Моделирование систем
2012. №1(31)

УДК 004.4

© 2012 г. **Г.В. Гренкин**
(Дальневосточный федеральный университет, Владивосток)

МЕТОДЫ ВЫЧИСЛИТЕЛЬНОЙ РЕАЛИЗАЦИИ РАНГОВОГО МЕТОДА КЛАСТЕРИЗАЦИИ

Предложена методика автоматизации процесса кластеризации эмпирических данных ранговым методом. Разработаны численные алгоритмы, в том числе получено обобщение алгоритма Грэхема построения выпуклой оболочки.
Ключевые слова: кластеризация, закон Ципфа, минимакс, выпуклая оболочка.

Введение

В различных областях знаний, содержащих большие массивы данных, нередко возникает необходимость группировать эти данные по тем или иным признакам, т.е. кластеризовать. Известными примерами подобных систематизаций являются иерархическая классификация растений и видов М. Адансона (1757), периодическая система элементов Д.И. Менделеева (1869). В то время способы классификации сводились к методу так называемой комбинационной группировки (все характеризующие объект признаки носят дискретный характер). Однако по мере развития электронно-вычислительной техники появилась возможность для этих целей использовать аппарат многомерного статистического анализа, который позволяет всю анализируемую совокупность объектов разбить на сравнительно небольшое число однородных в определенном смысле групп или классов [1].

Современные методы кластеризации довольно разнообразны, поскольку в них по-разному выбирается способ определения близости между объектами, а также используются различные алгоритмы вычислений. Оригинальный метод кластеризации, в котором не используется мера близости и который основан на модифицированном В.П. Масловым законе Ципфа, был предложен в статье М.А. Гузева и Е.В. Черныш [2]. Этот метод получил название рангового метода кластеризации. В статье [2] авторы рассматривают одномерные эмпирические данные: $w_1, w_2, ..., w_n$, – эти данные упорядочиваются по возрастанию, и каждому значению w ставится в соответствие порядковый номер – ранг r. Исходные данные анализируются с помощью соотношения

$$\ln w \cong -g \ln\left(\frac{N-r}{r}\right) + c \equiv -g \ln R + c, \qquad (1)$$

которое представляет собой модифицированный В.П. Масловым закон Ципфа (в

[2] принято $N = 2n + 1$). Авторами были использованы исследования В.П. Маслова, согласно которым для объектов, объединенных некоторым набором признаков, т.е. для определенной группы или кластера, существуют зависимости между соответствующими переменными модели, – например, в виде (1). При разбиении данных на кластеры ранговым методом на каждом из кластеров справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров (g, c), которые меняются при переходе от кластера к кластеру (рис. 1).

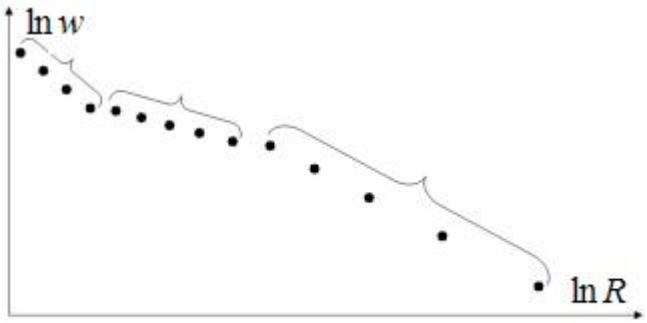

Рис. 1. Ранговый метод кластеризации.

Изначально задача кластеризации эмпирических данных ранговым методом является нечетко поставленной, так как в [2] указана только общая идея кластеризации. В связи с этим для одних и тех же данных выделять кластеры можно по-разному, т.е. разбиение определено неоднозначно. Поэтому разбиение данных на кластеры зачастую может быть субъективным. Таким образом, требуется автоматизировать процесс кластеризации эмпирических данных ранговым методом.

Для вычислительной реализации рангового метода кластеризации исходную постановку задачи необходимо формализовать, требуется построить математическую модель. Программная система должна принимать на вход исходные данные и выводить некоторую информацию, которая должна позволить пользователю анализировать данные с точки зрения рангового метода кластеризации.

