Approximation Algorithm for a Quadratic Euclidean Problem of Searching a Subset with the Largest Cardinality

Alexander A. Ageev[1] Alexander V. Kel'manov[1,2] Artem V. Pyatkin[1,2]
Sergey A. Khamidullin[1] Vladimir V. Shenmaier[1]

[1] Sobolev Institute of Mathematics, Acad. Koptyug avenue, 4, 630090 Novosibirsk, Russia
[2] Novosibirsk State University, Pirogova str. 1, 630090 Novosibirsk, Russia
E-mail: alexander.a.ageev@gmail.com, kelm@math.nsc.ru,
artem@math.nsc.ru, kham@math.nsc.ru, shenmaier@mail.ru

Abstract

We consider the problem of searching a subset in a finite set of points of Euclidean space. The problem is to find a cluster (subset) of the largest cardinality satisfying a given upper bound on the sum of squared distances between the cluster elements and their centroid. We prove that this problem is strongly NP-hard and present a polynomial-time 1/2-approximation algorithm.

1 Introduction

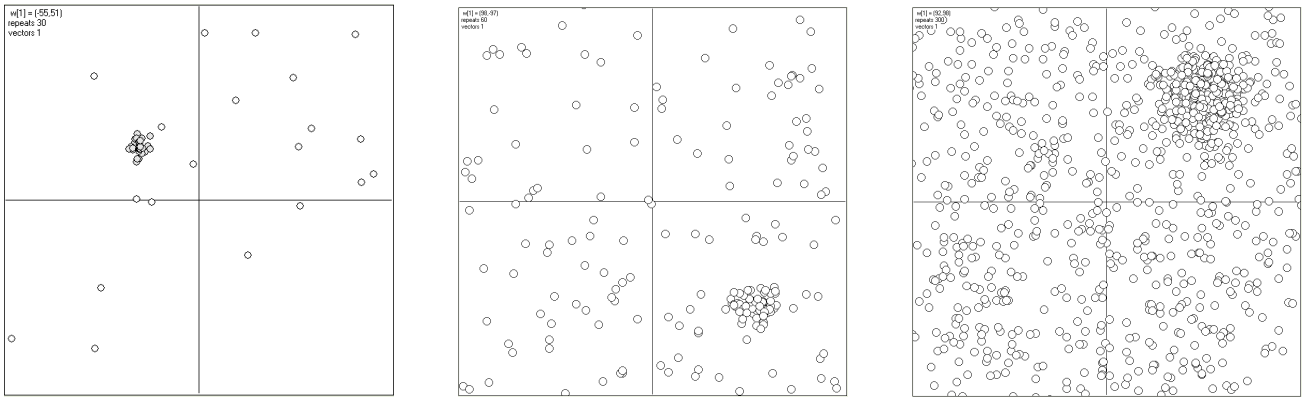
In the paper we consider the problem of finding a subset of largest cardinality in a finite set of points of the Euclidean space for which quadratic variation of points with respect to its unknown centroid does not exceed a given fraction of the quadratic variation of points of the input set with respect to its centroid. Our goal is to analyze the computational complexity of this problem and construct an algorithm for its solution.

The study is motivated, on the one hand, by the absence of published results on the complexity status of the problem and on efficient algorithms with guaranteed performance for this problem. On the other hand, it is motivated by relevance of the problem for a number of applications.

In particular, the problem under consideration is the typical problem for Editing and Cleaning data of "litter" in the form of inaccurate measurements of the characteristics of any objects (see for example [Waal et al., 2011], [Osborne, 2013], [Greco, 2015]). It is well-known that, in Machine learning problem, the cleaning of irrelevant "litter" from data is a necessary element [Bishop, 2006], [James et al., 2013], [Hastie et al., 2009], [Aggarwal, 2015]. If the values of the input data set are the results of measurements of the characteristics of an object and some of the measurements have been taken with instrumental error, the value of which exceeds some given threshold, then we need usually to find a subset (it's desirable, of largest cardinality), having no significant (exceeding the threshold) errors.

Figure 1 shows three examples (a, b, c) the sets of points on the plane (two-dimensional data). The examples correspond to the measurements of the characteristics of three objects. In each of the examples, one can see the dense and sparse subsets of points that correspond to the small and big instrumental errors.

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



(a) *(b)* *(c)*

Fig. 1.

