Some Algorithms With Guaranteed Accuracy for 2-Clustering Problems With Given Center of One Cluster

Alexander Kel'manov, Vladimir Khandeev

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

XIII International Asian School-seminar
"Problems of complex systems' optimization"

September 18-23, 2017, Novosibirsk, Russia

Introduction

The subject of study

are strongly NP-hard problems of partitioning a finite set of points of Euclidean space into two clusters.

The goal of study

is to present a short survey on some results of authors and their colleagues (from Sobolev Institute of Mathematics and Novosibirsk State University) for these problems.

Applications:

computer geometry, statistical analysis of data, pattern recognition.

Problem formulation

Problem 1

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

Find a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

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

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

Problem 2

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

Find a subset $\mathcal{C} \subseteq \mathcal{Y}$ that minimizes the objective function (1) under constraint $|\mathcal{C}| = M$.

Interpretation

Interpretation

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

Interpretation

It is assumed that:

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

2) the data contains some measurement errors;

3) the correspondence of the set element to some state of the object is not known in advance.

Interpretation

Interpretation

It is required:

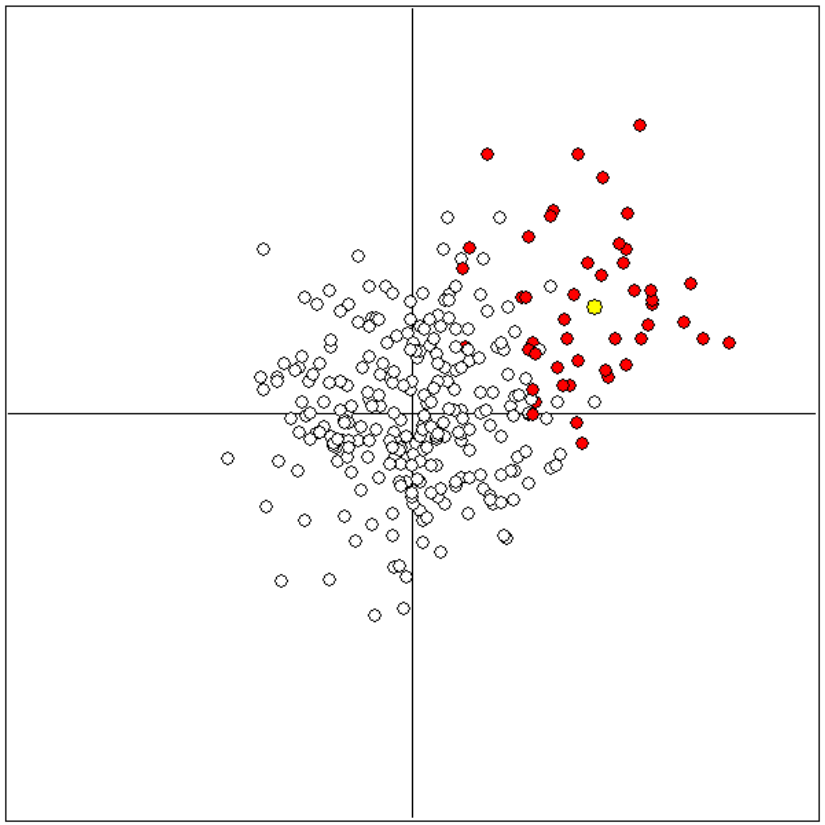
1) to divide the set into the two clusters (subsets) corresponding to the passive and active states of the object;

2) estimate the set of characteristics of the object in the active state.

Interpretation

Example

300 results of the measurements of a tuple of numerical characteristics of some object.
47 measurements correspond to the active state; 253 measurements correspond to the passive state.



Problem 1. Known results

Since

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

Problem 1 is polynomial-time equivalent to

Problem 3

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

Find a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\frac{1}{|\mathcal{C}|}\left\|\sum_{y\in\mathcal{C}} y\right\|^2 \to \max.$$

Problem 1. Known results

NP-hardness

1. Problem 3 is strongly NP-hard. Therefore, the problem admits neither exact polynomial, nor exact pseudopolynomial algorithm unless P=NP (Kel'manov, Pyatkin, 2008–2009).

Exact algorithms

2. Exact algorithm for Problem 3 with $\mathcal{O}(q^2 N^{2q})$ running time (Gimadi, Pyatkin, Rykov, 2010).

3. Exact algorithm for Problem 3 with $\mathcal{O}(qN^{q+1})$ running time (Shenmaier, 2016).

Approximation algorithms

4. An algorithm for Problem 3 which has the guaranteed approximation ratio $\varepsilon = (q-1)/(4L^2)$, where L is an integer parameter of the algorithm. The time complexity of the algorithm is $\mathcal{O}\big(Nq(q+\log N)(2L+1)^{q-1}\big)$ (Kel'manov, Pyatkin, 2009).

Problem 1. Obtained result

Obtained result

A 2-approximation algorithm for Problem 1 with $O(qN^2)$ running time.

The approach

1. For each point $x \in \mathcal{Y}$ construct

$$\mathcal{B}(x) = \{y \in \mathcal{Y} \mid 2(y, x) > \|x\|^2\}.$$

2. As the final solution, choose the tuple for which the value of

$$Q(\mathcal{B}(x), x) = \sum_{y \in \mathcal{B}(x)} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}(x)} \|y\|^2$$

is minimal.

