

УДК 519.176

АППРОКСИМАЦИОННАЯ СХЕМА ДЛЯ ОДНОЙ ЗАДАЧИ ПОИСКА ПОДМНОЖЕСТВА ВЕКТОРОВ *)

В. В. Шенмайер

Аннотация. Рассматривается следующая задача кластеризации: среди заданного множества векторов найти подмножество мощности k , обладающее минимальным квадратичным отклонением от своего среднего. Расстояния между векторами определяются евклидовой метрикой. Предлагается аппроксимационная схема (PTAS), позволяющая решать данную задачу с произвольной относительной погрешностью ε за время $O(n^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon}d)$, где n — число векторов в исходном множестве, d — размерность пространства.

Ключевые слова: выбор подмножества векторов, кластерный анализ, аппроксимационная схема, приближённый алгоритм.

Введение

Рассматривается следующая задача. Пусть X — конечное множество векторов в евклидовом пространстве и k — некоторое число, $1 \leq k \leq n$, где $n = |X|$. Требуется найти подмножество (кластер) $K \subseteq X$ мощности k такое, что сумма квадратов расстояний от векторов из K до вектора $\bar{c}(K) = \sum_{x \in K} x/k$ (центра кластера) минимальна:

$$\min_K \sum_{x \in K} \|x - \bar{c}(K)\|^2, \quad K \subseteq X, |K| = k.$$

Существует ряд содержательных трактовок задачи. Одна из них лежит в области распознавания образов. Имеется множество результатов измерений характеристик некоторых объектов. Результаты измерений представляют из себя многомерные вещественные векторы. Измерения

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 09-01-00032 и 10-07-00195), целевой программы № 2 Президиума РАН (проект № 227), а также целевой программы СО РАН (проект № 44).

имеют ошибку, соответствие между объектами и результатами измерений неизвестно, но известно, что k измерений соответствуют одному и тому же объекту. Требуется, используя критерий минимума суммы квадратов расстояний, найти подмножество результатов измерений, вероятнее всего соответствующих данному неизвестному объекту.

Другая интерпретация задачи лежит в области задач размещения. Имеется множество потребителей и одно предприятие мощности k (способное обслуживать k потребителей). Каждому потребителю соответствует точка в пространстве, в которой он находится. Место расположения предприятия не задано, оно может быть размещено в любой точке. Стоимость обслуживания потребителей определяется как сумма квадратов расстояний до них. Требуется определить место размещения предприятия и множество обслуживаемых потребителей таким образом, что стоимость их обслуживания минимальна.

В [1] установлено, что задача NP -трудна в сильном смысле. Отсюда следует, что при условии $P \neq NP$ не существует ни полиномиального алгоритма для её решения, ни псевдополиномиального алгоритма, ни полностью полиномиальной аппроксимационной схемы (FPTAS). В [2] предложен полиномиальный приближённый алгоритм, имеющий оценку точности 2. В [3] предложен псевдополиномиальный алгоритм для частного случая, когда размерность пространства фиксирована (трудоемкость данного алгоритма экспоненциально зависит от размерности пространства).

Предлагается приближённый алгоритм, имеющий относительную погрешность $1/t + 8\zeta(t, s)$, где $\zeta(t, s) = \sqrt{t-1}/s + (t-1)/s^2$, и трудоемкость $O(n^{t+1}s^{t-1}d)$, где $t, s \geq 1$ — произвольные целочисленные параметры, d — размерность пространства. В частности, при выборе $t = 2/\varepsilon$, где $\varepsilon > 0$, и $s = 9t^{3/2}$ имеем аппроксимационную схему (PTAS), позволяющую решать задачу с относительной погрешностью ε за время $O(n^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon}d)$.

1. Геометрические основы алгоритма

Заметим, что оптимальный кластер обладает свойством локальной оптимальности: он представляет из себя k векторов из X , ближайших к некоторой точке пространства (локальному центру). Идея алгоритма состоит в поиске этого локального центра среди векторов линейных оболочек всех наборов векторов множества X мощности t . При этом для дискретизации алгоритма на каждой линейной оболочке рассматривается $(t-1)$ -мерная сетка с шагом h , где h — константа, вычисляемая на предварительном шаге алгоритма.

