ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 23 № 4 2017

УДК 519.16 + 519.85

НЕУЛУЧШАЕМАЯ ГАРАНТИРОВАННАЯ ОЦЕНКА ТОЧНОСТИ ДЛЯ ЗАДАЧИ О k МЕДИАНАХ НА ОТРЕЗКЕ [0,1][1]

М. Ю. Хачай, Д. М. Хачай, В. С. Панкратов

Одномерная задача кластеризации k-medians рассматривается в контексте игры двух лиц с нулевой суммой. Множество стратегий первого игрока совпадает с совокупностью выборок фиксированной длины из отрезка $[0,1]$. Стратегиями второго игрока являются всевозможные разбиения произвольной выборки данной длины на заданное число кластеров. В качестве платежной выступает функция, оценивающая качество кластеризации, значение которой численно совпадает с суммой уклонений элементов выборки от центров ближайших к ним кластеров. Как нетрудно убедиться, за исключением редких случаев данная игра не имеет цены. Для произвольных натуральных n и k строится верхняя оценка $0.5n/(2k-1)$ нижней цены игры. Обосновывается достижимость найденной оценки при $k > 1$ и достаточно больших $n = n(k)$. Тем самым показывается, что для произвольной выборки длины n может быть построена кластеризация методом k медиан так, что значение платежной функции не превысит найденной оценки, причем данная оценка достижима при произвольном числе кластеров и выборок достаточно большой длины. Полученные результаты нашли применение в комбинаторной оптимизации при обосновании полиномиальной разрешимости подклассов труднорешаемых экстремальных задач

Ключевые слова: кластеризация, задача о k медианах, достижимая оценка точности.

M. Yu. Khachai, D. M. Khachai, V. S. Pankratov. Attainable best guarantee for the accuracy of k-medians clustering in [0,1].

The scalar k-medians clustering problem is considered in the context of a two-player zero-sum game. The set of strategies of the first player coincides with a family of fixed-length samples from the interval $[0,1]$. The strategies of the second player are all possible partitions of an arbitrary sample of a given length into a given number of clusters. The quality of the clustering is evaluated by the payoff function equal to the sum of deviations of the elements from the centers of clusters nearest to them. It is easy to see that the game has no value except for rare cases. For arbitrary positive integers n and k, we establish an upper bound $0.5n/(2k-1)$ for the lower value of the game and prove its attainability for $k > 1$ and sufficiently large $n = n(k)$. Thus, we show that a clustering of an arbitrary sample of length n can be constructed by the k medians method so that the payoff does not exceed the obtained bound, and the bound is attainable for an arbitrary number of clusters and for sufficiently long samples. These results are applicable in combinatorial optimization in the proof of polynomial solvability of subclasses of intractable extremal problems.

Keywords: clustering, k-medians problem, attainable accuracy guarantee.

MSC: 90C27, 90C05, 62H30

DOI: 10.21538/0134-4889-2017-23-4-301-310

Введение

Кластеризация, один из известных подходов к анализу данных, состоит в поиске разбиения исходного набора данных на заданное число непересекающихся подмножеств (*кластеров*), минимизирующем ту или иную функцию потерь (см., например, [1;3]). В большинстве известных постановок задачи кластеризации критерий оптимизации задается в виде функции уклонений элементов выборки от ближайших к ним центров кластеров.

Условие задачи о k *медианах* (k-*medians*) может быть задано следующим образом. Фиксируются натуральное число $k > 1$ и метрическое пространство (X, ρ). Для заданной конечной выборки $\xi = (x_1, \ldots, x_n)$, $x_i \in X$, требуется найти разбиение множества $\mathbb{N}_n = \{1, \ldots, n\}$ на

[1]Исследования поддержаны Российским фондом фундаментальных исследований, гранты 16-07-00266 и 17-08-01385.

k непустых подмножеств — *кластеров* $C_1, \ldots, C_k$ и для каждого j-го кластера указать точку $c_j \in X$, именуемую его *центром*, так что

$$\sum_{j=1}^{k} \sum_{i \in C_j} \rho(x_i, c_j) = \sum_{i=1}^{n} \min\{\rho(x_i, c_1), \ldots, \rho(x_i, c_k)\} \to \min. \tag{1}$$

Как следует из (1), для произвольной выборки и всякого j центр c_j удовлетворяет соотношению

$$c_j \in \arg\min\Big\{ \sum_{i \in C_j} \rho(x_i, c) \colon c \in X \Big\},$$

т. е. является *медианой* подвыборки $\xi_j = (x_i \colon i \in C_j)$.

Известно, что задача k-medians NP-трудна (при условии, что параметр k является частью ее условия) [5] даже для евклидовой метрики и не имеет PTAS при условии $P \neq NP$. Тем не менее в конечномерных евклидовых пространствах задача эффективно аппроксимируема. Так, в работе [9] для произвольных фиксированных числа кластеров k и размерности пространства d построена рандомизированная линейная приближенная схема с трудоемкостью $O(2^{(k/\varepsilon)^{O(1)}} dn)$. В статье [6] обосновывается рандомизированная приближенная схема, трудоемкость которой $O(n+g(\varepsilon,d)\cdot(k\log n)^{O(1)})$, где $g(\varepsilon,d) = \exp(O((1-\log\varepsilon)/\varepsilon)^{d-1})$, полиномиально зависит от числа кластеров. Более того, известно что задача k-medians полиномиально разрешима на вещественной прямой. По-видимому, наиболее эффективный точный алгоритм для этой постановки задачи предложен в работе [4] и обладает трудоемкостью $O(n \log n + kn)$.

