ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Май—июнь 2015. Том 22, № 3. С. 5–17

УДК 519.174 DOI: 10.17377/daio.2015.22.465

# РАНДОМИЗИРОВАННЫЙ АЛГОРИТМ ОТЫСКАНИЯ ПОДМНОЖЕСТВА ВЕКТОРОВ С МАКСИМАЛЬНОЙ ЕВКЛИДОВОЙ НОРМОЙ ИХ СУММЫ *)

*Э. Х. Гимади*[1,2], *И. А. Рыков*[1]
[1]Институт математики им. С. Л. Соболева,
пр. Коптюга, 4, 630090 Новосибирск, Россия
[2]Новосибирский гос. университет,
ул. Пирогова, 2, 630090 Новосибирск, Россия
e-mail: gimadi@math.nsc.ru, rykovweb@gmail.com

**Аннотация.** Представлен рандомизированный приближённый алгоритм для NP-трудной в сильном смысле задачи выбора из конечного семейства векторов в евклидовом пространстве заданного числа векторов с максимальной нормой суммы. Приведены условия его полиномиальности и асимптотической точности. Ил. 1, библиогр. 18.

**Ключевые слова:** поиск подмножества векторов, рандомизированный алгоритм, асимптотическая точность.

## Введение

В работе рассматривается следующая задача выбора подмножества векторов.

**Задача** MLSVS (Maximum of the Length of Sum of Vectors from a Subset). Пусть в евклидовом пространстве $\mathbb{R}^k$ задано конечное семейство векторов $V = \{v_1, \ldots, v_n\}$ и натуральное число $m < n$. В множестве $V$ требуется найти подмножество векторов $X$, состоящее из $m$ векторов и обладающее максимальной нормой суммы в рассматриваемом пространстве:

$$F(X) = \Big\| \sum_{v \in X} v \Big\|_2 \to \max_{X \subset V,\, |X|=m}. \qquad (1)$$

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 15–01–00462, 15–01–00976 и 13–07–00070).

Данная задача находится в русле исследований проблем кластерного анализа [11–13, 15]. В частности, она тесно связана с известным классом задач MSSC (Minimum Sum-of-Squares Clustering) [14, 17, 18], возникающих при решении проблем аппроксимации, компьютерной геометрии, статистики и распознавания образов. В их числе находится задача 1-MSSC-F [8–10], в которой требуется разбить конечное множество векторов евклидова пространства на два кластера, причём центр одного кластера неизвестен, а центр другого задан в начале координат, так, что сумма квадратов расстояний от элементов кластера до его геометрического центра минимальна. Оказалось, что оптимальные решения задачи 1-MSSC-F и рассматриваемой в данной статье задачи MLSVS совпадают. Отметим также, что данная задача является частным случаем задачи, возникающей при решении проблемы помехоустойчивого обнаружения повторяющегося фрагмента в сигнале [4].

В [2] доказано, что задача (1) NP-трудна в сильном смысле, и представлен приближённый детерминированный алгоритм с гарантированной относительной погрешностью, не превышающей $\frac{1}{8}(k-1)/L^2$, и временно́й сложностью $O(nk^2(2L+1)^{k-1})$, где $L$ — параметр алгоритма. При этом при фиксированной размерности пространства $k$ алгоритм в [2] даёт вполне полиномиальную аппроксиматизационную схему. В [5] представлен точный алгоритм решения задачи (1), имеющий временну́ю сложность $O(k^2n^{2k})$.

Указанные алгоритмы основаны на переборе конечных детерминированных множеств $W \subset \mathbb{R}^k$ направлений (векторов) в пространстве $\mathbb{R}^k$. Для каждого направления $w \in W$ за время $O(nk)$ строится допустимое решение $X(w)$ задачи (1) как набор из $m$ векторов в $V$, имеющих максимальные проекции на $w$, после чего в качестве приближённого решения принимается тот набор векторов, для которых соответствующая сумма проекций оказалась наибольшей по всем $w \in W$.

