1/2-Approximation Polynomial-Time Algorithm for a Problem of Searching for a Subset

A. Ageev, A. Kel'manov, A. Pyatkin, S. Khamidullin, V. Shenmaier

Sobolev Institute of Mathematics
Siberian Branch of the Russian Academy of Sciences,
Novosibirsk State University,
Novosibirsk, Russia

XIIIth International Conference
«**Problems of complex systems' optimization**»
Novosibirsk, September 18-23, 2017

Outline

Introduction (subject, goal and motivation of the investigation)

1. Problem formulation and related results

2. Problem complexity, NP-hardness

3. Approximation algorithm

Conclusion, open problems

Introduction

The subject of investigation is

one recently arised quadratic Euclidean clustering problem.

The goal is

to analyze the computational complexity of this problem and construct an algorithm for it solution.

The research is motivated by

poor research record on the problem and its relevance to many applications, in particular, to

(1) Geometric, approximation and statistical problems;

(2) Data clustering, Data mining, Machine learning, Big data;

(3) Applied problems in technical and medical diagnostics, remote monitoring, biometrics, bioinformatics, econometrics, criminology, processing of experimental data, processing and recognition of signals, etc.

One of the well-known (Fisher, 1958) data analysis problems is the MSSC (Minimum Sum-of-Squares Clustering) problem which is strongly NP-hard (Aloise D., Deshpande A., Hansen P., Popat P., 2009) and has the following formulation.

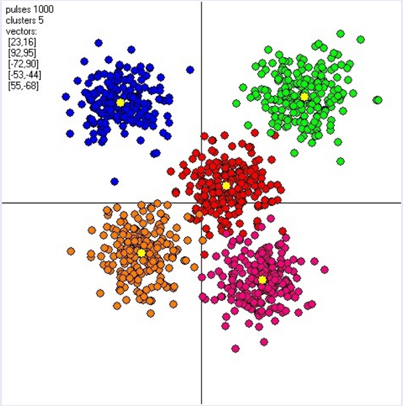
MSSC Problem (Minimum Sum-of-Squares Clustering)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and positive integer $J > 1$.
Find a partition of $\mathcal{Y}$ into non-empty clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_J\}$ such that

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of $\mathcal{C}_j$.

MSSC Problem. Two-dimensional example



1. Problem formulation, interpretation, closely related problems, known and our results

Problem 1 (Subset of points with the largest cardinality under a constraint on the total quadratic variation)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and number $\alpha \in (0, 1)$.
Find a subset $\mathcal{C} \subset \mathcal{Y}$ with the largest cardinality such that

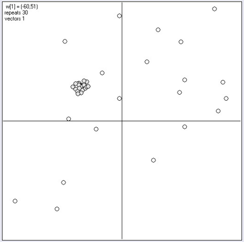
$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \leq \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ is the centroid (the geometrical center) of the subset $\mathcal{C}$, and $\overline{y}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} y$ is the centroid of the input set.

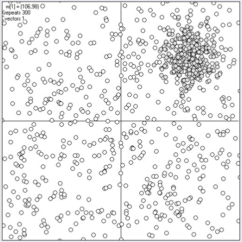
1. Problem formulation, interpretation, closely related problems, known and our results

Two-dimensional examples of the input sets

Example 1



Example 2



The problem has a simple interpretation, namely, searching for the largest by cardinality subset $\mathcal{C}$ of points, whose total quadratic deviation from the unknown centroid $\overline{y}(\mathcal{C})$ doesn't exceed the total quadratic deviation of the input set $\mathcal{Y}$ from its centroid $\overline{y}(\mathcal{Y})$ multiplied by α.

1. Problem formulation, interpretation, closely related problems, known and our results

Currently, there are no available algorithmic results for Problem 1. Some results are known for closest problem related to Problem 1, that is for

Problem 2 (M-Variance problem)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and positive integer number $M > 1$.
Find a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M such that

$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \longrightarrow \min$$

Known results for M-Variance problem

1. The strong NP-hardness of the problem (Kel'manov and Pyatkin, 2010).

2. An exact algorithm with $\mathcal{O}(qN^{q+1})$ running time, Aggarwal, Imai, Katoh, Suri (1991), Shenmaier, 2016.

Known results for M-Variance problem

3. A 2-approximation polynomial-time algorithm, $\mathcal{O}(qN^2)$,
Kel'manov, Romanchenko (2012).

4. An exact algorithm for the integer-valued variant of the data input. In the case of fixed space dimension the algorithm has $\mathcal{O}(N(MB)^q)$ running time, where B is the maximum absolute value of the coordinates of the input points,
Kel'manov, Romanchenko (2012).

