© 2011 г. О. Н. Граничин, д-р физ.-мат. наук,
Д. С. Шалымов, канд. физ.-мат. наук
(Санкт-Петербургский государственный университет)
Р. Аврос, канд. физ.-мат. наук,
З. Волкович, канд. физ.-мат. наук
(ОРТ Брауде Колледж, Израиль)

РАНДОМИЗИРОВАННЫЙ АЛГОРИТМ НАХОЖДЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ

Кластеризация активно изучается в таких областях, как статистика, распознавание образов, машинное обучение и др. Предлагается и обосновывается новый рандомизированный алгоритм нахождения количества кластеров в множестве данных, работоспособность которого демонстрируется примерами имитационного моделирования на синтетических данных с тысячами кластеров.

Ключевые слова: *кластеризация, устойчивость кластеризации, рандомизированные алгоритмы, рандомизированный алгоритм устойчивой кластеризации.*

© 2011 г. О. Н. Граничин, д-р физ.-мат. наук,
Д. С. Шалымов, канд. физ.-мат. наук
(Санкт-Петербургский государственный университет)
Р. Аврос, канд. физ.-мат. наук,
З. Волкович, канд. физ.-мат. наук
(ОРТ Брауде Колледж, Израиль)

РАНДОМИЗИРОВАННЫЙ АЛГОРИТМ НАХОЖДЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ

Кластеризация активно изучается в таких областях, как статистика, распознавание образов, машинное обучение и др. Предлагается и обосновывается новый рандомизированный алгоритм нахождения количества кластеров в множестве данных, работоспособность которого демонстрируется примерами имитационного моделирования на синтетических данных с тысячами кластеров.

1. Введение

На протяжении последних десятилетий в связи со стремительным развитием цифровых технологий наблюдается значительный рост объемов хранимых и обрабатываемых данных. Однако увеличение количества информации не означает непосредственного увеличения объемов знаний. В такой ситуации все более востребованными становятся новые математические методы, которые позволяли бы структурировать информацию, распознавать образы и находить объективные закономерности в больших объемах данных. Особенно это актуально при разработке систем автоматического управления. Среди методов интеллектуального анализа данных (data mining) важную роль играют способы выявления классов (кластеров), особенно те из них, которые способны работать в режиме реального времени. О популярности этих методов сегодня свидетельствует тот факт, что результат поиска по запросу термина “classification problem” в поисковой системе Google (на сентябрь 2010 г.) составил более девяти миллионов страниц.

Вместе с тем во многих областях все чаще используются новые "рандомизированные" подходы к решению задач организации работы сложных систем, которые дают удовлетворительные ответы с высокой степенью вероятности. Большое количество практически важных задач может быть сформулировано в виде: среди всех возможных значений некоторого набора параметров из определенного множества, удовлетворяющих большому набору ограничений (условий), выбрать одно или несколько, которым удовлетворяют дополнительные требования к некоторой целевой функции (например, минимизация, равенство нулю или принадлежность какому-то заданному множеству). На практике множество условий далеко не всегда можно предполагать конечным и даже ограниченным, что существенно увеличивает трудность решения задачи. Если оно все же конечное, но достаточно большое, то это несущественно облегчает решение, так как алгоритмы, основанные на переборе возможных вариантов не смогут закончить свою работу за реальное время. Невозможность детерминированного перебора всех вариантов приводит к стохастическим постановкам задач.

Истоки современных рандомизированных алгоритмов лежат в идеях метода статистического моделирования Монте-Карло, предложенного в Лос-Аламосе Джоном фон Нейманом, Николасом Метрополисом и Станиславом Уламом в ходе работ над Манхэттенским проектом [1]. Первоначально стохастический подход использовали для аппроксимации многомерных интегралов в уравнениях переноса, возникших в связи с задачей о движении нейтрона в изотропной среде. Многие практические задачи сводятся к решению проблемы робастной выпуклой оптимизации (Robast Convex Problems, RCP), в которых оптимизируются выпуклые функции и множество ограничений задается также с помощью выпуклых функций. В общем виде задачи робастной выпуклой оптимизации, как правило, NP-сложные, так как в них обычно для решения требуется проверка огромного количества условий-ограничений. Рандомизированный сценарный подход позволяет при задании малых положительных параметров уровня ϵ и конфиденциальности β априорно выбрать количество случайных проб N, для которых далее решается задача оптимизации. Полученное только для N ограничений решение, удовлетворяет всем остальным с вероятностью $(1-\beta)$. Исключение составляет множе-

ство, чья вероятность не превышает ϵ. В 2008 г. статья Дж. Калафиори и М. Кампи [2] с описанием основных идей сценарного подхода была выделена IEEE CSS как лучшая статья 2006 г.

Кластеризация — одно из наиболее часто проявляющихся свойств разнообразных наборов данных, процессов, систем, широко используемое при решении многообразных задач обработки данных, в том числе при распознавании образов, машинном обучении, автоматической классификации, выработке стратегий управления, исследовании свойств ДНК, а также при сравнении элементов биоразнообразия. В процессе кластеризации множество данных разбивается на группы. Принадлежность к группе математически обычно определяется с помощью функций (метрик), которые задают критерий схожести объектов. Результатом кластеризации является разбиение с наилучшим качеством.

В научных исследованиях и практических разработках массово применяются итеративные методы кластеризации, которые обычно базируются на априорном задании количества кластеров и некотором выборе первоначального разбиения. При этом качество результатов их применения существенно зависит от правильности оценки количества кластеров. В 2003 г. в [3] был предложен простой с вычислительной точки зрения рекуррентный рандомизированный алгоритм для автоматического разбиения множества данных на заранее заданное количество классов, теоретическое обоснование которого было сделано позже в [4]. В настоящее время алгоритмы кластеризации данных активно развиваются, но одной из сложнорешаемых проблем остается вопрос о нахождении истинного количества кластеров в множестве данных.