В случае приближённого алгоритма из [2] множество $W$ задавалось в виде конечной сетки точек, равномерно распределённых в некотором кубе с центром в начале координат. В точном алгоритме из [5] пространство разбивалось на конечное число областей, внутри которых решения для каждого из направлений совпадают. Множество $W$ составлялось как множество представителей этих областей (по одному направлению из каждой области).

В настоящей работе исследуется приближённый алгоритм $A$, который впервые представлен авторами в [6, 7]. В основу алгоритма положен вероятностный поиск наилучшего из множества $W$ направлений, которое

определяется точками, равномерно независимо выбираемыми на единичной сфере. Показано, что алгоритм является асимптотически точным при условии $|W| = O(\sqrt{k}(8/7\ln n)^k)$. Таким образом, данный подход позволяет получать решения, близкие к оптимальным, со значительно меньшей трудоёмкостью в сравнении с алгоритмами из [2, 5].

Ещё раз отметим, что в [9, 10] исследована задача 1-MSSC-F двухкластерного разбиения множества векторов по критерию минимума суммы квадратов отклонений от центров, полиномиально эквивалентная исследуемой задаче максимизации. Более точно, входные данные для этих задач совпадают, и оптимальное решение одной из них является оптимальным решением и для другой.

В [16] построен 2-приближённый алгоритм трудоёмкости $O(kn^2)$ решения задачи 1-MSSC-F, в [9] предложена полиномиальная аппроксиматизационная схема, а в [10] предложен рандомизированный алгоритм, обладающий линейной трудоёмкостью при константных оценках точности и вероятности несрабатывания и являющийся асимпотически точным при квадратичной трудоёмкости. Заметим, однако, что оценки точности этих алгоритмов не могут быть перенесены с задачи минимизации на задачу максимизации.

Таким образом, вопрос о построении полиномиальных алгоритмов с гарантированными оценками точности в общем случае задачи MLSVS (при нефиксированной размерности пространства $k$) остаётся открытым.

## 1. Приближённый рандомизированный алгоритм

В отличие от работ [2, 5], где используются детерминированные множества $W$, предлагается использование рандомизированного вспомогательного семейства направлений (векторов)

$$W = (W_i), \quad i = 1, \ldots, L,$$

когда каждый такой вектор выбирается случайно и независимо друг от друга.

Алгоритм $A$

Шаг 1. На поверхности единичного $k$-мерного шара в $\mathbb{R}^k$ выбрать независимо равномерно $L$ точек (векторов) $W_1, \ldots, W_L$.

Шаг 2. Для каждого вектора $W_i$ построить допустимое решение $X(W_i)$ как $m$-набор векторов из $V$, дающих максимальные проекции на $W_i$.

Шаг 3. В качестве приближённого решения задачи (1) выбрать такой набор $X(W_i)$, для которого $F(X)$ максимальна по всем $i = 1, \ldots, L$.

Наиболее естественным для рандомизированного алгоритма является проведение анализа в среднем. Одним из подходов к анализу в среднем, предложенным в [3], является нахождение $(\varepsilon, \delta)$-оценок алгоритма.

В соответствии с этим подходом будем говорить, что алгоритм $A$ приближённого решения задачи максимизации *имеет оценки* $(\varepsilon, \delta)$ *на множестве входов* $\mathfrak{I}$, если

$$\mathbb{P}\left\{\frac{f^*(I) - f_A(I)}{f^*(I)} < \varepsilon \mid I \in \mathfrak{I}\right\} \geqslant 1 - \delta,$$

где $f^*(I)$ — оптимальное значение целевой функции для входа $I$, а $f_A(I)$ — значение целевой функции на решении, найденном алгоритмом $A$ для входа $I$ ($\varepsilon$ называется *оценкой относительной погрешности*, а $\delta$ — *оценкой вероятности несрабатывания*).