Таким образом, необходимо ответить на вопрос – какую информацию нужно получить?

Математическая модель

Исходная формулировка рангового метода кластеризации представляет собой два требования к искомому разбиению: 1) точки каждого кластера близки к прямой $\ln w = -g \ln R + c$; 2) параметры g и c меняются при переходе от кластера к кластеру.

Возникают две задачи: оценка заданного разбиения данных на кластеры; нахождение разбиения.

В ранговом методе кластеризации кластер – это совокупность точек с последовательными рангами $r = a, ..., b$, т.е. промежуток $[a...b]$. Требуется формализовать приближенное соотношение (1) для кластера $[a...b]$. Введем меру отклонения точки $(\ln R, \ln w)$ от прямой $\ln w = -g \ln R + c$:

$$d(\ln R, \ln w, g, c) = \left|\ln w - \left(-g \ln R + c\right)\right|.$$

Зададим пороговое значение d_0 и потребуем, чтобы для всех точек кластера $[a...b]$ выполнялось условие:

$$d(\ln R_r, \ln w_r, g, c) \le d_0, r = a, ..., b.$$

Будем считать, что на кластере $[a...b]$ справедлив модифицированный В.П.

Масловым закон Ципфа с параметрами g, c при пороговом значении d_0, если

$$\max_{a \le r \le b} d(\ln R_r, \ln w_r, g, c) \le d_0 .$$

Введем величину

$$d_{\min}(0,\infty) = \inf_{(g,c)\in \mathrm{R}^2} \max_{a \le r \le b} d(\ln R_r, \ln w_r, g, c).$$

Будем считать, что на кластере $[a...b]$ справедлив модифицированный В.П. Масловым закон Ципфа при пороговом значении d_0, если

$$d_{\min}(0,\infty) \le d_0 .$$

Точку, для которой не выполняется неравенство

$$d(\ln R_r, \ln w_r, g, c) \le d_0 , \qquad (2)$$

будем называть аномальной. Потребуем, чтобы для всех точек кластера $[a...b]$ выполнялось неравенство (2), при этом допускается n_0 аномальных точек, для которых, однако, должно выполняться неравенство

$$d(\ln R_r, \ln w_r, g, c) \le d_0' .$$

Тройку $L = (d_0, n_0, d_0')$ назовем уровнем качества кластера. Будем считать, что на кластере $[a...b]$ справедлив модифицированный В.П. Масловым закон Ципфа на уровне L, если при некоторых значениях параметров g, c выполняется указанное требование.

Введем величину $d_{\min}(n_0, d_0')$ – минимальное значение порогового значения d_0, при котором на кластере $[a...b]$ справедлив модифицированный В.П. Масловым закон Ципфа на уровне $L = (d_0, n_0, d_0')$. Введем величину $n_{\min}(d_0, d_0')$ – минимальное количество аномальных точек n_0, при котором на кластере $[a...b]$ справедлив модифицированный В.П. Масловым закон Ципфа на уровне $L = (d_0, n_0, d_0')$.

Рассмотрим множество промежутков, на которых справедлив модифицированный В.П. Масловым закон Ципфа на уровне L. Те из них, которые нельзя расширить так, чтобы на расширенном промежутке также был справедлив этот закон, образуют множество максимальных промежутков на уровне L.

Пусть задан кластер $[a...b]$ и задана некоторая точка $(\ln R_0, \ln w_0)$, не принадлежащая этому кластеру. Введем расстояние от точки $(\ln R_0, \ln w_0)$ до кластера $[a...b]$. Рассмотрим множество значений параметров (g, c), – таких, что на кластере $[a...b]$ справедлив модифицированный В.П. Масловым закон Ципфа с данными значениями параметров. Выберем из этого множества те значения (g, c), при которых величина $d(\ln R_0, \ln w_0, g, c)$ принимает наименьшее значение. Это значение и будет расстоянием от точки $(\ln R_0, \ln w_0)$ до кластера $[a...b]$ (измеренным как отклонение точки от прямой). Также введем расстояние от точки до кластера, измеренное в количестве точек кластера. Зададим пороговое значение Δ_0 и рассмотрим множество значений параметров (g, c), для которых $d(\ln R_0, \ln w_0, g, c) \le \Delta_0$. Выберем из этого множества такие значения (g, c), что неравенство (2) не выполняется для наименьшего количества точек кластера $[a...b]$. Это количество точек и будет расстоянием от точки $(\ln R_0, \ln w_0)$ до кластера $[a...b]$.