Наряду с алгоритмическими вопросами, связанными с разработкой методов кластеризации, позволяющих для каждой отдельно взятой выборки эффективно строить разбиение ее на кластеры, не менее важными с точки зрения приложений в анализе данных и вычислительной геометрии [2;7] представляются вопросы обоснования универсальных оценок качества кластеризации, гарантированных для целого семейства постановок задачи k-medians. В данной статье находится такая оценка для задачи k-medians на вещественной прямой.

1. Постановка задачи

Рассмотрим антагонистическую игру двух лиц с нулевой суммой, порожденную задачей о k медианах. Зададимся натуральными числами n и k, большими единицы. Допустимыми стратегиями первого игрока являются выборки $\xi = (x_1, \ldots, x_n)$, $x_i \in [0,1]$. Стратегии второго игрока — конечные последовательности $\sigma = (c_1, \ldots, c_k)$, $c_i \in [0,1]$. Платежная функция $F(\xi, \sigma)$ задается соотношением $F(\xi,\sigma) = \sum\limits_{i=1}^{n} \min\{|x_i - c_1|, \ldots, |x_i - c_k|\}$. Пользуясь обычными обозначениями $\Phi(\xi) = \inf\limits_{\sigma \in [0,1]^k} F(\xi,\sigma)$ и $\Psi(\sigma) = \sup\limits_{\xi \in [0,1]^n} F(\xi,\sigma)$, полагаем, что цели первого и второго игроков определяются с точки зрения принципа гарантированного результата и состоят в поиске нижней $v_*(n,k) = \sup\limits_{\xi \in [0,1]^n} \Phi(\xi)$ и верхней $v^*(n,k) = \inf\limits_{\sigma \in [0,1]^k} \Psi(\sigma)$ цены игры соответственно.

Нетрудно убедиться в том, что для любых $k > 1$ и $n > 0$ игра не имеет цены, т. е. $v_*(n,k) < v^*(n,k)$. По многим причинам, восходящим к приложениям в анализе данных, комбинаторной оптимизации и вычислительной геометрии, представляется важным построение верхних оценок для $v_*(n,k)$, позволяющих охарактеризовать гарантированную точность решения задачи k-medians для произвольной выборки длины n. Конечно в качестве такой оценки всегда может быть выбрана верхняя цена игры $v^*(n,k)$, однако при больших значениях n эта оценка становится слишком неточной.

В статье строится более точная верхняя оценка $B(n,k) = 0.5n/(2k-1)$ и обосновывается ее достижимость при произвольном $k > 1$ и достаточно больших значениях n. Фактически нами

показывается, что произвольной выборке $\xi \in [0,1]^n$ может быть сопоставлена подходящая конечная последовательность $\sigma_\xi = (c_1, \ldots, c_k)$ центров кластеров так, что

$$\Phi(\xi) = \inf_{\sigma \in [0,1]^k} F(\xi, \sigma) = F(\xi, \sigma_\xi) \le B(n,k).$$

Более того, для произвольных числа кластеров k и достаточно большой длины выборки $n = n(k)$ мы описываем метод построения выборок $\xi^* = \xi^*(n,k)$ (являющихся оптимальными стратегиями первого игрока) таких, что $F(\xi^*, \sigma_{\xi^*}) = B(n,k)$.

2. Сведение к задаче линейного программирования

Для удобства изложения введем несколько предварительных допущений. Без ограничения общности всюду ниже будем полагать, что элементы произвольной выборки $\xi = (x_1, \ldots, x_n)$ упорядочены по возрастанию. Кроме того, будем полагать, что произвольный кластер $C = \{i_1, \ldots, i_m\} \subset \mathbb{N}_n$ наследует данное свойство, т. е. $i_1 < \ldots < i_m$ и $x_{i_1} \le \ldots \le x_{i_m}$. Как следствие, сумма уклонений представителей кластера C от его медианы c может быть представлена в виде

$$W(C) = \sum_{l=1}^{m} |x_{i_l} - c| = \sum_{l=1}^{\lfloor m/2 \rfloor} (c - x_{i_l}) \; + \sum_{l=\lceil m/2 \rceil + 1}^{m} (x_{i_l} - c) = - \sum_{l=1}^{\lfloor m/2 \rfloor} x_{i_l} \; + \sum_{l=\lceil m/2 \rceil + 1}^{m} x_{i_l}.$$

Подвыборки $(x_{i_l} \colon 1 \le l \le \lfloor m/2 \rfloor)$ и $(x_{i_l} \colon \lceil m/2 \rceil + 1 \le l \le, m)$ договоримся называть *нижней* и *верхней* половинами кластера C.

