ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 504, с. 3–27

ПОДРОБНЫЕ ПУБЛИКАЦИИ

УДК 51-7

РАНДОМИЗАЦИЯ И ЭНТРОПИЯ В МАШИННОМ ОБУЧЕНИИ И ОБРАБОТКЕ ДАННЫХ

© 2022 г. Академик РАН Ю. С. Попков[1,*]

Поступило 18.02.2022 г.
После доработки 26.02.2022 г.
Принято к публикации 04.03.2022 г.

Сочетание концепции рандомизации с энтропийными критериями позволяет получать решения в условиях максимальной неопределенности, что оказывается весьма эффективным в задачах машинного обучения и обработки данных. Демонстрируется применение этого подхода для энтропийно-рандомизированного оценивания функций на основе данных, рандомизированного "жесткого" и "мягкого" машинного обучения, кластеризации объектов, редукции размерности матрицы данных. Рассматриваются некоторые приложения задачи классификации, прогнозирования электрической нагрузки энергетической системы, рандомизированной кластеризации биологических объектов.

Ключевые слова: энтропия, рандомизация, машинное обучение, обработка данных, параметризация моделей, оценки условно-максимальной энтропии, балансовые уравнения, классификация, кластеризация, генерация случайных ансамблей

DOI: 10.31857/S2686954322030079

1. ВВЕДЕНИЕ

Обработка данных и машинное обучение являются весьма тесно связанными научными и технологическими направлениями, в рамках которых получено огромное количество результатов, опубликованных в огромном количестве статей и монографий и доложенных на международных конференциях. Ссылки на наиболее цитируемые из них будут сделаны в соответствующих разделах данной работы.

Современная концепция машинного обучения базируется на моделях с обозначенными параметрами, значения которых оцениваются различными методами (в основном математической статистики) с использованием данных с определенными, но гипотетическими, свойствами, и соответствующим образом обработанными и форматированными.

Весьма существенная особенность задач машинного обучения и обработки данных состоит в том, что их решение происходит при наличии недостоверности, неполноты и ошибок в данных, а также при недостаточности знаний об обучаемом объекте, которые проявляются в неадекватности используемых моделей.

Как же преодолеть этот барьер и повысить достоверность, надежность результатов машинного обучения?!

В данной работе предлагается: **во-первых**, использовать *рандомизированные* параметризованные модели и оценивать в результате машинного обучения не значения параметров, а их функции плотности распределения вероятности (ПРВ); и **во-вторых**, использовать не произвольную рандомизацию, а оптимальную, гарантирующую получение наилучших функций ПРВ при максимальной неопределенности. Последнее свойство формулируется в терминах условной максимизации информационной энтропии с учетом имеющихся реальных данных.

Основу предлагаемых методов составляет энтропийно-рандомизированное оценивание функций ПРВ, относительно которого рассматриваются математическая модель, алгоритм и его асимптотическая эффективность. Этот метод оценивания используется в процедурах "жесткого" и "мягкого" рандомизированного машинного обучения и кластеризации объектов.

Одной из проблем обработки данных является редукция их размерности. Предлагается для этой цели использовать энтропийные проекции, которые реализуюся как в детерминированном, так и в рандомизированном алгоритме.

Последний раздел работы посвящен прикладным задачам и иллюстративным примерам. Рас-

[1] *Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия*
**E-mail: popkov@isa.ru*

смотрены задачи рандомизированной бинарной классификации с использованием стохастических нейронных сетей, прогнозирования суточной электрической нагрузки энергетической системы, рандомизированной кластеризации биологических объектов.

2. НЕОПРЕДЕЛЕННОСТЬ И РАНДОМИЗАЦИЯ

Подавляющее большинство задач машинного обучения и обработки данных сопровождаются неопределенностью, проявляющейся в ошибках, неполноте, пропусках в массивах данных, в неадекватности математических моделей исследуемому объекту, в отсутствии надежных знаний о процессах, происходящих в нем, в непредсказуемости окружающей среды.

Декларирование существования неопределенности влечет за собой попытки ее моделирование, хотя бы вербальные, а затем и ее измерение. Со времен Л. Больцмана *измерения* связывают со статистической *энтропией* [1, 2] и впоследствии с ее информационной интерпретацией [3]. Известны многочисленные модификации энтропийных функций, связанные с включением в них некоторых особенностей как равновесных состояний макросистем, так и процессов их достижения. Одной из таких модификаций является энтропия Реньи [4], которая обобщает энтропийные функции больцмановского-шеннонского типа [5].

Использование энтропийных функционалов подразумевает некую вероятностную имитацию неопределенного события, т.е. неявно принимается стохастическая модель неопределенности.

Если неопределенное событие интерпретируется как непрерывная переменная, то информационная энтропия определяется через ее функцию плотности распределения вероятностей $P(\mathbf{x})$:

$$\mathcal{H}[P(\mathbf{x})] = -\int_{\mathcal{X}} P(\mathbf{x})\ln P(\mathbf{x})d\mathbf{x}, \quad \mathbf{x} \in \mathcal{X}. \tag{2.1}$$

Если неопределенное событие принадлежит дискретному множеству, то информационная энтропия определяется через вектор $\mathbf{p}$, характеризующий дискретную функцию распределения вероятностей:

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i. \tag{2.2}$$

Поскольку энтропия есть мера неопределенности, то ее максимизация при дополнительных условиях дает наилучшую оценку принятой вероятностной характеристики при максимальной неопределенности [6–8].

Итак, принимается стохастическая концепция неопределенности. Естественным воплощением ее является *рандомизированная* модель, которая представляет собой генератор ансамбля случайных событий, описываемый функциональной характеристикой, максимизирующей соответствующий энтропийный функционал.

Рандомизация как метод “погружения” исследуемого события (объекта) в ансамбль случайных событий с последующим анализом его числовых характеристик использовался давно и в разных прикладных областях. Прежде всего следует указать на задачи, в которых нужно формировать представительные выборки, например, в клинических исследованиях [9], социальных опросах [10], формировании рейтингов [11] и усредненных сетевых графиков [12] и др.

Рандомизация оказывается полезной в задачах, связанных с предсказанием неких событий с указанием вероятности их наступления. При этом вероятность параметризуется, и данные используются для оценивния указанных параметров [13, 14]. Найденные оценки позволяют вычислить размеры доверительных интервалов. Если эти данные относятся к конкретному индивиду, то найденные таким способом доверительные интервалы квалифицируются как индивидуальные [15].

Идеи рандомизации оказались весьма продуктивными в задачах, где используются нейронные сети. При этом нейронная сеть стала рандомизированной, т.е. содержащей случайные параметры в слоях, в функциях активации, и в количестве слоев [16]. Для обучения такой сети применяется модифицированный алгоритм random forest.

Применение рандомизации при конструировании алгоритмов оказалось весьма эффективным для улучшения их вычислительных свойств. При выполнении некоторых операций запускались соответствующие случайные механизмы, которые при последовательном их выполнении приводили к решению поставленной задачи [17]. Довольно много работ на эту тему в области автоматического управления, где многие задачи управления сводятся к выпуклой и не выпуклой оптимизации. Применение рандомизированных алгоритмов позволяет найти либо точное их решение, либо решение с вероятностью [18–20]. И наконец, следует упомянуть пласт работ по теории игр, где переход к рандомизированным (смешанным) стратегиям позволяет получить решение минимаксной задачи [21].

В данной работе делается очередной шаг в направлении расширения области применения концепции рандомизации. Мы будем синтезировать *рандомизированную модель* как математическую модель со случайными параметрами. В зависимости от структуры модели случайные параметры характеризуются либо функциями ПРВ, либо функциями РВ, которые определяются с учетом реальных данных и неопределенности путем максимизации энтропийных функционалов.

Энтропийно-оптимальные функции ПРВ или РВ сэмплируются, т.е. трансформируются в соответствующие последовательности случайных чисел, и генерируется реальный ансамбль выходов рандомизированной модели. Применяя стандартные методы математической статистики, вычисляются эмпирические числовые характеристики ансамбля: среднее, дисперсию, доверительные интервалы, квантили и др. В результате могут обнаружиться неожиданные свойства рандомизированной модели.

Приведем пример. Рассмотрим модель динамического объекта первого порядка.

$$\frac{dx(t)}{dt} = ax(t), \quad x(0) = x_0 > 0.$$

Решение этого уравнения

$$x(t) = x_0 \exp(at).$$

В зависимости от знака a траектория либо убывает, либо возрастает при $t \to \infty$.

Рассмотрим рандомизированную модель, в которой параметр a – случайный с функцией ПРВ

$$P(a) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(a-\overline{a})^2}{2\sigma^2}\right),$$

где $\overline{a}$ – среднее значение и σ^2 – дисперсия параметра a.

Рандомизированная модель генерирует ансамбль траекторий, средняя траектория в котором имеет вид:

$$\overline{x}(t) = x_0 \int_{-\infty}^{\infty} \exp(at)P(a)da = x_0 \exp\left(\overline{a}t + \frac{\sigma^2 t^2}{2}\right).$$

Отсюда следует, что при $\overline{a} \le 0$ траектория имеет минимум в точке $t^* = \overline{a}/\sigma^2$, что не реализуется в модели с неслучайным параметром a. Рисунок 1 иллюстрирует описанную ситуацию.

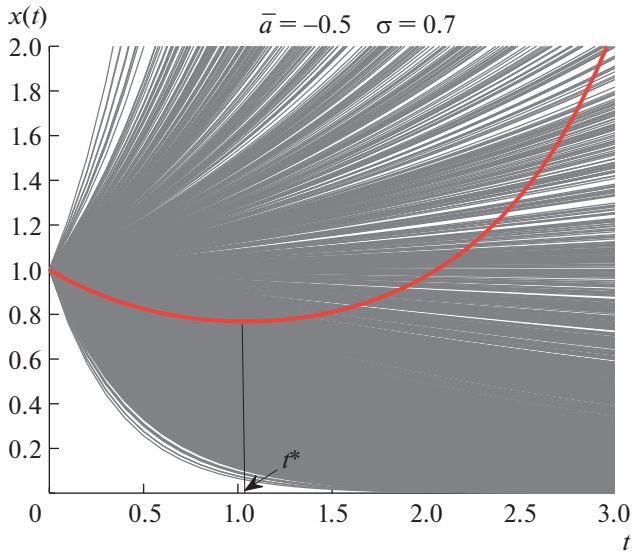

Рис. 1

3. ЭНТРОПИЙНО-РАНДОМИЗИРОВАННОЕ ОЦЕНИВАНИЕ (ЭРО) ФУНКЦИЙ ПРВ

3.1. Введение

Существующие методы оценивания функций ПРВ (максимальное правдоподобие, метод моментов, метод наименьших квадратов и др.) никак не учитывают неопределенность, сопровождаемую эти задачи, и требуют задания формы функции и ее параметризацию. Кроме того, приходится принимать весьма обременительные и непроверяемые гипотезы о свойствах данных как выборки из генеральной совокупности [22, 23].

3.2. Математическая формулировка метода ЭРО

Рассмотрим скалярную аналитическую функцию $\hat{y} = \varphi(x, \boldsymbol{\theta})$ с рандомизированными параметрами $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ интервального типа, т.е.

$$\boldsymbol{\theta} \in \mathscr{E} = [\boldsymbol{\theta}^-, \boldsymbol{\theta}^+]. \tag{3.1}$$

Вероятностные свойства параметров характеризуются функцией плотности распределения вероятностей (ПРВ) $P(\boldsymbol{\theta})$, определенной на множестве $\mathscr{E}$.

Поскольку параметры рандомизированные, то и переменная $\hat{y}$ – рандомизирована и принимает значения в множестве $\hat{\mathcal{Y}}$, размеры которого определяются функцией $\varphi(x, \boldsymbol{\theta})$ и вероятностными свойствами параметров.

Пусть имеются r измерений $\{x_1, \ldots, x_r\} = \mathbf{x}^{(r)}$, и $\{y_1, \ldots, y_r\} = \mathbf{y}^{(r)}$. В результате имеем следующую систему уравнений:

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta}), \tag{3.2}$$

где вектор-функция $\boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta})$ имеет компоненты $\varphi(x_t, \boldsymbol{\theta}), t = \overline{1, r}$.

Вектор $\hat{\mathbf{y}}$ (3.2) – рандомизированный. Рассмотрим в качестве его числовых характеристик r-мерный вектор с компонентами в виде нормированных моментов s-й степени:

$$\mathbf{m}^{(s)}[P(\boldsymbol{\theta})] = \left(\int_{\mathscr{E}} P(\boldsymbol{\theta})\boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})d\boldsymbol{\theta}\right)^{\frac{1}{s}}. \tag{3.3}$$

Заметим, что (3.3) является векторным функционалом от функции ПРВ $P(\boldsymbol{\theta})$.

Измеренные данные в виде вектора $\mathbf{y}^{(r)}$, будем приравнивать векторам нормированных моментов[1]:

$$\mathbf{m}^{(s)}[P(\boldsymbol{\theta})] = \mathbf{y}^{(r)}. \tag{3.4}$$

Из этих выражений следует, что совпадение моментных характеристик ансамбля рандомизированных переменных зависит от функции ПРВ $P(\boldsymbol{\theta})$.

Таким образом, *задача оценивания функции ПРВ параметров* формулируется следующим образом [25]:

$$\mathcal{H}[P(\boldsymbol{\theta})] = -\int\limits_{\mathscr{E}} P(\boldsymbol{\theta}) \ln P(\boldsymbol{\theta}) d\boldsymbol{\theta} \Rightarrow \max_{P(\boldsymbol{\theta})} \tag{3.5}$$

при ограничениях:

– нормировки функций ПРВ

$$\int\limits_{\mathscr{E}} P(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1; \tag{3.6}$$

– эмпирических балансов

$$\mathbf{m}^{(s)}[P(\boldsymbol{\theta})] = \mathbf{y}^{(r)}. \tag{3.7}$$

Задача (3.5)–(3.7) относится к классу ляпуновских [26, 27], которые характеризуются тем, что функционалы и ограничения интегрального типа и выпуклые.

3.3. Условия оптимальности

Условия оптимальности в задачах оптимизации ляпуновского типа формулируются в терминах вещественных множителей Лагранжа. При этом используются производные Гато интегральных функционалов [29].

Для задачи (3.5)–(3.7) функционал Лагранжа имеет вид:

$$\begin{aligned}\mathcal{L}[P(\boldsymbol{\theta}), \boldsymbol{\lambda}] = \mathcal{H}[P(\boldsymbol{\theta})] + \mu\left(1 - \int\limits_{\mathscr{E}} P(\boldsymbol{\theta}) d\boldsymbol{\theta}\right) + \\ + \langle \boldsymbol{\lambda}, (\mathbf{y}^{(r)} - \mathbf{m}^{(s)}[P(\boldsymbol{\theta})])\rangle.\end{aligned} \tag{3.8}$$

Технике получения условий оптимальности в терминах производной Гато посвящено Приложение А.

Используя условия оптимальности (1.1), получим оптимальную функцию ПРВ, параметризованную множителями Лагранжа $\boldsymbol{\theta}$:

$$\begin{aligned}P^*(\boldsymbol{\theta}|\boldsymbol{\lambda}) &= \frac{\exp(\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})\rangle)}{\mathcal{P}(\boldsymbol{\lambda})}, \\ \mathcal{P}(\boldsymbol{\lambda}) &= \int\limits_{\mathscr{E}} \exp(\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})\rangle) d\boldsymbol{\theta}.\end{aligned} \tag{3.9}$$

Из равенств (3.9) видно, что энтропийно-оптимальная функция ПРВ параметризована множителями Лагранжа $\boldsymbol{\lambda}$, которые определяются решением уравнений *эмпирических балансов*:

$$\mathcal{P}^{-1}(\boldsymbol{\lambda}) \int\limits_{\mathscr{E}} \boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta}) \exp(\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})\rangle) = \mathbf{y}^{(r)}. \tag{3.10}$$

Решение этих уравнений – неявная функция $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*(\mathbf{y}^{(r)}), \mathbf{x}^{(r)})$ зависит от измеренных данных ($\mathbf{y}^{(r)}$, $\mathbf{x}^{(r)}$), по которым строятся ЭРО функции ПРВ.

Важным частным случаем ЭРО является балансирование с данными средних характеристик ансамбля (3.3):

$$\mathbf{m}^{(1)}[P(\boldsymbol{\theta})] = \int\limits_{\mathscr{E}} P(\boldsymbol{\theta}) \boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3.11}$$

В этом случае в формулах (3.8)–(3.10) $s = 1$.

3.4. Существование неявной функции $\boldsymbol{\lambda}(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})$

Рассмотрим балансовые уравнения для средних числовых характеристик ансамбля (3.3):

$$\begin{aligned}\int\limits_{\mathscr{E}} (\boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta}) - \mathbf{y}^{(r)}) \exp(\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta})\rangle) d\boldsymbol{\theta} = \\ = \mathbf{W}(\boldsymbol{\lambda}|\mathbf{x}^r, \mathbf{y}^r) = \mathbf{0},\end{aligned} \tag{3.12}$$

где вектор-функция $\boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta}) = \{\varphi(x_1, \boldsymbol{\theta}), ..., \varphi(x_r, \boldsymbol{\theta})\}$.

Якобиан функции $\mathbf{W}$ имеет вид:

$$J_{\lambda}(\boldsymbol{\lambda}|\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) = \left[\frac{\partial W_t}{\partial \lambda_i}, |(t,i) = \overline{1,r}\right], \tag{3.13}$$

где элементы этой матрицы

$$\begin{aligned}\frac{\partial W_t}{\partial \lambda_i} = \int\limits_{\mathscr{E}} (\varphi(x_t, \boldsymbol{\theta}) - y_t) \varphi(x_i, \boldsymbol{\theta}) \times \\ \times \sum_{j=1}^{r} \exp\left(-\sum_{j=1}^{r} \lambda_j \varphi(x_j, \boldsymbol{\theta})\right) d\boldsymbol{\theta}.\end{aligned} \tag{3.14}$$

Теорема 1. *Пусть:*

• a) *функция* $\boldsymbol{\Phi}(\mathbf{x}^{(r)}, \boldsymbol{\theta})$ *непрерывна по совокупности переменных*;

• b) *для любых* $(\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) \in R^r \times R^r$ *выполняются следующие условия*

$$\det J_{\lambda}(\boldsymbol{\lambda}|\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) \neq 0, \tag{3.15}$$

$$\lim_{\|\boldsymbol{\lambda}\| \to \infty} \mathbf{W}(\boldsymbol{\lambda}|\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) = \pm\infty. \tag{3.16}$$

Тогда существует единственная неявная функция $\boldsymbol{\lambda}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$, *определенная на* $R^r \times R^r$.