Вероятность обычно рассчитывается по множеству входных данных, имеющих одинаковый размер (например, в случае задачи MLSVS по множеству всех примеров из $n$ векторов). При этом величины $\varepsilon$ и $\delta$ зависят от этого размера $n$. Алгоритм называется *асимптотически точным*, если значения $\varepsilon$ и $\delta$ стремятся к нулю при $n \to \infty$.

В определённом смысле можно говорить, что оценка относительной погрешности при наличии вероятности несрабатывания является «релаксацией» анализа в худшем, когда требуется, чтобы оценка относительной погрешности выполнялась для абсолютно всех входных данных. Пожертвовав частью входных примеров, можно существенно улучшить оценку относительной погрешности. Наконец, параметр $L$ алгоритма связывает оценки $(\varepsilon, \delta)$ с временно́й сложностью алгоритма $T$. Задача состоит в нахождении такого баланса величин $(\varepsilon, \delta, T)$, чтобы первые две стремились к нулю с ростом $n$, а последняя была как можно меньше.

Как и в [2], будем опираться при анализе на то, что при достаточно большом $L$ отклонение одного из рассмотренных направлений $w$ от направления оптимального вектора-суммы оказывается достаточно мало. Вектор-сумма выбранного для этого направления решения $X(w)$ будет при этом мало отличаться от оптимального вектора-суммы.

Для получения $(\varepsilon, \delta)$-оценок определим событие $\mathcal{B}$ *срабатывания алгоритма* $A$. Пусть $X^*$ — оптимальное решение задачи и $s(X^*)$ — вектор-сумма этого решения, $\varphi_0$ — некоторый параметр. Через $\mathcal{B}$ обозначим следующее событие: найдётся хотя бы один вектор $W_i$, образующий с вектором $s(X^*)$ угол, не превышающий $\varphi_0$.

Определим также $p_0$ как вероятность отклонения случайной (с равномерным распределением) точки на сфере от направления оптимального вектора-суммы на угол, меньший $\varphi_0$. Очевидно, что в силу независимости выбранных направлений

$$\delta = \Pr(\bar{\mathcal{B}}) = (1-p_0)^L \leqslant e^{-p_0 L}. \tag{2}$$

Для оценки величины $p_0$ как вероятности попадания в окрестность-«шапочку» оптимального направления используются формулы объёма $k$-мерного шара и площади поверхности $k$-мерной сферы, выражающиеся через гамма-функцию $\Gamma(x) = \int\limits_0^\infty t^{x-1}e^{-t}dt$.

Докажем вспомогательные леммы.

**Лемма 1.** *Для всякого натурального $k > 1$ верно неравенство*

$$\frac{\Gamma\left(\frac{k}{2}\right)}{2\sqrt{\pi}\,\Gamma\left(\frac{k+1}{2}\right)} \geqslant \frac{1}{\pi\sqrt{k}}. \tag{3}$$

Доказательство. Случай $k$ чётно: $k = 2j$. Тогда

$$\begin{aligned}\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} &= \frac{\Gamma(j)}{\Gamma\left(j+\frac{1}{2}\right)} = \frac{(j-1)!\,2^j}{\sqrt{\pi}(2j-1)!!} \\ &= \frac{1\cdot 2\cdots(j-1)\cdot 2^j}{\sqrt{\pi}\cdot 1\cdot 3\cdots(2j-1)} = \frac{2}{\sqrt{\pi}}\cdot\frac{2\cdot 4\cdots(2j-2)}{3\cdot 5\cdots(2j-1)} \\ &= \frac{2}{\sqrt{\pi}}\sqrt{\frac{2\cdot 2}{3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdot\frac{6\cdot 6}{5\cdot 7}\cdots\frac{(2j-2)(2j-2)}{(2j-3)(2j-1)}\cdot\frac{1}{(2j-1)}}.\end{aligned}$$

С учётом неравенства $s^2 > (s-1)(s+1)$ для всякого натурального $s > 1$ получим

$$\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} \geqslant \frac{2}{\sqrt{\pi(2j-1)}} = \frac{2}{\sqrt{\pi(k-1)}} \geqslant \frac{2}{\sqrt{\pi k}},$$

откуда следует справедливость (3) для чётного $k$.

