Стохастические системы

© 2019 г. А.А. БОЯРОВ (a.boiarov@spbu.ru),
О.Н. ГРАНИЧИН, д-р физ.-мат. наук (o.granichin@spbu.ru)
(Санкт-Петербургский государственный университет;
Институт проблем машиноведения РАН, Санкт-Петербург)

АЛГОРИТМ СТОХАСТИЧЕСКОЙ АППРОКСИМАЦИИ С РАНДОМИЗАЦИЕЙ НА ВХОДЕ ДЛЯ ОЦЕНИВАНИЯ ПАРАМЕТРОВ СМЕСИ ГАУССОВЫХ РАСПРЕДЕЛЕНИЙ БЕЗ УЧИТЕЛЯ ПРИ РАЗРЕЖЕННЫХ ПАРАМЕТРАХ[1]

Рассматриваются возможности применения алгоритмов стохастической аппроксимации с рандомизацией на входе в условиях неизвестных, но ограниченных помех при изучении кластеризации данных, порожденных смесью гауссовых распределений. Предлагаемый алгоритм, устойчивый ко внешним возмущениям, позволяет обрабатывать данные "на лету" и обладает высокой скоростью сходимости. Работа алгоритма иллюстрируется примерами его использования для кластеризации в различных сложных условиях.

Ключевые слова: кластеризация, обучение без учителя, рандомизация, стохастическая аппроксимация, смесь гауссовых распределений.

DOI: 10.1134/S0005231019080051

1. Введение

Большинство задач машинного обучения сводятся к многомерной оптимизации. В условиях существенно зашумленных данных наблюдений стандартные градиентные алгоритмы оптимизации демонстрируют значительное ухудшение качества работы. При этом алгоритмы стохастической аппроксимации с рандомизацией на входе во многих случаях остаются работоспособными.

Основной частью многих методов машинного обучения является решение задачи многомерной оптимизации [1]. Для этих алгоритмов крайне важны такие свойства, как точность, скорость работы и устойчивость к внешним возмущениям. При определении метода оптимизации в теории стандартным подходом является выбор близкой к реальным процессам математической модели и включение в нее различных помех, относящихся к грубости математической модели и характеризующих неконтролируемые внешние возмущения. Во многих прикладных задачах (например, при фильтрации спама или в

[1] Работа выполнена при частичной поддержке Российского научного фонда (грант № 16-19-00057).

потоковых сервисах) зашумленные данные поступают в систему последовательно, и обрабатывать их и принимать решение необходимо "на лету" (*онлайн*). Для обучения таких систем целесообразно использовать рекуррентные адаптивные алгоритмы обработки данных, среди которых часто используют подходы, основанные на стохастической аппроксимации (СА).

Алгоритм СА был впервые предложен Роббинсоном и Монро [2] и был развит для решения задачи оптимизации Кифером и Вольфовицем [3]. Этот подход, основанный на конечно-разностных аппроксимациях вектора-градиента функции потерь, был расширен до d-мерного (многомерного $d > 1$) случая в [4]. Метод использует $2d$ наблюдений на каждой итерации, чтобы построить последовательность оценок: по два наблюдения для аппроксимации каждой компоненты d-мерного вектора-градиента. Граничиным в [5, 6] и Поляком с Цыбаковым в [7] были предложены поисковые алгоритмы стохастической аппроксимации с рандомизацией на входе, которые используют на каждой итерации всего одно (или два) значение исследуемой функции в точке (или точках) на линии, проходящей через предшествующую оценку в случайно выбираемом направлении (как в алгоритме случайного поиска [8]). В англоязычной литературе схожие алгоритмы предложил Спал в [9], дав им название *SPSA* (*simultaneous perturbation stochastic approximation*). Когда в наблюдаемые данные добавлены неизвестные, но ограниченные помехи, качество классических методов, основанных на стохастическом градиенте, падает. Однако качество поисковых алгоритмов СА остается высоким [10, 11].

Для задач кластеризации Ллойд в [12] впервые описал классический метод k-средних (k-means), простота и стабильность которого сделали его популярным. Однако его основным недостатком является то, что он обрабатывает все данные одновременно, поэтому увеличение объема данных требует увеличения доступной памяти в компьютере. Кроме того, в худшем случае временна́я сложность алгоритма Ллойда экспоненциальна [13]. Чтобы исправить эти недостатки было предложено несколько подходов, основанных на идее обучения "на лету". Алгоритм [14] использует *мини-батчи* (подвыборки из тренировочных данных) для уменьшения вычислительного времени, необходимого для сходимости к локальному решению, при этом минимизируя ту же целевую функцию. Полученные таким образом результаты лишь немногим хуже, чем соответствующие результаты оригинального алгоритма. Вариант потокового метода k-средних описан в [13]. Он улучшает скорость сходимости, уменьшает требования к памяти и время работы. Другой метод онлайн кластеризации, основанный на ансамбле обучаемых агентов, рассмотрен в [15]. Более робастным вариантом k-средних является метод k-медоид (и его реализация Partitioning Around Medoids, PAM) [16]. Рандомизированный поисковый алгоритм СА, решающий задачу k-средних был предложен и обоснован в [17].

Смесь гауссовых распределений (СГР) (Gaussian mixture model, GMM) — это вероятностная модель, которая подразумевает, что все элементы данных порождены смесью конечного числа гауссовых распределений с неизвестными параметрами. Будем рассматривать СГР как обобщение кластеризации с помощью метода k-средних. Хорошо известный *EM* алгоритм (expectation-maximization) [18] традиционно используется для нахождения неизвестных

параметров СГР. Он основан на максимизации правдоподобия, когда модель зависит от скрытых переменных. Алгоритм *Variational Bayesian Gaussian mixture inference* является расширением алгоритма *EM*, которое может определять количество компонент в смеси (см. [19]). Этот алгоритм включает регуляризацию с помощью интегрирования информации об априорных распределениях, что делает его стабильнее, но медленнее, чем *EM*. Среди онлайн методов СГР кластеризации отметим потоковый метод, основанный на оценках плотностей [20].

В настоящей статье рассматривается рандомизированный алгоритм стохастической аппроксимации (РАСА), являющийся расширением и обобщением для модели СГР подхода из [17], предложенного ранее для случая k-средних. Предлагаемый алгоритм способен обрабатывать "на лету" потоковые данные и показывает высокую скорость работы. Сходимость центов кластеров к их истинным значениям гарантирована, даже если функционалы качества измерены с неизвестными, но ограниченными помехами. В статье дана оценка временно́й сложности предложенного метода. Работоспособность представленного РАСА сравнивается с другими современными алгоритмами кластеризации, такими как Affinity Propagation [21] и DBSCAN [22].

Статья организована следующим образом. В разделе 2 формулируется проблема кластеризации смеси гауссовых распределений как задача многомерной оптимизации. В разделе 3 предложен новый алгоритм РАСА для кластеризации и исследованы основные свойства его оценок. Раздел 4 содержит описания и результаты экспериментов при наличии и в отсутствие различных типов внешних помех. В разделе 5 описывается использование предложенного алгоритма для кластеризации в случае смеси гауссовых распределений с разреженными параметрами. В заключении подводится итог проделанной работы и обсуждаются направления дальнейших исследований.

2. Постановка задачи

Пусть определены натуральное число $k > 1$, множество входных данных $\mathbb{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots\}$, являющееся подмножеством Евклидова пространства $\mathbb{R}^d$, и заданное на $\mathbb{X}$ вероятностное распределение $\mathrm{P}(\mathbb{X})$. Обозначим через $1, \ldots, k$ множество индексов $\{1, 2, \ldots, k\}$. Будем считать, что множество входных данных $\mathbb{X}$ разбивается на k неизвестных подмножеств $\{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\}$: $\mathbb{X} = \cup_{i\in 1,\ldots,k}\mathbf{X}_i^\star$ таким образом, что вероятностное распределение $\mathrm{P}(\mathbb{X})$ можно представить с помощью смеси распределений: $\mathrm{P}(\mathbb{X}) = \sum_{i=1}^k p_i \mathrm{P}(\mathbf{X}_i^\star)$, где p_i ($p_i > 0$) и $\mathrm{P}(\mathbf{X}_i^\star)$, $i \in 1, \ldots, k$, — соответствующие вероятности и распределения.

Задача кластеризации заключается в нахождении оптимального разбиения $\mathcal{X}$ множества входных данных $\mathbb{X}$ на k непустых кластеров $\mathcal{X}(\mathbb{X}) = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$: $\mathbb{X} = \bigcup\limits_{i=1}^{k} \mathbf{X}_i$ и $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$, $i \neq j$. Разбиение $\mathcal{X}$ определяется функцией $\gamma_{\mathcal{X}} : \mathbb{X} \to 1 \ldots k$, которая ставит в соответствие каждой точке $\mathbb{X}$ индекс кластера. Таким образом, $\mathbf{X}_i = \{\mathbf{x} \in \mathbb{X} | \gamma_{\mathcal{X}}(\mathbf{x}) = i\}$. Задача кластеризации заключается в нахождении лучшего из них: $\mathcal{X}^\star = \{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\}$.

