Algorithms with Performance Guarantee for a Weighted 2-partition Problem

Alexander Kel'manov
Sobolev Institute of Mathematics
Acad. Koptyug avenue 4,
Novosibirsk State University
Pirogova str. 1,
630090 Novosibirsk, Russia
kelm@math.nsc.ru

Anna Motkova
Sobolev Institute of Mathematics
Acad. Koptyug avenue 4,
Novosibirsk State University
Pirogova str. 1,
630090 Novosibirsk, Russia
anitamo@mail.ru

Abstract

We consider the problem of partitioning a finite set of Euclidean points into two clusters minimizing the sum over both clusters the weighted sums of the squared intracluster distances from the elements of the clusters to their centers. The center of one of the clusters is unknown and determined as the average value over all points in the cluster, while the center of the other cluster is the origin. The weight factors for both intracluster sums are the cardinalities of the corresponding clusters. In this work, we present a short survey on the results for this problem and a new result: a 2-approximation algorithm.

1 Introduction

In this work, we consider a strongly NP-hard problem of discrete optimization. The goal of our work is to present a short survey on the results for these problems.

The problem under study is closely related to the well-known strongly NP-hard

***Weighted variance-based* 2-*clustering* Problem**

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 $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ and $\overline{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ are the geometric centers (centroids) of both clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$.

The main difference between the problem under study and this problem is that in the considered problem the center of one of the clusters is fixed in the origin (without loss of generality). It is obvious, that the considered problem and the Weighted variance-based 2-clustering problem are not equivalent and so both of them need an individual study. The importance of the problem for applications motivates to continue researches, e.g. importance for geometric problems, approximation problems, statistical problems of joint evaluations and

In: Yu. G. Evtushenko, M. Yu. Khachay, O. V. Khamisov, Yu. A. Kochetov, V.U. Malkova, M.A. Posypkin (eds.): Proceedings of the OPTIMA-2017 Conference, Petrovac, Montenegro, 02-Oct-2017, published at http://ceur-ws.org

testing hypotheses by nonuniform samples, problems of Data clustering, Data mining, Machine learning, Big data, applied problems in technical and medical diagnostics, etc.

2 Cardinality-weighted Variance-based 2-clustering with Given Center

Problem 1. *Given* a set $\mathcal{Y} = \{y_1, \dots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M. *Find* a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$F_1(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min, \tag{1}$$

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

Problem 1 is strongly NP-hard [Kel'manov & Pyatkin, 2015, Kel'manov & Pyatkin, 2016]. Therefore, by [Garey & Johnson, 1979], there are neither exact polynomial-time nor exact pseudopolynomial-time algorithms for this problem, unless P=NP.

Despite of this, a pseudopolynomial algorithm for the special case of Problem 1 exists. It proposed in [Kel'manov & Motkova, 2016a] and finds an optimal solution in the case of integer components of the points in the input set and fixed space dimension.

Algorithm $\mathcal{A}_1$ (exact pseudopolynomial algorithm).

Input: A set $\mathcal{Y}$ and some positive integer $M \le N$.

`Step 1.` Construct $\mathcal{G}$ — a multidimensional cubic uniform in each coordinate grid of size $2D$ with the distance $\frac{1}{M}$ between the nodes and the center at the origin, and D is the maximal modulus of the coordinates of input points.

`Step 2.` For each node $x \in \mathcal{G}$, calculate

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

find the subset $\mathcal{B}^x$ which consists of M points of the set $\mathcal{Y}$, at which the function $g^x(y)$ has the smallest values. Calculate $F_1(\mathcal{B}^x)$ by (1).

`Step 3.` In the family $F_1(\mathcal{B}^x), x \in \mathcal{G}$ constructed at Step 2, find the node $x_{\mathcal{A}_1} = \arg\min\limits_{x \in \mathcal{G}} F_1(\mathcal{B}^x)$ and the subset $\mathcal{B}^{x_{\mathcal{A}_1}}$. Put $\mathcal{C}_{A_1} = \mathcal{B}^{x_{\mathcal{A}_1}}$.

Output: The set $\mathcal{C}_{A_1}$.

Theorem 1. Assume that the elements of $\mathcal{Y}$ have integer values from the interval $[-D, D]$. Then algorithm $\mathcal{A}_1$ finds an optimal solution of Problem 1 in $\mathcal{O}(qN(2MD + 1)^q)$ time.