Случай $k$ нечётно: $k = 2j+1$. Тогда

$$\begin{aligned}&\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} = \frac{\Gamma\left(j+\frac{1}{2}\right)}{\Gamma(j+1)} = \frac{\sqrt{\pi}(2j-1)!!}{j!\,2^j} = \frac{\sqrt{\pi}\cdot 1\cdot 3\cdots(2j-1)}{1\cdot 2\cdots j\cdot 2^j} \\ &= \sqrt{\pi}\cdot\frac{3\cdot 5\cdots(2j-1)}{2\cdot 4\cdots 2j} = \sqrt{\pi}\sqrt{\frac{1}{2}\cdot\frac{3\cdot 3}{2\cdot 4}\cdot\frac{5\cdot 5}{4\cdot 6}\cdots\frac{(2j-1)(2j-1)}{(2j-2)2j}\cdot\frac{1}{2j}}.\end{aligned}$$

Аналогично предыдущему случаю имеем

$$\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} \geqslant \sqrt{\frac{\pi}{4j}} \geqslant \sqrt{\frac{\pi}{2(2j+1)}} = \sqrt{\frac{\pi}{2k}},$$

так что неравенство (3) справедливо и при нечётном $k$. Лемма 1 доказана.

Пусть $r = 2\sin\frac{\varphi_0}{2}$. Нетрудно видеть (рис. 1), что шапочка на сфере, состоящая из точек, отклоняющихся от заданной точки на угол не более $\varphi_0$, может быть определена как пересечение этой сферы с шаром, имеющим центр в заданной точке и радиус, равный $r$.

**Лемма 2.** *Для вероятности $p_0$ справедливо неравенство*

$$p_0 \geqslant \frac{\left(\frac{7}{8}r\right)^{k-1}}{\pi\sqrt{k}}.$$

Доказательство. Напомним формулы [24] объёма $k$-мерного шара $\mathbb{B}_k(R) \subset \mathbb{R}^k$ радиуса $R$:

$$V^{(k)}(R) = \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)}R^k$$

и площади поверхности $k$-мерной сферы $\mathbb{S}_k(R) \subset \mathbb{R}^{k+1}$ радиуса $R$:

$$S^{(k)}(R) = \frac{2\pi^{(k+1)/2}}{\Gamma\left(\frac{k+1}{2}\right)}R^k.$$

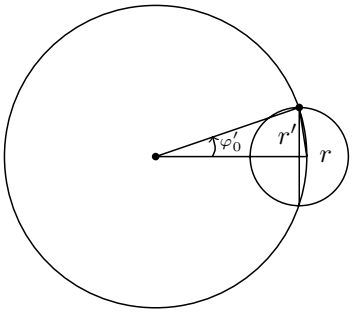


Рис. 1

Из рис. 1 нетрудно видеть, что объём шапочки на сфере в $k$-мерном пространстве, отсекаемой шаром радиуса $r$, не меньше объёма

$(k-1)$-мерного шара радиуса $\sqrt{r^2 - r^4/4}$ (на рисунке для $k = 2$ он представляет собой отрезок-хорду; в $k$-мерном пространстве этот шар образуется как часть гиперплоскости, отделяемая границей шапочки).

Таким образом, вероятность $p_0$ не меньше отношения объёма $(k-1)$-мерного шара $\mathbb{B}_{k-1}(r')$ радиуса $r' = \sqrt{r^2 - r^4/4}$ к площади единичной сферы $\mathbb{S}_{k-1}(1)$. Учитывая, что при $r < \frac{1}{2}$ верно $r' > \frac{7}{8}r$, получаем

$$p_0 > \frac{V^{(k-1)}(r')}{S^{(k-1)}(1)} = \frac{\pi^{(k-1)/2}\left(\frac{7}{8}r\right)^{k-1}\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)2\pi^{k/2}} = \frac{\left(\frac{7}{8}r\right)^{k-1}\Gamma\left(\frac{k}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{k+1}{2}\right)}.$$

Используя оценку из леммы 1, получаем требуемое неравенство. Лемма 2 доказана.