Доказательство этой теоремы приведено в Приложении *B*.

[1] В некоторых задачах энтропийного оценивания, в частности, в задачах ценообразования активов на финансовом рынке используются нормированные моменты в качестве характеристики качества финансовых инструментов [24].

Теорема 2. *Пусть выполнены условия теоремы* 1. *Тогда функция* $\lambda(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})$ *– аналитическая по совокупности перменных*.

Доказательство теоремы 2 приведено в Приложении B.

3.5. Асимптотика ЭРО

ЭРО дает энтропийно оптимальную ПРВ (3.10), (3.11) для наборов данных $\mathbf{x}^{(r)}$, $\mathbf{y}^{(r)}$ объемами r каждый.

Далее удобнее оперировать функциями ПРВ, параметризованными экспоненциальными множителями Лагранжа $\mathbf{z} = \exp(-\boldsymbol{\lambda})$. Тогда равенство (3.10) примет следующий вид:

$$P^*(\boldsymbol{\theta}, \mathbf{z}) = \frac{\prod_{j=1}^{r} [z_j]^{\varphi(x_j, \boldsymbol{\theta})}}{\mathcal{P}(\mathbf{z})}, \qquad (3.17)$$

$$\mathcal{P}(\mathbf{z}) = \int_{\mathscr{E}} \prod_{j=1}^{r} [z_j]^{\varphi(x_j, \boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Отсюда видно, что структура функции ПРВ зависит от значений экспоненциальных множителей Лагранжа $\mathbf{z}$, которые, в свою очередь, зависят от от коллекции данных $\mathbf{x}^{(r)}$, $\mathbf{y}^{(r)}$.

Определение. *Будем называть оценку функции ПРВ* $P^*(\boldsymbol{\theta}, \mathbf{z}^*)$ *асимптотически устойчивой*, *если*

$$\lim_{r\to\infty} P^*(\theta, \mathbf{z}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})) = P^*(\theta, \tilde{\mathbf{z}}), \qquad (3.18)$$

где

$$\tilde{\mathbf{z}} = \lim_{r\to\infty} \mathbf{z}(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}). \qquad (3.19)$$

Рассмотрим уравнения эмпирических балансов (3.12), перейдя в них к экспоненциальным множителям Лагранжа:

$$\Theta_t(\mathbf{z}, \mathbf{x}^{(r)}, \mathbf{y}^{r})) =$$

$$= \int_{\mathscr{E}} \prod_{j=1}^{r} [z_j(\mathbf{x}^{(r)}, \mathbf{y}^{r})]^{\varphi(x_j, \boldsymbol{\theta})} (\varphi(x_t, \boldsymbol{\theta}) - y_t)\, d\boldsymbol{\theta} = 0, \quad (3.20)$$

$$t = \overline{1, r}.$$

В предыдущем разделе было показано, что уравнения (3.12) определяют неявную аналитическую функцию $\boldsymbol{\theta}(\mathbf{x}^{r}), \mathbf{y}^{(r)})$.

В силу связи множителей и экспоненциальных множителей Лагранжа уравнения (3.20) определяют неявную аналитическую функцию $\mathbf{z} = \mathbf{z}(\mathbf{x}^{r})$, $\mathbf{y}^{(r)})$ для $(\mathbf{x}^{r}, \mathbf{y}^{(r)}) \in R^r \times R^r$.

Дифференцируем левую и правую часть этих уравнений по $\mathbf{x}^{(r)}$ и $\mathbf{y}^{(r)}$. Получим следующие уравнения:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}} = -\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1} \frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{x}^{(r)}}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{y}^{(r)}} = -\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1} \frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{y}^{(r)}}. \qquad (3.21)$$

Все матрицы в этих уравнениях квадратные, размера $(r \times r)$.

Переходя к нормам, получим следующие неравенства:

$$0 \le \left\|\frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}}\right\| \le \left\|\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{x}^{(r)}}\right\|,$$
$$0 \le \left\|\frac{\partial \mathbf{z}}{\partial \mathbf{y}^{(r)}}\right\| \le \left\|\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{y}^{(r)}}\right\|. \qquad (3.22)$$

В оба эти неравенства входит норма обратной матрицы $\left\|\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\|$.

Лемма 1. *Пусть квадратная матрица* A *– невырождена*, *т.е.* $\det A \ne 0$. *Тогда существует константа* $\alpha > 1$, *такая*, *что*

$$\frac{1}{\|A\|} \le \left\|A^{-1}\right\| \le \frac{\alpha}{\|A\|}. \qquad (3.23)$$

Доказательство леммы 1 приведено в Приложении C.

Применим неравенство (3.23) к норме обратной матрицы $\left\|\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\|$. Получим следующее неравенство:

$$\left(\left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right\|\right)^{-1} \le \left\|\left[\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \le \alpha \left(\left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right\|\right)^{-1}. \qquad (3.24)$$

В этих неравенствах (см. [28])

$$\left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{z}}\right\| = r \max_{t,j} \left|\frac{\partial \theta_t}{\partial z_j}\right| = r\theta. \qquad (3.25)$$

Лемма 2. *Пусть*

$$\left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{x}^{(r)}}\right\| \le \rho < \infty, \quad \left\|\frac{\partial \boldsymbol{\Theta}}{\partial \mathbf{y}^{(r)}}\right\| \le \omega < \infty. \qquad (3.26)$$

Тогда оценка функции ПРВ $P^*(\boldsymbol{\theta}, \mathbf{z}^*)$ *асимптотически устойчива в смысле* (3.18), (3.19)

Доказательство леммы 2 приведено в Приложении D.

4. РАНДОМИЗИРОВАННОЕ МАШИННОЕ ОБУЧЕНИЕ (РМО)

4.1. Введение

Машинное обучение является одним из трендовых направлений современной науки и технологий. Количество публикаций, посвященных машинному обучению, превышает десятки тысяч и продолжает расти. Основное их количество связано с разнообразными приложениями, особен-

ности которых порождают и новые задачи исследовательского характера [32, 34–36]. Несмотря на разнообразие работ по машинному обучению, их фундаментальная основа строится на методах математической статистики, а точнее, на методах оценивания вещественных параметров моделей.

Рандомизированное машинное обучение использует принципиально иной подход, связанный с оцениванием функций плотности распределения вероятностей с неизвестной структурой и параметрами. Развиваются методы оценивания функций ПРВ, основанные на условной (с учетом имеющихся данных) максимизации энтропии [37].

4.2. Постановка задачи

Рассмотрим РМО-процедуру применительно к задаче восстановления зависимостей. Предполагается, что имеются данные о входе $\mathbf{x}^{(r)}[k] \in R^n$ и выходе $\mathbf{y}^{(r)}[k] \in R^m$ объекта на временном интервале $\mathcal{T} = [0,T]$, где $k = \overline{0,T}$. На интервале наблюдения формируются блочные векторы входных и выходных данных:

$$\begin{aligned} \mathbf{X}^{(r)} &= \{\mathbf{x}^{(r)}[0],\ldots,\mathbf{x}^{(r)}[T]\}, \\ \mathbf{Y}^{(r)} &= \{\mathbf{y}^{(r)}[0],\ldots,\mathbf{y}^{(r)}[T]\}, \end{aligned} \tag{4.1}$$

размерности $n(T+1)$ и $m(T+1)$ соответственно.

Они сопровождаются случайными и независимыми измерительными шумами (ошибками) с заданными областями их значений (интервалами):

$$\begin{aligned} \boldsymbol{\eta}[k] \in \mathscr{E}_k &= [\boldsymbol{\eta}[k]^-, \boldsymbol{\eta}^+[k]], \\ \boldsymbol{\xi}[k] \in \mathscr{H}_k &= [\boldsymbol{\xi}[k]^-, \boldsymbol{\xi}^+[k]], \quad k = \overline{0,T}. \end{aligned} \tag{4.2}$$

Измерительные шумы на интервале $\mathcal{T}$ будем так же характеризовать соответствующими блочными векторами:

$$\begin{aligned} &\boldsymbol{\eta} \to \mathbf{E} = \{\boldsymbol{\eta}[0],\ldots,\boldsymbol{\eta}[T]\}^{\mathsf{T}} - n(T+1) - \text{вектор}, \\ &\boldsymbol{\xi} \to \mathbf{K} = \{\boldsymbol{\xi}[0],\ldots,\boldsymbol{\xi}[T]\}^{\mathsf{T}} - m(T+1) - \text{вектор}. \end{aligned} \tag{4.3}$$

Согласно (4.2) области значений этих векторов имеют вид:

$$\mathbf{E} \in \mathscr{E} = \bigcup_{k=0}^{T} \mathscr{E}_k, \quad \mathbf{K} \in \mathscr{H} = \bigcup_{k=0}^{T} \mathscr{H}_k. \tag{4.4}$$

Наблюдаемые входные $\hat{\mathbf{x}}[k]$ и выходные данные $\mathbf{v}[k]$ предположительно аддитивно связаны с измерительными шумами:

$$\begin{aligned} \hat{\mathbf{x}}[k] = \mathbf{x}^{(r)}[k] + \boldsymbol{\eta}[k], \quad \mathbf{v}[k] = \hat{\mathbf{y}}[k] + \boldsymbol{\xi}[k], \\ k = \overline{0,T}, \end{aligned} \tag{4.5}$$

где $\hat{\mathbf{y}}[k]$ – выход рандомизированной модели исследуемого объекта:

$$\hat{\mathbf{y}}[k] = \mathbf{f}(\mathbf{x}^{(r)}[k] + \boldsymbol{\eta}[k] | \mathbf{a}), \tag{4.6}$$

где: $\mathbf{f} \in R^m$, $\mathbf{a} \in R^r$ – вектор рандомизированных параметров интервального типа:

$$\mathbf{a} \in \mathscr{A} = [\mathbf{a}^-, \mathbf{a}^+]. \tag{4.7}$$

На интервале наблюдения $\mathcal{T}$ будем иметь:

$$\begin{aligned} \mathbf{V} &= \hat{\mathbf{Y}} + \mathscr{K}, \\ \hat{\mathbf{Y}} &= \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}), \end{aligned} \tag{4.8}$$

где

$$\begin{aligned} \hat{\mathbf{Y}} &= \{\hat{\mathbf{y}}[0],\ldots,\hat{\mathbf{y}}[T]\}^{\top} - \\ &- n(T+1) - \text{блочный вектор}, \\ \mathbf{F} &= \{\mathbf{f}(\mathbf{x}^{(r)}[0] + \boldsymbol{\eta}0]),\ldots,\mathbf{f}(\mathbf{x}^{(r)}[T] + \boldsymbol{\eta}[T])\}^{\top}. \end{aligned} \tag{4.9}$$

Поскольку параметры и измерительные ошибки являются случайными объектами, введем для их характеризации

• совместную функцию ПРВ $W(\mathbf{a},\mathbf{E})$, определенную на множестве

$$\mathscr{W} = \mathscr{A} \cup \mathscr{E}, \quad (\mathbf{a},\mathbf{E}) \in \mathscr{W}, \tag{4.10}$$

где множество $\mathscr{E}$ определено в (4.4);

• функцию ПРВ $Q(\mathbf{K})$, определенную на множестве $\mathscr{K}$ (4.4).

Таким образом, используя наборы реальных данных $\mathbf{X}^{(r)}$ и $\mathbf{Y}^{(r)}$, определить функции ПРВ $W(\mathbf{a},\mathbf{E})$ и $Q(\mathbf{K})$.

4.3. Алгоритм "жесткого" РМО для восстановления функций ПРВ параметров и шумов

Алгоритм РМО для случая 1-моментных эмпирических балансов ("жесткое" РМО) имеет вид:

$$\begin{aligned} &\mathscr{H}[W(\mathbf{a},\mathbf{E}), Q(\mathbf{K})] = \\ &= -\int_{\mathscr{W}} W(\mathbf{a},\mathbf{E}) \ln W(\mathbf{a},\mathbf{E}) d\mathbf{a} d\mathbf{E} - \\ &- \int_{\mathscr{K}} Q(\mathbf{K}) \ln Q(\mathbf{K}) d\mathbf{K} \Rightarrow \max, \end{aligned} \tag{4.11}$$

при ограничениях:

– нормировки функций ПРВ

$$\int_{\mathscr{W}} W(\mathbf{a},\mathbf{E}) d\mathbf{a} d\mathbf{E} = 1, \quad \int_{\mathscr{K}} Q(\mathbf{K}) d\mathbf{K} = 1; \tag{4.12}$$

– эмпирических балансов 1-й степени

$$\begin{aligned} &\int_{\mathscr{W}} W(\mathbf{a},\mathbf{E}) \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E} | \mathbf{a}) \otimes d\mathbf{a} \otimes d\mathbf{E} + \\ &+ \int_{\mathscr{K}} Q(\mathbf{K}) \mathbf{K} \otimes d\mathbf{K} = \mathbf{Y}^{(r)}, \end{aligned} \tag{4.13}$$

где $\otimes$ – знак покоординатного умножения векторов.

Применяя технику формирования условий оптимальности с помощью производных Гато (см.

Приложение В), получим следующие выражения для энтропийно-оптимальных функций ПРВ параметров и измерительных шумов, параметризованных множителями Лагранжа $\mathbf{\Lambda}$ и зависящих от входных $\mathbf{X}^{(r)}$ и выходных $\mathbf{Y}^{(r)}$ данных:

$$\begin{gathered} W^*(\mathbf{a},\mathbf{E}|\mathbf{\Lambda},\mathbf{X}^{(r)}) = \\ = \mathbf{W}^{-1}(\mathbf{\Lambda},\mathbf{X}^{(r)}) \otimes \overline{\exp}(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})\rangle), \\ Q^*(\mathbf{K}|\mathbf{\Lambda}) = \mathbf{Q}^{-1}(\mathbf{\Lambda}) \otimes \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big), \end{gathered} \tag{4.14}$$

где векторы

$$\begin{gathered} \mathbf{W}(\mathbf{\Lambda},\mathbf{X}^{(r)}) = \int\limits_{\mathscr{W}} \overline{\exp}(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})\rangle) \otimes d\mathbf{a} \otimes d\mathbf{E}, \\ \mathbf{Q}(\mathbf{\Lambda}) = \int\limits_{\mathscr{K}} \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big) \otimes d\mathbf{K}. \end{gathered} \tag{4.15}$$

В этих равенствах $\overline{\exp}$ – вектор с компонентами $\exp(\bullet)$.

Множители Лагранжа $\mathbf{\Lambda}$ определяются следующими уравнениями:

$$\begin{gathered} \mathbf{R}(\mathbf{\Lambda},\mathbf{X}^{(r)}) \otimes \mathbf{Q}(\mathbf{\Lambda}) + \mathbf{G}(\mathbf{\Lambda}) \otimes \mathbf{W}(\mathbf{\Lambda},\mathbf{X}^{(r)}) = \\ = \mathbf{Y}^{(r)} \otimes \mathbf{Q}(\mathbf{\Lambda}) \otimes \mathbf{W}(\mathbf{\Lambda},\mathbf{X}^{(r)}), \end{gathered} \tag{4.16}$$

где

$$\begin{gathered} \mathbf{R}(\mathbf{\Lambda},\mathbf{X}^{(r)}) = \int\limits_{\mathscr{W}} \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a}) \otimes \\ \otimes \overline{\exp}(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})\rangle) \otimes d\mathbf{a} \otimes d\mathbf{E}, \\ \mathbf{G}(\mathbf{\Lambda}) = \int\limits_{\mathscr{K}} \mathbf{K} \otimes \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big) \otimes d\mathbf{K}. \end{gathered} \tag{4.17}$$

Из приведенных выражений следует, что оценки функций ПРВ параметров и измерительных шумов, соответствующие максимальной (в единицах информационной энтропии) неопределенности, зависят не только от имеющихся данных о входе и выходе исследуемого объекта, но и от модели объекта. Причем в общем случае нелинейной модели вероятностные свойства рандомизированных параметров и измерительных шумов на входе характеризуются оптимальной оценкой совместной функции ПРВ. Последнее свидетельствует о связи вероятностных свойств энтропийно-рандомизированных параметров и входных измерительных шумов.

