Информатика и её применения

Том 17 Выпуск 2 Год 2023

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 20.06.23. Подписано в печать 29.06.23. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 12,79. Уч.-изд. л. 12,0. Тираж 100 экз.
Заказ № 1193
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 123103, наб. Новикова-Прибоя, д. 3, корп. 2

О ЗАДАЧАХ ОПТИМИЗАЦИИ, ВОЗНИКАЮЩИХ ПРИ ПРИМЕНЕНИИ ТОПОЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ К ПОИСКУ АЛГОРИТМОВ ПРОГНОЗИРОВАНИЯ С ФИКСИРОВАННЫМИ КОРРЕКТОРАМИ*

И. Ю. Торшин[1]

Аннотация: Корректирующие операции (корректоры) в мультиалгоритмических конструкциях алгебраического подхода могут строиться на основе известных физических моделей и/или многоуровневых описаний физических объектов. В рамках топологического подхода к анализу плохо формализованных задач поиск включаемых в корректор алгоритмов может рассматриваться как задача комбинаторной оптимизации либо как задача минимизации некой функции потерь. Исследование окрестностей цепей в решетке подмножеств объектов позволило получить ряд критериев ранговой оптимизации, перспективных для решения задач прогнозирования числовых таргетных переменных. Формализм апробирован на задаче взаимодействия лиганд–рецептор в рамках хемокиномного анализа лекарств (данные ProteomicsDB). Наилучшие результаты прогнозирования констант EC_{50} наблюдались именно при использовании полученных ранговых критериев: при усреднении по 300 биологическим активностям коэффициент корреляции на контроле составил $0{,}86 \pm 0{,}20$.

Ключевые слова: топологический анализ данных; теория решеток; задачи оптимизации; регрессия; хемоинформатика

DOI: 10.14357/19922264230201 **EDN:** IGSPEW

1 Введение

В рамках алгебраического подхода исследуются конструкции вида $\hat{A}_{(\theta_A)} = \hat{D}_{(\theta_D)} \circ \hat{C}_{(\theta_C)} \circ \hat{B}_{(\theta_B)}$, где $\hat{B}$ — распознающий оператор; $\hat{C}$ — корректирующая операция (корректор); $\hat{D}$ — решающее правило; θ_D, θ_C, θ_B и $\theta_A = (\theta_D, \theta_C, \theta_B)$ — векторы параметров [1]. Алгоритмы $\hat{A}_{(\theta_A)}$ применяются к входной *матрице информации* $M_{\text{вх}}$ для получения выходной *информационной матрицы* $M_{\text{вых}}$, причем в случае корректного алгоритма $M_{\text{вых}} = \hat{A}_{(\theta_A)} M_{\text{вх}}$. Обучение алгоритма по множеству прецедентов $Q = (M_{\text{вх}}, M_{\text{вых}})$ рассматривается как способ вычисления вектора $\theta_A(Q)$. Алгоритм, обученный по Q, *ε-корректен относительно контрольного* $Q' = (M'_{\text{вх}}, M'_{\text{вых}})$, если $L(M'_{\text{вых}}, \hat{A}_{(\theta_A(Q))} M'_{\text{вх}}) \le \varepsilon$, где L — та или иная функция потерь. Для оценки обобщающей способности используются разнообразные комбинаторные функционалы [2, 3].

Важным направлением исследований в научной школе Ю. И. Журавлёва – К. В. Рудакова стало изучение разрешимости и регулярности задач, где множества прецедентов $Q \subset I_i \times I_f$ определены над множествами *начальных* (I_i) и *конечных информаций* (I_f) [4]. При поиске ε-корректных алгоритмов $\hat{A}_{(\theta_A)}$ для решения разрешимых задач утверждается возможность настройки векторов θ_C и θ_B с получением ε-корректного алгоритма при произвольном $\hat{D}_{(\theta_D)}$ [1].

В некоторых задачах возможна фиксация корректора $\hat{C}$ в соответствии с проблемной областью. Например, в физической химии и в биохимии для описания взаимодействия лиганд–рецептор используется уравнение Хилла–Ленгмюра, в котором фигурирует концентрация лиганда C:

$$\frac{E_{\max}}{E(C)} = 1 + \left(\frac{\text{EC}_{50}}{C}\right)^s, \tag{1}$$

где E — измеряемое в эксперименте значение отклика, оценивающего связывание (например, интенсивность свечения флуоресцентной метки); $E_{\max}$ — максимальное значение отклика, EC_{50} — константа процесса (концентрация вещества, вызывающая 50% от максимального отклика, т. е. точка перегиба S-образной кривой $E(C)$). Выражение (1) — термодинамическое описание равновес-

*Работа выполнена при поддержке гранта РНФ (проект № 23-21-00154) с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, tiy135@yahoo.com

ного связывания лиганда L рецептором R в соответствии с уравнением квазихимической реакции

$$R + nL \leftrightarrow RL_n.$$

Приведение (1) к виду

$$\ln\left(\frac{E_{\max}}{E(C)} - 1\right) = -s\ln(C) + s\ln(\mathrm{EC}_{50}),$$

где s — коэффициент Хилла (угол наклона касательной в точке EC_{50}), указывает на возможность линейной аппроксимации вида

$$y_i(x_i) = bx_i + a\,,\ \ b = -s\,,\ \ a = s\ln(\mathrm{EC}_{50}),$$

что позволяет вычислять значения констант EC_{50} и s для n_e экспериментальных точек (например, методом наименьших квадратов):

$$b = \frac{n_e\sum_{i=1}^{n_e} y_i x_i - \sum_{i=1}^{n_e} y_i \sum_{i=1}^{n_e} x_i}{n_e\sum_{i=1}^{n_e} x_i^2 - \left(\sum_{i=1}^{n_e} x_i\right)^2};$$

$$a = \frac{1}{n}\left(\sum_{i=1}^{n_e} y_i - b\sum_{i=1}^{n_e} x_i\right).$$

Предположим, что молекулы лигандов заданы в виде множества хемогра́фов $\{G_j\}$, $j = 1,\ldots,N$, а для набора концентраций $\{C_i\}$, $i = 1,\ldots,n_e$, из физико-химических экспериментов получены значения $E_j(C_i)$ и вычислены значения констант $\mathrm{EC}_{50}(j)$. Если для такого набора данных можно построить алгоритмы хемометрического анализа [5, 6], которые на основании G_j позволяют прогнозировать $E_j(C_i)$, то выражение (1) может быть использовано как фиксированный «физический» корректор $\hat{C}$ алгоритма $\hat{A}$ при $D \equiv 1$. Для реализации этого «имитационного» алгоритма прогнозирования необходимо построить ε-корректные распознающие операторы $\hat{B}_{(\theta_{B_i}),i}$, $|E_j(C_i) - \hat{B}_{(\theta_{B_i}),i}G_j| \le \varepsilon$ [5].

Схема порождения алгоритма $\hat{A} = \hat{D}\circ\hat{C}\circ\hat{B}$ может использоваться не только для решения «финальных» задач $Z(M_{\text{вх}}, M_{\text{вых}})$, но и для важных промежуточных задач, таких как порождение более информативных «синтетических» признаковых описаний объектов (например, вида $\hat{B}_{(\theta_{B_i}),i}G_j$). В результате получаются вложенные алгоритмические структуры, которые могут описываться алгоритмами α-го уровня:

$$\hat{A}^{(\alpha)}_{(\theta^{\alpha}_{A})} = \hat{D}^{(\alpha)}_{(\theta^{\alpha}_{D})}\circ\hat{C}^{(\alpha)}_{(\theta^{\alpha}_{C})}\circ\hat{B}^{(\alpha)}_{(\theta^{\alpha}_{B})}.$$

Построение операторов $\hat{B}_{(\theta_{B_i}),i}$, $\hat{B}^{(\alpha}_{(\theta^{\alpha}_{B})}$, $\hat{C}^{(\alpha)}_{(\theta^{\alpha}_{C})}$ и др. целесообразно осуществлять в рамках топологического подхода к анализу данных [4, 7].

2 Исследование окрестностей цепей в решетке подмножеств объектов

В рамках топологического подхода алгоритм прогнозирования k-й таргетной переменной (например, представленный распознающим оператором вида $\hat{B}_{(\theta_{B_k}),k}$) находится в результате перебора цепей решетки в окрестности цепи, заданной k-й переменной [5], который может рассматриваться (1) как задача комбинаторной оптимизации (поиск множеств, формирующих соответствующую цепь решетки) или (2) задача минимизации того или иного функционала или «функции потерь» (невязка и др.). При задании метрики на элементах решетки (ρ_L) и определении расстояния между цепями (ρ_A) рассмотрим возможность сведения задачи комбинаторной оптимизации к задаче минимизации особой формы функционала (так называемая ранговая оптимизация).

Основы разрабатываемого топологического формализма изложены в [5, 6]. Вкратце: заданы множество исходных описаний объектов $\mathbf{X} = \{x_1,\ldots,x_{N_0}\}$, $\mathbf{X}\subseteq S$, множества значений признаков $I_k = \{\lambda_{k_1},\lambda_{k_2},\ldots,\lambda_{k_b},\ldots,\lambda_{k_{|I_k|}},\Delta\}$, $b = 1,\ldots,|I_k|$, функции $\Gamma_k : S\to I_k$, $k = 1,\ldots,n+l$, где n — число признаков; l — число таргетных (прогнозируемых) переменных; Δ — неопределенность. Тогда определены пространство допустимых признаковых описаний объектов $J_{\text{ob}}\subseteq I_i\times I_f$ ($I_i\subseteq I_1\times\cdots\times I_n$, $I_f\subseteq I_{n+1}\times\cdots\times I_{n+1}$), функции $D : S\to J_{\text{ob}}$, $D(x_\alpha) = (\Gamma_1(x_\alpha)\times\cdots\times\Gamma_k(x_\alpha)\times\cdots\times\Gamma_{n+1}(x_\alpha))_\Delta$ и $\varphi(\mathbf{X}) = \{D(x_\alpha)|x_\alpha\in\mathbf{X}\}$. Принимается, что множество $\mathbf{X}$ и *множество объектов* $Q = \varphi(\mathbf{X})$, $|Q| = N$, *регулярны* ($\exists\,\varphi^{-1} : \mathbf{X} = \varphi^{-1}(Q)$) [5].

Множество $\mathrm{U}(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}$ рассматривается как предбаза *топологии*

$$T(\mathbf{X}) = \{\varnothing, I, a\cup b, a\cap b : a,b\in\mathrm{U}(\mathbf{X})\},$$

где $I = \{\mathbf{X}\}$ — единичный элемент. Топологии $T(\mathbf{X})$ соответствует *решетка*

$$L(T(\mathbf{X})) = \{a\vee b, a\wedge b : a,b\in T(\mathbf{X}),\ (a\ge b)\text{ или }(a\le b)\}.$$

При регулярности $\mathbf{X}/Q$ решетка $L(T(\mathbf{X}))$ — булева, так что *булевы* признаки — вершины, *категориальные* признаки — антицепи, а *числовые* — цепи в $L(T(\mathbf{X}))$ [7]. Тогда k-му числовому признаку с множеством значений I_k, $\lambda_{k_{b-1}}\le\lambda_{k_b}\le\lambda_{k_{b+1}}$, соответствует цепь $A_k(\mathbf{X}) = A(I_k,\mathbf{X}) =$

$= \langle u(\lambda_{k_1}), \ldots, u(\lambda_{k_b}), \ldots \rangle$ в $L(T(\mathbf{X}))$, образованная множествами

$$u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta}); \;\; A_k(\mathbf{X}) \in \mathbf{A}(\mathbf{X}),$$

где $\mathbf{A}(\mathbf{X})$ — множество всех цепей решетки $L(T(\mathbf{X}))$. *Эмпирическая функция распределения* (э. ф. р.) k-го признака определяется через совокупность точек $\mathbf{cdf}(A_k(\mathbf{X})) = \{(\lambda_{k_b} \in I_k, |u(\lambda_{k_b})|/N)\}$, а значение э. ф. р. в точке λ вычисляется как $\mathrm{cdf}(\lambda, A_k(\mathbf{X})) = |u(\lambda_{k_b})|/N$, $\lambda_{k_{b-1}} \le \lambda \le \lambda_{k_b}$ методами кусочно-линейной аппроксимации и т. п.

Булевой решетке $L(T(\mathbf{X}))$ сопоставлено *метрическое пространство значений признаков* $M_L(L(T(\mathbf{X})), \rho_L)$ с метрикой $\rho_L : L^2 \to R^+$. Если $v : L \to R^+$ — *изотонная оценка* ($\forall_L a, b : v[a] + v[b] = v[a \wedge b] + v[a \vee b], \underset{L}{\forall} a, b : a \supseteq b \Rightarrow v[a] \ge v[b]$), то функция $\rho(x, y) = v[x \vee y] - v[x \wedge y]$ — ρ_L-метрика [7]. При заданной ρ_L *расстояние* $\rho_A : \mathbf{A}(\mathbf{X})^2 \to R^+$ *между цепями* $a = \langle a_1, \ldots, a_i, \ldots \rangle$ и $b = \langle b_1, \ldots, b_j, \ldots \rangle$ может быть определено метрикой Хаусдорфа или как функционал от совокупности расстояний между ближайшими элементами цепей:

$$\rho_A(a, b) = \min \left(\sum_{i=1,|a|} \rho_L(a_i, \arg\min_{b_j \in b} \rho_L(a_i, b_j)), \right.$$
$$\left. \sum_{i=1,|b|} \rho_L(b_j, \arg\min_{a_i \in a} \rho_L(b_j, a_i)) \right).$$

Важно, что при любом определении ρ_A подразумевается однозначное соответствие элементов цепей a и b. При заданных $A_k(\mathbf{X})$ и множестве допустимых цепей $A(\mathbf{X})_{1,n}$ искомый ε-корректный алгоритм соответствует решению задачи оптимизации [5]:

$$\arg\min_{a \in \mathrm{A}(\mathbf{X})_{1,n}} \rho_A\left(A_k(\mathbf{X}), a\right) | A(\mathbf{X})_{1,n} \subset \mathbf{A}(\mathbf{X}). \qquad (2)$$

Выражение (2) описывает задачу комбинаторной оптимизации, при решении которой перебираются цепи из множества цепей $A(\mathbf{X})_{1,n}$. Для решения (2) необходимо определить метрики ρ_L и ρ_A, множество $A(\mathbf{X})_{1,n}$ и способ перебора цепей, как это делается при прогнозировании классов значений числовых переменных [4]. В настоящей работе рассмотрен подход, в котором $\hat{B}_{(\theta_{B_k}),k} = f_{\theta_k} : I_i \to I_k$, а θ_k вычисляются в результате минимизации некоторой «функции потерь» на выборке обучения.

В рамках такого «регрессионного» подхода будем считать, что f_{θ_k} вычисляет значения из некоторого $I_{\theta_k} \subseteq I_k$, что соответствует цепи $A(I_{\theta_k}, \mathbf{X}) = \langle A_1^{\theta_k}, \ldots, A_\beta^{\theta_k}, \ldots \rangle$, $\beta = 1, \ldots, |I_{\theta_k}|$, т. е. подцепи $A(I_k, \mathbf{X}) = \langle A_1^k, \ldots, A_b^k, \ldots \rangle$, $b = 1, \ldots, |I_k|$. При $I_{\theta_k} \subseteq I_k$ существует функция перенумерации $\delta : \mathbf{N} \to \mathbf{N}$, так что $b = \delta(\beta)$ и э. ф. р. для $y = f_{\theta_k}(x)$ вычисляется как $\mathrm{cdf}(y, A_k(\mathbf{X}))$. Множество цепей $A(\mathbf{X})_{1,n}$ в (2) зададим как $\{A(I_{\theta_k}, \mathbf{X})\}$ для всех $\theta_k \in \Theta_k$. Введем дополнительное понятие *мощностно́й функции расстояния* и рассмотрим задачу (2) в контексте анализа э. ф. р. k-й таргетной переменной.

Определение 1. Пусть $A, B \in L(T(\mathbf{X}))$ — два произвольных множества. Выражение $r(A, B) = ||A| - |B||$ назовем мощностно́й функцией расстояния между множествами A и B.

Теорема 1. *$r(A, B)$ является метрикой.*

Доказательство. Для метрической функции расстояния должны быть выполнены аксиомы тождества, симметричности и треугольника. Выполнение аксиомы тождества $r(A, A) = 0$ очевидно вследствие тождественности множества A самому себе, а аксиомы симметричности $r(A, B) = r(B, A)$ — вследствие операции модуля в определении $r(\,)$. Докажем выполнимость аксиомы треугольника. Рассмотрим случай $|A| \le |B| \le |C|$, в котором длины трех сторон треугольника равны $|B| - |A|$, $|C| - |B|$ и $|C| - |A|$. Тестируя выполнение аксиомы для треугольника, вершины которого соответствуют множествам A, B и C, запишем три неравенства: $|B| - |A| + |C| - |B| \ge |C| - |A|$ (т. е. $|C| \ge |C|$), $|C| - |B| + |C| - |A| \ge |B| - |A|$ (т. е. $|C| \ge |B|$) и $|C| - |A| + |B| - |A| \ge |C| - |B|$ (т. е. $|B| \ge |A|$). Первое из неравенств выполнено всегда, а второе и третье соответствуют рассматриваемому случаю ($|A| \le |B| \le |C|$). Остальные варианты соотношений $|A|$, $|B|$ и $|C|$ сводятся к рассмотренному посредством подстановки переменных, так что аксиома треугольника выполнена и $r(\,)$ — метрика. Теорема доказана.

Следствие 1. *Метрика $r(A, B)$ отражает расстояния между слоями решетки, в которую входят рассматриваемые множества A и B.* По определению в слое решетки расположены элементы одной высоты в L.

Следствие 2. *Пусть в решетке L задана метрика $\rho_L(A, B) = v(A \cup B - v(A \cap B)$ на основе изотонной оценки $v(\,)$, равной высоте элемента в L, $v(A) = h(A) = |A|$. Тогда $r(A, B) = |\rho_L(A, \varnothing) - \rho_L(B, \varnothing)|$, где $\varnothing$ — нулевой элемент решетки L.* Соответствует следствию 1.

Следствие 3. *В отличие от метрик на основе изотонных оценок $r(A, B)$ не зависит от совместного вхождения объектов в множества A и B.* В определении 1 отсутствуют теоретико-множественные операции «$\cup$», «$\cap$» и др.

Следствие 4. *Значения метрик $\rho_L(A, B)$ и $r(A, B)$ равны для двух произвольных элементов одной цепи*

решетки L. Очевидно из следствия 2 и того, что каждый элемент цепи соответствует определенному слою решетки.

Следствие 5. *Пусть метрика* $\rho_L(A,B)$ *определена как в следствии 2,* $\rho_L(A,B) \le \varepsilon$*, а множества* A *и* B *принадлежат разным цепям. Тогда* $r(A,B) \le \le \rho_L(A,B)$ *и* $r(A,B) \le \varepsilon$. При заданных условиях $\rho_L(A,B) = |A\Delta B| = |A\backslash B| + |B\backslash A| \le \varepsilon$. Очевидно, что $|A| = |A\cap B| + |A\backslash B|$ и $|B| = |A\cap B| + |B\backslash A|$, так что $r(A,B) = ||A\backslash B| - |B\backslash A||$. Так как сумма двух неотрицательных чисел $|A\backslash B|$ и $|B\backslash A|$ в $\rho_L(A,B) \le \le \varepsilon$ не может быть больше их разности в $r(A,B)$, то и $r(A,B) \le \varepsilon$.

Следствие 6. $r(A,B) \le \varepsilon$ — *необходимое условие* $\rho_L(A,B) \le \varepsilon$.

Следствие 7. *Пусть* $A, B \subset \mathrm{X}$. *Если* $r(A,B) \le \le \varepsilon$*, то и* $r(X\backslash A, X\backslash B) \le \varepsilon$. Из дистрибутивности L и принципа двойственности очевидно, что $\rho_L(A,B) = \rho_L(X\backslash A, X\backslash B)$, поэтому при $\rho_L(A,B) \le \le \varepsilon$ и $r(A,B) \le \varepsilon$, и $r(X\backslash A, X\backslash B) \le \varepsilon$.

Следствие 8. *Верхняя оценка* $\rho_L(A,B)$ *равна* $|A|+|B|$*, а среднее верхней и нижней оценок* — $\max(|A|,|B|)$.

Теорема 2. *Задача комбинаторной оптимизации* (2) *сводима к задаче минимизации различий между э. ф. р. таргетной переменной и э. ф. р. искомого алгоритма* f_{θ_k} *при использовании нижней оценки* ρ_L *на основе высоты элемента и задании функции потерь посредством неравенств* $\rho_L(A,B) \le \varepsilon$.

Доказательство. В (2) фигурируют расстояния ρ_L между множествами, входящими в цепи $A(I_k, \mathbf{X})$ и $A(I_{\theta_k}, \mathbf{X})$. Для прогнозирования k-й переменной заданы множество исходных описаний $\mathbf{X}$, множество прецедентов $Q = \varphi(\mathbf{X}) = \{(x_i, y_i^k),\ x_i \in I_i,\ y_i^k \in I_k \subset R,\ i = 1,\ldots,N\}$, f_{θ_k} и набор векторов $\theta_k \in \Theta_k$, причем при любом θ_k из Θ_k область значений f_{θ_k} остается равной I_{θ_k}. Объекты из Q формируют цепь $A(I_k, \mathbf{X})$, а для f_{θ_k} и θ_k определена цепь $\mathrm{A}(I_{\theta_k}, \mathbf{X})$.

При любых ρ_A и ρ_L множества в составе цепей однозначно сопоставимы. Для произвольного $\beta = 1,\ldots,|I_{\theta_k}|$ рассмотрим два множества $A^k_{\delta(\beta)} \in \in A(I_k, \mathbf{X})$ и $A^{\theta_k}_{\beta} \in A(I_{\theta_k}, \mathbf{X})$. ε-корректному алгоритму соответствует условие

$$\forall\, \beta : \rho_L\left(A^k_{\delta(\beta)}, A^{\theta_k}_{\beta}\right) \le \varepsilon,$$

т. е.

$$\sum_{\beta=1}^{|I_{\theta_k}|} \rho_L\left(A^k_{\delta(\beta)}, A^{\theta_k}_{\beta}\right) \le \varepsilon |I_{\theta_k}|.$$

В соответствии с теоремой 1 и ее следствиями при использовании ρ_L на основе изотонной оценки, равной высоте элемента, $r(\,)$ — нижняя оценка ρ_L. Заменим ρ_L на $r(\,)$ и получим

$$\sum_{\beta=1}^{|I_{\theta_k}|} \left|\left|A^k_{\delta(\beta)}\right| - \left|A^{\theta_k}_{\beta}\right|\right| \le \varepsilon\, |I_{\theta_k}|\,.$$

Поделив обе части неравенства на $|Q| = |\mathbf{X}| = N$, получим

$$\sum_{\beta=1}^{|I_{\theta_k}|} \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\beta)}}, A_k(\mathbf{X})\right) - \mathrm{cdf}\left(\lambda_{k_\beta}, A(I_{\theta_k}, \mathbf{X})\right)\right| \le \le \frac{\varepsilon |I_{\theta_k}|}{N}\,.$$

При фиксированных I_{θ_k} и N это условие ε-корректного алгоритма можно рассматривать как регрессионную задачу, подразумевающую минимизацию значения ε по θ_k:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{\beta=1}^{|I_{\theta_k}|} \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\beta)}}, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(\lambda_{k_\beta}, A(I_{\theta_k}, \mathbf{X})\right)\right|. \quad (3)$$

Теорема доказана.

Заметим, что выражение (3) описывает параметрический критерий задачи оптимизации, где исследователь фиксирует значение параметра I_{θ_k}. В простейшем случае, когда $I_{\theta_k} = [0, \lambda]$, задача (3) редуцируется к задаче прогнозирования классов значений k-й числовой переменной [4].

По определению функции $\mathrm{cdf}(\,)$ ее значения соответствуют некоторой доле объектов множества $\mathbf{X}$, так что в (3) разность значений $\mathrm{cdf}(\,)$ для заданного λ_{k_β} равна доле объектов из $\mathbf{X}$, ошибочно классифицированных относительно класса значений $u(\lambda_{k_\beta})$. Поэтому можно перейти от задачи (3) к аналогичной задаче, в которой ошибка алгоритма f_{θ_k} оценивается не суммированием по значениям в I_{θ_k}, а суммированием ошибок в $\mathrm{cdf}(\,)$ по индивидуальным объектам:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{i=1}^{N} \left|\mathrm{cdf}\left(y_i^k, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(f_{\theta_k}(x_i), A(I_{\theta_k}, \mathbf{X})\right)\right|. \quad (4)$$

Поскольку $I_{\theta_k} \subseteq I_k$, удобней оценивать значения э. ф. р. от f_{θ_k} используя э. ф. р. таргетной переменной:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{i=1}^{N} \left|\mathrm{cdf}\left(y_i^k, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(f_{\theta_k}(x_i), A_k(\mathbf{X})\right)\right|. \quad (5)$$

Очевидно, что в задачах (3)–(5) подразумевается минимизация отклонений рангов/процентилей значений функции f_{θ_k} с функцией потерь в виде

модуля. При выборе в качестве ρ_L, например, квадрата $r(\,)$ в задаче типа (5) будет минимизироваться сумма квадратов невязок этих рангов/процентилей. Назовем (3), (4) и др. задачами *ранговой оптимизации*. Рассмотрим описанный сценарий, полученный в рамках топологического подхода к анализу данных, в контексте регрессии/аппроксимации функций.