Устойчивость кластеризации показывает, насколько различными получаются результирующие разбиения на группы после многократного применения алгоритмов кластеризации для одних и тех же данных. Небольшое расхождение результатов интерпретируется как высокая устойчивость. Количество кластеров, которое максимизирует кластерную устойчивость, может служить хорошей оценкой для их реального количества. В настоящее время используется несколько апробированных подходов к исследованию устойчивости кластеризации. Наиболее популярные геометрические подходы к определению количество кластеров основаны на рассмотрение разнообразных индексных функций [5–12]. В [13] вводится

понятие кластеров высокой плотности и количество кластеров определяется как общее число непересекающихся областей, чьи плотности превышают заданное значение. В [14, 15] измеряют устойчивость кластеризации с помощью соотношения количества случаев, когда пара элементов множества, принадлежащая одному кластеру до применения алгоритма кластеризации, оставалась в том же кластере после завершения работы алгоритма. В [16] используют функцию устойчивости кластеризации на основе измерения изоляций Ловингера. В [17] количество кластеров получают с помощью построения минимальных покрывающих деревьев и применения различных вероятностных метрик [18]. Другие подходы использует индекс корреляции внешнего разбиения в качестве измерения устойчивости, например, метод Clest [19] или перезапускаемая процедура с предсказаниями [20]. В [21] использовали дисперсии эмпирических распределений в качестве измерения устойчивости. В [22] для этой задачи строятся статистические модели. Общим недостатком перечисленных методов является то, что вычислительная сложность алгоритмов существенно растет при увеличении мощности исследуемого множества данных и, кроме того, большинство из них недостаточно математически обоснованы. Частое скачкообразное поведение тестовых данных, а также ограничения алгоритмов кластеризации приводят к возникновению шумов и потере информации. Чтобы избежать этого, как правило, производят большее количество итераций, что значительно повышает вычислительную сложность процесса кластеризации.

В [23] были приведены примеры возможного использования сценарного подхода для определения количества кластеров в множествах данных очень большого объема. Исследованный рандомизированный алгоритм поиска наиболее значительного скачка индексной функции типа [12] во многом базируется на предшествующей работе [24], в которой был предложен новый оригинальный алгоритм поиска разрыва у кусочно-непрерывной функции.

В этой статье предлагается новый рандомизированный алгоритм нахождения количества кластеров в множестве данных, который состоит в построении по небольшому количеству точек нескольких равномерных аппроксимаций индексной функции специального вида типа [12]. Новый алгоритм теоретически обосновывается при предположении, что

соответствующая индексная функция имеет единственный "скачок" в точке, соответствующей истинному количеству кластеров. В традиционных предшествовавших алгоритмах необходимо было вычислять соответствующую индексную функцию для всех возможных значений, что на практике при большой мощности исследуемого множества труднореализуемо. Имитационные эксперименты показывают работоспособность нового рандомизированного алгоритма устойчивой кластеризации при определении нескольких тысяч кластеров в условиях неопределенности, на порядок превышающей их действительное количество. При этом использование нового алгоритма позволяет существенно сократить общую вычислительную сложность процесса поиска кластеров.

Статья организована следующим образом. Во разделе 2 описывается способ построения индексной функции Сьюгер-Джеймса. В разделе 3 приводится новый рандомизированный алгоритм устойчивой кластеризации. В разделе 4 сформулирован и обоснован основной результат. Примеры имитационного моделирования, показывающие работоспособность нового алгоритма, приведены в разделе 5. В заключении подведены итоги и намечены пути дальнейшего развития.

2. Индексный метод нахождения количества кластеров

В процедурах разбиения множества на группы часто используются функции внутренних индексов. Правило останова определяет количество кластеров. В случае кластеризации это правило определяет кластеры и экстремальное значение, которое зависит от конкретного правила и определяет наиболее подходящее разбиение.

Исследователями К. Сьюгер и Г. Джеймсом был предложен индексный подход для нахождения количества кластеров, основанный на теории информационного кодирования [12]. В этом подходе для n-мерных данных строится кривая искажений. Предполагается, что множество имеет распределение, составленное из k^* компонент. При этом у всех компонент ковариационные матрицы одинаковые, обозначим их Г. Значение искажения определяется следующим образом:

$$G_k = \frac{1}{n} \min_{c_1,\ldots,c_k} E\,(x - c_x)^{\mathrm{T}}\,\Gamma\,(x - c_x)\,,$$

где $c_1, \ldots, c_k$ — множество из k центров кластеров, полученных после запуска какой-либо стандартной процедуры кластеризации, например алгоритма k-means [8] или из [4]; c_x — ближайший к точке x центр кластера. Здесь и далее E — символ математического ожидания; $\cdot^{\mathrm{T}}$ — символ транспонирования.

По искажениям G_k строится кривая скачков согласно правилу:

$$J(k) = G_k^{-\lambda} - G_{k-1}^{-\lambda},$$

где λ — задаваемая степень трансформации. Исходя из теории информационного кодирования в [12] советуют выбирать λ равной $n/2$.

При достаточно больших размерностях пространства (значений параметра n) значения функции трансформированных искажений $J(k)$ почти нулевые для малых $k < k^*$, далее при $k = k^*$ происходит скачок и кривая начинает медленно линейно расти для $k \geq k^*$. Алгоритм "скачка" Сьюгер–Джеймса использует такое поведение для определения наиболее вероятного значения кластеров k. Оцененное значение количества кластеров соответствует максимальному значению функции J_k. Пример кривой скачков, полученной по данным рис. 1, представлен на рис. 2.