Рассмотрим задачу оценки заданного разбиения данных на кластеры. Требуется оценить разбиение, т.е. определить, удовлетворяет ли оно двум требованиям (см. выше). Во-первых, для оценки каждого кластера в отдельности применим величины $d_{\min}$ и $n_{\min}$ (эти величины характеризуют качество кластера). Во-вторых, проанализируем, как изменяются данные величины при расширении и сужении каждого из кластеров разбиения. Для этого построим таблицу, содержащую значения величины $d_{\min}$ (или $n_{\min}$) для всех возможных промежутков $[a...b]$. В-третьих, для каждого кластера разбиения вычислим расстояния от нескольких точек слева и справа от кластера до этого кластера.

Рассмотрим задачу нахождения разбиения. Процесс кластеризации эмпирических данных ранговым методом мы будем рассматривать как процесс расширения кластеров. То есть если выделен некоторый кластер и он может быть расширен (в него можно включить соседствующие с ним точки), то он расширяется. В результате такого расширения кластеры, как правило, будут пересекаться, и это есть особенность кластеризации эмпирических данных ранговым методом. Итак, чтобы построить разбиение, найдем множество максимальных промежутков. Кроме того, для нахождения разбиения пользователю может быть полезна таблица, содержащая значения величины $d_{\min}$ (или $n_{\min}$) для всех возможных промежутков $[a...b]$.

Таким образом, построена математическая модель, произведена формальная постановка задачи. Требуется разработать численные алгоритмы решения поставленных задач.

Алгоритмы

Обозначим $x = \ln R$, $y = \ln w$. Рассмотрим задачу вычисления величины
$d_{\min}(0,\infty) = \inf\limits_{(g,c)\in \mathrm{R}^2} \max\limits_{a\le r\le b} \left| y_r - (-gx_r + c) \right|$.

Эту задачу можно рассматривать как задачу нахождения многочлена наилучшего равномерного приближения первой степени (см., например, [3]). Многочлен наилучшего равномерного приближения существует и единствен. Если $\left(-g^* x + c^*\right)$ – многочлен наилучшего равномерного приближения, то существуют три узла $x_{r_0} < x_{r_1} < x_{r_2}$, в которых разность $y_{r_k} - \left(-g^* x_{r_k} + c^*\right)$ достигает максимального по модулю значения с последовательной переменой знака (условие альтернанса). Из условия альтернанса вытекает, что прямая $y = -g^* x + c^*$ параллельна одному из ребер выпуклой оболочки множества точек $\{(x_r, y_r)\}$.

Выпуклой оболочкой множества точек называется наименьший выпуклый многоугольник, содержащий все точки данного множества. Для построения выпуклой оболочки в вычислительной геометрии есть алгоритм Грэхема и алгоритм Джарвиса [4].

Таким образом, для вычисления $d_{\min}(0,\infty)$ нужно построить выпуклую оболочку, перебрать все ее ребра и для каждого ребра провести прямую, ему параллельную, – такую, что максимальное из отклонений точек от этой прямой минимально.

Рассмотрим задачу вычисления $d_{\min}(0,\infty)$ с точки зрения проектирования точек на ось ординат во всех возможных направлениях.

Введем оператор проектирования $p_g(x,y)$. Проведем через точку (x, y) прямую с угловым коэффициентом $(-g)$, тогда проекция точки (x, y) на ось ординат – это точка пересечения данной прямой с осью ординат. Пусть $p_g(x,y)$ – координата проекции $p_g(x,y) = xg + y$.

