

ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ  
ЗАДАЧИ ВЫБОРА ТИПИЧНЫХ ПРЕДСТАВИТЕЛЕЙ  
В 2-РАЗБИЕНИИ КОНЕЧНОГО МНОЖЕСТВА ТОЧЕК  
МЕТРИЧЕСКОГО ПРОСТРАНСТВА

*И. А. Борисова*

<sup>1</sup> Институт математики им. С. Л. Соболева,  
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия

<sup>2</sup> Новосибирский гос. университет,  
ул. Пирогова, 2, 630090 Новосибирск, Россия

E-mail: [biamia@mail.ru](mailto:biamia@mail.ru)

**Аннотация.** Исследуется вычислительная сложность одной экстремальной задачи выбора подмножества  $p$  точек в заданном 2-разбиении конечного множества точек метрического пространства. При этом требуется, чтобы выбранное подмножество наилучшим образом описывало определяемые 2-разбиением кластеры с точки зрения некоторого геометрического критерия. Рассматриваемая задача является формализацией одной прикладной проблемы из анализа данных, заключающейся в отыскании подмножества типичных представителей выборки, состоящей из объектов двух классов с опорой на функцию конкурентного сходства. В статье доказывается, что рассматриваемая задача NP-трудна. Для этого к ней полиномиально сводится одна из хорошо известных NP-трудных в сильном смысле задач — задача о  $p$ -медиане. Библиогр. 15.

**Ключевые слова:** NP-трудная задача, выбор типичных объектов, функция конкурентного сходства, задача о  $p$ -медиане, анализ данных.

### Введение

Предметом исследования является экстремальная задача выбора из заданного конечного множества точек метрического пространства, разделённого на два непересекающихся кластера, подмножества  $p$  точек,

---

Исследование выполнено при финансовой поддержке Программы фундаментальных научных исследований СО РАН (проект № 0314–2019–0015).

© И. А. Борисова, 2020

наилучшим образом описывающих эти кластеры с точки зрения некоторого критерия. Цель исследования — определение вычислительной сложности этой задачи [1].

Рассматриваемая задача моделирует характерную для анализа данных проблему отыскания в разделённой на классы обучающей выборке наиболее типичных представителей этих классов для целей распознавания, поиска информативных признаков, очистки данных от шумов и пр.

Исследование мотивировано отсутствием опубликованных данных о вычислительной сложности этой задачи, а также её актуальностью в теоретическом плане и востребованностью во многих прикладных исследованиях, в том числе связанных с медицинской диагностикой.

### 1. Формулировка задачи, интерпретация и истоки

**Задача 1.** Даны два конечных непересекающихся множества  $X_1, X_2$  точек в метрическом пространстве размерности  $d$  и натуральное число  $p$ .

Найти непустые подмножества  $Y_1 \subseteq X_1, Y_2 \subseteq X_2$ , которые максимизируют целевую функцию

$$F(Y_1, Y_2) = \sum_{z \in X_1} \frac{\min_{y \in Y_2} r(z, y) - \min_{x \in Y_1} r(z, x)}{\min_{y \in Y_2} r(z, y) + \min_{x \in Y_1} r(z, x)} + \sum_{z \in X_2} \frac{\min_{y \in Y_1} r(z, y) - \min_{x \in Y_2} r(z, x)}{\min_{y \in Y_1} r(z, y) + \min_{x \in Y_2} r(z, x)}, \quad (1)$$

где  $r(a, b)$  — расстояние между точками  $a$  и  $b$ , при ограничении  $|Y_1| + |Y_2| = p$ .

Эта задача близка к таким известным задачам кластеризации, как  $k$ -means (или  $k$ -MSSC) [2, 3] и  $k$ -median [4] в их дискретной постановке [5, 6], когда центры кластеров отыскиваются не среди всех точек метрического пространства, а только среди тех, которые включены в исходное множество точек. Также похожие постановки есть в [7, 8], где в процессе кластеризации отыскиваются наиболее типичные элементы, в то время как элементы-выбросы в формируемые кластеры не включаются. Несколько вариантов постановок задачи выбора подмножества точек из заданного множества для различных геометрических критериев также рассматриваются в [9]. В отличие от всех вышеперечисленных задач, которые можно интерпретировать как задачи поиска центров неизвестных кластеров, в рассматриваемой постановке разбиение множества точек на два кластера задаётся изначально, но при этом каждый кластер может описываться несколькими точками вместо одного центра.