Проблему кластеризации можно описать следующим образом: элементы, принадлежащие одной группе (кластеру), являются более похожими, чем элементы, принадлежащие различным группам (кластерам). Для решения этой задачи вводится некоторая штрафная функция (функция качества) q_i, определяющая "близость" к кластеру i, $i \in 1, \ldots, k$. Тогда для получения оптимальной кластеризации необходимо минимизировать функционал:

$$F(\mathcal{X}) = \mathrm{E} f(\mathcal{X}, \mathbf{x}) \to \min_{\mathcal{X}}, \tag{1}$$

где E — символ математического ожидания и

$$f(\mathcal{X}, \mathbf{x}) = \sum_{i=1}^{k} \gamma_{\mathcal{X}}(\mathbf{x}) q_i(\mathcal{X}, \mathbf{x}).$$

Если векторы $\theta_i,\ i \in 1, \ldots, k$, интерпретировать как *центры кластеров* или *центроиды* и матрицы $\Gamma_i,\ i \in 1, \ldots, k$, — как *ковариационные матрицы*, тогда функционал качества кластеризации (1) принимает вид

$$F(\mathcal{X}) = \sum_{i=1}^{k} \int\limits_{\mathbf{X}_i} q_i(\theta_i, \Gamma_i, \mathbf{x}) \mathrm{P}(d\mathbf{x}) \to \min_{\mathcal{X}}. \tag{2}$$

Для $i \in 1, \ldots, k$ и фиксированного $\mathbf{x} \in \mathbb{X}$ каждая функция $q_i(\cdot, \cdot, \mathbf{x})$ зависит только от θ_i и Γ_i, т.е. $q_i(\cdot, \cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^{d \times d} \times \mathbb{X} \to \mathbb{R}$. Тогда можно выбрать правило разбиения

$$\mathbf{X}_i(\Theta, \Gamma) = \Big\{ \mathbf{x} \in \mathbb{X} :\ q_i(\theta_i, \Gamma_i, \mathbf{x}) < q_j(\theta_j, \Gamma_j, \mathbf{x}),\ j \in 1, \ldots, i-1;$$
$$q_i(\theta_i, \Gamma_i, \mathbf{x}) \leqslant q_j(\theta_j, \Gamma_j, \mathbf{x}),\ j \in i+1 \ldots k \Big\}, \quad i \in 1, \ldots, k,$$

которое минимизирует (1). Здесь $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ — $(d \times k)$-матрица, и Γ — множество, состоящее из k матриц $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$, где $\Gamma_i \in \mathbb{R}^{d \times d}$, $i \in 1, \ldots, k$. Следовательно, (2) можно переписать следующим образом:

$$F(\Theta, \Gamma) = \int\limits_{\mathbb{X}} \langle \mathbf{j}(\Theta, \Gamma, \mathbf{x}), \mathbf{q}(\Theta, \Gamma, \mathbf{x}) \rangle \, \mathrm{P}(d\mathbf{x}) \to \min_{\Theta, \Gamma}, \tag{3}$$

где $\mathbf{j}(\Theta, \Gamma, \mathbf{x}) \in \mathbb{R}^k$ является вектором, состоящим из нулей и единиц, соответствующих значениям характеристических функций $\mathbf{1}_{\mathbf{X}_i(\Theta,\Gamma)}(\Theta, \Gamma, \mathbf{x})$, $i \in 1, \ldots, k$, и $\mathbf{q}(\Theta, \Gamma, \mathbf{x}) \in \mathbb{R}^k$ — вектор значений $q_i(\theta_i, \Gamma_i, \mathbf{x})$, $i \in 1, \ldots, k$.

Важный частный случай соответствует равномерному распределению $\mathrm{P}(\cdot)$ и штрафной функции, представляющей собой квадрат расстояния Махаланобиса

$$q_i(\theta_i, \Gamma_i, \mathbf{x}) = (\mathbf{x} - \theta_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x} - \theta_i). \tag{4}$$

2.1. Смесь гауссовых распределений

В качестве модели описания данных будем использовать одну из самых распространенных: *Смесь гауссовых распределений* (*Gaussian Mixture Model*) (СГР, GMM):

$$f(\mathcal{X}, \mathbf{x}) = f(\Theta, \Gamma, \mathbf{x}) = \sum_{i=1}^{k} p_i G(\mathbf{x}|\boldsymbol{\theta}_i, \Gamma_i), \tag{5}$$

где $G(\mathbf{x}|\boldsymbol{\theta}_i, \Gamma_i)$ — плотность гауссового распределения со средним $\boldsymbol{\theta}_i \in \mathbb{R}^d$ и ковариационной матрицей Γ_i, $i \in 1, \ldots, k$.

Рассмотрим *задачу*: *По последовательности входных данных* $\{\mathbf{x}^1, \mathbf{x}^2, \ldots\}$ *и заданному значению* k *найти параметры* $\boldsymbol{\theta}_i \in \mathbb{R}^d$ *и* Γ_i, $i \in 1, \ldots, k$, *гауссовых распределений, смесь которых породила последовательность входных данных.* Такая задача подходит под введенную выше задачу кластеризации.

Существует много возможных разбиений множества $\mathbb{X}$ на k подмножеств. Задача о нахождении неизвестных параметров $\boldsymbol{\theta}_i \in \mathbb{R}^d$ и Γ_i, $i \in 1 \ldots k$, тесно связана с задачей кластеризации, которая согласно (2) заключается в нахождении:

$$F(\mathcal{X}) = \sum_{i=1}^{k} \int_{\mathbf{X}_i} (\mathbf{x} - \boldsymbol{\theta}_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x} - \boldsymbol{\theta}_i) \mathrm{P}(d\mathbf{x}) \to \min_{\mathcal{X}}. \tag{6}$$

Для последовательности входных данных $\{\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^n\}$ функционал (6) принимает вид

$$F(\Theta, \Gamma) = \sum_{i=1}^{k} \sum_{\mathbf{x}^j \in \mathbf{X}_i} (\mathbf{x}^j - \boldsymbol{\theta}_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x}^j - \boldsymbol{\theta}_i) \to \min_{\Theta, \Gamma}, \quad j \in 1, \ldots, n. \tag{7}$$

Эта связь заключается в том, что при модели данных СГР и при введенной выше постановке задачи кластеризации нахождение оптимальных параметров СГР эквивалентно нахождению оптимальных параметров кластеризации согласно следующей лемме.

Лемма. Пусть последовательность $\{\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^n\}$ *порождена Смесью гауссовых распределений* (5) *с параметрами* $\Theta^\star$ *и* $\Gamma^\star$. *Тогда в заданных выше условиях при* $n \to \infty$ *функционал* (7) *стремится к минимуму при* $\Theta^\star$ *и* $\Gamma^\star$.

Доказательство леммы приводится в Приложении.

Подходы, основанные на максимизации функции правдоподобия, традиционно используются для решения соответствующей задачи кластеризации. *EM* алгоритм [18] или *Variational Bayesian Gaussian mixture inference* алгоритм [19] основаны на построении оценки скрытых переменных. Смесь гауссовых распределений является обобщением метода кластеризации k-средних, которое использует предположение о ковариационных структурах данных так же, как и о центрах скрытых классов.

2.2. Алгоритм k-средних

Рассмотрим один из самых популярных методов кластеризации: алгоритм k-средних, описанный в [12]. Он ищет такое разбиение $\mathcal{X}$, которое минимизирует сумму квадратов внутрикластерных расстояний. Каждый кластер характеризуется соответствующим центроидом. Все матрицы Γ будем считать единичными.

Алгоритм k-средних.

Вход: $\mathbb{X}$, k, *максимальное число итераций*

Выход: *оценка центроидов* $\widehat{\Theta}$, $\mathcal{X}$

1: **Инициализация**: $n := 0$. *Случайно выбирается* k *начальных центроидов* $\widehat{\Theta}^0 = \left(\widehat{\theta}_1^0, \widehat{\theta}_2^0, \ldots, \widehat{\theta}_k^0\right)$ *из элементов* $\mathbb{X}$

2: **Классификация**: $i \in 1 \ldots k$
$\mathbf{X}_i^n = \left\{\mathbf{x} \in \mathbb{X} : \|\widehat{\theta}_i^n - \mathbf{x}\|^2 \leqslant \|\widehat{\theta}_j^n - \mathbf{x}\|^2,\ j \in 1 \ldots k\right\}$

3: **Минимизация**: $\widehat{\theta}_i^{n+1} = \dfrac{1}{|\mathbf{X}_i^n|} \sum_{\mathbf{x}^j \in \mathbf{X}_i^n} \mathbf{x}_j$

4: **Итерация**: $n := n + 1$. *Шаги* 2 *и* 3 *повторяются, пока центроиды не перестанут меняться или не будет достигнуто максимальное число итераций.*

3. Алгоритм СА с рандомизацией на входе для кластеризации

Пусть $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n, \ldots$ — последовательность входных данных, сгенерированных в соответствии с вероятностным распределением (5) с параметрами $\Theta^\star$ и $\Gamma^\star$. В дальнейшем верхний индекс n будем использовать для обозначения номера итерации.

Опишем алгоритм стохастической аппроксимации с рандомизацией на входе для задачи кластеризации со штрафной функцией (3) (квадрат расстояния Махаланобиса). Покажем, что при такой штрафной функции задача (3) имеет решение.