In the case of fixed space dimension q the algorithm is pseudopolynomial.

In [Kel'manov & Pyatkin, 2015, Kel'manov & Pyatkin, 2016] it is also shown that for Problems 1 there does not exist any fully polynomial time approximation scheme (FPTAS), unless P=NP. In [Kel'manov & Motkova, 2016b] was proposed such approximation scheme for a special case of the Problem 1 when the space dimension q is fixed.

Algorithm $\mathcal{A}_2$ (approximation scheme).

Input: a set $\mathcal{Y}$ and numbers M and ε.

For each point $y \in \mathcal{Y}$ Steps 1–6 are executed.

`Step 1.` Compute the values $g^y(z)$, $z \in \mathcal{Y}$, using formula (2); find a subset $\mathcal{B}^y \subseteq \mathcal{Y}$ with M smallest values $g^y(z)$, compute $F(\mathcal{B}^y)$ using formula (1).

`Step 2.` If $F(\mathcal{B}^y) = 0$, then put $\mathcal{C}_{\mathcal{A}_3} = \mathcal{B}^y$; exit.

`Step 3.` Compute $H = H(y) = \frac{1}{M}\sqrt{F_1(\mathcal{B}^y)}$ and $h = h(y, \varepsilon) = \frac{1}{M}\sqrt{\frac{2\varepsilon}{q} F_1(\mathcal{B}^y)}$.

`Step 4.` Construct the lattice $\mathcal{G}(y, h, H + h/2)$ using formula

$$\mathcal{G}(y, h, H + h/2) = \{d \in \mathbb{R}^q \,|\, d = y + h \cdot (i_1, \dots, i_q), i_k \in \mathbb{Z}, |hi_k| \le H + h/2, k \in \{1, \dots, q\}\}.$$

`Step 5.` For each node x of the lattice $\mathcal{G}(y, h, H + h/2)$ compute the values $g^x(y)$, $y \in \mathcal{Y}$, using formula (2) and find a subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with M smallest values $g^x(y)$. Compute $F_1(\mathcal{B}^x)$ using formula (1), remember this value and the set $\mathcal{B}^x$.

`Step 6.` If $F(\mathcal{B}^x) = 0$, then put $\mathcal{C}_{\mathcal{A}_2} = \mathcal{B}^x$; exit.

Step 7. In the family $\{\mathcal{B}^x \,|\, x \in \mathcal{G}(y, h, H + h/2), y \in \mathcal{Y}\}$ of candidate sets that have been constructed in Steps 1–6, choose as a solution $\mathcal{C}_{\mathcal{A}_2}$ the set $\mathcal{B}^x$ for which $F_1(\mathcal{B}^x)$ is minimal.

Output: the set $\mathcal{C}_{\mathcal{A}_2}$.

Theorem 2. For any fixed $\varepsilon > 0$ algorithm $\mathcal{A}_2$ finds a $(1 + \varepsilon)$-approximate solution of Problem 1 in $\mathcal{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}} + 2\right)^q\right)$ time.

Corollary 1. In the case when the dimension q of space is bounded by a constant value, this algorithm is an FPTAS.

Later in [Kel'manov et al., 2017] was proposed the similar approximation scheme with the same time complexity for the generalization of Problem 1 in which the weight factors are given as input (positive real numbers). Also there was presented an improved algorithm that allows to find $(1 + \varepsilon)$-approximate solution of generalized problem in $\mathcal{O}\left(\sqrt{q}N^2\left(\frac{\pi e}{2}\right)^{q/2}\left(\sqrt{\frac{2}{\varepsilon}} + 2\right)^q\right)$ time. In the case when the dimension q of space is bounded by a constant value, this algorithm is also an FPTAS. In addition, improved algorithm remains polynomial even when the dimension q of the space is bounded by the value $C \log N$, where C is the positive constant. It is clear that in this case algorithm implements a PTAS with time complexity $N^{O(\log \frac{1}{\varepsilon})}$.

The last proposed algorithm for Problem 1 until this moment is a 2-approximation algorithm. It was proposed in [Kel'manov & Motkova, 2017] and allows us to find an approximate solution in the polynomial time.

Algorithm $\mathcal{A}_3$ (2-approximation algorithm).

Input: N-elements set $\mathcal{Y} \subset \mathbb{R}^q$, natural number $M \leq N$.

