ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2020, том 60, № 1, с. 151–158

УДК 519.72

О СЛОЖНОСТИ НЕКОТОРЫХ КВАДРАТИЧНЫХ ЗАДАЧ РАЗБИЕНИЯ КОНЕЧНОГО МНОЖЕСТВА ТОЧЕК ЕВКЛИДОВА ПРОСТРАНСТВА НА СБАЛАНСИРОВАННЫЕ КЛАСТЕРЫ[1)]

© 2019 г. А. В. Кельманов[1,2]*, А. В. Пяткин[1,2,], В. И. Хандеев[1,2,***]**

[1] *630090 Новосибирск, пр-т акад. Коптюга, 4, Ин-т матем. им. С.Л. Соболева, Россия;*

[2] *630090 Новосибирск, ул. Пирогова, 2, Новосибирский гос. ун-т, Россия*

**e-mail: kelm@math.nsc.ru*

***e-mail: artem@math.nsc.ru*

****e-mail: khandeev@math.nsc.ru*

Поступила в редакцию 20.05.2019 г.
Переработанный вариант 20.05.2019 г.
Принята к публикации 18.09.2019 г.

Рассматриваются три родственные между собой задачи разбиения N-элементного множества точек d-мерного евклидова пространства на два кластера так, чтобы сбалансировать значения: (1) внутрикластерного квадратичного разброса, нормированного на размер кластера, в первой задаче; (2) внутрикластерного квадратичного разброса, во второй задаче; (3) мощностно-взвешенного внутрикластерного разброса, в третьей задаче. Доказано, что все эти задачи NP-полны. Библ. 17.

Ключевые слова: евклидово пространство, сбалансированное разбиение, квадратичный разброс, нормированный на размер кластера разброс, мощностно-взвешенный разброс, NP-полнота.

DOI: 10.31857/S0044466919110061

1. ВВЕДЕНИЕ

Предметом исследования настоящей работы являются проблемы дискретной оптимизации. Мы анализируем три родственные задачи сбалансированного по внутрикластерному разбросу разбиения конечного множества точек евклидова пространства. Цель работы – анализ статуса вычислительной сложности задач.

Исследование мотивировано открытостью указанного вопроса, а также важностью рассматриваемых задач как для математической теории оптимизации и вычислительной математики, так и для приложений, среди которых, в частности, анализ данных (Data analysis), интерпретация данных (Data mining) и статистика (Statistics). Примеры приложений приведены в следующем разделе.

Статья имеет следующую структуру. В разд. 2 мы приводим несколько замечаний, поясняющих мотивацию нашего исследования. Здесь же даны математические формулировки задач и показана их связь с актуальными и прикладными проблемами. В следующем разделе мы доказываем, что все рассматриваемые задачи NP-полны.

2. ФОРМУЛИРОВКИ ЗАДАЧ И СВЯЗАННЫЕ С НИМИ ПРОБЛЕМЫ

Классическая математическая статистика и относительно новая дисциплина Data mining [1], [2], [3], [4] имеют похожие цели: и та, и другая ориентированы на анализ структур экспериментальных данных. Тем не менее имеется существенное отличие этих дисциплин. Оно обусловлено отличием математических задач, на решение которых фокусируются эти дисциплины, и, как следствие, – отличием создаваемых и применяемых математических инструментов в виде методов и алгоритмов.

[1)]Работа выполнена при финансовой поддержке РФФИ, проекты 19-01-00308 и 18-31-00398, программы ФНИ РАН, проекты 0314-2019-0015 и 0314-2019-0014, а также программы Тор-5-100 Министерства образования и науки РФ.

Статистика ориентирована исключительно на анализ однородных выборочных данных (т.е. данных, принадлежащих одному и тому же распределению) или сопоставление (сравнение) выборочных данных, принадлежащих разным однородным распределениям. Data mining ставит своей целью выяснение (дословно – "раскопку") структуры данных, которые, как правило, принадлежат разным распределениям, в условиях отсутствия информации о соответствии данных распределению [1], [2], [3], [4]. По сути, статистические методы как инструменты пригодны лишь для применения к частным случаям разнообразных проблем в Data mining.

