An Approximation Polynomial-Time Algorithm for a Cardinality-Weighted 2-clustering Problem

Alexander Kel'manov, Anna Motkova

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

XIII International Asian School-seminar
"Problems of complex systems' optimization"
in the scope of International multi-conference IEEE SIBIRCON 2017
Novosibirsk, Russia
September 18 - 23, 2017

Outline

1. Introduction:
   1. The problem formulation
   2. Some closely related problems
   3. Existing and our reached results
2. Approximation algorithm
3. Conclusion

Introduction

The subject of the study

is a problem of partitioning a finite set of points in Euclidean space into two subsets (clusters).

The goal of our study

is to substantiate an approximation algorithm for this problem.

Motivation and applications:

Our research is motivated by

(1) insufficient study of the problem into an algorithmic direction and

(2) its importance for some applications including, for example, data mining, data clustering, pattern recognition, machine learning, approximation, statistical problems of joint evaluation and hypotheses testing with heterogeneous samples.

Problem 1. Cardinality-weighted variance-based 2-clustering with given center

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$, a positive integer $M \leq N$.

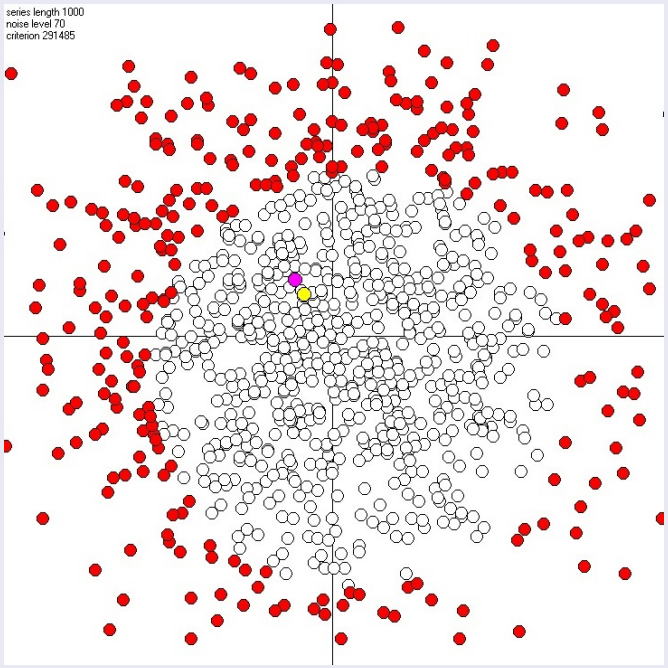
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$F(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min, \qquad (1)$$

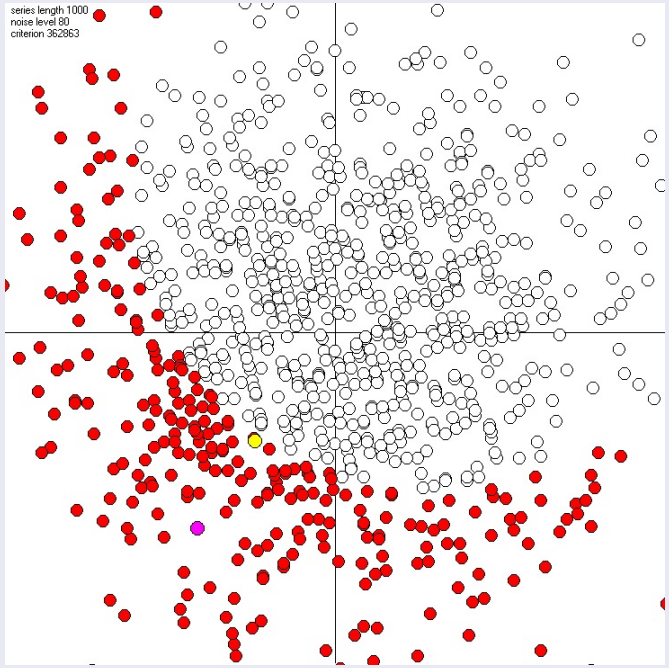
where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ is the geometric center (centroid) of $\mathcal{C}$ and such that $|\mathcal{C}| = M$.

Introduction: The problem formulation

2-dimensional example



2-dimensional example



Introduction: Closely related problems (without given center)

The well-known strongly NP-hard (Aloise D., Deshpande A., Hansen P., Popat P., **2009**):

Minimum sum-of-squares 2-clustering (2-MSSC)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \longrightarrow \min,$$

where $\overline{y}(\mathcal{C})$ and $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ are the centroids of $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$.