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 $F_1(\mathcal{B}^y)$ using formula (1).

Step 2. If $F_1(\mathcal{B}^y) = 0$, then put $\mathcal{C}_{\mathcal{A}_2} = \mathcal{B}^y$; 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 $\mathcal{C}_{\mathcal{A}_3}$ the set $\mathcal{B}^x$, for which $F_1(\mathcal{B}^x)$ is minimal.

Output: the set $\mathcal{C}_{\mathcal{A}_3}$.

Two examples of an input set (of 1000 points) and admissible solutions (i.e. 300-elements subset $\mathcal{B}^y$) found by Algorithm $\mathcal{A}_3$ at step 1 are presented at Fig.1 *(a)* and *(b)*.

(a) *(b)*

Fig. 1.

Let us substuntiate this algorithm below.

The following two lemmas are well known. (see, for example, [Kel'manov & Romanchenko, 2012, Kel'manov & Romanchenko, 2014]).

Lemma 1. For an arbitrary point $x \in \mathbb{R}^q$, a finite set $\mathcal{Z} \subset \mathbb{R}^q$ and $\overline{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ ($\overline{z}$ is the centroid of $\mathcal{Z}$), it is true that

$$\sum_{z \in \mathcal{Z}} \|z - x\|^2 = \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 + |\mathcal{Z}| \cdot \|x - \overline{z}\|^2 .$$

Lemma 2. For a finite set $\mathcal{Z} \subset \mathbb{R}^q$, if a point $u \in \mathbb{R}^q$ is closer (in terms of distance) to the centroid $\overline{z}$ of $\mathcal{Z}$ than any point in $\mathcal{Z}$, then

$$\sum_{z \in \mathcal{Z}} \|z - u\|^2 \leq 2 \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 .$$

Lemma 3. Let

$$S(\mathcal{C}, x) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - x\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \, \mathcal{C} \subseteq \mathcal{Y}, \, x \in \mathbb{R}^q .$$

Then the next statements are true:

(1) for any nonempty fixed set $\mathcal{C} \subseteq \mathcal{Y}$ the minimum of the function $S(\mathcal{C}, x)$ over $x \in \mathbb{R}^q$ is reached at the point $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$;

(2) if $|\mathcal{C}| = M = const$, then for any fixed point $x \in \mathbb{R}^q$ the minimum of function $S(\mathcal{C}, x)$ over $\mathcal{C} \subseteq \mathcal{Y}$ is reached at the subset $\mathcal{B}^x$ that consists of M points of the set $\mathcal{Y}$, at which the function (2) has the smallest values.

Proof. The first statement follows from Lemma 2 and the definition of the functions S and F. Since $|\mathcal{Y}| = N$ and $|\mathcal{C}| = M$, the second statement follows from the next chain of equalities:

$$S(\mathcal{C}, x) = M \sum_{y \in \mathcal{C}} \|y - x\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 = M \sum_{y \in \mathcal{C}} \|y\|^2 - 2M \sum_{y \in \mathcal{C}} \langle y, x \rangle + M^2 \|x\|^2 +$$

$$(N - M) \Big(\sum_{y \in \mathcal{Y}} \|y\|^2 - \sum_{y \in \mathcal{C}} \|y\|^2 \Big) = \sum_{y \in \mathcal{C}} \Big\{ (2M - N) \|y\|^2 - 2M \langle y, x \rangle \Big\} + M^2 \|x\|^2 + (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 .$$

It remains to note that in the last two equalities the last two addends do not depend on $\mathcal{C}$.

Lemma is proved.

Theorem 3. An Algorithm $\mathcal{A}_3$ finds a 2-approximate solution of Problem 1 in time $\mathcal{O}\left(qN^2\right)$.

Proof. Let $\mathcal{C}^*$ be an optimal subset and $t = \arg\min\limits_{y \in \mathcal{C}^*} ||y - \overline{y}(\mathcal{C}^*)||^2$ be a point from a subset $\mathcal{C}^*$ that is the closest to its centroid.

Step-by-step, the algorithm goes through every points y of the set $\mathcal{Y}$. Since $t \in \mathcal{Y}$, a permissible solution also will be formed for the point t: the subset B^t, defined by Lemma 3 (if $x = t$).