Пусть $f(y, K) = \sum_{x \in K} \|x - y\|^2$ — значение целевой функции в точке y на кластере K . Положим $f(K) = f(\bar{c}(K), K)$. Легко показать, что при фиксированном кластере K минимальное значение $f(y, K)$ достигается в точке $\bar{c}(K)$, при этом

$$f(y, K) = f(K) + k \|y - \bar{c}(K)\|^2. \quad (1)$$

Для доказательства равенства (1) достаточно рассмотреть случай, когда центр кластера совпадает с началом координат: $\bar{c}(K) = (0, \dots, 0)$, а $y = (1, 0, \dots, 0)$. В этом случае $\|x - y\|^2 = \|x\|^2 - 2x(1) + 1$, где $x(1)$ — первая координата вектора x . Отсюда

$$f(y, K) = \sum_{x \in K} \|x\|^2 - 2 \sum_{x \in K} x(1) + k.$$

Но $\sum_{x \in K} \|x\|^2 = f(K)$, $\sum_{x \in K} x(1) = 0$ и $k = k \|y - \bar{c}(K)\|^2$ согласно выбору векторов $\bar{c}(K)$ и y . Таким образом, равенство (1) доказано.

Обозначим через $\varepsilon(y, K)$ относительную погрешность решения y локальной задачи поиска оптимального центра кластера K :

$$\varepsilon(y, K) = \frac{f(y, K) - f(K)}{f(K)}.$$

Заметим, что если K^* — оптимальный кластер, а кластер K^y состоит из k векторов множества X , ближайших к точке y , то величина $\varepsilon(y, K^*)$ является верхней оценкой относительной погрешности решения K^y исходной глобальной задачи поиска подмножества векторов. Действительно, $f(K^y) \leq f(y, K^y) \leq f(y, K^*)$, следовательно,

$$\frac{f(K^y) - f(K^*)}{f(K^*)} \leq \frac{f(y, K^*) - f(K^*)}{f(K^*)} = \varepsilon(y, K^*).$$

Это наблюдение и следующая теорема дают геометрическое обоснование алгоритма.

Теорема 1. Пусть K — произвольное множество векторов мощности k и $1 \leq t \leq k$. Тогда линейная оболочка одного из подмножеств K мощности t содержит вектор y_t такой, что $\varepsilon(y_t, K) \leq 1/t$.

Доказательство теоремы основывается на следующем факте.

Лемма 1. Пусть x — произвольный вектор евклидова пространства и $y = y(x, K)$ — ближайший к $\bar{c}(K)$ вектор, лежащий на лучах,

проведённых из x во все векторы кластера K . Тогда

$$\varepsilon(y, K) \leq \frac{\varepsilon(x, K)}{1 + \varepsilon(x, K)}.$$

ДОКАЗАТЕЛЬСТВО. Не нарушая общности, будем считать, что вектор $\bar{c}(K)$ совпадает с началом координат $O = (0, \dots, 0)$. Можно также считать, что координатные оси повернуты таким образом, что вектор x лежит на первой из них: $x = (x(1), 0, \dots, 0)$, для определённости $x(1) \geq 0$. Аналогично можно считать, что вектор y находится в плоскости, образованной первой и второй координатными осями.

Поскольку вектор O — центр K , среди векторов кластера найдутся лежащие в полупространстве $X(1) \leq 0$. Следовательно, луч xy пересекает гиперплоскость $X(1) = 0$ в некоторой точке $A = (0, a, 0, \dots, 0)$. При этом если одно из чисел $x(1)$ или a равно нулю, то вектор y совпадает с оптимальным решением, и, следовательно, утверждение леммы очевидно. Поэтому, не нарушая общности, будем предполагать, что $a > 0$ и $x(1) > 0$.

Заметим, что согласно выбору луча xy все векторы кластера K находятся вне конуса C , образованного вектором x и шаром, находящимся в гиперплоскости $X(1) = 0$, радиуса a и с центром в начале координат (рис. 1).