This problem is related to classical work by Fisher (1958) and also called 2-*Means*.

Thousands of publications are dedicated to this problem and its applications.

Introduction: Closely related problems (without given center)

Cardinality-Weighted variance-based 2-clustering (Cardinality-Weighted 2-MSSC)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$|\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \longrightarrow \min,$$

where $\overline{y}(\mathcal{C})$ and $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ are the centroids of clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$.

This problem is NP-hard and equivalent to

Min-sum all-pairs 2-clustering

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$\sum_{x \in \mathcal{C}} \sum_{z \in \mathcal{C}} \|x - z\|^2 + \sum_{x \in \mathcal{Y} \setminus \mathcal{C}} \sum_{z \in \mathcal{Y} \setminus \mathcal{C}} \|x - z\|^2 \longrightarrow \min .$$

Introduction: Closely related problems (without given center)

The most important results for this problems were presented in:

Cardinality-Weighted variance-based 2-clustering (Cardinality-Weighted 2-MSSC)

1. Sahni S., GonzalezT.: P-Complete Approximation Problems, 1976
2. Brucker P.: On the Complexity of Clustering Problems, 1978
3. Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J.: Efficient Algorithms for Variance-Based k-Clustering, 1993
4. Inaba M., Katoh N., Imai H.: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering: (extended abstract), 1994
5. de la Vega F., Kenyon C.: A Randomized Approximation Scheme for Metric Max-Cut, 2001
6. de la Vega F., Karpinski M., Kenyon C., Rabani Y.: Polynomial Time Approximation Schemes for Metric Min-Sum Clustering, 2002
7. Kel'manov A.V., Pyatkin A.V.: NP-Hardness of Some Quadratic Euclidean 2-clustering Problems, 2015

Introduction: Closely related problems (without given center)

M-variance

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

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

where $\overline{y}(\mathcal{C})$ is the centroid of $\mathcal{C}$.

Aggarval, Imai, Katoh, Suri (exact algoritm, 1991),

Kel'manov, Pyatkin (strong NP-hardness, 2010),

Kel'manov, Romanchenko:
- 2-approximation polynomial-time algorithm, 2011,
- exact pseudopolynomial-time algorithm for the special case, 2012,
- FPTAS for the special case, 2014,

Shenmaier:
PTAS, 2012, exact algorithm, 2016.

Introduction: Closely related problems (without given center)

Normalized variance-based 2-clustering (Normalized 2-MSSC)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$\frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \longrightarrow \min,$$

where $\overline{y}(\mathcal{C})$ and $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ are the centroids of clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$.

Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J. (1993, 1994)

The complexity status of the problems seemed to be unclear up to now.

Introduction: Existing and our reached results for Problem 1

Existing results

1. The strong NP-hardness of the problem was proved. (Kel'manov, Pyatkin, **2015**).

2. An exact algorithm for the case of integer components of the input points was presented. If the dimension q of the space is bounded by a constant, then this algorithm has a pseudopolynomial $\mathcal{O}(N(MB)^q)$-time complexity, where B is the maximum absolute value of the coordinates of the input points. (Kel'manov, Motkova, **2015**)

3. An approximation algorithm was presented. It allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qN^2(\sqrt{\frac{2q}{\varepsilon}}+2)^q)$ time for a given relative error ε. If the space dimension q is bounded by a constant this algorithm implements a fully polynomial-time approximation scheme with $\mathcal{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$-time complexity. (Kel'manov, Motkova, **2016**)

Introduction: Existing and our reached results for Problem 1

Our new result (Kel'manov, Motkova)

We present an approximation algorithm. It allows to find a 2-approximate solution of the problem in $\mathcal{O}\left(qN^2\right)$ time.