Тогда

$$d_{\min}(0,\infty) = \inf_{(g,c)\in R^2} \max_{a\le r\le b} \left| p_g(x_r,y_r) - c \right| = \min_{g\in R} \min_{c\in R} \max_{a\le r\le b} \left| p_g(x_r,y_r) - c \right| =$$

$$= \frac{1}{2} \min_{g\in R} \left(\max_{a\le r\le b} p_g(x_r,y_r) - \min_{a\le r\le b} p_g(x_r,y_r) \right).$$

Рассмотрим следующие функции:

$$j_r(g) = p_g(x_r,y_r) = x_r g + y_r;\ j_{\max}(g) = \max_{a\le r\le b} j_r(g);\ j_{\min}(g) = \min_{a\le r\le b} j_r(g).$$

Требуется найти минимум функции $j(g) = j_{\max}(g) - j_{\min}(g)$. Функции $j_{\max}(g)$ и $j_{\min}(g)$ кусочно-линейные, следовательно, $j(g)$ – кусочно-линейная функция. Значит, минимум $j(g)$ достигается в одной из точек излома.

Возникает задача нахождения точек излома функций $j_{\max}(g)$ и $j_{\min}(g)$. Графики этих функций – это ломаные.

Рассмотрим алгоритм построения ломаной – графика функции $j_{\max}(g)$ (рис. 2).

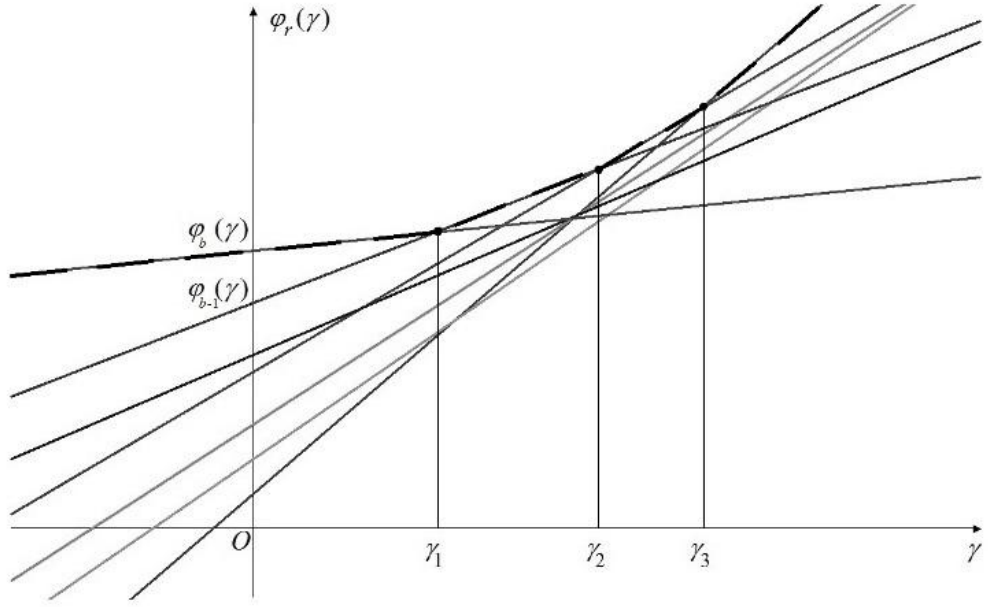

Рис. 2. Алгоритм Джарвиса (пунктиром выделен график функции $j_{\max}(g)$).

Начинаем строить ломаную с прямой – графика функции $j_b(g)$ (эта прямая – текущий максимум). Находим точки пересечения данной прямой с прямыми – графиками функций $j_{b-1}(g)$, …, $j_a(g)$. Выбираем точку пересечения с наименьшей абсциссой g_1. Пусть это точка пересечения с графиком функции $j_{k_1}(g)$. Пря-

мая – график функции $j_{k_1}(g)$ – становится текущим максимумом. Далее находим точки пересечения этой прямой с прямыми – графиками функций $j_{k_1-1}(g)$, …, $j_a(g)$. Выбираем точку пересечения с наименьшей абсциссой g_2. Пусть это точка пересечения с графиком функции $j_{k_2}(g)$. Прямая – график функции $j_{k_2}(g)$ – становится текущим максимумом. Процесс продолжается, пока текущим максимумом не станет прямая – график функции $j_a(g)$.