Введем следующие *предположения*:

Предположение 1. Функции $q_i(\cdot,\cdot,\mathbf{x}) : \mathbb{R}^d \to \mathbb{R},\ i \in 1, \ldots, k,$ – дифференцируемы при любом $\mathbf{x} \in \mathbb{X}$ и их градиенты удовлетворяют условию Липшица, т.е. $\|\nabla_\theta q_i(\theta_1, \Gamma_i, \mathbf{x}) - \nabla_\theta q_i(\theta_2, \Gamma_i, \mathbf{x})\| \leqslant M\|\theta_1 - \theta_2\|$, с некоторой константой M, не зависящей от $\mathbf{x}$. Кроме того, каждая из функций $F_i(\theta_i, \Gamma_i) = \int_{\mathbf{X}_i} q_i(\theta_i, \Gamma_i, \mathbf{x})\mathrm{P}(d\mathbf{x})$, $i \in 1, \ldots, k$, имеет единственный минимум в некоторой точке $\theta_i^\star$ и $\langle \theta - \theta_i^\star, \nabla F_i(\theta)\rangle \geqslant \mu\|\theta - \theta_i^\star\|^2, \forall\theta \in \mathbb{R}^d$, с некоторой константой $\mu > 0$ (условие сильной выпуклости);

Предположение 2. Все матрицы Γ_i, $i \in 1, \ldots, k$, являются симметричными положительно определенными, и их собственные значения ограничены $\lambda_i^j \leqslant C_\lambda$, $i \in 1, \ldots, k$, $j \in 1, \ldots, d$;

Предположение 3. Кластеры $\mathbf{X}_i(\Theta^\star, \Gamma^\star)$, $i \in 1, \ldots, k$, значимо разделены между собой: если для некоторого $i \in 1, \ldots, k$, $\mathbf{x} \in \mathbf{X}_i(\Theta^\star, \Gamma^\star)$ и для θ_i, Γ_i

неравенство

$$|q_i(\theta_i, \Gamma_i, \mathbf{x})| \leqslant d_{\max} = \max_{i \in 1 \ldots k} \max_{\mathbf{x} \in \mathbf{X}_i(\Theta^\star, \Gamma^\star)} |q_i(\theta_i^\star, \Gamma_i^\star, \mathbf{x})|$$

выполняется, тогда для $\forall j \neq i$, $j \in 1, \ldots, k$, следующее неравенство верно:

$$|q_i(\theta_i, \Gamma_i, \mathbf{x}')| > d_{\max} + 2c_v \ \ \forall \mathbf{x}' \in \mathbf{X}_j(\Theta^\star, \Gamma^\star). \tag{8}$$

Для всех точек входных данных $\mathbf{x}^n$ и для любой выбранной пары Θ, Γ можем получить зашумленные измерения штрафной функции

$$y_i^n(\Theta, \Gamma) = q_i(\theta_i, \Gamma_i, \mathbf{x}^n) + v_i^n, \quad i \in 1, \ldots, k, \tag{9}$$

где шум v_i^n ограничен: $|v_i^n| \leqslant c_v$, и если шум является случайным, то он не зависит от выбора Θ, Γ, и $\mathrm{E}\{v_i^n\} < \infty$, $\mathrm{E}\{{v_i^n}^2\} \leqslant \sigma^{n2}$.

Обозначим k-векторы значений $y_i^n(\Theta, \Gamma)$ и v_i^n через $\mathbf{y}^n(\Theta, \Gamma)$ и $\mathbf{v}^n$ соответственно; $\mathbf{j}^n(\Theta, \Gamma)$ — k-вектор характеристических функций $\mathbf{j}(\Theta, \Gamma, \mathbf{x})$, найденных с зашумленными измерениями $\mathbf{y}^n(\Theta, \Gamma)$, $n = 1, 2, \ldots$; $\widehat{\Theta}^n$, $\widehat{\Gamma}^n$ — оценки центров и ковариационных матриц кластеров n-го шага алгоритма (т.е. для $\mathbf{x}^n$) соответственно; $l^n = \operatorname{argmax} \mathbf{j}^n(\Theta, \Gamma)$ — индекс кластера, к которому отнесен элемент данных $\mathbf{x}^n$.

Пусть $\Delta^n \in \mathbb{R}^d$, $n = 1, 2, \ldots$ — вектор, состоящий из независимых случайных величин, порожденных распределением Бернулли (стандартно используется параметр распределения Бернулли 0,5), называемых *пробными случайными возмущениями*, k — число кластеров, $\widehat{\Theta}^0 \in \mathbb{R}^{d \times k}$ — матрица начальных значений центроидов, $\widehat{\Gamma}^0$ — набор начальных ковариационных матриц, $\{\alpha^n\}$ и $\{\beta^n\}$ — последовательности положительных чисел.

Для оценки ковариационных матриц кластеров будем использовать матрицы разброса из оценки по методу максимума правдоподобия [23] и параметризованное кумулятивное скользящее среднее оценок ковариационных матриц с регуляризацией. Пусть λ — натуральное число (параметр регуляризации); ω^n — последовательность положительных чисел. Тогда алгоритм СА с рандомизацией на входе для кластеризации строит следующие оценки:

$$\begin{cases} \mathbf{y}_{\pm}^n = \mathbf{y}^n \left(\widehat{\Theta}^{n-1} \pm \beta^n \Delta^n {\mathbf{j}^n}^{\mathrm{T}}, \widehat{\Gamma}^{n-1} \right), \\ \widehat{\Theta}^n = \widehat{\Theta}^{n-1} - {\mathbf{j}^n}^{\mathrm{T}} \alpha^n \dfrac{\mathbf{y}_+^n - \mathbf{y}_-^n}{2\beta^n} \Delta^n {\mathbf{j}^n}^{\mathrm{T}}, \end{cases} \tag{10}$$

$$\Xi_{l^n} = \begin{cases} \omega^n \dfrac{(\widehat{\theta}_{l^n}^{n-1} - \mathbf{x}^n)(\widehat{\theta}_{l^n}^{n-1} - \mathbf{x}^n)^{\mathrm{T}} - \widehat{\Gamma}_{l^n}^{n-1}}{n}, & n > \lambda, \\ I_d \qquad \text{иначе}, \end{cases}$$

$$\widehat{\Gamma}_{l^n}^n = \widehat{\Gamma}_{l^n}^{n-1} + \Xi_{l^n}, \tag{11}$$

где I_d — единичная $(d \times d)$-матрица.

Теорема. Пусть предположения 1–3 и следующие условия выполнены:

1. *Обучающая последовательность* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n, \ldots$ *состоит из одинаково распределенных независимых случайных векторов, таких что они с ненулевой вероятностью принимают значения в каждом из* k *классов в пространстве* $\mathbb{X}$;

2. $\forall n \geqslant 1$ *случайные векторы* $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n$ *и* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{n-1}$ *не зависят от* $\mathbf{x}^n$ *и* Δ^n, *и случайный вектор* $\mathbf{x}^n$ *не зависит от* Δ^n;

3. $\sum_n \alpha^n = \infty$ *и* $\alpha^n \to 0$, $\beta^n \to 0$, $\alpha^n(\beta^n)^{-2} \to 0$ *при* $n \to \infty$, $\omega^n \to 1$ *при* $n \to \infty$, $\lambda < C$.

Если последовательности оценок $\{\widehat{\Theta}^n\}$ *и* $\{\widehat{\Gamma}^n\}$, *полученные с помощью алгоритмов* (10) *и* (11), *удовлетворяют отношению*

$$\overline{\lim_{n\to\infty}} \left\langle \mathbf{j}^n \left(\widehat{\Theta}^{n-1}, \widehat{\Gamma}^{n-1} \right), \mathbf{q} \left(\widehat{\Theta}^{n-1}, \widehat{\Gamma}^{n-1}, \mathbf{x}^n \right) \right\rangle \leqslant d_{\max} + c_v, \tag{12}$$

тогда последовательность $\{\widehat{\Theta}^n\}$ *сходится в среднеквадратичном смысле:* $\lim_{n\to\infty} \mathrm{E}\{\|\widehat{\Theta}^n - \Theta^\star\|^2\} = 0$, *и последовательность* $\{\widehat{\Gamma}^n\}$ *сходится по вероятности:* $\widehat{\Gamma}^n \xrightarrow{p} \Gamma^\star$. *Более того, если* $\sum_n \alpha^n(\beta^n)^2 + (\alpha^n)^2(\beta^n)^{-2} < \infty$, *тогда* $\widehat{\Theta}^n \to \Theta^\star$ *при* $n \to \infty$ *с вероятностью единица.*

Доказательство теоремы приводится в Приложении.

К основным преимуществам описанного РАСА для кластеризации относятся:

- Алгоритм является итеративным, т.е. реализует идею онлайн обучения:
  - адаптивность, обработка новых данных “на лету”;
  - экономия памяти, не нужно хранить весь набор данных в памяти.
- Высокая скорость работы алгоритма.
- Алгоритм сохраняет работоспособность при росте размерности оцениваемых параметров (в отличие от, например, метода k-средних).
- Нечувствительность к шуму при измерении штрафной функции в точках входных данных.