В литературе также встречаются похожие постановки задачи выбора подмножества точек из 2-разбиения, однако для других критериев. Так, например, в [10, 11] выбирается подмножество минимальной мощности таким образом, что для любого элемента исходного 2-разбиения ближайший элемент этого подмножества будет относиться к тому же кластеру. В такой постановке задача выбора подмножества точек из 2-разбиения NP-трудна. В [12, 13] выбирается подмножество точек, определяющих полосу максимальной ширины, отделяющую элементы одного кластера от элементов другого. Эта задача является задачей квадратичного программирования и решается за полиномиальное время.

В прикладном плане рассматриваемая задача может трактоваться как формализация проблемы выбора типичных объектов из классифицированной выборки в метрическом пространстве. Она возникает в распознавании образов и машинном обучении при анализе выборок, предварительно разделённых экспертами на несколько классов, когда есть необходимость выделить наиболее типичных представителей каждого класса, позволяющих наилучшим образом эти классы отличать друг от друга. Так, при анализе медицинских данных объектами могут быть описания пациентов с их симптомами, классами — различные заболевания, диагностированные у этих пациентов, а типичными представителями — пациенты с симптоматикой, наиболее характерной для различных форм каждого заболевания. Принципы, в соответствии с которыми осуществляется отбор типичных представителей классифицированной выборки при машинном обучении, могут меняться в зависимости от того, как формализуется понятие типичного объекта.

В рассматриваемой постановке типичными считаются те объекты классифицированной выборки, на которые максимально похожи объекты из того же класса и непохожи объекты других классов. В качестве меры похожести-непохожести для объектов, представленных в виде точек многомерного метрического пространства, используется функция конкурентного сходства или FRiS-функция (Function of Rival Similarity) [14], которая позволяет вычислять величину конкурентного сходства на основе расстояний между объектами.

## 2. Описание математической модели

Рассмотрим моделируемую задачу анализа данных, в которой дана выборка  $A$ , состоящая из объектов двух классов  $\{1, 2\}$ ,  $A = A_1 \cup A_2$ ,  $A_1 \cap A_2 = \emptyset$ . Для всех троек объектов из такой выборки задан способ вычислять конкурентное сходство  $f(z, x, y)$  таким образом, что чем больше похож  $z$  на объект  $x$  и непохож на объект  $y$ , тем  $f(z, x, y)$  больше. Тогда типичными с точки зрения конкурентного сходства считаются объекты, сходство с которыми объектов своего класса как можно больше,

а сходство объектов чужого класса как можно меньше. В этом случае задача выбора подмножества  $p$  типичных объектов из этой выборки заключается в отыскании таких непустых подмножеств  $B_1 \subseteq A_1$ ,  $B_2 \subseteq A_2$ ,  $|B_1| + |B_2| = p$ , которые обеспечат максимальное сходство всех объектов выборки с типичными объектами своего класса при минимальном сходстве с типичными объектами другого класса:

$$F(B_1, B_2) = \sum_{z \in A_1} \max_{x \in B_1} \min_{y \in B_2} f(z, x, y) + \sum_{z \in A_2} \max_{x \in B_2} \min_{y \in B_1} f(z, x, y) \rightarrow \max_{\substack{B_1 \subseteq A_1, B_2 \subseteq A_2, \\ |B_1| + |B_2| = p, \\ |B_1| > 0, |B_2| > 0}}. \quad (2)$$

В большинстве случаев при решении прикладных задач каждый объект из множества  $A$  рассматривается как точка в пространстве описывающих характеристик, на котором определена некоторая метрика  $r$ . Это позволяет вычислять расстояние  $r(x, y)$  между любой парой объектов  $x, y \in A$ . В заданном метрическом пространстве в качестве величины конкурентного сходства  $f(z, x, y)$  используется FRiS-функция [11], определяемая по формуле

$$\text{FRiS}(z, x, y) = \frac{r(z, y) - r(z, x)}{r(z, y) + r(z, x)}.$$

Для получения критерия из задачи (1) осталось подставить формулу для вычисления FRiS-функции в максимизируемый функционал задачи (2), в результате чего его слагаемые преобразуются следующим образом:

$$\begin{aligned} \max_{x \in B_1} \min_{y \in B_2} \text{FRiS}(z, x, y) &= \max_{x \in B_1} \min_{y \in B_2} \frac{r(z, y) - r(z, x)}{r(z, y) + r(z, x)} = \\ &= \frac{\min_{y \in B_2} r(z, y) - \min_{x \in B_1} r(z, x)}{\min_{y \in B_2} r(z, y) + \min_{x \in B_1} r(z, x)}, \\ \max_{x \in B_2} \min_{y \in B_1} \text{FRiS}(z, x, y) &= \frac{\min_{y \in B_1} r(z, y) - \min_{x \in B_2} r(z, x)}{\min_{y \in B_1} r(z, y) + \min_{x \in B_2} r(z, x)}. \end{aligned}$$

### 3. Вычислительная сложность задачи выбора подмножества типичных объектов

Для установления сложностного статуса задачи 1 покажем, что к ней сводится задача о  $p$ -медиане в следующей постановке.

**Задача 2** (задача о  $p$ -медиане). Даны два множества индексов  $I = \{1, \dots, m\}$  — множество клиентов,  $J = \{1, \dots, n\}$  — множество потенциальных мест открытия предприятий и числовая матрица  $(h_{ij})_{mn}$  потенциальной прибыли от удовлетворения спроса  $i$ -го клиента  $j$ -м предприятием,  $h_{ij} \geq 0$  и натуральное число  $p \geq 2$ . Найти подмножество предприятий  $S \subseteq J$  мощности  $p$  таким образом, чтобы максимизировать общую прибыль  $Q(S)$  от обслуживания всех клиентов:

$$Q(S) = \sum_{i \in I} \max_{j \in S} h_{ij}. \quad (3)$$

Задача о  $p$ -медиане принадлежит классу NP-трудных в сильном смысле задач, поскольку к ней сводится задача о вершинном покрытии [15].

**Теорема 1.** Задача о  $p$ -медиане полиномиально сводится к задаче 1.

**ДОКАЗАТЕЛЬСТВО.** Предположим, что в задаче о  $p$ -медиане матрица прибыли невырожденная, т. е. в ней есть ненулевые элементы, и отыщем среди них максимальный  $H = \max_{i \in I, j \in J} h_{ij} > 0$ . Для удобства дальнейшего изложения перепишем задачу 1 в следующем виде.

Даны два множества индексов  $I_1 = \{1, 2, \dots, m_1\}$ ,  $I_2 = \{m_1 + 1, \dots, m_1 + m_2\}$ , числовая матрица  $(r_{ij})_{(m_1+m_2) \times (m_1+m_2)}$ , для элементов которой выполняются условия  $r_{ij} \geq 0$ ,  $r_{ii} = 0$ ,  $r_{ij} = r_{ji}$ ,  $r_{ij} \leq r_{ik} + r_{kj}$ , и натуральное число  $p \geq 2$ . Найти два непустых подмножества индексов  $S_1 \subseteq I_1$  и  $S_2 \subseteq I_2$ , которые максимизируют целевую функцию

$$\text{Fr}(S_1, S_2) = \sum_{i \in I_1} \frac{\min_{k \in S_2} r_{ik} - \min_{j \in S_1} r_{ij}}{\min_{k \in S_2} r_{ik} + \min_{j \in S_1} r_{ij}} - \sum_{i \in I_2} \frac{\min_{k \in S_2} r_{ik} - \min_{j \in S_1} r_{ij}}{\min_{k \in S_2} r_{ik} + \min_{j \in S_1} r_{ij}} \quad (4)$$

при ограничении  $|S_1| + |S_2| = p$ .

В этой постановке индексами нумеруются точки метрического пространства, а числовая матрица  $(r_{ij})$  задаёт попарные расстояния между этими точками.