**Теорема 1.** *Алгоритм A приближённого решения задачи* (1) *имеет оценки относительной погрешности*

$$\varepsilon_A \leqslant \varphi_0^2/2,$$

*вероятности несрабатывания*

$$\delta_A \leqslant \exp\left(-\frac{\left(7/4\sin\frac{\varphi_0}{2}\right)^{k-1}}{\pi\sqrt{k}}L\right)$$

*и временно́й сложности* $T_A = O(nkL)$.

Доказательство. Оценим относительную погрешность алгоритма A.

Пусть $X^A$ — решение, найденное алгоритмом, $X^*$ — оптимальное решение, $c \in W$ — направление, ближайшее среди рассмотренных направлений к вектору-сумме оптимального решения, $X^c = X(c) = \left\{X_1^c, \ldots, X_m^c\right\}$ — соответствующее допустимое решение. Обозначим через $s(X)$ сумму векторов из $X$, т. е. $F(X) = \|s(X)\|$. Верна следующая цепочка неравенств:

$\|s(X^A)\| \geqslant \|s(X^c)\|$, поскольку алгоритм выбирает наилучшее из направлений из $W$;

$\|s(X^c)\| \geqslant \sum\limits_{i=1}^{m} X_i^c * c$, поскольку длина вектора не меньше его проекции на любое направление и проекция вектора-суммы равна сумме проекций;

$\sum\limits_{i=1}^{m} X_i^c * c \geqslant \sum\limits_{i=1}^{m} X_i^* * c$, поскольку $m$ векторов из $X(c)$ дают наибольшие проекции $c$ среди всех входных векторов;

$\sum\limits_{i=1}^{m} X_i^* * c = \|s(X^*)\| * \cos\varphi \geqslant \|s(X^*)\| * \cos\varphi_0$, поскольку угол $\varphi$ между $s(X^*)$ и $c$ не превосходит $\varphi_0$ при срабатывании алгоритма.

Поэтому

$$\varepsilon_A = 1 - \frac{\|s(X^A)\|}{\|s(X^*)\|} \leqslant 1 - \cos\varphi_0 = 2\left(\sin\frac{\varphi_0}{2}\right)^2 \leqslant \frac{\varphi_0^2}{2}.$$

Вероятность несрабатывания алгоритма оценивается величиной $(1-p_0)^L$, отсюда в силу леммы 2 и неравенства 2 следует, что

$$\delta_A \leqslant \exp\left(-\frac{\left(7/4\sin\frac{\varphi_0}{2}\right)^{k-1}}{\pi\sqrt{k}}L\right).$$

Оценка временно́й сложности $T_A = O(nkL)$ следует из описания алгоритма $A$. Действительно, для каждого вспомогательного направления $W_i,\ i = 1, 2, \ldots, L$, за время $O(nk)$ вычисляются проекции (скалярные произведения) $n$ входных векторов $v_1, v_2, \ldots, v_n$ на вектор $W_i$, после чего из вычисленных $n$ скаляров осуществляется выбор фиксированного числа (равного $m$) наибольших значений, что можно сделать за время $O(n)$ [1]. Теорема 1 доказана.

В следующей теореме представлен пример условий асимптотической точности алгоритма $A$, следующий непосредственно из теоремы 1.

**Теорема 2.** *Алгоритм $A$ с параметрами*

$$\varphi_0 = \frac{1}{\ln n} \text{ и } L = \left\lceil \left(\frac{8/7}{1-1/(12\ln^2 n)}\right)^{k-1} \pi\sqrt{k}\ln^k n \right\rceil \tag{4}$$

*за время*

$$T_A = O\left(k^{3/2} n \ln n \left(\frac{8/7\ln n}{1-1/(12\ln^2 n)}\right)^{k-1}\right)$$