Besides, a value $F(\mathcal{B}^t)$ of objective function for the Problem 1 will be calculated. For this value we have this inequalities from definitions of the functions F and S and Lemma 3 and a definition of the point t:

$$F(\mathcal{B}^t) = S^{\overline{y}(\mathcal{B}^t)}(\mathcal{B}^t) \leq S^t(\mathcal{B}^t) \leq S^t(\mathcal{C}^*). \tag{3}$$

Applying Lemma 2 to the set $\mathcal{Z} = \mathcal{C}^*$ and the point $u = t$, we have

$$\sum_{y \in \mathcal{C}^*} ||y - t||^2 \leq 2 \sum_{y \in \mathcal{C}^*} ||y - \overline{y}(\mathcal{C}^*)||^2.$$

Using this inequality and definition (2), we find an estimation for a right part of (3)

$$S^t(\mathcal{C}^*) = M \sum_{y \in \mathcal{C}^*} ||y - t||^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} ||y||^2$$

$$\leq 2M \sum_{y \in \mathcal{C}^*} ||y - \overline{y}(\mathcal{C}^*)||^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} ||y||^2 \leq 2F(\mathcal{C}^*). \tag{4}$$

Combining (3) and (4), we have

$$F(\mathcal{B}^t) \leq \; 2F(\mathcal{C}^*). \tag{5}$$

From Step 3 we know that the algorithm finds a solution of Problem 1 in the next form:

$$\mathcal{C}_{\mathcal{A}} = \arg\min_{y \in \mathcal{Y}} F(\mathcal{B}^y). \tag{6}$$

Since $\mathcal{B}^t \in \{\mathcal{B}^y | y \in \mathcal{Y}\}$, from (6) we have an inequality

$$F(\mathcal{C}_A) \leq F(\mathcal{B}^t). \tag{7}$$

Finally, (5) and (7) implies an estimate

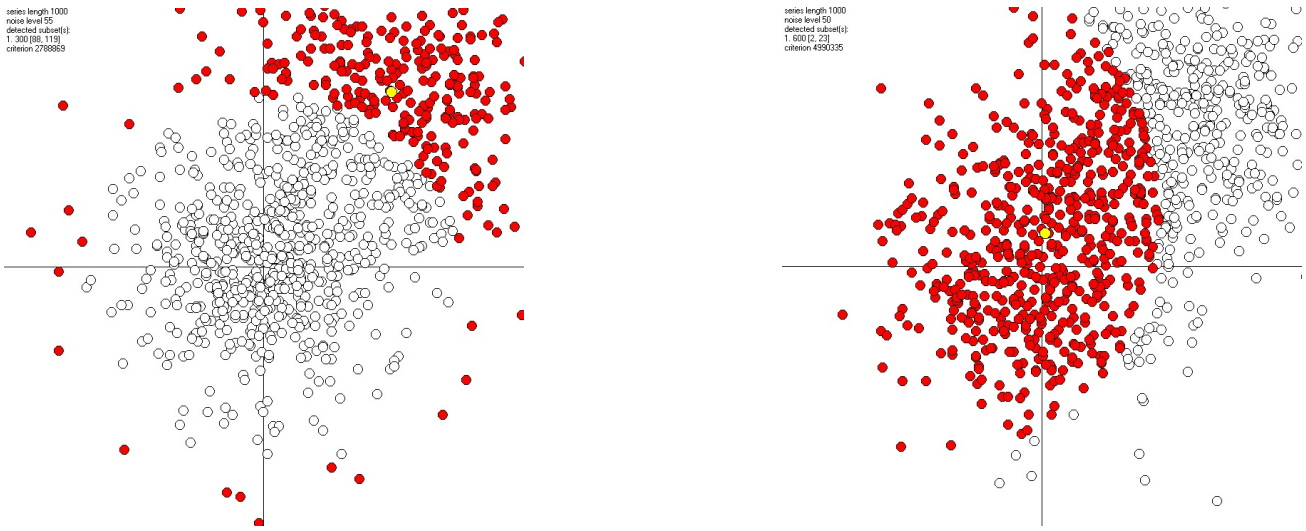
$$F(\mathcal{C}_A) \leq 2F(\mathcal{C}^*). \tag{8}$$

Let us consider the case when at Step 2 of the algorithm the condition $F(\mathcal{B}^y) = 0$ is executed for some input point $y \in \mathcal{Y}$. For every subset $\mathcal{C} \subseteq \mathcal{Y}$ inequality $F(\mathcal{C}) \geq 0$ is correct, so the subset $\mathcal{C}_A = \mathcal{B}^y \subseteq \mathcal{Y}$ is an optimal solution of Problem 1. Inequality (8) is correct for this solution too. It means that a subset $\mathcal{C}_A$ is a 2-approximation solution of Problem 1.