Для доказательства утверждения теоремы рассмотрим задачу выбора  $p + 1$  индексов для частного случая, когда второе множество индексов состоит из одного индекса, а мощность первого множества индексов равна  $m + n$ , т. е. будем рассматривать два набора индексов  $I_1 = \{1, 2, \dots, m + n\}$ ,  $I_2 = \{m + n + 1\}$ , а числовую матрицу попарных расстояний  $(r_{ij})_{(m+n+1) \times (m+n+1)}$  зададим следующим образом:

$$\begin{aligned} r_{i(m+n+1)} &= r_{(m+n+1)i} = 1, \quad i = 1, \dots, m, \\ r_{i(m+n+1)} &= r_{(m+n+1)i} = 0,5, \quad i = m + 1, \dots, m + n, \\ r_{ij} &= 1, \quad i = 1, \dots, m, j = 1, \dots, m, i \neq j, \end{aligned}$$

$$\begin{aligned}
r_{ij} &= 0,5, \quad i = m+1, \dots, m+n, j = m+1, \dots, m+n, i \neq j, \\
r_{i(j+m)} &= \frac{3H - h_{ij}}{3H + h_{ij}}, \quad i = 1, \dots, m, j = 1, \dots, n, \\
r_{(i+m)j} &= \frac{3H - h_{ji}}{3H + h_{ji}}, \quad i = 1, \dots, n, j = 1, \dots, m, \\
r_{ii} &= 0, \quad i = 1, \dots, m+n+1.
\end{aligned}$$

Таким образом,

$$\begin{aligned}
& (r_{ij})_{(m+n+1) \times (m+n+1)} \\
&= \begin{pmatrix} \begin{pmatrix} 0 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 0 \end{pmatrix}_{m \times m} & \begin{pmatrix} \frac{3H-h_{ij}}{3H+h_{ij}} \end{pmatrix}_{m \times n} & (1)_{m \times 1} \\ \begin{pmatrix} \frac{3H-h_{ij}}{3H+h_{ij}} \end{pmatrix}_{n \times m}^\top & \begin{pmatrix} 0 & \dots & 0,5 \\ \vdots & \ddots & \vdots \\ 0,5 & \dots & 0 \end{pmatrix}_{n \times n} & (0,5)_{n \times 1} \\ (1)_{1 \times m} & (0,5)_{1 \times n} & 0 \end{pmatrix}.
\end{aligned}$$

Нетрудно убедиться, что, будучи определённой таким образом, матрица удовлетворяет всем аксиомам расстояний. Первым двум — по определению, а выполнение же неравенства треугольника следует из того факта, что так как  $0,5 \leq \frac{3H-h_{ij}}{3H+h_{ij}} \leq 1$ , для любых  $i \neq j$  выполняется  $0,5 \leq r_{ij} \leq 1$ , и для любой тройки расстояний  $r_{ij}$ ,  $r_{ik}$  и  $r_{kj}$  из этого диапазона имеет место  $r_{ij} \leq 1 \leq r_{ik} + r_{kj}$ .

Множество  $I_2$  состоит из одного индекса, а на  $S_2 \subseteq I_2$  накладывается требование, чтобы оно было непустым, а значит,  $S_2 = I_2 = \{m+n+1\}$ . В результате максимизируемый функционал из задачи (4) упрощается:

$$\begin{aligned}
& \text{Fr}(S_1, \{m+n+1\}) \\
&= \sum_{i \in I_1} \frac{r_{i(m+n+1)} - \min_{j \in S_1} r_{ij}}{r_{i(m+n+1)} + \min_{j \in S_1} r_{ij}} - \frac{r_{(m+n+1)(m+n+1)} - \min_{j \in S_1} r_{(m+n+1)j}}{r_{(m+n+1)(m+n+1)} + \min_{j \in S_1} r_{(m+n+1)j}} \\
&= \sum_{i=1}^m \frac{1 - \min_{j \in S_1} r_{ij}}{1 + \min_{j \in S_1} r_{ij}} + \sum_{i=m+1}^{m+n} \frac{0,5 - \min_{j \in S_1} r_{ij}}{0,5 + \min_{j \in S_1} r_{ij}} + 1 \\
&= \sum_{i=1}^m \max_{j \in S_1} \frac{1 - r_{ij}}{1 + r_{ij}} + \sum_{i=m+1}^{m+n} \max_{j \in S_1} \frac{0,5 - r_{ij}}{0,5 + r_{ij}} + 1.
\end{aligned}$$