Далее, для произвольного разбиения на кластеры $C_1 \cup \ldots \cup C_k = \mathbb{N}_n$ введем обозначение $m_j = |C_j| > 0$ и договоримся для каждых $j_1 < j_2$ и произвольных $i_1 \in C_{j_1}$ и $i_2 \in C_{j_2}$ полагать выполненным соотношение $i_1 < i_2$ и, как следствие, $x_{i_1} \le x_{i_2}$.

В наших предположениях значение $\Phi(\xi)$ для произвольной выборки ξ может быть выражено исключительно в терминах разбиений $C_1 \cup \ldots \cup C_k = \mathbb{N}_n$ и задается соотношением

$$\Phi(\xi) = \min\Big\{ \sum_{j=1}^{k} \sum_{i \in C_j} |x_i - c_j| \colon C_1 \cup \ldots \cup C_k = \mathbb{N}_n \Big\} = \min\Big\{ \sum_{j=1}^{k} \Big(- \sum_{i=1}^{\lfloor m_j/2 \rfloor} x_{i+m_1+\ldots+m_{j-1}}$$

$$+ \sum_{i=\lceil m_j/2 \rceil + 1}^{m_j} x_{i+m_1+\ldots+m_{j-1}} \Big) \colon m_1, \ldots, m_k > 0, \; m_1 + \ldots + m_k = n \Big\}.$$

Таким образом, искомая нижняя цена игры $v_*(n,k) = \sup\limits_{\xi \in [0,1]^n} \Phi(\xi)$ совпадает с оптимальным значением задачи линейного программирования (2)

$$v_*(n,k) = \max u :$$

$$\sum_{j=1}^{k} \Big(- \sum_{i=1}^{\lfloor m_j/2 \rfloor} x_{i+m_1+\ldots+m_{j-1}} \; + \sum_{i=\lceil m_j/2 \rceil + 1}^{m_j} x_{i+m_1+\ldots+m_{j-1}} \Big) \ge u \qquad \Big(\sum_{j=1}^{k} m_j = n \Big), \tag{2}$$

$$0 \le x_1 \le \ldots \le x_n \le 1.$$

Задача (2) разрешима при произвольных рассматриваемых значениях параметров n и k. Ее система ограничений определена в пространстве $n+1$ переменной; число входящих в нее неравенств определяется количеством разбиений числа n на k (ненулевых) натуральных слагаемых, быстро растущим с ростом n. Опираясь на свойства симметрии и выпуклость множества оптимальных решений задачи ЛП, размерность задачи (2) удается сократить, как, в частности показывается в следующей простой лемме.

Лемма. *Для произвольных $n, k > 1$ задача* (2) *обладает оптимальным решением, удовлетворяющим соотношению $x_i + x_{n-i+1} = 1$.*

Доказательство. В самом деле, пусть $[x', u']$, где $x' = [x_1, x_2, \dots x_n]$, — произвольное оптимальное решение задачи (2). Очевидная симметрия влечет оптимальность решения $[x'', u]$, в котором $x'' = [1 - x_n, 1 - x_{n-1}, \dots, 1 - x_1]$. Следовательно, вектор $[y, u]$, где $y = (x' + x'')/2$, также является оптимальным решением в силу выпуклости оптимального множества задачи линейного программирования. Поскольку y удовлетворяет следующему простому соотношению $y_{n-i+1} = (x_{n-i+1} + (1 - x_i))/2 = 1 - (x_i + (1 - x_{n-i+1}))/2 = 1 - y_i$, решение $[y, u]$ является искомым. Лемма доказана.

Данная лемма позволяет ограничиться рассмотрением только симметричных решений, удовлетворяющих соотношению $x_i + x_{n-i+1} = 1$, и, соответственно, сократить число переменных в задаче (2) вдвое. Опираясь на аналогичные соображения симметрии, без ограничения общности можно полагать выполненными соотношения $m_i \le m_{k-i+1}$ для произвольного $i \le k/2$, что позволяет существенно сократить число ограничений задачи (2) и ускорить поиск точного значения нижней цены игры $v_*(n, k)$.

3. Верхняя оценка для $v_*(n, k)$

Для построения верхней оценки нижней цены игры $v_*(n, k)$ мы воспользуемся соотношениями двойственности для задачи (2).

Теорема 1. *Для произвольных $k > 1$, $n > 1$ и выборки $\xi = (x_1, \dots, x_n) \in [0, 1]^n$ существует конечная последовательность $\sigma_\xi = (c_1, \dots, c_k) \in [0, 1]^k$, для которой*

$$F(\xi, \sigma_\xi) \le \frac{n}{2(2k-1)}.$$

Частный случай теоремы 1 для $k = 2$ опубликован ранее, например, в работе [8].

Доказательство. Рассуждения удобно проводить, рассматривая длину выборки n по модулю $2k - 1$. В самом деле, пусть $n = (2k - 1)t + r$ для некоторого натурального t и остатка $r \in [0, \dots, 2k - 2]$. Покажем, что для произвольного значения r в системе ограничений задачи (2) удается найти подходящие неравенства, неотрицательная линейная комбинация которых влечет соотношение $u \le B(n, k) = 0.5n/(2k - 1)$. Мы остановимся на доказательстве для остатков $r = 0$ и $r = 1$. Для других значений r доказательство может быть построено по аналогии.