Если предполагается, что входные данные измеряются точно, т.е. $\mathbf{\eta} = 0$, процедура формирования оценок функций ПРВ упрощается. В этом случае имеем:

$$\hat{\mathbf{y}}[k] = \mathbf{f}(\mathbf{x}^{(r)}[k]\|\mathbf{a}). \tag{4.18}$$

В принятых выше обозначениях функции ПРВ параметров и шумов приобретают следующий вид:

$$\begin{gathered} W^*(\mathbf{a}|\mathbf{\Lambda},\mathbf{X}^{(r)}) = \mathbf{W}^{-1}(\mathbf{\Lambda},\mathbf{X}^{(r)}) \otimes \overline{\exp}\Big(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)}|\mathbf{a})\rangle\Big), \\ Q^*(\mathbf{K}|\mathbf{\Lambda}) = \mathbf{Q}^{-1}(\mathbf{\Lambda}) \otimes \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big), \end{gathered} \tag{4.19}$$

где векторы

$$\begin{gathered} \mathbf{W}(\mathbf{\Lambda},\mathbf{X}^{(r)}) = \int\limits_{\mathscr{W}} \overline{\exp}(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)}|\mathbf{a})\rangle) \otimes d\mathbf{a}, \\ \mathbf{Q}(\mathbf{\Lambda}) = \int\limits_{\mathscr{K}} \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big) \otimes d\mathbf{K}. \end{gathered} \tag{4.20}$$

Множители Лагранжа определяются из уравнения (4.16), где

$$\begin{gathered} \mathbf{R}(\mathbf{\Lambda},\mathbf{X}^{(r)}) = \\ = \int\limits_{\mathscr{W}} \mathbf{F}(\mathbf{X}^{(r)}|\mathbf{a}) \otimes \overline{\exp}(\langle \mathbf{\Lambda}, \mathbf{F}(\mathbf{X}^{(r)}|\mathbf{a})\rangle) \otimes d\mathbf{a}, \\ \mathbf{G}(\mathbf{\Lambda}) = \int\limits_{\mathscr{K}} \mathbf{K} \otimes \overline{\exp}\big(\langle \mathbf{\Lambda}, \mathbf{K}\rangle\big) \otimes d\mathbf{K}. \end{gathered} \tag{4.21}$$

4.4. Алгоритм "мягкого" РМО для восстановления функций ПРВ

В некоторых задачах РМО не требуется строгое выполнение балансов между числовыми характеристиками ансамбля выхода модели и данными. Учитывая (4.5), (4.6), воспользуемся евклидовым расстоянием между наблюдаемым выходом модели и данными в виде:

$$\begin{gathered} \rho(\mathbf{V},\mathbf{Y}^{(r)}) = \|\mathbf{V} - \mathbf{Y}^{(r)}\|_E = \\ = \|\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a}) + \mathbf{K} - \mathbf{Y}^{(r)}\|_E \le \\ \le \|\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})\|_E + \|\mathbf{K} - \mathbf{Y}^{(r)}\|_E = \\ = \varrho_1(\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})) + \varrho_2(\mathbf{K} - \mathbf{Y}^{(r)}) = \overline{\varrho}(\mathbf{a},\mathbf{E},\mathbf{K}). \end{gathered} \tag{4.22}$$

Из этих равенств следует, что верхняя граница расстояния $\overline{\varrho}$ является функцией случайных параметров $\mathbf{a}$ и измерительных шумов $\mathbf{E}$, $\mathbf{K}$.

Определим функционал $\mathscr{R}$, являющийся средним функции $\overline{\varrho}$:

$$\begin{gathered} \mathscr{R}[W(\mathbf{a},\mathbf{E}),Q(\mathbf{K})] = \\ = \int\limits_{\mathscr{W}\times\mathscr{E}} W(\mathbf{a},\mathbf{E})\varrho_1(\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a}) \otimes d\mathbf{a} \otimes d\mathbf{E} + \\ + \int\limits_{\mathscr{K}} Q(\mathbf{K})\varrho_2(\mathbf{K} - \mathbf{Y}^{(r)}) \otimes d\mathbf{K}. \end{gathered} \tag{4.23}$$

Алгоритм "мягкого" РМО имеет вид:

$$\begin{gathered} \mathscr{J}[W(\mathbf{a},\mathbf{E}),Q(\mathbf{K})] = -\mathscr{H}[W(\mathbf{a},\mathbf{E}),Q(\mathbf{K})] - \\ - \mathscr{R}[W(\mathbf{a},\mathbf{E}),Q(\mathbf{K})] \Rightarrow \max, \end{gathered} \tag{4.24}$$

при ограничениях нормировки ПРВ:

$$\int\limits_{\mathscr{W}\times\mathscr{E}} W(\mathbf{a},\mathbf{E})d\mathbf{a} \otimes d\mathbf{E} = 1, \quad \int\limits_{\mathscr{K}} Q(\mathbf{K})d\mathbf{K} = 1. \tag{4.25}$$

Задача оптимизации (4.24), (4.25) также ляпуновского типа, и ее решение имеет вид:

$$W^*(\mathbf{a},\mathbf{E}) = \mathbb{W}^{-1}\exp(\varrho_1(\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})),$$
$$Q^*(\mathbf{K}) = \mathbb{Q}^{-1}\exp(\varrho_2(\mathbf{K} - \mathbf{Y}^{(r)})), \quad (4.26)$$

где нормировочные константы

$$\mathbb{W} = \int\limits_{\mathcal{W}\times\mathcal{E}} \exp(\varrho_1(\mathbf{F}(\mathbf{X}^{(r)} + \mathbf{E}|\mathbf{a})) \otimes d\mathbf{a} \otimes d\mathbf{E},$$
$$\mathbb{Q} = \int\limits_{\mathcal{K}} \exp(\varrho_2(\mathbf{K} - \mathbf{Y}^{(r)})) \otimes d\mathbf{K}. \quad (4.27)$$

Алгоритм "мягкого" РМО позволяет получать аналитические выражения для энтропийно-оптимальных ПРВ параметров и шумов, не требующие решения балансовых уравнений. Из (4.26) следует, что эти ПРВ экспоненциального класса, но их морфология определяется не только структурой математической модели, но и принятыми векторными нормами.

5. КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ НА ОСНОВЕ ЭРО

Метод энтропийно-рандомизированного оценивания, изложенный в первом разделе, позволяет восстанавливать функции ПРВ случайных параметров. До сих пор рассматривались непрерывно-дифференцируемые функции ПРВ. Однако существует класс задач, где присутствуют дискретные рандомизированные объекты. Ансамбли таких объектов характеризуются дискретными функциями распределения вероятностей, значения которых принадлежат интервалу [0, 1], или дискретными функциональными формами распределения вероятностей, значения которых принадлежат неотрицательному интервалу $[0,\infty)$. Одними из таковых являются задачи кластеризации.

5.1. Введение

Кластеризация объектов различной природы является одним из направлений машинного обучения, в котором метки учителя заменяются какими-то внутренними характеристиками объектов или внешними характеристиками кластеров. К внутренним относятся расстояния между объектами внутри кластера [50, 51], характристики сходства объектов [52], а к внешним – расстояния между кластерами [53]. Как математическая задача, кластеризация не имеет универсальной формулировки, и поэтому алгоритмы кластеризиции носят, как правило, эвристический характер [54, 55].

Весьма развитое направление связано с кластеризацией больших массивов текстов. Ему обычно предшествуют процедуры выявления скрытых признаков, основанные на латентном семантическом анализе [56], которые затем используются для кластеризации [57, 58].

Большинство алгоритмов кластеризации используют расстояние между объектами, измеряемое в принятой метрике, и переборные алгоритмы с эвристическим управлением [59]. Результаты кластеризации существенно зависят от принятой метрики. Поэтому числовая оценка качества кластеризации оказывается весьма важной [60–62].

Предлагаемый метод кластеризации основан на рандомизированном представлении ансамбля возможных кластеров, характеризуемым функцией распределения вероятностей.

5.2. Принцип рандомизированной кластеризации

Рассмотрим n объектов, состояние каждого характеризуется вектором $\mathbf{x}^{(i)} \in R^m$. В указанном пространстве множество объектов отображается в облако из n точек. Пусть, для начала, это множество точек нужно разделить на два подмножества – кластера $\mathcal{K}_s$ и $\mathcal{K}_{n-s}$ с заданными объемами (количеством объектов) s и $n - s$.

Поскольку количество объектов конечно и объем кластера $\mathcal{K}_s$ задан, то формально можно образовать конечное количество кластеров типа $\mathcal{K}_s$ объемом s, но с различным составом объектов. Отбор объектов в кластеры будем производить случайным образом и независимо друг от друга. Количество таких кластеров равно числу сочетаний $C_n^s = \dfrac{n!}{s!(n-s)!}$, но состав их образован случайным механизмом. Из оставшихся объектов образуются кластеры типа $\mathcal{K}_{n-s}$.

Таким образом, сформирован ансамбль рандомизированных кластеров $\mathcal{K}_s$, объема C_n^s, который характеризуется пока неизвестной дискретной функцией распределения вероятностей $p(k, s)$, где $k \to \{i_1,\ldots,i_s\}$ – номер конкретного кластера в ансамбле; $i_1,\ldots,i_s$ – номера объектов, входящих в k-кластер.

Для определения функции $p(k, s)$ воспользуемся методом энтропийно-рандомизированного оценивания (ЕРО). Согласно этому методу свойства оптимальности функции $p(k, s)$ характеризуются максимумом информационной энтропии при условии, что принятый *средний* показатель качества ансамбля кластеров принимает заданное значение.

Тогда наиболее вероятный кластер в ансамбле определяется максимумом функции $p^*(k^*, s) = \max_k p^*(k, s)$. Поскольку информационная энтропия определяется функцией распределения вероятностей, то вычисляется ее значение $H^*(s)$ для $p^*(k, s)$, и определяется $s^* = \max_s H^*(s)$.

5.3. Числовая характеристика множества объектов

Рассмотрим множество из n объектов, каждый из которых характеризуеся вектором $\mathbf{x}^{(i)}(i=\overline{1,n})$ из пространства R^m. Математическим образом совокупности объектов является блочный вектор следующего вида:

$$\mathbf{X}^{(1,\ldots,n)}=\begin{pmatrix}\mathbf{x}^{(1)}\\ \cdots\\ \mathbf{x}^{(n)}\end{pmatrix} \tag{5.1}$$

Определим расстояние между его компонентами – векторами $\mathbf{x}^{(i)}(i=\overline{1,n})$ в виде:

$$\varrho(\mathbf{x}^{(i)},\mathbf{x}^{(j)})=\|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\|_{R^m}, \tag{5.2}$$

где $\|\bullet\|_{R^m}$ – норма в пространстве R^m. Сформируем матрицу расстояний

$$D_{(n\times n)}=\begin{pmatrix}0 & \varrho(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) & \cdots & \varrho(\mathbf{x}^{(1)},\mathbf{x}^{(n)})\\ \varrho(\mathbf{x}^{(2)},\mathbf{x}^{(1)}) & 0 & \cdots & \varrho(\mathbf{x}^{(2)},\mathbf{x}^{(n)})\\ \cdots & \cdots & \cdots & \\ \varrho(\mathbf{x}^{(n)},\mathbf{x}^{(1)}) & \varrho(\mathbf{x}^{(n)},\mathbf{x}^{(2)}) & \cdots & 0\end{pmatrix}. \tag{5.3}$$

Определим в качестве числовой характеристики вектора $\mathbf{X}^{(1,\ldots,n)}$ среднее значение элементов матрицы расстояний $D_{(n\times n)}$, которое будем обозначать $\mathrm{dis}(\mathbf{X}^{(1,\ldots,n)})$ и называть *индикатором* вектора $\mathbf{X}^{(1,\ldots,n)}$:

$$\mathrm{dis}(\mathbf{X}^{(1,\ldots,n)})=\frac{1}{n^2}\sum_{(i,j)=1}^{n}\varrho(\mathbf{x}^{(i)},\mathbf{x}^{(j)}). \tag{5.4}$$

Важной характеристикой вектора $\mathbf{X}^{(1,\ldots,n)}$ являются минимальный и максимальный элементы матрицы расстояний $D_{(n\times n)}$:

$$\begin{aligned}\inf(D_{(n\times n)})&=\min_{i,j}\varrho(\mathbf{x}^{(i)},\mathbf{x}^{(j)}),\\ \sup(D_{(n\times n)})&=\max_{i,j}\varrho(\mathbf{x}^{(i)},\mathbf{x}^{(j)}).\end{aligned} \tag{5.5}$$

Заметим, что элементы матриц расстояний формируемых кластеров должны принадлежать интервалу

$$\mathcal{I}=[\inf(D),\sup(D)]. \tag{5.6}$$

5.4. Алгоритм рандомизированной бинарной кластеризации

Задача бинарной кластеризации состоит в распределении n объектов по двум кластерам $\mathcal{H}_{(s^*)}$, $\mathcal{H}_{(n-s^*)}$ с объемами s^* и $(n-s^*)$ объектов соответственно:

$$\begin{gathered}\mathcal{H}_{(s^*)}=\{i_1,\ldots,i_{s^*}\},\quad \mathcal{H}_{(n-s^*)}=\{j_1,\ldots,j_{(n-s^*)}\};\\ (i_1,\ldots,i_s)\neq(j_1,\ldots,j_{(n-s^*)}),\quad (i_\alpha,j_\beta)=\overline{1,n},\\ \alpha=\overline{1,s^*},\quad \beta=\overline{1,n-s^*}.\end{gathered} \tag{5.7}$$

1. $s=\tilde{s}$ – задано. При каждом фиксированном объеме кластера, т.е. значении $\tilde{s}$, процедура его формирования состоит в выделении в векторе $\mathbf{X}^{(1,\ldots,n)}$ подвектора

$$\mathbf{x}_{(\tilde{s})}=\mathbf{x}_{(\tilde{s})}^{(i_1,\ldots,i_{\tilde{s}})}=\begin{pmatrix}\mathbf{x}^{(i_1)}\\ \cdots\\ \mathbf{x}^{(i_{\tilde{s}})}\end{pmatrix}, \tag{5.8}$$

размера $\tilde{s}<n$.

Если подвектор $\mathbf{X}_{(\tilde{s})}^{(i_1,\ldots,i_{\tilde{s}})}$ выделен, то оставшиеся компоненты образуют подвектор $\mathbf{X}_{(n-\tilde{s})}=\mathbf{X}_{(n-\tilde{s})}^{j_1,\ldots,j_{n-\tilde{s}}}$, а совокупность номеров оставшихся компонент образуют кластер $\mathcal{H}_{(n-\tilde{s})}$.

Согласно *принципу рандомизированной бинарной кластеризации* вектор $\mathbf{x}_{(\tilde{s})}$ объявляется случайным. Перенумеруем набор

$$\{i_1,\ldots,i_{\tilde{s}}\}\to k=\overline{1,K(\tilde{s})};\quad K(\tilde{s})=C_{\tilde{s}}^{n}. \tag{5.9}$$

Таким образом, в результате рандомизации генерируется конечный ансамбль случайных векторов:

$$\mathcal{X}_{(\tilde{s})}=\{\mathbf{X}_{(\tilde{s})}^{1},\ldots,\mathbf{X}_{(\tilde{s})}^{K(\tilde{s})}\}. \tag{5.10}$$

Поскольку векторы в этом ансамбле случайные, то существуют вероятности реализации элементов этого ансамбля, т.е. функция распределения вероятностей $p(\tilde{s},k)$, где $\tilde{s}$ – объем кластера, а k – номер его реализации:

$$\mathbf{X}_{(\tilde{s})}^{(k)}\text{ с вероятностью }p(\tilde{s},k),\quad k=\overline{1,K(\tilde{s})}. \tag{5.11}$$

Итак, задача рандомизированной бинарной кластеризации сводится к определению подходящей в каком-то смысле функции дискретного распределения вероятностей $p(\tilde{s},k),k=\overline{1,K(\tilde{s})}$.

Рассмотрим кластер $\mathcal{H}_1$ объемом $\tilde{s}$ и соответствующий ему блочный вектор:

$$\mathbf{X}_{(\tilde{s})}^{(i_1,\ldots,i_{\tilde{s}})}=\begin{pmatrix}\mathbf{x}^{(i_1)}\\ \cdots\\ \mathbf{x}^{(i_{\tilde{s}})}\end{pmatrix}=\mathbf{X}_{(\tilde{s})}^{(k)},\quad \{i_1,\ldots,i_{\tilde{s}}\}\to k, \tag{5.12}$$

и матрицу расстояний:

$$\begin{gathered}D_{(\tilde{s})}^{(i_1,\ldots,i_{\tilde{s}})}=\\ =\begin{pmatrix}0 & \varrho^{(k)}(\mathbf{x}^{i_1},\mathbf{x}^{i_2}) & \cdots & \varrho^{(k)}(\mathbf{x}^{i_1},\mathbf{x}^{i_{\tilde{s}}})\\ \cdots & \cdots & \cdots & \\ \varrho^{(k)}(\mathbf{x}^{i_{\tilde{s}}},\mathbf{x}^{i_1}) & \varrho^{(k)}(\mathbf{x}^{i_{\tilde{s}}],}\mathbf{x}^{i_2}) & \cdots & 0\end{pmatrix}=D_{(\tilde{s})}^{(k)}.\end{gathered} \tag{5.13}$$

Воспользуемся понятием индикатора матрицы (5.4) и определим его для кластера $\mathcal{K}_s$ в виде:

$$\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)}) = \frac{1}{\tilde{s}^2}\sum_{(t,h)=1}^{\tilde{s}} \varrho^{(k)}(\mathbf{x}^{i_t}, \mathbf{x}^{i_h}). \tag{5.14}$$

Поскольку векторы $\mathbf{X}_{(\tilde{s})}^{(k)}$ предполагаются случайными объектами, на их ансамбле существует функция распределения $p(\tilde{s},k)$. Определим *средний индикатор* в виде:

$$\mathcal{M}\{\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)})\} = \sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)}). \tag{5.15}$$

Сформулируем *алгоритм бинарной кластеризации* при фиксированном значении $s = \tilde{s}$ в следующем виде:

– максимизация энтропийного функционала Больцмана-Шеннона:

$$\mathcal{H}_B[p(\tilde{s},k)|\tilde{s}] = -\sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\ln p(\tilde{s},k) \Rightarrow \max_{p(\tilde{s},k)}, \tag{5.16}$$

– при ограничениях:

$$0 \le p(\tilde{s},k) \le 1, \quad k = \overline{1, K(\tilde{s})}, \tag{5.17}$$

$$\inf(D) \le \sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)}) \le \sup(D), \tag{5.18}$$

где нижняя $\inf(D)$ и верхняя $\sup(D)$ границы значений элементов матрицы расстояний для исходного вектора определены в (5.6), и индикатор $\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)})$ определен равенством (5.14).