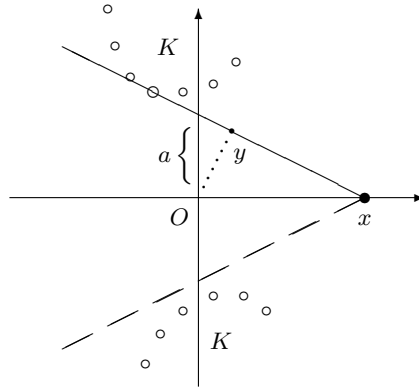


Рис. 1

Согласно равенству (1) имеем

$$\varepsilon(x, K) = \frac{\|x\|^2 k}{f(O, K)} = \frac{x(1)^2 k}{f(O, K)}. \quad (2)$$

Из геометрических соображений

$$\varepsilon(y, K) = \frac{\|y\|^2 k}{f(O, K)} = \frac{a^2 x(1)^2}{a^2 + x(1)^2} \frac{k}{f(O, K)}. \quad (3)$$

Оценим величину $f(O, K)$. Применим для этого технику упрощений, заключающуюся в переходе к более простой геометрической ситуации с сохранением оцениваемой величины. Заметим, что $f(O, K) = f(O', K')$, где O' — двумерный нулевой вектор, а множество K' состоит из двумерных векторов вида $u' = (u(1), \sqrt{\|u\|^2 - u(1)^2})$, где $u \in K$. Действительно, каждый вектор u' совпадает по первой координате с u , а вторая координата u' равна расстоянию от u до первой оси координат. Таким образом, $\|u'\| = \|u\|$.

Далее в силу того, что среднее квадратов не меньше квадрата среднего, имеем $f(O', K') \geq k \|\bar{u}'\|^2$, где $\bar{u}' = \sum_{u' \in K'} u' / k$. Но поскольку векторы исходного кластера K лежат вне конуса C , векторы множества K' лежат над прямой, соединяющей двумерные векторы $x' = (x(1), 0)$ и $A' = (0, a)$. Следовательно, средний вектор \bar{u}' также лежит над данной прямой. При этом поскольку $\bar{u}'(1) = 0$, имеем $\|\bar{u}'\| \geq a$. Отсюда

$$f(O, K) \geq k a^2. \quad (4)$$

Подставим оценку (4) в равенство (3):

$$\varepsilon(y, K) \leq \frac{x(1)^2}{a^2 + x(1)^2}. \quad (5)$$

С другой стороны, объединяя равенства (2) и (3), получим

$$\varepsilon(y, K) = \frac{a^2 \varepsilon(x, K)}{a^2 + x^2(1)}. \quad (6)$$

Рассмотрим выражения (5) и (6) как функции от аргумента a . Первая из них монотонно убывает от 1 до 0, вторая монотонно возрастает от 0 до $\varepsilon(x, K)$. Следовательно, минимум из этих двух функций достигает максимума в точке их пересечения, определяемой соотношением $x^2(1) = a^2 \varepsilon(x, K)$. Таким образом,

$$\varepsilon(y, K) \leq \frac{a^2 \varepsilon(x, K)}{a^2 + a^2 \varepsilon(x, K)} = \frac{\varepsilon(x, K)}{1 + \varepsilon(x, K)}.$$

Лемма 1 доказана.

ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 1 проводится индукцией по t .

БАЗА ИНДУКЦИИ: $t = 1$. Пусть вектор y_1 — ближайший к центру из всех векторов кластера K . Тогда $f(\bar{c}(K), K) \geq k \|y_1 - \bar{c}(K)\|^2$. Следовательно,

$$f(y_1, K) = f(\bar{c}(K), K) + k \|y_1 - \bar{c}(K)\|^2 \leq 2f(\bar{c}(K), K).$$

Таким образом, $\varepsilon(y_1, K) \leq 1$.

ИНДУКТИВНЫЙ ПЕРЕХОД. Рассмотрим в качестве y_{t+1} вектор $y(y_t, K)$ из леммы 1. Заметим, что данный вектор y_{t+1} лежит в линейной оболочке $(t+1)$ -го вектора кластера K . При этом согласно лемме 1 имеем

$$\varepsilon(y_{t+1}, K) \leq \frac{\varepsilon(y_t, K)}{1 + \varepsilon(y_t, K)}.$$

По индукции $\varepsilon(y_t, K) \leq 1/t$, следовательно,

$$\varepsilon(y_{t+1}, K) = \frac{1}{1/\varepsilon(y_t, K) + 1} \leq \frac{1}{t+1}.$$

Теорема 1 доказана.