Случай $r = 0$. Пусть $n = (2k - 1)t$. Полагая без ограничения общности t и k четными, рассмотрим неравенство

$$-\sum_{i=1}^{t/2} x_i + \sum_{i=t/2+1}^{t} x_i - \sum_{i=t+1}^{2t} x_i + \sum_{i=2t+1}^{3t} x_i - \dots - \sum_{i=(k-1)t+1}^{kt-t/2} x_i - \sum_{i=kt-t/2+1}^{kt} x_i$$

$$+ \sum_{i=kt+1}^{(k+1)t} x_i + \dots - \sum_{i=(2k-3)t+1}^{(2k-2)t} x_i + \sum_{i=(2k-2)t+1}^{(2k-1)t} x_i \ge u, \tag{3}$$

соответствующее разбиению заданной выборки на кластеры мощности $m_1 = t$; $m_2 = 2t$; $\dots$; $m_k = 2t$. Пользуясь соотношением $x_i + x_{n-i+1} = 1$, преобразуем неравенство (3) к виду

$$-\sum_{i=1}^{t/2} x_i + \sum_{i=t/2+1}^{t} x_i - \sum_{i=t+1}^{2t} x_i + \sum_{i=2t+1}^{3t} x_i - \dots - \sum_{i=(k-1)t+1}^{kt-t/2} x_i - \sum_{i=(k-1)t+1}^{kt-t/2} (1 - x_i)$$

$$+ \sum_{i=(k-2)t+1}^{(k-1)t} (1 - x_i) + \dots - \sum_{i=t+1}^{2t} (1 - x_i) + \sum_{i=1}^{t} (1 - x_i) \ge u. \tag{4}$$

Мощности кластеров, порождающих искомую подсистему

m_1	m_2	...	m_k
t+1	$2t$	...	$2t$
t	$2t$+1	...	$2t$
...	...	...	...
t	$2t$	...	$2t$+1

После приведения подобных имеем

$$u + \sum_{i=1}^{t/2} x_i \le t/2, \tag{5}$$

и, следовательно,

$$u \le t/2 = \frac{n}{2(2k-1)} \tag{6}$$

ввиду неотрицательности переменных x_i.

С л у ч а й $r = 1$. Пусть $n = (2k-1)t + 1$. Покажем, что для произвольных $k > 1$ и $t > 1$ подсистема из k неравенств, порождаемых кластеризациями, состоящих из кластеров, мощности которых приведены в таблице выше, влечет справедливость неравенства

$$u \le \frac{(2k-1)t+1}{2(2k-1)}. \tag{7}$$

Конкретнее, покажем, что неравенство (7) является следствием линейной комбинации неравенств данной подсистемы с коэффициентами $1, 2, \ldots, 2$ соответственно.

Доказательство проведем индукцией по k, как и ранее, без ограничения общности полагая t четным.

База индукции. При $k = 2$ искомая подсистема состоит из неравенств

$$-\sum_{i=1}^{1/2} x_i + \sum_{i=t/2+2}^{t+1} x_i - \sum_{i=t+2}^{2t+1} x_i + \sum_{i=2t+2}^{3t+1} x_i \ge u, \tag{8}$$

$$-\sum_{i=1}^{1/2} x_i + \sum_{i=t/2+1}^{t} x_i - \sum_{i=t+1}^{2t} x_i + \sum_{i=2t+2}^{3t+1} x_i \ge u, \tag{9}$$

соответствующих кластеризациям $m_1 = t+1,\ m_2 = 2t$ и $m_1 = t,\ m_2 = 2t+1$ соответственно. Пользуясь соотношением $x_i + x_{3t+1-i+1} = 1$ и проводя преобразования по аналогии с (3)–(6), приводим неравенства (8) и (9) к виду

$$u + 2\sum_{i=1}^{1/2} x_i + x_{t/2+1} - 2x_{t+1} \le t/2 - 1/2, \tag{10}$$

$$u + 2\sum_{i=1}^{t/2} x_i + x_{t+1} \le t/2 + 1/2. \tag{11}$$

Сворачивая подсистему (10), (11) с коэффициентами 1 и 2 соответственно, убеждаемся в том, что неравенство $u \le t/2 + 1/6 = (3t+1)/6$ является ее следствием, чем завершаем обоснование базы индукции.

Шаг индукции. Допустим, для k предположение индукции верно. Докажем его истинность при $k+1$. Проведенные ниже рассуждения предполагают четность k и легко могут быть адаптированы на случай нечетных значений параметра.