Описанный алгоритм представляет собой алгоритм Джарвиса построения выпуклой оболочки. Прямые, образующие ломаную, соответствуют вершинам выпуклой оболочки, а абсциссы вершин ломаной соответствуют угловым коэффициентам прямых, содержащих ребра выпуклой оболочки. Время выполнения алгоритма – $O(hM)$, где M – количество точек, h – число вершин выпуклой оболочки.

Опишем алгоритм Грэхема, который имеет оптимальное время выполнения $O(M)$ (рис. 3).

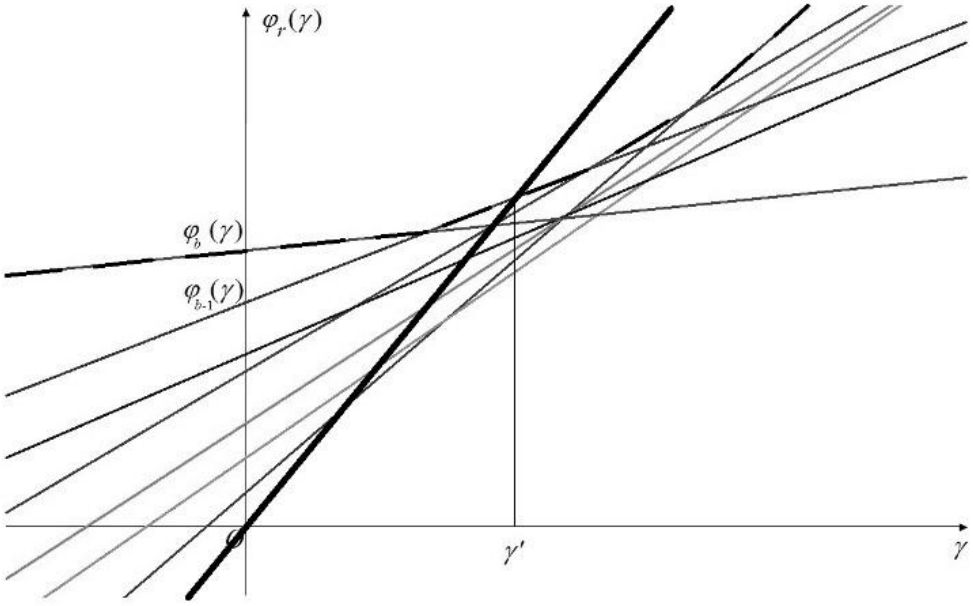

Рис. 3. Алгоритм Грэхема (пунктиром выделен график функции $\max\limits_{q+1\le r\le b} j_r(g)$, жирная прямая – график функции $j_q(g)$).

Будем добавлять прямые по одной: сначала $j_b(g)$, затем $j_{b-1}(g)$, …, $j_a(g)$. Будем последовательно находить $\max\limits_{b-1\le r\le b} j_r(g)$, $\max\limits_{b-2\le r\le b} j_r(g)$, …, $\max\limits_{a\le r\le b} j_r(g)$.

Предположим, что ломаная – график функции $\max\limits_{q+1\le r\le b} j_r(g)$ – уже построена. Укажем, как построить ломаную – график функции $\max\limits_{q\le r\le b} j_r(g)$.

Можно показать, что прямая – график функции $j_q(g)$ – пересекает ломаную – график функции $\max\limits_{q+1\le r\le b} j_r(g)$ – ровно в одной точке с абсциссой g'.

Ломаную – график функции $\max\limits_{q\le r\le b} j_r(g)$ – можно получить так: на проме-

жутке $(-\infty,g']$ эта ломаная совпадает с ломаной – графиком функции $\max\limits_{q+1\le r\le b} j_r(g)$, а на промежутке $[g',+\infty)$ – с прямой – графиком функции $j_q(g)$.

Рассмотрим задачу вычисления величины

$$d_{\min}(n_0,\infty)=\inf_{(g,c)\in R^2}\max_{a\le r\le b}\left|y_r-(-gx_r+c)\right|,$$

где $\max\limits_{a\le r\le b} z_r$ – j-й по максимальности элемент последовательности $z_a, \ldots, z_b$.