Сравним временну́ю сложность простой базовой версии метода k-средних, описанного в подразделе (2.2) (как специальной версии *EM* алгоритма), и РАСА для кластеризации. В общем случае метод k-средних является NP-сложной оптимизационной задачей. Обозначим через t число итераций алгоритма (максимальное установленное или необходимое для сходимости). Для одной итерации метода k-средних (классификация, минимизация) необходимо kd операций. Тогда временная сложность равна $\mathcal{O}(tNkd)$, так как алгоритм использует все N элементов данных на каждой итерации.

Для сравнения с методом k-средних будем считать, что все матрицы Γ_i, $i \in 1, \ldots, k$, в (4) являются единичными. РАСА для кластеризации необходимо kd операций для вычисления $\mathbf{j}^{n\mathrm{T}}$, kd операций для вычисления каждой из функций $\mathbf{y}^n_-$ и $\mathbf{y}^n_+$. Таким образом, временна́я сложность каждой итерации РАСА для кластеризации составляет $\mathcal{O}(3kd)$. Число итераций равно N. Тогда оценка временно́й сложности всего алгоритма равна $\mathcal{O}(3Nkd)$.

Из полученных оценок временны́х сложностей можно сделать вывод, что в заданных условиях при $t > 3$ алгоритм СА с рандомизацией на входе для кластеризации является более быстрым, чем алгоритм k-средних. Отметим, что в практических задачах t, чаще всего, значительно больше 3.

Метод построения оценок Г (11) добавляет $\mathcal{O}((N-\lambda)d^2)$ к сложности основного алгоритма.

4. Эксперименты в задаче кластеризации

Представим результаты экспериментов, проведенных для сравнения предложенного алгоритма с классическими подходами в задаче кластеризации. Для наглядности в этих экспериментах используются синтетические данные малой размерности. В качестве метрики качества кластеризации выбран индекс Ранда (Adjusted Rand Index, ARI) [24]. Этот критерий позволяет оценить качество кластеризации при заранее известном разбиении данных на кластеры. ARI отображает влияние как неверных положительных (false positive), так и неверных отрицательных (false negative) решений, принятых во время кластеризации. Таким образом, этот критерий позволяет показать, на сколько построенное разбиение данных $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$ близко к эталонному разбиению $\mathcal{X}^\star = \{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\}$.

В начале рассмотрим смесь гауссовых распределений со следующими параметрами: объем данных $N = 5000$, $\Theta = \begin{pmatrix} 0 & 2 & -3 \\ 0 & 2 & 6 \end{pmatrix}$, $\Gamma_1 = \begin{pmatrix} 1 & -0{,}7 \\ -0{,}7 & 1 \end{pmatrix}$, $\Gamma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Gamma_3 = \begin{pmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{pmatrix}$, вероятности в смеси $\mathbf{p} = col(0{,}4,\ 0{,}4,\ 0{,}2)$.

В РАСА для кластеризации были выбраны следующие параметры: $\gamma = 1/6$, $\alpha^n = 0{,}25/n^\gamma$, $\beta^n = 15/n^{\frac{\gamma}{4}}$ и $\omega^n = \mathrm{th}(\frac{n}{\lambda})$. Этот выбор основан на результатах о быстрейшей сходимости СА из [11].

4.1. k-средние

Рисунок 1 иллюстрирует результаты использования РАСА для кластеризации в случае единичных матриц Γ_i, $i \in 1, \ldots, k$, (случай k-средних). Он отображает L_2 расстояния между истинными центроидами и их оценками, полученными на каждом шаге алгоритма, и движение оценок центроидов в зависимости от итераций алгоритма.

Для сравнения качества работы был использован метод k-средних с мини-батчами из [14] с размером мини-батча, равным единице (таким образом, этот

Таблица 1. Средний ARI для эксперимента k-средних

Алгоритм	Среднее ARI
k-средних	0,858
Онлайн k-средних	0,819
PAM	0,84
Affinity Propagation	0,17
DBSCAN	0,41
РАСА для кластеризации	0,857

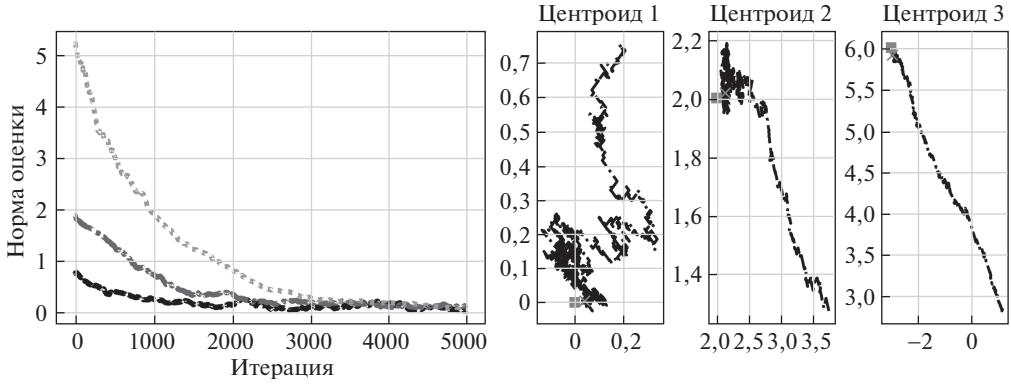

Рис. 1. (Слева) Сходимость L_2-норм оценок центроидов, полученных с помощью PACA для кластеризации в эксперименте с k-средними. (Справа) Движение оценок центроидов, полученных с помощью PACA для кластеризации, (точки) и истинные значения центров кластеров (квадраты) в эксперименте с k-средними.

метод можно назвать онлайн k-средних). Таблица 1 показывает средние значения ARI для каждого алгоритма после 100 запусков. Из этих результатов видно, что качество работы предложенного в статье метода близко к качеству работы методов k-средних, онлайн k-средних и Partitioning Around Medoids (PAM), а также превосходит Affinity Propagation и DBSCAN.

4.2. Смесь гауссовых распределений

Рисунок 2 показывает L_2 расстояния между истинными центроидами и их оценками, полученными с помощью PACA для кластеризации с параметром $\lambda = 1000$, и движение оценок центроидов в случае СГР эксперимента.

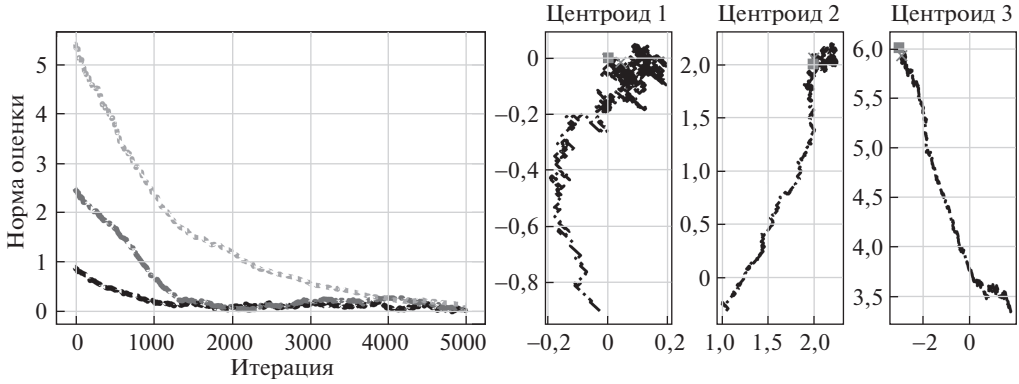

Рис. 2. (Слева) Сходимость L_2-норм оценок центроидов, полученных с помощью PACA для кластеризации в СГР эксперименте. (Справа) Движение оценок центроидов, полученных с помощью PACA для кластеризации, (точки) и истинные значения центров кластеров (квадраты) в СГР эксперименте.

Таблица 2. Средний ARI для СГР эксперимента

Алгоритм	Среднее ARI
EM	0,903
Variational Bayesian Gaussian mixture inference	0,915
Affinity Propagation	0,14
DBSCAN	0,41
PACA для кластеризации	0,909

В этом случае PACA для кластеризации сравним с алгоритмами *EM* и *Variational Bayesian Gaussian mixture inference*. Таблица 2 представляет средние значения ARI для каждого алгоритма после 100 запусков. По этим значениям видно, что предложенный метод близок к двум классическим подходам для обучения СГР, но отметим еще раз, что PACA для кластеризации обладает существенно меньшей временно́й сложностью. Также представленный метод превосходит Affinity Propagation и DBSCAN.

4.3. Внешние возмущения

В подразделах 4.1 и 4.2 проведено сравнение предложенного алгоритма с другими при внешних возмущениях v_i^n в (9), равных нулю. Теперь проведем эксперименты с различными видами шума для всех $i \in 1, \ldots, k$, $n = 1, 2, \ldots$: $v_i^n \sim \mathcal{N}(0,1)$; $v_i^n \sim \mathcal{N}(0, \sqrt{2})$; $v_i^n \sim \mathcal{N}(1,1)$; $v_i^n \sim \mathcal{N}(1, \sqrt{2})$; случайным: $v_i^n = 10\times$ $\times(\text{rand}(\,) \cdot 4 - 2)$; иррегулярным: $v_i^n = 0{,}1 \cdot \sin(n) + 19 \cdot \text{sign}(50 - \text{n} \mod 100)$; константным: $v_i^n = 20$.