Рассмотрим подсистему, построенную на предыдущем шаге индукции. Через μ обозначим номер кластера, содержащего центральный элемент $kt - t/2 + 1$ множества индексов $\{1, 2, \ldots, (2k-1)t+1\}$ кластеризуемых точек. Нетрудно убедиться, что $\mu = k/2 + 1$. Зафиксируем произвольное неравенство рассматриваемой подсистемы, определяемое l-й строкой таблицы, приведенной выше, и порождаемой ей кластеризацией

$$|C_1| = t, \ldots, |C_l| = 2t + 1, \ldots, |C_k|. \tag{12}$$

Допустим $l > \mu$ и, следовательно, $|C_\mu| = 2t$. Переходя к случаю $k+1$, представим произвольную выборку ξ длины $(2k+1)t+1$ в виде

$$\xi = (x_1, \ldots, x_{kt-t/2}, y_{\alpha_1}, \ldots, y_{\alpha_t}, x_{kt-t/2+1}, y_{\alpha_{t+1}}, \ldots, y_{\alpha_{2t}}, x_{kt-t/2+2}, \ldots, x_{(2k-1)t+1}),$$

где $\xi' = (x_1, \ldots, x_{(2k-1)t+1})$ — порождающая ее выборка длины $(2k-1)t+1$. Кластеризации (12) выборки ξ' сопоставим кластеризацию выборки ξ, построенную по следующим правилам (см. рис. 1).

1. Кластеры $C_1, \ldots, C_{\mu-1}, C_{\mu+1}, \ldots, C_k$ сохраняют прежние значения.

2. Кластер C_μ замещается парой новых кластеров, назовем их C'_μ и C''_μ, где

$$C'_\mu = \{i \in C_\mu \colon (k-1)t + 1 \le i \le kt - t/2\} \cup \{\alpha_1, \ldots, \alpha_t\} \cup \{kt - t/2 + 1\} \cup \{\alpha_{t+1}, \ldots, \alpha_{3t/2-1}\},$$

$$C''_\mu = \{\alpha_{3t/2}, \ldots, \alpha_{2t}\} \cup \{i \in C_\mu \colon k - t/2 + 2 \le i \le (k+1)t\}.$$

По построению неравенство, соответствующее построенной кластеризации, обладает следующими свойствами:

1) коэффициенты при переменных $x_1, \ldots, x_{kt-t/2}, x_{kt-t/2+2}, \ldots, x_{(2k-1)t+1}$ наследуются из исходного неравенства;

2) переменные $y_{\alpha_1}, \ldots, y_{\alpha_{2t}}$ симметричны относительно $x_{kt-t/2+1}$ и, следовательно, удовлетворяют соотношению $y_{\alpha_j} + y_{\alpha_{2t-j+1}} = 1$;

3) переменная $x_{kt-t/2+1} = 1/2$ входит в верхнюю половину кластера C'_μ, в то время как в исходной кластеризации она входила в нижнюю половину кластера C_μ.

Опуская слагаемые, соответствующие переменным x_i, выпишем левую часть нового неравенства:

$$\ldots - \sum_{j=1}^{t/2} y_{\alpha_j} + \sum_{j=t/2+1}^{t} y_{\alpha_{j+1}} + \sum_{j=t+1}^{3t/2-1} y_{\alpha_j} - \sum_{j=3t/2}^{2t} y_{\alpha_j} + \ldots$$

$$= \ldots - \sum_{j=1}^{t/2} y_{\alpha_j} + \sum_{j=t/2+1}^{t} y_{\alpha_j} + \sum_{j=t/2+2}^{t} (1 - y_{\alpha_j}) - \sum_{j=1}^{t/2+1} (1 - y_{\alpha_j}) + 1 \ldots = \ldots - 2y_{\alpha_{t/2+1}} \ldots .$$

После приведения подобных имеем

$$u + \mathbf{lhs}(l) - 2y_{\alpha_{t/2+1}} \le \mathbf{rhs}(l) - 1,$$

где $u + \mathbf{lhs}(l)$ и $\mathbf{rhs}(l)$ обозначают левую и правую части исходного l-го неравенства соответственно.

Проводя рассуждения по аналогии в случае $l < \mu$, получим неравенство

$$u + \mathbf{lhs}(l) + 2y_{\alpha_{t/2+1}} \le \mathbf{rhs}(l) + 1.$$

Осталось рассмотреть случай $l = \mu$. Здесь $|C_\mu| = 2t + 1$ и возможны два варианта:
1) $|C'_m| = 2t$, $|C''_m| = 2t + 1$ и 2) $|C'_m| = 2t + 1$, $|C''_m| = 2t$.