Задача (5.16)–(5.18) является конечно-мерной. Ограничения (5.17) можно опустить, если в качестве целевого функционала использовать энтропию Ферми [49]. После преобразований будем иметь следующую конечно-мерную задачу энтропийно-линейного программирования [64]:

$$H(\mathbf{p}|\tilde{s}) = -\sum_{k=1}^{K(\tilde{s})} p_{\tilde{s},k}\ln p_{\tilde{s},k} + \\ + (1-p_{\tilde{s},k})\ln(1-p_{\tilde{s},k}) \Rightarrow \max_{0\le p_{\tilde{s},k}\le 1},$$

$$\sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\overline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) \le -1, \quad \overline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) = -\frac{\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)})}{\inf(D)} \tag{5.19}$$

$$\sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\underline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) \le 1, \quad \underline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) = \frac{\mathrm{dis}(\mathbf{X}_{(\tilde{s})}^{(k)})}{\sup(D)}.$$

Функция Лагранжа имеет вид:

$$L[\mathbf{p},\lambda_1,\lambda_2|\tilde{s}] = H(\mathbf{p}|\tilde{s}) + \\ + \lambda_1\left(-1 - \sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\overline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)})\right) + \\ + \lambda_2\left(1 - \sum_{k=1}^{K(\tilde{s})} p(\tilde{s},k)\underline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)})\right). \tag{5.20}$$

Согласно теоремы Куна–Таккера [65], условия оптимальности имеют вид:

$$\nabla_{\mathbf{p}} L(\mathbf{p}^*\lambda_1^*,\lambda_2^*|\tilde{s}) = \mathbf{0}, \quad \frac{\partial L(\mathbf{p}^*,\lambda_1^*,\lambda_2^*|\tilde{s})}{\partial\lambda_i} \ge 0,$$

$$\lambda_i\frac{\partial L(\mathbf{p}^*,\lambda_1^*,\lambda_2^*|\tilde{s})}{\partial\lambda_i} = 0, \quad \lambda_i \ge 0, \quad i = 1,2.$$

Первая группа условий аналитически разрешима относительно компонент вектора $\mathbf{p}$:

$$p^*(\tilde{s},k|\lambda_1,\lambda_2) = \\ = \frac{\exp(-\lambda_1\overline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}) - \lambda_2\underline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}))}{1+\exp(-\lambda_1\overline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}) - \lambda_2\underline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}))}, \\ k = \overline{1, K(\tilde{s})}. \tag{5.21}$$

Вторая группа условий преобразуется в следующие неравенства:

$$L_{\lambda_1}(p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*) = \\ = -1 - \sum_{k=1}^{K(\tilde{s})} p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*)\overline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}) \ge 0, \\ L_{\lambda_2}(p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*) = \\ = 1 - \sum_{k=1}^{K(\tilde{s})} p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*)\underline{\mathrm{dis}}(X_{(\tilde{s})}^{(k)}) \ge 0. \tag{5.22}$$

Третья группа условий сводится к следующим уравнениям:

$$\lambda_1^* L_{\lambda_1}(p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*) = 0,$$

$$\lambda_2^* L_{\lambda_2}(p^*(\tilde{s},k|\lambda_1^*,\lambda_2^*) = 0, \tag{5.23}$$

$$\lambda_1^* \ge 0, \quad \lambda_2^* \ge 0.$$

Для определения неотрицательного решения указанных уравнений можно применить мультипликативный алгоритм следующего вида [49]:

$$\lambda_1^{q+1} = \lambda_1^q(1+\gamma L_{\lambda_1}(p^*(\tilde{s},k|\lambda_1^q,\lambda_2^q)), \\ \lambda_2^{q+1} = \lambda_2^q(1+\gamma L_{\lambda_2}(p^*(\tilde{s},k|\lambda_1^q,\lambda_2^q)), \\ (\lambda_1^0,\lambda_2^0) > 0. \tag{5.24}$$

Здесь $\gamma > 0$ – параметр, выбираемый из условий $\mathfrak{G}$-сходимости итерационного процесса (5.24).

Алгоритм (5.24) называется $\mathfrak{G}$-*сходящимся*, если в пространстве R_+^2 существует множество $\mathfrak{G}$ и скаляры $a(\mathfrak{G})$ и γ такие, что для всех $(\lambda_1^0, \lambda_2^0) \in \mathfrak{G}$ и $0 < \gamma \le a(\mathfrak{G})$ он сходится к решению λ_1^*, λ_2^* уравнения (5.23), причем сходимость в окрестности λ_1^*, λ_2^* – линейная.

Теорема 3. *Алгоритм* (5.24) $\mathfrak{G}$-*сходится к решению задачи* (5.24).

Доказательство приведено в Приложении Е.

Таким образом, определена функция распределения вероятностей

$$p^*(\tilde{s}, k \,|\, \lambda_1^*, \lambda_2^*) = \frac{\exp(-\lambda_1^* \overline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) - \lambda_2^* \underline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}))}{1 + \exp(-\lambda_1^* \overline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}) - \lambda_2^* \underline{\mathrm{dis}}(\mathbf{X}_{(\tilde{s})}^{(k)}))}, \quad k = \overline{1, K(s)}. \tag{5.25}$$

Естественно предположить, что, согласно общему принципу статистической механики, реализуемый кластер (при фиксированном объеме $\tilde{s}$) соответствует максимуму функции распределения вероятностей, т.е.

$$\mathcal{H}_{\tilde{s},k^*} \Rightarrow (\tilde{s}, k^*) = \max_k p^*(\tilde{s}, k \,|\, \lambda_1^*, \lambda_2^*). \tag{5.26}$$

2. $s = \overline{1, n-1}$. Рассмотрим случай, когда объем кластера не задан, т.е $\tilde{s} = s$, которое принимает значения в интервале $[1, n-1]$. При этом образуется последовательность максимальных значений информационной энтропии

$$H_B(s) = -\sum_{k=1}^{K(s)} p^*(s, k^*) \ln p^*(s, k^*). \tag{5.27}$$

Оптимальное значение объема кластера определяется максимальным элементом в этой последовательности:

$$s^* = \operatorname*{argmax}_s H_B(s). \tag{5.28}$$

6. ЭНТРОПИЙНЫЕ ПРОЕКЦИИ ДЛЯ РЕДУКЦИИ РАЗМЕРНОСТИ МАТРИЦЫ ДАННЫХ

Во многих прикладных задачах обработки данных последние форматируются в виде прямоугольных матриц $U_{(m\times s)}$. Без ограничения общности будем считать m – количество объектов (прецендентов), s – количество признаков.

По разным причинам возникает необходимость "сжать" матрицу данных, т.е. трансформировать ее в матрицу, размерности $(m \times r)$ или $(n \times r)$, $n < m$, $r < s$.

Данная проблема вложена в более общую: приближение заданного набора многомерных точек маломерным аффинным многообразием [38]. Здесь следует отметить метод главных компонент (МГК) [39] и его робастные версии [40], а так же метод случайных проекций [41, 42].

В [43] был предложен энтропийный метод одномерного (столбцы или строки) детерминированного сжатия матрицы данных (EDR-метод), основанный на "прямом" и "обратном" проектировании. Матрицы-проекторы определяются путем минимизации кросс-энтропийного функционала.

Здесь EDR-метод развивается для параллельного сжатия матрицы данных с учетом их информационной емкости, реализуемого на базе условных энтропийных проекций с детерминированными и рандомизированными матрицами-проекторами. В последнем случае применяется принцип сохранения среднего расстояния между многомерными и маломерными точками в соответствующих пространствах.

6.1. Параллельное детерминированное проектирование с ограничениями информационной емкости (расширенный EDR-метод)

Параллельная реализация процедуры "прямого" и "обратного" проектирования, примененная к матрице данных $U_{(m\times s)} > 0$, приводит к следующей цепочке матричных равенств:

- "прямая" проекция

$$U_{(m\times s)} Q_{(s\times r)} = Y_{(m\times r)}, \quad B_{(n\times m)} Y_{(m\times r)} = Z_{(n\times r)}, \tag{6.1}$$

- "обратная" проекция

$$Z_{(n\times r)} W_{(r\times s)} = D_{(n\times s)}, \quad E_{(m\times n)} D_{(n\times s)} = X_{(m\times s)}. \tag{6.2}$$

Матрицы-проекторы Q, B, W, E – неотрицательные. Равенства (6.1) преобразуют матрицу $U_{(m\times s)}$ в "сжатую" матрицу $Z_{(n\times r)}$, где $n < m$, $r < s$. Равенство (6.2) преобразует матрицу $Z_{(n\times r)}$ в матрицу $X_{m\times s}$ той же размерности, что и исходная матрица данных $U_{(m\times s)}$.

Из равенств (6.1)–(6.2) имеем:

$$X_{(m\times s)} = E_{(m\times n)} \left\{ \left[B_{(n\times m)} \left(U_{(m\times s)} Q_{(s\times r)} \right) \right] W_{(r\times s)} \right\} > 0. \tag{6.3}$$

Скобки в этом равенстве указывают на последовательность операций проектирования: $(\bullet) \to [\bullet] \to \{\bullet\}$.

Элементы матрицы-проекции $Z_{(n\times s)}$ имеют вид:

$$z_{\mu,\nu} = \sum_{\beta=1}^{m} b_{\mu,\beta} \sum_{\alpha=1}^{r} u_{\beta,\alpha} q_{\alpha,\nu}, \quad \mu = \overline{1, n}, \quad \nu = \overline{1, r}. \tag{6.4}$$

Элементы матрицы $X_{(m\times s)}$ имеют вид:

$$x_{ij} = \sum_{\mu=1}^{n} e_{i,\mu} \sum_{\nu=1}^{r} w_{\nu,j} \sum_{\beta=1}^{m} b_{\mu,\beta} \sum_{\alpha=1}^{s} u_{\beta,\alpha} q_{\alpha,\nu} > 0, \quad i = \overline{1,m}, \quad j = \overline{1,s}. \tag{6.5}$$

Для измерения отклонения преобразованной матрицы $X_{(m\times s)}$ от исходной $U_{(m\times s)}$ воспользуемся *информационной кросс-энтропией* [44]

$$\mathcal{H}(X|U) = \sum_{i=1}^{m}\sum_{j=1}^{s} s_{ij}(X|U), \tag{6.6}$$

где

$$s_{ij} = x_{ij} \ln \frac{x_{ij}}{u_{ij}}. \tag{6.7}$$

С учетом равенства (6.5) не трудно видеть, что информационная кросс-энтропия (6.6) есть скалярная функция от матрицы данных $U > 0$ и матриц-проекторов $(Q, B, W, E) \geq 0$, т.е.

$$\mathcal{H} = \mathcal{H}(Q, B, W, E|U). \tag{6.8}$$

Важным показателем качества процедуры редукции является оптимальное снижение информационной емкости редуцированной матрицы $Z_{(n\times r)}$ по сравнению с информационной емкостью исходной матрицы данных $U_{(m\times s)}$ [45].

Информационная емкость измеряется в энтропийных терминах:

$$\mathcal{I}_Z = \sum_{(i,j)=1}^{n,r} z_{ij}(Q,B) \ln z_{ij}(Q,B) + e^{-1} nr,$$
$$\mathcal{I}_U = \sum_{(i,j)=1}^{m,s} u_{ij} \ln u_{ij} + e^{-1} ms. \tag{6.9}$$

Различие в указанных информационных емкостях будем характеризовать квадратичным функционалом

$$\mathcal{J}(Q,B|U) = \left(\sum_{(i,j)=1}^{n,r} z_{ij}(Q,B) \ln z_{ij}(Q,B) - A \right)^2, \tag{6.10}$$

где

$$A = e^{-1}(ms - nr) + \sum_{(i,j)=1}^{m,s} u_{ij} \ln u_{ij}. \tag{6.11}$$

Образуем обобщенный функционал

$$\mathcal{F}(Q,B,W,E|U) = \mathcal{H}(Q,B,W,E|U) + \mathcal{J}(Q,B|U), \tag{6.12}$$

и оптимальные значения неотрицательных элементов матриц-проекторов будем определять, минимизируя функционал $\mathcal{F}(Q,B,W,E|U)$:

$$(Q^*, B^*, W^*, E^*) = \arg \min_{(Q,B,W,E)\geq 0} \mathcal{F}(Q,B,W,E|U). \tag{6.13}$$

Замечание. В задаче (6.13) условие близости информационных емкостей матрицы данных и редуцированной матрицы может быть реализовано в виду соответствующего ограничения. Тогда оптимальные значения неотрицательных элементов матриц-проекторов определяются решением следующей задачи:

$$(Q^*, B^*, W^*, E^*) = \arg \min_{(Q,B,W,E)\in\Omega} \mathcal{H}(Q,B,W,E|U),$$
$$\Omega = \{(Q,B,W,E) : (Q,B,W,E) \geq 0; \quad \mathcal{I}_Z(Q,B) \geq \delta \mathcal{I}_U\}, \quad \delta \in (0,1). \tag{6.14}$$

Допустимый уровень снижения информационной емкости редуцированной матрицы регулируется параметром δ.

Алгоритм параллельной редукции. Задача (6.13) является задачей минимизации функционала на неотрицательном ортанте. Для ее решения применим метод проекций градиента, предварительно осуществив векторизацию соответствующих матриц [46].

Введем блочные векторы $\mathbf{v} = \{\mathbf{q}, \mathbf{b}\}$ и $\mathbf{c} = \{\mathbf{w}, \mathbf{e}\}$, каждый размерности

$$N = (sr + nm),$$

где векторы $\mathbf{q}$, $\mathbf{b}$ являются результатами векторизации матриц Q, B, а векторы $\mathbf{w}$, $\mathbf{e}$ являются результатами векторизации матриц W, E соответственно.

Представим (6.13) в следующем виде:

$$\mathcal{F}(\mathbf{v},\mathbf{c}|\mathbf{u}) = \mathcal{H}(\mathbf{v},\mathbf{c}|\mathbf{u}) + \mathcal{J}(\mathbf{v}|\mathbf{u}) \Rightarrow \min,$$
$$\mathcal{H}(\mathbf{v},\mathbf{c}|\mathbf{u}) = \langle \mathbf{x}(\mathbf{v},\mathbf{c}), \mathbf{y}(\mathbf{v},\mathbf{c}|\mathbf{u}) \rangle_{R^{ms}},$$
$$\mathcal{J}(\mathbf{v}) = \langle \mathbf{z}(\mathbf{v}), \mathbf{g}(\mathbf{v}) \rangle_{R^{nr}}, \tag{6.15}$$
$$\mathbf{v} \geq \mathbf{0}, \quad \mathbf{c} \geq \mathbf{0}.$$

Здесь приняты следующие обозначения :

- вектор $\mathbf{u}$ – результат векторизации матрицы данных U, размерности (ms); и вектор $\mathbf{x}$, размерности (ms), с компонентами (6.5);
- вектор $\mathbf{y}$, размерности (ms) с компонентами

$$y_k = \ln \frac{x_k}{u_k}, \quad k = \overline{1, ms} \tag{6.16}$$

- вектор $\mathbf{g}$, размерности (nr) с компонентами

$$g_k = \ln z_k, \quad k = \overline{1, nr}. \tag{6.17}$$

В параллельной процедуре вектор $\mathbf{v}$ объединяет элементы матриц Q, B, с помощью которых производится “сжатие” матрицы данных по одному измерению. В вектор $\mathbf{c}$ входят элементы матриц W, E, с помощью которых производится “сжатие” по второму измерению. Такое разделение векторов удобно для применения по-координатного алгоритма.

Итерационный шаг по-координатной схемы метода проекций градиента состоит из двух последовательно реализуемых этапов: на одном осуществляется итерация по $\mathbf{v}$-проекциям градиента,

а на другом – по $\mathbf{c}$-проекциям градиента функционала $\mathcal{F}(\mathbf{v},\mathbf{c}|\mathbf{u})$. Обозначим градиенты по этим векторам:

$$\nabla_{\mathbf{v}}\mathcal{F} = \nabla_{\mathbf{v}}\mathcal{H} + \nabla_{\mathbf{v}}\mathcal{J},$$
$$\nabla_{\mathbf{c}}\mathcal{F} = \nabla_{\mathbf{c}}\mathcal{H}. \tag{6.18}$$

Для численного решения этой задачи применим по-координатную схему метода проекций градиента. Алгоритм минимизации функционала $\mathcal{F}$ имеет следующий вид:

а) *начальный шаг*

$$\mathbf{v}^0 > \mathbf{0}, \quad \mathbf{c}^0 > \mathbf{0};$$

б) *i-й итерационный шаг*

$$X^i = E^i\{[B^i(UQ^i)]W^i\};$$
$$\mathcal{F}^i = \mathcal{H}^i(\mathbf{v}^i,\mathbf{c}^i|\mathbf{u}) + \mathcal{J}^i(\mathbf{v}^i|\mathbf{u});$$
$$\mathbf{v}^{(i+1)} = \begin{cases} \mathbf{v}^i + \gamma_{\mathbf{v}}(\nabla_{\mathbf{v}}\mathcal{H}^i(\mathbf{v}^i,\mathbf{c}^i|\mathbf{u}) + \nabla_{\mathbf{v}}\mathcal{J}^i(\mathbf{v}^i)), \\ \text{если} \quad \mathbf{v}^{(i+1)} \ge \mathbf{0}, \\ \mathbf{v}^i, \quad \text{если} \quad \mathbf{v}^{(i+1)} < \mathbf{0}. \end{cases}$$
$$\mathbf{v}^{(i+1)} \Rightarrow Q^{(i+1)}, B^{(i+1)};$$
$$\mathbf{c}^{(i+1)} = \begin{cases} \mathbf{c}^n + \gamma_{\mathbf{c}}\nabla_{\mathbf{c}}\mathcal{H}(\mathbf{v}^i,\mathbf{c}^i|\mathbf{u}), \quad \text{если} \quad \mathbf{c}^{(i+1)} \ge \mathbf{0}, \\ \mathbf{c}^i, \quad \text{если} \quad \mathbf{c}^{(i+1)} < \mathbf{0}; \end{cases}$$
$$\mathbf{c}^{(n+1)} \Rightarrow W^{(i+1)}, E^i(i+1);$$
$$X^{(i+1)} = E^{(i+1)}\{[B^{(i+1)}(UQ^{(i+1)})]W^{(i+1)}\};$$
$$\mathcal{F}^{(i+1)} = \mathcal{H}^{i+1}(\mathbf{v}^{i+1},\mathbf{c}^{i+1}|\mathbf{u}) + \mathcal{J}^{i+1}(\mathbf{v}^{i+1}|\mathbf{u})$$

в) *условие остановки*

$$\text{если} \quad \mathcal{F}^{(i+1)} - \mathcal{F}^{(i)} \le \Delta, \Rightarrow \text{STOP}.$$

6.2. Энтропийно-рандомизированные проекции (REDR-метод)

Рандомизированное проектирование с целью снижения размерности исходной матрицы данных основано на существовании линейного преобразования, мало меняющего среднее расстояние между точками исходного и редуцированного пространств (лемма Джонсона-Линденштраусса [66, 67].