В этих условиях были сравнены четыре алгоритма: k-средних (как более стабильный, чем онлайн версия), предложенный в статье метод с единичными ковариационными матрицами (обозначим его PACA единичный), *EM* алгоритм и предложенный в статье метод с оценкой ковариационных матриц с параметром $\lambda = 3000$ (обозначим его PACA ковариационный). Средние значения ARI для каждого алгоритма после 10 запусков представлены в табл. 3.

Как видно из этих результатов, PACA для кластеризации сильно превосходит метод k-средних почти для всех вариантов помех. Версия предложенного алгоритма с оценкой ковариационных матриц также почти во всех случаях превосходит и *EM*, и версию с единичными матрицами.

Таблица 3. Средний ARI для эксперимента с помехами

Шум	k-средние	PACA единич.	*EM*	PACA ков.
$\mathcal{N}(0,1)$	0,608	0,768	0,654	0,815
$\mathcal{N}(0,\sqrt{2})$	0,246	0,546	0,463	0,738
$\mathcal{N}(1,1)$	0,612	0,829	0,657	0,774
$\mathcal{N}(1,\sqrt{2})$	0,246	0,601	0,371	0,612
Случайный	0	0,418	0,318	0,434
Иррегулярный	0,758	0,854	0,863	0,856
Константный	0,812	0,861	0,807	0,860

5. РАСА при разреженных параметрах смеси гауссовых распределений

В [25] была предложена интересная модель смеси гауссовых распределений с *разреженными* (*sparse*) параметрами (разреженная СГР, sparse GMM). Эта модель имеет среднее

$$\theta_i \in \mathbb{R}^d, \quad \theta_{il} \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \sim \mathcal{C}_+(0,1), \quad l \in 1 \dots d,$$

где $\mathcal{C}_+(0,1)$ означает распределение Коши, ограниченное на позитивной оси (принимает только позитивные значения), с параметрами 0 и 1. Диагональная ковариационная матрица задается с помощью

$$\Gamma_i = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2), \quad \sigma_j \sim \mathcal{C}_+(0;0{,}5), \quad j \in 1 \dots d. \tag{13}$$

Веса $p_i \sim \mathcal{D}(e_0, \dots, e_0)$, $i \in 1 \dots k$, предполагаются принадлежащими распределению Дирихле, где параметр $e_0 \sim \mathcal{G}(\alpha_p, k\alpha_p)$ получен из гамма-распределения со средним k^{-1} и дисперсией $(\alpha_p k^2)^{-1}$, $\alpha_p = 10$.

В разреженной СГР можно выделить три основных типа поведения кластеров (рассмотрим случай $k = 3$, $d = 2$) (рис. 3). При такой модели данных часто возникают ситуации, когда данные из одного кластера лежат внутри другого, что делает процедуру кластеризации сложной.

Авторы [25] использовали этот подход для моделирования множества различных интересных распределений шума. РАСА для кластеризации, оценивающий только центроиды кластеров Θ (как алгоритм k-средних) показывает

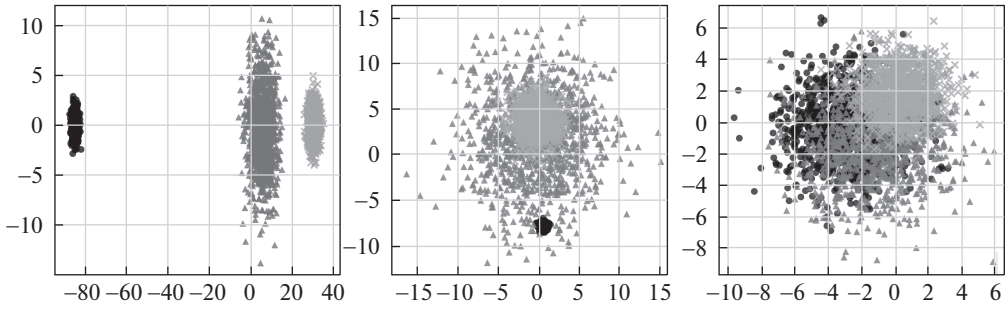

Рис. 3. Типы разреженной СГР: 1, 2, 3.

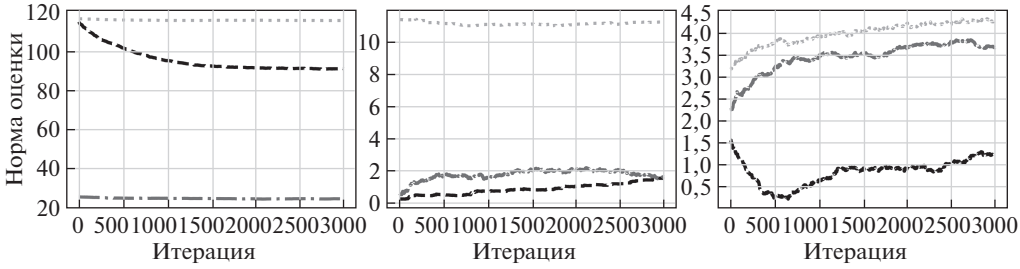

Рис. 4. РАСА k-средних сходимость центроидов: типы 1, 2, 3.

слабые результаты для данных, полученных из модели разреженной СГР (см. подраздел 5.2). Рисунок 4 демонстрирует графики L_2-норм расстояний между истинными центрами кластеров и их оценками, полученными на каждом этапе алгоритма. Как видно из этих графиков, РАСА k-средних демонстрирует слабую сходимость.

5.1. РАСА для кластеризации при разреженной СГР

Рассмотрим подходы для улучшения качества при разреженной СГР. Во-первых, алгоритмы (10) и (11) позволяют построить более робастные оценки в случае sparse GMM.

Во-вторых, добавление нового L_2 регуляризатора в функцию потерь (7) позволяет сделать оценки центроидов $\widehat{\theta}_i$, $i = 1 \dots k$, ближе к центрам разреженной СГР:

$$\psi^n \left\| \theta_{l^n} - \xi \right\|^2, \tag{14}$$

где $l^n = \operatorname{argmax} \mathbf{j}^n(\Theta, \Gamma)$ – индекс кластера, к которому отнесена точка данных $\mathbf{x}^n$; $\xi \in \mathbb{R}^n$, $\xi_l \sim \mathcal{N}(0, \sigma^2)$, $\sigma \sim \mathcal{C}_+(0,1)$, $l \in 1, \dots, d$; ψ^n – последовательность возрастающих положительных чисел.

В-третьих, добавление L_1 регуляризатора в функцию потерь (7) по каждой размерности центроидов позволяет сделать их ближе к разреженному среднему $\theta_{il} \sim \mathcal{N}(0, \sigma_i^2)$, $\sigma_i \sim \mathcal{C}_+(0,1)$, $i \in 1 \dots k$, $j \in 1 \dots d$:

$$\tau^n \sum_{i=1}^{d} \sum_{j=1}^{k} |\widehat{\Theta}_{ij}|, \tag{15}$$

где τ^n – последовательность возрастающих положительных чисел.

5.2. Эксперименты

Для экспериментов были выбраны значения $k = 3$, $d = 2$, $N = 3000$. Параметры РАСА для кластеризации: $\gamma = 1/6$, $\alpha^n = 0{,}25/n^\gamma$, $\beta^n = 15/n^{\frac{\gamma}{4}}$, $\omega^n = \operatorname{th}(\frac{n}{\lambda})$, $\psi^n = 1e^{-5}$, $\tau^n = 1e^{-4} \operatorname{th}(\frac{n}{\lambda})$. В качестве метрик для сравнения алгоритмов были выбраны ARI и среднее L_2-расстояние между центроидами, полученными с помощью алгоритма, и их истинными значениями.

Сравнение РАСА k-средних и РАСА для кластеризации при разреженной СГР после 1000 запусков экспериментов представлены в табл. 4.

Для разреженной СГР типа 1 модифицированный РАСА дает: ARI = 1,0, L_2-расстояние = 0,130 (рис. 5).

Таблица 4. ARI и L_2-расстояние

Алгоритм	Среднее ARI	Среднее L_2-расстояние
РАСА k-средних	0,134	1,830
РАСА для разреженной СГР	0,518	1,617

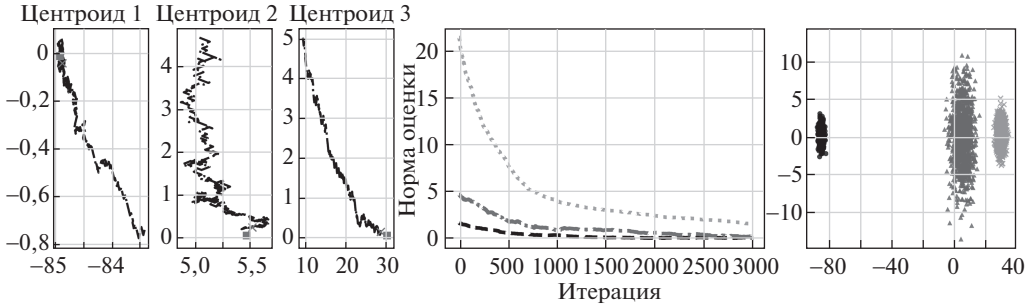

Рис. 5. Тип 1: сходимость центроидов, движение оценок центроидов, кластеризация PACA для разреженной СГР.