Легко видеть, что вариант 1) аналогичен случаю $l > \mu$, и полученное неравенство после приведения подобных примет вид

$$u + \mathbf{lhs}(\mu) - 2y_{\alpha_{t/2+1}} \le \mathbf{rhs}(\mu) - 1.$$

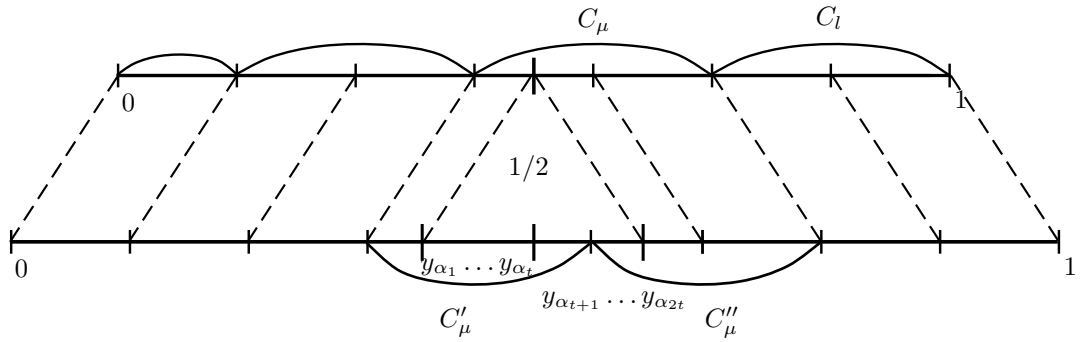

Рис. 1. Схема индуктивного перехода.

Вариант 2) заслуживает отдельного рассмотрения. Как показано выше, неравенство (3) соответствует разбиению выборки $\xi' = (x_1, \dots, x_{(2k-1)t})$ на кластеры с мощностями $m_1 = t,\ m_2 = \dots = m_k = 2t$. Выборка ξ может быть получена из выборки ξ' путем пополнения ее $2t+1$ переменными

$$\xi = (x_1, \dots, x_{kt-t/2}, y_{\alpha_1}, \dots, y_{\alpha_t}, z = 1/2, y_{\alpha_{t+1}}, \dots, y_{\alpha_{2t}}, x_{kt-t/2+1}, \dots, x_{(2k-1)t}).$$

Учитывая эквивалентность неравенств (3) и (5) и проводя рассуждения по аналогии со случаем $l > \mu$, показываем, что рассматриваемому случаю соответствует неравенство

$$u + 2\sum_{i=1}^{t/2} x_i + y_{\alpha_{t/2+1}} \le t/2 + 1/2.$$

Таким образом, в системе ограничений задачи (2) для $n = (2k+1)t+1$ и $k+1$ нами найдена подсистема из $k+1$ неравенства

$$\begin{array}{rcll} u + \mathbf{lhs}(l) + 2y_{\alpha_{t/2+1}} & \le & \mathbf{rhs}(l) + 1 & (1 \le l \le \mu - 1) \\ u + \mathbf{lhs}(\mu) - 2y_{\alpha_{t/2+1}} & \le & \mathbf{rhs}(\mu) - 1 & \\ u + 2\sum\limits_{i=1}^{t/2} x_i + y_{\alpha_{t/2+1}} & \le & t/2 + 1/2 & \\ u + \mathbf{lhs}(l) - 2y_{\alpha_{t/2+1}} & \le & \mathbf{rhs}(l) - 1 & (\mu + 1 \le l \le k). \end{array} \tag{13}$$

По предположению индукции линейная свертка подсистемы (13) с коэффициентами $1, 2, \dots, 2$ примет вид

$$(2k+1)u + \sum_{i=1}^{kt-t/2} a_i x_i + b y_{\alpha_{t/2+1}} \le t + (2k-1)\frac{(2k-1)t+1}{2(2k-1)} = ((2k+1)t+1)/2,$$

причем $a_i \ge 0$. Кроме того, нетрудно проверить, что $b = 0$. Тем самым

$$u \le \frac{(2k+1)t+1}{2(2k+1)} = B((2k+1)t+1, k+1)$$

в силу неотрицательности переменных x_i, что завершает обоснование индукции и доказательство теоремы в целом.

4. Достижимость

Теорема 2. *Для произвольного $k > 1$ и $n \ge n_0(k)$ оценка $B(n,k)$ достижима.*

Д о к а з а т е л ь с т в о проведем для простейшего нетривиального случая $k = 3$. При больших значениях k рассуждения могут быть проведены по аналогии. Для произвольного достаточно большого значения n мы укажем такую выборку ξ^* длины n, для которой справедливо соотношение $F(\xi^*, \sigma_{\xi^*}) = B(n, 3)$. Приведем возможные варианты построения искомых выборок, представляя их длину по модулю 6.

С л у ч а й $n = 6q + 2s$ для $0 \le s \le 2$. Сконцентрируем элементы выборки ξ^* в позициях $0, \frac{1}{5}, \ldots, \frac{4}{5}, 1$ с кратностями $q, q, q+s, q+s, q, q$ соответственно, как указано на рис. 2:

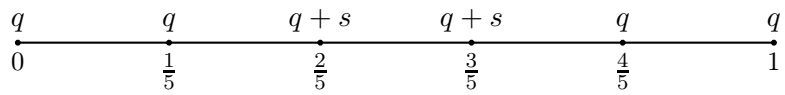

Рис. 2. Случай четных остатков.

К сожалению, для нечетных остатков общей схемы построения выборки найти не удается. Рассмотрим каждый нечетный остаток в отдельности.