3. Рандомизированный алгоритм

Основываясь на критерии "искажений" Сьюгер-Джеймса [12], задачу поиска истинного количества кластеров можно интерпретировать как отдельный случай более общей задачи, а именно: задачи локализации точек разрыва кусочно-непрерывной функции. Пусть $k_{\max}$ — это максимальное количество кластеров, для которых строилась функции $J(k)$ трансформированных "искажений". Для удобства отмасштабируем аргументы в интервал $[0; 1]$. Обозначим $I(k/k_{\max}) = J(k)$ и назовем $I(\cdot)$ индексной функцией. Для определения точки, в которой происходит скачок, можно использовать рандомизированный сценарный подход, описанный в [24] для следующей более общей задачи "о разладке". Пусть вещественная функция f на интервале $[0; 1]$ имеет не более одного скачка, причем точка скачка $x^* \in (0; 1)$. Обсуждаемая в [24] задача заключается в нахождении доверительного интервала для x^*, если функция f удовлетворяет следующим условиям:

1. Функция $f(\cdot)$ является липшицевой с константой Липщица C на интервалах $(0; x^*)$ и $(x^*; 1)$;

2. Если существует скачок в точке x^* (точка разрыва), то его размер не превышает в этой точке определенной константы $B > 0$.

Константа C представляет "гладкость" индексной функции на тех частях интервала, где она непрерывна. Константа B характеризует возможный скачок индексной функции в точке x^*, которая соответствует здесь истинному значению кластеров. Очевидно, что случай, когда $B \gg C$, представляется наиболее интересным, поскольку тогда отмасштабированная индексная функция $I(x)$ должна вести себя существенно иначе возле точки x^*, чем во всех других точках.

Распространим определенную выше функцию $I(x)$ на весь интервал $[0; 1]$, введя кусочно-линейную отмасштабированную функцию f в виде:

$$f_I(0) = 0;\ f_I(x) = I(k/k_{\max}) \text{ для } x = 1/k_{\max}, 2/k_{\max}, \ldots, 1;$$

$$f_I(x) = f_I(k/k_{\max}) + (x - k/k_{\max})(f_I(k+1) - f_I(k))$$

для $k/k_{\max} \le x \le (k+1)/k_{\max},\ k = 0, \ldots, k^* - 2, k^*, \ldots, k_{\max} - 1;$

$$f_I(x) = f_I(k^* - 1) \text{ для } (k^* - 1)/k_{\max} \le x \le k^*/k_{\max}.$$

Для функции $f_I(\cdot)$ ограничения 1–2 удовлетворяются, если константы C и B выбрать из условий:

$$\mathbf{A1}: \quad C \ge \max_{j=2,\ldots,k^*-1,k^*+1,\ldots,k_{\max}} |J(j) - J(j-1)|,$$

$$\mathbf{A2}: \qquad B \le |J(k^*) - J(k^* - 1)|.$$

Следуя общей логике [24], для поиска истинного значения количества кластеров можно воспользоваться алгоритмом:

1. Выбрать параметр надежности $\beta \in (0, 1)$.

2. Выбрать параметр M, представляющий максимальную степень семейства многочленов Чебышева, которыми будет аппроксимироваться функция f_I:

$$p_m(x) = \cos(m \arccos x),\ m = 0, 1, 2, \ldots, M.$$

3. Выбрать число $N \geq M$ и количество групп точек $T > 1$:

$$T = \left[\frac{1}{N}\left(\frac{4C}{\beta BN} - 1\right)\right] + 1. \quad (1)$$

Здесь $[\cdot]$ — функция целой части числа.

4. Выбрать случайным образом T групп по N точек из интервала $(0; 1)$:

$$Z_t = \{z_{tn},\ n = 1, ..., N\},\ t = 1, \ldots, T.$$

Обозначим

$$Z = \bigcup_t Z_t.$$

(В разделе 4 будет показано, что наибольшее расстояние между двумя соседними точками из Z с вероятностью $(1 - \beta)$ не превосходит $\frac{B}{4C}$).

5. Выберем константу $D > 0$. По каждой группе точек Z_t, $t = 1, \ldots, T$, построим T равномерных аппроксимаций для $f_I(x)$:

$$g_t(x) = \sum_{m=0}^{M} d_{tm} p_m(x),\ t = 1, \ldots, T, \quad (2)$$

минимизирующих погрешности

$$\gamma_t = \max_{x \in Z_t} |g_t(x) - f_I(x)|$$

при условии

$$|d_{tm}| \leq D,\ m = 0, \ldots, M,\ \ t = 1, \ldots, T.$$

Для решения этой задачи выпуклой оптимизации можно воспользоваться, например, Matlab ToolBox (YALMIP, SeDuMi или cvx).

Если какую-либо из задач аппроксимации не удалось решить, то необходимо вернуться на предшествующие шаги и выбрать другие параметры M, N, T, D.

6. Определить результирующую функцию

$$\chi(x) = \max_{t=1,\dots,T} g_t(x) - \min_{t=1,\dots,T} g_t(x),\ x \in (0;1) \tag{3}$$

и

$$h(x) = \max_{z\in[z_l(x),z_r(x)]} \max_{t=1,\dots,T} |g_t'(z)|, \tag{4}$$

где

$$z_l(x) = \arg\max\{z\ \in Z : z \le x\},\ x \in (0;1).$$

и

$$z_r(x) = \arg\min\{z\ \in Z : z > x\},\ x \in (0;1).$$

7. Вычислить

$$\gamma = \max_t \gamma_t \tag{5}$$

и определить линию уровня принятия решения

$$L(x) = \frac{3B}{4} - \frac{B}{4C}h(x) - 2\gamma.$$

С вероятностью

$$P = (1-\beta)$$

интервал

$$\Delta = \{\tilde{x} = xk_{\max} : x \in (0;1),\ \chi(x) > L(x)\} \tag{6}$$

— непустой и значение истинного количества кластеров содержится в Δ.