3 Ранговая оптимизация и задачи регрессии/аппроксимации

Рассмотрим случай одномерной вещественной функции, заданной набором точек $\mathrm{Pr} = \{(x_i, y_i)\}$, $i = 1, \ldots, N$, $x_i \in [a, b]$, $y_i \in [c, d]$. Пусть Pr соответствует некоторой «истинной» функции f_t, $y_i = f_t(x_i)$. Будем аппроксимировать набор Pr функцией f_θ, $y_i = f_\theta(x_i) + o(x_i, \theta)$. Пусть функции f_t и f_θ непрерывны, дифференцируемы и интегрируемы на интервале $[a, b]$. Необходимо найти вектор параметров θ, минимизирующий абсолютные значения $o(x_i, \theta)$ в соответствии с заданным критерием оптимизации.

Критерии оптимизации можно формулировать исходя из идеи минимизации площади $S(f_t, f_\theta)$, соответствующей суммарному отличию f_t и f_θ на $[a, b]$. В идеальном случае $S = 0$ ($o(x) \equiv 0$), иначе $f_t(x) = f_\theta(x) + o(x)$, $o(x) > 0$, и площадь S может быть оценена корнем из интеграла $\int_a^b (f_\theta\, dx - f_t\, dx)^2$, соответствующего критерию $\min_\theta \sum_{i=1}^N (y_i - f_\theta(x_i))^2$. Аналогично можно получить критерии минимизации взвешенной суммы невязок, критерий метода наименьших модулей и др. Также возможна оценка площади S взятием интеграла по Лебегу, когда суммируются ошибки в x при заданном y, и т. п.

Пусть при заданном Pr площадь S оценивается как сумма некоторых вкладов «w» отдельных точек, $S = \sum_{i=1}^N w(x_i, y_i) + o(\mathrm{Pr})$. Примем, что каждой точке сопоставлен одинаковый вклад μ в площадь S (с точностью до $o(\mathrm{Pr})/N$), так что $S \sim \mu N$. Такое допущение может быть оправдано при достаточно большом N и при достаточно равномерном распределении точек вдоль соответствующих осей (ситуация, характерная для «больших данных», производимых современными физико-химическими технологиями сбора данных). Тогда можно рассмотреть приближенные оценки площади S (и соответствующие критерии минимизации), основанные не на *разностях* в значениях y_i и $f_\theta(x_i)$, а на *подсчете числа точек* (*объектов*) *в* Pr *и в его подмножествах*.

Вернемся к случаю произвольного множества $\mathbf{X}$, $Q = \varphi(\mathbf{X}) = \{(x_i, y_i^k), x_i \in I_i, y_i^k \in I_k\}$, $I_k = (\lambda_{k_b}), k = 1, \ldots, n$. Пусть задано подмножество $\zeta = (\lambda_{k_\alpha}) \subseteq I_k$, $\alpha = 1, \ldots, m$, $m = |I_k|$. В одномерном случае каждому значению λ_{k_α} соответствует горизонтальная линия, параллельная оси абсцисс, а в случае произвольного $x_i \in I_i$ — гиперплоскость. Для заданного λ_{k_α} вычислим число объектов со значениями y_i^k ниже λ_{k_α}, $n_\alpha^{t\leq} = |u(\lambda_{k_\alpha})|$, и y_i^k выше λ_{k_α}, $n_\alpha^{t>} = |\mathbf{X} \backslash u(\lambda_{k_\alpha})|$. Определим аналогичные числа для λ_{k_α} в цепи $A(I_{\theta_k}, \mathbf{X})$, производимой алгоритмом f_{θ_k}, $n_\alpha^{\theta_k \leq} = |\{(x, y) \in Q | f_{\theta_k}(x) \leq \lambda_{k_\alpha}\}|$ и $n_\alpha^{\theta_k >} = |\{(x, y) \in Q | f_{\theta_k}(x) > \lambda_{k_\alpha}\}|$. При вкладе каждой из точек, равном μ, оценим значение площади S_α для выбранного λ_{k_α} как разность в числе точек (объектов), у которых значение y_i^k ниже порогового значения λ_{k_α}, и объектов, у которых значение y_i^k выше λ_{k_α}:

$$S_\alpha = \left(\left|n_\alpha^{t\leq} - n_\alpha^{\theta_k \leq}\right| + \left|n_\alpha^{t>} - n_\alpha^{\theta_k >}\right|\right) \mu\,.$$

По построению

$$n_\alpha^{t\leq} + n_\alpha^{t>} = n_\alpha^{\theta_k \leq} + n_\alpha^{\theta_k >} = |Q|,$$

так что

$$S_\alpha = 2\left|n_\alpha^{t\leq} - n_\alpha^{\theta_k \leq}\right| \mu\,,$$

что эквивалентно

$$S_\alpha = \\ = 2\mu N \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\alpha)}}, A_k(\mathbf{X})\right) - \mathrm{cdf}\left(\lambda_{k_\alpha}, A(I_{\theta_k}, \mathbf{X})\right)\right|.$$

Оценим площадь $S(f_t, f_\theta)$ как математическое ожидание S_α по всем элементам множества ζ:

$$S = \frac{1}{m} \sum_{\alpha=1}^{m} S_\alpha.$$

Считая μ, N и m константами и минимизируя S по θ_k, приходим к задаче (3).

Таким образом, задачи ранговой оптимизации в рамках топологического подхода (минимизация расстояния между цепями решетки) также могут быть получены исходя из специфического способа оценки различий $S(f_t, f_\theta)$ между «истинной» функцией f_t и ее аппроксимацией f_θ. Перспективно использовать комбинации критериев (3)–(5), что позволит одновременно минимизировать и отличия индивидуальных объектов, и отличия значений э. ф. р. в паре «переменная–алгоритм».

Получаемые критерии оптимизации относятся к параметрическим — в качестве параметра выступает подмножество множества I_k. При определенном выборе подмножества ζ получаются конструкции, идеологически близкие к задачам квантильной регрессии [8]. При назначении весов значениям λ_{k_α} и исследовании э. ф. р. значений S_α можно получить более сложные критерии.

4 Результаты экспериментального тестирования формализма

Формализм апробирован на задаче взаимодействия лиганд–рецептор, в которой значения $\mathrm{EC}_{50}(j)$ прогнозируются исходя из химической структуры молекул. Решения задач в постановках (3), (5) позволяют прогнозировать не только сами значения $\mathrm{EC}_{50}(j)$, но и значения откликов $E_j(C_i)$, для которых затем используется корректор в виде (1). При прогнозировании $\mathrm{EC}_{50}(j)$ и $E_j(C_i)$ на основе хемогра́фа G_j в качестве множества начальных информаций I_i использовалось множество хемоинвариантов над алфавитом специальных меток (см. ниже). Алгоритмы $f_{\theta_k} : I_i \to R$ строились в виде композиций вложенных корректирующих функций нижнего уровня (порождение синтетических признаков) для фиксированного числа моделей $n_{\mathrm{mod}} : f_{\theta_k} = g(f_1(\sum \omega_k^j x_k), \ldots, f_l(\sum \omega_k^j x_k), \ldots)$, $l = 1, \ldots, n_{\mathrm{mod}}$, где g — внешняя корректирующая функция; f_l — внутренние корректирующие функции (модели порождения синтетических числовых признаков); n_{mod} — их число. Суммирование $\sum \omega_j^k x_j$ проводится по компонентам вектора $\mathbf{x} \in I_i$, $k = 1, \ldots, |\mathbf{x}|$. Использовались линейные, нелинейные, монотонные и немонотонные функции-корректоры g и f_l (более 20 монотонных и немонотонных преобразований, в том числе описанных в работе [6]). Векторы параметров настраивались мультистартовой стохастической оптимизацией в рамках кросс-валидационного дизайна [6].

При прогнозировании EC_{50} исходя из $E_j(C_i)$ использовались вложенные алгоритмические структуры, описываемые алгоритмами 2-го уровня:

$$\hat{A}^{(2)}_{(\theta^2_A)} = \hat{C}^{(2)}_{(\theta^2_C)} \circ \hat{B}^{(2)}_{(\theta^2_B)}.$$

В качестве корректирующей операции $\hat{C}^{(2)}_{(\theta^2_C)}$ использовалось уравнение Хилла (1), а в качестве распознающих операторов $\hat{B}^{(2)}_{(\theta^2_B)}$ — функции $g(f_1(\sum \omega_k^j x_k), \ldots)$, настраиваемые на множествах откликов $E_j(C_i)$.

Для хемографа $X \in \mathbf{\Gamma}$ ($\mathbf{\Gamma}$ — множество всех хемографов, основанное на алфавите меток Y) хемоинварианты порождались на основании множеств χ-цепей длины $\tilde{Y}^m(X)$ и χ-узлов $\hat{Y}(X)$ [6]. Вкратце: пусть задано множество подграфов (χ-цепей и χ-узлов) $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\} \subset \mathbf{\Gamma}$. Определим оператор вхождения подграфа π в хемограф X как

$$\hat{\beta}[X]\pi = \left(|\pi \cap \mathbf{\Pi}(X)| > 0\right), \mathbf{\Pi}(X) = \tilde{Y}^m(X) \cup \hat{Y}(X),$$

а последовательное применение $\hat{\beta}$ к $\boldsymbol{\pi}$ — булев вектор

$$\hat{\boldsymbol{\beta}}[X]\boldsymbol{\pi} = \left(\hat{\beta}[X]\boldsymbol{\pi}_1, \hat{\beta}[X]\boldsymbol{\pi}_2, \ldots, \hat{\beta}[X]\boldsymbol{\pi}_n\right).$$

Для множества хемографов $\mathbf{X}$ множество начальных информаций

$$I_i = \bigcup_{k=1}^{|\mathbf{X}|} \hat{\boldsymbol{\beta}}[X_k] \left(\tilde{Y}^m(X_k) \cup \hat{Y}(X_k)\right), \; m = 5$$

(соответствует оптимальным результатам тестирования регулярности по Журавлёву [6]).

Тестирование алгоритмов f_{θ_k} и $\hat{A}^{(2)}_{(\theta^2_A)}$ проводилось на выборке данных из ProteomicsDB (https://www.proteomicsdb.org), содержащей данные по C_i ($C_i = 1$, 3, 10, 100, 1000, 3000, 30 000 нмоль/л), $E_j(C_i)$ и $\mathrm{EC}_{50}(j)$ для 300 ферментов-киназ (так называемый кино́м человека, часть протеома) и ряда лекарств. Киназы представляют собой таргетные белки известных и перспективных лекарственных средств. Наилучшие результаты прогнозирования $\mathrm{EC}_{50}(j)$ получались при (1) пренебрежении эффектами атомов водорода (т. е. при использовании более простых Y-алфавитов), (2) использовании линейного распознающего оператора в сочетании с немонотонными корректорами (нейронные сети, решающие деревья, полиномиальные функции и др.), 3) совместном использовании критериев (3) и (5). Результаты экспериментов суммированы в таблице.

Как при использовании линейной, так и нейросетевой $g(\,)$, применение критериев ранговой регрессии (3) и (5) способствовало снижению различий между значениями коэффициента корреляции на обучении и контроле. Наиболее выраженный эффект наблюдался для (5), в то время как минимизация отклонений э. ф. р. по (3) имела вспомогательное значение. Наилучший результат получен при использовании нейросетевой g, настраиваемой в соответствии с обоими ранговыми критериями ($r_c = 0{,}86 \pm 0{,}20$).

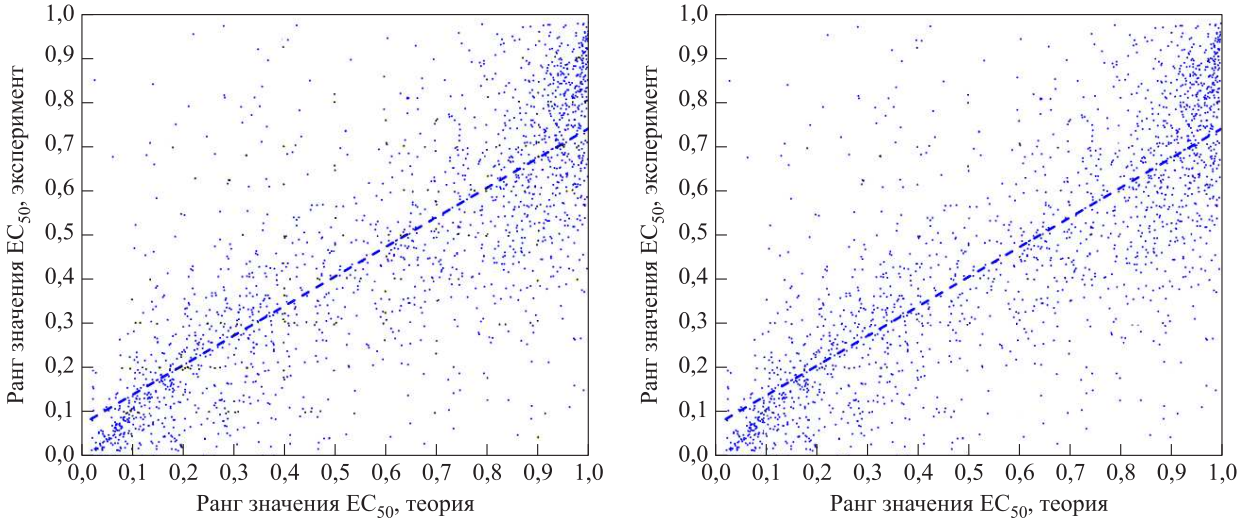
В отличие от описанного выше «прямого» прогнозирования EC_{50} прогнозирование $E_j(C_i)$ с использованием корректора (1) оказалось менее успешным (см. рисунок). Точность прогнозирования отдельных значений $E_j(C_i)$ была сопоставима с приведенной в таблице ($r_c \sim 0{,}85 \pm 0{,}21$), но данная схема прогнозирования отличалась существенно более низким качеством на контроле ($r_c \sim 0{,}63 \pm 0{,}24$).

В целом критерии (5) и (3) могут успешно использоваться не только для хемокиномного анализа, но и для хемотранскриптомного анализа лигандов, поиска эффективных и безопасных средств для фармакотерапии COVID-19, в хемомикробиомном анализе и др. (см. ресурс www.chemoinformatics.ru).

Тестирование расчетов $EC_{50}(j)$ для 300 киназ человека

Эксперимент	r	r_c	CO	CO_c
Обычная регрессия, g — линейная	$0{,}67 \pm 0{,}25$	$0{,}45 \pm 0{,}26$	$0{,}24 \pm 0{,}15$	$0{,}25 \pm 0{,}14$
Ранговая регрессия по (5), g — линейная	$0{,}68 \pm 0{,}23$	$0{,}48 \pm 0{,}25$	$0{,}20 \pm 0{,}19$	$0{,}22 \pm 0{,}20$
Ранговая регрессия по (3), g — линейная	$0{,}67 \pm 0{,}25$	$0{,}47 \pm 0{,}25$	$0{,}23 \pm 0{,}21$	$0{,}22 \pm 0{,}22$
Ранговая регрессия по (3) и (5), g — линейная	$0{,}68 \pm 0{,}23$	$0{,}47 \pm 0{,}25$	$0{,}20 \pm 0{,}20$	$0{,}21 \pm 0{,}20$
Обычная регрессия, g — нейросеть	$0{,}89 \pm 0{,}13$	$0{,}79 \pm 0{,}13$	$0{,}18 \pm 0{,}12$	$0{,}13 \pm 0{,}11$
Ранговая регрессия по (5), g — нейросеть	$\mathbf{0{,}88 \pm 0{,}15}$	$\mathbf{0{,}83 \pm 0{,}28}$	$\mathbf{0{,}05 \pm 0{,}03}$	$\mathbf{0{,}05 \pm 0{,}03}$
Ранговая регрессия по (3), g — нейросеть	$0{,}89 \pm 0{,}13$	$0{,}81 \pm 0{,}16$	$0{,}18 \pm 0{,}17$	$0{,}17 \pm 0{,}17$
Ранговая регрессия по (5) и (3), g — нейросеть	$\mathbf{0{,}88 \pm 0{,}15}$	$\mathbf{0{,}86 \pm 0{,}20}$	$\mathbf{0{,}03 \pm 0{,}02}$	$\mathbf{0{,}04 \pm 0{,}03}$

Примечания: r — коэффициент ранговой корреляции на обучении, r_c — на контроле; CO — стандартное отклонение на обучении, CO_c — на контроле; кросс-валидационный дизайн (10 разбиений, «случай–контроль» 6 : 1). В качестве нейросети использовалась 2-слойная сеть с функцией активации softmax.



(*а*) (*б*)

Прогнозирование EC_{50} с использованием схемы «прямого» прогнозирования ($y = 0{,}67x + 0{,}0706$, $R^2 = 0{,}622$) (*а*) и схемы прогнозирования $E_j(C_i)$ с корректором в виде уравнения Хилла–Ленгмюра ($y = 0{,}5959x + 0{,}018$, $R^2 = 0{,}4027$) (*б*). Приведены результаты для контрольной выборки

5 Заключение

Перспективным направлением повышения точности работы алгоритмов машинного обучения стала разработка математического инструментария, позволяющего порождать проблемно-ориентированные теории для решения конкретных прикладных задач. В частности, топологический подход к распознаванию позволяет систематически анализировать различные способы порождения признаковых описаний плохо формализованных задач распознавания/классификации. Как показали результаты настоящей работы, выбор определенных метрик ρ_L и ρ_A в рамках топологического подхода соответствует порождению специфических критериев рангового характера, оптимизация которых позволяет улучшить показатели обучения соответствующих алгоритмов. С разработанными критериями обучение моделей и оценка качества прогнозирования проводятся на основе сохранения отношений порядка значений, а не самих значений таргетной переменной (что крайне важно, например, с точки зрения теоретической химии). В зависимости от конкретных задач э. ф. р. могут быть представлены в дискретной форме (как в настоящей работе) либо в виде аппроксимаций посредством непрерывных функций. Возможна разработка более сложных ранговых критериев на основе статистических функционалов.

Литература

1. *Журавлёв Ю. И.* Избранные научные труды. — М.: Магистр, 1998. 420 с.
2. *Torshin I. Y., Rudakov K. V.* Combinatorial analysis of the solvability of the problems of recognition, completeness of algorithmic models. Part 1: Factorization approach //

Pattern Recognition Image Analysis, 2017. Vol. 27. No. 1. P. 16–28. doi: 10.1134/S1054661817010151.

3. *Torshin I. Y., Rudakov K. V.* Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 2: Metric approach within the framework of the theory of classification of feature values // Pattern Recognition Image Analysis, 2017. Vol. 27. No. 2. P. 184–199. doi: 10.1134/S1054661817020110.
4. *Torshin I. Y., Rudakov K. V.* On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables // Pattern Recognition Image Analysis, 2019. Vol. 29. No. 3. P. 654–667. doi: 10.1134/S1054661819040175.
5. *Торшин И. Ю.* О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул I: Основы проблемно ориентированной теории // Информатика и её применения, 2022. Т. 16. Вып. 1. С. 39–45. doi: 10.14357/19922264220106.
6. *Торшин И. Ю.* О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул II: Сопоставление формализма с конструктами квантовой механики и экспериментальная апробация предложенных алгоритмов // Информатика и её применения, 2022. Т. 16. Вып. 2. С. 35–43. doi: 10.14357/19922264220205.
7. *Torshin I. Y., Rudakov K. V.* On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification // Pattern Recognition Image Analysis, 2015. Vol. 25. No. 4. P. 577–587. doi: 10.1134/S1054661815040252.
8. *Koenker R., Bassett G.* Regression quantiles // Econometrica, 1978. Vol. 46. No. 1. P. 33–50. doi: 10.2307/1913643.

Поступила в редакцию 05.10.22

ON OPTIMIZATION PROBLEMS ARISING FROM THE APPLICATION OF TOPOLOGICAL DATA ANALYSIS TO THE SEARCH FOR FORECASTING ALGORITHMS WITH FIXED CORRECTORS

I. Yu. Torshin

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Corrective operations (correctors) in multialgorithmic constructions of the algebraic approach can be based on known physical models and/or multilevel descriptions of physical objects. At the same time, within the framework of the topological approach to the analysis of poorly formalized problems, the search for algorithms included in the corrector can be considered as a combinatorial optimization problem or as a problem of minimizing a certain loss function. The study of the neighborhoods of chains in the lattice of subsets of objects made it possible to obtain a number of rank optimization criteria that are promising for solving the problems of predicting numerical target variables. The formalism was tested on the problem of ligand–receptor interaction within the framework of the chemokine analysis of drug molecules (data from ProteomicsDB). The best results of predicting constants were observed when using the obtained rank criteria (correlation coefficient on a sliding control 0.86 ± 0.20 averaging over 300 biological activities).

Keywords: topological data analysis; lattice theory; optimization problems; regression; chemoinformatics

DOI: 10.14357/19922264230201 **EDN:** IGSPEW

Acknowledgments

The research was funded by the Russian Science Foundation, project 23-21-00154. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Zhuravlev, Yu. I. 1998. *Izbrannye nauchnye trudy* [Selected scientific works]. Moscow: Magistr. 420 p.
2. Torshin, I. Y., and K. V. Rudakov. 2017. Combinatorial analysis of the solvability of the problems of recognition, completeness of algorithmic models. Part 1: Factorization approach. *Pattern Recognition Image Analysis* 27(1):16–28. doi: 10.1134/S1054661817010151.
3. Torshin, I. Y., and K. V. Rudakov. 2017. Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 2: Metric approach within the framework of the theory of clas-

sification of feature values. *Pattern Recognition Image Analysis* 27(2):184–199. doi: 10.1134/ S1054661817020110.

4. Torshin, I. Yu., and K. V. Rudakov. 2019. On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables. *Pattern Recognition Image Analysis* 29(4):654–667. doi: 10.1134/S1054661819040175.
5. Torshin, I. Yu. 2022. O primenenii topologicheskogo podkhoda k analizu plokho formalizuemykh zadach dlya postroeniya algoritmov virtual'nogo skrininga kvantovomekhanicheskikh svoystv organicheskikh molekul I: Osnovy problemno orientirovannoy teorii [On the application of a topological approach to analysis of poorly formalized problems for constructing algorithms for virtual screening of quantum-mechanical properties of organic molecules I: The basics of the problem-oriented theory]. *Informatika i ee Primeneniya — Inform Appl.* 16(1):39–45. doi: 10.14357/19922264220106.
6. Torshin, I. Yu. 2022. O primenenii topologicheskogo podkhoda k analizu plokho formalizuemykh zadach dlya postroeniya algoritmov virtual'nogo skrininga kvantovomekhanicheskikh svoystv organicheskikh molekul II: Sopostavlenie formalizma s konstruktami kvantovoy mekhaniki i eksperimental'naya aprobatsiya predlozhennykh algoritmov [On the application of a topological approach to analysis of poorly formalized problems for constructing algorithms for virtual screening of quantum-mechanical properties of organic molecules II: Comparison of formalism with constructions of quantum mechanics and experimental approbation of the proposed algorithms]. *Informatika i ee Primeneniya — Inform Appl.* 16(2):35–43. doi: 10.14357/19922264220205.
7. Torshin, I. Y., and K. V. Rudakov. 2015. On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification. *Pattern Recognition Image Analysis* 25(4):577–587. doi: 10.1134/ S1054661815040252.
8. Koenker, R., and G. Bassett. 1978. Regression quantiles. *Econometrica* 46(1):33–50. doi: 10.2307/1913643.

Received October 5, 2022

Contributor

Torshin Ivan Y. (b. 1972) — Candidate of Science (PhD) in physics and mathematics, Candidate of Science (PhD) in chemistry, senior scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; tiy135@yahoo.com

МОНАДА ДИАГРАММ КАК МАТЕМАТИЧЕСКАЯ МЕТАМОДЕЛЬ СИСТЕМНОЙ ИНЖЕНЕРИИ

С. П. Ковалёв[1]

Аннотация: Рассматриваются вопросы разработки перспективных математических методов системной инженерии, способных лечь в основу компьютерных инструментов автоматического синтеза и анализа систем и процессов. В русле современных тенденций в качестве аппарата для этих методов выбрана теория категорий. Ее применение отталкивается от представления структуры систем, процессов, требований и других результатов системного проектирования диаграммами в категориях, объектами которых служат алгебраические модели составных частей, а морфизмы описывают взаимосвязи между частями. При помощи фундаментальной уплощающей конструкции Гротендика описано явное построение категорий диаграмм, монады диаграмм, монады и комонады диаграмм с отмеченной точкой. Указаны области приложения этих конструкций в процедурах системной инженерии. Предложен подход к реализации высокоавтоматизированных технологий типа порождающего проектирования для сложных многоуровневых систем.

Ключевые слова: теория категорий; монада диаграмм; конструкция Гротендика; копредел; системная инженерия; система систем; порождающее проектирование

DOI: 10.14357/19922264230202 **EDN:** FDUVDU

1 Введение

Эффективность традиционных узкопрофильных инженерных дисциплин (механика, гидравлика, электроника и др.), в особенности в условиях цифровизации, обусловлена широким применением математического аппарата, включающего аналитическую геометрию, дифференциальные уравнения, математическое программирование и др. Однако для решения задач системной инженерии математические методы развиты слабо, они проигрывают «практикам» и «техникам» [1]. Поэтому успешные системные проекты плохо поддаются тиражированию и сборке в мегапроекты: не хватает формальных абстрактных схем выполненных процедур и правил расчета влияния контекста на достигнутые результаты. Тем более трудно передать проектирование систем компьютеру: приходится ограничиваться средствами редактирования и несложного анализа «наглядных» цифровых схем, изображающих придуманные инженером структуры систем и процессов.

В ряде публикаций (см., например, [2–4]) предлагается решить эту проблему путем привлечения теории категорий — раздела высшей алгебры, направленного на унифицированное представление объектов различной природы и отношений между ними. В основе приложения теории категорий лежит представление структуры систем, процессов, требований и др. диаграммами в категориях типа «каталогов», объектами которых служат алгебраические модели составных частей, а морфизмы описывают те или иные взаимосвязи между частями. Известны категории-каталоги геометрических форм, сценариев поведения частей, энергетических ресурсов и т. д. [5].