*находит асимптотически точное решение задачи* (1) *с оценками относительной погрешности $\varepsilon_A \leqslant \frac{1}{2\ln^2 n}$ и вероятности несрабатывания $\delta_A \leqslant \frac{1}{n}$.*

Доказательство. Действительно, обозначив $x = \frac{1}{2\ln n}$, с учётом неравенства $\frac{\sin x}{x} \geqslant 1 - x^2/3$ имеем

$$\begin{aligned}\delta_A &\leqslant \exp\left(-\frac{\left(7/4\sin\frac{\varphi_0}{2}\right)^{k-1}}{\pi\sqrt{k}}L\right) \\ &= \exp\left(-\frac{(7/4\sin x)^{k-1}}{\pi\sqrt{k}}\left(\frac{8/7}{1-x^2/3}\right)^{k-1}\pi\sqrt{k}\ln^k n\right) \\ &= \exp\left(-\left(\frac{\frac{\sin x}{x}}{1-x^2/3}\right)^{k-1}\ln n\right) \leqslant \exp(-\ln n) = \frac{1}{n}.\end{aligned}$$

**Следствие.** *Алгоритм $A$ с параметрами* (4) *при фиксированной размерности $k$ пространства $\mathbb{R}^k$ асимптотически точен и полиномиален.*

### Заключение

Для задачи выбора подмножества векторов заданной мощности с максимальной длиной вектора-суммы построен рандомизированный алгоритм и найдены его оценки временно́й сложности, относительной погрешности и вероятности несрабатывания. Указаны значения параметров алгоритма, при котором алгоритм имеет существенно меньшую трудоёмкость в сравнении с известными точным и приближённым алгоритмами и при этом является асимптотически точным.

Интерес представляет построение полиномиального алгоритма с гарантированной оценкой точности для указанной задачи.

## ЛИТЕРАТУРА


1. **Ахо А., Хопкрофт Дж., Ульман Дж.** Построение и анализ вычислительных алгоритмов. М.: Мир, 1979. 536 с.
2. **Бабурин А. Е., Гимади Э. Х., Глебов Н. И., Пяткин А. В.** Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. операций. Сер. 2. 2007. Т. 14, № 1. С. 32–42.
3. **Гимади Э. Х., Глебов Н. И., Перепелица В. А.** Алгоритмы с оценками для задач дискретной оптимизации // Пробл. кибернетики. 1975. Вып. 31. С. 35–42.
4. **Гимади Э. Х., Кельманов А. В., Кельманова М. А., Хамидуллин С. А.** Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. журн. индустр. математики. 2006. Т. 9, № 1. С. 55–74.
5. **Гимади Э. Х., Пяткин А. В., Рыков И. А.** О полиномиальной разрешимости некоторых задач выбора подмножества векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. операций. 2008. Т. 15, № 6. С. 11–19.
6. **Гимади Э. Х., Рыков И. А.** Приближённый рандомизированный алгоритм отыскания подмножества векторов с максимальной нормой суммы в многомерном евклидовом пространстве //Дискретная оптимизация и исследование операций: Мат. конф. (Алтай, 27 июня–3 июля 2010 г.). Новосибирск: Изд-во Ин-та математики, 2010. С. 102.
7. **Гимади Э. Х., Рыков И. А.** Рандомизированный алгоритм отыскания подмножества векторов с максимальной нормой суммы в евклидовом пространстве // Тр. XV Байк. междунар. школы-семинара «Методы оптимизации и их приложения». Т. 4. Дискретная оптимизация. Иркутск: РИО ИДСТУ СО РАН, 2011. С. 76–81.
8. **Долгушев А. В., Кельманов А. В.** Приближённый алгоритм решения одной задачи кластерного анализа // Дискрет. анализ и исслед. операций. 2011. Т. 18, № 2. С. 29–40.