4. Основной результат

Т е о р е м а 1. Если выполнены условия **A1** *и* **A2**, *тогда с вероятностью* $p = (1-\beta)$ *множество* Δ *непусто и содержит точку* $x^*k_{\max}$, *равную истинному количеству кластеров.*

Д о к а з а т е л ь с т в о. Рассмотрим случайную величину ζ, определяемую расстояниями между двумя соседними точками в Z. Несложно убедиться, что ζ положительна и ее математическое ожидание равно:

$$E\zeta = \frac{1}{NT+1}.$$

Для вероятности события $\{|\zeta| > B/(4C)\}$ в силу неравенства Маркова и условия (1) имеем:

$$\mathsf{P}\left\{|\zeta| > \frac{B}{4C}\right\} \leq \frac{4C}{B(NT+1)} \leq \beta.$$

Следовательно, с вероятностью $(1-\beta)$ существуют две точки z_{i_l} и z_{j_r} в Z:

$$z_{i_l} < x^* \leq z_{j_r}$$

и

$$|z_{j_r} - z_{i_l}| \leq \frac{B}{4C}.$$

Рассмотрим функции g_i или g_j на интервалах $\bar{\Delta}_l = [z_{i_l}; x^*]$ и $\bar{\Delta}_r = [x^*; z_{j_r}]$. Из определения (5) следует, что

$$|f_I(z_{i_l}) - g_i(z_{i_l})| + |f_I(z_{j_r}) - g_j(z_{j_r})| \leq 2\gamma.$$

Следующие соотношения могут быть получены последовательно из приведенных формул и условий алгоритма:

$$\chi(x^*) \geq |g_j(x^*) - g_i(x^*)| \geq |g_j(z_{j_r}) - g_i(z_{i_l})| - (|\bar{\Delta}_l| + |\bar{\Delta}_r|)H \geq$$

$$\geq |f_I(z_{j_r}) - f_I(z_{i_l})| - 2\gamma - (|\bar{\Delta}_l| + |\bar{\Delta}_r|)H \geq B - 2\gamma - (|\bar{\Delta}_l| + |\bar{\Delta}_r|)(H + C) \geq$$

$$\geq B - 2\gamma - \frac{B}{4C}(H + C),$$

где H является максимальной производной $g_i(\cdot)'$ на интервале $[z_{i_l}, z_{j_r}]$. В итоге, учитывая уравнение (4), получаем

$$\chi(x^*) \geq \frac{3B}{4} - \frac{B}{4C}h(x^*) - 2\gamma,$$

откуда следует утверждение теоремы 1.

5. Примеры

1. Первый набор тестовых данных доступен по адресу

   $http://archive.ics.uci.edu/ml/datasets/Libras+Movement.$

Эти данные содержат 15 кластеров, каждый из которых содержит по 24 точки. Каждый кластер характеризует один тип движения руки в официальном бразильском языке жестов LIBRAS. Из видеопотока, в котором записано движение, на основе равномерного распределения выбираются 45 кадров. В каждом кадре находится центральная точка, соответствующая положению руки. Вместе по всем кадрам эти 45 точек образуют дискретную версию кривой F. Все кривые, полученные по видео потокам, нормализованы в унитарном пространстве. Для того чтобы эти кривые можно было использовать в алгоритмах кластеризации, было использовано отображение, сопоставляющее каждой такой кривой вектор, состоящий из 90 свойств.

- Количесто элементов: 360.
- Количество свойств у каждого элемента: 91.

Рассмотрим интервал $[1; 30]$, который предположительно содержит настоящее количество кластеров. Для каждой точки вычислим значение функции трансформированных искажений Сьюгер-Джеймса $J(k)$ на основе алгоритма PAM (Partition Around Medoids) [25]. Полученная кривая изображена на рис. 3. Как видно, она имеет скачок в точке $k^* = 15$ с $B = 0,534$ и

$$C = \max_{k \neq 15}(J(k) - J(k-1)) = 0,049.$$

Таким образом, найдено настоящее количество кластеров. Общее количество раз, когда потребовалось вычислять значение функции $J(k)$, в этом случае было равно 30.

Благодаря описанному в статье методу количество точек, для которых вычисляется значение индексной функции, может быть сокращено. Пусть из каких-либо соображений априорно известно, что $B > 0,5$, а $C < 0,05$, т. е. значения параметров отличаются на порядок. Если выбрать $\beta = 0,95$; $M = 4$; $N = 4$; $T = 3$ и $D = 0,9$, то условие (1) выполняется и для получения доверительного интервала согласно теореме 1 потребуется вычислить значения индексной функции только в 12 точках. В результате получаем три значения $\{0,034;\ 0,022; 0,021\}$ для γ_t, которые соответствуют трем аппроксимирующим кривым $g_t(\cdot)$. Эти кривые представлены на рис. 4 вместе с результирующей функцией $\chi(\cdot)$.

Результирующая функция $\chi(\cdot)$ и уровень принятия решений изображены на рис. 5, на котором выделен 95-процентный доверительный интервал $[11, 21]$. Можно видеть, что функция $\chi(\cdot)$ имеет пик, расположенный вблизи точки $x^* = 15/30$. С ростом параметра B соответствующий достоверный интервал уменьшается.

2. Для того чтобы проверить, насколько эффективен предлагаемый алгоритм в задачах с большим количеством кластеров, были сгенерированы синтетические данные, сгруппированные в 1024 кластера, каждый из которых содержит от 8 до 16 точек. Точки в каждом кластере были сгенерированы согласно равномерному распределению в кругах с радиусами от 10 до 30 (выбиралось случайное значение для каждого кластера).