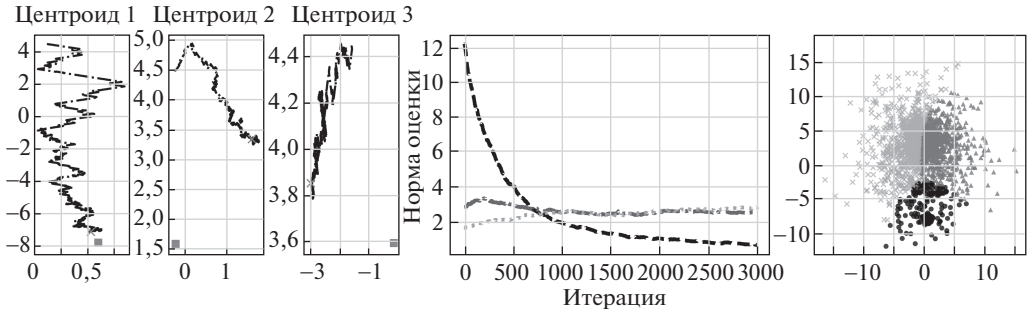

Рис. 6. Тип 2: сходимость центроидов, движение оценок центроидов, кластеризация PACA для разреженной СГР.

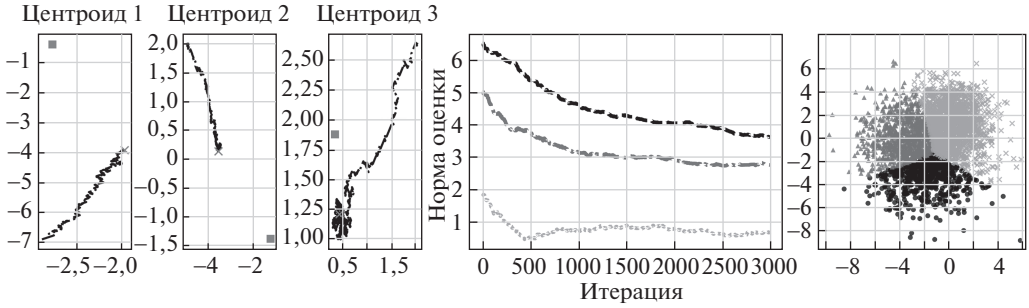

Рис. 7. Тип 3: сходимость центроидов, движение оценок центроидов, кластеризация PACA для разреженной СГР.

Для разреженной СГР типа 2 модифицированный PACA дает: ARI = = 0,521, L_2-расстояние = 0,923 (рис. 6).

Для разреженной СГР типа 3 модифицированный PACA дает: ARI = = 0,277, L_2-расстояние = 2,051 (рис. 7).

Также были проведены эксперименты с PACA-оценкой σ в (13). Результаты были получены многообещающие, но нестабильные.

5.3. Результат

Практический результат заключается в том, что использование (11), (14) и (15) в РАСА для кластеризации позволяет улучшить результаты для данных, порожденных смесью гауссовых распределений с разреженными параметрами. Теоретический результат заключается в том, что (11), (14) и (15) могут ослабить предположение 3 теоремы о значимой разделимости кластеров. Причиной для такого предположения является результат, полученный выше, так как при разреженной СГР кластеры могут лежать друг в друге (тип 2 и тип 3).

6. Заключение

В статье представлен алгоритм СА с рандомизацией на входе для кластеризации и описано его теоретическое обоснование. Продемонстрирована робастность этого метода в случае, когда данные измерены с почти произвольными внешними помехами. В дополнение к этому свойству представленный алгоритм реализует идею обучения "на лету" и обладает высокой скоростью работы. Продемонстрировано применение предложенного метода для кластеризации смеси гауссовых распределений, в том числе и при разреженных параметрах этой смеси. В серии экспериментов показаны преимущества алгоритма перед классическими подходами.

ПРИЛОЖЕНИЕ

Доказательство леммы.

Если последовательность $\{\mathbf{x}^n\}$ порождена СГР с параметрами $\Theta^\star$ и $\Gamma^\star$, то тогда с этими параметрами достигается минимум минус логарифм функции правдоподобия $-\sum_{\mathbf{x}\in\mathbb{X}} \ln\left(\sum_{i=1}^k p_i G(\mathbf{x}|\theta_i,\Gamma_i)\right)$. По определению плотности нормального распределения это выражение может быть переписано как

$$\text{(П.1)}\quad \sum_{i=1}^{k}\sum_{\mathbf{x}^j\in\mathbf{X}_i}\left(-\ln p_i(2\pi)^{-\frac{d}{2}}|\Gamma_i|^{-1}+\frac{1}{2}(\mathbf{x}^j-\theta_i)^{\mathrm{T}}\Gamma_i^{-1}(\mathbf{x}^j-\theta_i)\right),\quad j\in 1,\ldots,n.$$

Обозначим первое слагаемое в (П.1) через L, а второе – через R, тогда

$$\arg\min_{\Theta,\Gamma} L=\arg\min_{\Theta,\Gamma} R=(\Theta^\star,\Gamma^\star)\quad\text{при}\quad n\to\infty.$$

Функционал (7) по условию может быть переписан через (2) и (4) как

$$F(\Theta,\Gamma)=\sum_{i=1}^{k}|\mathbf{X}_i|^{-1}\sum_{\mathbf{x}^j\in\mathbf{X}_i}(\mathbf{x}^j-\theta_i)^{\mathrm{T}}\Gamma_i^{-1}(\mathbf{x}^j-\theta_i),\quad j\in 1,\ldots,n.$$

Таким образом, $\arg\min_{\Theta,\Gamma} F(\Theta,\Gamma)=\arg\min_{\Theta,\Gamma} R=(\Theta^\star,\Gamma^\star)$ при $n\to\infty$.

Доказательство леммы закончено.

Доказательство теоремы.

1. На первом шаге доказательства зафиксируем Γ и докажем сходимость $\{\widehat{\Theta}^n\}$. Выберем некоторое r из $1, 2, \ldots, k$ и такую подпоследовательность $\{\widehat{\Theta}^{n_j}\}$, при построении которой изменениям подверглись только оценки центра r-го кластера.

Необходимо доказать, что $\exists\ N_r$ такое, что $\mathbf{x}^{n_j} \in \mathbf{X}_r(\Theta^\star, \Gamma^\star)\ \ \forall j \geqslant N_r$. Предположим, что это не так. Тогда $\exists$ бесконечная возрастающая подпоследовательность $\{n_{j_t}\}$, для которой $\mathbf{x}^{n_{j_t}} \in \mathbf{X}_r(\Theta^\star, \Gamma^\star)$ и $\mathbf{x}^{n_{j_t+1}} \notin \mathbf{X}_r(\Theta^\star, \Gamma^\star)$. Сравним значения $q(\widehat{\theta}_r^{(n_{j_t}-1)}, \Gamma_r, \mathbf{x}^{n_{j_t}})$ и $q(\widehat{\theta}_r^{(n_{j_t})}, \Gamma_r, \mathbf{x}^{n_{j_t+1}})$. Сначала оценим $\|\widehat{\theta}_r^{(n_{j_t}-1)} - \widehat{\theta}_r^{n_{j_t}}\|$. В силу алгоритма (10), теоремы о среднем, компактности множества Θ и условия предположения 1:

$$\left\|\widehat{\theta}_r^{(n_{j_t}-1)} - \widehat{\theta}_r^{n_{j_t}}\right\| \leqslant$$

$$\leqslant \alpha^{n_{j_t}} m \left(2C_v + \max_{s\in[-1;1]} \left|\nabla_\theta q\left(\widehat{\theta}_r^{n_{j_t}-1} + s\alpha^{n_{j_t}}\Delta^{n_{j_t}}, \Gamma_r, \mathbf{x}^{n_{j_t}}\right)\right|\right) \leqslant \alpha^{n_{j_t}} C$$

с некоторой константой C. Следовательно, для достаточно больших t разность $|q(\widehat{\theta}_r^{n_{j_t}}, \Gamma_r, \mathbf{x}^{n_{j_t+1}}) - q(\widehat{\theta}_r^{(n_{j_t}-1)}, \Gamma_r, \mathbf{x}^{n_{j_t+1}})| < C_v/3$. Так как $\mathbf{x}^{n_{j_t+1}} \notin \notin \mathbf{X}_r(\Theta^\star, \Gamma^\star)$, то по (8) $|q(\widehat{\theta}_r^{(n_{j_t}-1)}, \Gamma_r, \mathbf{x}^{n_{j_t+1}})| > d_{\max} + 2C_v$. Из этих отношений для достаточно большого t имеем неравенство

$$\left|q\left(\widehat{\theta}_r^{n_{j_t}}, \Gamma_r, \mathbf{x}^{n_{j_t+1}}\right)\right| \geqslant \left|q\left(\widehat{\theta}_r^{(n_{j_t}-1)}, \Gamma_r, \mathbf{x}^{n_{j_t+1}}\right)\right| -$$

$$- \left|q\left(\widehat{\theta}_r^{n_{j_t}}, \Gamma_r, \mathbf{x}^{n_{j_t+1}}\right) - q\left(\widehat{\theta}_r^{(n_{j_t}-1)}, \Gamma_r, \mathbf{x}^{n_{j_t+1}}\right)\right| > d_{\max} + 5/3C_v.$$

С другой стороны, по условию (12) для достаточно большого t

$$\left|q\left(\widehat{\theta}_r^{(n_{j_t-1})}, \Gamma_r, \mathbf{x}^{n_{j_t}}\right)\right| < d_{\max} + \frac{4}{3}C_v.$$

Получили противоречие.