9. **Долгушев А. В., Кельманов А. В., Шенмайер В. В.** Приближённая полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: Сб. докл. 9-й междунар. конф. (Республика Черногория, г. Будва, 16–22 сентября 2012 г.). М.: Торус Пресс, 2012. С. 242–244.
10. **Кельманов А. В., Хандеев В. И.** Рандомизированный алгоритм для одной задачи двухкластерного разбиения множества векторов // Журн. вычисл. математики и мат. физики. 2015. Т. 55, № 2. С. 335–344.
11. **Bern M., Eppstein D.** Approximation algorithms for geometric problems // Approximation algorithms for NP-hard problems. Boston: PWS Publ. Co., 1997. P. 296–345.
12. **Bishop C. M.** Pattern recognition and machine learning. New York: Springer, 2006. 738 p.
13. **James G., Witten D., Hastie T., Tibshirani R.** An introduction to statistical learning. New York: Springer, 2013. 426 p.
14. **Jain A. K.** Data clustering: 50 years beyond $K$-means // Pattern. Recognit. Lett. 2010. Vol. 31. P. 651–666.
15. **Flach P.** Machine learning: The art and science of algorithms that make sense of data. New York: Cambridge Univ. Press, 2012. 396 p.
16. **Huber G.** Notes: gamma function derivation of $n$-sphere volumes // Amer. Math. Monthly. 1982. Vol. 89, No. 5. P. 301–302.
17. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Stat. Probab. Vol. 1. Berkeley, CA: Univ. California Press, 1967. P. 281–297.
18. **Rao M. R.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. 1971. Vol. 66. P. 622–626.

*Гимади Эдуард Хайрутдинович*
*Рыков Иван Александрович*

Статья поступила
21 октября 2014 г.
Исправленный вариант —
2 марта 2015 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
May–June 2015. Volume 22, No. 3. P. 5–17

UDC 519.174 DOI: 10.17377/daio.2015.22.465

# A RANDOMIZED ALGORITHM FOR THE VECTOR SUBSET PROBLEM WITH THE MAXIMAL EUCLIDEAN NORM OF ITS SUM

*E. Kh. Gimadi*[1,2], *I. A. Rykov*[1]
[1]Sobolev Institute of Mathematics,
4 Koptyug Ave., 630090 Novosibirsk, Russia
[2]Novosibirsk State University,
2 Pirogov St., 630090 Novosibirsk, Russia
e-mail: gimadi@math.nsc.ru, rykovweb@gmail.com

**Abstract.** We present a randomized approximation algorithm for the problem of finding a subset of a finite set of vectors in the Euclidean space with the maximal norm of the sum vector. We show that with an appropriate choice of parameters, the algorithm is polynomial for the problem with any fixed dimension and asymptotically optimal. Il. 1, bibliogr. 18.

**Keywords:** search for vector subset, randomized algorithm, asymptotical exactness.

## REFERENCES

1. **A. V. Aho, J. E. Hopcroft,** and **J. D. Ullman,** *The Design and Analysis of Computer Algorithms,* Addison-Wesley, Boston, 1974. Translated under the title *Postroenie i analiz vychislitel'nykh algoritmov,* Mir, Moscow, 1979.
2. **A. E. Baburin, E. Kh. Gimadi, N. I. Glebov,** and **A. V. Pyatkin,** The problem of finding a subset of vectors with the maximum total weight, *Diskretn. Anal. Issled. Oper., Ser. 2,* **14**, No. 1, 32–42, 2007. Translated in *J. Appl. Ind. Math.,* **2**, No. 1, 32–38, 2008.
3. **E. Kh. Gimadi, N. I. Glebov,** and **V. A. Perepelitsa,** Algorithms with estimates for discrete optimization problems, in S. V. Yablonskii, ed., *Problemy kibernetiki* (Problems of Cybernetics), Vol. 31, pp. 35–42, Nauka, Moscow, 1976.
4. **E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova,** and **S. A. Khamidullin,** A posteriori detection of a quasiperiodic fragment with a given number of repetitions in a numerical sequence, *Sib. Zh. Ind. Mat.,* **9**, No. 1, 55–74, 2006.