Problem 2. Known results

Since

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

Problem 2 is polynomial-time equivalent to

Problem 4

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

Find a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\left\|\sum_{y\in\mathcal{C}} y\right\| \to \max,$$

under constraint $|\mathcal{C}| = M$.

Problem 2. Known results

NP-hardness

1. Problem 4 is strongly NP-hard. Therefore, the problem admits neither exact polynomial, nor exact pseudopolynomial algorithm unless P=NP (Gimadi, Kel'manov, Kel'manova, Khamidullin, 2006–2008; Baburin, Gimadi, Glebov, Pyatkin, 2008).

Exact algorithms

2. Exact algorithm for Problem 4 with $\mathcal{O}(q^2 N^{2q})$ running time (Gimadi, Pyatkin, Rykov, 2010).

3. Exact algorithm for Problem 4 with $\mathcal{O}(qN^{q+1})$ running time (Shenmaier, 2016).

Problem 2. Known results

Exact algorithms

4. Exact algorithm for the case of Problem 4 with integer input. Running time of the algorithm is $\mathcal{O}(qMN(2MD)^{q-1})$, where D is the maximum absolute coordinate value of the points in the input set (Baburin, Gimadi, Glebov, Pyatkin, 2008).

5. Exact algorithm for the case of Problem 4 with integer input. Running time of the algorithm is $\mathcal{O}(Nq^{q+1}(MD)^{q-1})$, where D is the maximum absolute coordinate value of the points in the input set (Gimadi, Glazkov, Rykov, 2009).

Approximation algorithms

6. A randomized algorithm for Problem 4 which allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qNL)$ time with probability $1-\delta$, where $\varepsilon \le \varphi_0^2/2$, $\delta \le \exp\left(-\frac{\left(7/4 \sin\frac{\varphi_0}{2}\right)^{q-1}}{\pi\sqrt{q}} L\right)$, and L and φ_0 are parameters of the algorithm. Also, parameters of the algorithm were found for which in the case of fixed space dimension the algorithm is asymptotically exact and polynomial (Gimadi, Rykov, 2015).

Problem 2. Known results

Approximation algorithms

7. A 2-approximation algorithm for Problem 2 with $O(qN^2)$ running time (Dolgushev, Kel'manov, 2011).

8. A polynomial-time approximation scheme for Problem 2 with a $O(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$-time complexity, where ε is an arbitrary relative error (Dolgushev, Kel'manov, Shenmaier, 2015).

Problem 2. Obtained results

Obtained results

1. A pseudopolynomial algorithm which finds an optimal solution of Problem 2 in the case of integer components of the points in the input set and fixed space dimension. The running time of the algorithm is $\mathcal{O}(N(MD)^q)$, where D is the maximum absolute coordinate value of the points in the input set.

The approach

1. Construct a multidimensional grid (lattice) $\mathcal{G}$ with rational step uniform with respect to each coordinate. The grid step is chosen so that one of the grid nodes coincide with the geometric center of one of the clusters being optimized.

2. For each node $x \in \mathcal{G}$ find a subset $\mathcal{B}_M(x)$ of M elements of the original set that have the largest projections onto this node.

3. In the family $\{\mathcal{B}_M(x) \mid x \in \mathcal{G}\}$ of subsets choose as the final solution the tuple for which the value of $Q(\mathcal{B}_M(x), x)$ is minimal.

Problem 2. Obtained results

Obtained results

2. Unless P=NP, in the general case of Problem 2 there is no fully polynomial-time approximation scheme (FPTAS).
Such a scheme is presented for the case of fixed space dimension. The running time of the algorithm is $\mathcal{O}(N^3(1/\varepsilon)^{q/2})$, where ε is an arbitrary relative error.

The approach

1. For each point of the input set, construct a multidimensional grid (lattice) with adaptive step and size such that the center of the desired subset necessarily belongs to the domain of one of these grids.

2. For each node x of every grid find a subset $\mathcal{B}_M(x)$ of M elements of the original set that have the largest projections onto this node.

3. In the family of found subsets choose as the final solution the tuple for which the objective function of Problem 2 is minimal.

Problem 2. Obtained results

Obtained results

3. A randomized algorithm for Problem 2. The running time of the algorithm for the fixed failure probability, relative error of the solution and for the certain value of parameter k is $\mathcal{O}(2^k q(k+N))$.
The algorithm is asymptotically exact and has $\mathcal{O}(qN^2)$-time complexity for the special values of the parameters.

The approach

1. A finite multiset $\mathcal{T}$ is formed by random and independent choice (with replacement) of k elements from $\mathcal{Y}$, where k is the parameter of the algorithm.

2. For each nonempty $\mathcal{H} \subseteq \mathcal{T}$ compute the centroid $\overline{y}(\mathcal{H})$ and form a subset of M elements of the original set that have the largest projections onto this centroid.

3. In the family of found subsets choose as the final solution the tuple for which the objective function of Problem 2 is minimal.

Conclusion

In the paper we present a short survey on some results for two problems of partitioning a finite set of points in Euclidean space into two clusters.

A task of much interest is to construct faster algorithms for the problems and to find special cases of the problems for which construction of linear and sublinear randomized algorithms is possible.

Thank you for your attention!