Complexity of Normalized K-Means Clustering Problems

Alexander Ageev

Sobolev Institute of Mathematics,
4 Koptyug Ave., 630090 Novosibirsk, Russia
ageev@math.nsc.ru

We study the computational complexity of the following two clustering problems.

Problem 1 (*Normalized K-Means Clustering*). Given a set $\mathcal{Y}$ of N points in $\mathbb{R}^d$ and a positive ineger $K \geq 2$, find a partition of $\mathcal{Y}$ into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ minimizing

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

where $\overline{y}(\mathcal{C}_k)$ is a centroid of cluster $\mathcal{C}_k$.

Problem 2 (*Normalized K-Means clustering with a given center*). Given a set $\mathcal{Y}$ of N points in $\mathbb{R}^d$ and a positive ineger $K \geq 2$, find a partition of $\mathcal{Y}$ into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ minimizing

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

where $\overline{y}(\mathcal{C}_k)$ is a centroid of cluster $\mathcal{C}_k$.

The problems are important, in particular, in applied statistics, data mining and machine learning.

The complexity status of the problems seemed to be unclear up to now. In this paper we prove that Problem 1 is strongly NP-hard for each fixed $K \geq 3$ and Problem 2 is strongly NP-hard for each fixed $K \geq 4$.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research (project 15-01-00462).