- Количесто элементов: 11245.
- Количество свойств у каждого элемента: 2.

Искусственные данные представлены на рис. 6.

Рассмотрим интервал $[1; 3100]$, который содержит истинное количество кластеров. Для каждой точки вычислим кривую трансформированных искажений $J(k)$. Результаты представлены на рис. 7.

Описанный выше сценарный подход позволяет существенно сократить количество запусков алгоритма кластеризации. Предполагая, что $B > 0,7$ и $C < 0,07$, выберем $\beta = 0,98$; $M = 8$; $N = 10$; $T = 3$ и $D = 0,7$. Нам потребуется вычислить только 30 значений функции $J(k)$ вместо 3100 для получения 98% доверительного интервала. На рис. 8 изображены три аппроксимирующие кривые $g_t(\cdot)$ вместе с результирующей функцией $\chi(\cdot)$.

Уровень принятия решений изображен на рис. 9 вместе с результирующей функцией $\chi(\cdot)$. Около точки $x^* = 1024/3100$ можно увидеть пик. Если выбрать в качестве доверительного интервала отрезок $[950; 1358]$, то для конечного решения исходной задачи необходимо всего произвести 408 дополнительных вычислений индексной функции $J(k)$. Общее количество вычислений индексной функции в таком случае составит 438, что значительно меньше изначального числа 3100.

6. Заключение

В статье предложен и обоснован новый метод устойчивой кластеризации, базирующийся на идее использования рандомизированного сценарного подхода в комбинации с хорошо известными методами определения количества кластеров, основанными на применении индексных функций. Основная идея заключается в вычислении небольшого количества значений функции искажений и определения местоположения ее скачка на основе аппроксимаций с помощью фиксированного набора полиномов Чебышева. Интервал, содержащий истинное количество кластеров, может быть получен с помощью относительно небольшого количества вычислений функции искажений. Существенное сокращение вычислений доказано теоретически в достаточно общем случае и демонстрируется примерами имитационного моделирования. В дальнейшем авторам хотелось бы использовать новый метод в реальных задачах с неизвестным заранее существенным количеством кластеров.

СПИСОК ЛИТЕРАТУРЫ

1. *Metropolis N., Ulam S.* The Monte Carlo Method // J. Amer. statistical assoc. 1949. V. 44. No. 247. P. 335–341.

2. *Calafiore G., Campi M.C.* The scenario approach to robust control design // IEEE Trans. Automat. Control. May 2006. V. 51. No. 5. P. 742–753.

3. *Граничин О.Н., Поляк Б.Т.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. М.: Наука. 2003.

4. *Граничин О.Н., Измакова О.А.* Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения // АиТ. 2005. № 8. С. 52–63.

5. *Dunn J.C.* Well Separated Clusters and Optimal Fuzzy Partitions // J. Cybern. 1974. No. 4. P. 95–104.

6. *Hubert L., Schultz J.* Quadratic assignment as a general data-analysis strategy // J. Math. Statist. Psychol. 1974. No. 76. P. 190–241.

7. *Calinski R., Harabasz J.* A dendrite method for cluster analysis // Commun. Statistics. No. 3. 1974. P. 1–27.

8. *Hartigan J.* Statistical theory in clustering // J. Classification. 1985. V. 2. P. 63–76.

9. *Krzanowski W., Lai Y* A criterion for determining the number of groups in a dataset using sum of squares clustering // Biometrics. 1985. No. 44. P. 23–34.

10. *Gordon A.D.* Identifying genuine clusters in a classification // Comput. Statistics and Data Analysis. 1994. No. 18. P. 516–581.

11. *Milligan G., Cooper M.* An examination of procedures for determining the number of clusters in a data set // Psychometrika. 1985. No. 50. P. 159–179

12. *Sugar C., James G.* Finding the number of clusters in a data set: An information theoretic approach // J. America Statistical Assoc. 2003. No. 98. P. 750–763.

13. *Hartigan J.* Clustering Algorithms. New York: John Wiley. 1975.

14. *Levine E., Domany E.* Resampling method for unsupervised estimation of cluster validity // Neural Computation. 2001. No. 13. P. 2573–2593.

15. *Ben-Hur A., Elisseeff A., Guyon. I.* Statistical learning theory and randomized algorithms for control // IEEE Control Systems. 1998. No. 12. P. 69–85.

16. *Mufti G.B., Bertrand P., Moubarki L.* Determining the number of groups from measures of cluster validity // Proc. of ASMDA2005. 2005. P. 404–414.

17. *Barzily Z., Volkovich Z., Akteke-Ozturk B., Weber G.-W.* On a minimal spanning tree approach in the cluster validation problem // Informatica. 2009. No. 20. P. 187–202.

18. *Volkovich Z., Barzily Z.* On application of probability metrics in the cluster stability problem // 1st Europ. Conf. on Data Mining (ECDM07), Lisbon, Portugal. 2007. P. 5–7.

19. *Dudoit S., Fridlyand J.* A prediction-based resampling method for estimating the number of clusters in a dataset // Genome Biol. 2002. No. 3. P. 112–129.

20. *Lange T., Roth V., Braun L.M., Buhmann J.M.* Stability-based validation of clustering solutions // Neural Computation. 2004. No. 16. P. 1299–1323.

21. *Jain A.K., Moreau J.V.* Bootstrap technique in cluster analysis // Pattern Recognition. 1987. No. 20. P. 547–568.

22. *Volkovich Z., Barzily Z., Morozensky L.* A statistical model of cluster stability // Pattern Recognition. 2008. No. 41. P. 2174–2188.

23. *Avros R., Granichin O., Shalymov D., Volkovich V.* A randomized algorithm for estimation number of clusters // 24th Europ. Conf. on Operation Research. July 11-14, 2010. Lisbon, Portugal. P. 53.