Однако диаграммы чаще всего рассматриваются фрагментарно и специфично для конкретного прикладного контекста. Отрывочность и разнобой наблюдаются даже на уровне метамодели — языка, на котором описываются и верифицируются свойства и преобразования диаграмм в ходе моделирования объектов и процедур системной инженерии. При этом терминологически выверенные глубокие исследования диаграммных конструкций, написанные «чистыми» математиками, совершенно непостижимы для инженеров, да и не очень им интересны.

Настоящая работа нацелена на преодоление этого недостатка, вводя универсальный (в строгом теоретико-категориальном смысле) язык, основанный на известной конструкции монады диаграмм [6, 7]. Основным рабочим инструментом служат элементарные универсальные конструкции в «категории» **CAT**, состоящей из всех категорий и всех функторов. Такие конструкции (в том числе компоненты рассматриваемых в работе монад) определены с точностью до изоморфизма, но для ясности изложения предполагается, что у них задан некоторый канонический вид. Описано явное постро-

[1] Институт проблем управления им. В. А. Трапезникова Российской академии наук, kovalyov@sibnet.ru

ение и области приложения категорий диаграмм, монады диаграмм, монады и комонады диаграмм с отмеченной точкой.

2 Категории структур систем

Предполагается, что читатель знаком с основами теории категорий. Используются элементарные теоретико-категориальные конструкции и обозначения, введенные в работах [7, 8]. Начнем с конструкции универсального расслоения (universal bundle) для категорий — это канонический функтор $\mathbf{B} : \dot{\mathbf{C}}\mathbf{AT} \to \mathbf{CAT}$, где через $\dot{\mathbf{C}}\mathbf{AT}$ обозначена «категория» всех категорий с отмеченной точкой, в которой объектом служит любая пара (A, C), $C \in \mathrm{Ob}\,\mathbf{CAT}$, $A \in \mathrm{Ob}\,C$, а морфизмом (A, C) в (A', C') служит любая пара $(f : GA \to A', G : C \to C')$; функтор $\mathbf{B}$ «забывает» отмеченную точку. Для произвольного функтора $F : D \to \mathbf{CAT}$ декартов квадрат с универсальным расслоением известен как уплощающая конструкция Гротендика $\int F$ (Grothendieck flattening construction) [9] (рис. 1).

В явном виде объектом категории $\int F$ служит любая пара (A, X), $X \in \mathrm{Ob}\,D$, $A \in \mathrm{Ob}\,FX$, а морфизмом пары (A, X) в (A', X') служит любая пара $(f : (Fg)A \to A', g : X \to X')$ (с законом композиции вида $(f, g) \circ (h, q) = (f \circ (Fg)h, g \circ q)$). Можно принять такое описание конструкции Гротендика за ее определение [10, п. 12.2.10] (и построить универсальное расслоение как $\int 1_{\mathbf{CAT}}$).

Отображение $(A, X) \mapsto X$ задает канонический «забывающий» функтор из $\int F$ в D, представленный левой вертикальной стрелкой в вышеприведенном декартовом квадрате. Полный прообраз (декартов квадрат) этого функтора относительно любого элемента $\ulcorner X \urcorner : \mathbf{1} \to D$ (где $\mathbf{1}$ — сингулярная категория, терминальный объект в $\mathbf{CAT}$) задает каноническое вложение $\boldsymbol{i}_X : FX \hookrightarrow \int F : A \mapsto (A, X)$, $f \mapsto (f, 1_X)$. В сумме получается вложение $[\boldsymbol{i}_X]_{X \in \mathrm{Ob}\,D} : \coprod_{X \in \mathrm{Ob}\,D} FX \hookrightarrow \int F$, биективное на объектах. Если категория D дискретна, то это суммарное вложение является изоморфизмом.

Для контравариантного функтора $F : D^{\mathrm{op}} \to \mathbf{CAT}$ в качестве правой вертикальной стрелки вышеприведенного декартового квадрата используется универсальное «op-расслоение» $\mathbf{B} : \dot{\mathbf{C}}\mathbf{AT}^{\mathrm{op}} \to \mathbf{CAT}^{\mathrm{op}}$ (а функтор F рассматривается как действующий из D в $\mathbf{CAT}^{\mathrm{op}}$). Применим такую контравариантную конструкцию Гротендика к функтору $C^- : \mathbf{Cat}^{\mathrm{op}} \to \mathbf{CAT} : X \mapsto C^X$ для некоторой категории C. (Универсальный характер этого функтора вытекает из того, что экспонента C^X вычисляется посредством эндофунктора, правого сопряженного к эндофунктору произведения $C \mapsto X \times C$ [8, § IV.6], которое, в свою очередь, сводится к декартову квадрату функторов $!_X : X \to \mathbf{1}$ и $!_C : C \to \mathbf{1}$, каждый из которых определен единственным образом.) В результате получается категория диаграмм, которая обозначается через $\mathbf{D}C$ и служит «нестрогим ко-пополнением» (lax cocompletion) [6] категории C. Класс объектов категории $\mathbf{D}C$ состоит из всех C-диаграмм, а морфизмом диаграммы $\Delta : X \to C$ в $\Delta' : X' \to C$ служит любая пара вида $\langle \gamma, \mathfrak{f} \rangle$, состоящая из функтора $\mathfrak{f} : X \to X'$ и естественного преобразования $\gamma : \Delta \to \Delta' \mathfrak{f}$; закон композиции имеет вид $\langle \gamma, \mathfrak{f} \rangle \circ \langle \varphi, \mathfrak{g} \rangle = \langle \gamma \mathfrak{g} \circ \varphi, \mathfrak{f}\mathfrak{g} \rangle$. Имеется каноническое вложение $[\boldsymbol{i}_I]_{I \in \mathrm{Ob}\,\mathbf{Cat}} : \coprod_{I \in \mathrm{Ob}\,\mathbf{Cat}} C^I \hookrightarrow \mathbf{D}C$, биективное на объектах.

Процедура построения категории $\mathbf{D}C$ функториальна по C: любой функтор $F : C \to C'$ индуцирует функтор

$$F- : \mathbf{D}C \to \mathbf{D}C' : \Delta \mapsto F\Delta, \langle \gamma, \mathfrak{f} \rangle \mapsto \langle F\gamma, \mathfrak{f} \rangle.$$

Тем самым определен эндофунктор

$$\mathbf{D} : \mathbf{CAT} \to \mathbf{CAT} : C \mapsto \mathbf{D}C,\ F \mapsto F-$$

(так что можно построить категорию *всех* диаграмм $\int \mathbf{D}$). Отметим, что $\mathbf{D1} \cong \mathbf{Cat}$ и левая вертикальная стрелка в декартовом квадрате конструкции Гротендика для $\mathbf{D}C$ представляет собой канонический функтор формы диаграмм (рис. 2):

$$\mathbf{D}!_C : \mathbf{D}C \to \mathbf{Cat} : \Delta \mapsto \mathrm{dom}\,\Delta, \langle \gamma, \mathfrak{f} \rangle \mapsto \mathfrak{f}.$$

В приложениях в области системной инженерии диаграммы представляют структуры систем, а их морфизмы описывают структурные преобразования систем на алгебраическом языке [5]. Функтор $\mathbf{D}$ описывает естественный переход от каталогов объектов к каталогам структур систем, которые можно составить из объектов. Подходящие подкатегории в $\mathbf{D}C$ могут служить пространствами

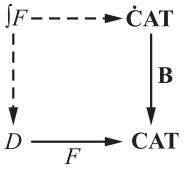


Рис. 1 Уплощающая конструкция Гротендика

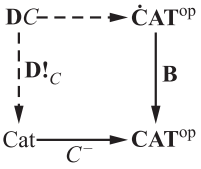


Рис. 2 Канонический функтор формы диаграмм

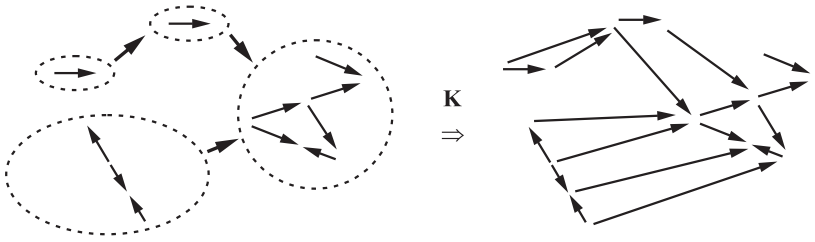


Рис. 3 Отрисовка — умножение в монаде диаграмм

проектирования (design space) для автоматического поиска суб- и Парето-оптимальных структур систем. Для такой оптимизации целевые функции, изначально заданные заинтересованными сторонами систем, преобразуются в функторы, действующие из пространств проектирования в линейно упорядоченные множества значений, рассматриваемые как категории [11]. Для эффективной навигации в пространствах проектирования вдоль морфизмов диаграмм могут привлекаться средства компьютерной алгебры. Открывается возможность реализации высокоавтоматизированных технологий типа порождающего проектирования (generative design) [12] для сложных систем.

3 Монада диаграмм

Как известно [6, 7], функтор $\mathbf{D}$ определяет в $\mathbf{CAT}$ монаду, которая по ряду свойств аналогична монаде степени $\langle 2^{-}, \{-\}, \cup\rangle$ в категории множеств $\mathbf{Set}$. Компонента единицы монады D, соответствующая категории C — это полное вложение C в $\mathbf{D}C$, которое переводит объект A в диаграмму-точку $\ulcorner A \urcorner$. Умножение задает «отрисовку» $\mathbf{D}C$-диаграммы в виде C-диаграммы: отрисовка $\mathbf{K}\Xi$ диаграммы $\Xi : I \to \mathbf{D}C$ порождается заменой каждой точки $i \in \mathbf{Ob}\, I$ C-диаграммой Ξi и разделением каждого морфизма диаграмм Ξk, $k \in \operatorname{Mor} I$, на составляющие его морфизмы из категории C. Таким образом, точка диаграммы $\mathbf{K}\Xi$ — это пара (l, i), $i \in \operatorname{Ob} I$, $l \in \operatorname{Ob} \operatorname{dom} \Xi i$, помеченная объектом $(\Xi i)l$, а стрелка из нее в (l', i') — это пара $(q : \mathfrak{h}l \to l', k : i \to i')$, помеченная морфизмом $(\Xi i')q \circ \theta_l$, где $\langle \theta, \mathfrak{h}\rangle = \Xi k$ (рис. 3).

Формально отрисовка вычисляется при помощи ковариантной конструкции Гротендика для композиции Ξ с функтором формы $\mathbf{D}!_C$ и каноническим вложением $\mathbf{Cat}$ в $\mathbf{CAT}$. Приведем композицию соответствующих декартовых квадратов (рис. 4). Здесь правый декартов квадрат задает универсальное расслоение для малых категорий. Центральный декартов квадрат приводит к категории диаграмм с отмеченной точкой $\dot{\mathbf{D}}C$ — это конструкция Гротендика для конструкции Гротендика $\mathbf{D}C$. Ее объектом служит любая пара (x, Δ), состоящая из диаграммы $\Delta : X \to C$ и точки $x \in \operatorname{Ob} X$, а морфизмом такой пары в $(x', \Delta' : X' \to C)$ служит любая тройка $\langle q, \gamma, \mathfrak{f}\rangle$, где $\langle \gamma, \mathfrak{f}\rangle : \Delta \to \Delta'$ — морфизм диаграмм и $q : \mathfrak{f}x \to x'$ — стрелка в X'; закон композиции морфизмов имеет вид:

$$\langle q, \gamma, \mathfrak{f}\rangle \circ \langle r, \varphi, \mathfrak{g}\rangle = \langle q \circ \mathfrak{f}r, \gamma\mathfrak{g} \circ \varphi, \mathfrak{f}\mathfrak{g}\rangle.$$

Имеется канонический функтор $\boldsymbol{b}_C : \dot{\mathbf{D}}C \to \mathbf{D}C$, забывающий отмеченную точку. Кроме того, имеется функтор означения $\boldsymbol{e}_C : \dot{\mathbf{D}}C \to C$, переводящий пару (x, Δ) в Δx и морфизм $\langle q, \gamma, \mathfrak{f}\rangle : (x, \Delta) \to (x', \Delta')$ в $\Delta' q \circ \gamma_x : \Delta x \to \Delta' x'$. Отложим более подробное рассмотрение конструкции $\dot{\mathbf{D}}C$ до следующего раздела; приведенных здесь сведений о ней достаточно для проверки того, что левый декартов квадрат задает искомую отрисовку диаграммы Ξ.

Имеется вложение диаграмм $[\langle 1_{\Xi i}, \boldsymbol{i}_i : \operatorname{dom} \Xi i \hookrightarrow \operatorname{dom} \mathbf{K}\Xi\rangle]_{i \in \operatorname{Ob} I} : \coprod_{i \in \operatorname{Ob} I} \Xi i \hookrightarrow \mathbf{K}\Xi$, биективное на точках. Если диаграмма I дискретна, то это вложение является изоморфизмом, а отрисовка — вершиной копредела диаграммы Ξ.

Определим отрисовку произвольного морфизма $\mathbf{D}C$-диаграмм $\phi : \Xi \to \Xi'$. Рассмотрим $\mathbf{DD}C$-диаграмму со схемой $\mathbf{2}$ (это граф $0 \to 1$), единственная нетождественная стрелка которой помечена морфизмом ϕ. Двукратная отрисовка этой диаграммы дает C-диаграмму со схемой, снабженной забывающим функтором в $\mathbf{2}$, прообраз нетождественной стрелки относительно которого позволяет восстановить морфизм $\mathbf{K}\phi : \mathbf{K}\Xi \to \mathbf{K}\Xi'$. В явной

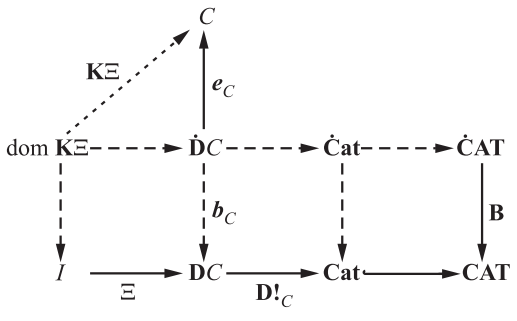


Рис. 4 Вычисление отрисовки

форме пусть $\phi = \langle \varphi, \mathfrak{g} \rangle$, где естественное преобразование $\varphi : \Xi \to \Xi' \mathfrak{g}$ составлено из компонентов $\varphi_i = \langle \gamma^i, \mathfrak{f}^i \rangle : \Xi i \to \Xi' \mathfrak{g} i$, $i \in \mathrm{Ob}\, I$. Рассмотрим функтор $\mathfrak{e} : \mathrm{dom}\, \mathbf{K}\Xi \to \mathrm{dom}\, \mathbf{K}\Xi'$, сопоставляющий точке (l, i) точку $(\mathfrak{f}^i l, \mathfrak{g} i)$, а стрелке $(q, k) : (l, i) \to (l', i')$ — пару $(\mathfrak{f}^{i'} q, \mathfrak{g} k)$, которая задает стрелку в диаграмме $\mathbf{K}\Xi'$, направленную из $\mathfrak{e}(l, i)$ в $\mathfrak{e}(l', i')$, ввиду соотношения $\langle \gamma^{i'}, \mathfrak{f}^{i'} \rangle \circ \Xi k = \Xi' \mathfrak{g} k \circ \langle \gamma^i, \mathfrak{f}^i \rangle$. Семейство морфизмов $\gamma^i_l : (\Xi i) l \to (\Xi' \mathfrak{g} i)(\mathfrak{f}^i l)$, $(l, i) \in \in \mathrm{Ob}\, \mathrm{dom}\, \mathbf{K}\Xi$, составляет естественное преобразование $\mathbf{K}\Xi$ в $(\mathbf{K}\Xi')\mathfrak{e}$, которое в паре с функтором $\mathfrak{e}$ и образует морфизм $\mathbf{K}\phi$. Непосредственно проверяется, что таким путем действительно задается функтор $\mathbf{K}$ и все такие функторы образуют естественное преобразование $\mathbf{DD}$ в $\mathbf{D}$, удовлетворяющее аксиомам монады.

Рассмотрим алгебры монады диаграмм. Свободные алгебры порождаются компонентами отрисовки $\mathbf{K}$. Единица также вносит свой вклад: если ее компонента, соответствующая категории C, имеет левый сопряженный функтор (а правого сопряженного она иметь не может), то единица этого сопряжения состоит из копределов C-диаграмм, так что категория C кополна и, более того, указанный левый сопряженный $\mathrm{colim} : \mathbf{D}C \to C$ задает алгебру [7]. Поскольку функтор, сопряженный к заданному, определяется однозначно с точностью до изоморфизма [8, § IV.1], получается, что конструкция копредела «закодирована» в тривиальной процедуре построения одноточечных диаграмм.

Другим примером алгебры монады $\mathbf{D}$ служит $\int : \mathbf{D}(\mathbf{CAT}) \to \mathbf{CAT} : \Delta \mapsto \int \Delta$, устроенная аналогично свободной алгебре над $\mathbf{1}$. При помощи отрисовки строятся и другие алгебры, неизоморфные ни копределам, ни свободным алгебрам. Например, обозначим через $\mathbf{PrCAT}$ полную подкатегорию в $\mathbf{CAT}$, состоящую из всех тонких категорий (предпорядков). Она рефлективна: ее вложение в $\mathbf{CAT}$ имеет левый сопряженный $\mathbf{P} : \mathbf{CAT} \to \mathbf{PrCAT}$ с тождественной коединицей (для произвольной категории C, $\boldsymbol{P}C$ — это класс $\mathrm{Ob}\, C$, предупорядоченный бинарным предикатом $\mathrm{Mor}(-, -) \neq \varnothing$). Далее, пусть $\check{\mathbf{D}} : \mathbf{CAT} \to \mathbf{CAT}$ — подфунктор в $\mathbf{D}$, сопоставляющий категории C полную подкатегорию в $\mathbf{D}C$, состоящую из всех тонких C-диаграмм, и действующий на функторы так же, как $\mathbf{D}$. Если C — тонкая категория, то имеется алгебра $\boldsymbol{P}(\mathbf{K}-) : \mathbf{D}\check{\mathbf{D}}C \to \check{\mathbf{D}}C$.

Алгебры строятся и при помощи морфизмов монады $\mathbf{D}$ в другие монады [8, упражнение 6.2.3(б)]. Так, $\mathbf{D}$ имеет две подмонады-ретракта: одна из них сопоставляет произвольной категории C категорию всех дискретных C-диаграмм, а другая — категорию всех C-диаграмм в форме группоидов. Следовательно, если в C есть суммы, то существуют две в общем случае неизоморфные друг другу алгебры над C: одна сопоставляет произвольной C-диаграмме сумму всех ее вершин, а вторая — сумму всех вершин ее скелета.

Что касается приложений монады диаграмм, то компонента ее единицы задает представление любого объекта как бесструктурной сингулярной системы, а левый сопряженный к ней (при его наличии) представляет сборку систем в цельные объекты. Функтор отрисовки строит детальную «плоскую» структуру многоуровневых систем, состоящих из систем (systems of systems, SoS), путем разрешения взаимосвязей верхнего уровня, причем ассоциативность умножения монады гарантирует независимость итоговой структуры от порядка анализа промежуточных уровней. Алгебры задают шаблоны различных процедур «упаковки» систем в сложные объекты с соблюдением условий естественности относительно модификаций [7]. Из «материала» монады диаграмм строятся и другие конструкции.

4 Монада и комонада систем систем

Существует сопряжение функторов, связывающее конструкцию Гротендика с функтором формы диаграмм [6]. Рассмотрим категорию $\mathbf{CAT}/\mathbf{Cat}$, в которой объектами служат все функторы вида $F : D \to \mathbf{Cat}$, а морфизмом такого функтора в $F' : D' \to \mathbf{Cat}$ служит любой функтор $G : D \to D'$, такой что $F'G = F$ (эту категорию можно построить посредством декартова квадрата в $\mathbf{CAT}$ [13]). Поскольку $!_{C'} H = !_C$ для произвольного функтора $H : C \to C'$, функтор формы диаграмм индуцирует функтор

$$\mathbf{D}_! : \mathbf{CAT} \to \mathbf{CAT}/\mathbf{Cat} : C \mapsto \mathbf{D}!_C, \quad H \mapsto \mathbf{D}H.$$

В свою очередь, любой функтор $F : D \to \mathbf{Cat}$ в композиции с каноническим вложением категории $\mathbf{Cat}$ в $\mathbf{CAT}$ дает функтор, обозначаемый через $\hat{}F : D \to \mathbf{CAT}$, для которого определена конструкция Гротендика $\int \hat{} F$. Отображение $F \mapsto \int \hat{} F$ служит функцией объектов функтора $\int \hat{} : \mathbf{CAT}/\mathbf{Cat} \to \mathbf{CAT}$, который переводит морфизм G в единственную стрелку из $\int \hat{} F$ в $\int \hat{} F'$, делающую коммутативной диаграмму, в которой все пунктирные стрелки изображают ребра декартовых квадратов (рис. 5).

В явном виде имеем

$$\int \hat{} G : \int \hat{} F \to \int \hat{} F' : (A, X) \mapsto (A, GX),$$
$$(f, g) \mapsto (f, Gg).$$

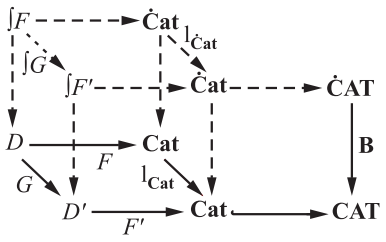


Рис. 5 Конструкция Гротендика как функтор

Как показывает диаграмма (см. рис. 4), описывающая построение отрисовки $\mathbf{K}\Xi$ (см. композицию правого и центрального декартовых квадратов), категория $\int\hat{}(\mathbf{D}!_C)$ совпадает с категорией диаграмм с отмеченной точкой $\dot{\mathbf{D}}C$. Тем самым определен эндофунктор

$$\dot{\mathbf{D}} = \int\hat{}\,\mathbf{D}_! : \mathbf{CAT} \to \mathbf{CAT} : C \mapsto \dot{\mathbf{D}}C\,,$$

действующий на функторы так же, как $\mathbf{D}$, не затрагивая отмеченные точки диаграмм. Более того, функтор $\int\hat{}$ сопряжен слева к функтору $\mathbf{D}_!$, причем коединица этого сопряжения $\boldsymbol{e} : \int\hat{}\,\mathbf{D}_! = \dot{\mathbf{D}} \to 1_{\mathbf{CAT}}$ состоит из функторов означения $\boldsymbol{e}_C$, $C \in \mathrm{Ob}\,\mathbf{CAT}$ (так что верхний треугольник на вышеупомянутой диаграмме — это частный случай треугольного тождества сопряжения, что свидетельствует об универсальности конструкции отрисовки). Компонента единицы $\boldsymbol{t} : 1_{\mathbf{CAT}/\mathbf{Cat}} \to \mathbf{D}_! \int\hat{}$, соответствующая функтору $F : D \to \mathbf{Cat}$, представляет собой функтор $\boldsymbol{t}_F : D \to \mathbf{D}\int\hat{}F$, сопоставляющий объекту X диаграмму $\boldsymbol{i}_X : FX \hookrightarrow \int\hat{}F$, а морфизму $g : X \to X'$ — морфизм $\langle \rho, Fg\rangle : \boldsymbol{i}_X \to \boldsymbol{i}_{X'}$, где естественное преобразование $\rho : \boldsymbol{i}_X \to \boldsymbol{i}_{X'}(Fg)$ составлено из компонентов $\rho_A = (1_{(Fg)A}, g) : (A, X) \to ((Fg)A, X')$, $A \in \mathrm{Ob}\,FX$; ясно, что $(\mathbf{D}!_{\int\hat{}F})\boldsymbol{t}_F = F$.

Это сопряжение стандартным способом [8, § VI.1] задает комонаду $\dot{\mathbf{D}}$ в $\mathbf{CAT}$, коединица которой состоит из функторов $\boldsymbol{e}_C$, а коумножение состоит из «размножителей диаграмм» вида

$$\mathbf{M} = \int\hat{}\,\boldsymbol{t}_{\mathbf{D}!_C} : \dot{\mathbf{D}}C \to \dot{\mathbf{D}}\dot{\mathbf{D}}C : (x, \Delta : X \to C) \mapsto$$
$$\mapsto (x, \boldsymbol{i}_\Delta : X \hookrightarrow \dot{\mathbf{D}}C), \langle q, \gamma, \mathfrak{f}\rangle \mapsto \langle q, \gamma^{\mathbf{M}}, \mathfrak{f}\rangle,$$

где $\gamma^{\mathbf{M}} : \boldsymbol{i}_\Delta \to \boldsymbol{i}_{\Delta'}\mathfrak{f}$ — естественное преобразование, сопоставляющее каждому $x \in \mathrm{Ob}\,X$ морфизм $\langle 1_{\mathfrak{f}x}, \gamma, \mathfrak{f}\rangle : (x, \Delta) \to (\mathfrak{f}x, \Delta')$. Тем самым $\mathbf{M}$ размещает в каждой точке x диаграммы Δ копию Δ, у которой точка x отмечена (рис. 6).