С л у ч а й $n = 6q + 5$. Как и для четных остатков разместим элементы выборки в фиксированном, независящем от q, множестве позиций (см. рис. 3):

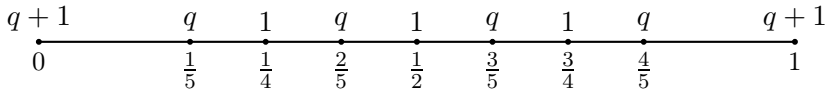

Рис. 3. Случай $n = 6q + 5$.

С л у ч а й $n = 6q + 3$. Как и в предыдущем случае, разместим элементы выборки, как указано на рис. 4:

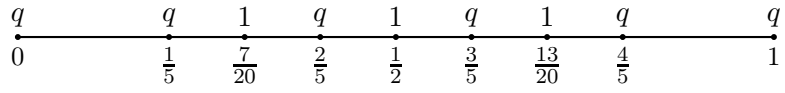

Рис. 4. Случай $n = 6q + 3$.

С л у ч а й $n = 6q + 1$ удобно рассмотреть, разбив его на два подслучая, учитывающие четность q. При $q = 2t$ разместим элементы выборки согласно рис. 5:

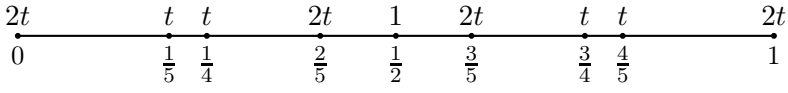

Рис. 5. Случай $n = 6q + 1$ при четном q.

а при $q = 2t + 1$ так, как указано на рис. 6:

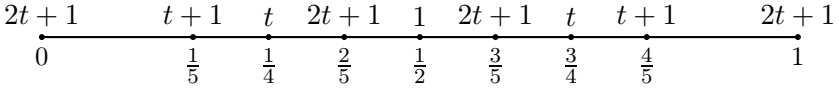

Рис. 6. Случай $n = 6q + 1$ при нечетном q.

Непосредственной проверкой убеждаемся в том, что в каждом из рассмотренных случаев выполняется соотношение $F(\xi, \sigma_\xi) = B(n, 3)$, при произвольном $n \ge 19$. Более того,

можно показать, что число перебираемых вариантов определяется исключительно числом различных позиций и не изменяется с ростом n. А именно при обосновании того, что для произвольной кластеризации C_1, C_2, C_3 с центрами $\sigma = (c_1, c_2, c_3)$ выполняется соотношение $F(\xi, \sigma) \geq B(n, 3)$, достаточно ограничиться лишь теми вариантами, в которых каждая из указанных на рис. 2–6 точек входит (или не входит) в кластер в соответствии с ее кратностью.

В самом деле, пусть точка p имеет кратность q и входит в два граничащих друг с другом кластера, например, C_1 и C_2 с кратностями q_1 и q_2, где $q_1 + q_2 = q$, соответственно. Обозначим через c_1 и c_2 медианы кластеров и предположим для определенности, что $p - c_1 \geq c_2 - p$. Перенеся точку p во второй кластер в размере ее кратности, имеем для полученных кластеров $\tilde{C}_1$ и $\tilde{C}_2$

$$W(\tilde{C}_1) + W(\tilde{C}_2) \leq \sum\{|x_i - c_1| \colon i \in C_1, x_i \neq p\} + \sum\{|x_i - c_2| \colon i \in C_2, x_i \neq p\} + q \cdot |c_2 - p|$$

$$\leq \sum\{|x_i - c_1| \colon i \in C_1\} + \sum\{|x_i - c_2| \colon i \in C_2\} = W(C_1) + W(C_2).$$

Следовательно, кластеризация C_1, C_2, C_3 мажорирует кластеризацию $\tilde{C}_1, \tilde{C}_2, C_3$.

Теорема доказана.

Заключение

В статье для произвольных натуральных n и k построена верхняя оценка $0.5n/(2k-1)$ для нижней цены антагонистической игры двух лиц, порожденной задачей кластеризации методом k медиан, и обоснована ее достижимость при произвольном числе кластеров $k > 1$ и достаточно большой длине выборки $n = n(k)$. Полученные результаты нашли приложение в задачах комбинаторной оптимизации, в частности при обосновании полиномиальной разрешимости одной геометрической постановки обобщенной задачи коммивояжера [7]. Открытым остается вопрос о распространении полученных результатов на случай пространств большей размерности.