Перенумеруем для удобства последовательность $\{\widehat{\theta}_r^{n_j}\}$ с $j = N_r$ в $\{\widehat{\theta}_r^{i}\}$. По (10) имеем $\|\widehat{\theta}_r^{i} - \theta_r^\star\|^2 \leqslant \|\widehat{\theta}_r^{(i-1)} - \theta_r^\star - \frac{\alpha^i}{2\beta^i}\left(y_+^i - y_-^i\right)\Delta^i\|^2$.

Обозначим через $\mathcal{F}_r^{n-1}$ σ-алгебру событий, порождаемых случайными величинами $\widehat{\theta}_r^0, \widehat{\theta}_r^1, \ldots, \widehat{\theta}_r^{(n-1)}$, полученными по алгоритму (10). Тогда получаем, что

$$\text{(П.2)} \qquad \mathrm{E}_{\mathcal{F}_r^{i-1}}\|\widehat{\theta}_r^{i} - \theta_r^\star\|^2 \leqslant \|\widehat{\theta}_r^{i-1} - \theta_r^\star\|^2 -$$

$$- \alpha^i\left\langle\widehat{\theta}_r^{(i-1)} - \theta_r^\star, (\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i\left(y_{r,+}^i - y_{r,-}^i\right)\right\rangle +$$

$$+ \frac{(\alpha^i)^2}{4(\beta^i)^2}\mathrm{E}_{\mathcal{F}_r^{i-1}}\|\Delta^i\|^2\left(y_{r,+}^i - y_{r,-}^i\right)^2.$$

Оценим последнее слагаемое в правой части неравенства (П.2). Функции $q(\cdot,\Gamma_i,\mathbf{x}_i)$ удовлетворяют условию Липшица, $\|\Delta^i\|^2=m$, тогда, используя теорему о среднем и неравенство Коши–Буняковского, последовательно получаем, что

$$\begin{gathered}
\mathrm{E}_{\mathcal{F}_r^{i-1}}\|\Delta^i\|^2\left(y^i_{r,+}-y^i_{r,-}\right)^2=2m\Big(\mathrm{E}_{\mathcal{F}_r^{i-1}}\left(v^i_{r,+}-v^i_{r,-}\right)^2+\\
+\mathrm{E}_{\mathcal{F}_r^{i-1}}\left(q\left(\widehat{\theta}_r^{(i-1)}+\beta^i\Delta^i,\Gamma_r,\mathbf{x}^i\right)-q\left(\widehat{\theta}_r^{(i-1)}-\beta^i\Delta_i,\Gamma_r,\mathbf{x}^i\right)\right)^2\Big)\leqslant\\
\leqslant 2m\mathrm{E}_{\mathcal{F}_r^{i-1}}(v^i_r)^2+2m\mathrm{E}_{\mathcal{F}_r^{i-1}}\Big(\left|q\left(\widehat{\theta}_r^{(i-1)}+\beta^i\Delta^i,\Gamma_r,\mathbf{x}^i\right)-q\left(\theta^\star_r,\Gamma^\star_r,\mathbf{x}^i\right)\right|+\\
+\left|q\left(\widehat{\theta}_r^{(i-1)}-\beta^i\Delta^i,\Gamma_r,\mathbf{x}^i\right)-q\left(\theta^\star_r,\Gamma^\star_r,\mathbf{x}^i\right)\right|\Big)^2\leqslant\\
\leqslant 2m\mathrm{E}_{\mathcal{F}_r^{i-1}}\Big((M+0{,}5)\left(\left\|\widehat{\theta}_r^{(i-1)}+\Delta^i\beta^i\right\|^2+\left\|\widehat{\theta}_r^{(i-1)}-\Delta^i\beta^i\right\|^2\right)+\\
+\left\|\nabla_\theta q\left(\theta^\star_r,\Gamma^\star_r,\mathbf{x}^i\right)\right\|^2\Big)^2+2m\mathrm{E}_{\mathcal{F}_r^{i-1}}(v^i_r)^2.
\end{gathered}$$

В силу равномерной ограниченности функции $\nabla_\theta q(\theta^\star_r,\Gamma^\star_r,\cdot)$ получаем неравенство

$$\mathrm{E}_{\mathcal{F}_r^{i-1}}\|\Delta^i\|^2\left(y^i_{r,+}-y^i_{r,-}\right)^2\leqslant C_1\mathrm{E}_{\mathcal{F}_r^{i-1}}(v^i_r)^2+C_2(\beta^i)^2\|\widehat{\theta}_r^{(i-1)}-\theta^\star_r\|^2+o((\beta^i)^2),$$

где $C_i,\ i=1,2\ldots,$ — некоторая положительная константа.

Рассмотрим второе слагаемое правой части (П.2):

$$\begin{gathered}
(\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)=(\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i v^i_r+(\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i\times\\
\times\left(q\left(\widehat{\theta}_r^{(i-1)}+\beta^i\Delta^i,\Gamma_r,\mathbf{x}^i\right)-q\left(\widehat{\theta}_r^{(i-1)}-\beta^i\Delta^i,\Gamma_r,\mathbf{x}^i\right)\right).
\end{gathered}\tag{П.3}$$

Первое слагаемое в правой части (П.3) равно нулю в силу свойств Δ^i и независимости Δ^i и $v^i_{r,\pm}$. Используя теорему о среднем, равномерную ограниченность функции $\nabla_\theta q(\theta,\Gamma,\cdot)$ и определение $F_r(\Theta,\Gamma)$, преобразуем второе слагаемое правой части (П.3) к виду

$$\begin{gathered}
(\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\left(\Delta^i\left(y^i_{r,+}-y^i_{r,-}\right)\right)=2\nabla F_r(\widehat{\theta}_r^{(i-1)})+\\
+\mathrm{E}_{\mathcal{F}_r^{i-1}}\left(\Delta^i\left\langle\Delta^i,\nabla_\theta q\left(\theta_r^{(i-1)^+},\Gamma_r,\mathbf{x}^i\right)-\nabla_\theta q\left(\widehat{\theta}_r^{(i-1)},\Gamma_r,\mathbf{x}^i\right)\right\rangle\right)+\\
+\mathrm{E}_{\mathcal{F}_r^{i-1}}\left(\Delta^i\left\langle\Delta^i,\nabla_\theta q\left(\theta_r^{(i-1)^-},\Gamma_r,\mathbf{x}^i\right)-\nabla_\theta q\left(\widehat{\theta}_r^{(i-1)},\Gamma_r,\mathbf{x}^i\right)\right\rangle\right),
\end{gathered}$$

здесь

$$\theta_r^{(i-1)^\pm}\in\left[\theta_r^{(i-1)},\theta_r^{(i-1)}\pm\beta^i\Delta^i\right].$$

Функция $q(\cdot,\Gamma,\mathbf{x})$ удовлетворяет условию Липшица и функция $F_r(\cdot,\Gamma)$ строго выпукла, тогда, используя справедливое при любом $\varepsilon>0$ неравенство

$\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\| \leqslant (\varepsilon^{-1}\beta^i + \varepsilon(\beta^i)^{-1}\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\|^2)/2$, выводим

$$-\alpha^i \left\langle \widehat{\theta}_r^{(i-1)} - \theta_r^\star, (\beta^i)^{-1}\mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i \left(y_{r,+}^i - y_{r,-}^i\right)\right\rangle =$$

$$= -2\alpha^i \left\langle \widehat{\theta}_r^{(i-1)} - \theta_r^\star, \nabla F_r\left(\widehat{\theta}_r^{(i-1)}, \Gamma_r\right)\right\rangle -$$

$$- \alpha^i \left\langle \widehat{\theta}_r^{(i-1)} - \theta_r^\star, \mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i \left\langle \Delta^i, \nabla_\theta q\left(\theta_r^{(i-1)^+}, \Gamma_r, \mathbf{x}^i\right) - \nabla_\theta q\left(\widehat{\theta}_r^{(i-1)}, \Gamma_r, \mathbf{x}^i\right)\right\rangle\right\rangle -$$

$$- \alpha^i \left\langle \widehat{\theta}_r^{(i-1)} - \theta_r^\star, \mathrm{E}_{\mathcal{F}_r^{i-1}}\Delta^i \left\langle \Delta^i, \nabla_\theta q\left(\theta_r^{(i-1)^-}, \Gamma_r, \mathbf{x}^i\right) - \nabla_\theta q\left(\widehat{\theta}_r^{(i-1)}, \Gamma_r, \mathbf{x}^i\right)\right\rangle\right\rangle \leqslant$$

$$\leqslant -2\mu\alpha^i \left\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\right\|^2 + Mm^{3/2}\alpha^i\beta^i \left(\varepsilon^{-1}\beta^i + \varepsilon(\beta^i)^{-1}\left\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\right\|^2\right).$$