Рассмотрим коалгебры комонады $\dot{\mathbf{D}}$. Воспользуемся конструкцией сравнивающего функтора [8, § VI.3]: таковым служит функтор K, действующий из $\mathbf{CAT}/\mathbf{Cat}$ в категорию коалгебр, сопоставляя функтору $F : D \to \mathbf{Cat}$ коалгебру $\int\hat{}\,\boldsymbol{t}_F : \int\hat{}F \to \dot{\mathbf{D}}\int\hat{}F$. Любая коалгебра комонады $\dot{\mathbf{D}}$ изоморфна коалгебре такого вида [6]. В частности, $K(\mathbf{D}!_C) = \mathbf{M} : \dot{\mathbf{D}}C \to \dot{\mathbf{D}}\dot{\mathbf{D}}C$ (это косвободная коалгебра). А если F переводит любой объект в $\mathbf{1}$, т. е. $F = \ulcorner\mathbf{1}\urcorner!_D$, то конструкция Гротендика дает D и коалгебра $K(\ulcorner\mathbf{1}\urcorner!_D) : D \to \dot{\mathbf{D}}D$ переводит объект A в диаграмму-точку $\ulcorner A\urcorner$ с единственно возможной отмеченной точкой. В свою очередь, если F — это некоторая диаграмма-точка $\ulcorner I\urcorner : \mathbf{1} \to \mathbf{Cat}$, то $\int\hat{}\ulcorner I\urcorner = I$ и $K(\ulcorner I\urcorner) : I \to \dot{\mathbf{D}}I$: $i \mapsto (i, 1_I)$, $k \mapsto \mapsto \langle k, 1_{1I}, 1_I\rangle$.

Кроме комонады, эндофунктор $\dot{\mathbf{D}}$ определяет в $\mathbf{CAT}$ и монаду, устроенную аналогично монаде диаграмм: компонента единицы совпадает с $K(\ulcorner\mathbf{1}\urcorner!_C)$, а умножение переводит $\dot{\mathbf{D}}C$-диаграмму Ξ с отмеченной точкой i в отрисовку $\mathbf{K}(\boldsymbol{b}_C\Xi)$ с отмеченной точкой (l, i), где l — отмеченная точка Ξi (т. е. $\Xi i = (l, \boldsymbol{b}_C(\Xi i))$). Тем самым семейство функторов $\boldsymbol{b}_C$, $C \in \mathrm{Ob}\,\mathbf{CAT}$, образует морфизм монады $\dot{\mathbf{D}}$ в $\mathbf{D}$. В частности, любая алгебра $a : \mathbf{D}C \to C$ монады $\mathbf{D}$ порождает алгебру $a\boldsymbol{b}_C : \dot{\mathbf{D}}C \to C$ монады $\dot{\mathbf{D}}$. Легко проверить, что алгебру задает также любой функтор означения $\boldsymbol{e}_C : \dot{\mathbf{D}}C \to C$.

С прикладной точки зрения конструкция $\dot{\mathbf{D}}$ формально выражает первый принцип системной инженерии: целевая система (system of interest, SoI) рассматривается в контексте как элемент некоторой известной объемлющей системы [14]. Именно из такого рассмотрения возникают требования к це-

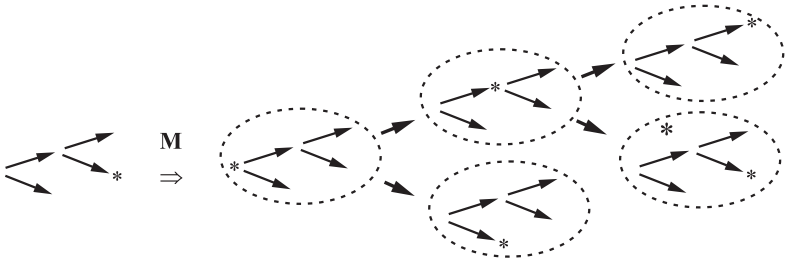


Рис. 6 Коумножение в комонаде диаграмм с отмеченной точкой

левой системе и начинается ее жизненный цикл. А структурная схема объемлющей системы с составляющими из каталога C, снабженная указанием места целевой системы в ней — это и есть объект категории $\dot{\mathbf{D}}C$. Разумеется, целевая система, в свою очередь, служит объемлющей для своих подсистем, так что возникают приложения конструкций вида $\dot{\mathbf{D}}^n C$, где $n > 1$ — число рассматриваемых системных уровней, в инженерии систем систем (SoS).

Монада $\dot{\mathbf{D}}$ показывает, как корректно учесть вхождение целевой системы при проведении с объемлющими системами процедур структурного анализа и синтеза из разд. 3. А комонада $\dot{\mathbf{D}}$ предоставляет специфические средства для работы с системой в контексте: коединица извлекает из объемлющей систему целевую, «забывая» все остальное, а коумножение порождает самоподобные многоуровневые системы, в которых каждый элемент воспроизводит структуру предыдущего уровня с сохранением своей идентичности. Самоподобие относится к ключевым свойствам систем — как природных, вплоть до Вселенной как целого [15], так и технических, например в сфере телекоммуникаций [16]. Генерация самоподобных структур расширяет возможности автоматического синтеза систем, в том числе бионического типа.

5 Заключение

При помощи универсальных конструкций в **CAT** (в основном декартовых квадратов, в том числе конструкции Гротендика, и сопряжений) в настоящей работе построены математические объекты, в которых «закодированы» ключевые процедуры системной инженерии. Показана основополагающая роль монады диаграмм. Целесообразно рекомендовать ее и производные конструкции к реализации в «умных» инструментах цифровой системной инженерии, способных автоматически генерировать оптимальные структуры систем и процессов.

Литература

1. *Левенчук А. И.* Системноинженерное мышление. — М.: TechInvestLab, 2015. 305 с.
2. *Mabrok M. A., Ryan M. J.* Category theory as a formal mathematical foundation for model-based systems engineering // Appl. Math. Inform. Sci., 2017. Vol. 11. No. 1. P. 43–51. doi: 10.18576/amis/110106.
3. *Breiner S., Subrahmanian E., Jones A.* Categorical foundations for system engineering // Disciplinary convergence in systems engineering research / Eds. A. Madni, B. Boehm, R. Ghanem, D. Erwin, D. Wheaton. — Springer, 2018. P. 449–463. doi: 10.1007/978-3-319-62217-0_32.
4. *Watson M. D.* Future of systems engineering // INCOSE INSIGHT, 2019. Vol. 22. Iss. 1. P. 8–12. doi: 10.1002/inst.12231.
5. *Ковалёв С. П.* Методы теории категорий в модельно-ориентированной системной инженерии // Информатика и её применения, 2017. Т. 11. Вып. 3. С. 42–50. doi: 10.14357/1992226417030.
6. *Guitart R., van den Bril L.* Décompositions et lax- complétions // Cahiers Topologie Géométrie Différentielle Catégoriques, 1977. Vol. 18. No. 4. P. 333–407.
7. *Ковалёв С. П.* Теория категорий как математическая прагматика модельно-ориентированной системной инженерии // Информатика и её применения, 2018. Т. 12. Вып. 1. С. 95–104. doi: 10.14357/19922264180112.
8. *Маклейн С.* Категории для работающего математика / Пер. с англ. — М.: Физматлит, 2004. 352 с. (*Mac Lane S.* Categories for the working mathematician. — Springer, 1978. 317 p.)
9. Grothendieck construction. — nLab, 2020. https://ncatlab.org/nlab/show/Grothendieck+construction.
10. *Barr M., Wells C.* Category theory for computing science. — London: Prentice Hall, 1990. 538 p.
11. *Ковалёв С. П.* Алгебраические методы порождающего проектирования крупномасштабных технических систем // Управление развитием крупномасштабных систем: Мат-лы XII Междунар. конф. — М.: ИПУ РАН, 2019. С. 384–386.
12. *Kowalski J.* CAD is a lie: Generative design to the rescue. — San Rafael, CA, USA: Autodesk, 2016. https://www.autodesk.com/redshift/generative-design.
13. Comma category. — nLab, 2019. https://ncatlab.org/nlab/show/comma+category.
14. *Hitchins D.* What are the general principles applicable to systems? // Incose Insight, 2009. Vol. 12. Iss. 4. P. 59–64. doi: 10.1002/INST.200912459.
15. *Iovane G., Laserra E., Tortoriello F. S.* Stochastic self-similar and fractal universe // Chaos Soliton. Fract., 2004. Vol. 20. Iss. 3. P. 415–426. doi: 10.1016/j.chaos.2003.08.004.
16. *Larijani H.* Local area networks and self-similar traffic // Network performance engineering / Ed. D. D. Kouvatsos. — Lecture notes in computer science ser. — Springer, 2011. Vol. 5233. P. 174–190. doi: 10.1007/978-3-642-02742-0_8.

Поступила в редакцию 25.02.21

THE MONAD OF DIAGRAMS AS A MATHEMATICAL METAMODEL OF SYSTEMS ENGINEERING

S. P. Kovalyov

V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 65 Profsoyuznaya Str., Moscow 117997, Russian Federation

Abstract: The paper addresses issues associated with the development of advanced mathematical methods for systems engineering suitable as the basis of computer tools for automatic synthesis and analysis of systems and processes. Following recent trends, category theory is employed as the framework for the methods. Its application is based on representing the structure of systems, processes, requirements, and other system design results as diagrams in categories whose objects are the algebraic models of parts and morphisms describe relationships between parts. Applying the fundamental Grothendieck flattening construction, the following constructions are described explicitly: categories of diagrams, the monad of diagrams, and the monad and the comonad of pointed diagrams. Application areas of these constructions in systems engineering procedures are identified. An approach is proposed to implement highly automated technologies of the generative design kind for complex multilevel systems.

Keywords: category theory; monad of diagrams; Grothendieck construction; colimit; systems engineering; system of systems; generative design

DOI: 10.14357/19922264230202 **EDN:** FDUVDU

References

1. Levenchuk, A. I. 2015. *Sistemnoinzhenernoe myshlenie* [Systems engineering thinking]. Moscow: TechInvestLab. 305 p.
2. Mabrok, M. A., and M. J. Ryan. 2017. Category theory as a formal mathematical foundation for model-based systems engineering. *Appl. Math. Inform. Sci.* 11(1):43–51. doi: 10.18576/amis/110106.
3. Breiner, S., E. Subrahmanian, and A. Jones. 2018. Categorical foundations for system engineering. *Disciplinary convergence in systems engineering research*. Eds. A. Madni, B. Boehm, R. Ghanem, D. Erwin, and D. Wheaton. Springer. 449–463. doi: 10.1007/978-3-319-62217-0_32.
4. Watson, M. D. 2019. Future of systems engineering. *INCOSE INSIGHT* 22(1):8–12. doi: 10.1002/inst.12231.
5. Kovalyov, S. P. 2017. Metody teorii kategoriy v model'no-orientirovannoy sistemnoy inzhenerii [Methods of category theory in model-based systems engineering]. *Informatika i ee Primeneniya — Inform. Appl.* 11(3):42–50. doi: 10.14357/1992226417030.
6. Guitart, R., and L. van den Bril. 1977. Decompositions et lax-completions. *Cahiers Topologie Geometrie Differentielle Categoriques* 18(4):333–407.
7. Kovalyov, S. P. 2018. Teoriya kategoriy kak matematicheskaya pragmatika model'no-orientirovannoy sistemnoy inzhenerii [Category theory as a mathematical pragmatics of model-based systems engineering]. *Informatika i ee Primeneniya — Inform. Appl.* 12(1):95–104. doi: 10.14357/19922264180112.
8. Mac Lane, S. 1978. *Categories for the working mathematician*. New York, NY: Springer. 317 p.
9. nLab. 2020. Grothendieck construction. Available at: https://ncatlab.org/nlab/show/Grothendieck+construction/ (accessed May 29, 2023).
10. Barr, M., and C. Wells. 1990. *Category theory for computing science*. London: Prentice Hall. 538 p.
11. Kovalyov, S. P. 2019. Algebraicheskie metody porozhdayushchego proektirovaniya krupnomasshtabnykh tekhnicheskikh sistem [Algebraic methods of generative design of large-scale technical systems]. *12th Conference (International) "Management of Large-Scale System Development" Proceedings*. Moscow: IPU RAN. 384–386.
12. Kowalski, J. 2015. CAD is a lie: Generative design to the rescue. San Rafael, CA: Autodesk. Available at: https://www.themanufacturer.com/articles/cad-is-a-lie-generative-design-to-the-rescue/ (accessed May 29, 2023).
13. nLab. 2019. Comma category. Available at: https://ncatlab.org/nlab/show/comma+category/ (accessed May 29, 2023).
14. Hitchins, D. 2009. What are the general principles applicable to systems? *Incose Insight* 12(4):59–64. doi: 10.1002/INST.200912459.
15. Iovane, G., E. Laserra, and F. S. Tortoriello. 2004. Stochastic self-similar and fractal universe. *Chaos Soliton. Fract.* 20(3):415–426. doi: 10.1016/j.chaos.2003.08.004.
16. Larijani, H. 2011. Local area networks and self-similar traffic. *Network performance engineering*. Ed. D. D. Kouvatsos. Lecture notes in computer science ser. Springer. 5233:174–190. doi: 10.1007/978-3-642-02742-0_8.

Received February 25, 2021

Contributor

Kovalyov Sergey P. (b. 1972) — Doctor of Science in physics and mathematics, leading scientist, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 65 Profsoyuznaya Str., Moscow 117997, Russian Federation; kovalyov@sibnet.ru

КОМПОЗИЦИОНАЛЬНОЕ ПРЕДСТАВЛЕНИЕ СТРУКТУРЫ ИГРЫ МНОГИХ ЛИЦ В МОНОИДАЛЬНОЙ КАТЕГОРИИ БИНАРНЫХ ОТНОШЕНИЙ

Н. С. Васильев[1]

Аннотация: Предложен системный подход к решению игры многих лиц, отвечающий современным сетевым технологиям. Он позволяет оптимизировать функционирование мультиагентных систем. Моноидальная категория бинарных отношений применяется как средство описания правил игры, исследования и модификации поведения игроков. Игровая проблема состоит в том, чтобы по возможности максимизировать отношения предпочтения всех участников игры. В соответствии с правилами игры их композиция определяет результирующее отношение игры (РОИ). Поиск рационального поведения игроков сведен к нахождению максимальных элементов РОИ. Формализовано использование разнообразных классов допустимых стратегий, процессов обмена информацией между игроками и формирование коалиций. Доказано существование РОИ и изучена структура его максимальных элементов, сокращающая поиск. Выяснено значение отношений предшествования ходов и абсолютно оптимальных предпочтений игроков в процессе формирования коалиций.

Ключевые слова: отношения предпочтения: абсолютно оптимальное, гарантированное, предшествования ходов; граф игры; допустимая стратегия; рациональное решение; характеристическое отношение коалиции; результирующее отношение игры; моноидальная категория; композициональность

DOI: 10.14357/19922264230203 **EDN:** GPMZTS

1 Введение

Игровой задаче со многими участниками присуще большое разнообразие постановок, которые приходится динамически уточнять в процессе исследования, проводя анализ результатов возможных рациональных решений [1–5]. Востребовано новое, композициональное, представление игры, отвечающее сетевым технологиям мультиагентных систем, программно реализуемое для оптимизации их работы [6]. Предложено формализовать игру средствами моноидальной категории бинарных отношений [7, 8], так как традиционные нормальная форма и развернутое представление игры не обладают нужным качеством [2, 9]. В отличие от применения платежных функций теперь решения принимаются на базе отношений на множестве ситуаций игры, которые учитывают правила игры и динамику поведения партнеров. Сходными соображениями руководствуются в многокритериальных задачах [3] и экономических приложениях теории игр [4, 10, 11].

Переход от нормальной к композициональной форме игры

Пусть каждый участник игровой задачи $i \in I = \{1, 2, \ldots, n\}$ стремится по возможности максимизировать свой критерий эффективности $w_i : X \to \to R, X \equiv X_1 \times X_2 \times \cdots \times X_n$, выбирая свой контролируемый фактор $x_i \in X_i$, $i \in I$ [1, 2, 5]. Поиск рационального решения игры на множестве ситуаций $x \in X$ проводится в условиях неопределенности, обусловленной различием интересов игроков и невозможностью точного прогноза результата игры из-за незнания действий партнеров. Поэтому агенты вынуждены расширять классы допустимых стратегий $\tilde{X}_i$, $X_i \subset \tilde{X}_i$, чтобы учитывать поступающую информацию и коалиционное поведение игроков [2, 5].

Сравнение ситуаций игры можно проводить не только с помощью критериев эффективности $w_i = = w_i(x)$, но и применяя отношения предпочтения игроков $\rho_i \subset X^2$, $i \in I$. Точное знание интересов моделируется линейным порядком, а нечеткие представления — бинарными отношениями общего вида. По-прежнему игроки стремятся по возможности выбирать максимально предпочтительные для них ситуации игры.

Рациональное поведение агентов определено правилами игры, порождающими из исходных ρ_i, $i \in I$, вспомогательные бинарные отношения, которые выражают принципы оптимального поведения, открытые в теории игр [1–5], и руководят действиями игроков. Производные отноше-

[1] Московский государственный технический университет им. Н. Э. Баумана, nik8519@yandex.ru

ния учитывают кооперацию и информированность игроков. Они конструируются посредством алгебраических операций $A = (\circ, \cup, \cap, {}^{\mathrm{op}}, \times; \sigma, \varnothing)$ и композиции морфизмов категории бинарных отношений REL [7, 8]. Так, обмен информацией между агентами всегда приводит к сужению их отношений предпочтения $\rho|_A = \rho \cap A^2$ на некоторую часть A множества ситуаций игры.

Пример 1.1. В бескоалиционной игре $\Gamma = \Gamma^0$ участники не имеют никакой информации о стратегиях партнеров и отсутствует договоренность об очередности ходов. Тогда *результирующим* отношением, которое в конечном итоге требуется оптимизировать, будет следующая дизъюнктивная сумма [7, 8]:

$$\rho^{\Gamma} = \coprod_{i \in I} \rho_i. \qquad (1)$$

Коалиции игроков C создаются за счет коммуникации. Им отвечают некоторые подыгры Γ' исходной игры Γ, в которых фиксированы стратегии агентов $i \neq C$. Интересы коалиции представлены *характеристическим* отношением ρ_C, совпадающим с РОИ Γ'. Конструкция ρ_C определяется исходя из правил игры и принципа рационального поведения участников коалиции.

Пример 1.2. В коалиции Парето $C = C^P$ все участники обладают полной информацией о действиях друг друга и совместно выбирают общий контролируемый фактор $x_C = (x_i, i \in C)$, стремясь к максимизации следующего отношения [3]:

$$\rho^{\Gamma'} = \rho_C = \bigcap_{i \in C} \rho_i \,. \qquad (2)$$

Абсолютно оптимальным отношением i-го игрока назовем сужение его отношения предпочтения $a_i = \rho_i|_{x_i}$ при фиксированных значениях всех остальных параметров $x_j \in X_j$, $j \neq i$. Включение $(x, x') \in a_i$ означает достижимость более предпочтительной ситуации x' из x ходом i-го игрока. Для игры в нормальной форме этому понятию соответствует применение игроком абсолютно оптимальной стратегии [2].

Пример 1.3. В игре $\Gamma^{0,n}$ участники коалиции $C = C(\pi, \varnothing)$, $\pi = \{(l, l+1), l = 1, \ldots, n-1\}$, не обмениваясь данными ($D = \varnothing$), выполняют ходы в линейном порядке

$$\pi : \boxed{1} - - \to \boxed{2} - - \to \cdots - - \to \boxed{n}. \qquad (3)$$

Тогда предпочтения коалиции задаются характеристическим отношением

$$\rho^{\Gamma^{0,n}} = \rho_{C(\pi,\varnothing)} = a_n \circ a_{n-1} \circ \cdots \circ a_1 \,. \qquad (4)$$

В (4) использована композиция морфизмов категории REL [8]. Рациональное поведение всякого игрока $i \in C(\pi, \varnothing)$ — это использование абсолютно оптимальной стратегии, отвечающей стремлению выбрать ситуацию из множества $\mathrm{MAX}\, a_i$. Коалиция $C(\pi, \varnothing)$ сможет окончательно выбрать оптимальное решение x^*_C лишь на последнем $(n-1)$-м шаге, завершив построение суперпозиции (4). При этом произведение морфизмов берется в противоположном порядке π^{op}.

Категориальный подход согласует правила игры, процессы формирования коалиций, классы допустимых стратегий и поиск рационального решения (см. (1)–(4)).

Наличие отношения эквивалентности $x \sim x'$ на множестве X позволяет упростить игру за счет факторизации [7, 8]. Разбиение ситуаций игры на классы эквивалентности строится с помощью критериев $x \sim x' \Leftrightarrow w_i(x) = w_i(x')$. То же самое делается с применением отношения толерантности $\tau \subset X^2$, которое по определению обладает свойствами рефлексивности и симметричности.

Лемма 1.1. *Пусть семейство $\{x : x \tau y\}$ состоит из открытых множеств. Тогда компакт $X \subset R^n$ можно заменить конечным множеством.*

Игровая задача как максимизация результирующего отношения игры

Интересы игроков $i \in I$ представлены бинарными отношениями предпочтения $\rho_i \subset X^2$, $i \in I$, заданными на конечном множестве ситуаций игры X. Они рефлексивны и транзитивны. Все агенты ведут себя рационально. Их стратегии включают обмен данными с другими участниками конфликта, решение о вступлении в одну или несколько коалиций и выбор момента выполнения своего хода [2, 5]. *Ход* игрока — это принятие решения вида $\tilde{x}_i : X \to \to X_i$, сопровождаемого, возможно, сообщением стратегии $\tilde{x}_i$ одному или нескольким партнерам по игре.

Коллективные действия игроков порождают правила игры, которые согласуют классы допустимых стратегий, процессы формирования коалиций, вводят *отношение предшествования* ходов π и уточняют содержание данных, которыми обмениваются игроки [5] (см. примеры 1.1 и 1.3). Согласованность означает отсутствие противоречия в процессе принятия решений. Благодаря этому существует результирующее отношение игры $\rho^{\Gamma(S)}$. Композициональная структура $\rho^{\Gamma(S)} \subset X^2(S)$ формализует процесс поиска рационального решения. В нее входят модули, отвечающие подыграм Γ_r, $r = 1, 2, \ldots, R$, проходящим в коалициях $C = C_r$, представленных характеристическими отношениями ρ_{C_r}, $r = 1, 2, \ldots, R$. Все подыгры *наследуют*

правила исходной игры, в частности отношение предшествования ходов π_C.

Рациональное решение игры $\Gamma(S)$ в классе стратегий S — это выбор ситуации $x^* \in X(S)$, являющейся максимальным элементом *результирующего* отношения

$$x^* \in \mathrm{MAX}\,\rho^{\Gamma(S)}. \tag{5}$$

На множестве классов $\widehat{X}(S)$ эквивалентных ситуаций $x \sim x' \Leftrightarrow (x,x'),(x',x) \in \rho^{\Gamma(S)}$, введем фактор-отношение $\widehat{\rho}^{\Gamma(S)} \subset \widehat{X}^2(S)$ [7, 8].

Теорема 1.1. *Пусть отношение $\rho^{\Gamma(S)}$ рефлексивно и транзитивно. Тогда с точностью до эквивалентности существует рациональное решение игры.*

Доказательство. В частично упорядоченном множестве $(\widehat{X},\widehat{\rho}),\ \widehat{\rho} = \widehat{\rho}^{\Gamma(S)}$, возьмем произвольный элемент $\widehat{x} \in \widehat{X}$. Если $\widehat{x} = \mathrm{MAX}\,\widehat{\rho}^{\Gamma(S)}$, то теорема доказана. Искомая ситуация x^*, $x^* \in \widehat{x}$. Иначе решением игры будет последний элемент $\widehat{x}_k = \widehat{x}^*$ максимальной цепи $\mathrm{Ch} = \left(\widehat{x}_0, \widehat{x}_1, \ldots, \widehat{x}_k\right)$, в которой $\widehat{x}_0 = \widehat{x}$ и $\left(\widehat{x}_l\,\widehat{x}_{l+1}\right) \in \widehat{\rho}^{\Gamma(S)}$ для всех $l = 0,1,\ldots,k-1$.

Следствие 1.1. *Конечное ациклическое антисимметричное отношение содержит максимальный и минимальный элементы.*

Наряду с эффективностью (5) в игровых задачах используется принцип устойчивости выбираемой ситуации. *Равновесием* в коалиционной игре $\Gamma(S)$ назовем ситуацию

$$x^* \in \bigcap_r \mathrm{MAX}\,\rho_{\rho_{C_r}}. \tag{6}$$

Вообще говоря, принципы (5), (6) противоречат друг другу [1–5].

2 Характеристическое отношение коалиции

Всякий раз, когда имеется нетривиальное отношение предшествования ходов, возникают иерархически организованные коалиции [1, 2, 5] (см. пример 1.3). При этом их участники, вообще говоря, обмениваются информацией D. Иерархия внутри коалиции порождает подкоалиции. Коалиции передают другим игрокам $i \notin C$ некоторые данные $x_C \in D$, которые ранее сообщали их участники $k \in C$. Сообщение стратегий-функций будем выражать в форме включения $\tilde{x}_C \in \tilde{D}$.

Пример 2.1. В игре $\Gamma^{1,n}$ образуется коалиция $C(\pi, D)$, $D = \{\overline{x}_1, \ldots, \overline{x}_{n-1}\}$, в результате договоренности о порядке ходов (3) и передачи игроком $l = 1,\ldots,n-1$ величины $\overline{x}_l \in D$ партнеру $l+1$. Характеристическим отношением коалиции будет

$$\rho^{\Gamma^{1,n}} = \rho_{C(\pi,D)} = a_1 \circ a_2 \circ \cdots \circ a_n\,. \tag{7}$$

В обосновании формулы (7) лежат те же причины, что и у схемы (4). Чтобы определиться с выбором передаваемой величины $\overline{x}_l$, $l = 1,2,\ldots,n-1$, всякий игрок l учитывает рациональность поведения следующего игрока $l+1$. Поэтому суперпозиция (7) строится в порядке ходов π. Формулы (4) и (7) отвечают принципу *динамического* программирования.