Заметим, что

$$d_{\min}(n_0,\infty)=\inf_{(g,c)\in R^2}\max_{a\le r\le b}\left|p_g(x_r,y_r)-c\right|=\min_{g\in R}\min_{c\in R}\max_{a\le r\le b}\left|p_g(x_r,y_r)-c\right|=$$

$$=\min_{g\in R}\left(\frac{1}{2}\min_{0\le n\le n_0}\left(\max_{a\le r\le b}^{n+1} p_g(x_r,y_r)-\min_{a\le r\le b}^{n_0-n+1} p_g(x_r,y_r)\right)\right)=$$

$$=\frac{1}{2}\min_{0\le n\le n_0}\min_{g\in R}\left(\max_{a\le r\le b}^{n+1} p_g(x_r,y_r)-\min_{a\le r\le b}^{n_0-n+1} p_g(x_r,y_r)\right).$$

Рассмотрим следующие функции:

$$j_{\max}^{(n)}(g)=\max_{a\le r\le b}^{n+1} j_r(g),$$

$$j_{\min}^{(n)}(g)=\min_{a\le r\le b}^{n+1} j_r(g),$$

$$j_{(n)}(g)=j_{\max}^{(n)}(g)-j_{\min}^{(n_0-n)}(g).$$

Требуется найти минимумы функций $j_{(n)}(g), n=0,1,\ldots,n_0$; $j_{(n)}(g)$ – кусочно-линейная функция, минимум достигается в точке излома.

Возникает задача нахождения точек излома функций $j_{\max}^{(n)}(g)$ и $j_{\min}^{(n)}(g)$, $n=0,1,\ldots,n_0$.

Рассмотрим алгоритм построения ломаных – графиков функций $j_{\max}^{(0)}(g)$, $j_{\max}^{(1)}(g)$, …, $j_{\max}^{(n_0)}(g)$ (см. рис. 4). Будем добавлять прямые по одной: сначала $j_b(g)$, затем $j_{b-1}(g)$, …, $j_a(g)$. Будем последовательно находить $\max\limits_{q\le r\le b}^{n+1} j_r(g)$ $(n=0,1,\ldots,n_0)$ для $q=b-1,b-2,\ldots,a$.

Предположим, что ломаные – графики функций $\max\limits_{q+1\le r\le b}^{n+1} j_r(g)$ $(n=0,1,\ldots,n_0)$ – уже построены. Укажем, как построить ломаные – графики функций $\max\limits_{q\le r\le b}^{n+1} j_r(g)$ $(n=0,1,\ldots,n_0)$.

Можно показать, что прямая – график функции $j_q(g)$ – пересекает каждую из ломаных – графиков функций $\max\limits_{q+1\le r\le b}^{n+1} j_r(g)$ $(n=0,1,\ldots,n_0)$ – ровно в одной точке с абсциссой $g'_{(n)}$.

Ломаные — графики функций $\max_{q\le r\le b}{}_{n+1}\,j_r(g)$ $(n=0,1,\ldots,n_0)$ – можно получить следующим образом. Ломаная – график функции $\max_{q\le r\le b}{}_1\,j_r(g)$ – на промежутке $(-\infty, g'_{(0)}]$ совпадает с ломаной – графиком функции $\max_{q+1\le r\le b}{}_1\,j_r(g)$, а на промежутке $[g'_{(0)}, +\infty)$ – с прямой – графиком функции $j_q(g)$. Ломаная – график функции $\max_{q\le r\le b}{}_2\,j_r(g)$ – на промежутке $(-\infty, g'_{(1)}]$ совпадает с ломаной – графиком функции $\max_{q+1\le r\le b}{}_2\,j_r(g)$, на промежутке $[g'_{(1)}, g'_{(0)}]$ – с прямой – графиком функции $j_q(g)$, а на промежутке $[g'_{(0)}, +\infty)$ – с ломаной – графиком функции $\max_{q+1\le r\le b}{}_1\,j_r(g)$. Ломаная – график функции $\max_{q\le r\le b}{}_{n+1}\,j_r(g)$ – на промежутке $(-\infty, g'_{(n)}]$ совпадает с ломаной – графиком функции $\max_{q+1\le r\le b}{}_{n+1}\,j_r(g)$, – на промежутке $\left[g'_{(n)}, g'_{(n-1)}\right]$ – с прямой – графиком функции $j_q(g)$, – а на промежутке $[g'_{(n-1)}, +\infty)$ – с ломаной – графиком функции $\max_{q+1\le r\le b}{}_n\,j_r(g)$.