In Data mining and Big data problems, the key element is approximation of the available data by a mathematical model, allowing adequately interpret the data and reasonably explain their origin in terms of the model (see for example, [Waal et al., 2011], [Osborne, 2013], [Greco, 2015], [Bishop, 2006], [James et al., 2013], [Hastie et al., 2009], [Aggarwal, 2015]). In particular, the following statistical hypothesis on the origin of the data could be verified: is it true that the input set is an inhomogeneous sample from several probability distributions while a part of the elements of this set is a sample from a single probability distribution with an unknown average (it is assumed that the correspondence between the sample elements and the distribution is unknown). To verify this conjecture, we first need to find the sample-subset from a single probability distribution. Only after that we can use the classical results from the field of statistical hypothesis testing and investigate the properties of the found set.

2 Problem Formulation, Known and Obtained Results

Everywhere below denote by $\mathbb{R}$ the set of real numbers and by $\|\cdot\|$ the Euclidean norm.

Consider the following problems.

Problem 1 (Subset of points with the largest cardinality).

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

$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \leq \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2, \tag{1}$$

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.

The problem has a simple interpretation, namely, searching the largest by cardinality subset $\mathcal{C}$ of points, whose total quadratic variation from the unknown centroid $\overline{y}(\mathcal{C})$ doesn't exceed the total quadratic variation of the input set $\mathcal{Y}$ from its centroid $\overline{y}(\mathcal{Y})$ by $1/\alpha$ times. If the coordinates of the points of the input set $\mathcal{Y}$ are the results of measuring the characteristics of some object, there could be instrumental errors; if their magnitude (in the form of a quadratic variation) exceeds a certain predetermined threshold (the right-hand side of (1)), then the solution of the problem 1 yields a subset $\mathcal{C}$ of the largest cardinality that does not contain data with a significant (exceeding the threshold) error. The level of the error significance (threshold value) can be regulated by changing α.

The closest problem related to the problem 1 is the following

Problem 2 (M-Variance).

Given: a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and positive integer $M > 1$.

Find: a subset $\mathcal{C} \subseteq \mathcal{Y}$ with cardinality M such that

$$F(\mathcal{C}) \to \min.$$

This problem, like problem 1, is also typical for the applications mentioned above. Unlike the problem 1, in this problem it is required to find a subset of points of a given cardinality, in which the scatter of points from the unknown centroid is minimal. For this problem, in contrast to the problem 1, some results are known.

Apparently, the first formulation of this extremal problem was given in [Aggarwal, 1991], and its strong NP-hardness is established in [Kelmanov, 2011].

In [Aggarwal, 1991] and [Shenmaier, 2016] the solvability of the problem in time $\mathcal{O}(qN^{q+1})$, polynomial in the case for the fixed (bounded from above by a constant) dimension q of the space was shown. The polynomial-time 2-approximation algorithm with $\mathcal{O}(qN^2)$ running time was proposed in [Kelmanov, 2012]. The polynomial-time approximation scheme (PTAS) is developed in [Shenmaier, 2012]. This scheme solves the problem with an arbitrary relative error ε in time $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$.

For the case of a fixed dimension q of space and integer coordinates of points, an exact pseudopolynomial algorithm is constructed [Kelmanov & Romanchenko, 2012] having the $\mathcal{O}(N(MD)^q)$ time complexity, where D is the maximum absolute value of the coordinates of the input points.

In [Kelmanov & Romanchenko, 2014] it is proved that for problem 2 having a numerical input there is no fully polynomial-time approximation scheme (FPTAS), unless $P = NP$ and in the same paper such a scheme was developed for the case when the space dimension is bounded from above by a constant. This scheme solves the problem with an arbitrary relative error ε in time $\mathcal{O}(N^2(M/\varepsilon)^q)$.

In this paper we show that the problem 1 is NP-hard in the strong sense and present an 1/2-approximation polynomial-time algorithm.

3 Analysis of the Problem Complexity

Before analyzing the computational complexity of problem 1, note that the right-hand side (1) does not depend on $\mathcal{C}$ and for a specified input is a constant. Therefore, the problem 1 in the property verification form looks as follows.

Problem 1A (Subset of points with the largest cardinality).

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

Question: is there a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality at least M such that

$$F(\mathcal{C}) \le A\,? \tag{2}$$

Remind that the following problem is NP-complete in the strong sense [Kelmanov, 2011].

Problem 2A (M-Variance).