5. PTAS of complexity $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, where ε is a guaranteed relative error,
Shenmaier (2012).

6. $(1+\varepsilon)$-Approximation algorithm, which implements an FPTAS with $\mathcal{O}(N^2(M/\varepsilon)^q)$-time complexity, in the case of fixed space dimension,
Kel'manov, Romanchenko (2014).

1. Problem formulation, interpretation, closely related problems, known and our results

Our results (Ageev, Kel'manov, Pyatkin, Khamidullin, Shenmaier, 2017)

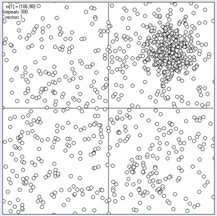
1. Problem 1 is strongly NP-hard.

2. 1/2-approximation polynomial-time algorithm with running time

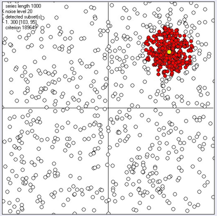
$$\mathcal{O}(N^2(q + \log N)$$

.

example of an input set



found subset



2. Problem complexity

Problem 1A (Problem 1 in a property verification form)

Input: a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$, positive real A and integer M.
Question: is there a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality at least M, such that

$$F(\mathcal{C}) \leq A. \tag{1}$$

Remind that the following problem (Problem 2 in a property verification form) belongs to the class of NP-complete problems in the strong sense.

Problem 2A - M-Variance (Problem 2 in a property verification form)

Input: a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$, a positive integer M, and a positive real B.
Question: is there a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M such that

$$F(\mathcal{C}) \leq B.$$

2. Problem complexity, NP-hardness

Note that the function F has the following property:

$$if \mathcal{C}_1 \subseteq \mathcal{C}_2, \quad then \quad F(\mathcal{C}_1) \leq F(\mathcal{C}_2).$$

Therefore, if in the problem 1A the answer is positive, then there is a subset of cardinality M satisfying the inequality (1).

Thus, problems 1A and 2A are equivalent and obviously we have the following

Statement 1

The problem 1A is NP-complete in the strong sense.

It follows from statement 1 that Problem 1 is an NP-hard problem in the strong sense, that is, it is not easier than Problem 2.

The idea of approximation algorithm

1. For each point y of the input set, a subset consisting of the maximum number of closest to y (in the sense of Euclidean distance) points from the input set is constructed such that the sum of the squares of the distances from y to the points of the subset does not exceed a given threshold (that is the fraction of the quadratic scatter of points of the input set).

2. Among the found subsets the one with the largest cardinality is taken as an output.

3. Approximation algorithm

Algorithm $\mathcal{A}$

Input: set $\mathcal{Y}$ and number $\alpha \in (0,1)$.
Step 1. Compute the value $A = \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2$.

For each point $y \in \mathcal{Y}$ perform steps 2 and 3.
Step 2. Compute the distances from the point y to all points in $\mathcal{Y}$ and sort the set $\mathcal{Y}$ in the nondecreasing order according to these distances. Denote this sequence by $y_1, \ldots, y_N$.

Step 3. Find the subsequence $y_1, \ldots y_M$ of maximum length such that

$$\sum_{i=1}^{M} \|y - y_i\|^2 \leq A.$$

Define the subset $\mathcal{C}^y = \{y_1 \ldots, y_M\}$.

Step 4. In the family $\{\mathcal{C}^y | \, y \in \mathcal{Y}\}$ of admissible subsets constructed in step 3 choose as the output $\mathcal{C}_A$ any subset $\mathcal{C}^y$ of the largest cardinality.

Output: subset $\mathcal{C}_A$.

To justify the accuracy bound for this algorithm, we need two facts.

Statement 2

Let a sequence $0 \leq a_1 \leq \ldots \leq a_k$ and a positive number $\beta \leq 1$ be given. Then, $g(\lfloor k\beta \rfloor) \leq \beta g(k)$, where $g(i) = a_1 + \ldots + a_i$, and $g(0) = 0$.

