TRUDY INSTITUTA MATEMATIKI I MEKHANIKI UrO RAN

Vol. 25 No. 4 **2019**

MSC: 68W25, 68Q25

DOI: 10.21538/0134-4889-2019-25-4-69-78

QUADRATIC EUCLIDEAN 1-MEAN AND 1-MEDIAN 2-CLUSTERING PROBLEM WITH CONSTRAINTS ON THE SIZE OF THE CLUSTERS: COMPLEXITY AND APPROXIMABILITY

A. V. Kel'manov, A. V. Pyatkin, V. I. Khandee

We consider the problem of partitioning a set of N points in d-dimensional Euclidean space into two clusters minimizing the sum of the squared distances between each element and the center of the cluster to which it belongs. The center of the first cluster is its centroid (the geometric center). The center of the second cluster should be chosen among the points of the input set. We analyze the variant of the problem with given sizes (cardinalities) of the clusters; the sum of the sizes equals the cardinality of the input set. We prove that the problem is strongly NP-hard and there is no fully polynomial-time approximation scheme for its solution.

Keywords: Euclidean space, clustering, 2-partition, quadratic variation, center, centroid, median, strong NP-hardness, nonexistence of FPTAS, approximation-preserving reduction.

REFERENCES

1. Aloise D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 2009, vol. 75, no. 2, pp. 245–248. doi: 10.1007/s10994-009-5103-0 .
2. Kariv O., Hakimi S. An algorithmic approach to network location problems. Part II: The p-Medians. *SIAM J. Appl. Math.*, 1979, vol. 37, no. 3, pp. 539–560. doi: 10.1137/0137041 .
3. Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A. A posteriori detection of a quasiperiodic fragment with a given number of repetitions in a numerical sequence. *Sib. Zh. Ind. Mat.*, 2006, vol. 9, no. 1, pp. 55–74 (in Russian).
4. Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A. A posteriori detecting a quasiperiodic fragment in a numerical sequence. *Pattern Recognition and Image Analysis*, 2008, vol. 18, no. 1, pp. 30–42. doi: 10.1134/S1054661808010057 .
5. Baburin A.E., Gimadi E.Kh., Glebov, N.I., Pyatkin, A.V. The problem of finding a subset of vectors with the maximum total weight. *J. Appl. Industr. Math.*, 2008, vol. 2, no. 1, pp. 32–38. doi: 10.1007/s11754-008-1004-3 .
6. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning.* N Y: Springer Science+Business Media, LLC, 2013, 426 p. ISBN: 978-1461471370 .
7. Bishop C.M. *Pattern Recognition and Machine Learning.* N Y: Springer Science+Business Media, LLC, 2006, 738 p. ISBN: 978-0-387-31073-2 .
8. Shirkhorshidi A.S., Aghabozorgi S, Wah T,Y., and Herawan T. Big data clustering: A review. In: Murgante B. et al. (eds), Computational Science and Its Applications (ICCSA 2014), *Lecture Notes in Computer Science*, 2014, vol. 8583, pp. 707–720. doi: 10.1007/978-3-319-09156-3_49 .
9. Aggarwal C.C. *Data mining: The textbook.* Cham: Springer, 2015, 734 p. doi: 10.1007/978-3-319-14142-8 .
10. Edwards A.W.F., Cavalli-Sforza L.L. A method for cluster analysis. *Biometrics*, 1965, vol. 21, pp. 362–375. doi: 10.2307/2528096 .
11. Garey M.R., Johnson D.S. *Computers and intractability: A guide to the theory of NP-completeness.* San Francisco: Freeman, 1979, 338 p. ISBN: 0716710447 .
12. Papadimitriou C.H. *Computational complexity.* N Y: Addison-Wesley, 1994, 523 p. ISBN: 0-201-53082-1 .
13. Vazirani V.V. *Approximation Algorithms.* Berlin; Heidelberg; N Y: Springer-Verlag, 2003, 380 p. doi: 10.1007/978-3-662-04565-7 .

14. Dolgushev A.V., Kel'manov A.V. An approximation algorithm for solving a problem of cluster analysis. *J. Appl. Indust. Math.*, 2011, vol. 5, no. 4, pp. 551–558. doi: 10.1134/S1990478911040107 .
15. Dolgushev A.V., Kel'manov A.V., Shenmaier V.V. A polynomial-time approximation scheme for one problem of cluster analysis In: K. V. Vorontsov (ed.) Intelligent Data Processing: Proc. of the 9th Internat. Conf. (Republic of Montenegro, Budva, September 16–22, 2012), Moscow: Torus Press, 2012, pp. 242–244 (in Russian).
16. Kel'manov A.V., Khandeev V.I. A Randomized algorithm for two-cluster partition of a set of vectors. *Comput. Math. Math. Phys.*, 2015, vol. 55, no. 2, pp. 330–339. doi: 10.1134/S096554251502013X .
17. Kel'manov A.V., Motkova A.V., Shenmaier V.V. An approximation scheme for a weighted two-cluster partition problem. Analysis of Images, Social Networks and Texts - 6th Internat. Conf. (AIST 2017), Revised Selected Papers, *Lecture Notes in Computer Science*, 2018. Vol. 10716. P. 323–333. doi: 10.1007/978-3-319-73013-4_30 .

Received August 12, 2019
Revised September 10, 2019
Accepted September 16, 2019

Funding Agency: This work was supported by the Russian Foundation for Basic Research (project nos. 19-01-00308 and 18-31-00398), by Program I.5.1 for Fundamental Research of the Siberian Branch of the Russian Academy of Sciences (project nos. 0314-2019-0014 and 0314-2019-0015), and by the Ministry of Education and Science of the Russian Federation within the Russian Academic Excellence Project.

Alexander Vasil'evich Kel'manov, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: kelm@math.nsc.ru .

Artem Valer'evich Pyatkin, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: artem@math.nsc.ru .

Vladimir Il'ich Khandeev, Cand. Sci. (Phys.-Math.), Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: khandeev@math.nsc.ru .

Cite this article as: A. V. Kel'manov, A. V. Pyatkin, V. I. Khandeev. Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with constraints on the size of the clusters: Complexity and approximability, *Trudy Instituta Matematiki i Mekhaniki URO RAN*, 2019, vol. 25, no. 4, pp. 69–78 .