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

Question: is there a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M such that

$$F(\mathcal{C}) \le B\,?$$

In [Kelmanov, 2011], the well-known [Papadimitriou, 1994] strongly NP-complete problem Clique in a regular graph whose degree is not fixed was reduced to problem 2A.

Note that the function F has the following property: if $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $F(\mathcal{C}_1) \le F(\mathcal{C}_2)$. Therefore, if in the problem 1A the answer is positive, then there is a subset of cardinality M satisfying the inequality (2). 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 the problem 2.

4 Approximation Algorithm

The idea of the proposed approximation algorithm for the problem 1 is following. For each point y of the input set, a subset consisting of the maximum number of closest (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). Among the found subsets the one with the largest cardinality is taken as an output. This approach is realized in the following algorithm.

Algorithm $\mathcal{A}$.

Input: set $\mathcal{Y}$ and number α.

Step 1. Compute the value $A = \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2$ (the right part of the inequality (1)).

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

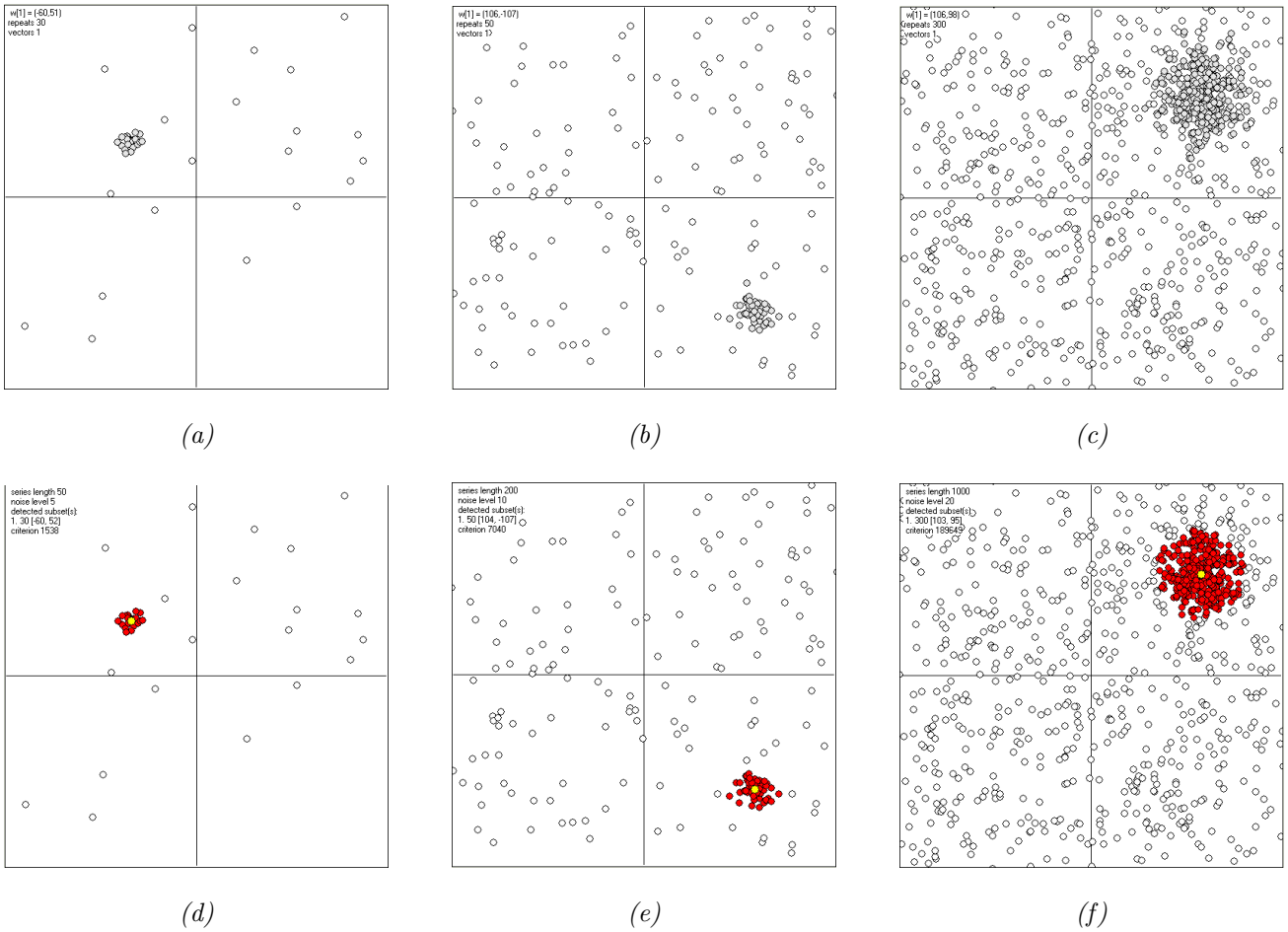
Step 4. In the family $\{\mathcal{C}^y | \, y \in \mathcal{Y}\}$ of admissible subsets constructed in step 3 choose as the output $\mathcal{C}_A$ any subset $\mathcal{C}^y$ of the largest cardinality.