Для удобства введём следующие обозначения:

$$g_{ij} = \frac{1 - r_{ij}}{1 + r_{ij}}, \quad i = 1, \dots, m, j = 1, \dots, m + n,$$

$$g_{ij} = \frac{0,5 - r_{ij}}{0,5 + r_{ij}}, \quad i = m + 1, \dots, m + n, j = 1, \dots, m + n.$$

Тогда вся задача сводится к нахождению непустого подмножества индексов  $S_1 \subseteq I_1$  такого, что

$$\text{Fr}(S_1, \{m + n + 1\}) = \sum_{i=1}^{m+n} \max_{j \in S_1} g_{ij} + 1 \rightarrow \max_{\substack{S_1 \subseteq I_1, \\ |S_1|=p}}. \quad (5)$$

Вычислим  $(g_{ij})_{(m+n) \times (m+n)}$  по матрице расстояний  $(r_{ij})_{(m+n+1) \times (m+n+1)}$ :

$$g_{ii} = \frac{1 - 0}{1 + 0} = 1, \quad i = 1, \dots, m,$$

$$g_{ij} = \frac{1 - 1}{1 + 1} = 0, \quad i = 1, \dots, m, j = 1, \dots, m, i \neq j,$$

$$g_{i(j+m)} = \frac{1 - \frac{3H - h_{ij}}{3H + h_{ij}}}{1 + \frac{3H - h_{ij}}{3H + h_{ij}}} = \frac{h_{ij}}{3H}, \quad i = 1, \dots, m, j = 1, \dots, n,$$

$$g_{(i+m)(i+m)} = \frac{0,5 - 0}{0,5 + 0} = 1, \quad i = 1, \dots, n,$$

$$g_{(i+m)(j+m)} = \frac{0,5 - 0,5}{0,5 + 0,5} = 0, \quad i = 1, \dots, n, j = 1, \dots, n, i \neq j,$$

$$g_{(i+m)j} = \frac{0,5 - \frac{3H - h_{ji}}{3H + h_{ji}}}{0,5 + \frac{3H - h_{ji}}{3H + h_{ji}}} = \frac{3h_{ji} - 3H}{9H - h_{ji}}, \quad i = 1, \dots, n, j = 1, \dots, m.$$

В итоге имеем

$$(g_{ij})_{(m+n) \times (m+n)} = \begin{pmatrix} \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}_{m \times m} & \begin{pmatrix} h_{ij} \\ 3H \end{pmatrix}_{m \times n} \\ \begin{pmatrix} 3h_{ij} - 3H \\ 9H - h_{ij} \end{pmatrix}_{n \times m}^\top & \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}_{n \times n} \end{pmatrix}.$$

Множество  $I_1$  разделим на подмножества  $I_1^- = \{i \mid i \in I_1, i \leq m\}$  и  $I_1^+ = \{i \mid i \in I_1, i > m\}$ , а множество  $S_1$  разобьём на подмножества  $S_1^- = \{i \mid i \in S_1, i \leq m\}$  и  $S_1^+ = \{i \mid i \in S_1, i > m\}$ .

Заметим, что  $g_{ij} = \frac{3h_{j(i-m)} - 3H}{9H - h_{j(i-m)}} \leq 0$  для всех  $i \in I_1^+$ ,  $j \in I_1^-$ , в то время как  $g_{ij} = \frac{h_{i(j-m)}}{3H} \geq 0$  для всех  $i \in I_1^-$ ,  $j \in I_1^+$ . Значит,  $g_{ij^-} \leq g_{ij^+}$  для всех  $i \in I_1$ ,  $j^- \in I_1^-$ ,  $j^+ \in I_1^+$ ,  $i \neq j^-$ , следовательно, для любого  $i \notin S_1$

$$\max_{j \in S_1^-} g_{ij} \leq \max_{j \in S_1^+} g_{ij}. \quad (6)$$

Кроме того, для всех  $i \in S_1$  выполняется

$$\max_{j \in S_1} g_{ij} = 1. \quad (7)$$

Из (7) следует, что если  $S_1 \subseteq I_1^-$ , т. е.  $S_1^+ = \emptyset$ , то

$$\begin{aligned} \text{Fr}(S_1^-, \{m+n+1\}) &= \sum_{i=1}^m \max_{j \in S_1^-} g_{ij} + \sum_{i=m+1}^{m+n} \max_{j \in S_1^-} g_{ij} + 1 \\ &= p + \sum_{i=m+1}^{m+n} \max_{j \in S_1^-} \frac{3h_{j(i-m)} - 3H}{9H - h_{j(i-m)}} + 1 \leq p + 1. \end{aligned}$$