Замечание 1. При $t \geq k$ утверждение теоремы 1 также справедливо, поскольку вектор $\bar{c}(K)$ принадлежит линейной оболочке k векторов из K и при этом $\varepsilon(\bar{c}(K), K) = 0$.

Замечание 2. Доказанная оценка относительной погрешности является достижимой при любых t . Для того чтобы убедиться в этом, достаточно рассмотреть в качестве множества K набор из k единичных ортов пространства \mathbb{R}^k и устремить k в бесконечность.

Действительно, расстояние от начала координат $O = (0, \dots, 0)$ до центра кластера равно $\sqrt{1/k}$. Отсюда и из (1) получаем $f(K) = f(O, K) - k/k = k - 1$. С другой стороны, линейная оболочка любых t единичных ортов находится на расстоянии $\sqrt{1/t}$ от начала координат. Следовательно, согласно неравенству треугольника расстояние от точки y_t до центра кластера не меньше величины $\sqrt{1/t} - \sqrt{1/k}$. Отсюда в силу равенства (1) имеем

$$f(y_t, K) - f(K) \geq k(1/t - 2/\sqrt{tk} + 1/k).$$

Таким образом,

$$\varepsilon(y_t, K) \geq (1/t - 2/\sqrt{tk} + 1/k) / (1 - 1/k),$$

что стремится к величине $1/t$ с ростом k .

Теорема 1 гарантирует, что если в качестве локального центра искомого кластера рассмотреть векторы линейных оболочек всех t -векторников из X , то один из них приведёт к решению с относительной погрешностью $1/t$.

2. Дискретизация алгоритма

Для эффективного нахождения хорошего приближения искомой точки y_t определим ограниченные области в пространстве, в которых она может находиться.

Пусть K^* — оптимальный кластер, а векторы y_1, \dots, y_t последовательно построены с помощью леммы 1 применительно к K^* . Тогда из геометрических соображений

$$\|y_t - \bar{c}(K^*)\| \leq \|y_1 - \bar{c}(K^*)\| \leq \sqrt{f(K^*)/k}.$$

Величина $f(K^*)$ неизвестна, но она может быть оценена сверху величиной $f_1 = \min_{x \in X} f(K^x)$ — значением целевой функции на приближённом решении, полученном перебором всех векторов множества X в качестве локальных центров искомого кластера. Отсюда

$$\|y_t - \bar{c}(K^*)\| \leq A, \tag{7}$$

где $A = \sqrt{f_1/k}$. Таким образом, в силу неравенства треугольника для нахождения вектора y_t достаточно рассматривать окрестности радиуса $2A$ первых векторов рассматриваемых t -векторников. Данный радиус можно вычислить на предварительном этапе алгоритма.

Пусть x_1, \dots, x_t — произвольный набор векторов из X . Рассмотрим $(t-1)$ -мерную сетку на линейной оболочке множества $\{x_1, \dots, x_t\}$. В качестве центра сетки возьмём вектор x_1 , базиса сетки — ортонормированный базис, получаемый по правилам линейной алгебры из векторов $x_2 - x_1, \dots, x_t - x_1$, а шага сетки — величину $h = 4A/s$, где s — целочисленный параметр алгоритма (наряду с t). Заметим, что евклидов шар радиуса $2A$ покрывают не более чем s^{t-1} ячеек данной сетки, поэтому достаточно рассмотреть s^{t-1} её узлов.

Оценим качество наилучшего сеточного решения. Пусть x_1, \dots, x_t — набор векторов из K^* , определённый в соответствии с леммой 1, в линейной оболочке которого лежит искомый вектор y_t , а y'_t — ближайший к y_t узел соответствующей сетки.

Лемма 2. *Справедливо неравенство*

$$\varepsilon(y'_t, K^*) \leq \varepsilon(y_t, K^*) + 8\zeta(t, s),$$

где $\zeta(t, s) = \sqrt{t-1}/s + (t-1)/s^2$.