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

The following theorem holds

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

Figure 2 presented below shows the effectiveness of the algorithm for the problem of search for a set of similar elements in a data collection. The subsets (points of darker color) found by the algorithm (for some given α) are presented in Fig. 2(d), Fig. 2(e), Fig. 2(f) correspondingly for the data presented in Fig. 2(a), Fig. 2(b), Fig. 2(c). The complementary subsets can be interpreted as the collections of unlike elements, which can be removed during editing or cleaning.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

Fig. 2.

5 Conclusion

The results presented in the paper have two aspects. On the one hand, they are an investigation of a new discrete optimization problem, and on the other hand, they give a new effective algorithmic tool for solving one of the actual problems in data analysis.

In terms of analysis of large-scale data, faster algorithms with guaranteed accuracy bounds, solving the problem in linear and sublinear time, are of considerable interest. The construction of such algorithms, as well as polynomial approximation schemes for the considered problem, seems to be a matter of immediate prospects.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research, project nos. 15-01-00462, 15-01-00976, and 16-07-00168, and by the grant of Presidium RAS (program 5, project 227), and by the Ministry of Science and Education of the Russian Federation under the 5-100 Excellence Programme.

References

[Waal et al., 2011] de Waal,T., Pannekoek, J., & Scholtus S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley and Sons, Inc. Hoboken, New Jersey.

[Osborne, 2013] Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data. 1st Edition*. Los Angeles: SAGE Publication, Inc.

[Greco, 2015] Greco, L. (2015). *Robust Methods for Data Reduction Alessio Farcomeni*. Chapman and Hall/CRC.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, LLC.

[James et al., 2013] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) *An Introduction to Statistical Learning*. New York: Springer Science+Business Media, LLC.

[Hastie et al., 2009] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning (2nd edition)*. Springer-Verlag.

[Aggarwal, 2015] Aggarwal, C. (2015). *Data Mining: The Textbook*. Springer International Publishing.

[Aggarwal, 1991] Aggarwal, A., Imai, H., Katoh, N., & Suri S. (1991). Finding k points with minimum diameter and related problems. *J. Algorithms, 12(1)*, 38-56.

[Kelmanov, 2011] Kelmanov, A. V., & Pyatkin, A. V. (2011). NP-Completeness of Some Problems of Choosing a Vector Subset. *J. Appl. Indust. Math., 5(3)*, 352-357. doi:10.1134/S1990478911030069

[Shenmaier, 2016] Shenmaier, V. V. (2016). Solving Some Vector Subset Problems by Voronoi Diagrams. *J. Appl. Indust. Math., 10(4)*, 560-566. doi:10.1134/S199047891604013X

[Kelmanov, 2012] Kelmanov A. V., & Romanchenko S. M. (2012). An Approximation Algorithm for Solving a Problem of Search for a Vector Subset. *J. Appl. Indust. Math., 6(1)*, 90-96. doi:10.1134/S1990478912010097

[Shenmaier, 2012] Shenmaier, V. V. (2012). An Approximation Scheme for a Problem of Search for a Vector Subset. *J. Appl. Indust. Math., 6(3)*. 381-386. doi:10.1134/S1990478912030131

[Kelmanov & Romanchenko, 2012] (2012). Kelmanov, A. V., & Romanchenko S. M. Pseudopolynomial Algorithms for Certain Computationally Hard Vector Subset and Cluster Analysis Problems. *Automation and Remote Control, 73(2)*. 349-354. doi:10.1134/S0005117912020129

[Kelmanov & 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. doi:10.1134/S1990478914030041

[Papadimitriou, 1994] Papadimitriou, C. H. (1994). *Computational Complexity*. New-York: Addison-Wesley.