Если же  $S_1^+ \neq \emptyset$ , то максимизируемая сумма с учётом (6) и (7) может быть расписана следующим образом:

$$\begin{aligned} \text{Fr}(S_1, \{m+n+1\}) &= \sum_{i=1}^{m+n} \max\{\max_{j \in S_1^-} g_{ij}, \max_{j \in S_1^+} g_{ij}\} + 1 \\ &= \sum_{\substack{i \in I_1, \\ i \notin S_1}} \max_{j \in S_1^+} g_{ij} + p + 1 = \sum_{\substack{i \in I_1^-, \\ i \notin S_1^-}} \max_{j \in S_1^+} \frac{h_{i(j-m)}}{3H} + p + 1 > p + 1. \end{aligned}$$

Таким образом показано, что решение задачи (5) должно содержать не менее одного индекса из  $I_1^+$ , при этом её решение эквивалентно решению следующей задачи:

$$\sum_{\substack{i \in I_1^-, \\ i \notin S_1^-}} \max_{j \in S_1^+} h_{i(j-m)} \rightarrow \max_{\substack{S_1^- \subseteq I_1^-, S_1^+ \subseteq I_1^+, \\ |S_1^-| + |S_1^+| = p, \\ S_1^+ \neq \emptyset}}. \quad (8)$$

Для задачи (8), в свою очередь, очевидно, что для любого  $s^+ \in I_1^+$ ,  $s^+ \notin S_1^+$  выполняется

$$\sum_{\substack{i \in I_1^-, \\ i \notin S_1^-}} \max_{j \in S_1^+} h_{i(j-m)} \leq \sum_{\substack{i \in I_1^-, \\ i \notin S_1^-}} \max_{j \in S_1^+ \cup \{s^+\}} h_{i(j-m)},$$

а для любого  $s^- \in S_1^-$  имеем



$$\sum_{\substack{i \in I_1^-, \\ i \notin S_1^-}} \max_{j \in S_1^+} h_{i(j-m)} \leq \sum_{\substack{i \in I_1^-, \\ i \notin S_1^- / \{s^-\}}} \max_{j \in S_1^+} h_{i(j-m)}.$$

Из этого следует, что любое множество  $S^*$ , являющееся решением задачи (8) и содержащее индексы из  $I_1^-$ , может быть преобразовано в решение  $S^{+*}$  этой же задачи, содержащее только индексы из  $I_1^+$ . Для этого достаточно заменить в нём каждый  $s^- \in \{I_1^- \cap S^*\}$  произвольным ещё не включённым в решение  $s^+ \in \{I_1^+ \setminus S^*\}$ . Полученное таким образом множество  $S^{+*}$  также является решением следующей задачи:

$$\sum_{i \in I_1^-} \max_{j \in S_1^+} h_{i(j-m)} \rightarrow \max_{\substack{S_1^+ \subseteq I_1^+, \\ |S_1^+|=p}}. \quad (9)$$

Осталось заметить, что так как множество индексов  $I_1^-$  совпадает с  $I$ , а множество  $I_1^+$  с точностью до сдвига совпадает с множеством индексов  $J$  из задачи о  $p$ -медиане, то (9) — это не что иное, как эквивалентная запись задачи о  $p$ -медиане. Таким образом, решив задачу (5) в эквивалентной записи (8) и заменив в решении все индексы из  $I_1^-$  на не попавшие в решение индексы из  $I_1^+$ , мы тем самым получим решение задачи (3). Теорема 1 доказана.

**Следствие 1.** *Задача 1 выбора подмножества из множества точек в метрическом пространстве, разделённого на два класса, NP-трудна и остаётся таковой, даже если одно из множеств одноэлементно.*

**ДОКАЗАТЕЛЬСТВО.** Утверждение следствия выполняется в силу того, что к рассматриваемой задаче за полиномиальное время сводится задача о  $p$ -медиане, которая NP-трудна в сильном смысле. Следствие 1 доказано.