ДОКАЗАТЕЛЬСТВО. Расстояние от вектора y_t до y'_t в $(t-1)$ -мерном евклидовом пространстве не превосходит величины $v = \sqrt{t-1}h/2$. Отсюда и из равенства (1) получаем

$$f(y'_t, K^*) - f(y_t, K^*) \leq k((a+v)^2 - a^2),$$

где $a = \|y_t - \bar{c}(K^*)\|$. В силу неравенства (7) имеем $a \leq A$, следовательно,

$$\begin{aligned} (a+v)^2 - a^2 &\leq 2Av + v^2 = A\sqrt{t-1}h + (t-1)h^2/4 \\ &= 4A^2\sqrt{t-1}/s + 4A^2(t-1)/s^2 = 4A^2\zeta(t, s). \end{aligned}$$

Таким образом,

$$f(y'_t, K^*) - f(y_t, K^*) \leq 4kA^2\zeta(t, s) = 4f_1\zeta(t, s).$$

Из теоремы 1 следует, что $f_1 \leq 2f(K^*)$ (данное соотношение также доказано в работе [2]). Отсюда $f(y'_t, K^*) - f(y_t, K^*) \leq 8f(K^*)\zeta(t, s)$. Поделив данное выражение на $f(K^*)$, получим требуемое неравенство. Лемма 2 доказана.

Замечание 3. Арифметические операции над векторами можно реализовать таким образом, что погрешность вычисления узлов сетки будет пренебрежимо мала по сравнению с шагом сетки и, следовательно, не повлияет на оценку погрешности алгоритма.

Теорема 1 и лемма 2 гарантируют, что алгоритм, заключающийся в переборе $n^t s^{t-1}$ кандидатов на роль локального центра искомого кластера, приведёт к решению с относительной погрешностью $1/t + 8\zeta(t, s)$. Поскольку выбор k ближайших к локальному центру векторов множества X занимает не более n действий (например, с помощью предложенного в [4] алгоритма поиска k -го наименьшего числа в массиве из n чисел), а все арифметические операции с векторами линейно зависят от размерности пространства, трудоёмкость данного алгоритма оценивается величиной $O(n^{t+1} s^{t-1} d)$.

Свойство 1. При $t = 1$ алгоритм имеет относительную погрешность 1 (другими словами, относительную точность 2) и совпадает с алгоритмом из [2]. Трудоёмкость алгоритма оценивается величиной $O(n^2 d)$.

Свойство 2. При $t = 2$ алгоритм имеет относительную погрешность $1/2 + \varepsilon$ и трудоёмкость $O(n^3 d/\varepsilon)$, $\varepsilon \in (0, 1]$. В этом случае в качестве s достаточно взять величину $9/\varepsilon$.

Свойство 3. При $t = 2/\varepsilon$, где $\varepsilon > 0$, и $s = 9 t^{3/2}$ алгоритм позволяет решать задачу за время $O(n^{2/\varepsilon+1} (9/\varepsilon)^{3/\varepsilon} d)$ с относительной погрешностью ε .

Действительно, при выбранном s имеем

$$\zeta(t, s) \leq \frac{1}{9t} + \frac{1}{81t^2} \leq \frac{1}{8t},$$

значит, относительная погрешность алгоритма не превосходит величины $2/t = \varepsilon$. Оценка трудоёмкости алгоритма следует из того, что

$$s^{t-1} = (9 t^{3/2})^{t-1} \leq (9 (2/\varepsilon)^{3/2})^{2/\varepsilon} \leq (9/\varepsilon)^{3/\varepsilon}.$$

Таким образом, получена аппроксимационная схема (PTAS).

ЛИТЕРАТУРА

1. Кельманов А. В., Пяткин А. В. NP-полнота некоторых задач выбора подмножества векторов // Дискрет. анализ и исслед. операций. — 2010. — Т. 17, № 5. — С. 37–45.
2. Кельманов А. В., Романченко С. М. Приближённый алгоритм решения одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. — 2011. — Т. 18, № 1. — С. 61–69.
3. Кельманов А. В., Романченко С. М. Псевдополиномиальные алгоритмы для некоторых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. — 2011. — В печати.
4. Wirth Н. Algorithms + data structures = programs // New Jersey: Prentice Hall, 1976. — 366 p.

Шенмайер Владимир Владимирович,
e-mail: shenmaier@mail.ru

Статья поступила
15 июня 2011 г.

Переработанный вариант —
8 сентября 2011 г.