Proof. Let $m = \lfloor k\beta \rfloor$. Then, since the sequence a_i does not decrease, we have

$$g(k) = g(m) + \sum_{i=m+1}^{k} a_i \geq g(m) + (k-m)a_{m+1}$$

$$\geq g(m) + \frac{k-m}{m} g(m) = g(m)\frac{k}{m} \geq \frac{g(m)}{\beta}.$$

Recall that in Problem 1,

$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2, \quad \mathcal{C} \subseteq \mathcal{Y} \subset \mathbb{R}^q.$$

Put

$$f(x, \mathcal{Z}) = \sum_{y \in \mathcal{Z}} \|y - x\|^2, \quad x \in \mathbb{R}^q, \; \mathcal{Z} \subset \mathbb{R}^q.$$

The following statement is well-known

Lemma 1

Let $\overline{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ be the centroid of the finite set $\mathcal{Z} \subset \mathbb{R}^q$, let a point $x \in \mathbb{R}^q$ satisfy the condition $\|x - \overline{z}\| \leq \|z - \overline{z}\|$ for every $z \in \mathcal{Z}$. Then

$$F(\mathcal{Z}) \leq f(x, \mathcal{Z}) \leq 2F(\mathcal{Z}).$$

Theorem 1

Algorithm finds a 1/2-approximate solution of Problem 1 in $\mathcal{O}(N^2(q + \log N)$ time.

Proof. Let $\mathcal{C}^*$ be the cluster of the maximal cardinality (in Problem 1) and $\overline{y}(\mathcal{C}^*)$ be the centroid of $\mathcal{C}^*$. Let y be the point from $\mathcal{C}^* \subseteq \mathcal{Y}$, closest to $\overline{y}(\mathcal{C}^*)$.

Then, by Lemma 1 we have

$$f(y, \mathcal{C}^*) \leq 2F(\mathcal{C}^*) \leq 2A.$$

Further, let $\mathcal{C} = \mathcal{C}^*$ if $|\mathcal{C}^*|$ is even; let $\mathcal{C} = \mathcal{C}^* \setminus \{y\}$ otherwise. Note that $f(y, \mathcal{C}) = f(y, \mathcal{C}^*)$ in any case.

Proof of Theorem 1

In the conditions of Statement 2, let $k = |\mathcal{C}|$, $\beta = 1/2$, and choose as a_i, $i = 1, \ldots, k$, the squares of the distances from point y to points $y_i \in \mathcal{C}$. Note that $g(k) = f(y, \mathcal{C})$ and $\lfloor k/2 \rfloor = k/2$ because k is even.

Denote by $\mathcal{C}'$ a cluster composed of the $k/2$ closest to y points from $\mathcal{C}$. Let $\mathcal{C}_0 = \mathcal{C}' \cup \{y\}$. Then $|\mathcal{C}_0| \geq M^*/2$, and, in this case,

$$f(y, \mathcal{C}_0) = g(k/2) \leq g(k)/2 = f(y, \mathcal{C}^*)/2 \leq A$$

by Statement 2; i.e., $\mathcal{C}_0$ is an admissible solution of Problem 1 with a cluster of cardinality $M^*/2$.

But then the condition $M \geq M^*/2$ holds also for the cluster $\mathcal{C}^y$ consisting of the maximal number M of the closest points to y and satisfying the inequality $f(y, \mathcal{C}^y) \leq A$.

Proof of Theorem 1

It remains to note that at Step 4 in the collection $\{\mathcal{C}^y | \, y \in \mathcal{Y}\}$, the closest point y to the centroid of the optimal cluster, and the subset corresponding to it, will be clearly considered. Consequently, the solution found by the algorithm $\mathcal{A}$ contains at least $M^*/2$ elements, and is a $1/2$-approximate solution of Problem 1.

Let us estimate the **time complexity** of the algorithm.

Step 1 requires $\mathcal{O}(qN)$ operations.

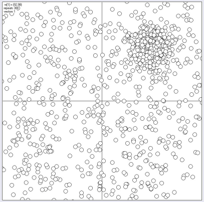
For each point y, **Steps 2, 3** need $\mathcal{O}(qN + N \log N)$ time, where $\mathcal{O}(N \log N)$ is the sorting complexity,

Step 4 is performed in $\mathcal{O}(N)$ time.

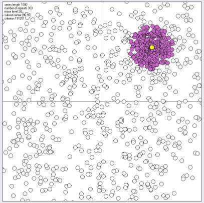
Therefore, the time complexity of the algorithm is $O(N^2(q + \log N))$.

Numerical simulation. Examples

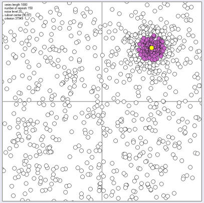
example of an input set, 1000 points



found subset, 303 points, $\alpha = 0.01$



found subset, 150 points, $\alpha = 0.002$



Thank you for your attention!