К сожалению, весь хорошо развитый аппарат математической статистики оказывается совершенно непригодным для применения на практике в ситуациях, когда соответствие выборочных (экспериментальных) данных распределению отсутствует. В этих ситуациях, типичных для Data mining, требуется сначала найти подходящее (адекватное) разбиение данных на однородные группы (кластеры). Лишь после получения такого разбиения на однородные кластеры классические методы математической статистики становятся применимыми корректно с математической точки зрения. Так называемый разведочный поиск разбиений, адекватных данным, на кластеры является одной из ключевых проблем в Data mining. Вопросы адекватности разбиений (т.е. вопросы соответствия этих разбиений и экспериментальных данных) лежат вне рамок настоящей работы.

В этой работе мы фокусируемся на другой не менее важной проблеме, а именно, на изучении сложностного статуса задач, которые индуцируются поиском оптимальных разбиений, позволяющих на следующем этапе анализа понять (раскрыть, объяснить или интерпретировать) структуру экспериментальных данных. Иными словами, нас интересует вычислительная сложность задач, которые необходимо решить, чтобы инструментально (алгоритмически) обеспечить решение прикладных проблем Data mining.

Существует множество оптимизационных задач разбиения конечного множества объектов (точек) на кластеры по разнообразным критериям. Большинство из известных практически важных для Data mining задач разбиения данных на кластеры NP-трудны (см. цитируемые ниже работы). Однако у еще большей части задач статус вычислительной сложности остается не раскрытым. В настоящей работе мы рассматриваем несколько таких задач, которые по смыслу тесно связаны как с классическими задачами проверки статистических гипотез, так и с некоторыми прикладными проблемами (см. далее).

Заметим еще, что рассматриваемые задачи разбиения на кластеры не эквивалентны ни одной из хорошо известных труднорешаемых кластеризационных задач k-*means* (или k-MSSC) [5], [6], k-*median* [7], [8], k-*center* [9], [10], k-*Variance* [11] и др. [12], [13], [14], [15]. Насколько нам известно, рассматриваемые задачи также не эквивалентны ни одной из других труднорешаемых квадратичных евклидовых задач разбиения, выявленных в последние годы. Этот факт, как и сказанное выше, мотивировал наше исследование.

Всюду далее $\|\cdot\|$ – евклидова норма. Рассматриваемые задачи имеют следующие формулировки в форме верификации свойств.

Задача 1 (сбалансированное 2-разбиение по дисперсионному критерию). Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$, что

$$\left|\frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}}\|y-\bar{y}(\mathcal{C})\|^2-\frac{1}{|\mathcal{Y}\setminus\mathcal{C}|}\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-\bar{y}(\mathcal{Y}\setminus\mathcal{C})\|^2\right|\le\varepsilon, \tag{1.1}$$

где $\bar{y}(\mathcal{C})=\frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}}y$ и $\bar{y}(\mathcal{Y}\setminus\mathcal{C})=\frac{1}{|\mathcal{Y}\setminus\mathcal{C}|}\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}y$ – центроиды (геометрические центры) кластеров $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$ соответственно?

Задача 2 (сбалансированное 2-разбиение по критерию суммарного квадратичного разброса). Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$, что

$$\left|\sum_{y\in\mathcal{C}}\|y-\bar{y}(\mathcal{C})\|^2-\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-\bar{y}(\mathcal{Y}\setminus\mathcal{C})\|^2\right|\le\varepsilon? \tag{1.2}$$

Задача 3 (сбалансированное 2-разбиение по критерию мощностно-взвешенного суммарного квадратичного разброса). Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y}\backslash\mathcal{C}$, что

$$\left|\left|\mathcal{C}\right|\sum_{y\in\mathcal{C}}\left\|y-\bar{y}(\mathcal{C})\right\|^2-\left|\mathcal{Y}\backslash\mathcal{C}\right|\sum_{y\in\mathcal{Y}\backslash\mathcal{C}}\left\|y-\bar{y}(\mathcal{Y}\backslash\mathcal{C})\right\|^2\right|\le\varepsilon? \tag{1.3}$$