Следовательно, имеем

$$\mathrm{E}_{\mathcal{F}_r^{i-1}}\left\|\widehat{\theta}_r^{\,i} - \theta_r^\star\right\|^2 \leqslant \left\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\right\|^2 \left(1 - 2\mu\alpha^i + Mm^{3/2}\alpha^i\varepsilon + C_2(\alpha^i)^2/4\right) +$$

$$+ Mm^{3/2}\alpha^i(\beta^i)^2\varepsilon^{-1} + (\alpha^i)^2/(4(\beta^i)^2)\left(C_1\mathrm{E}_{\mathcal{F}_r^{i-1}}{v_r^i}^2 + o((\beta^i)^2)\right).$$

Выберем ε настолько малым, чтобы $Mm^{3/2}\varepsilon < 2\mu$, и пусть i достаточно велико. По условиям теоремы получаем

$$\mathrm{E}_{\mathcal{F}_r^{i-1}}\left\|\widehat{\theta}_r^{\,i} - \theta_r^\star\right\|^2 \leqslant \left\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\right\|^2 \left(1 - C_3\alpha^i\right) +$$

$$+ C_4\left(\alpha^i(\beta^i)^2 + (\alpha^i)^2(\beta^i)^{-2}(1 + \mathrm{E}_{\mathcal{F}_r^{i-1}}v_r^i)^2\right).$$

Тогда, по лемме Роббинса–Сигмунда [1] $\widehat{\theta}_r^{\,i} \to \theta_r^\star$ при $i \to \infty$ с вероятностью единица

$$\mathrm{E}\left\{\|\widehat{\theta}_r^{\,i} - \theta_r^\star\|^2\right\} \leqslant$$

$$\leqslant \mathrm{E}\left\{\|\widehat{\theta}_r^{(i-1)} - \theta_r^\star\|^2\right\}(1 - C_5\alpha^n) + C_6\left(\alpha^n(\beta^n)^2 + (\alpha^n)^2(\beta^n)^{-2}(1 + \sigma^{n2})\right).$$

Сходимость в среднеквадратичном смысле к точке $\theta_r^\star$ последовательности оценок $\{\widehat{\theta}_r^{\,i}\}$ следует из [1].

2. Пусть теперь Θ фиксирована. Рассмотрим оценки $\{\widehat{\Gamma}_l{}^n = \{\hat{g}_{ij}^n\}\}$, $l \in$ $\in 1, \ldots, k$. Обозначим $\hat{s}_{ij}^n = \{(\widehat{\theta}_l - \mathbf{x}^n)(\widehat{\theta}_l - \mathbf{x}^n)^{\mathrm{T}}\}_{ij}$ — (i,j)-й элемент l-й матрицы разброса, тогда

$$\hat{g}_{ij}^n = \left(\lambda + \sum_{r=\lambda+1}^{n}\omega^r\hat{s}_{ij}^r\right)n^{-1}.$$

По условиям теоремы $\lambda n^{-1} \to 0$, $\omega^r \to 1$, $\hat{s}_{ij}^r n^{-1} \to \Gamma_l$ при $n \to \infty$. Таким образом, при $n \to \infty$ оценка $\widehat{\Gamma}^n$ является оценкой по методу максимизации правдоподобия, значит $\widehat{\Gamma}^n \xrightarrow{p} \Gamma^\star$.

Оценки по алгоритмам (10) и (11) строятся итеративно, т.е. на n-м шаге алгоритма сначала строится оценка $\widehat{\Theta}^{\,n}$ при фиксированном $\widehat{\Gamma}^{\,n-1}$, которая затем фиксируется и строится $\widehat{\Gamma}^{\,n}$. Таким образом, для каждого из методов (10) и (11) доказаны в пп. 1 и 2 соответствующие сходимости оценок, а функция $\mathbf{q}(\Theta, \Gamma, \mathbf{x})$ задается с помощью (4), отсюда следует результат теоремы. Это завершает доказательство теоремы.

СПИСОК ЛИТЕРАТУРЫ

1. *Поляк Б.Т.* Введение в оптимизацию. М.: Наука, 1983.
2. *Robbins H., Monro S.* A Stochastic Approximation Method // Ann. Math. Stat. 1951. P. 400–407.
3. *Kiefer J., Wolfowitz J.* Stochastic Estimation of the Maximum of a Regression Function // Ann. Math. Stat. 1952. V. 23. No. 3. P. 462–466.
4. *Blum J.R.* Multidimensional Stochastic Approximation Methods // Ann. Math. Stat. 1954. V. 25. No. 4. P. 737–744.
5. *Граничин О.Н.* Об одной стохастической рекуррентной процедуре при зависимых помехах в наблюдении, использующей на входе пробные возмущения // Вестн. Ленингр. ун-та. Сер. 1. 1989. №1 (4). С. 19–21.

   Granichin O.N. A Stochastic Recursive Procedure with Correlated Noises in the Observation, that Employs Trial Perturbations at the Input // Vestnik Leningrad University: Mathematics (Vestnik Leningradskogo Universita. Matematika). 1989. V. 22. No. 1. P. 27–31.
6. *Граничин О.Н.* Процедура стохастической аппроксимации с возмущением на входе // АиТ. 1992. № 2. С. 97–104.

   Granichin O.N. Procedure of Stochastic Approximation with Disturbances at the Input // Autom. Remote Control. 1992. V. 53. No. 2. P. 232–237.
7. *Поляк Б.Т., Цыбаков А.Б.* Оптимальные порядки точности поисковых алгоритмов стохастической аппроксимации // Проблемы передачи информации. 1990. Т. 26. С. 126–133.

   Polyak B.T., Tsybakov A.B. Optimal orders of accuracy for search algorithms of stochastic optimization // Prob. Inform. Transm+. 1990. V. 26. No. 2. P. 126–133.
8. *Растригин Л.А.* Статистические методы поиска. М.: Наука, 1968.
9. *Spall J.C.* Multivariate Stochastic Aproximation Using a Simultaneous Perturbation Gradient Aproximation // IEEE Trans. Autom. Control. 1992. V. 37. No. 3. P. 332–341.
10. *Граничин О.Н.* Поисковые алгоритмы стохастической аппроксимации с рандомизацией на входе // АиТ. 2015. № 5. С. 43–59.

    Granichin O.N. Stochastic Approximation Search Algorithms with Randomization at the Input // Autom. Remote Control. 2015. V. 76. No. 5. P. 762–775.
11. *Granichin O., Volkovich Z., Toledano-Kitai D.* Randomized Algorithms in Automatic Control and Data Mining. Springer, 2015.
12. *Lloyd S.* Least Squares Quantization in PCM // IEEE Trans. Inform. Theory. 1982. V. 28. No. 2. P. 129–136.
13. *Shindler M., Wong A., Meyerson A.* Fast and Accurate k-means For Large Datasets // Proc. 24th NIPS Conf. 2011.
14. *Sculley D.* Web Scale K-Means clustering // Proc. 19th WWW Conf. 2010.

15. *Katselis D., Beck C.L., van der Schaar M.* Ensemble Online Clustering through Decentralized Observations // Proc. 53rd IEEE CDC. 2014. P. 910–915.
16. *Kaufman L., Rousseeuw P.* Finding Groups in Data: An Introduction to Cluster Analysis'. N.Y.: John Wiley & Sons Inc., 1990.
17. *Граничин О.Н., Измакова О.А.* Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения // АиТ. 2005. № 8. С. 52–63.

    Granichin O.N., Izmakova O.A. A Randomized Stochastic Approximation Algorithm for Self-Learning // Autom. Remote Control. 2005. V. 66. No. 8. P. 1239–1248.
18. *Dempster A., Laird N., Rubin D.* Maximum Likelihood from Incomplete Data via the EM Algorithm // J. Royal Stat. Soc. Ser. B. 1977. V. 39. No. 1. P. 1–38.
19. *Bishop C.M.* Pattern Recognition and Machine Learning. Springer, 2006.
20. *Song M., Wang H.* Highly Efficient Incremental Estimation of GMM for Online Data Stream Clust // Proc. SPIE III. 2005. P. 174–184.
21. *Frey B.J., Dueck D.* Clustering by Passing Messages between Data Points // Science. 2007. No. 315 (5814). P. 972–976.
22. *Ester M., Kriegel H.P., Sander J., Xu X.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, 1996. P. 226–231.
23. *Huber P.J.* Robust Statistics. Wiley, 1981.
24. *Hubert L., Arabie P.* Comparing partitions // J. Classif. 1985. V. 2. No. 1. P. 193–218.
25. *Dahlin J., Wills A., Ninness B.* Sparse Bayesian ARX Models with Flexible Noise Distributions // Proc. 18th IFAC Sympos. on System Identification. 2018. P. 25–30.

Статья представлена к публикации членом редколлегии П.С. Щербаковым.

Поступила в редакцию 01.06.2017

После доработки 19.12.2018

Принята к публикации 07.02.2019