5. **E. Kh. Gimadi, A. V. Pyatkin,** and **I. A. Rykov,** On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension, *Diskretn. Anal. Issled. Oper.,* **15**, No. 6, 11–19, 2008. Translated in *J. Appl. Ind. Math.,* **4**, No. 1, 48–53, 2010.
6. **E. Kh. Gimadi** and **I. A. Rykov,** Approximation randomized algorithm for finding a vector subset with the maximal norm of sum in a Euclidean space, in *Materialy Rossiiskoi konferentsii "Diskretnaya optimizatsiya i issledovanie operatsii"* (Proc. Russian Conf. "Discrete Optimization and Operation Research"), *Altay, Russia, June 27 – July 3, 2010,* p. 102, Izdatel'stvo Inst. Mat., Novosibirsk, 2010.
7. **E. Kh. Gimadi** and **I. A. Rykov,** Randomized algorithm for finding a vector subset with the maximal norm of sum in a Euclidean space, in *Trudy XV Baikal'skoi mezhdunarodnoi shkoly-seminara "Metody optimizatsii i ikh prilozheniya"* (Proc. XV Baikal Int. School-Seminar "Optimization Methods and Their Applications"), *Listvyanka, Irkutsk Reg., Russia, June 23–29, 2011,* Vol. 4, pp. 76–81, RIO IDSTU SO RAN, Irkutsk, 2011.
8. **A. V. Dolgushev** and **A. V. Kel'manov,** An approximation algorithm for solving a problem of cluster analysis, *Diskretn. Anal. Issled. Oper.,* **18**, No. 2, 29–40, 2011. Translated in *J. Appl. Ind. Math.,* **5**, No. 4, 551–558, 2011.
9. **A. V. Dolgushev, A. V. Kel'manov,** and **V. V. Shenmaier,** A polynomial approximation scheme for a problem of cluster analysis, in *Doklady 9 mezhdunarodnoi konferentsii "Intellektualizatsiya obrabotki informatsii"* (Proc. 9th Int. Conf. "Intellectualization of Information Processing"), *Budva, Montenegro, Sept. 16–22, 2012,* pp. 242–244, Torus Press, Moscow, 2012.
10. **A. V. Kel'manov** and **V. I. Khandeev,** A randomized algorithm for two-cluster partition of a set of vectors, *Zh. Vychisl. Mat. Mat. Fiz.,* **55**, No. 2, 335–344, 2015. Translated in *Comput. Math. Math. Phys.,* **55**, No. 2, 330–339, 2015.
11. **M. Bern** and **D. Eppstein,** Approximation algorithms for geometric problems, in D. S. Hochbaum, ed., *Approximation Algorithms for NP-hard Problems,* pp. 296–345, PWS Publ. Co., Boston, 1997.
12. **C. M. Bishop,** *Pattern Recognition and Machine Learning,* Springer, New York, 2006.
13. **G. James, D. Witten, T. Hastie,** and **D. Tibshirani,** *An Introduction to Statistical Learning. With Applications in R,* Springer, New York, 2013.
14. **A. K. Jain,** Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett.,* **31**, No. 8, 651–666, 2010.
15. **P. Flach,** *Machine Learning. The Art and Science of Algorithms that Make Sense of Data,* Cambridge Univ. Press, Cambridge, 2012.
16. **G. Huber,** Gamma function derivation of $n$-sphere volumes, *Am. Math. Mon.,* **89**, No. 5, 301–302, 1982.
17. **J. B. MacQueen,** Some methods for classification and analysis of multivariate observations, in L. M. Le Cam and J. Neyman, eds., *Proc. 5th Berkeley*

*Symp. Math. Stat. Probab., Berkeley, USA, June 21 – July 18, 1965 and Dec. 27, 1965 – Jan. 7, 1966,* Vol. 1, pp. 281–297, Univ. of California Press, Berkeley, 1967.

**18. M. R. Rao,** Cluster analysis and mathematical programming, *J. Am. Stat. Assoc.,* **66**, 622–626, 1971.

*Edward Kh. Gimadi*
*Ivan A. Rykov*

Received
21 October 2014
Revised
2 March 2015