### Заключение

В работе доказано, что известная NP-трудная в сильном смысле задача о  $p$ -медиане сводится к задаче выбора из заданного разделённого на два непересекающихся кластера конечного множества точек метрического пространства подмножества  $p$  типичных представителей этих кластеров, наилучшим образом описывающих их с точки зрения некоторого геометрического критерия. Тем самым доказано, что эта задача NP-трудна, и в рамках парадигмы  $P \neq NP$  невозможно предложить точный полиномиальный алгоритм решения данной задачи.

### ЛИТЕРАТУРА

1. Garey M. R., Johnson D. S. Computers and intractability: A guide to the theory of NP-completeness. San Francisco: Freeman. 1979.

2. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // *Mach. Learn.* 2009. Vol. 75, No. 2. P. 245–248.
3. **Dasgupta S.** The hardness of  $k$ -means clustering. Tech. Rep. CS2007-0890. Univ. California, 2008. P. 1–6.
4. **Papadimitriou C. H.** Worst-case and probabilistic analysis of a geometric location problem // *SIAM J. Comput.* 1981. Vol. 10, No. 3. P. 542–557.
5. **Har-Peled S., Mazumdar S.** Coresets for  $k$ -means and  $k$ -median clustering and their applications // *Proc. 36th Annu. ACM Sympos. Theory Comput.* (Chicago, IL, USA, June 13–15, 2004). New York: ACM, 2004. P. 291–300.
6. **Kaufman L., Rousseeuw P. J.** Clustering by means of medoids // *Statistical data analysis based on the  $L_1$ -norm and related methods*. Amsterdam: North Holland. 1987. P. 405–416.
7. **Кельманов А. В., Пяткин А. В., Хандеев В. И.** О сложности некоторых максиминных задач кластеризации // *Тр. Ин-та математики и механики УрО РАН*. 2018. Т. 24, № 4. С. 189–198.
8. **Кельманов А. В., Пяткин А. В.** NP-трудность некоторых евклидовых задач разбиения конечного множества точек // *Журн. вычисл. математики и мат. физики*. 2018. Т. 58, № 5. С. 852–856.
9. **Aggarwal H., Imai N., Katoh N., Suri S.** Finding  $k$  points with minimum diameter and related problems // *J. Algorithms*. 1991. Vol. 12, No. 1. P. 38–56.
10. **Zukhba A. V.** NP-completeness of the problem of prototype selection in the nearest neighbor method // *Pattern Recognit. Image Anal.* 2010. Vol. 20, No. 4. P. 484–494.
11. **Banerjee S., Bhore S., Chitnis R.** Algorithms and hardness results for nearest neighbor problems in bicolored point sets // *Proc. 13th Latin Amer. Theor. Inform. Symp.* (Buenos Aires, Argentina, Apr. 16–19, 2018). Cham: Springer, 2018. P. 80–93. (Lect. Notes Comput. Sci.; Vol. 10807).
12. **Вапник В. Н.** Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979. 448 с.
13. **Burges C. J. C.** A Tutorial on support vector machines for pattern recognition // *Data Mining Knowl. Discov.* 1998. Vol. 2, No. 2. P. 121–167.
14. **Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Methods of recognition based on the function of rival similarity // *Pattern Recognit. Image Anal.* 2008. Vol. 18, No. 1. P. 1–6.
15. **Kariv O., Hakimi S.** An algorithmic approach to network location problems. II: The  $p$ -medians // *SIAM J. Appl. Math.* 1979. Vol. 37. P. 539–560.

Борисова Ирина Артёмовна

Статья поступила  
10 сентября 2018 г.  
После доработки —  
24 декабря 2019 г.  
Принята к публикации  
19 февраля 2020 г.