В статистике хорошо известен критерий Фишера сравнения дисперсий (F-критерий) по выборочным данным из двух распределений [16]. Если рассматривать кластеры $\mathcal{C}$ и $\mathcal{Y}\backslash\mathcal{C}$ как выборки из двух нормальных распределений с неизвестными средними, то этот критерий позволяет сравнить (проверить на равенство) выборочные дисперсии

$$\frac{1}{\left|\mathcal{C}\right|-1}\sum_{y\in\mathcal{C}}\left\|y-\bar{y}(\mathcal{C})\right\|^2 \tag{1.4}$$

и

$$\frac{1}{\left|\mathcal{Y}\backslash\mathcal{C}\right|-1}\sum_{y\in\mathcal{Y}\backslash\mathcal{C}}\left\|y-\bar{y}(\mathcal{Y}\backslash\mathcal{C})\right\|^2 \tag{1.5}$$

этих распределений по их отношению, которое близко к 1 в случае равенства.

Формулы (1.4) и (1.5) в статистике известны как несмещенные оценки дисперсии по выборочным данным. В задаче 1 фигурируют смещенные оценки, которые отличаются от несмещенных лишь знаменателями. Однако это отличие, как известно, несущественно в асимптотическом смысле, т.к. обе оценки являются асимптотически несмещенными.

Легко видеть, что в задаче 1 требуется разбить входное множество $\mathcal{Y}$ на два кластера по критерию сбалансированных выборочных дисперсий. Статистическая интерпретация задачи 1: можно ли разбить неоднородную выборку $\mathcal{Y}$ на две части (подвыборки) $\mathcal{C}$ и $\mathcal{Y}\backslash\mathcal{C}$ с выборочными дисперсиями, которые отличаются не более, чем на некоторое заданное число $\varepsilon > 0$?

Задачи 2 и 3 аналогичны по смыслу задаче 1, но отличаются критериями разбиения. В задаче 2 критерием является суммарный квадратичный разброс относительно среднего (т.е. центроида). В задаче 3 таким критерием является мощностно-взвешенный разброс относительно центроида или сумма квадратов попарных расстояний, так как для любого конечного множества $\mathcal{X}\subset\mathbb{R}^d$ справедливо легко проверяемое равенство

$$\left|\mathcal{X}\right|\sum_{z\in\mathcal{X}}\left\|z-\bar{z}(\mathcal{X})\right\|^2=\frac{1}{2}\sum_{z\in\mathcal{X}}\sum_{x\in\mathcal{X}}\left\|z-x\right\|^2, \tag{1.6}$$

где $\bar{z}(\mathcal{X})$ — центроид множества $\mathcal{X}$.

Статистические трактовки задач 2–3 в виде 2-разбиений неоднородной выборки $\mathcal{Y}$ аналогичны статистической трактовке задачи 1.

Заметим, что сформулированным задачам 2-разбиения легко дать иные интерпретации, в частности, физического и социального плана. Если считать, что точка пространства определяет вектор (силу), направленную из начала координат в эту точку, то, например, задачу 2 можно интерпретировать как поиск сбалансированного (уравновешенного) по суммарному разбросу 2-разбиения разнонаправленных (противодействующих) сил. Если же координату точки рассматривать как численную характеристику мнения (по какому-либо вопросу) человека из некоторой социальной совокупности, то, например, задачу 1 можно трактовать как поиск 2-разбиения этой совокупности на группы, сбалансированные в среднем по разбросу различных мнений. Сходные социально-физические трактовки можно дать задаче 3 с той лишь разницей, что в этой задаче требуется найти 2-разбиение, которое сбалансировано одновременно и по суммарному разбросу элементов в кластерах, и по размеру (мощности) кластеров.

Ниже мы, фактически, показываем, что эти простые по смыслу прикладные проблемы индуцируют труднорешаемые математические задачи. Вопросы алгоритмической аппроксимируемости задач, как и упомянутые выше вопросы адекватности разбиений, находятся вне рамок настоящей работы.

3. АНАЛИЗ СЛОЖНОСТИ

Для выяснения статуса вычислительной сложности сформулированных задач рассмотрим сначала следующую задачу, которая объединяет задачи 1–3.