Пример 2.2. В двухуровневой иерархической системе ведущий игрок 1 первым делает свой ход [1]. «Подчиненные» игроки $2,\ldots,n$ между собой не обмениваются информацией. В зависимости от того, передает первый агент значение $\overline{x}_1$ или нет, результирующее отношение соответствующей игры равно (см. (1), (4) и (7)):

$$D_1 = \varnothing \Rightarrow \rho^{\Gamma^{0;1,n-1}} = \left(\coprod_{l=2}^{n} a_l\right) \circ a_1$$

$$\text{либо } D_1 \neq \varnothing \Rightarrow \rho^{\Gamma^{1;1,n-1}} = a_1 \circ \left(\coprod_{l=2}^{n} a_l\right). \tag{8}$$

Подчиненные игроки могли бы объединиться в коалицию Парето C, сводя решаемую проблему к иерархической игре двух лиц (1 и C) с результирующим отношением $\rho_{\{1,C\}(\pi,\varnothing)} = a_C \circ a_1$ или $\rho_{\{1,C\}(\pi,D)} = a_1 \circ a_C$ в зависимости от того, множество $D_1 = \varnothing$ или нет. Здесь $a_C = \rho_C|_{x_C}$ — абсолютно оптимальное отношение коалиции (см. (2)).

Рассмотрим теперь игры $\Gamma^{2,n}$, $n \geq 2$, в которых партнерам последовательно передаются данные вида $\tilde{x}_i \in \tilde{D}_i$, $\tilde{x}_i : X \to X_i$, $i = 1,\ldots n-1$ [2, 5]. Обратная связь может приводить к ситуациям *равновесия* (6).

Пример 2.3. В обобщенной игре Гермейера $\Gamma^{2,2}$ [2] граф ходов и характеристическое отношение коалиции $C(\pi,\tilde{D})$ имеют следующий вид:

$$\pi,\tilde{D}: \boxed{1} \overset{\tilde{x}_1}{\to} \boxed{2} \sim \rho^{\Gamma^{2,2}} = \rho_{C(\pi,\tilde{D})} =$$

$$= \rho_2 \circ \left(\rho_1 \cup \rho_2^G\right), \quad \rho_2^G = \rho_2|_{\mathrm{MIN}\,\rho_2|_{X_1}}. \tag{9}$$

В формуле (9) использовано *гарантированное* отношение второго игрока $\rho_2^G \subset \rho_2$, с помощью которого сравниваются результаты «послойной» минимизации отношения $\rho_2|_{X_1}$ при любом фиксированном параметре $x_2 \in X_2$:

$$\rho_2^G \triangleq \{(x,x') : x\rho_2 x';\ \ x,x' \in \mathrm{MIN}\,\rho_2|_{X_1}\}. \tag{10}$$

В теореме 3.2 изучен общий случай системы $\Gamma^{2,n}$, $n \geq 2$, и доказана формула (9).

Пример 2.4. В двухуровневой иерархической игре $\Gamma^{1;1,n-1}$, $\pi = \{(1,2),\dots,(1,n-1)\}$, ведущий игрок 1 не может сообщить стратегию-функцию $\tilde{x}_1 : X_2 \times \cdots \times X_n \to X_1$ своим партнерам $\{2,3,\dots,n\}$. Отсутствие коммуникации между ними ведет к противоречию в процессе принятия решений: в подкоалициях $\{1,2\},\dots,\{1,n-1\}$ нельзя одновременно разыграть игры $\Gamma^{2,2}$. Класс функции $\tilde{D}_1$ недопустим. Поэтому требуется изменить правила игры: игрок 1 сообщает партнерам лишь значения $\overline{x}_1 \in D_1 \subset \tilde{D}_1$. Тогда в подкоалициях решаются подыгры $\Gamma^{1,2}$, а в целом — игра $\Gamma^{1;1,n-1}$ (см. (7) и (8)). Следовательно, интересы коалиции $C = C(\pi, \tilde{D}_1)$ характеризуются отношением

$$\rho^{\Gamma^{1;1,n-1}} = \rho_C = a_1 \circ a_2 \coprod \cdots \coprod a_1 \circ a_n\,, \qquad (11)$$

поэтому необходимо контролировать допустимость применяемых стратегий.

3 Композициональность игры

Развиваемый подход базируется на свойстве композициональности бинарных отношений, применяемых в игровых операциях. Основой служит моноидальная категория бинарных отношений REL [7, 8]. Объектами REL выступают конечные множества $X, Y, Z, \dots$, а морфизмами — бинарные отношения $\alpha : X \to Y, \beta : Y \to Z, \dots$, заданные на произведениях их областей и кообластей $X \times Y, Y \times Z, \dots$ Напомним [8], что композицией морфизмов $\alpha : X \to Y, \beta : Y \to Z$ в категории REL является суперпозиция $\alpha \circ \beta \subset X \times Z, \alpha \circ \beta : X \to Z$ этих отношений, равная

$$\alpha \circ \beta = \{(x,z) : \exists_{y \in Y}\, x\alpha y \wedge y\beta z\}\,. \qquad (12)$$

Вместо записи $(x,y) \in \alpha \subset X \times Y$ здесь использовано инфиксное обозначение $x\alpha y$. Единичными морфизмами для операции произведения (12) служат тривиальные порядки σ_X и σ_Y, а $\sigma_Z = \{(z,z) : z \in \in Z\}$ для любого множества Z.

В категории REL определена операция дизъюнктивной суммы морфизмов $\alpha \coprod \beta : (X \coprod Y) \to \to (Y \coprod Z)$, называемая также моноидальным произведением:

$$\alpha \coprod \beta = \{(x,y;1) : x\alpha y\} \cup \{(y,z;2) : y\beta z\}. \qquad (13)$$

Операции (12) и (13) ассоциативны, а сумма (13) обладает свойством коммутативности и имеет единицу $\varnothing$. Поэтому REL представляет собой моноидальную категорию [8].

В рассматриваемых игровых задачах у каждого морфизма совпадают области и кообласти, например это так у исходных отношений предпочтения игроков $\rho : X \to X$. При построении вспомогательных бинарных отношений посредством моноидального произведения $\rho_1 \coprod \rho_2$ изменяются области и кообласти получаемых морфизмов. Тем не менее всегда можно считать допустимыми произвольные композиции морфизмов. Так, под левыми частями выражений

$$\Big(\gamma_1 \coprod \gamma_2\Big) \circ \gamma = \gamma_1 \circ \gamma \coprod \gamma_2 \circ \gamma\,;$$
$$\gamma \circ \Big(\gamma_1 \coprod \gamma_2\Big) = \gamma \circ \gamma_1 \coprod \gamma \circ \gamma_2$$

следует понимать следующие композиции морфизмов соответственно:

$$\Big(\gamma_1 \coprod \gamma_2\Big) \circ \Big(\gamma \coprod \gamma\Big)\,;$$
$$\Big(\gamma \coprod \gamma\Big) \circ \Big(\gamma_1 \coprod \gamma_2\Big)\,.$$

Вообще говоря, РОИ $\rho^{\Gamma(S)} \subset Z^2$ определено на множестве $Z = X(S) \neq X$. Поэтому найденное решение задачи (5) нужно дополнительно преобразовать для получения допустимой ситуации игры $x^* \in X$. Это делается проектированием $Z \to X$ (см. определение (13)).

Задание правил игры

Договоренности между игроками могут приводить к формированию коалиций до начала их ходов. *Заранее* объявляемые коалиции Парето (см. пример 1.2) или иерархические коалиции (см. примеры 2.1–2.4) имеют приоритет при выполнении ходов, приводящий к изменению исходного отношения π. После этого коалиции трактуются как отдельные игроки с отношениями предпочтения ρ_C (см. (2) и (4)).

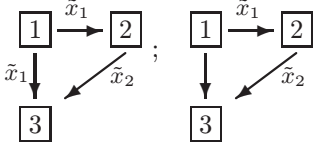
Выбираемое агентами отношение предшествования ходов π управляет созданием иерархически организованных коалиций в процессе игры Оно должно обладать свойствами антисимметричности, рефлексивности и ацикличности [5]. В конечном итоге игра проходит между коалициями разных типов, интересы которых представлены характеристическими отношениями. Проходящие внутри коалиций C подыгры наследуют отношение предшествования ходов, обозначаемое π_C.

Всякая коалиция C действует как игрок, передающий информацию $D_C = \{\tilde{x}_i, \overline{x}_j, i,j \in C, i \neq j\}$ одновременно всем участникам конфликта, которые были адресатами сообщений ее членов. Если адресаты $r,k,\dots$ вошли в некоторые коалиции $C' = C$, то данные $x_C = \{\tilde{x}_i, \overline{x}_j\}$ поступают игроку C' и распределяются по назначению.

В понятие стратегии входит участие игрока в разработке правил игры, инициировании момента *хода*, выборе контролируемого фактора и, возможно, коммуникации — сообщении информации некоторым из партнеров. Размеченный передаваемыми данными граф G_π отношения предшествования ходов назовем *графом игры* (см. (3) и (9)). Так как все участники конфликта заинтересованы в выборе рационального решения, то они используют лишь *допустимые* классы стратегий. Выбор графа игры должен сопровождаться анализом классов стратегий, которые намерены применять игроки, и в случае необходимости их корректировать. Изменение правил игры направлено на соблюдение требования непротиворечивости процесса поиска решения (см. пример 2.4).

Противоречие в обмене данными устраняется с помощью *допустимой* разметки графа игры. Никто из игроков не может одновременно сообщить партнерам и процедуру выбора фактора $\tilde{x}$, и его значение $\overline{x}$. Для адресатов $j \in T_i = \{j : i\pi j, i \neq j\}$ игрока i либо дуги $i \to j$ графа G_π вовсе не помечаются, если $\exists_l l \in T_i \wedge j\pi l$ (см., например, (3)), либо имеют разметку $\overline{x}_i$, если мощность $\#T_i > 1$, или $\tilde{x}_i$, если $\#T_i = 1$.

Пример 3.1. В левой части формулы



(14)

изображена недопустимая разметка графа G_π. В правой части (14) показан «исправленный» граф игры. Разметка $\tilde{x}_1$ дуги $1 \to 3$ допустима, но игнорирует заданный порядок ходов $(2, 3) \in \pi$.

Построение результирующего отношения игры

Пусть граф игры обладает допустимой разметкой дуг. Докажем, что имеется РОИ $\rho^{\Gamma(S)}$, которое однозначно определено в соответствии с правилами игры, при этом оно представляет собой композицию исходных и производных бинарных отношений в категории REL. Индуктивное конструирование $\rho^{\Gamma(S)}$ опирается на отношение предшествования ходов π, моделируя процесс формирования иерархических коалиций вследствие коммуникации игроков.

Теорема 3.1. *Во всякой игре существует результирующее отношение.*

Доказательство. Применим метод математической индукции, проводимой по числу участников игры. При $n \leq 2$ РОИ существует (см. примеры 1.1–1.3 и 2.1–2.4). Если $\pi = \sigma$, то для всех значений $n = 1, 2, \ldots$ результирующее отношение вычисляется по формуле (1).

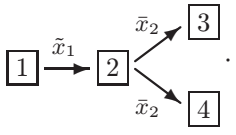
Пусть $\pi \neq \sigma$. Согласно следствию 1.1 и ацикличности отношения π найдется элемент $i \in \mathrm{MIN}\,\pi$, для которого $\exists_j (j \neq i) \wedge i\pi j$. Сформируем коалицию $C_1 = \{j : i\pi j\}$. По свойству рефлексивности π игрок $i \in C_1$. Воспользуемся предположением индукции: существует характеристическое отношение ρ_{C_1}. Не исключено, что при следующих ходах внутри коалиции C_1 будут образованы подкоалиции C_1^k, $k = 1, 2, \ldots, K$, с числом участников $r_k < n$. Тогда ρ_{C_1} строится композицией морфизмов $\rho_{C_1^k}$ категории REL.

Построим новый граф игры $G'_{\pi'}$. Пусть из вершины $i \in G_\pi$ исходит l, $l \geq 2$, дуг, которые имеют допустимую разметку. Заменим одним агентом коалицию C_1. Исключим из графа игры G_π все дуги $(i, j) \in \pi$, отвечающие коммуникациям внутри коалиции C_1. Из новой вершины $C_1 \in G'_{\pi'}$ проведем дуги $(C_1, r) \in \pi'$, где $(j, r) \in \pi$, $j \in C_1$. Таким образом, построен наследник π' отношения предшествования ходов π. Разметим дуги (C_1, r) так же, как и $(j, r) \in \pi$, обеспечив допустимость графа игры $G'_{\pi'}$. В Γ' число участников $m < n$. По предположению индукции существует результирующее отношение $\rho^{\Gamma'}$, строящееся композицией соответствующих морфизмов.

Композиционный процесс завершается построением бескоалиционной игры $\pi^{(k)} = \sigma$, результирующим отношением которой выступает моноидальное произведение (1). Проиллюстрируем теорему 3.1 на примере графа игры (14). Сначала в подкоалиции $C = \{1, 2\}$ строится характеристическое отношение (9), а затем для игры $C \to 3$ — результирующее (см. формулу (4)):

$$\rho^{\Gamma^{2,3(\pi')}} = a_3 \circ \left(\rho_2 \circ \left(\rho_1 \cup \rho_2^G\right)\right). \qquad (15)$$

Пример 3.2. Рассмотрим граф игры Γ, описывающий функционирование трехуровневой иерархической системы



.

На первом шаге построения отношения ρ^Γ формируется иерархическая коалиция $C_1 = \{1, 2\}$, в которой разыгрывается операция $\Gamma^{2,2}$ с параметрами x_3 и x_4. Ее характеристическое отношение ρ_{C_1} задается формулой (9). При втором ходе идет игра трех лиц $\{C_1, 3, 4\}$ (см. примеры 2.2 и 2.4 и формулу (11)).

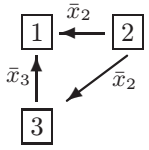
Поэтому

$$\rho^G = \rho_{C_1} \circ a_C = \\ = \left(\rho_2 \circ \left(\rho_1 \cup \rho_2^G\right)\right) \circ \left(\rho_3 \coprod \rho_4\right)\Big|_{X_3 \times X_4}. \quad (16)$$

Если в правилах игры оговорить заранее, что игроки 3 и 4 создадут коалицию Парето C^P (см. (2)), то второй ход приведет к игре двух лиц $\Gamma' = \left\{C_1, C^P, \pi' : C_1 \to C^P\right\}$. Неопределенность выбора рационального решения уменьшится по сравнению с оптимизацией (16), так как теперь исследуется результирующее отношение

$$\rho^{\Gamma'} = \rho_{C_1} \circ a_{C^P} = \left(\rho_2 \circ \left(\rho_1 \cup \rho_2^G\right)\right) \circ \rho_3 \cap \rho_4.$$

Пример 3.3. На первом шаге построения РОИ



участники объединяются в одну иерархическую коалицию C_1 с ведущим игроком 2. На втором шаге формируется подкоалиция $C_1^1 = \{1, 3\}$ с игроком 1. Согласно примеру 2.1, $\rho_{C_1^1} = a_3 \circ a_1$. Характеристическое отношение коалиции $C_1 = \{2, C_1^1\}$ равно $\rho^\Gamma = \rho_{C_1} = a_2 \circ \rho_{C_1^1} = a_2 \circ a_3 \circ a_1$ (см. (7)).

Обобщение игры Гермейера

Докажем формулы (9) и (10). Под *гарантированным* отношением игрока 2 будем понимать следующее сужение его отношения предпочтения:

$$\rho_2^\Gamma = \rho_2|_{\mathrm{MIN}\,\rho_2|_{X_1}}. \quad (17)$$

Пусть $\pi_{X_1} : X \to X_1$ — проектирование. Любая функция $\tilde{x}_1^{\text{н}} : X_2 \to \pi_{X_1}\mathrm{MIN}\,\rho_2|_{X_1}$ называется стратегией *наказания* 2-го игрока [2], а $x_2^G \in \in \pi_{X_2}\mathrm{MAX}\,\rho_2^G$ — *гарантирующей* стратегией.

Теорема 3.2. *Пусть ρ_2 — транзитивное отношение. Тогда в игре двух лиц $\Gamma^{2,2}$ результирующее отношение равно*

$$\rho^{\Gamma^{2,2}} = \rho_2 \circ \left(\rho_1 \cup \rho_2^G\right). \quad (18)$$

Доказательство. Пусть $x_2^* = \pi_{X_2} x^*$, $x^* \in \in \mathrm{MAX}\,\rho_1$. Если игрок 2 выберет $x_2 \neq x_2^*$, то игрок 1 применит стратегию наказания $\tilde{x}_1^{\text{н}}$. Поэтому перед игроком 2 стоит альтернатива: либо применять $x_2 = x_2^*$, либо быть наказанным. Угроза действенна лишь при условии $(\tilde{x}_1^{\text{н}}(x_2^G), x_2^G)\rho_2 x^*$, когда не помогает использование игроком 2 гарантирующей стратегии $x_2 = x_2^G$. В самом деле, определение ρ_2^G и транзитивность ρ_2 дают

$$(\tilde{x}_1^{\text{н}}(x_2), x_2)\,\rho_2\left(\tilde{x}_1^{\text{н}}(x_2^G), x_2^G\right) \wedge \left(\tilde{x}_1^{\text{н}}(x_2^G), x_2^G\right)\rho_2 x^* \Rightarrow \\ \Rightarrow (\tilde{x}_1^{\text{н}}(x_2), x_2)\,\rho_2 x^*.$$

«Выгоднее» для игрока 2 выбрать x_2^*, а не x_2^G. Отсюда $\rho^{\Gamma^{2,2}}$ имеет вид (18) (см. (9)).

Следствие 3.1. *Иерархическая игра $\Gamma^{2,n}$ с $n = 2k$ или $n = 2k+1$ участниками*

$$\boxed{1} \overset{\tilde{x}_1}{\to} \cdots \overset{\tilde{x}_r}{\to} \boxed{r+1} \overset{\tilde{x}_{r+1}}{\to} \cdots \overset{\tilde{x}_{n-1}}{\to} \boxed{n} \quad (19)$$

имеет следующие результирующие отношения:

$$\rho^{\Gamma^{2,2k}} = \rho_{2k} \circ \left(a_{2k-1} \circ \rho^{\Gamma^{2,2k-2}} \cup \rho_{2k}^G\right); \\ \rho^{\Gamma^{2,2k+1}} = a_{2k+1} \circ \rho^{\Gamma^{2,2k}}; \\ \rho^{\Gamma^{2,0}} = \sigma, \;\; k = 1, 2, \ldots \quad (20)$$

Доказательство. В соответствии с теоремой 3.1 с каждым ходом игроков происходит расширение коалиции $C_{k+1} = C_k \cup \{k+1\}$, $C_1 = \{1\}$, и поочередное применение формул (4) и (18) вычисления характеристических отношений.

Пример 3.4. Пусть в игре (19) имеются предварительные договоренности о формировании подкоалиций $C_{2k} = \{2k-1, 2k\}$, $\rho_{C_{2k}} = \rho^{\Gamma^{2,2}}|_{X_{2k-1} \times X_{2k}}$ (см. (18)). Тогда

$$\rho_C^{\Gamma^{2,2k}} = \rho_{2k} \circ \left(\rho_{2k-1} \cup \rho_{2k}^G\right) \circ \cdots \circ \rho_2 \circ \left(\rho_1 \cup \rho_2^G\right); \\ \rho_C^{\Gamma^{2,2k+1}} = a_{2k+1} \circ \rho_C^{\Gamma^{2,2k}}.$$

4 Поиск рационального решения игры

Композиционное строение отношения $\rho^{\Gamma(S)}$ сокращает перебор в задаче (5).

Теорема 4.1. *Имеют место свойства*

$$\left.\begin{array}{l} \mathrm{MAX}\rho_2|_{\mathrm{MAX}\,\rho_1} \subset \mathrm{MAX}\left(\rho_2 \circ \rho_1\right); \\ \mathrm{MAX}\left(\rho_1 \coprod \rho_2\right) = \mathrm{MAX}\,\rho_1 \cup \mathrm{MAX}\,\rho_2. \end{array}\right\} \quad (21)$$

Если отношения ρ_1 и ρ_2 транзитивны и рефлексивны, то $\mathrm{MAX}\,\rho_2|_{\mathrm{MAX}\,\rho_1} = \mathrm{MAX}\left(\rho_2 \circ \rho_1\right)$.

Доказательство. Вторая из формул (21) непосредственно следует из определения (13) дизъюнктивной суммы. Докажем первую. По определению максимального элемента $x^* = \mathrm{MAX}\,\rho_2|_{\mathrm{MAX}\,\rho_1}$ имеем

$$\forall_{x_1^*} \forall_x \left((x_1^*, x) \in \rho_1 \Rightarrow x = x_1^*\right) \wedge \\ \wedge \left((x^*, x_1^*) \in \rho_2 \Rightarrow x_1^* = x^*\right),$$

что эквивалентно формуле

$$\forall_{x_1^*}\forall_x\left((x_1^*,x)\in\rho_1\right)\wedge\left((x^*,x_1^*)\in\rho_2\right)\Rightarrow x=x^*.$$

Переписывая левую часть импликации в форме суперпозиции отношений, приходим к следующему выводу:

$$\forall_x\,(x^*,x)\in\rho_2\circ\rho_1\Rightarrow x=x^*.$$

Значит, доказано, что

$$\operatorname{MAX}\rho_2|_{\operatorname{MAX}\rho_1}\subset\operatorname{MAX}(\rho_2\circ\rho_1).$$

Пусть отношения ρ_1 и ρ_2 транзитивны и рефлексивны. Включение $x^*\in\operatorname{MAX}(\rho_2\circ\rho_1)$ равносильно формуле

$$\forall_x\,(x^*,x)\in\rho_2\circ\rho_1\Rightarrow x=x^*.$$

Согласно определению композиции (12), имеем

$$\forall_x\exists_{x_1^*}\,(x^*,x_1^*)\in\rho_2\wedge(x_1^*,x)\in\rho_1\Rightarrow x=x^*.$$

Как следствие, имеем

$$\forall_x\forall_{x_1^*}\,(x^*,x_1^*)\in\rho_2\wedge(x_1^*,x)\in\rho_1\Rightarrow x=x^*.$$

Опираясь на рефлексивность отношения ρ_2, подставим сюда $x_1^*=x^*$. Тогда верна формула

$$(x^*,x)\in\rho_1\Rightarrow x=x^*,$$

означающая $x^*\in\operatorname{MAX}\rho_1$. Рассуждая аналогично, подстановкой $x=x_1^*$ получим, что $x^*=\operatorname{MAX}\rho_2$ и тем более $x^*\in\operatorname{MAX}\rho_2|_{\operatorname{MAX}\rho_1}$. Итак, противоположное вложение $\operatorname{MAX}(\rho_2\circ\rho_1)\subset\operatorname{MAX}\rho_2|_{\operatorname{MAX}\rho_1}$ также имеет место.

Из соображений двойственности вытекает следующий результат.

Следствие 4.1. *Справедливы соотношения*:

$$\left.\begin{aligned}&\operatorname{MIN}\rho_2|_{\operatorname{MIN}\rho_1}\subset\operatorname{MIN}(\rho_1\circ\rho_2),\\&\operatorname{MIN}\rho_1\coprod\rho_2=\operatorname{MIN}\rho_1\cup\operatorname{MIN}\rho_2.\end{aligned}\right\}\qquad(22)$$

Если ρ_1 и ρ_2 — транзитивные и рефлексивные отношения, то $\operatorname{MIN}\rho_2|_{\operatorname{MIN}\rho_1}=\operatorname{MIN}(\rho_1\circ\rho_2)$.

В теореме 4.1 и следствии 4.1 обобщен метод динамического программирования применительно к бинарным отношениям. Воспользуемся свойством ассоциативности композиции морфизмов и формулами (4), (7), (18), (20)–(22) (см. также пример 2.4), и методом математической индукции докажем следующее

Следствие 4.2. *Рациональные решения игр* $\Gamma^{0,n}$, $\Gamma^{1,n}$; $\Gamma^{0;1,n-1}$, $\Gamma^{1;1,n-1}$, $\Gamma^{2,n}$ *удовлетворяют следующим свойствам*:

$$\left.\begin{aligned}&\operatorname{MAX}a_n|_{\operatorname{MAX}a_{n-1}|\cdots|_{\operatorname{MAX}a_1}}\subset\operatorname{MAX}\rho^{\Gamma^{0,n}};\\&\operatorname{MAX}a_1|_{\operatorname{MAX}a_2|\cdots|_{\operatorname{MAX}a_n}}\subset\operatorname{MAX}\rho^{\Gamma^{1,n}};\\&\bigcup_{l=2}^{n}\operatorname{MAX}a_l|_{\operatorname{MAX}a_1}=\operatorname{MAX}\rho^{\Gamma^{0;1,n-1}};\\&\operatorname{MAX}a_1|_{\bigcup_{l=2}^{n}\operatorname{MAX}a_l}=\operatorname{MAX}\rho^{\Gamma^{1;1,n-1}};\\&\operatorname{MAX}a_{2k+1}|_{\operatorname{MAX}\rho^{\Gamma^{2,2k}}}\subset\operatorname{MAX}\rho^{\Gamma^{2,2k+1}};\\&\operatorname{MAX}\rho_{2k}\big|_{\operatorname{MAX}a_{2k-1}\cup\operatorname{MAX}\rho_{2k}^{G}\big|_{\operatorname{MAX}\rho^{\Gamma^{2,2k-2}}}}\subset\\&\qquad\subset\operatorname{MAX}\rho^{\Gamma^{2,2k+2}}.\end{aligned}\right\}\qquad(23)$$

Пример 4.1. Рассмотрим игру $\Gamma^{2,3}$, в которой предпочтения участников являются рефлексивными транзитивными замыканиями следующих отношений:

$$\begin{aligned}\rho_1&=\{(0,1),(0,5),(4,0),(3,4),(3,2),(7,6)\};\\\rho_2&=\{(0,2),(1,0),(2,4),(3,2),3,5),(4,5),(6,4),\\&\qquad(6,7)\};\\\rho_3&=\{(2,0),(4,0),(3,5),(3,1),(6,4),(7,6)\},\end{aligned}$$

заданных на бинарном кубе $\underline{8}\simeq\{0,1\}^3$. Ситуации $(x_1,x_2,x_3)\in\underline{8}$ представлены числами $x\simeq x_1+2x_2+4x_3$, записанными в двоичной системе. Результирующее отношение игры имеет вид (15) (см. (20), $k=1$). Найдем рациональное решение x^*, сократив перебор возможных вариантов в задаче (5) с помощью формул (23).