COMPUTATIONAL COMPLEXITY OF THE PROBLEM  
OF CHOOSING TYPICAL REPRESENTATIVES  
IN A 2-CLUSTERING OF A FINITE SET OF POINTS  
IN A METRIC SPACE

I. A. Borisova

<sup>1</sup> Sobolev Institute of Mathematics,  
4 Acad. Koptuyug Avenue, 630090 Novosibirsk, Russia<sup>2</sup> Novosibirsk State University,  
2 Pirogov Street, 630090 Novosibirsk, Russia

E-mail: biamia@mail.ru

**Abstract.** We consider the computational complexity of one extremal problem of choosing a subset of  $p$  points from some given 2-clustering of a finite set in a metric space. The chosen subset of points has to describe the given clusters in the best way from the viewpoint of some geometric criterion. This is a formalization of an applied problem of data mining which consists in finding a subset of typical representatives of a dataset composed of two classes based on the function of rival similarity. The problem is proved to be NP-hard. To this end, we polynomially reduce to the problem one of the well-known problems NP-hard in the strong sense, the  $p$ -median problem. Bibliogr. 15.

**Keywords:** NP-hard problem, typical representative, rival similarity,  $p$ -median problem, data mining.

## REFERENCES

1. **M. R. Garey** and **D. S. Johnson**, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).
2. **D. Aloise**, **A. Deshpande**, **P. Hansen**, and **P. Popat**, NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.* **75** (2), 245–248 (2009).
3. **S. Dasgupta**, The hardness of  $k$ -means clustering, in *Technical report CS2007-0890* (Univ. California, San Diego, 2008), pp. 1–6.

---

This research is supported by the Programme for Fundamental Scientific Research of SB RAS (Project 0314–2019–0015).

English version: Journal of Applied and Industrial Mathematics **14** (2), 242–248 (2020), DOI 10.1134/S1990478920020039.

4. **C. H. Papadimitriou**, Worst-case and probabilistic analysis of a geometric location problem, *SIAM J. Comput.* **10** (3), 542–557 (1981).
5. **S. Har-Peled** and **S. Mazumdar**, Coresets for  $k$ -means and  $k$ -median clustering and their applications, in *Proc. 36th Annu. ACM Sympos. Theory Comput., Chicago, IL, USA, June 13–15, 2004* (ACM, New York, 2004), pp. 291–300.
6. **L. Kaufman** and **P. J. Rousseeuw**, Clustering by means of medoids, in *Statistical Data Analysis Based on the  $L_1$ -Norm and Related Methods* (North Holland, Amsterdam, 1987), pp. 405–416.
7. **A. V. Kel'manov**, **A. V. Pyatkin**, and **V. I. Khandeev**, On the complexity of some max–min clustering problems, *Tr. Inst. Mat. Mekh. UrO RAN* **24** (4), 189–198 (2018) [Russian].
8. **A. V. Kel'manov** and **A. V. Pyatkin**, NP-hardness of some Euclidean problems of partition of a finite set of points, *Zh. Vychisl. Mat. Mat. Fiz.* **58** (5), 852–856 (2018) [Russian].
9. **H. Aggarwal**, **N. Imai**, **N. Katoh**, and **S. Suri**, Finding  $k$  points with minimum diameter and related problems, *J. Algorithms* **12** (1), 38–56 (1991).
10. **A. V. Zukhba**, NP-completeness of the problem of prototype selection in the nearest neighbor method, *Pattern Recognit. Image Anal.* **20** (4), 484–494 (2010).
11. **S. Banerjee**, **S. Bhore**, and **R. Chitnis**, Algorithms and hardness results for nearest neighbor problems in bicolored point sets, in *Proc. 13th Latin Amer. Theor. Inform. Symp., Buenos Aires, Argentina, Apr. 16–19, 2018* (Springer, Cham, 2018), pp. 80–93 (Lect. Notes Comput. Sci., Vol. 10807).
12. **V. N. Vapnik**, *The Restoration of Dependencies from Empirical Data* (Nauka, Moscow, 1974) [Russian].
13. **C. J. C. Burges**, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowl. Discov.* **2** (2), 121–167 (1998).
14. **N. G. Zagoruiko**, **I. A. Borisova**, **V. V. Dyubanov**, and **O. A. Kutnenko**, Methods of recognition based on the function of rival similarity, *Pattern Recognit. Image Anal.* **18** (1), 1–6 (2008).
15. **O. Kariv** and **S. Hakimi**, An algorithmic approach to network location problems. II: The  $p$ -medians, *SIAM J. Appl. Math.*, **37** (3), 539–560 (1979).

Irina A. Borisova

Received September 10, 2018

Revised December 24, 2019

Accepted February 19, 2020