Задача $\Pi(g(x))$**.** Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и вещественное число $\varepsilon > 0$. *Вопрос*: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y}\backslash\mathcal{C}$, что

$$\left|\frac{1}{2}g(|\mathcal{C}|)\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2-\frac{1}{2}g(|\mathcal{Y}\backslash\mathcal{C}|)\sum_{y\in\mathcal{Y}\backslash\mathcal{C}}\sum_{z\in\mathcal{Y}\backslash\mathcal{C}}\|y-z\|^2\right|\le\varepsilon? \tag{1.7}$$

С учетом (1.6), будем считать, что весовой коэффициент $g(x)$ у сумм в неравенстве (1.7) объединенной задачи равен $1/x^2$ для задачи 1, $1/x$ – для задачи 2 и 1 – для задачи 3, где x – мощность соответствующего кластера.

Докажем NP-полноту задачи $\Pi(g(x))$ для каждой функции $g(x)$ из множества $\left\{\frac{1}{x^2},\frac{1}{x},1\right\}$. Воспользуемся известным NP-полным вариантом классической задачи **Разбиение**, которая имеет следующую формулировку [17].

Задача РРД (разбиение с равными долями). Дано мультимножество из $2K$ целых неотрицательных чисел $a_1, \ldots, a_{2K}$, сумма которых равна $2W$. Вопрос: существует ли разбиение этого мультимножества на два мультиподмножества по K элементов каждое с суммой элементов в каждом из мультиподмножеств, равной W?

Справедлива

Теорема 1. *Для каждой функции* $g(x)\in\left\{\frac{1}{x^2},\frac{1}{x},1\right\}$ *задача* $\Pi(g(x))$ *NP-полна.*

Доказательство. Рассмотрим произвольный пример задачи РРД, т.е. мультимножество $A=\{a_1, \ldots, a_{2K}\}$, сумма элементов которого равна $2W$. Можно считать, что $K>3$, а также, что $W\ge 9$, так как в противном случае задача, очевидно, решается за линейное время с помощью полного перебора или с помощью динамического программирования соответственно.

Положим $\varepsilon=\frac{1}{3K}$ и выберем рациональные числа b_i, $i=1, \ldots, 2K$, удовлетворяющие соотношению

$$\sqrt{a_i}\le b_i\le\sqrt{a_i}+\delta, \tag{1.8}$$

где

$$\delta=\frac{\varepsilon}{K(K-1)Wg(K)}. \tag{1.9}$$

По произвольному входу задачи РРД построим пример входа задачи $\Pi(g(x))$. Положим $N=2K$, $d=4K$, $\mathcal{Y}=\{y_1, \ldots, y_{2K}\}$, где для всех $i=1, \ldots, 2K$ точка y_i содержит рациональное число (параметр) $M>0$ в компоненте i и число b_i в компоненте $2K+i$, а все остальные компоненты этой точки равны 0.

Значения параметра M уточним далее в зависимости от функции $g(x)\in\left\{\frac{1}{x^2},\frac{1}{x},1\right\}$.

Зафиксируем некоторые свойства элементов семейства $\mathcal{Y}$ точек в построенном примере входа задачи $\Pi(g(x))$.

Свойство 1. Пусть подмножество $I\subset\{1,\ldots,2K\}$ содержит номера точек из кластера $\mathcal{C}\subset\mathcal{Y}$. Тогда для точек из $\mathcal{C}$ справедливо равенство:

$$\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2=2|\mathcal{C}|(|\mathcal{C}|-1)M^2+2(|\mathcal{C}|-1)\sum_{i\in I}b_i^2, \tag{1.10}$$

которое следует из того, что в построенном примере для всех $i\ne j$, очевидно, выполнено

$$\|y_i-y_j\|^2=2M^2+b_i^2+b_j^2.$$

Далее, заметим, что согласно (1.8) имеет место оценка

$$a_i \le b_i^2 \le a_i + \delta^2 + 2\delta\sqrt{a_i} \le a_i + \delta(\delta + 2\sqrt{2W}) \le a_i + W\delta, \tag{1.11}$$

поскольку $W > \delta + 2\sqrt{2W}$ при $W \ge 9$ и $\delta < \frac{1}{2}$.