Рассмотрим снова матрицу данных $U_{(m\times s)}$. В пространстве R^s ее отображает множество точек $\mathfrak{U} = \{\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}\}$. Определим, как это производилось в разделе 5, *индикатор* этой группы (матрицы-данных) в виде:

$$\text{dis}(U) = \frac{1}{m^2}\sum_{(\alpha,\beta)=1}^{m}\varrho(\mathbf{u}^\alpha,\mathbf{u}^\beta). \tag{6.19}$$

Матрицу данных $U_{(m\times s)}$ трансформируем в редуцированную матрицу $Z_{(n\times r)}$, $n < m$, $r < s$ с помощью случайных, интервальных левых $B_{(n\times m)}$ и правых $Q_{(s\times r)}$ матриц-проекторов:

$$Z_{(n\times r)} = B_{(n\times m)}U_{(m\times s)}Q_{(s\times r)},$$
$$Q \in \mathfrak{Q} = [Q^-,Q^+], \quad B \in \mathfrak{B} = [B^-,B^+]. \tag{6.20}$$

Вероятностные свойства матриц проекторов характеризуются совместной функцией ПРВ $P(Q,B)$ (ПРВ), которая определена на носителе $\mathcal{V}$:

$$(Q,B) \in \mathcal{V} = \mathfrak{Q} \cap \mathfrak{B}. \tag{6.21}$$

Элементы редуцированной матрицы Z имеют вид:

$$z_{\mu,\nu}(Q,B) = \sum_{\beta=1}^{m} b_{\mu,\beta}\sum_{\alpha=1}^{r} u_{\beta,\alpha}q_{\alpha,\nu},$$
$$\mu = \overline{1,n}, \quad \nu = \overline{1,r}. \tag{6.22}$$

По аналогии с (6.19) определим индикатор редуцированной матрицы $Z_{(n\times s)}$ в виде:

$$\text{dis}(Z(Q,B)) = \frac{1}{n^2}\sum_{(\eta,\kappa)=1}^{n}\varrho(\mathbf{z}^{(\eta)}(Q,B),\mathbf{z}^{(\kappa)}(Q,B)). \tag{6.23}$$

Поскольку элементы матриц-проекторов – случайные, индикатор $\text{dis}(Z(Q,B))$ является функцией случайных переменных. Его математическое ожидание

$$\mathcal{M}\{\text{dis}(Z(Q,B))\} = G[P(Q,B)] =$$
$$= \int_{\mathcal{V}} P(Q,B)\text{dis}(Z(Q,B))dQdB. \tag{6.24}$$

Для определения функции ПРВ $P(Q,B)$ будем использовать оценку максимальной энтропии (см. раздел 1):

$$\mathcal{H}[\mathcal{P}(\mathfrak{Q},\mathfrak{B})] = -\int_{\mathfrak{L}} P(Q,B)\ln P(Q,B)dQdB \Rightarrow \max, \tag{6.25}$$

при ограничениях:

$$\int_{\mathfrak{L}} P(Q,B)dQdB = 1;$$
$$\int_{\mathcal{V}} P(Q,B)\text{dis}(Z(Q,B))dQdB = \delta\text{dis}(U), \tag{6.26}$$
$$0 < \delta < \theta < 1.$$

Задача (6.25)–(6.26) относится к классу ляпуновских задач [48], для которых условия оптимальности формулируются в терминах стационарности функционала Лагранжа

$$\mathcal{L}[P(Q,B),\lambda] =$$
$$= \mathcal{H}[\mathcal{P}(\mathfrak{Q},\mathfrak{B})] + \lambda\left(\delta\text{dis}(U) - G[P(Q,B)]\right), \tag{6.27}$$

где λ – скалярный множитель Лагранжа.

Получим энтропийно-оптимальную ПРВ

$$P^*(Q,B)=\frac{\exp\left(-\lambda \mathrm{dis}(Z(Q,B))\right)}{\mathcal{P}(\lambda)}, \tag{6.28}$$

где

$$\mathcal{P}(\lambda)=\int\limits_{\mathcal{L}}\exp\left(-\lambda \mathrm{dis}(Z(Q,B))\right)dQdB. \tag{6.29}$$

Множитель Лагранжа λ определяется из следующего уравнения:

$$\int\limits_{V}\exp\left(-\lambda \mathrm{dis}(Z(Q,B))\right)\times \times\left(\mathrm{dis}(Z(Q,B))-\delta \mathrm{dis}(U)\right)dQdB=0. \tag{6.30}$$

Таким образом, энтропийно-оптимальная функция ПРВ $P^*(Q,B)$ (6.28) позволяет, путем ее сэмплирования, генерировать матрицы-проекторы Q, B, сохраняющие "в среднем" расстояние между точками (векторами $\mathbf{z}^{(\alpha)}$) редуцированной матрицы $\mathcal{L}$.

6.3. Рандомизированные матрицы-проекторы с заданными значениями элементов

Рассмотрим матрицу данных $U_{(m\times s)}$, которую нужно "сжать" по переменной s до размера r:

$$Y_{(m\times r)}=U_{(m\times s)}Q_{(s\times r)}. \tag{6.31}$$

Матрица $U_{(m\times s)}\geq 0$ и имеет нормированные элементы ($0\leq u_{ij}\leq 1$). Определим индикатор редуцированной матрицы $Y_{(m\times r)}$ в виде:

$$\mathrm{dis}(Y(Q))=\frac{1}{m^2}\sum_{(i,j)=1}^{m}\rho(\mathbf{y}^{(i)}(Q),\mathbf{y}^{(j)}(Q)). \tag{6.32}$$

Рассмотрим случай, когда элементы матрицы $Q_{(s\times r)}$ могут принимать значения 0 или 1, и размещение их в матрице – случайное. Количество различных матриц такого типа равно $N=2^{rs}$:

$$Q^{(1)},\ldots,Q^{(N)}. \tag{6.33}$$

Полагая, что реализации – случайные, их вероятностные свойства будем характеризовать дискретной функцией распределения вероятностей (ДРВ)

$$W(a)=w_a,\quad 0\leq w_a\leq 1,\quad a=\overline{1,N}. \tag{6.34}$$

Математическое ожидание индикатора (6.32)

$$\mathcal{M}\{\mathrm{dis}(Y(Q))\}=\sum_{a=1}^{N}w_a\mathrm{dis}(Y(Q^a)). \tag{6.35}$$

Функцию ДРВ $W(a)$ будем искать в классе функций, максимизирующих функцию информационной энтропии Ферми [49]:

$$H_F(\mathbf{w})=-\sum_{a=1}^{N}w_a\ln w_a+ +(1-w_a)\ln(1-w_a)\Rightarrow \max, \tag{6.36}$$

при ограничении математического ожидания индикатора (6.32):

$$\sum_{a=1}^{N}w_a\mathrm{dis}(Y(Q^a))=\delta \mathrm{dis}(U),\quad 0<\varepsilon\leq\delta\leq 1. \tag{6.37}$$

Задача (6.36)–(6.37) является конечно-мерной задачей максимизации с вогнутой целевой функцией и нелинейным ограничением.

Рассмотрим функцию Лагранжа

$$L(\mathbf{w},\lambda)=H_F(\mathbf{w})+ +\lambda\left(\delta \mathrm{dis}(U)-\sum_{a=1}^{N}w_a\mathrm{dis}(Y(Q^a))\right). \tag{6.38}$$

Условия стационарности этой функции имеют вид:

$$\frac{\partial L}{\partial w_a}=-\ln\frac{w_a}{1-w_a}-\lambda \mathrm{dis}(Y(Q^a))=0,\quad a=\overline{1,N};$$
$$\frac{\partial L}{\partial \lambda}=\left(\delta \mathrm{dis}(U)-\sum_{a=1}^{N}w_a\mathrm{dis}(Y(Q^a))\right)=0. \tag{6.39}$$

Отсюда получаем, что энтропийно-оптимальное распределение вероятностей имеет вид:

$$w_a^*=\frac{\exp(-\lambda \mathrm{dis}(Y(Q^a))))}{1+\exp(-\lambda \mathrm{dis}(Y(Q^a)))},\quad a=\overline{1,N}, \tag{6.40}$$

где параметр λ определяется из следующего уравнения:

$$\sum_{a=1}^{N}\frac{\exp(-\lambda \mathrm{dis}(Y(Q^a))))\mathrm{dis}(Y(Q^a))}{1+\exp(-\lambda \mathrm{dis}(Y(Q^a)))}=\delta \mathrm{dis}(U). \tag{6.41}$$

Таким образом, равенство (6.40) определяет распределение вероятностей матриц-проекторов с элементами {0, 1}. Имеет смысл выбрать матрицу-проектор

$$Q^{(a^*)}\Rightarrow a^*=\max_{1\leq a\leq N}w_a^*, \tag{6.42}$$

хотя возможны и другие стратегии.

7. ПРИКЛАДНЫЕ ЗАДАЧИ И ИЛЛЮСТРАТИВНЫЕ ПРИМЕРЫ

7.1. Рандомизированная бинарная классификация

Проблема классификации объектов является весьма актуальной в современной теоретической и прикладной науке. При рандомизированной бинарной классификации (РБК) применяется модель решающего правила со случайными параметрами, оценки *функции ПРВ* которых определяются путем ЭРО-оценивания (см. раздел 3) с

учетом обучающих последовательностей данных. Принципиальное отличие РБК от существующих процедур состоит в генерации эмпирической функции принадлежности классам обученной на реальных данных.

7.1.1. Процедура РБК

1. Структура обучающих данных. Пусть имеется обучающая коллекция из n объектов, характеризуемых векторами $\{\mathbf{e}^{(1)},\ldots,\mathbf{e}^{(n)}\}$ из признакового пространства R^m и вектором-ответов $\mathbf{y} = \{0,1,...,1\}$. Этот вектор размерности n, и его компоненты 0 и 1 являются метками принадлежности объекта классу 1 (“1”), или классу 2 (“0”).

2. Модель решающего правила. Используем однослойную нейронную сеть с сигмоидной функцией активации [63]:

$$\hat{y}^{(i)}(\mathbf{a}) = \mathrm{sigm}(\langle \mathbf{E}^{(i)}, \mathbf{b}\rangle), \quad i = \overline{1,n}, \tag{7.1}$$

где

$$\mathrm{sigm}(x_i) = \frac{1}{1+\exp[-\alpha(x_i-\Delta)]},$$
$$x_i = (\langle \mathbf{E}^{(i)}, \mathbf{b}\rangle), \tag{7.2}$$
$$\mathbf{a} = \{\mathbf{b}, \alpha, \Delta\}.$$

На рис. 2 показан график сигмоидной функции с параметрами “крутизны” α и “порога” Δ. Значения функции $\mathrm{sigm}(x)$ в интервале $\left[\frac{1}{2},1\right]$ соответствуют первому классу, и значения в интервале $\left[0,\frac{1}{2}\right]$ – второму классу.

В рандомизированной модели (7.1), (7.2) параметры $\mathbf{a} = \{a_1,\ldots,a_{(n+2)}\}$ – интервального типа:

$$a_k \in \mathscr{A}_k = [a_k^-, a_k^+], \quad k = \overline{1,n+2},$$
$$\mathscr{A} = \bigcup_{k=1}^{n+2} \mathscr{A}_k. \tag{7.3}$$

Их вероятностные свойства характеризуются *PDF* $P(\mathbf{a})$, которая определена на множестве $\mathscr{A}$.

Итак, алгоритм рандомизированной бинарной классификации представляется в следующем виде:

$$\mathscr{H}[P(\mathbf{a})] = -\int_{\mathscr{A}} P(\mathbf{a})\ln P(\mathbf{a})d\mathbf{a} \Rightarrow \max, \tag{7.4}$$

при условиях

$$\int_{\mathscr{A}} P(\mathbf{a})d\mathbf{a} = 1, \tag{7.5}$$

$$\int_{\mathscr{A}} P(\mathbf{a})\hat{y}^i(\mathbf{a})d\mathbf{a} = y_i, \quad i = \overline{1,h}. \tag{7.6}$$

Энтропийно-оптимальная ПРВ

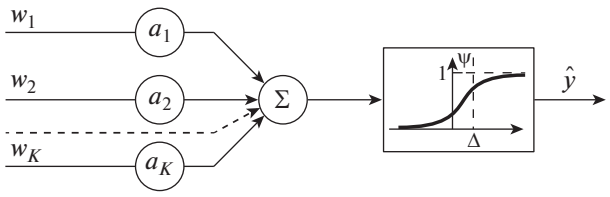

Рис. 2. Структура однослойной нейронной сети

$$P^*(\mathbf{a}|\theta) = \frac{\exp\left[-\sum_{i=1}^{h}\theta_i\hat{y}^{(i)}(\mathbf{a})\rangle\right]}{\mathscr{P}(\theta)}, \tag{7.7}$$

где $\hat{y}^{(i)}(\mathbf{a})$ определяется равенством (7.1), и

$$\mathscr{P}(\theta) = \int_{\mathscr{A}} \exp\left[-\sum_{i=1}^{h}\theta_i\hat{y}^{(i)}(\mathbf{a})\right]. \tag{7.8}$$

Множители Лагранжа определяются из уравнений (7.6).

3. Реализация рандомизированной бинарной классификации. Рассмотрим коллекцию из M объектов, характеризуемых векторами $\mathbf{t}^{(i)}$, $i = \overline{1,M}$ и подлежащих бинарной классификации. Воспользуемся найденной энтропийно-оптимальной функцией ПРВ $P^*(\mathbf{a})$ модели решающего правила (7.1) для определения эмпирических вероятностей принадлежности объекта k классу 1 или 2. Обозначим эти вероятности $p_1^{(k)}$ и $p_2^{(k)}$ соответственно, $k = \overline{1,M}$.

Функция ПРВ $P(\mathbf{a})$ сэмплируется, т.е. генерируется модифицированным методом Монте-Карло соответствующая последовательность случайных векторов-параметров модели $\mathbf{a}^{(i)}, i = \overline{1,N}$, где N – количество МК-испытаний. Пусть в результате этих испытаний оказалось, что первый объект N_1 раз был отнесен к первому классу и $N - N_1$ раз – ко второму; ...., k-й объект N_k раз отнесен к первому классу и $N - N_k$ раз – ко второму классу, и т.д. При достаточно большом числе испытаний определяются эмпирические вероятности принадлежности

$$p_1^{(1)} = \frac{N_1}{N},\ldots,p_1^{(k)} = \frac{N_k}{N},\ldots,\frac{N_M}{N},$$
$$p_2^{(1)} = \frac{N-N_1}{N},\ldots,p_2^{(k)} = \frac{N-N_k}{N},\ldots,\frac{N-N_M}{N},$$

которые являются результатом рандомизированной бинарной классификации.

Если требуется “жесткий” вариант классификации, то необходимо задать порог δ вероятности, и считать объекты, вероятность принадлежности которых превышает порог, относящимися к соответствующему классу. В результате “жест-

Таблица 1

i	$e_1^{(i)}$	$e_2^{(i)}$	$e_3^{(i)}$	$e_4^{(i)}$
1	0.11	0.75	0.08	0.21
2	0.91	0.65	0.11	0.81
3	0.57	0.17	0.31	0.91

Таблица 2. Матрица данных

№	x_1	x_2	Вид	№	x_1	x_2	Вид	№	x_1	x_2	Вид
1	4.5	1.5	2	8	3.9	1.4	2				
2	4.6	1.5	2	9	4.5	1.3	2	15	1.4	0.2	1
3	4.7	1.4	2	10	4.6	1.3	2	16	1.5	0.2	1
4	1.7	0.4	1	11	1.4	0.2	1	17	1.5	0.1	1
5	1.3	0.2	1	12	4.7	1.6	2	18	4.9	1.5	2
6	1.4	0.3	1	13	4.0	1.3	2	19	3.3	1.0	2
7	1.5	0.2	1	14	1.4	0.2	1	20	1.4	0.2	1

кой" классификации возникает вектор $\mathbf{u} = \{0,1,1, ..., 1\}$, номер компоненты которого соответствует номеру объекта, а 0 или 1 отображает принадлежность его классу 1 или 2.

Далее мы воспользуемся этим вектором для оценки точности "жесткой" классификации, выполненной рандомизированной бинарной процедурой.

7.1.2. Модельные примеры

1. Рандомизированная классификация четырехмерных объектов. Рассмотрим объекты, характеризуемые 4 признаками.

1.1. Обучение. Обучающая коллекция состоит из трех объектов, значения признаков которых показаны в табл. 1.

Рандомизированная модель решающего правила (7.1), (7.2) имеет параметры: $\alpha = 1,0$ и $\Delta = 0$. Выход модели $\hat{\mathbf{y}} = \{0,18; 0,81; 0,43\}$ ($y_i < 0,5$ соответствует классу 2, $y_i \geq 0,5$ соответствует классу 1).

Множители Лагранжа для энтропийно-оптимальной *PDF* (7.7) имеют следующие значения: $\bar{\theta}^* = \{0.2524; 1.7678; 1.6563\}$. Параметры $a_j \in [-10, 10]$, $j = \overline{1,4}$. Энтропийно-оптимальная для данной обучающей коллекции функция *PDF* имеет вид:

$$P^*(\mathbf{a}, \bar{\theta}) = \frac{\exp\left(-\sum_{i=1}^{3}\theta_i y_i(\mathbf{a})\right)}{\mathcal{P}(\bar{\theta})}, \qquad y_i(\mathbf{a}) = \left(1 + \exp\left(-\sum_{k=1}^{4} e_k^{(i)}, a_k\right)\right)^{(-1)}. \tag{7.9}$$

На рис. 3 показано двумерное сечение *PDF* $P^*(\mathbf{a}, \bar{\theta}^*)$.

1.2. Реализация. На этом этапе используется коллекция из r модельных объектов, где каждый объект характеризуется вектором $\mathbf{t}^{(j)} \in R^{(4)}$. Генерируется массив (500 × 4) четырехмерных случайных, независимых векторов $\mathbf{t}^{(i)}, i = \overline{1,500}$ с независимыми компонентами, равномерно распределенными в интервалах [0, 1]. Далее применяется процедура рандомизированной бинарной классификации. На рис. 4 показаны эмпирические вероятности $p_1^{(i)}, p_2^{(i)}$ принадлежности t_i-объекта классу 1 и 2.