24. *Граничин О.Н., Халидов В.И.* Рандомизированный подход к обнаружению разрывов функции // Стохастическая оптимизация в информатике. 2005. Т. 1(1). С. 73–80.

25. *Kaufman L., Rousseeuw P.* Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc. 1990.

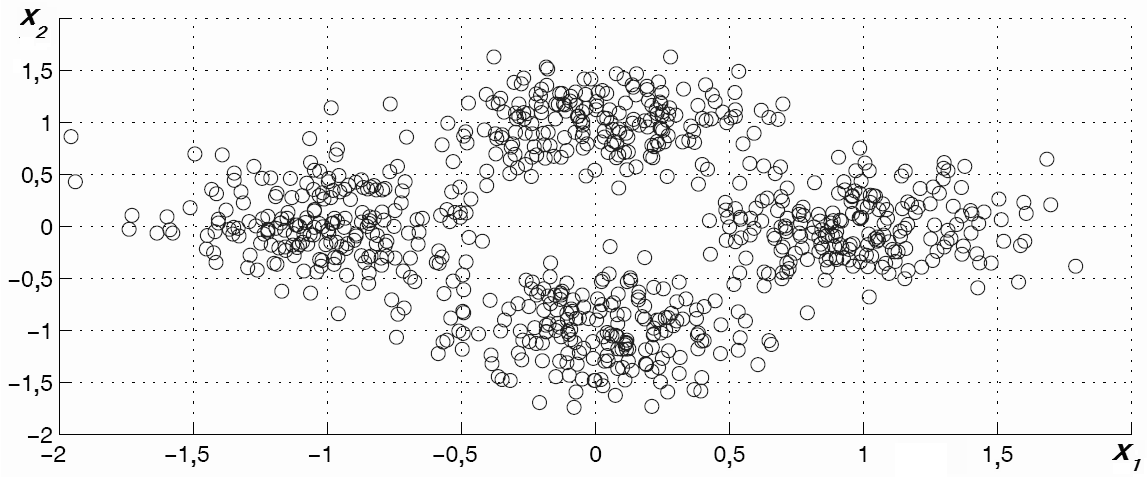

Рис. 1. Изображение данных, состоящих из четырех компонент.

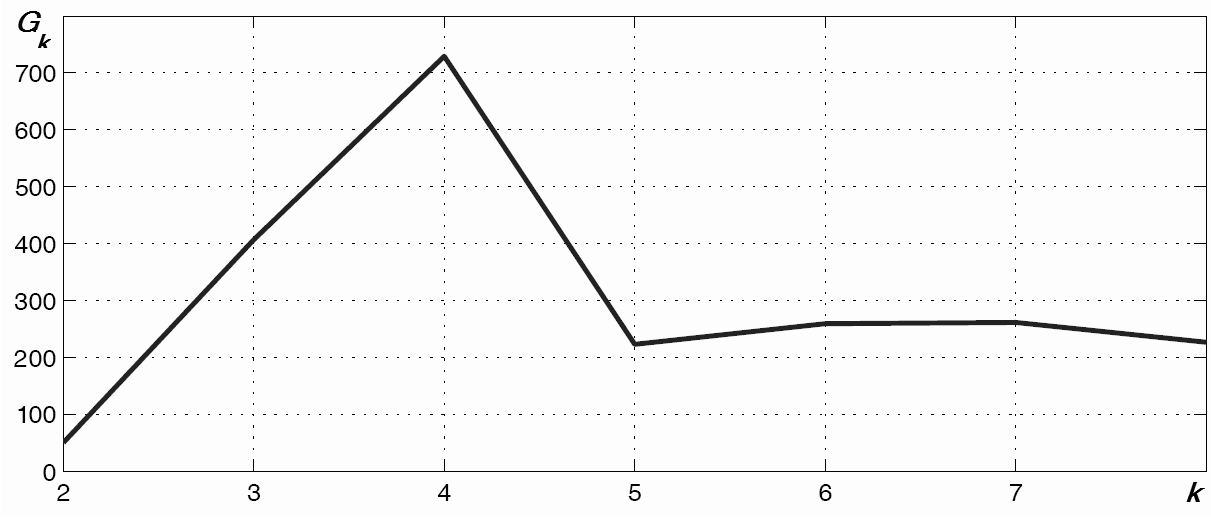

Рис. 2. График кривой искажений.

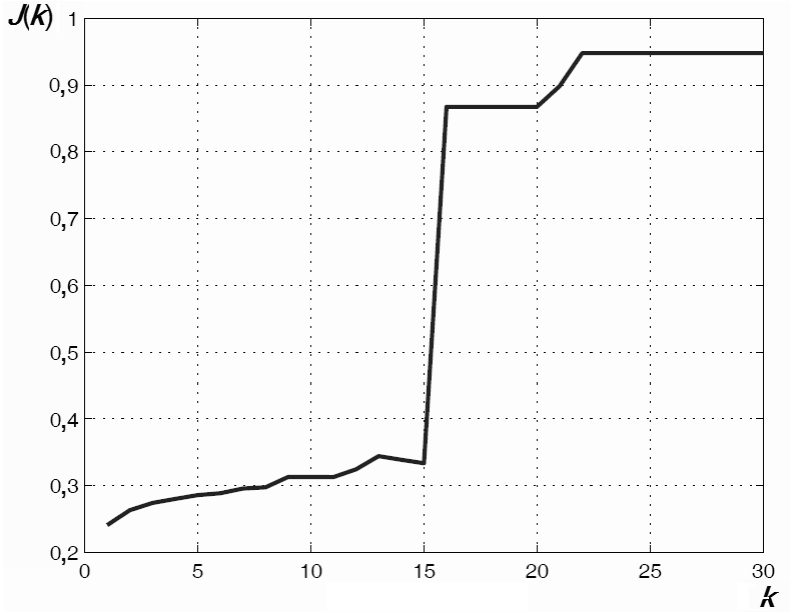

Рис. 3. Функция искажений Сьюгер–Джеймса $J(k)$, вычисленная для официального бразильского языка жестов LIBRAS.