Approximation algorithm

Here and below $\mathcal{B}^x$ is a subset of $\mathcal{Y}$ with M smallest values of the function

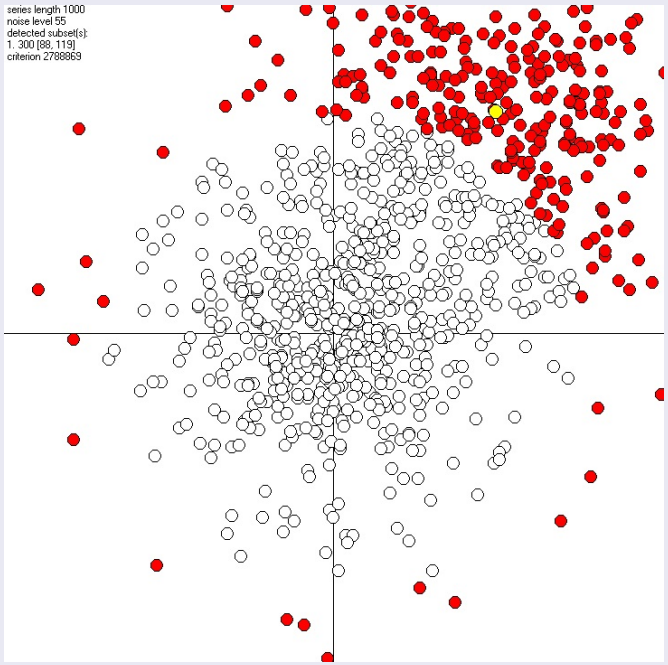
$$g^x(y) = (2M - N)\,\|y\|^2 - 2M\,\langle y, x\rangle\,,\ \ y \in \mathcal{Y}\,. \tag{2}$$

Algorithm $\mathcal{A}$

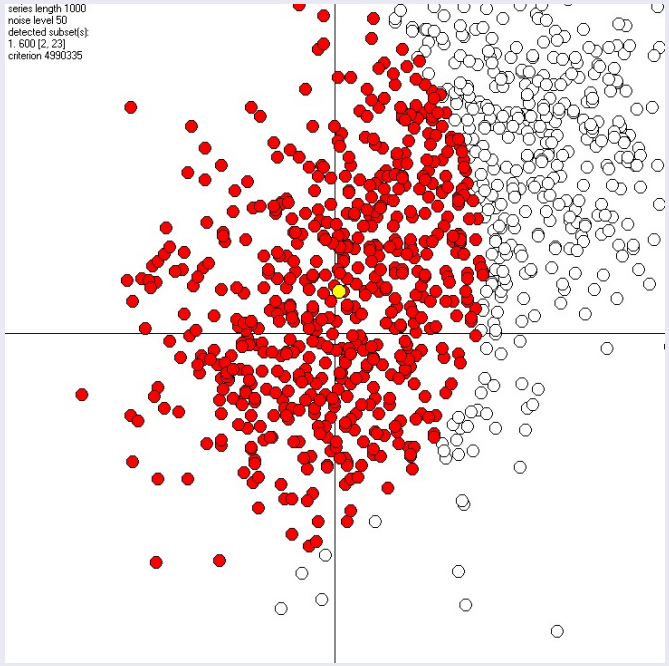
Input: N-elements set $\mathcal{Y} \subset \mathbb{R}^q$, natural number $M \le N$.
For each point $y \in \mathcal{Y}$ Steps 1–2 are executed.
Step 1. Compute the values $g^y(z),\ z \in \mathcal{Y}$, using formula (2); find an M-elements subset $\mathcal{B}^y \subseteq \mathcal{Y}$ with the smallest values $g^y(z)$, compute $F(\mathcal{B}^y)$ using formula (1).
Step 2. If $F(\mathcal{B}^y) = 0$, then put $\mathcal{C}_{\mathcal{A}} = \mathcal{B}^y$; exit.
Step 3. In the family $\{\mathcal{B}^y|,\ y \in \mathcal{Y}\}$ of candidate sets that have been constructed in Steps 1–2, choose as a solution $\mathcal{C}_{\mathcal{A}}$ the set $\mathcal{B}^x$, for which $F(\mathcal{B}^x)$ is minimal.
Output: the set $\mathcal{C}_{\mathcal{A}}$.

Examples of an input set and 2-approximate solutions found by Algorithm $\mathcal{A}$

400-elements subset $C_{\mathcal{A}}$



600-elements subset $C_{\mathcal{A}}$



Approximation algorithm

Theorem 1.

An Algorithm $\mathcal{A}$ finds a 2-approximate solution of Problem 1 in time $\mathcal{O}\left(qN^2\right)$.

Conclusion

- We presented a 2-approximation algorithm for a quadratic Euclidian problem of weighted partitioning a finite set of points into two clusters with given center of one cluster. This algorithm is the polynomial one and can be interesting for applications.
- It seems important to continue studying the questions on algorithmical approximability of the problem.

Thank you for your attention!