., Pyatkin, A.V.: On Complexity of Some Problems of Cluster Analysis of Vector Sequences. J. of Applied and Industrial Mathematics. 2013. Vol. 7(3). P. 363–369.