СПИСОК ЛИТЕРАТУРЫ

1. **Aggarwal C. C., Reddy C. K.** Data clustering: algorithms and applications. Bosa Roca: Taylor & Francis Inc., 2013. 652 p. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Ser.) ISBN: 9781466558212 .
2. **Ben-David S.** Computational feasibility of clustering under clusterability assumptions [e-resource]. CoRR abs/1501.00437, 2015. URL: http://arxiv.org/abs/1501.00437.
3. **Duda R. O., Hart P. E., Stork D. G.** Pattern classification. N. Y.: Wile, 2001. 680 p. ISBN: 978-0-471-05669-0 .
4. Fast exact k-means, k-medians and bregman divergence clustering in 1d [e-resource] / A. Grønlund, K. G. Larsen, A. Mathiasen, J. S. Nielsen // CoRR abs/1701.07204, 2017. URL: http://arxiv.org/abs/1701.07204 .
5. **Guruswami V., Indyk P.** Embeddings and non-approximability of geometric problems // Proc. of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03). Philadelphia: Society for industrial and applied mathematics, 2003. P. 537–538. ISBN: 0-89871-538-5 .
6. **Har-Peled S., Mazumdar S.** On coresets for k-means and k-median clustering // Proc. of the Thirty-Sixth Annual ACM Symposium on Theory of Computing (STOC '04). N. Y.: ACM, 2004. P. 291–300. doi: 10.1145/1007352.1007400 .
7. **Khachay M., Neznakhina K.** Generalized pyramidal tours for the generalized traveling salesman problem // Lecture Notes in Computer Science. 2017. Vol. 10627. P. 265–277. doi: 10.1007/978-3-319-71150-8-23 .
8. **Khachay M., Pankratov V., Khachay D.** Attainable best guarantee for the accuracy of k-medians clustering in [0,1] // 8^{th} International Conf. Optimization and Applications (OPTIMA2017) / eds. Y. G. Evtushenko, et al. Aachen: CEUR Workshop Proceedings, 2017. P. 322–327.
9. **Kumar A., Sabharwal Y., Sen S.** Linear-time approximation schemes for clustering problems in any dimensions // J. ACM. Feb. 2010. Vol. 57(2). P. 5:1–5:32. doi: 10.1145/1667053.1667054 .

Хачай Михаил Юрьевич
д-р физ.-мат. наук, проф. РАН, зав. отделом
Институт математики и механики им. Н. Н. Красовского УрО РАН
Уральский федеральный университет, Омский государственный технический университет
e-mail: mkhachay@imm.uran.ru

Поступила 22.09.17

Хачай Даниил Михайлович
студент, математик
Институт математики и механики им. Н. Н. Красовского УрО РАН
Уральский федеральный университет
e-mail: dmx@imm.uran.ru

Панкратов Василий Сергеевич
аспирант, математик
Институт математики и механики им. Н. Н. Красовского УрО РАН
e-mail: pankratov.vs@gmail.com

REFERENCES

1. Aggarwal C. C., Reddy C. K. *Data clustering: algorithms and applications*. Bosa Roca: Taylor & Francis Inc., 2013, Chapman & Hall/CRC Data Mining and Knowledge Discovery Ser., 652 p. ISBN: 9781466558212 .
2. Ben-David S. Computational feasibility of clustering under clusterability assumptions, CoRR abs/1501.00437, 2015, 27 p. Available at: http://arxiv.org/abs/1501.00437.
3. Duda R. O., Hart P. E., Stork D. G. *Pattern classification*. NY: Wile, 2001, 680 p. ISBN: 978-0-471-05669-0 .
4. Grønlund A., Larsen K. G., Mathiasen A., Nielsen J. S. Fast exact k-means, k-medians and bregman divergence clustering in 1d. CoRR abs/1701.07204, 2017, 12 p. URL: http://arxiv.org/abs/1701.07204 .
5. Guruswami V., Indyk P. Embeddings and non-approximability of geometric problems. *Proc. of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA '03). Philadelphia: Society for Industrial and Applied Mathematics, 2003, pp. 537–538. ISBN: 0-89871-538-5 .
6. Har-Peled S., Mazumdar S. On coresets for k-means and k-median clustering. *Proc. of the Thirty-Sixth Annual ACM Symposium on Theory of Computing* (STOC '04). NY: ACM, 2004, pp. 291–300. doi: 10.1145/1007352.1007400 .
7. Khachay M., Neznakhina K. Generalized pyramidal tours for the generalized traveling salesman problem. *Lecture Notes in Computer Science*, 2017, vol. 10627, pp. 265–277. doi: 10.1007/978-3-319-71150-8-23 .
8. Khachay M., Pankratov V., Khachay D. Attainable best guarantee for the accuracy of k-medians clustering in [0,1]. *8^{th} International Conf. Optimization and Applications* (OPTIMA2017), Y. G. Evtushenko, et al. eds., Aachen: CEUR Workshop Proceedings, 2017, pp. 322–327.
9. Kumar A., Sabharwal Y., Sen S. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM. Feb.*, 2010, vol. 57(2), pp. 5:1–5:32. doi: 10.1145/1667053.1667054 .

The paper was received by the Editorial Office on September 22, 2017.

Mikhail Yur'evich Khachai, Dr. Phys.-Math. Sci., Prof., Krasovskii Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, Yekaterinburg, 620990 Russia; Ural Federal University, Ekaterinburg, 620002 Russia; Omsk State Technical University, Omsk, 644050 Russia, e-mail: mkhachay@imm.uran.ru .

Daniil Mikhailovich Khachai, graduate student, Krasovskii Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, 620990 Russia; Institute of Mathematics and Computer Science, Ural Federal University, Ekaterinburg, 620002 Russia,
e-mail: dmx@imm.uran.ru .

Vasiliy Sergeevich Pankratov, doctoral student, Krasovskii Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, 620990 Russia,
e-mail: pankratov.vs@gmail.com .