Просуммировав неравенства (1.11), получим

Свойство 2. Для суммы квадратов координатных элементов (1.8) справедлива оценка

$$\sum_{i=1}^{2K} b_i^2 \le 2W(1 + K\delta). \tag{1.12}$$

Свойство 3. Для любого разбиения множества $\{1, \ldots, 2K\}$ индексов на подмножества I_1, I_2 одинаковой мощности K имеют место неравенства

$$\left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| - KW\delta \le \left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \le \left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| + KW\delta, \tag{1.13}$$

так как

$$\sum_{i\in I_1} a_i^2 \le \sum_{i\in I_1} b_i^2 \le \sum_{i\in I_1} a_i^2 + KW\delta$$

и

$$\sum_{i\in I_2} a_i^2 \le \sum_{i\in I_2} b_i^2 \le \sum_{i\in I_2} a_i^2 + KW\delta$$

в соответствии с (1.11).

Допустим теперь, что подмножество I_1 индексов содержит номера точек из $\mathscr{C}$, а подмножество I_2 – номера точек из $\mathscr{Y}\backslash\mathscr{C}$, причем $|I_1| = c$. Тогда в задаче $\Pi(g(x))$ неравенство (1.7) при $x = c$ в соответствии с (1.10) принимает вид:

$$\begin{aligned}&\Bigg|(c(c-1)g(c) - (2K-c)(2K-c-1)g(2K-c))M^2 + \\ &+ \ (c-1)g(c)\sum_{i\in I_1} b_i^2 - (2K-c-1)g(2K-c)\sum_{i\in I_2} b_i^2\Bigg| \le \varepsilon.\end{aligned} \tag{1.14}$$

Если в задаче РРД существует разбиение мультимножества A на два требуемых подмножества, то для соответствующих подмножеств индексов выполняется равенство

$$c = |I_1| = |I_2| = K \tag{1.15}$$

и равенство

$$\sum_{i\in I_1} a_i = \sum_{i\in I_2} a_i. \tag{1.16}$$

Тогда, подставив (1.15) в (1.14) и учитывая (1.13) и (1.16), легко заметить, что для выполнения неравенства (1.7) в задаче $\Pi(g(x))$ достаточно, чтобы

$$(K-1)g(K)\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \le K(K-1)g(K)W\delta \le \varepsilon,$$

что выполняется, если имеет место равенство (1.9).

Таким образом, если в задаче РРД существует требуемое разбиение, то и в задаче $\Pi(g(x))$ требуемое разбиение также существует при выборе параметра δ по формуле (1.9).

Для доказательства обратной импликации укажем значение параметра M для каждой из трех функций $g(x) \in \left\{\frac{1}{x^2}, \frac{1}{x}, 1\right\}$ в объединенной задаче $\Pi(g(x))$. Рассмотрим три случая.

I. При $g(x) = 1$ (что соответствует задаче 3) из (1.9) имеем

$$\delta = \frac{\varepsilon}{K(K-1)W}.$$

При этом неравенство (1.14) в задаче $\Pi(g(x))$ принимает вид

$$\varepsilon \geq \left|(4Kc - 4K^2 + 2K - 2c)M^2 + (c-1)\sum_{i\in I_1} b_i^2 - (2K - c - 1)\sum_{i\in I_2} b_i^2\right|.$$

Оценивая модуль суммы в правой части этого неравенства, с учетом (1.12) получим

$$\begin{gathered}\varepsilon \geq \left|2(K-c)(1-2K)M^2\right| - (c-1)\sum_{i\in I_1} b_i^2 - (2K-c-1)\sum_{i\in I_2} b_i^2 \geq \\ \geq 2\left|K-c\right|(2K-1)M^2 - 4KW(K\delta+1) \geq \left|K-c\right| M^2 - 4KW(K\delta+1),\end{gathered} \tag{1.17}$$

так как $2(2K-1) > 1$.