Характеристическое отношение коалиции $\{1,2\}$ равно $\rho_2\circ(\rho_1\cup\rho_2^G)$ (см. (17) и (18)). Сначала вычислим $\rho_1\cup\rho_2^G=\{(0,1),(0,5),(4,0),(3,4),(3,2),(7,6)\}\cup\{(1,4),(3,4),(6,4)\}$, а затем — максимальные элементы отношений, входящих в композицию $\rho^{\Gamma^{2,3}}$:

$$\begin{aligned}M_1&\triangleq\operatorname{MAX}\rho_1\cup\rho_2^G=\{2,5\};\\M_2&\triangleq\operatorname{MAX}\rho_2|_{M_1}=\{5\};\\\operatorname{MAX}a_3|_{M_2}=\{5\}&\Rightarrow x^*=5\simeq(1,0,1).\end{aligned}$$

5 Заключение

Композициональная структура РОИ многих лиц выражает формализацию правил игры, процессов формирования коалиций и выбора игроками допустимых стратегий. Подобное представление игры сокращает перебор при поиске рациональных решений. С его помощью получено обобщение классической теоремы Гермейера. Модульность композиции результирующего отношения упрощает разработку и оптимизацию мультиагентных

систем. Предложенный метод исследования и численного решения игровой задачи нуждается в алгоритмизации.

Литература

1. *Моисеев Н. Н.* Элементы теории оптимальных систем. — М.: Наука, 1974. 526 с.
2. *Гермейер Ю. Б.* Игры с непротивоположными интересами. — М.: Наука, 1976. 326 с.
3. *Подиновский В. В., Ногин В. Д.* Парето-оптимальные решения многокритериальных задач. — М.: Наука, 1982. 256 с.
4. *Розен В. В.* Применение теории бинарных отношений к общей теории игр // Математические методы решения экономических задач. — Новосибирск: Наука, 1982. С. 127–152.
5. *Васильев Н. С.* Коалиционно устойчивые эффективные равновесия в моделях коллективного поведения с обменом информацией // Информатика и её применения, 2015. Т. 9. Вып. 2. С. 2–13. doi: 10.14357/19922264150201.
6. *Bai Q., Ren F., Fujita K., Znang M.* Multi-agent and complex systems. — Studies in computational intelligence ser. — Luxembourg: Springer, 2016. 210 p.
7. *Скорняков Л. А.* Элементы общей алгебры. — М.: Наука, 1983. 272 с.
8. *Маклейн С.* Категории для работающего математика / Пер. с англ. — М.: Физматлит, 2004. 352 с. (*Mac Lane S.* Categories for the working mathematician. — Berlin – Heidelberg – New York: Springer, 1978. 317 p.)
9. *Shoham Y., Leyton-Brown R.* Multiagent systems: Algorithmic, game-theoretic, and logical foundations. — Cambridge University Press, 2010. 532 p.
10. *Dixit A. K., Natebuff B. J.* The art of strategy. — New York, London: W. W. Norton & Co., 2008. 446 p.
11. *Dixit A. K., Skeath S., Reiley D. H., Jr.* Games of strategy. — New York, London: W. W. Norton & Co., 2017. 880 p.

Поступила в редакцию 12.03.23

MULTIPLAYERS' GAMES COMPOSITIONAL STRUCTURE IN THE MONOIDAL CATEGORY OF BINARY RELATIONS

N. S. Vasilyev

N. E. Bauman Moscow State Technical University, 5-1 Baumanskaya 2nd Str., Moscow 105005, Russian Federation

Abstract: The system approach is suggested for multiplayers' games solution that meets up-to-date network technologies. It allows to optimize the functionality of multiagent systems. The monoidal category of binery relations is applied to make games rules description and players' behavior study and modification. The game problem is to maximize, if possible, the preference relations of all participants in the game. Their composition in the monoidal binary relations category in correspondence with games rules defines resulting game relation (RGR). Players' rational behavior search is reduced to RGR maximum elements choice. The author formalizes the use of various classes of permissible strategies, information exchange processes, and coalitions formation. The RGR existence is proved and maximum RGR elements structure is studied. Moves priority and absolutely optimal preference relations significance are clarified for the coalitions formation process.

Keywords: player's preference relations: absolutely optimal relation, guarantied relation, moves priority relation; game graph; permissible strategy; rational solution; coalition characteristic relation; resulting game relation; monoidal category; compositionality

DOI: 10.14357/19922264230203 **EDN:** GPMZTS

References

1. Moiseev, N. N. 1975. *Elementy teorii optimal'nykh sistem* [Elements of optimal systems theory]. Moscow: Nauka. 527 p.
2. Germeyer, Yu. B. 1976. *Igry s neprotivopolozhnymi interesami* [Games with non-opposite interests]. Moscow: Nauka. 326 p.
3. Podinovskiy, V. V., and V. D. Nogin. 1982. *Pareto-optimal'nye resheniya mnogokriterial'nykh zadach* [Pareto optimal solutions in multicriteria problems]. Moscow: Nauka. 256 p.
4. Rozen, V. V. 1982. Primenenie teorii binarnykh otnosheniy k obshchey teorii igr [Application of the theory of binary relations to general game theory]. *Matematicheskie metody resheniya ekonomicheskikh zadach* [Mathematical methods for solving economic problems]. Novosibirsk: Nauka. 127–152.
5. Vasilyev, N. S. 2014. Koalitsionno ustoychivye effektivnye ravnovesiya v modelyakh kollektivnogo povedeniya s obmenom informatsiey [On availability of Pareto effective equilibrium situations in collective behavior models with data exchange]. *Informatika i ee Primeneniya — Inform. Appl.* 9(2):2–13. doi: 10.14357/19922264150201.

6. Bai, Q., F. Ren, K. Fujita, and M. Znang. 2016. *Multi-agent and complex systems*. Studies in computational intelligence ser. Luxembourg: Springer. 210 p.
7. Skornyakov, L. A. 1983. *Elementy obshchey algebry* [Elements of general algebra]. Moscow: Nauka. 272 p.
8. Mac Lane, S. 1978. *Categories for the working mathematician*. Berlin – Heidelberg – New York: Springer. 317 p.
9. Shoham, Y., and R. Leyton-Brown. 2010. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press. 532 p.
10. Dixit, A. K., and B. J. Nalebuff. 2008. *The art of strategy*. New York, London: W. W. Norton & Co. 446 p.
11. Dixit, A. K., S. Skeath, and D. H. Reiley, Jr. 2017. *Games of strategy*. New York, London: W. W. Norton & Co. 880 p.

Received March 12, 2023

Contributor

Vasilyev Nikolai S. (b. 1952) — Doctor of Science in physics and mathematics, professor, N. E. Bauman Moscow State Technical University, 5-1 Baumanskaya 2nd Str., Moscow 105005, Russian Federation; nik8519@yandex.ru

РЫНОК С МАРКОВСКОЙ СКАЧКООБРАЗНОЙ ВОЛАТИЛЬНОСТЬЮ I: МОНИТОРИНГ ЦЕНЫ РИСКА КАК ЗАДАЧА ОПТИМАЛЬНОЙ ФИЛЬТРАЦИИ*

А. В. Борисов[1]

Аннотация: Первая часть цикла посвящена исследованию задачи определения рыночной цены риска в финансовой системе, содержащей безрисковый актив — банковский вклад, базовые рисковые инструменты и их деривативы. Модель эволюции базовых активов, описываемая стохастической дифференциальной системой (СДС), содержит стохастическую волатильность, порожденную скрытым марковским скачкообразным процессом (МСП). Исследуемый рынок предполагается неполным с отсутствием возможности арбитража. Рыночная цена риска, соответствующая преобладающей мартингальной мере, выражается через скрытый марковский процесс и не может быть восстановлена точно. Тем не менее она может быть оценена оптимальным образом по наблюдениям цен активов. Исходя из наличия преобладающей мартингальной меры в статье получена СДС в частных производных, описывающая эволюцию во времени цены деривативов и являющаяся аналогом классического уравнения Блэка–Шоулза. Задача оценки рыночной цены риска сформулирована в терминах оптимальной фильтрации состояния СДС наблюдения. Также обсуждаются вопросы численной реализации решения данной задачи оценивания.

Ключевые слова: марковский скачкообразный процесс; оптимальная фильтрация; стохастическая волатильность; рыночная цена риска; преобладающая мартингальная мера

DOI: 10.14357/19922264230204 **EDN:** GAXCHQ

1 Введение

Одна из практических проблем, привлекающих решение задачи фильтрации состояний стохастических систем, — мониторинг внутренней волатильности базовых финансовых инструментов по разнородным наблюдениям их цен [1–3]. Помимо цен основных инструментов в качестве доступной статистической информации присутствует богатейший объем данных о торгах производными инструментами (деривативами). В работе [4] предложен один из способов использования такого массива данных. Для этого была выбрана концепция неполных безарбитражных рынков [5, 6]. Зависимость цены деривативов от текущих цен базовых активов и времени до погашения определялась относительно преобладающей мартингальной меры с помощью формулы Фейнмана–Каца. Эта мера определялась в терминах рыночной цены риска (Market Price of Risk, MPR) — некоторого «внешнего» ненаблюдаемого процесса, связывающего мгновенную процентную ставку и внутреннюю волатильность. Таким образом, решение задачи оценивания MPR в реальном масштабе времени позволяет получить и оценки внутренней волатильности.

Автор [4] распространяет концепцию MPR, ранее использовавшуюся преимущественно для описания колебаний процентных ставок по бескупонным облигациям, ставок по кредитам и пр. (см., например, [7] и ссылки внутри), на класс базовых рисковых активов. В качестве возможных моделей MPR были выбраны диффузионные процессы — решения СДС с винеровскими процессами в правой части. Основной результат [4] заключается в формулировке задачи оценивания MPR по наблюдениям базовых активов и деривативов в виде задачи фильтрации состояний СДС в самом общем виде. Эта проблема весьма сложна для ее как аналитического, так и численного решения.

Представляемая работа посвящена исследованию упомянутой задачи в случае, когда MPR определяется поведением внешнего ненаблюдаемого МСП с конечным числом состояний. Цель статьи заключается в исследовании задачи мониторинга MPR по имеющимся наблюдениям цен базовых активов и деривативов в рамках математического аппарата стохастического анализа. Статья имеет следующую структуру. В разд. 2 представлено описание исследуемого класса финансовых систем, включающего банковский депозит, базовые рисковые активы и их деривативы. Ставка по депози-

*Работа выполнена с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук; aborisov@frccsc.ru

ту является детерминированной, а цены базовых активов и деривативов описываются наблюдаемыми случайными процессами. Мгновенные ставки и внутренние волатильности базовых активов зависят от скрытого МСП. Уравнения динамики деривативов отсутствуют.

Относительно исследуемого класса систем известно, что он соответствует неполному безарбитражному рынку. Данные условия гарантируют существование некоторой преобладающей мартингальной меры, относительно которой все базовые активы имеют одинаковую процентную ставку, совпадающую с депозитной. Существование такой меры позволило характеризовать временну́ю эволюцию деривативов в форме решения системы линейных дифференциальных уравнений в частных производных — некоторого обобщения уравнения Блэка–Шоулза [8]. Таким образом, корректно построена стохастическая дифференциальная модель наблюдения, в которой в качестве ненаблюдаемого состояния выступает скрытый фактор-процесс и модели наблюдаемых цен базовых активов и их деривативов полностью определены. Оценивание MPR в реальном масштабе времени эквивалентно фильтрации состояний скрытого МСП. Эти выкладки содержатся в разд. 3.

Полученная система наблюдения обладает некоторыми свойствами, препятствующими непосредственному применению известных методов оптимальной нелинейной фильтрации для построения оценок МСП: поток σ-подалгебр, порожденный наблюдениями, не является непрерывным справа, шумы в наблюдениях вырождены и пр. Раздел 4 статьи посвящен перспективам теоретических исследований и численной реализации полученной задачи фильтрации.

2 Описание модели финансовой системы

На отрезке времени $[0,T]$ рассматривается финансовая система, состоящая из банковского вклада, N базовых финансовых инструментов S и M деривативов F.

Ставка r_t банковского вклада известна и неслучайна. Эволюция цен базовых активов определяется на базисе с фильтрацией $(\Omega,\mathcal{F},\mathcal{P},\{\mathcal{F}_t\}_{t\in[0,T]})$ как наблюдаемый случайный процесс $S_t \triangleq \triangleq \operatorname{col}\left(S_t^1,\ldots,S_t^N\right)$ — единственное сильное решение СДС:

$$dS_t = \operatorname{diag} S_t a\,(t,Z_t)\,dt + \operatorname{diag} S_t\sigma(t,Z_t)\,dw_t,\qquad t\in(0,T],\ \ S_0\sim\pi_0(s),\quad (1)$$

где

- S_0 — $\mathcal{F}_0$-измеримое начальное условие;
- $w_t = \operatorname{col}\left(w_t^1,\ldots,w_t^K\right)\in\mathbb{R}^K$ — $\mathcal{F}_t$-согласованный стандартный винеровский процесс ($K\geqslant N$);
- $a(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$ — $(N\times 1)$- и $(N\times K)$-мерные функции мгновенной процентной ставки и внутренней волатильности базовых инструментов S; $\sigma(\cdot,\cdot)$ имеет полный строчный ранг $\mathcal{P}$-п. н. почти везде по мере Лебега на $[0,T]$;
- Z_t — $\mathcal{F}_t$-согласованный ненаблюдаемый (скрытый) процесс, описывающий действие на финансовую систему неконтролируемых внешних факторов.

В (1) $Z_t = \operatorname{col}\left(Z_t^1,\ldots,Z_t^\ell\right)\in\mathbb{S}^L = \{e_1,\ldots,e_\ell\}$ представляет собой неоднородный МСП с конечным множеством состояний — набором координатных ортов евклидова пространства $\mathbb{R}^L$, известной матричнозначной функцией интенсивностей переходов $\Lambda(\cdot)$ и начальным распределением π_0^Z. Дополнительно о Z предполагается, что все компоненты его распределения строго положительны на $[0,T]$. Известно [9], что МСП Z_t — единственное сильное решение СДС

$$dZ_t = \Lambda^\top(t)Z_t\,dt + dM_t,\ \ t\in(0,T],\ \ Z_0\sim\pi_0^Z,\quad (2)$$

где $M_t = \operatorname{col}\left(M_t^1,\ldots,M_t^L\right)\in\mathbb{R}^L$ — некоторый $\mathcal{F}_t$-согласованный мартингал.

Для исследуемой финансовой системы предполагается отсутствие арбитража [10] и существование на измеримом пространстве $(\Omega,\mathcal{F})$ преобладающей мартингальной меры $\mathcal{Q}$, $\mathcal{Q}\sim\mathcal{P}$, для которой выполнены следующие условия.

1. Процесс M_t является мартингалом относительно меры $\mathcal{Q}$.
2. Относительно $\mathcal{Q}$ процесс S_t — единственное сильное решение СДС

$$dS_t = r_tS_t\,dt + \operatorname{diag} S_t\sigma(t,Z_t)\,dw_t^{\mathcal{Q}},\qquad t\in(0,T],\ \ S_0\sim\pi_0(s),\quad (3)$$

в которой $w_t^{\mathcal{Q}}\in\mathbb{R}^K$ — $\mathcal{F}_t$-согласованный стандартный винеровский процесс.

Согласно обобщенной теореме Гирсанова [11] винеровский процесс $w_t^{\mathcal{Q}}$ связан с исходным w_t равенством

$$dw_t^{\mathcal{Q}} = \theta_t\,dt + dw_t,\quad (4)$$

где $\theta_t\in\mathbb{R}^K$ — недоступный наблюдению $\mathcal{F}_t$-согласованный процесс MPR.

Момент погашения всех деривативов совпадает с T и соответствует платежному требованию

$$H(S_T) = \operatorname{col}(H^1(S_T), \ldots, H^M(S_T)),$$

где $H(s) = \operatorname{col}(H^1(s), \ldots, H^M(s))$ — известная детерминированная вектор-функция. Эволюция цен производных финансовых инструментов является наблюдаемым $\mathcal{F}_t$-согласованным специальным семимартингалом $F_t = \operatorname{col}(F_t^1, \ldots, F_t^M)$, таким что $F_T = H(S_T))$ $\mathcal{P}$-п. н.

Данный рынок является неполным и означает неединственность мартингальной вероятностной меры. Точное знание MPR θ_t эквивалентно знанию преобладающей вероятностной меры $\mathcal{Q}$ и позволяет решить ряд актуальных финансовых задач, в том числе задачу построения хеджирующего портфеля [5, 8]. В рассматриваемой задаче MPR недоступна прямому безошибочному наблюдению. Все данные, доступные на момент времени t, описываются σ-алгеброй наблюдений

$$\mathcal{O}_t \triangleq \sigma\{S_u, F_u : \ 0 \leqslant u \leqslant t\},$$

поэтому при наличии этой информации очевидным представляется поиск условного среднего MPR относительно имеющихся наблюдений:

$$\widehat{\theta}_t \triangleq \mathsf{E}_{\mathcal{P}}\left\{\theta_t | \mathcal{O}_t\right\}.$$

3 Сведение задачи определения цены риска к задаче оптимальной фильтрации

Из уравнений (1), (3) и (4) можно определить связь между функциями r, a и σ, с одной стороны, и MPR θ — с другой. Подстановкой (4) в (3) можно получить СДС (1) в виде представления специального семимартингала. В силу единственности этого представления равенство

$$\operatorname{diag} S_t a(t, Z_t) = r_t S_t + \operatorname{diag} S_t a(t, Z_t)\sigma(t, Z_t)\theta_t$$

верно $\mathcal{P}$-п. н. почти везде по мере Лебега на $[0, T]$. Так как все компоненты вектора S_t $\mathcal{P}$-п. н. строго положительны, последнее равенство преобразуется к виду

$$\sigma(t, Z_t)\theta_t = a(t, Z_t) - r_t\mathbf{1}, \tag{5}$$

где $\mathbf{1}$ — вектор-столбец подходящей размерности, составленный из единиц.

Далее, в силу того что $Z_t \in \mathbb{S}^L$, функции a и σ представимы в виде

$$a(t, Z_t) = \sum_{\ell=1}^{L} Z_t^\ell a^\ell(t); \ \ \sigma(t, Z_t) = \sum_{\ell=1}^{L} Z_t^\ell \sigma^\ell(t),$$

где $\{a^\ell(t)\}_{\ell=\overline{1,L}}$ и $\{\sigma^\ell(t)\}_{\ell=\overline{1,L}}$ — наборы известных неслучайных функций — «алфавиты» мгновенных ставок и внутренних волатильностей:

$$a^\ell(t) \triangleq a(t, e_\ell); \ \ \sigma^\ell(t) \triangleq \sigma(t, e_\ell).$$

В этом случае система (5) принимает вид:

$$\sum_{\ell=1}^{L} Z_t^\ell \sigma^\ell(t)\theta_t = \sum_{\ell=1}^{L} Z_t^\ell a^\ell(t) - r_t\mathbf{1},$$

откуда в силу условий $\pi^Z(t) > 0$ и $\operatorname{rk}(\sigma(\cdot)) \equiv K$ следует, что MPR θ имеет вид:

$$\theta_t = \sum_{\ell=1}^{L} Z_t^\ell \left(\sigma^\ell(t)\right)^+ \left(a^\ell(t) - r_t\mathbf{1}\right) + \\ + \sum_{\ell=1}^{L} Z_t^\ell \left(I - \left(\sigma^\ell(t)\right)^+ \sigma^\ell(t)\right)\xi_t, \tag{6}$$

где I — единичная матрица подходящей размерности; A^+ — матрица, псевдообратная матрице A; $\xi_t \in \mathbb{R}^L$ — произвольный $\mathcal{F}_t$-согласованный случайный процесс. Второе слагаемое в (6) аннулируется умножением справа на $\sigma(t, Z_t)$ и не влияет на превалирующую меру $\mathcal{Q}$, поэтому ниже в работе считается, что оно в (6) отсутствует. Таким образом, MPR θ_t является линейным преобразованием МСП Z_t.

Вид функций $F^m(t, S_t, Z_t)$ ($m = \overline{1, M}$) цен деривативов определяется, исходя из отсутствия в исследуемой финансовой системе арбитража и наличия мартингальных вероятностных мер, как дисконтированное условное среднее платежного требования H на момент погашения T, вычисленное относительно преобладающей меры $\mathcal{Q}$:

$$F^m(t, S_t, Z_t) = e^{-\int_t^T r_s\, ds}\mathsf{E}_{\mathcal{Q}}\left\{H^m(S_T)|\mathcal{F}_t\right\}.$$

Заметим, что $\Pi_t^m \triangleq \mathsf{E}_{\mathcal{Q}}\left\{H^m(S_T)|\mathcal{F}_t\right\}$ — мартингал относительно $\mathcal{Q}$ — представим в виде

$$\Pi^m = e^{\int_t^T r_s ds}\sum_{\ell=1}^{L} Z_t^\ell F^{m\ell}(t, S_t),$$

где $F^{m\ell}(t, S_t) \triangleq F^m(t, S_t, e_\ell)$.

Предположим, что $F^{m\ell}$ обладает достаточной степенью гладкости по своим переменным, тогда Π_t^m допускает следующий стохастический дифференциал (ниже в выкладках зависимость функций от своих аргументов для краткости опущена):

$$d\Pi^m = e^{\int_t^T r\, ds}\left(-r\sum_{\ell=1}^{L} Z^\ell F^{m\ell}\, dt + \right. \\ \left. + \sum_{\ell=1}^{L} dZ^\ell F^{m\ell} + \sum_{\ell=1}^{L} Z^\ell\, dF^{m\ell}\right).$$

Рассмотрим отдельно второе и третье слагаемое в скобках последнего выражения:

$$\sum_{\ell=1}^{L} dZ^\ell F^{m\ell} = \sum_{\ell=1}^{L} e_\ell^\top \left(\Lambda^\top Z\,dt + dM\right) F^{m\ell} =$$
$$= \sum_{i,j,\ell=1}^{L} e_\ell^i \Lambda^{ji} Z^j F^{m\ell}\,dt + \sum_{\ell=1}^{L} F^{m\ell}\,dM^\ell =$$
$$= \sum_{\ell=1}^{L} Z^\ell \sum_{j=1}^{L} \Lambda^{\ell j} F^{mj}\,dt + \sum_{\ell=1}^{L} F^{m\ell}\,dM^\ell;$$

$$\sum_{\ell=1}^{L} Z^\ell\,dF^{m\ell} = \sum_{\ell=1}^{L} Z^\ell \left(F_t^{m\ell}\,dt + \sum_{n=1}^{N} F_{s^n}^{m\ell}\,dS^n + \right.$$
$$\left. + \frac{1}{2}\sum_{i,j=1}^{N} F_{s^i,s^j}^{m\ell}\,d\langle S^i, S^j\rangle \right) = \sum_{\ell=1}^{L} Z^\ell \left(F_t^{m\ell}\,dt + \right.$$
$$+ \sum_{n=1}^{N} F_{s^n}^{m\ell} \left(S^n a_n^\ell\,dt + S^n \sum_{k=1}^{K} \sigma_{nk}^\ell\,dw^k \right) +$$
$$\left. + \frac{1}{2}\sum_{i,j=1}^{N} F_{s^i,s^j}^{m\ell}\,d\langle S^i, S^j\rangle \right),$$

где

$$F_t^{m\ell} \triangleq \left.\frac{\partial F^{m\ell}(t,s)}{\partial t}\right|_{(t,S_t)};$$
$$F_{s^n}^{m\ell} \triangleq \left.\frac{\partial F^{m\ell}(t,s)}{\partial s^n}\right|_{(t,S_t)};$$
$$F_{s^i,s^j}^{m\ell} \triangleq \left.\frac{\partial^2 F^{m\ell}(t,s)}{\partial s^i \partial s^j}\right|_{(t,S_t)}.$$

Используем в третьем слагаемом формулы (4), (6) и обозначение $\nabla_s F^{m\ell} \triangleq \operatorname{col}\left(F_{s^1}^{m\ell}, \ldots, F_{s^N}^{m\ell}\right)$:

$$\sum_{\ell=1}^{L} Z^\ell \sum_{n=1}^{N}\sum_{k=1}^{K} F_{s^n}^{m\ell} S^n \sigma_{nk}^\ell\,dw^k =$$
$$= \sum_{\ell=1}^{L} Z^\ell \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)\sigma^\ell\,dw =$$
$$= \sum_{\ell=1}^{L} Z^\ell \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)\sigma^\ell (dw^{\mathcal{Q}} - \theta\,dt) =$$
$$= \sum_{\ell=1}^{L} Z^\ell \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)(r\mathbf{1} - a^\ell)\,dt +$$
$$+ \sum_{\ell=1}^{L} Z^\ell \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)\sigma^\ell\,dw^{\mathcal{Q}}.$$

Если

$$B^\ell(t) \triangleq \sigma^\ell(t)\left(\sigma^\ell(t)\right)^\top = \|B_{ij}^\ell(t)\|_{i,j=\overline{1,K}},$$

то Π^m допускает дифференциал:

$$d\Pi^m = e^{\int_t^T r\,ds}\left(\sum_{\ell=1}^{L} F^{m\ell}\,dM^\ell + \right.$$
$$\left. + \sum_{\ell=1}^{L} Z^\ell \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)\sigma^\ell\,dw^{\mathcal{Q}} \right) +$$
$$+ e^{\int_t^T r\,ds} \sum_{\ell=1}^{L} Z^\ell \left[-rF^{m\ell} + \sum_{j=1}^{L} \Lambda^{\ell j} F^{mj} + F_t^{m\ell} + \right.$$
$$+ \left(\nabla_s F^{m\ell}\right)^\top \operatorname{diag}(S)(r\mathbf{1} - a^\ell) +$$
$$\left. + \frac{1}{2}\sum_{i,j=1}^{N} F_{s^i,s^j}^{m\ell} S^i S^j B_{ij}^\ell \right] dt.$$

Первое слагаемое в нем представляет собой дифференциал мартингала относительно меры $\mathcal{Q}$, второе — дифференциал процесса с ограниченной вариацией. Мартингальное свойство процессов $\{\Pi^m\}$ относительно $\mathcal{Q}$, а также поэлементное выполнение неравенств $\pi^Z(t) > 0$ на отрезке $[0,T]$ позволяют определить функции цены $\{F^{m\ell}(t,s)\}$ как решение системы уравнений Колмогорова [12]:

$$\left.\begin{aligned} &F_t^{m\ell} = rF^{m\ell} - \sum_{j=1}^{L} \Lambda^{\ell j} F^{mj} - \\ &- \sum_{n=1}^{N} F_{s^n}^{m\ell} s^n (r - a_n^\ell) - \frac{1}{2}\sum_{i,j=1}^{N} s^i s^j F_{s^i,s^j}^{m\ell} B_{ij}^\ell, \\ &\qquad \ell = \overline{1,M},\ \ m = \overline{1,M},\ \ t \in [0,T]; \\ &F^{m\ell}(T,s) = H^m(s). \end{aligned}\right\} \quad (7)$$

Таким образом, деривативы F допускают стохастический дифференциал

$$dF^m = \sum_{\ell=1}^{L} Z^\ell \left[rF^{m\ell} + \sum_{n=1}^{N} F_{s^n}^{m\ell} S^n (a_n^\ell - r) \right] dt +$$
$$+ \sum_{\ell=1}^{L} F^{m\ell}\,dM^\ell + \sum_{\ell=1}^{L} Z^\ell \sum_{n=1}^{N}\sum_{k=1}^{K} F_{s^n}^{m\ell} S^n \sigma_{nk}^\ell\,dw^k,$$
$$m = \overline{1,M},\ \ \ell = \overline{1,L},\ \ F^{m\ell} = F^{m\ell}(t,S_t). \quad (8)$$

Итак, определение MPR θ_t в описанной финансовой системе сводится к решению задачи оптимальной фильтрации состояний МСП Z (2) по совокупности наблюдений S (1) и F (8).