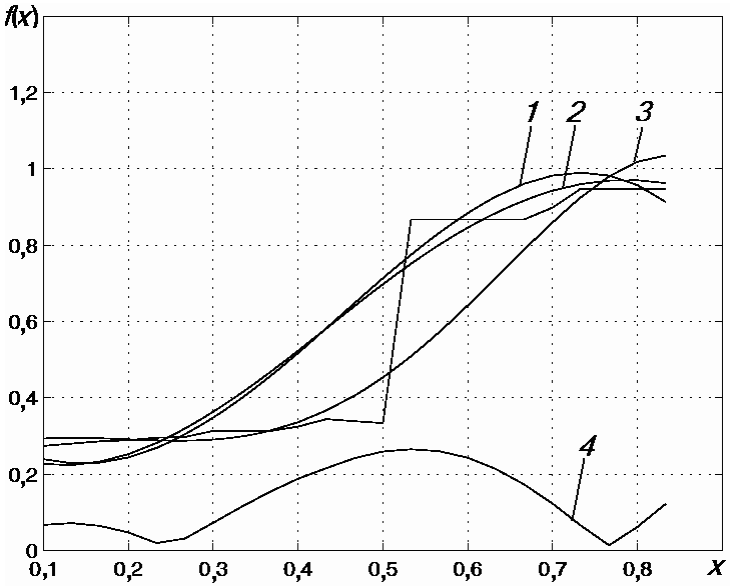

Рис. 4. Аппроксимирующие кривые: 1 — $g_1(\cdot)$, 2 — $g_2(\cdot)$, 3 — $g_3(\cdot)$; результирующая функция $\chi(\cdot)$ — 4.

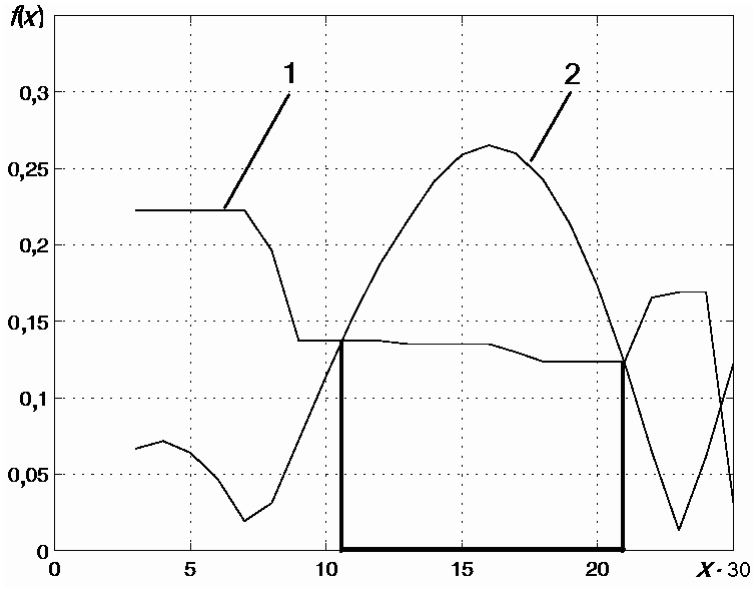

Рис. 5. Уровень принятия решений $L(x) - 1$; результирующая функция $\chi(\cdot) - 2$.

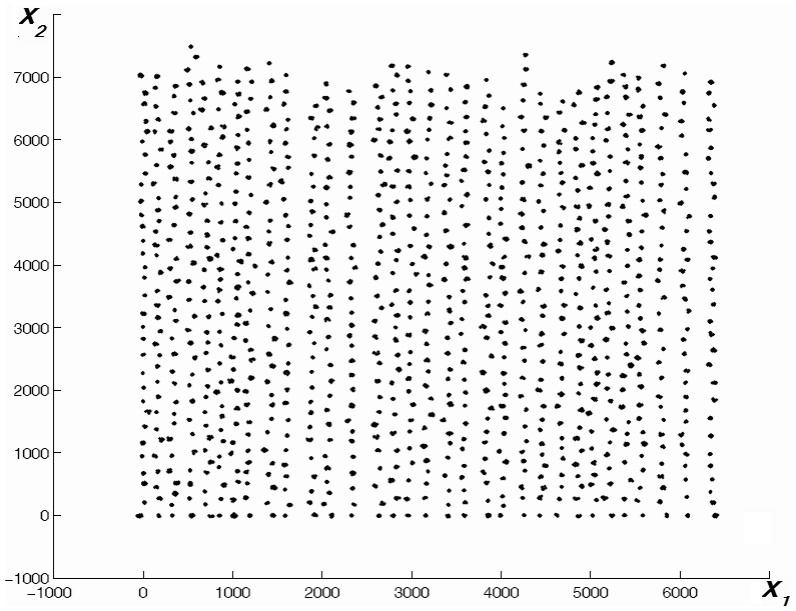

Рис. 6. Синтетические данные, сгруппированные в 1024 кластера.