Выберем параметр M так, что

$$M^2 > \varepsilon + 4KW(K\delta+1).$$

Тогда коэффициент при M^2 в правой части (1.17) должен быть равен 0. Действительно, если $K \neq c$, то $\left|K-c\right| \geq 1$, и из неравенства (1.17) следует

$$\varepsilon \geq M^2 - 4KW(K\delta+1) > \varepsilon,$$

противоречие. Таким образом, $\left|K-c\right| = 0$, а значит, $K = c$.

Далее, из целочисленности элементов мультимножества A следует: если в задаче РРД

$$\sum_{i\in I_1} a_i \neq \sum_{i\in I_2} a_i, \tag{1.18}$$

то

$$\left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| \geq 1. \tag{1.19}$$

Поэтому из неравенства (1.14) с учетом (1.13) имеем

$$\varepsilon \geq (K-1)\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \geq (K-1)(1-KW\delta) = K-1-\varepsilon,$$

что противоречит условию $\varepsilon = \dfrac{1}{3K}$. Следовательно, в задаче РРД выполнено равенство (1.16).

II. Если $g(x) = 1/x$ (что соответствует задаче 2), то из (1.9) имеем

$$\delta = \frac{\varepsilon}{(K-1)W}.$$

При этом неравенство (1.14) имеет вид

$$\varepsilon \geq \left|(c-1-(2K-c-1))M^2 + \frac{c-1}{c}\sum_{i\in I_1} b_i^2 - \frac{2K-c-1}{2K-c}\sum_{i\in I_2} b_i^2\right|.$$

Оценивая модуль суммы в правой части этого неравенства, с учетом (1.12) получаем

$$\varepsilon \geq \left|2(c-K)\right| M^2 - \sum_{i=1}^{2K} b_i^2 \geq 2\left|K-c\right| M^2 - 2W(K\delta+1). \tag{1.20}$$

Выберем параметр M так, что

$$M^2 > \varepsilon + 2W(K\delta+1).$$

Тогда при этом выборе параметра M из (1.20) имеем $K = c$, как и в случае I.

Далее, как и в рассмотренном случае I, если в задаче РРД имеет место неравенство (1.18), то из этого неравенства следует (1.19). Поэтому из (1.14) с учетом (1.13) имеем

$$\varepsilon \geq \frac{K-1}{K}\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \geq \frac{K-1}{K}(1-KW\delta) = \frac{K-1}{K} - \varepsilon,$$

что противоречит условиям $\varepsilon = \frac{1}{3K}$ и $K > 3$. Следовательно, в задаче РРД выполнено равенство (1.16).

III. Наконец, при $g(x) = 1/x^2$ (что соответствует задаче 1) из (1.9) имеем

$$\delta = \frac{K\varepsilon}{(K-1)W}.$$

При этом неравенство (1.14) имеет вид

$$\varepsilon \geq \left|\left(\frac{c-1}{c} - \frac{2K-c-1}{2K-c}\right)M^2 + \frac{c-1}{c^2}\sum_{i\in I_1} b_i^2 - \frac{2K-c-1}{(2K-c)^2}\sum_{i\in I_2} b_i^2\right|.$$

Оценивая модуль суммы в правой части этого неравенства, с учетом (1.12) получаем

$$\varepsilon \geq \left|\frac{2(K-c)}{c(2K-c)}\right| M^2 - \sum_{i=1}^{2K} b_i^2 \geq \frac{|K-c|}{2K^2} M^2 - 2W(K\delta+1). \tag{1.21}$$

Выберем теперь параметр M так, что

$$M^2 > 2K^2(\varepsilon + 2W(K\delta+1)).$$

Тогда из (1.21) имеем $K = c$, как и в случаях I и II.

Далее, как и в рассмотренных случаях I и II, если в задаче РРД имеет место неравенство (1.18), то из этого неравенства следует (1.19). Поэтому из (1.14) с учетом (1.13) имеем

$$\varepsilon \geq \frac{K-1}{K^2}\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \geq \frac{K-1}{K^2}(1-KW\delta) = \frac{K-1}{K^2} - \varepsilon,$$

т.е. $K - 1 \leq 2K^2\varepsilon = 2K/3$, что неверно при $K > 3$. Следовательно, в задаче РРД выполнено равенство (1.16).