4 Заключение

Формулы (2), (1) и (8) задают СДС наблюдения с МСП в качестве состояния и мультипликативными шумами в наблюдениях. Несмотря на то что теоретическое решение похожей задачи фильтра-

ции представлено в [13], а комплекс алгоритмов численного решения — в [14, 15], при решении поставленной в данной статье задачи возникает ряд теоретических и реализационных проблем. Прежде всего, обратимся к теоретическим вопросам.

Первое: поток σ-алгебр $\{\mathcal{O}_t\}_{t\in[0,T]}$, порожденный наблюдениями (1) и (8), не является непрерывным справа. Для корректного использования аппарата стохастического анализа можно «сгладить справа» исходные σ-алгебры наблюдений, т. е. рассматривать $\mathcal{O}_{t+}$ вместо $\mathcal{O}_t$, и трансформировать задачу фильтрации в задачу прогнозирования на малый фиксированный шаг $\delta > 0$, т. е. в задачу поиска $\mathsf{E}_{\mathcal{P}}\left\{Z_t|\mathcal{O}_{(t-\delta)+}\right\}$.

Второе: система наблюдения (2), (1) и (8) формально не принадлежит классу систем наблюдения, для которых задача фильтрации состояния МСП решена. Уравнение (8) цен деривативов нелинейно: коэффициенты сноса и диффузии зависят от случайного процесса S_t. Тем не менее результат [13] может быть распространен и на данную систему: процесс S_t имеет $\mathcal{P}$-п. н. траектории и доступен наблюдению, поэтому вместо детерминированных коэффициентов можно использовать их наблюдаемые случайные аналоги. Уместна аналогия с условно-гауссовскими процессами [16], оптимальная фильтрация состояний которых может быть выполнена с помощью фильтра Калмана—Бьюси со случайными, но наблюдаемыми параметрами.

Третье: из уравнений (1) и (8) следует, что шумы в этих наблюдениях вырожденные, и путем линейных преобразований из них можно выделить наблюдаемые компоненты, не содержащие шумов. В [13] предложен прием, преобразующий подобные наблюдения в совокупность считающих процессов и косвенных наблюдений МСП, полученных в известные детерминированные моменты времени.

Следующие вопросы относятся к области численной реализации решения поставленной задачи фильтрации. Продолжим «сквозную» нумерацию проблем.

Четвертое: численные алгоритмы фильтрации [15] разработаны для систем наблюдения, в которых параметры динамики и наблюдений не зависят от времени. В рассмотренной финансовой задаче это принципиально не так: с приближением к моменту погашения деривативов их цена изменяется даже при условии постоянства цены базовых активов. Предложенные в [15] численные алгоритмы применимы и в этом случае, однако из-за нестационарности системы наблюдения они будут порождать дополнительные ошибки, наличие которых следует принимать во внимание.

Пятое: определение функции F цены деривативов как решение системы (7) представляется самостоятельной нетривиальной задачей. Аналитическое решение ее отсутствует, а применяемые численные методы должны обеспечивать достаточную точность аппроксимации не только самих функций $F^{m\ell}$, но и их частных производных $F^{m\ell}_{s^n}$, так как они входят в матрицу наблюдений (8).

Шестое: наблюдения F (8) — это сумма непрерывных и скачкообразных процессов, которую теоретически легко разделить. Однако непрерывные наблюдения являются идеализацией. В действительности доступны наблюдения, дискретизованные по времени, либо результаты высокочастотных измерений в случайные моменты времени [17]. В таких наблюдениях выделить скачки не представляется возможным, нужны новые варианты алгоритмов численной фильтрации состояний МСП.

Перечисленные проблемы не представляются непреодолимыми, основа их решения заложена в работе [13]. Последующие части цикла будут посвящены вопросам численной реализации решения задачи мониторинга MPR, начиная с моделирования цен соответствующих активов, эволюционирующих в соответствии со скачкообразным изменением MPR, и заканчивая алгоритмами фильтрации скрытой MPR по имеющимся разнородным наблюдениям цен базовых бумаг и деривативов.

Литература

1. *Elliott R., Malcolm W., Tsoi A.* HMM volatility estimation // 41st IEEE Conference on Decision and Control Proceedings, 2002. — Piscataway, NJ, USA: IEEE, 2003. Vol. 1. P. 398–404. doi: 10.1109/CDC.2002.1184527.
2. *Голдентаер Л., Клебанер Ф., Липцер Р.* Слежение за функцией волатильности // Проблемы передачи информации, 2005. Т. 41. Вып. 3. С. 32–50.
3. *Cvitanić J., Liptser R., Rozovskii B.* A filtering approach to tracking volatility from prices observed at random times // Ann. Appl. Probab., 2006. Vol. 16. No. 3. P. 1633–1652. doi: 10.1214/105051606000000222.
4. *Runggaldier W.* Estimation via stochastic filtering in financial market models // Contemp. Math., 2004. Vol. 351. P. 309–318. doi: 10.1090/conm/351/06412.
5. *Björk T.* Arbitrage theory in continuous time. — New York, NY, USA: Oxford University Press, 1998. 324 p.
6. *Duffie D.* Dynamic asset pricing theory. — Princeton, NJ, USA: Princeton University Press, 2001. 472 p.
7. *Dai Q., Singleton K.* Specification analysis of affine term structure models // J. Financ., 2000. Vol. 55. No. 5. P. 1943–1978. doi: 10.1111/0022-1082.00278.

8. *Shiryayev A.* Essentials of stochastic finance: Facts, models, theory. — New Jersey, NJ, USA: World Scientific, 1999. 834 p.
9. *Elliott R., Moore J., Aggoun L.* Hidden Markov models: Estimation and control. — New York, NY, USA: Springer, 2010. 382 p.
10. *Criens D.* No arbitrage in continuous financial markets // Math. Financ. Econ., 2020. Vol. 14. P. 461–506.
11. *Cohen S., Elliott R.* Stochastic calculus and applications. — New York, NY, USA: Birkhäuser, 2015. 666 p.
12. *Gihman I., Skorohod A.* The theory of stochastic processes III. — New York, NY, USA: Springer, 1979. 388 p.
13. *Borisov A., Sokolov I.* Optimal filtering of Markov jump processes given observations with state-dependent noises: Exact solution and stable numerical schemes // Mathematics, 2020. Vol. 8. No. 4. Art. No. 506.
14. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям II: случай аддитивных шумов // Информатика и её применения, 2020. Т. 14. Вып. 1. С. 17–23.
15. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям III: случай мультипликативных шумов // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 10–18.
16. *Liptser R., Shiryaev A.* Statistics of random processes II: Applications. — Berlin/Heidelberg: Springer, 2001. 402 p.
17. *Королев В., Черток А., Корчагин А., Горшенин А.* Вероятностно-статистическое моделирование информационных потоков в сложных финансовых системах на основе высокочастотных данных // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 12–21.

Поступила в редакцию 05.10.22

MARKET WITH MARKOV JUMP VOLATILITY I: PRICE OF RISK MONITORING AS AN OPTIMAL FILTERING PROBLEM

A. V. Borisov

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The first part of series is devoted to investigating the market price of risk in a financial system. It contains riskless bank deposits, risky base assets, and their derivatives. The model of the underlying price evolution represents a stochastic differential system with stochastic volatility which is a hidden Markov jump process. The investigated market is incomplete and has no arbitrage possibilities. The market price of risk, which corresponds to a prevailing martingale measure, can be characterized via the hidden Markov jump process but can not be restored precisely. However, it can be estimated optimally using the observations of both the derivative and underlying prices. Using the concept of the prevailing martingale measure existence, one can derive a system of the partial differential equations which describes an evolution of the derivative prices and represents some analog of the classic Black–Sholes equation. Then, one can convert the calculation problem for the market price of risk to the optimal state filtering in a differential stochastic observation system. The paper also discusses various aspects of the numerical realization for the stated estimation problem.

Keywords: Markov jump process; optimal filtering; diffusion and counting observations; multiplicative observation noise; numerical approximation accuracy

DOI: 10.14357/19922264230204 **EDN:** GAXCHQ

Acknowledgments

The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Elliott, R., W. Malcolm, and A. Tsoi. 2002. HMM volatility estimation. *41st IEEE Conference on Decision and Control Proceedings*. Piscataway, NJ: IEEE. 1:398–404. doi: 10.1109/CDC.2002.1184527.
2. Goldentayer, L., F. K. Klebaner, and R. Sh. Liptser. 2005. Tracking volatility. *Probl. Inf. Transm.* 41(3):212–229. doi: 10.1007/s11122-005-0026-2.
3. Cvitanić, J., R. Liptser, and B. Rozovskii. 2006. A filtering approach to tracking volatility from prices observed

at random times. *Ann. Appl. Probab.* 16:1633–1652. doi: 10.1214/105051606000000222.

4. Runggaldier, W. 2004. Estimation via stochastic filtering in financial market models. *Contemp. Math.* 351:309–318. doi: 10.1090/conm/351/06412.
5. Björk, T. 1998. *Arbitrage theory in continuous time*. New York, NY: Oxford University Press. 324 p.
6. Duffie, D. 2001. *Dynamic asset pricing theory*. Princeton, NJ: Princeton University Press. 472 p.
7. Dai, Q., and K. Singleton. 2000. Specification analysis of affine term structure models. *J. Financ.* 55(5):1943–1978. doi: 10.1111/0022-1082.00278.
8. Shiryayev, A. 1999. *Essentials of stochastic finance: Facts, models, theory*. New Jersey, NJ: World Scientific. 834 p.
9. Elliott, R., J. Moore, and L. Aggoun. 2010. *Hidden Markov models: Estimation and control*. New York, NY: Springer. 382 p.
10. Criens, D. 2020. No arbitrage in continuous financial markets. *Math. Financ. Econ.* 14:461–506. doi: 10.1007/s11579-020-00262-1.
11. Cohen, S., and R. Elliott. 2015. *Stochastic calculus and applications*. New York, NY: Birkhäuser. 666 p.
12. Gihman, I., and A. Skorohod. 1979. *The theory of stochastic processes III*. New York, NY: Springer. 388 p.
13. Borisov, A., and I. Sokolov. 2020. Optimal filtering of Markov jump processes given observations with state-dependent noises: Exact solution and stable numerical schemes. *Mathematics* 8(4):506. doi: 10.3390/math8040506.
14. Borisov, A. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam II: sluchay additivnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations II: Additive noise case]. *Informatika i ee primeneniya — Informatics and Applications* 14(1): 117–23. doi: 10.14357/19922264200103.
15. Borisov, A. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam III: sluchay mul'tiplikativnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations III: Multiplicative noises case] *Informatika i ee primeneniya — Informatics and Applications* 14(2)10–18. doi: 10.14357/19922264200202.
16. Liptser, R., and A. Shiryaev. 2001. *Statistics of random processes II: Applications*. Berlin/Heidelberg: Springer. 402 p.
17. Korolev, V., A. Chertok, A. Korchagin, and A. Gorshenin. 2013. Veroyatnostno-statisticheskoe modelirovanie informatsionnykh potokov v slozhnykh finansovykh sistemakh na osnove vysokochastotnykh dannykh [Probability and statistical modeling of information flows in complex financial systems based on high-frequency data]. *Informatika i ee primeneniya — Informatics and Applications* 7(1)12–21.

Received October 5, 2022

Contributor

Borisov Andrey V. (b. 1965) — Doctor of Science in physics and mathematics, principal scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; aborisov@frccsc.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2023. Т. 17. Вып. 2. С. 34–40

СРЕДНЕКВАДРАТИЧНЫЙ РИСК FDR-ПРОЦЕДУРЫ В УСЛОВИЯХ СЛАБОЙ ЗАВИСИМОСТИ

М. О. Воронцов[1], О. В. Шестаков[2]

Аннотация: Во многих прикладных областях возникает задача обработки больших массивов данных. При этом часто перед обработкой массив данных подвергается некоторому преобразованию, приводящему к «разреженному», или «экономному», представлению, при котором абсолютное значение большинства элементов массива равно нулю (или достаточно мало). Кроме того, в результате помех при получении и передаче данных в них попадает шум, который при дальнейшей обработке желательно некоторым образом удалить. Возникающая при этом задача математически эквивалентна некоторым задачам множественной проверки гипотез. Ранее для решения указанной задачи в условиях нормальности, независимости и разреженности данных была предложена процедура, основанная на методе контроля средней доли ложных отклонений (False Discovery Rate, FDR) гипотез. В настоящей работе исследуется асимптотика риска указанной процедуры в случае наличия слабой зависимости в данных.

Ключевые слова: пороговая обработка; множественная проверка гипотез; среднеквадратичный риск

DOI: 10.14357/19922264230205 **EDN:** AVJZDX

1 Введение

В современных приложениях статистики зачастую требуется обрабатывать большие массивы зашумленных данных – источниками шума могут выступать помехи и несовершенство оборудования. Примерами служат исследования в области генетики с возникающими в них задачами множественной проверки гипотез [1], задачи обработки изображений с высоким разрешением [2] и другие прикладные проблемы. В связи с этим рассмотрим задачу нахождения оценки неизвестного вектора μ как функции x в модели данных

$$x_i = \mu_i + z_i, \;\; i = 1, \ldots, n,$$

где $\mu_i \in \mathbb{R}$, $z_i \sim N(0, \sigma^2)$ для всех i.

Приведенная задача может рассматриваться как частный случай задачи множественной проверки гипотез, а именно: пусть построено n статистик x_i для проверки нулевых гипотез $H_{0,i}$ против альтернатив $H_{1,i}$, причем при верной гипотезе $H_{0,i}$ (соответственно $H_{1,i}$) распределение x_i известно и равно $N(0, \sigma^2)$ (соответственно $N(\mu_i, \sigma^2)$, $\mu_i \neq 0$ и неизвестно). Принятие гипотезы $H_{0,i}$ в такой постановке равносильно заключению $\mu_i = 0$.

В работе [3] для решения рассматриваемой задачи в условиях независимости компонент вектора x и разреженности вектора μ была предложена процедура построения оценки $\hat{\mu}_F$ вектора μ, основанная на методе контроля средней доли ложных отклонений (FDR) гипотез при помощи алгоритма Бенжамини–Хочберга, и было проведено исследование асимптотики риска построенной оценки.

В то же время в определенных приложениях, например при анализе полученных в результате использования ДНК-микрочипов данных [4], исследовании геофизических процессов и анализе помех в телекоммуникационных каналах, условие независимости компонент вектора x может не выполняться. В данной работе исследуется асимптотика риска предложенной в [3] оценки $\hat{\mu}_F$ в случае, когда компоненты вектора x слабо зависимы, а μ принадлежит классу разреженности

$$l_0[\eta] = \{\mu \in \mathbb{R}^n \,:\, ||\mu||_0 \leq \eta n\}, \;\; \eta \in (0; 1).$$

2 Обработка вектора данных с помощью FDR-процедуры

Предложенная в [3] процедура заключается в жесткой пороговой обработке компонент вектора x порогом $\hat{t}_F = \hat{t}_F(x)$, и ее результат — оценка $\hat{\mu}_F$ вектора μ с компонентами

[1]Факультет вычислительной математики и кибернетики, Московский государственный университет имени М. В. Ломоносова; Московский центр фундаментальной и прикладной математики, m.vtsov@mail.ru

[2]Факультет вычислительной математики и кибернетики, Московский государственный университет имени М. В. Ломоносова; Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Московский центр фундаментальной и прикладной математики, oshestakov@cs.msu.ru

$$(\hat{\mu}_F)_i = p(x_i, \hat{t}_F) \equiv \begin{cases} x_i, & |x_i| > \hat{t}_F; \\ 0, & |x_i| \le \hat{t}_F, \end{cases}$$

где

$$\hat{t}_F = \sigma z\left(\frac{q\hat{k}_F}{2n}\right).$$

Здесь $z(\alpha)$ — квантиль уровня $(1-\alpha)$ стандартного нормального распределения; $q \in (0;1)$ — управляющий параметр FDR-метода;

$$\hat{k}_F = \max\left\{k \,:\, |x|_{(k)} \ge t_k\right\},$$

где $|x|_{(k)}$ — k-й элемент вектора, получаемого в результате упорядочения вектора $|x|$ по невозрастанию:

$$|x|_{(1)} \ge |x|_{(2)} \ge \cdots \ge |x|_{(n)},$$

$$t_k = \sigma z\left(\frac{qk}{2n}\right).$$

Далее полагается, что $q \equiv q_n$ зависит от n.

В [3] для среднеквадратичного риска

$$\rho(\hat{\mu}_F, \mu) = \mathsf{E}||\hat{\mu}_F - \mu||^2 = \mathsf{E}\sum_{i=1}^{n}\left(p(x_i, \hat{t}_F) - \mu_i\right)^2$$

оценки $\hat{\mu}_F$ получен следующий результат.

Теорема 1. *Пусть x_i, $i = 1, \ldots, n$, независимы, $\liminf q_n \ln n > 0$, $\limsup q_n < 1$, а также η_n лежит в интервале $[n^{-1}\ln^5 n; n^{-\delta}]$, $\delta > 0$. Тогда при $n \to \infty$*

$$\sup_{\mu \in l_0[\eta_n]} \rho(\hat{\mu}_F, \mu) \le \le R_n(l_0[\eta_n])\left(1 + 2q_n(1-q_n)^{-1} + o(1)\right),$$

где

$$R_n(l_0[\eta_n]) = \inf_{\hat{\mu}=\hat{\mu}(x)} \sup_{\mu \in l_0[\eta_n]} \rho(\hat{\mu}, \mu).$$

В работе также приведена асимптотика

$$R_n(l_0[\eta_n]) \sim cn\eta_n \ln \eta_n^{-1},$$

где $c = c(\sigma)$.

При пороговой обработке иногда также используется так называемый универсальный порог $T_U = \sigma\sqrt{2\ln n}$, предложенный в работе [5]. Исследования в [6, 7] показали, что порог T_U в определенном смысле максимальный и рассматривать пороги выше него не имеет смысла. Более того, нетрудно показать, что $t_k < T_U$ для всех k и всех достаточно больших n. В связи с этим всюду далее полагаем, что порог $\hat{t}_F$ выбирается на отрезке $[0; T_U]$.

3 Асимптотика среднеквадратичного риска FDR-процедуры в условиях слабой зависимости

Перейдем к исследованию асимптотики риска оценки $\hat{\mu}_F$ в случае, когда компоненты вектора x слабо зависимы, а именно: имеют достаточно быстро убывающий коэффициент сильного перемешивания [8]

$$\alpha(k) = \sup_{1 \le m \le n} \alpha\left(\sigma(x_i, i \le m), \sigma(x_i, i \ge m+k)\right), \quad k = 1, \ldots, n-1,$$

где

$$\alpha(\mathcal{B}, \mathcal{C}) = \sup_{B \in \mathcal{B}, C \in \mathcal{C}} |\mathsf{P}(BC) - \mathsf{P}(B)\mathsf{P}(C)|.$$

Отметим, что для любой измеримой функции $f(\cdot)$ коэффициент сильного перемешивания набора $f(x_1), \ldots, f(x_n)$ не больше коэффициента сильного перемешивания набора $x_1, \ldots, x_n$.

Покажем справедливость следующего вспомогательного утверждения.

Утверждение 1. *Пусть для набора действительных случайных величин $X_1, \ldots, X_n$ с коэффициентом сильного перемешивания $\alpha(\cdot)$ выполняется $\mathsf{E}X_i = 0$, $|X_i| \le b$, $i = 1, \ldots, n$. Тогда для любого целого числа $m \in [1; n/2]$ и любого $\varepsilon > 0$ справедливо*

$$\mathsf{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \varepsilon\right) \le 2\exp\left\{-\frac{\varepsilon^2}{32v_0} B\left(\frac{nb\varepsilon}{8mv_0}\right)\right\} + + 2\exp\left\{-\frac{\varepsilon^2}{32v_1} B\left(\frac{nb\varepsilon}{8mv_1}\right)\right\} + + 22\left(1 + \frac{4bn}{\varepsilon}\right)^{1/2} m\alpha\left(\left[\frac{n}{2m}\right]\right), \quad (1)$$

где

$$v_0 = \sum_{j=1}^{m} \mathsf{E}\left(([2(j-1)p] + 1 - 2(j-1)p) \times \times X_{[2(j-1)p]+1} + X_{[2(j-1)p]+2} + \cdots + X_{[(2j-1)p]} + + ((2j-1)p - [(2j-1)p])X_{[(2j-1)p]+1}\right)^2;$$

$$v_1 = \sum_{j=1}^{m} \mathsf{E}\left(([(2j-1)p] + 1 - (2j-1)p) \times \times X_{[(2j-1)p]+1} + X_{[(2j-1)p]+2} + \cdots + X_{[2jp]} + + (2jp - [2jp])X_{[2jp]+1}\right)^2;$$

$p = \dfrac{n}{2m}$; $B(\lambda) = 2\lambda^{-2}((1+\lambda)\ln(1+\lambda)-\lambda)$, $\lambda > 0$.

Доказательство. При доказательстве теоремы 1.3 из [8] показано, что

$$\mathsf{P}\left(\left|\sum_{j=1}^{m} V_j\right| > \frac{\varepsilon}{2}\right) \le \mathsf{P}\left(\left|\sum_{j=1}^{m} W_j\right| > \frac{\varepsilon}{4}\right) + 11\left(1+\frac{4bn}{\varepsilon}\right)^{1/2} m\alpha\left(\left[\frac{n}{2m}\right]\right),$$

где

$$V_j = ([2(j-1)p]+1-2(j-1)p)\times \times X_{[2(j-1)p]+1} + X_{[2(j-1)p]+2} + \cdots + X_{[(2j-1)p]} + + ((2j-1)p - [(2j-1)p])X_{[(2j-1)p]+1},$$

$$W_j \stackrel{d}{=} V_j,\ j = 1,\ldots,m,$$

а случайные величины $W_1,\ldots,W_m$ независимы. Применяя для случайных величин $W_1,\ldots,W_m$ неравенство Беннета [9], получим

$$\mathsf{P}\left(\left|\sum_{j=1}^{m} W_j\right| > \frac{\varepsilon}{4}\right) \le 2\exp\left\{-\frac{\varepsilon^2}{32v_0}\,B\left(\frac{pb\varepsilon}{4v_0}\right)\right\}.$$

Проводя аналогичные рассуждения для случайных величин

$$V'_j = ([(2j-1)p]+1-(2j-1)p)\times \times X_{[(2j-1)p]+1} + X_{[(2j-1)p]+2} + \cdots + X_{[2jp]} + + (2jp - [2jp])X_{[2jp]+1},\ j = 1,\ldots,m,$$

и объединяя результаты, с учетом

$$\mathsf{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \varepsilon\right) \le \le \mathsf{P}\left(\left|\sum_{i=1}^{m} V_i\right| > \frac{\varepsilon}{2}\right) + \mathsf{P}\left(\left|\sum_{i=1}^{m} V'_i\right| > \frac{\varepsilon}{2}\right)$$

получаем требуемое. □

Замечание. Из непрерывности правой части неравенства (1) по ε следует, что в левой части можно заменить знак $>$ на $\ge$.

Введем следующие обозначения:

$$k_n = [\eta_n n];\ \gamma_n = (\ln\ln n)^{-1};\ \kappa_n = (1-q_n-\gamma_n)^{-1}k_n;$$

$$p_i = \mathsf{P}(|x_i| \ge t_k),\ X_i = \mathbb{1}(|x_i| \ge t_k) - p_i,\ i = 1,\ldots,n;$$

$$N(t_k) = \#\{i : |x_i| \ge t_k\};\ M = \mathsf{E}N(t_k) = \sum_{i=1}^{n} p_i.$$

Заметим, что $\mathsf{E}X_i = 0$, $|X_i| < 1$, $\mathsf{D}X_i = p_i(1-p_i)$ для всех i, k.