Let us estimate the time complexity of the algorithm. At Step 1 we need no more than $\mathcal{O}(qN)$ operations to calculate values $g^y(z)$. Searching of M the smallest elements in the set of N elements requires $\mathcal{O}(N)$ operations (for example, using an algorithm of finding n-th smallest value in an unordered array [Wirth, 1976]). Step 2 need constant time. So for any point $y \in \mathcal{Y}$ total execution time of Steps 1 and 2 is $\mathcal{O}(qN)$.

As Steps 1 and 2 are executed N times, total time complexity of this steps is $\mathcal{O}(qN^2)$. Time complexity of Step 3 is estimated by value $\mathcal{O}(N)$. So, the time complexity of the algorithm is $\mathcal{O}(qN^2)$. Theorem 3 is proved.

Two examples of an input set (of 1000 points) and 2-approximate solutions found by Algorithm $\mathcal{A}_3$ are presented at Fig.2 *(a)* (i.e. 400-elements subset $\mathcal{C}_A$) and *(b)* (i.e. 600-elements subset $\mathcal{C}_A$).



(a) *(b)*

Fig. 2.

3 Conclusion

In this work, we presented a short survey on the results for the problem of 2-partitioning of points in Euclidean space into two clusters. Also we presented the new result: a 2-approximation algorithm.

It seems important to continue studying and the issue of great interest is the substantiation of randomized algorithms for this problem with linear or sub-linear time complexity.

Acknowledgements

The research for algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ was supported by the Russian Foundation for Basic Research, Projects 16-31-00186, 16-07-00168. Research for algorithm $\mathcal{A}_3$ was supported by the Russian Science Foundation, Project 16-11-10041.

References

[Kel'manov & Pyatkin, 2015] Kel'manov, A. V., & Pyatkin, A. V. (2015). NP-Hardness of Some Quadratic Euclidean 2-Clustering Problems. *Dokl.Akad.Nauk, 464(5)*, 535–538.

[Kel'manov & Pyatkin, 2016] Kel'manov, A. V., & Pyatkin, A. V. (2016). On the Complexity of Some Quadratic Euclidean 2-Clustering Problems. *Comput. Math. Math. Phys., 56(3)*, 491–497.

[Garey & Johnson, 1979] Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco: Freeman.

[Kel'manov & Motkova, 2016a] Kel'manov, A. V., & Motkova, A. V. (2016). Exact Pseudopolynomial Algorithms for a Balanced 2-Clustering Problem. *J. of Appl. and Ind. Math., 10(3)*, 349–355.

[Kel'manov & Motkova, 2016b] Kel'manov, A. V., & Motkova, A. V. (2016). A Fully Polynomial-Time Approximation Scheme for a Special Case of a Balanced 2-Clustering Problem. *LNCS, 9869*, 182–192.

[Kel'manov et al., 2017] Kel'manov, A. V., & Motkova, A. V., & Shenmaier, V. V. (2017). Approximation Schemes for Some Quadratic Problems of Weighted 2-Partitioning a Set of Points. (in Russian) *Tr. Inst. Mat. Mekh., 23(3)*, 159–170.

[Kel'manov & Motkova, 2017] Kel'manov, A. V., & Motkova, A. V. (2017). An Approximation Polynomial-Time Algorithm for a Weighted 2-Clustering Problem with Restriction on Clusters Cardinalities. (in Russian) *Comput. Math. Math. Phys.* (accepted)

[Kel'manov & Romanchenko, 2012] Kel'manov, A. V., & Romanchenko, S. M. (2012). An Approximation Algorithm for Solving a Problem of Search for a Vector Subset. *J. Appl. Ind. Math., 6(1)*, 90–96.

[Kel'manov & Romanchenko, 2014] Kel'manov, A. V., & Romanchenko, S. M. (2014). An FPTAS for a Vector Subset Search Problem. *J. Appl. Indust. Math., 8(3)*, 329–336.

[Wirth, 1976] Wirth, N. (1976). *Algorithms + Data Structures = Programs*. New Jersey: Prentice Hall.