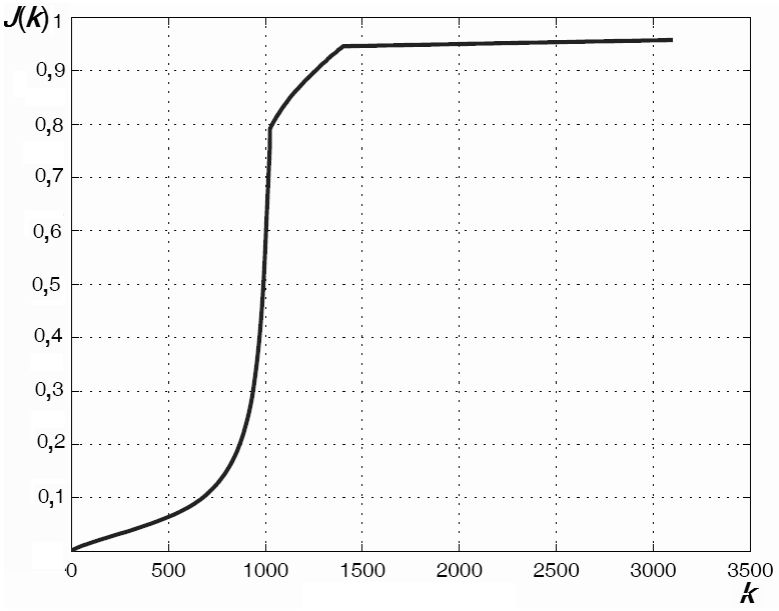

Рис. 7. Функция искажений Сьюгер–Джеймса для синтетический данных.

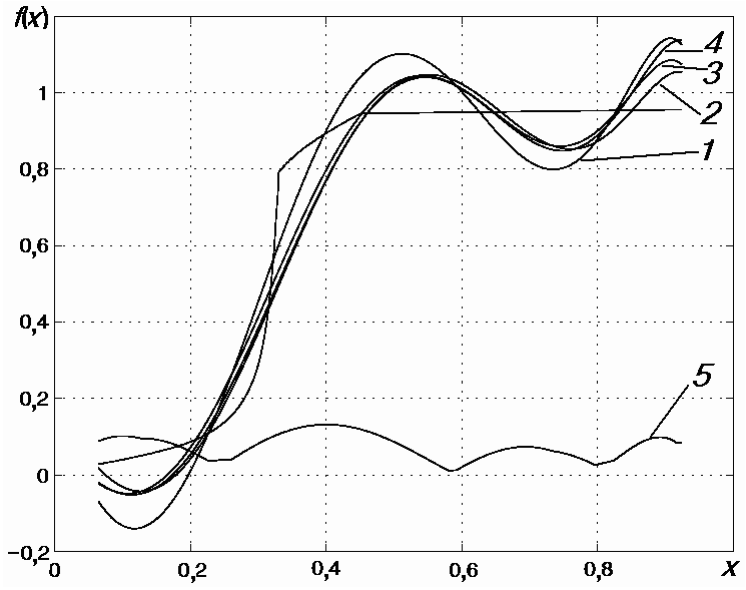

Рис. 8. Для синтетических данных: 1–4 — аппроксимирующие кривые $g_t(\cdot),\ t = 1, 2, 3, 4$; 5 — результирующая функция $\chi(\cdot)$.

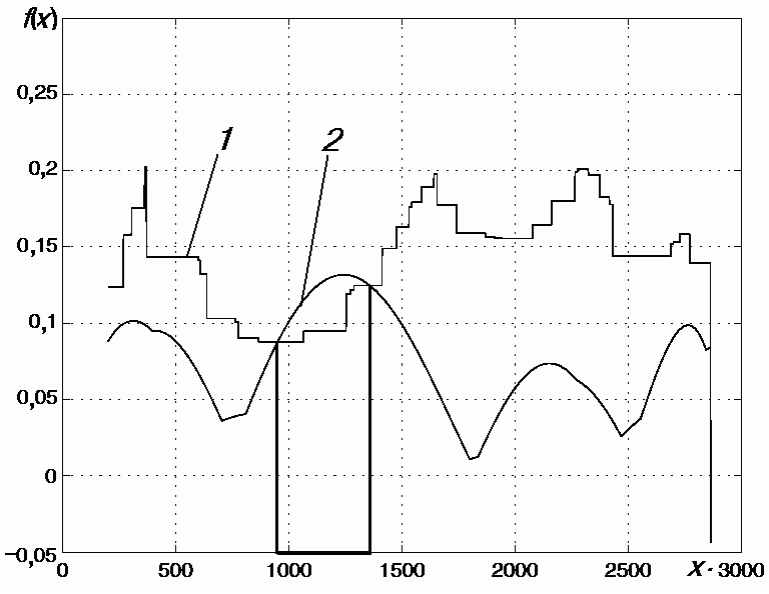

Рис. 9. Для синтетических данных: 1 — уровень принятия решений $L(x)$; 2 — результирующая функция $\chi(\cdot)$.

Подписи к рисункам

Рис. 1. Изображение данных, состоящих из четырех компонент.

Рис. 2. График кривой искажений.

Рис. 3. Функция искажений Сьюгер–Джеймса $J(k)$, вычисленная для официального бразильского языка жестов LIBRAS.

Рис. 4. Аппроксимирующие кривые: 1 — $g_1(\cdot)$, 2 — $g_2(\cdot)$, 3 — $g_3(\cdot)$; результирующая функция $\chi(\cdot)$ — 4.

Рис. 5. Уровень принятия решений $L(x)$ — 1; результирующая функция $\chi(\cdot)$ — 2.

Рис. 6. Синтетические данные, сгруппированные в 1024 кластера.

Рис. 7. Функция искажений Сьюгер–Джеймса для синтетический данных.

Рис. 8. Для синтетических данных: 1–4 — аппроксимирующие кривые $g_t(\cdot),\ t = 1, 2, 3, 4$; 5 — результирующая функция $\chi(\cdot)$.

Рис. 9. Для синтетических данных: 1 — уровень принятия решений $L(x)$; 2 — результирующая функция $\chi(\cdot)$.