Данный алгоритм представляет собой обобщение алгоритма Грэхема. Время выполнения алгоритма – $O(M(n_0+1))$.

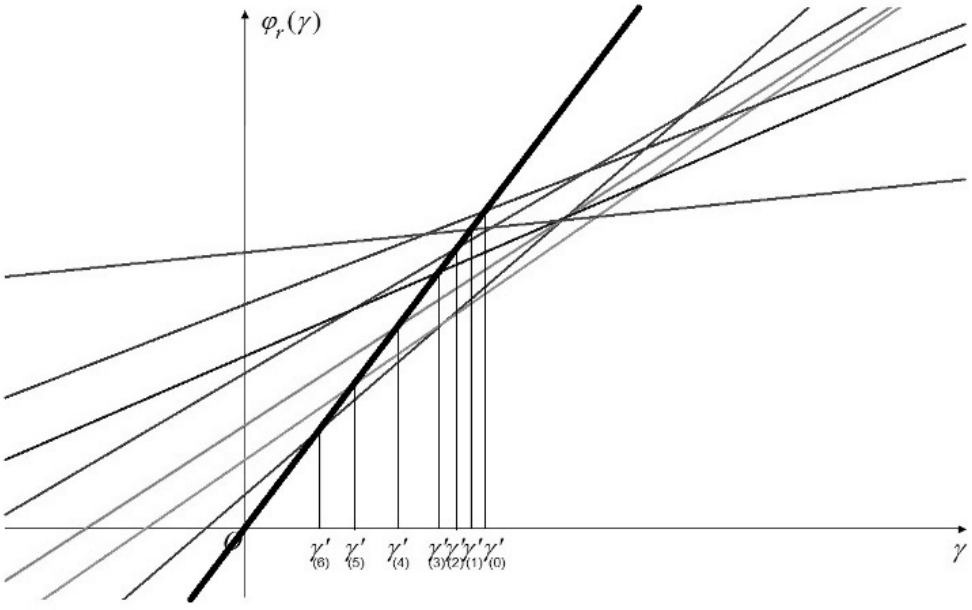

Рис. 4. Обобщенный алгоритм Грэхема.

Заключение

В работе сформулированы требования к программной системе (т.е. какую информацию она должна выводить). Для оценки заданного разбиения данных на

кластеры предлагается оценить каждый кластер в отдельности, вычислив $d_{\min}$ и $n_{\min}$ для каждого кластера, а также проанализировать изменение данных величин при расширении и сужении кластеров разбиения и вычислить расстояния от соседствующих с кластерами точек до кластеров. Для нахождения разбиения можно найти множество максимальных промежутков, а также построить таблицу $d_{\min}$ и $n_{\min}$ для всех возможных промежутков. Данная информация может позволить пользователю анализировать исходные данные с точки зрения рангового метода кластеризации.

Разработаны численные алгоритмы. Применены алгоритмы Джарвиса и Грэхема построения выпуклой оболочки, а также получено обобщение алгоритма Грэхема.

На основе данных результатов может быть разработана программная система, в которой могут быть реализованы описанные методы.

ЛИТЕРАТУРА

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. и др.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989.
2. *Гузев М.А., Черныш Е.В.* Ранговый анализ в задачах кластеризации // Информатика и системы управления. – 2009. – №3(21). – С.13-19.
3. *Демьянов В.Ф., Малоземов В.Н.* Введение в минимакс. – М.: Наука, 1972.
4. *Препарата Ф., Шеймос М.* Вычислительная геометрия: Введение. – М.: Мир, 1989.

Статья представлена к публикации членом редколлегии М.А. Гузевым.

E-mail:
Гренкин Глеб Владимирович – glebgrenkin@gmail.com.