Таким образом, в задаче $\Pi(g(x))$ для каждой функции $g(x)$ выполнение неравенства (1.7) влечет в задаче РРД существование разбиения множества A на две равные по мощности доли с одинаковыми суммами элементов.

Из теоремы 1 и тождества (1.6) непосредственно вытекает основной результат настоящей работы:

Следствие. Задачи 1–3 NP-полны.

4. ЗАКЛЮЧЕНИЕ

В настоящей работе мы доказали NP-полноту трех родственных между собой квадратичных задач разбиения конечного множества точек евклидова пространства на сбалансированные по разбросу точек кластеры. Мы показали тесную связь рассмотренных задач с проблемами статистики, анализа данных и их интерпретации.

Очевидно, что рассмотренные задачи 1–3 можно обобщить на случай, когда число кластеров больше, чем 2. В этих обобщениях задач вопросом является – можно ли разбить входное множество так, чтобы соответствующие неравенства (1.1), (1.2) и (1.3) выполнялись для всех пар кластеров. Ясно, что, когда число кластеров является частью входа, то эти обобщения тоже являются NP-полными задачами. Вопрос о статусе сложности параметрического случая этих задач, когда число кластеров не является частью входа, остается открытым. Построение приближенных эффективных алгоритмов с теоретическими гарантиями качества (точности и трудоемкости) представляется важным делом ближайшей перспективы.

СПИСОК ЛИТЕРАТУРЫ

1. *Aggarwal C.C.* Data Mining: The Textbook. Berlin: Springer, 2015. 734 p.
2. *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2001. 533 p.
3. *Han J., Kamber M., Pei J.* Data Mining: Concepts and Techniques (3rd ed.). Waltham: Morgan Kaufmann Publishers is an inprint of Elsevier, 2012. 703 p.
4. *Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T.* Big Data Clustering: A Review // Lecture Notes in Computer Science. 2014. V. 8583. P. 707–720.
5. *Aloise D., Deshpande A., Hansen P., Popat P.* NP-hardness of Euclidean sum-of-squares clustering // Machine Learning. 2009. V. 75. № 2. P. 245–248.
6. *Mahajan M., Nimbhorkar P., Varadarajan K.* The planar k-means problem is NP-hard // Theoretical Computer Science. 2012. V. 442. P. 13–21.
7. *Arora S., Raghavan P., Rao S.* Approximation schemes for Euclidean k-medians and related problems // Proceedings of the 30th Annual ACM Symposium on Theory of Computing. 1998. P. 106–113.
8. *Papadimitriou C.H.* Worst-Case and Probabilistic Analysis of a Geometric Location Problem // SIAM J. Comput. 1981. V. 10. № 3. P. 542–557.
9. *Masuyama S., Ibaraki T., Hasegawa T.* The computational complexity of the m-center problems in the plane // IEEE Trans. IECE Jpn. 1981. V. 64. № 2. P. 57–64.
10. *Hochbaum D.S., Shmoys D.B.* A best possible heuristic for the k-center problem // Mathematics of Operations Research. 1985. V. 10. № 2. P. 180–184.
11. *Aggarwal H., Imai N., Katoh N., Suri S.* Finding k points with minimum diameter and related problems // J. Algorithms. 1991. V. 12. № 1. P. 38–56.
12. *Brucker P.* On the Complexity of Clustering Problems // Lecture Notes in Economics and Mathematical Systems. 1978. V. 157. P. 45–54.
13. *Indyk P.* A sublinear time approximation scheme for clustering in metric space // Proc. of the 40th Ann. IEEE Symp. on Foundations of Computer Science (FOCS). New York: IEEE, 1999. P. 154–159.
14. *Hansen P., Jaumard B., Mladenovich N.* Minimum sum of squares clustering in a low dimensional space // J. Classification. 1998. V. 15. P. 37–55.
15. *Hansen P., Jaumard B.* Cluster analysis and mathematical programming // Mathematical Programming. 1997. V. 79. P. 191–215.
16. *Snedecor G.W., Cochran W.G.* Statistical Methods, Eighth Edition. Iowa State University Press, 1989. 503 p.
17. *Garey M.R., Johnson D.S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman, 1979. 314 p.