2. Рандомизированная классификация двумерных объектов. Рассмотрим объекты, характеризуемые 2 признаками.

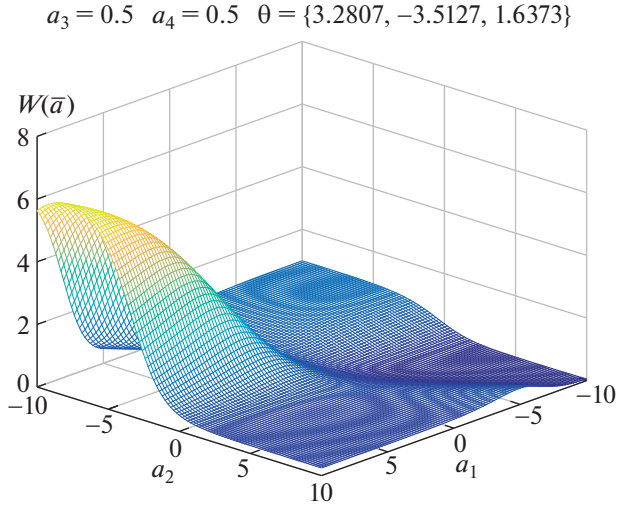

Рис. 3. Двумерное сечение *PDF* $P^*(\mathbf{a}, \bar{\theta}^*)$.

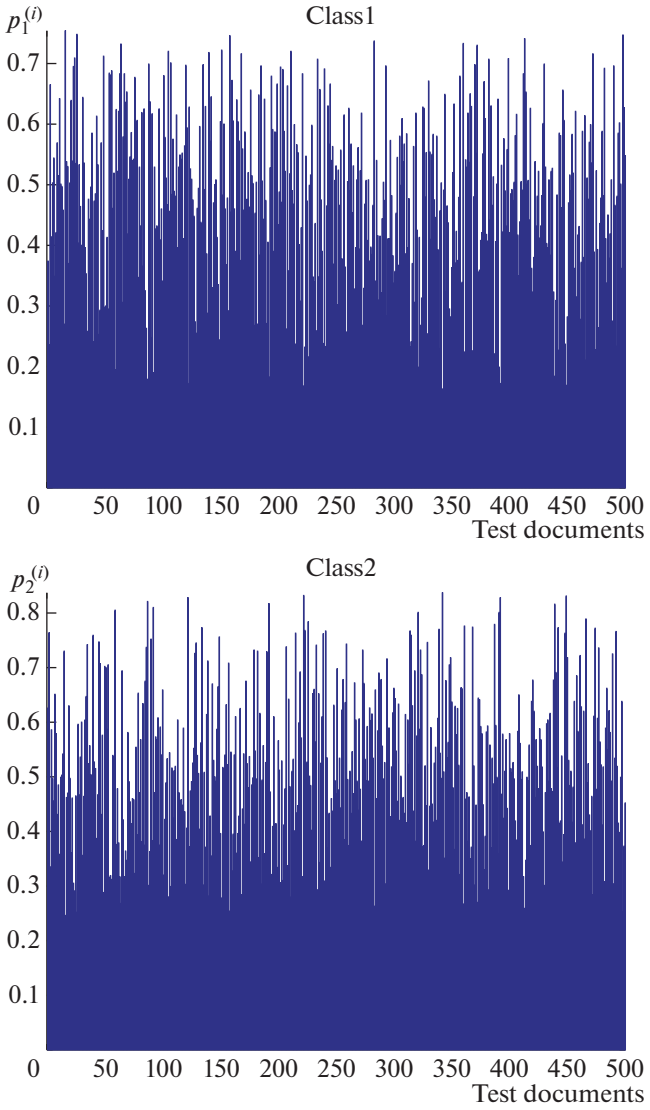

Рис. 4. Эмпирические вероятности принадлежности классам.

2.1. Обучение. Обучающая коллекция состоит из трех объектов, каждый из которых описывается двумя признаками, значения которых показаны в табл. 2. Значения параметров α, Δ и интервалов для случайных параметров $\mathbf{a}$ соответствуют примеру 1. Множители Лагранжа для энтропийно-оптимальной *PDF* (7.7) имеют следующие значения: $\bar{\theta}^* = \{9.6316; -18.5996; 16.7502\}$. Энтропийно-оптимальная для данной обучающей коллекции функция $P^*(\mathbf{a}|\bar{\theta})$ имеет вид:

$$P^*(\mathbf{a}) = \frac{\exp\left(-\sum_{i=1}^{3}\theta_i y_i(\mathbf{a})\right)}{\mathcal{P}(\bar{\theta})}, \qquad (7.10)$$

$$y_i(\mathbf{a}) = \left(1 + \exp\left(-\sum_{k=1}^{2} e(i)_k, a_k\right)\right)^{(-1)}.$$

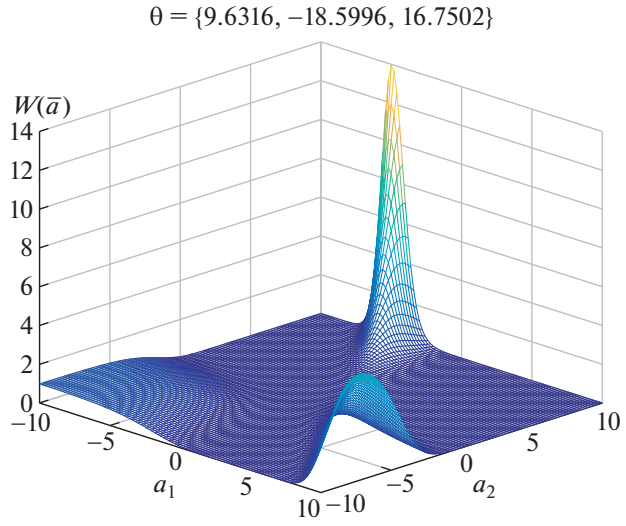

Рис. 5. Энтропийно-оптимальная *PDF* $P^*(\mathbf{a}, \bar{\theta}^*)$.

На рис. 5 показана функция $P^*(\mathbf{a}, \bar{\theta}^*)$.

2.2. Реализация. Все параметры этого примера соответствуют примеру 2. На рис. 6 показаны эмпирические вероятности $p_1^{(i)}$, $p_2^{(i)}$ принадлежности t_i-го объекта классу 1 и 2 ($i = \overline{1,500}$).

7.2. Рандомизированная кластеризация биологичеких объектов

Рассмотрим бинарную кластеризацию цветков ириса, используя базу Fisheriris (цветки ириса: ширина x_1 и длина x_2 лепестков). База содержит данные по указанным признакам трех видов цветков: “setosa” (1), “versicolor” (2), “virginica” (3), в количестве 50-ти двумерных точек на каждый вид. Далее будем рассматривать два вида (1, 2) и по 10 данным на каждый вид. В иллюстративных примерах удобнее представлять характеристики цветков в матричном виде.

Пример 1

Матрица данных содержит числовые значения двух признаков для 1-го и 2-го видов и представлена в табл. 2.

На рис. 7 показано расположение объектов-точек на плоскости.

Матрица расстояний $D_{(20\times20)}$ показана в табл. 3, 4.

Минимальный и максимальный элементы:

$$\inf(D_{(20\times20)}) = 0, \quad \sup(D_{(20\times20)}) = 3.73. \qquad (7.11)$$

Ансамбль возможных кластеров имеет объем $K(10) = 184786$. Кластер с номером $k = 256, i_1 = 1, i_2 = 2, i_3 = 3, i_4 = 4, i_5 = 5, i_6 = 6, i_7 = 7, i_8 = 14, i_9 = 15,$

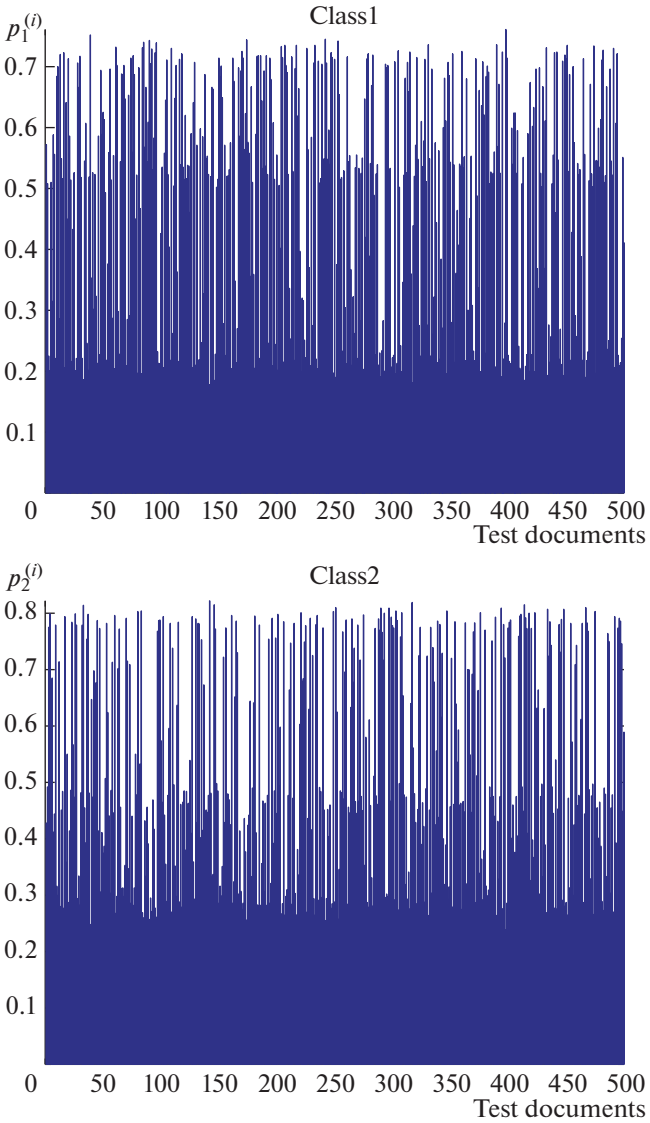

Рис. 6. Вероятности принадлежности классам.

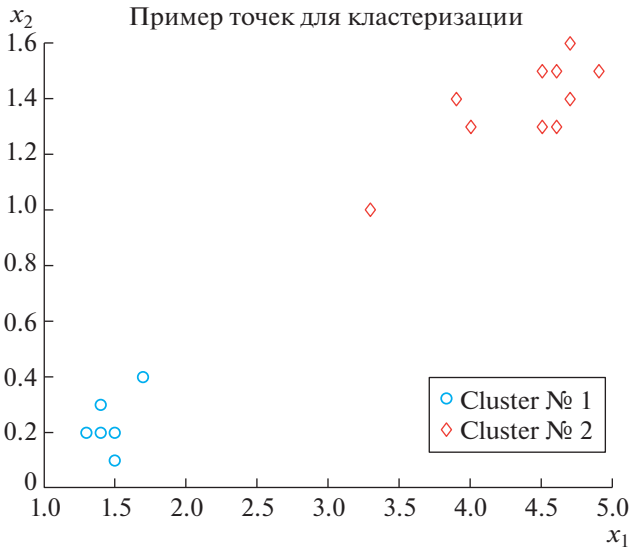

Рис. 7. Изображение точек на двумерном графике.

$i_{10} = 20$. Матрица расстояний $D_{(10)}^{(256)}$ приведена в табл. 5.

Индикатор матрицы $X_{(10)}^{(256)}$, соответствующей кластеру $\mathcal{K}_{(10)}^{(256)}$,

$$\operatorname{dis}(X_{(10)}^{(256)}) = 1.5021. \qquad (7.12)$$

Значения индикаторов для кластеров от $k = 1$ до $k = 184786$ показаны на рис. 8.

Энтропийно-оптимальная функция распределения вероятностей для $s = 10$ имеет вид:

$$p^*(k\,|10) = \frac{\exp(-\lambda^*\operatorname{dis}(X_{(10)}^{(k)}))}{1+\exp(-\lambda^*\operatorname{dis}(X_{(10)}^{(k)}))}, \qquad (7.13)$$
$$\lambda^* = 12.1153.$$

Кластер $\mathcal{K}_1$ с максимальной вероятностью имеет номер

$$k^* = 166922,$$
$$\mathcal{K}_1 = \{4,5,6,7,11,14,15,16,17,20\}, \qquad (7.14)$$
$$\operatorname{dis}\mathcal{K}_1 = 0.1354.$$

Кластер $\mathcal{K}_2$ образует следующие объекты-точки: $\{1,2,3,8,9,10,12,13,18,19\}$.

Расположение кластеров $\mathcal{K}_1$ и $\mathcal{K}_2$ показано на рис. 9.

Сравнивая с рис. 7, видно, что совпадение 10/10, т.е. ошибка кластеризации равна нулю.

Пример 2

Рассмотрим другую матрицу данных из той же базы. Расположение объектов-точек показано на рис. 10.

Построим матрицу расстояний $D_{(20\times20)}$ и определим минимальны и максимальный элементы:

$$\inf(D_{(20\times20)}) = 0, \quad \sup(D_{(20\times20)}) = 3.73. \qquad (7.15)$$

Ансамбль возможных кластеров имеет объем $K(10) = 184\,786$. Значения индикаторов для кластеров от $k = 1$ до $k = 184\,786$ показаны на рис. 11.

Энтропийно-оптимальная функция распределения вероятностей для $s = 10$ имеет вид:

$$p^*(k\,|10) = \frac{\exp(-\lambda^*\operatorname{dis}(X_{(10)}^{(k)}))}{1+\exp(-\lambda^*\operatorname{dis}(X_{(10)}^{(k)}))}, \quad \lambda^* = 100. \qquad (7.16)$$

Кластер $\mathcal{K}_1$ с максимальной вероятностью имеет номер

$$k^* = 177\,570,$$
$$\mathcal{K}_1 = \{5,6,7,8,11,14,15,16,17,20\}, \qquad (7.17)$$
$$\mathrm{dis}\mathcal{K}_1 = 0.4420.$$

Его образуют следующие объекты-точки: 5, 6, 7, 8,11,14,15, 16, 17, 20. Кластер $\mathcal{K}_2$ образуют следующие объекты-точки: $1,2,3,4,9,10,12,13,18,19$.

Расположение кластеров $\mathcal{K}_1$ и $\mathcal{K}_2$ показано на рис. 12. Сравнивая с рис. 7, видно, что совпадение 8/10.

ПРИЛОЖЕНИЕ А:

УСЛОВИЯ ОПТИМАЛЬНОСТИ В ЛЯПУНОВСКИХ ЗАДАЧАХ ОПТИМИЗАЦИИ

Функция $P(\boldsymbol{\theta})$ – непрерывно-дифференцируемая, т.е. принадлежит классу C^1 [29]. Выберем в этом классе произвольную функцию $h(\boldsymbol{\theta})$ и представим ее в виде:

$$P(\boldsymbol{\theta}) = P^*(\boldsymbol{\theta}) + \beta h(\boldsymbol{\theta}),$$

где функция $P^*(\boldsymbol{\theta})$ является решением задачи (3.5)–(3.7), и β – вещественный параметр.

Подставим указанное выше представление функции ПРВ в (3.8). Полагая, что все функции из класса C^1 – фиксированы, получим, что функционал Лагранжа зависит от параметра β. Тогда условия стационарности функционала (3.8) в терминах производной Гато приобретают следующий вид:

$$\left.\frac{d\mathcal{L}}{d\beta}\right|_{\beta=0} = 0.$$

В результате получаем следующее интегральное уравнение:

$$\int_{\mathscr{E}} h(\boldsymbol{\theta})\Omega(\boldsymbol{\theta})d\boldsymbol{\theta} = 0,$$

где

$$\Omega(\boldsymbol{\theta}) = \ln P(\boldsymbol{\theta}) - \mu - \langle \boldsymbol{\lambda}, \boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})\rangle,$$

где вектор-функция $\boldsymbol{\Phi}^s(\mathbf{x}^{(r)}, \boldsymbol{\theta})$ определена в (3.2).

Интегральное уравнение выполняется при любых функциях $h(\boldsymbol{\theta})$ из C^1, если

$$\Omega(\boldsymbol{\theta}) = 0. \qquad (1.1)$$

ПРИЛОЖЕНИЕ В: ДОКАЗАТЕЛЬСТВО ТЕОРЕМ 1, 2

Доказательство теоремы 1. В силу условия *а)* функция $\mathbf{W}(\boldsymbol{\lambda}|\mathbf{y}^{(r)}, \mathbf{x}^{(r)})$, непрерывная по совокупности переменных, порождает в $R^r \times R^r$ векторное поле $\Psi_{(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})}(\boldsymbol{\lambda}) = \mathbf{W}(\boldsymbol{\lambda}|\mathbf{y}^{(r)}, \mathbf{x}^{(r)})$.

Таблица 3. Матрица расстояний для матрицы данных

№	1	2	3	4	5	6	7	8	9	10
1	0	0.1	0.22	3.01	3.45	3.32	3.27	0.61	0.2	0.22
2	0.1	0	0.14	3.1	3.55	3.42	3.36	0.71	0.22	0.2
3	0.22	0.14	0	3.16	3.61	3.48	3.42	0.8	0.22	0.14
4	3.01	3.1	3.16	0	0.45	0.32	0.28	2.42	2.94	3.04
5	3.45	3.55	3.61	0.45	0	0.14	0.2	2.86	3.38	3.48
6	3.32	3.42	3.48	0.32	0.14	0	0.14	2.73	3.26	3.35
7	3.27	3.36	3.42	0.28	0.2	0.14	0	2.68	3.2	3.29
8	0.61	0.71	0.8	2.42	2.86	2.73	2.68	0	0.61	0.71
9	0.2	0.22	0.22	2.94	3.38	3.26	3.2	0.61	0	0.1
10	0.22	0.2	0.14	3.04	3.48	3.35	3.29	0.71	0.1	0
11	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
12	0.22	0.14	0.2	3.23	3.68	3.55	3.49	0.82	0.36	0.32
13	0.54	0.63	0.71	2.47	2.92	2.79	2.73	0.14	0.5	0.6
14	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
15	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
16	3.27	3.36	3.42	0.28	0.2	0.14	0	2.68	3.2	3.29
17	3.31	3.4	3.45	0.36	0.22	0.22	0.1	2.73	3.23	3.32
18	0.4	0.3	0.22	3.38	3.83	3.7	3.64	1	0.45	0.36
19	1.3	1.39	1.46	1.71	2.15	2.02	1.97	0.72	1.24	1.33
20	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38

Таблица 4. Матрица расстояний для матрицы данных (продолжение)

№	1	2	3	4	5	6	7	8	9	10
1	0	0.1	0.22	3.01	3.45	3.32	3.27	0.61	0.2	0.22
2	0.1	0	0.14	3.1	3.55	3.42	3.36	0.71	0.22	0.2
3	0.22	0.14	0	3.16	3.61	3.48	3.42	0.8	0.22	0.14
4	3.01	3.1	3.16	0	0.45	0.32	0.28	2.42	2.94	3.04
5	3.45	3.55	3.61	0.45	0	0.14	0.2	2.86	3.38	3.48
6	3.32	3.42	3.48	0.32	0.14	0	0.14	2.73	3.26	3.35
7	3.27	3.36	3.42	0.28	0.2	0.14	0	2.68	3.2	3.29
8	0.61	0.71	0.8	2.42	2.86	2.73	2.68	0	0.61	0.71
9	0.2	0.22	0.22	2.94	3.38	3.26	3.2	0.61	0	0.1
10	0.22	0.2	0.14	3.04	3.48	3.35	3.29	0.71	0.1	0
11	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
12	0.22	0.14	0.2	3.23	3.68	3.55	3.49	0.82	0.36	0.32
13	0.54	0.63	0.71	2.47	2.92	2.79	2.73	0.14	0.5	0.6
14	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
15	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38
16	3.27	3.36	3.42	0.28	0.2	0.14	0	2.68	3.2	3.29
17	3.31	3.4	3.45	0.36	0.22	0.22	0.1	2.73	3.23	3.32
18	0.4	0.3	0.22	3.38	3.83	3.7	3.64	1	0.45	0.36
19	1.3	1.39	1.46	1.71	2.15	2.02	1.97	0.72	1.24	1.33
20	3.36	3.45	3.51	0.36	0.1	0.1	0.1	2.77	3.29	3.38

Таблица 5. Матрица расстояний для кластера $\mathcal{K}(256)$

№	1	2	3	4	5	6	7	8	9	10
1	0	0.1	0.22	3.01	3.45	3.32	3.27	3.36	3.36	3.36
2	0.1	0	0.14	3.1	3.55	3.42	3.36	3.45	3.45	3.45
3	0.22	0.14	0	3.16	3.61	3.48	3.42	3.51	3.51	3.51
4	3.01	3.1	3.16	0	0.45	0.32	0.28	0.36	0.36	0.36
5	3.45	3.55	3.61	0.45	0	0.14	0.2	0.1	0.1	0.1
6	3.32	3.42	3.48	0.32	0.14	0	0.14	0.1	0.1	0.1
7	3.27	3.36	3.42	0.28	0.2	0.14	0	0.1	0.1	0.1
8	3.36	3.45	3.51	0.36	0.1	0.1	0.1	0	0	0
9	3.36	3.45	3.51	0.36	0.1	0.1	0.1	0	0	0
10	3.36	3.45	3.51	0.36	0.1	0.1	0.1	0	0	0

Таблица 6. Матрица данных

№	x_1	x_2	Вид	№	x_1	x_2	Вид	№	x_1	x_2	Вид
1	6.4	3.2	2	8	5.2	2.7	2				
2	6.5	2.8	2	9	5.7	2.8	2	15	4.9	3.0	1
3	7.0	3.2	2	10	6.6	2.9	2	16	5.0	3.4	1
4	5.4	3.9	1	11	5.1	3.5	1	17	4.9	3.1	1
5	4.7	3.2	1	12	6.3	3.3	2	18	6.9	3.1	2
6	4.6	3.4	1	13	5.5	2.3	2	19	4.9	2.4	2
7	4.6	3.1	1	14	4.4	2.9	1	20	5.0	3.6	1

Выберем в $R^r \times R^r$ произвольный вектор $\mathbf{u}$ и определим векторное поле

$$\Pi_{\mathbf{u}}(\boldsymbol{\lambda}) = \Psi_{(\mathbf{y}^{(r)},\mathbf{x}^{(r)})}(\boldsymbol{\lambda}) - \mathbf{u}.$$

В силу условия b) при фиксированном $\mathbf{u}$ поле $\Pi_{\mathbf{u}}(\boldsymbol{\lambda})$ не имеет нулей на сферах $\|\boldsymbol{\lambda}\| = \varrho$ достаточно большего радиуса ϱ. Поэтому на сферах $\|\boldsymbol{\lambda}\| = \varrho$ определено *вращение* [30].

Рассмотрим два векторных поля

$$\Pi_{\mathbf{u}^{(1)}}(\boldsymbol{\lambda}) = \Psi_{(\mathbf{y}^{(r)},\mathbf{x}^{(r)})}(\boldsymbol{\lambda}) - \mathbf{u}^{(1)},$$
$$\Pi_{\mathbf{u}^{(2)}}(\boldsymbol{\lambda}) = \Psi_{(\mathbf{y}^{(r)},\mathbf{x}^{(r)})}(\boldsymbol{\lambda}) - \mathbf{u}^{(2)}.$$

Эти векторные поля *гомотопны* [30] на сферах достаточно большего радиуса. Поэтому поле

$$\Omega(\boldsymbol{\lambda}) = \alpha\Pi_{\mathbf{u}^{(1)}}(\boldsymbol{\lambda}) + (1-\alpha)\Pi_{\mathbf{u}^{(2)}}(\boldsymbol{\lambda}) =$$
$$= \Psi_{(\mathbf{y}^{(r)},\mathbf{x}^{(r)})}(\boldsymbol{\lambda}) - [\alpha\mathbf{u}^{(1)} + (1-\alpha)\mathbf{u}^{(2)}]$$

не имеет нулей на сферах достаточно большого радиуса при любых $\alpha \in [0,1]$. Гомотопные поля имеют одинаковые вращения:

$$\gamma(\Pi_{\mathbf{u}^{(1)}}(\boldsymbol{\lambda})) = \gamma(\Pi_{\mathbf{u}^{(2)}}(\boldsymbol{\lambda})).$$

На сферах достаточно большого радиуса векторные поля $\Pi_{\mathbf{u}^{(1)}}(\boldsymbol{\lambda}), \Pi_{\mathbf{u}^{(2)}}(\boldsymbol{\lambda})$ невырожденные, но в шаре $\|\boldsymbol{\lambda}\| \le \varrho_1 < \varrho$ каждое из них может иметь некоторое количество особых точек. Обозначим $\kappa(\mathbf{u}^{(1)})$ и $\kappa(\mathbf{u}^{(2)})$ – количество особых точек векторных полей $\Pi_{\mathbf{u}^{(1)}}(\boldsymbol{\lambda})$, $\Pi_{\mathbf{u}^{(2)}}(\boldsymbol{\lambda})$ соответственно. Поскольку векторные поля гомотопны, то

$$\kappa(\mathbf{u}^{(1)}) = \kappa(\mathbf{u}^{(2)}) = \kappa.$$

В силу условия (3.15) эти особые точки изолированы.

Воспользуемся теперь понятием *индекса* особой точки:

$$\mathrm{ind}(\boldsymbol{\lambda}^0) = (-1)^{\beta(\boldsymbol{\lambda}^0)},$$

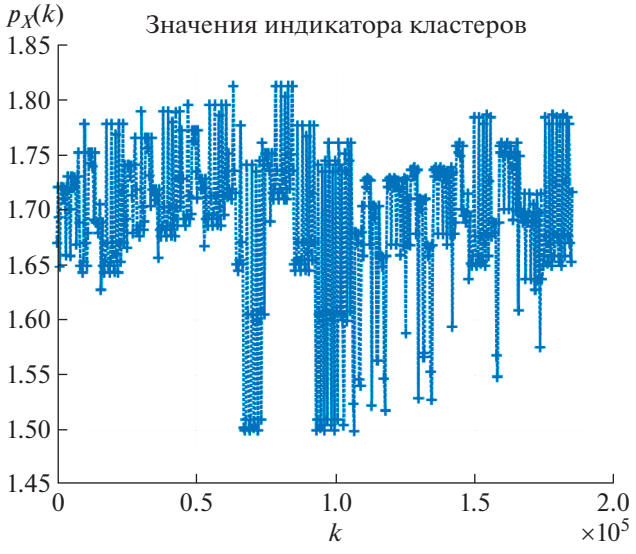

Рис. 8. Значения индикаторов для $k \in [1,184\,786]$.

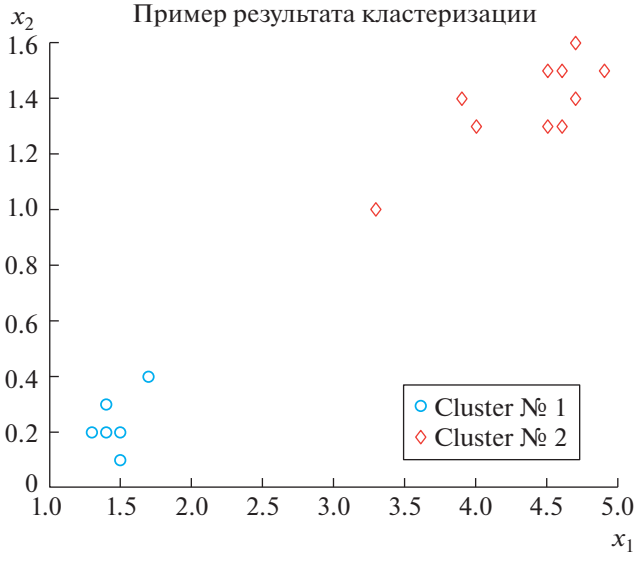

Рис. 9. Результат рандомизированной кластеризации.

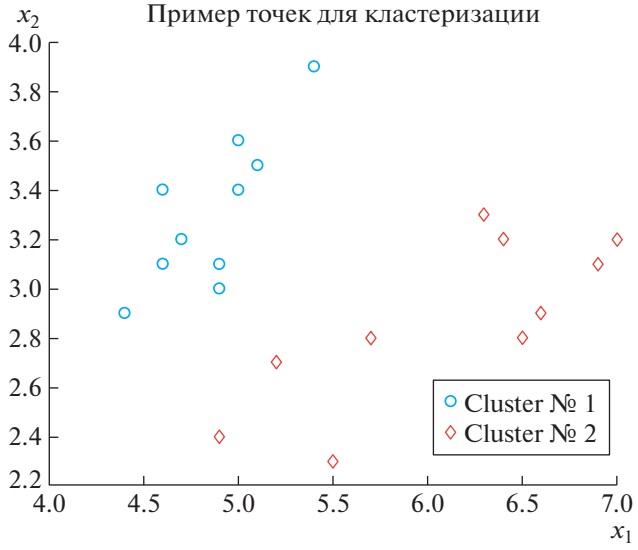

Рис. 10. Изображение точек на двумерном графике.

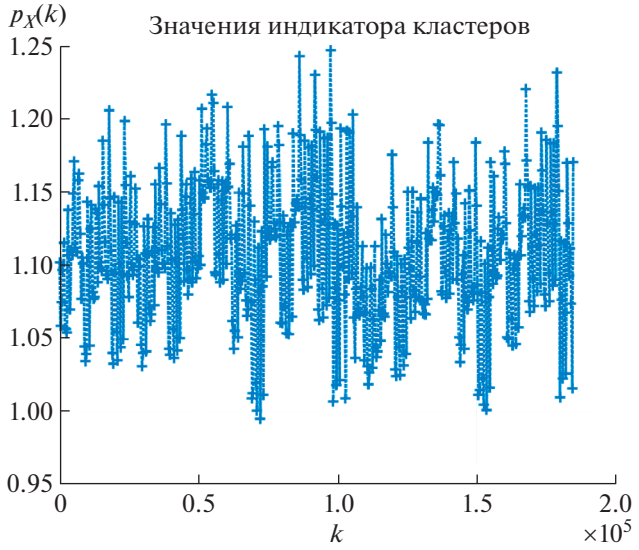

Рис. 11. Значения индикаторов для $k = [1,184\,786]$.

где $\beta(\boldsymbol{\lambda}^0)$ – количество собственных чисел матрицы $\Pi'_{\mathbf{u}}(\boldsymbol{\lambda}^0) = J_{\lambda}(\boldsymbol{\lambda}^0|, \mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ с отрицательной вещественной частью. Из определения индекса видно, что на его значение влияет четность величины $\beta(\boldsymbol{\lambda}^0)$, а не ее абсолютное значение. В силу условия (3.15) четность всех особых точек одинаковая. Действительно, поскольку $J_{\lambda}(\boldsymbol{\lambda}^0|\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) \neq 0$, то при любых $(\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) \in R^r \times R^r$ собственные числа матрицы $J_{\lambda}(\boldsymbol{\lambda}^0|\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ могут переходить из левой полуплоскости в правую только парами: вещественные числа трансформируются в пару комплексно-сопряженных, которые затем переходят мнимую ось.

Учитывая это обстоятельство, получим, что вращение гомотопных полей

$$\gamma(\Pi_{\mathbf{u}}) = \kappa(-1)^{\beta},$$

где β – количество собственных чисел матрицы $\Pi'_{\mathbf{u}}(\boldsymbol{\lambda})$ для какой-нибудь особой точки.

Покажем, что векторное поле $\Pi_{\mathbf{u}}(\boldsymbol{\lambda})$ имеет единственную особую точку в шаре $\|\boldsymbol{\lambda}\| \leq \rho_1 < \rho$. Рассмотрим уравнение

$$\Pi_{\mathbf{u}}(\boldsymbol{\lambda}) = \Phi_{(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})}(\boldsymbol{\lambda}) - \mathbf{u} = 0.$$

Пусть это уравнение при каждой фиксированной паре $\mathbf{x}^{(r)}$, $\mathbf{y}^{(r)}$ имеет κ особых точек, т.е. функций $\boldsymbol{\lambda}^{(1)}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)}), \ldots, \boldsymbol{\lambda}^{(\kappa)}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$. Следовательно, оно определяет многозначную функцию $\boldsymbol{\lambda}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$, κ ветвей которой изолированы (что следует из изолированности особых точек). Каждая из ветвей $\boldsymbol{\lambda}^{(i)}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ определяет в $R^r \times R^r$ открытое множество (в силу условия b), а

$$\bigcup_{i=1}^{\kappa} \boldsymbol{\lambda}^{(i)}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) = R^r \times R^r.$$

Это возможно только тогда, когда $\kappa = 1$. Следовательно, для каждой пары $\mathbf{x}^{(r)}, \mathbf{y}^{(r)}$ из $R^r \times R^r$ существует единственная функция $\boldsymbol{\lambda}^*(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$, обращающая в ноль функцию $\mathbf{W}(\boldsymbol{\lambda}|\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$. ■

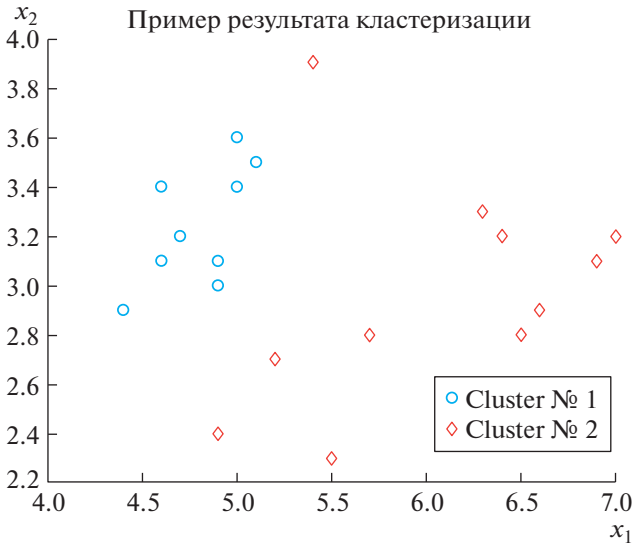

Рис. 12. Результат рандомизированной кластеризации.

Доказательство теоремы 2. Из (3.15) и условия *а)* теоремы 1 следует, что функция $\mathbf{W}(\lambda|\mathbf{x}^{(r)},\mathbf{y}^{(r)})$ – аналитическая по совокупности переменных. Поэтому левую часть уравнения (4.5) можно представить обобщенным рядом Тейлора [31], и строить решение также в виде обобщенного ряда Тейлора. Степенные блоки этого ряда определяются с помощью рекуррентной процедуры. ■.

ПРИЛОЖЕНИЕ C: ДОКАЗАТЕЛЬСТВО ЛЕММЫ 1

Доказательство. Элементы обратной матрицы

$$a_{ik}^{(-1)} = \frac{A_{ki}}{\det A}, \quad (k,i) = \overline{1,r}$$

в силу условия леммы 1 ограничены, т.е.

$$a_{ik}^{(-1)} \le M < \infty, \quad \left\|A^{-1}\right\| < \infty.$$

Следовательно, существует константа $\alpha > 1$, для которой выполняется неравенство (3.23). ■

ПРИЛОЖЕНИЕ D: ДОКАЗАТЕЛЬСТВО ЛЕММЫ 2

Согласно неравенству (3.10) имеем

$$\left\|\frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}}\right\| \le \left\|\left[\frac{\partial \mathbf{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \left\|\frac{\partial \mathbf{\Theta}}{\partial \mathbf{x}^{(r)}}\right\|,$$

$$\left\|\frac{\partial \mathbf{z}}{\partial \mathbf{y}^{(r)}}\right\| \le \left\|\left[\frac{\partial \mathbf{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \left\|\frac{\partial \mathbf{\Theta}}{\partial \mathbf{y}^{(r)}}\right\|.$$

Полагая, что матрица $\left[\frac{\partial \mathbf{\Theta}}{\partial \mathbf{z}}\right]$ невырожденная (следует из невырожденности Якобиана (3.15)), следуя лемме 1 и учитывая (3.26), норму обратной матрицы можно оценить следующим образом:

$$\left\|\left[\frac{\partial \mathbf{\Theta}}{\partial \mathbf{z}}\right]^{-1}\right\| \le \alpha \left\|\left[\frac{\partial \mathbf{\Theta}}{\partial \mathbf{z}}\right]\right\|^{-1} = \frac{\alpha}{r\theta}.$$

При выполнении условий леммы 2

$$\left\|\frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}}\right\| \le \frac{\alpha\rho}{r\theta}, \quad \left\|\frac{\partial \mathbf{z}}{\partial \mathbf{y}^{(r)}}\right\| \le \frac{\alpha\omega}{r\theta}.$$

Отсюда видно, что при росте объема выборки ($r \to\to \infty$) нормы соответствующих якобианов стремятся к нулю. Поэтому $\tilde{\mathbf{z}} = \mathbf{z}(\mathbf{x}^{\infty},\mathbf{y}^{\infty})$. ■

ПРИЛОЖЕНИЕ E: ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 3

Рассмотрим вспомогательную систему дифференциальных уравнений, получаемую из (5.24) при $\gamma \to 0$:

$$\frac{d\lambda_i}{dt} = \lambda_i L_{\lambda_i}(p^*(s,k|\lambda_1,\lambda_2)), \quad i = 1,2.$$

На первом этапе доказывается ее устойчивость "в целом", т.е. при любых начальных отклонениях из R_+^2.

На втором этапе показывается, что алгоритм (5.24) представляет собой схему Эйлера для вспомогательных дифференциальных уравнений.

Рассмотрим некоторые детали доказательства и определим на R_+^2 функцию:

$$V(\lambda_1,\lambda_2) = \sum_{i=1}^{2}(\lambda_i - \lambda_i^*) - \lambda_i^*(\ln\lambda_i - \ln\lambda_i^*).$$

Функция $V \ge 0$ строго выпукла на R_+^2, принимает минимальное значения в точке $(\lambda_1^*,\lambda_2^*)$, ее гессиан в этой точке невырожден. Определим ее производную по времени в силу исходных уравнений:

$$\frac{dV}{dt} = -\lambda_1^* L_{\lambda_1}(p^*(s,k)|\lambda_1,\lambda_2) - \\ - \lambda_2^* L_{\lambda_2}(p^*(s,k)|\lambda_1,\lambda_2).$$

Отсюда имеем:

$$\frac{dV}{dt} = \begin{cases} < 0, & \text{если } \lambda_1^* > 0,\ \lambda_2^* > 0 \\ = 0, & \text{если } \lambda_1^* = \lambda_2^* = 0 \end{cases}$$

Следовательно, функция V является функцией Ляпунова для вспомогательных дифференциальных уравнений на пространстве R_+^2. Все решения этих уравнений асимптотически устойчивы из любых начальных условий $\lambda_1^0 > 0$, $\lambda_2^0 > 0$.

Алгоритм (5.24) представляет собой разностную схему Эйлера. В силу асимптотической устойчивости решений дифференциальных уравнений всегда существуют шаг $\gamma > 0$ и область начальных условий, для которых схема Эйлера сходится. ■

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование поддержано Министерством науки и высшего образования РФ, проект № 075-15-2020-799.

СПИСОК ЛИТЕРАТУРЫ

1. *Boltzmann L.* Vorlesungen uber Gastheory. Leipzig, 1896. V. 1, J.A. Barth; 1898. V. 2, J.A. Barth.
2. *Jaynes E.T.* Gibbs vs Boltzmann entropy. American Journal of Physics. 1965. V. 33. P. 391–398.
3. *Shannon C.E.* Mathematical Theory of Communication. 1948, The Bell System Technical Journal. V. 27. P. 373–423, 623–656, July, Oct.
4. *Renyi A.* Probability Theiry, North Holland, Amsterdam, 1970.

5. *Bashkirov A.G.* Renyi-entropy is statistical entropy for a large scale systems, Theoretical and Mathematicol Physics., 2006. V. 149. № 2. P. 299–317.
6. *Jaynes E.T.* Information theory and statistical Mechanics. Physical Review, 1957. V. 104(4). P. 620–630.
7. *Rosenkrantz R.D., Jaynes E.T.* Paper on Probability, Statistics,and Statistical Physics. Kluwer Academic Pablishers, 1989.
8. *Jaynes E.T.* Probability theory: the logic of science. Cambridge Uni. Press, 2003.
9. *Schulz K.F., Grimes D.A.* Generation of allocation sequences in randomized trials. The Lancet, 2002, Feb. 09. https://doi.org/10.1016/S0140-6736(02)07683-3
10. *Maravelakis P.* The use of Statistics in Social Sciences. Journal of Humanities and Applied Social Sciences. 2019. V. 1. № 2. P. 87–97. https://doi.org/10.1108/JHASS-08-2019-0038
11. *Bonaccorsi A., Cicero T., Ferraro A., Malgarini M.* Journal raiting as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise. F1000. 2015, 4:196. P. 1–12.
12. *Gauvin L., Genois M., Karsal M., Kivela M., Takaguchi T., Valdano E., Vesterguard C.L.* Randomized reference models for temporal network. 2020, arXiv: 1806.04032v3[physics.soc-ph], 11 Feb.
13. *Vovk V., Shafer G.* Good randomized sequential probability forecasting is always possible. Journal of Royal Statistical Society B. 2005. V. 65. Pt. 5. P. 747–763.
14. *Vyugin V.* On calibration error of randomized forecasting algorithm. theoretical Computer Science. 2009. V. 410. P. 1781–1795.
15. *Zhao S., Ma T., Ermon S.* Indsividual Calibration with Randomized Forecasting. 2020, arXiv:2006.10288v3 [stat.ML]9 Sept 2020.
16. *Mancini T., Calvo-Pardo H., Olmo J.* Extremly randomized neural networks for constructing prediction interval. Neural Networks. 2021. V. 144. P. 113–128.
17. *Motwani R., Raghaven P.* Randomized Algorithms. Cambridge Uni. Press, NY, 1995.
18. *Vidyasagar M.* Randomized Algorithms for robust controller synthesis using statistical learning theory. Automatica. 2001. V. 37(10). P. 1515–1528.
19. *Tempo R., Calafiory G., Dabbene F.* Randomized Algorithms for Analysis and Control of Uncertain Systems, Springer, 2013.
20. *Granichin O., Volkovich Z., Toledano-Kitai D.* Randomized Algorithms in Automatic Control and Data Mining. Springer, 2015.
21. *Osborn M.J., Rubinstein A.* A Course in Game Theory. 1994, Cambridge, MA, MIT Press.
22. *Borovkov A.A.* Mathematical Statistics. CTC Press, 1999.
23. *Larsen R.J., Marx M.L.* An Introduction to Mathematical Statistics. 2012, Prentice Hall.
24. *Avellaneda M.* Minimum relative-entropy calibration of asset-pricing model. Journal of theoretical and applied finance. 1998. V. 1(04). P. 447–472.
25. *Popkov A.Yu., Popkov Yu.S.* New Methods of Entropy-Robust Estimation for Randomized Models under Limited Data // Entropy. 2014. № 16. P. 675–698. https://doi.org/10.3390/e16020675
26. *Ioffe A.D., Tikhomirov V.M.* Theory of extremal problems. Elsevier North Holland. 1979. P. 456.
27. *Alexeev V.M., Tikhomirov V.M., Fomin S.V.* Optimal Control. Springer-Verlag, Boston, MA. 1987. P. 277.
28. *Voevodin V.V., Kuznetzov Yu.A.* Matrices and Computing. Nauka, Moscow, 1984.
29. *Kaashoek M.A., van der Mee Cornelis* Recent Advances in Operator Theory and Its Applications, Springer Science & Business Media, 2006. V. 160.
30. *Parr R.G., Yang W.* Density-functional Theory of Atoms and Molecules. NY, Oxford Uni Press. 1989.
31. *Coulson T. et al.* Estimating the functional form for the density dependence from life history data. Ecology. 2008. V. 89(6). P. 1661–1674.
32. *Beckman M.* On the distribution of urban rent and residential density. Journal of Economic Theory, 1969. V. 1. P. 60–67.
33. *Abrams P.A.* Determining the functional form of density Dependence: Deduction Approaches for Cosumer-Resource Systems. American Naturalist. 2009. V. 179. № 3. P. 321–330.
34. *Hyvarinen A.* Estimation of unnormalized statistical models by score matching. Journal of Machine Learning Research. 2005. V. 6. P. 695–709.
35. *Gutman M.U., Hyvarinen A.* Noise-constrastive estimation of unnormalized statistical models? with applications to natural image statistics. Journal of Machine Learning Research. 2012. V. 13. P. 307–361.
36. *Matsuda T., Uehara M., Hyvarinen A.* Information criteria for non-normalized models. Journal of Machine Learning Research. 2021. V. 22. P. 1–33.
37. *Vapnik V.N.* Statistical Learning Theory. Wiley, 1998. 768 p.
38. *Witten I.H., Eibe F.* Data Mining: Practical machine learning tools and tehniques. Morgan Kaufmann, 2005.
39. *Bishop C.M.* Pattern Recognition and Machine Learning. Springer, Series: Information Theory and Statistics, 2006.
40. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer Series in Statistics, Berlin, 2009. 698 p.
41. *Воронцов К.В.* Математические методы обучения по прецендентам. Курс лекций МФТИ. 2013.
42. *Попков Ю.С., Попков А.Ю., Дубнов Ю.А.* Рандомизированное машинное обучение при ограниченных объемах данных. УРРС, 2018.
43. *Bruckstein A.M., Donoho D.L., Elad M.* From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM review. 2009. V. 51 (1). P. 34–81.
44. *Jolliffe I.* Principal component analysis. New York: Springer. 2011. https://doi.org/10.1007/b98835
45. *Поляк Б.Т., Хлебников М.В.* Метод главных компонент: робастные версии. Автоматика и телемеханика. 2017. V. 3. P. 130–148.
46. *Bingham E., Mannila H.* Random projection in dimensionality reduction: applications to image and text data. Proceedings of the seventh ACM SIGKDD interna-

tional conference on Knowledge discovery and data mining. 2001. P. 245–250.

47. *Vempala S.S.* The random projection method. *American Mathematical Soc.* 2005. V. 65.
48. *Popkov Y.S., Dubnov Y.A., Popkov A.Y.* Entropy Dimension Reduction Method for Randomized Machine Learning Problems. *Automation and Remote Control.* 2018. V. 79(11). P. 2038–2051. https://doi.org/10.1134/S0005117918110085
49. *Kullback S., Leibler R.A.* On information and sufficiency. The annals of mathematical statistics. 1951. V. 22(1). P. 79–86.
50. *Попков Ю.С., Попков А.Ю.* Кросс-энтропийная оптимальная редукция размерности матрицы данных с ограничением информационной емкости. Доклады академии наук. 2019. V. 488. P. 21–23. https://doi.org/10.31857/S0869-5652488121-23
51. *Magnus J.R., Neudecker H.* Matrix differential calculus with applications in statistics and econometrics. Wiley. 1988.
52. *Попков Ю.С.* Оценки максимальной энтропии непрерывных функций и их асимптотическая эффективность. Доклады Академии наук в печати.
53. *Иоффе А.Д., Тихомиров В.М.* Теория экстремальных задач. М.: Наука. 1974.
54. *Popkov Yu.S.* Macrosystems theory and its applications (Lecture notes in control and information sciences) 203. Springer. 1995.
55. *Goldberger A.S.* A Course in Econometrics. Harvard University Press, 1991. 437 p.
56. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М., Юнити, 1998.
57. *Лагутин М.Б.* Наглядная математическая статистика. М., БИНОМ, Лаборатория знаний, 2013.
58. *Rousses G.* A Course of the Mathematical Statistics. Academic Press Inc. 2015. 600 p.
59. *Golan A., Judge G.G., Miller D.* Maximum Entropy Econometrics: Robust Estimation with Limited Data. John Wiley and Sons Ltd., Chichester, UK, 1996.
60. *Мандель И.Д.* Кластерный анализ. М., Финансы и Статистика, 1988.
61. *Загоруйко Н.Г.* Когнитивный анализ данных. Академическое издательство “ГЕО”, 2012. 203 с.
62. *Загоруйко Н.Г., Барахнин В.Б., Борисова И.А., Ткачев Д.А.* Кластеризация текстовых документов из электронной базы публикаций алгоритмом FRiS-Tax. Вычислительные технологии. 2013. Т. 6. № 18. С. 62–74.
63. *Jain A., Murty M., Flynn P.* Data clastering: a review. ACM Computing surveys. 1990. V. 31. № 3. P. 264–323.
64. *Воронцов К.В.* Лекции по алгоритмам кластеризации имногомерному шкалированию. http: // www.ccas.ru/voron/download/Clastering.pdf
65. *Lescovec J., Rajaraman A., Ullman J.* Mining of Massive Datasets. Cambridge University Press, 2014. 511 p.
66. *Deerwester S., Dumias S.T., Furnas G.W., Landauer T.K., Harshman R.* Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 1999. V. 41. P. 391–407.
67. *Zamir O.E.* Clastering Web Documents: A Phrase-Based Methods for Grouping Search Engine Results. Uni. of Washington, USA, 1999. P. 65–117.
68. *Cao G., Song D., Bruza P.* Suffix-Tree Clastering on Post-retrieval Documents Information. The Uni. of Queensland, 2003.
69. *Jain A., Dubs R.* Clastering Methods and Algorithms. Prentice-Hall Inc. 1988.
70. *Pal N.R., Biswas J.* Claster validation using graph theoretic consept. Pattern Recognition. 1997. V. 30(6). P. 847–857.
71. *Halkidi M., Batistakis Y., Vazirgiannis M.* On clastering validation techniques. Journal of Intelligent Information Systems. 2001. V. 17(2-3). P. 107–145.
72. *Han J., Kamber M., Pei J.* Data Mining Consept and Techniques. Third Edition, Morgan Kaufmann Publishers, 2012, 703 p.
73. *Popkov Yu.S.* Macrosystem theory and application. Springer, LNIS, 1995.
74. *Aggarval A.* Neural Networks and Deap Learning. Springer International Publishing AG, part os Springer Nature, 2018.
75. *Popkov Yu.S.* New class of multiplicative algorithms for solving of entropy-linear programs. European Journal of Operation Research. 2006. № 174. P. 1368–1379.
76. *Polyak B.T.* Introduction to Optimization, Optimization Software, 1987.
77. *Johnson W.B., Lindenstrauss J.* Extensions of Lipshitz mapping into Hilbert Space. Modern Analysis and Probability. 1984. V. 26. P. 189–206.
78. *Achlioptas D.* Database-friendly random projections. Amer. Math. Soc., 2001, PODS’01. P. 274–281.
79. *Malout R.* A comparison of algorithms for maximum entropy parameters estimation. Proc. of the 6th conference on Natural Language learning, 2002. V. 20. P. 1–7.
80. *Borwein J., Choksi R., Marechal P.* Probability Distribution of Assets Inferred from Option Prices via Principle Maximum Entropy. SIAM Journal of Optimization, 2003. V. 14. № 2. P. 464–478.
81. *Golan A.* Information andentropy econometrics – a review and synthesis. Foundations and trends in econometrics. 2008. V. 2(1-2). P. 1–145.
82. *Csiszar I., Matus F.* On minimization of entropy functionals under moment constraints. IEEE International Symposium on Information Theory, 2008.
83. *Loubes J-M., Rochet* Approximate maximum entropy on the mean for instrumental variable regression. Statistics and Probability Letters, 2012. V. 82. Issue 5. P. 972–978.
84. *Borwein J.M., Lewis A.S.* Partially-Finite Programming in L_1 and Existance of Maximum Entropy Estimates. SIAM Journal of Optimisation, 1993. V. 3. № 2. P. 248–267.
85. *Burg J.P.* The relationship between maximum entropy spectra and maximum likelyhood spectra. Geophysics. 1972. V. 37. P. 375–376.
86. *Christakos G.* A Bayesian/maximum entropy viewto the spatial estimation problem. Mathematical Geology, 1990. V. 22. P. 763–777.
87. *Singh V.P., Guo H.* Parameter estimation for 3-parameter generalized Pareto distribution by the principle

maximum entropy. Hydrological Sciences Journal, 1994. P. 165–181.

88. *Popkov Yu.S., Dubnov Yu.A., Popkov A.Yu.* Randomized machine learning: Statement, solution, applications // IEEE 8th International Conference on Intelligent Systems (IS) (2016). https://doi.org/10.1109/IS.2016.7737456.
89. *Kolmogorov A.N., Fomin S.V.* Elemets of the Theory of Functions and Functional Analysis. Dover Publication, 1999.
90. *Krasnosel'skii M.A., Zabreiko P.P.* Geometrical Methods of Nonlinear Analysis. A Series of Comprehensive Studies in Mathematics, 263, Springer-Verlag, 1984, Berlin – Heldenberg – NY, 490 p.
91. *Riordan B., Verbula D., McGruire A.D.* Shrinking ponds in subarctic Alaska based on 1950-2002 remotely sensed images. Journal of Geophysic Researches. 2006. V. 111, G04002.
92. *Kirpotin S., Polishchuk Y., Bruksina N.* Abrupt changes of thermokarst lakes in Western Siberia: impacts of climatic warming on permafrost melting. International Journal of Environmental Studies. 2009. V. 66. № 4. P. 423–431.
93. Электронный ресурс: https://cloud.uriit.ru/index.php/s/0DOrxL9RmGqXsV0.

RANDOMIZATION AND ENTROPY IN MACHINE LEARNING AND DATA PROCESSING

Academician of RAS Yu. S. Popkov[a]

[a] *Federal Research Center "Informatics Russian Academy of Sciences", Moscow, Russian Federation*

Combining the concept of randomization with entropic criteria allows solutions to be obtained in the conditions of maximum uncertainty, which is very effective in machine learning and data processing. The application of this approach for the Data-based, pyo-randomized evaluation of functions, randomized "hard" and "soft" machine learning, object clustering, data matrix dimension reduction. Some applications of the task of classification, forecasting the electric load of the energy system, randomized clustering of biological objects are considered.

Keywords: entropy, randomization, machine learning, data processing, parameterization of models, estimates of conditional maximum entropy, balance equations, classification, clustering, random ensemble generation