Лемма 1. *Пусть $\eta_n \le b < 1$, $m \in [1; n/2] \cap \mathbb{N}$, а $\alpha(\cdot)$ — коэффициент сильного перемешивания компонент вектора x. Для некоторого $N \in \mathbb{N}$ при $n \ge N$ справедливо*

$$\sup_{\mu\in l_0[\eta_n]} \mathsf{P}\left(\hat{k}_F \ge \kappa_n\right) \le 4n\exp\left\{-\frac{(1-b)m}{64n}\kappa_n q_n \gamma_n^2\right\} + + 22\left(1+\frac{4n}{(1-b)\kappa_n q_n \gamma_n}\right)^{1/2} nm\alpha\left(\left[\frac{n}{2m}\right]\right).$$

Доказательство. Фиксируем $\mu \in l_0[\eta_n]$. Имеем

$$\mathsf{P}(\hat{k}_F \ge \kappa_n) \le \sum_{k\ge\kappa_n} \mathsf{P}(N(t_k) \ge k). \qquad (2)$$

Фиксируем $k \ge \kappa_n$; задача – ограничить вероятность $\mathsf{P}(N(t_k) \ge k)$ сверху. Ниже показано, что $M < k$ для всех $k \ge \kappa_n$ и всех достаточно больших n. По утверждению 1 имеем:

$$\mathsf{P}(N(t_k) \ge k) = \mathsf{P}\left(\sum_{i=1}^{n} X_i \ge k - M\right) \le \le \mathsf{P}\left(\left|\sum_{i=1}^{n} X_i\right| \ge k - M\right) \le \le 2\exp\left\{-\frac{(k-M)^2}{32v_0}\,B\left(\frac{n(k-M)}{8mv_0}\right)\right\} + + 2\exp\left\{-\frac{(k-M)^2}{32v_1}\,B\left(\frac{n(k-M)}{8mv_1}\right)\right\} + + 22\left(1+\frac{4n}{(k-M)}\right)^{1/2} m\alpha\left(\left[\frac{n}{2m}\right]\right). \qquad (3)$$

Для произвольного набора центрированных случайных величин $\xi_1,\ldots,\xi_l$ с конечными дисперсиями справедливо

$$\mathsf{E}(\xi_1+\cdots+\xi_l)^2 \le l\sum_{i=1}^{l} \mathsf{D}\xi_i,$$

откуда

$$v_{0,1} \le \left(\left[\frac{n}{2m}\right]+1\right)\sum_{i=1}^{n} p_i(1-p_i) \le \frac{nM}{m}.$$

Рассмотрим первое слагаемое в (3). Пусть сначала $n(k-M)/(8mv_0) \le 1$. Так как функция $B(\lambda)$ убывает по λ и $v_0 \le nM/m$, то

$$\frac{(k-M)^2}{32v_0}\,B\left(\frac{n(k-M)}{8mv_0}\right) \geq \frac{(k-M)^2 m}{32nM}\,B(1) = \frac{mM}{32n}\left(\frac{k}{M}-1\right)^2 B(1).$$

Если же $n(k-M)/(8mv_0) > 1$, то, поскольку $\lambda B(\lambda)$ возрастает по λ при $\lambda \geq 1$,

$$\frac{(k-M)^2}{32v_0}\,B\left(\frac{n(k-M)}{8mv_0}\right) \geq \frac{(k-M)m}{4n}\,B(1) = \frac{mM}{4n}\left(\frac{k}{M}-1\right) B(1).$$

Объединяя данные результаты, с учетом $B(1) > 1/2$ получим

$$2\exp\left\{-\frac{(k-M)^2}{32v_0}\,B\left(\frac{n(k-M)}{8mv_0}\right)\right\} \leq 2\exp\left\{-\frac{mM}{64n}\min\left\{\left(\frac{k}{M}-1\right)^2, \left(\frac{k}{M}-1\right)\right\}\right\}$$

и аналогично для слагаемого с v_1.

Перейдем к поиску границ возможных значений M. Вспомним, что в векторе $\mu \in l_0[\eta_n]$ не более $k_n = [\eta_n n] \leq bn$ ненулевых, а следовательно, и не менее $n - k_n$ нулевых компонент, откуда при $k \geq \kappa_n$ для M получим оценку снизу:

$$M = \sum_{i=1}^{n} p_i \geq k_n \cdot 0 + (n-k_n)\frac{kq_n}{n} \geq \frac{(n-k_n)}{n}\,kq_n \geq (1-b)\kappa_n q_n.$$

С другой стороны,

$$M \leq k_n \cdot 1 + (n-k_n)\frac{kq_n}{n} = k_n + \left(1-\frac{k_n}{n}\right)kq_n.$$

Рассмотрим функцию

$$g(x) = \frac{x}{k_n + (1-k_n/n)xq_n}.$$

Тогда $k/M \geq g(k)$ для любого k. Заметим, что функция $g(x)$ возрастает по x, поэтому при $k \geq \kappa_n$ имеем

$$g(k) \geq g(\kappa_n) = \frac{\kappa_n}{k_n + (1-k_n/n)\kappa_n q_n} = \left(1-\gamma_n - \frac{k_n q_n}{n}\right)^{-1} > 1+\gamma_n,$$

откуда

$$\left(\frac{k}{M}-1\right) > \gamma_n.$$

Также здесь показано, что $k > M$.

Наконец, заметим, что неравенство

$$\left(\frac{k}{M}-1\right)^2 > \frac{k}{M}-1$$

выполняется лишь в случае

$$\frac{k}{M}-1 > 1\,.$$

Но тогда тем более

$$\frac{k}{M}-1 > \gamma_n^2,$$

откуда

$$\min\left\{\left(\frac{k}{M}-1\right)^2, \left(\frac{k}{M}-1\right)\right\} > \gamma_n^2.$$

Объединяя выписанные неравенства, получим

$$\mathsf{P}(N(t_k) \geq k) \leq 4\exp\left\{-\frac{(1-b)m}{64n}\,\kappa_n q_n \gamma_n^2\right\} + 22\left(1+\frac{4n}{(1-b)\kappa_n q_n \gamma_n}\right)^{1/2} m\alpha\left(\left[\frac{n}{2m}\right]\right),$$

что вместе с (2) дает утверждение леммы. □

Обозначим $T_1 = \sigma\sqrt{2\ln \eta_n^{-1}}$.

Лемма 2. *Пусть* $\eta_n \geq n^{-\delta_1}$, $\delta_1 < 1$; $\lim \eta_n = 0$; $q_n \leq 1/2$; $\liminf q_n \ln n \geq 2c_3 > 0$; *а также существуют такие константы* $c_1, c_2 > 0$, *что* $\alpha(k) \leq c_1 k^{-1-(9/2)\delta_1/(1-\delta_1)-c_2}$ *для любого* $k \in \mathbb{N}$. *Тогда при* $n \to \infty$

$$\sup_{\mu \in l_0[\eta_n]} \mathsf{P}\left(\hat{t}_F \leq T_1\right) = o\left(\eta_n^2\right).$$

Доказательство. Используя требование $q_n \leq 1/2$ и свойства квантилей нормального распределения [3], можно показать, что при всех достаточно больших n справедливо

$$t_{\kappa_n} \equiv \sigma z\left(\frac{q_n\kappa_n}{2n}\right) > T_1,$$

откуда

$$\mathsf{P}\left(\hat{t}_F \leq T_1\right) \leq \mathsf{P}\left(\hat{t}_F \leq t_{\kappa_n}\right) = \mathsf{P}\left(\hat{k}_F \geq \kappa_n\right).$$

Заметив, что $\gamma_n > \ln^{-1} n$, $\kappa_n > \eta_n n/2$, $q_n > c_3 \ln^{-1} n$ для всех достаточно больших n, и применив лемму 1 с $m = n^{\delta_1}\ln^5 n$, получим требуемое. □

Следующее утверждение доказано в [3] для $\sigma = 1$ и $T_1 \geq 3^{1/4}$, приведенное ниже обобщение элементарно.

Лемма 3. *Пусть $\hat{t}$ — произвольный случайный порог, $\eta_n \le b < 1$, $x_i \sim N(\mu_i, \sigma^2)$, $(\hat{\mu})_i = p(x_i, \hat{t})$, $i = 1, \ldots, n$. Тогда с некоторой константой $c \equiv c(\sigma, b)$*

$$\mathsf{E}\,||\hat{\mu} - \mu||^2\, \mathbb{1}(\hat{t} \le T_1) \le cT_1^2 n \left(\mathsf{P}(\hat{t} \le T_1)\right)^{1/2}.$$

Перейдем, наконец, к основному утверждению работы.

Теорема 2. *Пусть выполнены требования леммы* 2. *При $n \to \infty$*

$$\sup_{\mu \in l_0[\eta_n]} \rho(\hat{\mu}_F, \mu) \le n\eta_n T_U^2 (1 + o(1)).$$

Доказательство. Пусть $\mu \in l_0[\eta_n]$. Имеем

$$\rho(\hat{\mu}_F, \mu) = \mathsf{E}\,||\hat{\mu}_F - \mu||^2\, \mathbb{1}(\hat{t}_F \le T_1) + \mathsf{E}\,||\hat{\mu}_F - \mu||^2\, \mathbb{1}(\hat{t}_F > T_1). \quad (4)$$

Используя леммы 2 и 3, для первого слагаемого в (4) получим

$$\mathsf{E}\,||\hat{\mu}_F - \mu||^2\, \mathbb{1}(\hat{t}_F \le T_1) \le n\, o(\eta_n) \ln \eta_n^{-1}. \quad (5)$$

Заметим, что

$$(p(x_i, t) - \mu_i)^2 = \begin{cases} (x_i - \mu_i)^2, & |x_i| > t; \\ \mu_i^2, & |x_i| \le t. \end{cases}$$

Отсюда для второго слагаемого в (4)

$$\mathsf{E}\,||\hat{\mu}_F - \mu||^2\, \mathbb{1}(\hat{t}_F > T_1) = \mathsf{E}\sum_{i=1}^n \left(p(x_i, \hat{t}_F) - \mu_i\right)^2 \mathbb{1}(T_1 < \hat{t}_F \le T_U) = \mathsf{E}\sum_{i=1}^n \left((x_i - \mu_i)^2 \mathbb{1}(|x_i| > \hat{t}_F) + \mu_i^2 \mathbb{1}(|x_i| \le \hat{t}_F)\right) \mathbb{1}(T_1 < \hat{t}_F \le T_U) \le$$

$$\le \mathsf{E}\sum_{i=1}^n \left((x_i - \mu_i)^2 \mathbb{1}(|x_i| > T_1) + \mu_i^2 \mathbb{1}(|x_i| \le T_U)\right) \equiv E_1 + E_2,$$

где

$$E_1 = \mathsf{E}\sum_{i:|\mu_i|>0} \left((x_i - \mu_i)^2 \mathbb{1}(|x_i| > T_1) + \mu_i^2 \mathbb{1}(|x_i| \le T_U)\right);$$

$$E_2 = \mathsf{E}\sum_{i:|\mu_i|=0} x_i^2 \mathbb{1}(|x_i| > T_1).$$

Пусть $\xi \sim N(0,1)$, $x > 0$, тогда

$$\mathsf{E}\xi^2 \mathbb{1}(|\xi| > x) \le 2\left(x + \frac{1}{x}\right)\varphi(x),$$

где использовано неравенство $1 - \Phi(x) \le \varphi(x)/x$, $x > 0$ ($\Phi(x)$ и $\varphi(x)$ — соответственно функция распределения и плотность $N(0,1)$). Отсюда

$$E_2 \le \sqrt{\frac{2}{\pi}}\, n \frac{T_1}{\sigma}\, e^{-T_1^2/(2\sigma^2)}(1 + o(1)) = O\left(n\eta_n \sqrt{\ln \eta_n^{-1}}\right). \quad (6)$$

Пусть далее $\xi \sim N(\mu, \sigma)$, тогда если $|\mu| \le T_U$, то $\mu^2 \mathsf{P}(|\xi| \le T_U) \le T_U^2$. Если же $\mu > T_U$ (для $\mu < -T_U$ аналогично), используя $2(1 - \Phi(x)) \le e^{-x^2/2}$ для $x \ge 0$, получим

$$\mu^2 \mathsf{P}(|\xi| \le T_U) < \mu^2 \left(1 - \Phi\left(\frac{\mu - T_U}{\sigma}\right)\right) \le \le \frac{\mu^2}{2} e^{-(\mu - T_U)^2/(2\sigma^2)} \le \frac{T_U^2}{2} + O(T_U),$$

где последнее неравенство можно получить, исследуя выражение в левой части на экстремум по μ. Из приведенных соотношений следует, что

$$E_1 \le n\eta_n \sigma^2 + \sum_{i:|\mu_i|>0} \mu_i^2 \mathsf{P}(|x_i| \le T_U) \le n\eta_n T_U^2 (1 + o(1)). \quad (7)$$

Объединяя (5)–(7), получаем утверждение теоремы. □

От степени разреженности вектора μ (скорости убывания η_n) зависит асимптотический порядок верхней границы риска, полученной в теореме 2. Например, при $\eta_n = n^{-\delta}$, $\delta \in (0,1)$, получим

$$\sup_{\mu \in l_0[\eta_n]} \rho(\hat{\mu}_F, \mu) \le 2\sigma^2\, n^{1-\delta} \ln n\, (1 + o(1));$$

если же $\eta_n = (\ln n)^{-r}$, $r > 0$, то

$$\sup_{\mu \in l_0[\eta_n]} \rho(\hat{\mu}_F, \mu) \le 2\sigma^2\, n (\ln n)^{1-r}\, (1 + o(1)).$$

Литература

1. *Menyhart O., Weltz B., Gyorffy B.* MultipleTesting.com: A tool for life science researchers for multiple hypothesis testing correction // PLoS One, 2021. Vol. 16. No. 6. Art. 0245824.
2. *Krylov V. A., Moser G., Serpico S. B., Zerubia J.* False discovery rate approach to unsupervised image change detection // IEEE T. Image Process., 2016. Vol. 25. No. 10. P. 4704–4718.
3. *Abramovich F., Benjamini Y., Donoho D., Johnstone I.* Adapting to unknown sparsity by controlling the false discovery rate // Ann. Stat., 2006. Vol. 34. No. 2. P. 584–653.

4. *Farcomeni A.* Some results on the control of the false discovery rate under dependence // Scand. J. Stat., 2007. Vol. 34. No. 2. P. 275–297.

5. *Donoho D., Johnstone I.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.

6. *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Am. Stat. Assoc., 1995. Vol. 90. P. 1200–1224.

7. *Marron J. S., Adak S., Johnstone I. M., Neumann M. H., Patil P.* Exact risk analysis of wavelet regression // J. Comput. Graph. Stat., 1998. Vol. 7. P. 278–309.

8. *Bosq D.* Nonparametric statistics for stochastic processes: Estimation and prediction. — Lecture notes in statistics ser. — New York, NY, USA: Springer, 1996. Vol. 110. 188 p.

9. *Pollard D.* Convergence of stochastic processes. — Springer ser. in statistics. — New York, NY, USA: Springer, 1984. 215 p.

Поступила в редакцию 05.12.22

MEAN-SQUARE RISK OF THE FDR PROCEDURE UNDER WEAK DEPENDENCE

M. O. Vorontsov[1,2] and O. V. Shestakov[1,2,3]

[1]M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2]Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[3]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: In many application areas, the problem of processing large amounts of data arises. In this case, before processing, the data array is often subjected to some transformation leading to a "sparse" or "economical" representation in which the absolute value of most elements of the array is equal to zero (or sufficiently small). In addition, as a result of interference when receiving and transmitting data, they become corrupted with noise and it is desirable to remove this noise during further processing. The resulting task is mathematically equivalent to some multiple hypothesis testing problems. Previously, to solve this problem under conditions of normality, independence, and sparsity of data, a procedure based on the method of controlling the average proportion of erroneously rejected hypotheses was proposed (False Discovery Rate, FDR). In this paper, the authors study the asymptotics of the mean-square risk of this procedure in the case of a weak dependence in the data.

Keywords: thresholding; multiple hypothesis testing; mean-square risk

DOI: 10.14357/19922264230205 **EDN:** AVJZDX

References

1. Menyhart, O., B. Weltz, and B. Gyorffy. 2021. MultipleTesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLoS One* 16(6):0245824. doi: 10.1371/journal.pone.0245824.
2. Krylov, V. A., G. Moser, S. B. Serpico, and J. Zerubia. 2016. False discovery rate approach to unsupervised image change detection. *IEEE T. Image Process.* 25(10):4704–4718. doi: 10.1109/TIP.2016.2593340.
3. Abramovich, F., Y. Benjamini, D. Donoho, and I. M. Johnstone. 2006. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Stat.* 34(2):584–653. doi: 10.1214/009053606000000074.
4. Farcomeni, A. 2007. Some results on the control of the false discovery rate under dependence. *Scand. J. Stat.* 34(2):275–297. doi: 10.1111/j.1467-9469.2006.00530.x.
5. Donoho, D., and I. M. Johnstone. 1994. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3):425–455. doi: 10.1093/biomet/81.3.425.
6. Donoho, D., and I. M. Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* 90(432):1200–1224. doi: 10.1080/01621459.1995.10476626.
7. Marron, J. S., S. Adak, I. M. Johnstone, M. H. Neumann, and P. Patil. 1998. Exact risk analysis of wavelet regression. *J. Comput. Graph. Stat.* 7(3):278–309. doi: 10.1080/10618600.1998.10474777.
8. Bosq, D. 1996. *Nonparametric statistics for stochastic processes: Estimation and prediction*. Lecture notes in statistics ser. New York, NY: Springer Verlag. 188 p.
9. Pollard, D. 1984. *Convergence of stochastic processes.* Springer ser. in statistics. New York, NY: Springer. 215 p.

Received December 5, 2022

Contributors

Vorontsov Mikhail O. (b. 1996) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; mathematician, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; m.vtsov@mail.ru

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; leading scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; oshestakov@cs.msu.su

ИССЛЕДОВАНИЕ РОБАСТНОСТИ ЧИСЛЕННЫХ АППРОКСИМАЦИЙ ФИЛЬТРА ВОНЭМА*

А. В. Босов[1]

Аннотация: Исследуются свойства решения задачи оптимальной фильтрации состояния непрерывной цепи Маркова по линейным наблюдениям, зашумленным винеровским процессом, в предположении неполной информации о его интенсивности. Неопределенность системы наблюдения задается верхней границей интенсивности шума. Численная реализация оптимального решения в постановке с полной информацией, обеспечиваемого фильтром Вонэма, не гарантирует устойчивости. Показано, что фильтр Вонэма в постановке с неопределенностью является робастным по отношению к интенсивности шума, если параметры модели не приводят к его расходимости. В общем случае неустойчивость численной схемы Эйлера–Маруямы фильтра Вонэма сохраняется. Простые эвристические приемы, обеспечивающие устойчивые аппроксимации фильтра Вонэма, показывают работоспособность для более широкого набора параметров. Однако в постановке с неопределенностью удается привести примеры, когда такие эвристические фильтры показывают неприемлемо низкое качество. Лучшее решение дают дискретизованные фильтры — аппроксимации фильтра Вонэма, реализованные для дискретной модели, аппроксимирующей исходную непрерывную систему наблюдения. На серии численных экспериментов показано, что данные фильтры обладают робастностью для всех наборов параметров. Если в расчетах среди смоделированных траекторий фильтра Вонэма нет расходящихся, то дискретизованные фильтры немного проигрывают. Если расходящиеся траектории есть, то выигрыш дискретизованных фильтров носит абсолютный характер.

Ключевые слова: марковский скачкообразный процесс; стохастическая фильтрация; робастное оценивание; фильтр Вонэма; численная схема Эйлера–Маруямы; дискретизованные фильтры

DOI: 10.14357/19922264230206 **EDN:** BGILKR

1 Введение

Задача стохастической фильтрации — оценивания состояния динамической системы по косвенным наблюдениям — одна из самых востребованных задач стохастического анализа. Наиболее яркие решения, такие как фильтр Калмана–Бьюси [1], фильтр Вонэма [2], условно-гауссовский фильтр [3], были и остаются источниками актуальных постановок для фундаментальных и прикладных исследований. Одно из наиболее известных направлений развития классических методов оценивания дают постановки с неопределенностью, т. е. неполной априорной информацией о параметрах системы наблюдения. В этих постановках, как правило, исследуются методы, обеспечивающие устойчивость оценок по отношению к неопределенности, т. е. обладающие свойством робастности. Среди робастных оценок значительное место занимают те, которые обладают минимаксным свойством. Эти оценки дают наилучший результат в предположении, что реализуется наихудший сценарий неопределенности, что и позволяет считать их робастными.

В фокусе данной работы неопределенность, связанная с отсутствием точной априорной информации об интенсивности шума в наблюдениях. Самые известные и содержательные результаты для такого типа неопределенности дает фильтр Калмана [4–12]. Вне зависимости от того, как задана неопределенность и какова постановка (фильтрация, управление, оценивание по интегральному критерию), решение минимаксной задачи обеспечивается исключительными свойствами фильтра Калмана — линейностью оценки и уравнением Риккати для ковариации ее ошибки.

Предполагалось исследовать аналогичные свойства в постановке, отличающейся от фильтрации Калмана–Бьюси моделью состояния тем, что вместо линейного гауссовского процесса состояние задается дискретной цепью Маркова, т. е. решается задача фильтрации Вонэма [2]. Несмотря на большое сходство фильтров Калмана–Бьюси и Вонэма,

*Исследование выполнено за счет гранта Российского научного фонда (проект 22-28-00588) с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, ABosov@frccsc.ru

между ними есть принципиальная разница. Последний — существенно нелинейный, и записать его точность не удается. По этой причине непосредственное решение минимаксной задачи не представляется возможным. Но с практической точки зрения это обстоятельство не играет особой роли, поскольку в связи с применением фильтра Вонэма становится гораздо актуальнее задача обеспечения устойчивости его численной реализации. Именно этот аспект ставится предметом исследования в статье и детализирован в следующем разделе с постановкой задачи. Поскольку возможности теоретического аппарата здесь ограничены, то акцент сделан на практическое исследование.

2 Постановка задачи робастной фильтрации

На каноническом вероятностном пространстве $(\Omega, \mathcal{F}, \mathcal{P}, \mathcal{F}_t)$, $t \in [0, T]$, рассмотрим стохастическую систему наблюдения с вектором состояния y_t и линейными наблюдениям z_t следующего вида:

$$dy_t = \Lambda_t' y_t\, dt + d\Lambda_t^y, \qquad y_0 = Y\,; \quad (1)$$

$$dz_t = a_t y_t\, dt + b_t z_t\, dt + \sigma_t\, dw_t,\ z_0 = Z\,. \quad (2)$$

Уравнение (1) определяет марковский скачкообразный процесс — цепь с конечным числом состояний и значениями в множестве $\{e_1, \ldots, e_{n_y}\}$, состоящем из единичных координатных векторов в евклидовом пространстве $\mathbb{R}^{n_y}$. Предполагается, что начальное состояние Y имеет известное распределение π, Λ_t — матрица интенсивностей переходов, Λ_t' — транспонированная матрица Λ_t, а Λ_t^y — $\mathcal{F}_t$-согласованный мартингал [13]. Уравнение (2) представляет косвенные наблюдения $z_t \in \mathbb{R}^{n_z}$ за состоянием цепи y_t, порождаемую ими σ-алгебру будем обозначать $\mathcal{F}_t^z$ и предполагать, что $\mathcal{F}_t^z \subseteq \mathcal{F}_t \subseteq \subseteq \mathcal{F}$. Далее, $w_t \in \mathbb{R}^{n_w}$ — не зависящий от Λ_t^y, Y и Z стандартный векторный винеровский процесс; $Z \in \mathbb{R}^{n_z}$ — гауссовский случайный вектор с известными моментами $\mathbb{E}\{Z\}$ и $\mathbb{E}\{ZZ'\}$; матричные функции $a_t \in \mathbb{R}^{n_z \times n_y}$, $b_t \in \mathbb{R}^{n_z \times n_z}$ и $\sigma_t \in \mathbb{R}^{n_z \times n_w}$ предполагаются ограниченными: $|a_t| + |b_t| + |\sigma_t| \le C$ для всех $0 \le t \le T$, что обеспечивает существование решения уравнения (2). Кроме того, ошибки наблюдений предполагаются невырожденными, $\sigma_t \sigma_t' > 0$. Через $\mathbb{E}\{\cdot\}$ и $\mathbb{E}\{\cdot|\cdot\}$ обозначены операторы безусловного и условного математического ожидания соответственно.

Основное условие рассматриваемой задачи — предположение об отсутствии точной информации об интенсивности ошибок наблюдений, выражаемое неравенством $\sigma_t \sigma_t' \le \Sigma_t$, в котором известна положительно определенная матрица Σ_t, имеющая смысл верхней границы ковариации шума. Задачей ставится анализ свойств алгоритмов оценивания состояния цепи y_t (фильтров) по $\mathcal{F}_t^z$. Точное решение задачи минимаксной фильтрации

$$\hat{y}_t^{\min\max} = \arg\min_{\hat{y}_t} \max_{\sigma_t : \sigma_t \sigma_t' \le \Sigma_t} \mathbb{E}\left\{ |\hat{y}_t - y_t|^2 \right\}$$

получить затруднительно, но для дальнейших целей можно исходить из интуитивного предположения, что

$$\hat{y}_t^{\min\max} \approx \hat{y}_t^W(\Sigma_t),$$

где $\hat{y}_t^W(\sigma_t \sigma_t') = \mathbb{E}\{y_t | \mathcal{F}_t^z, \sigma_t\}$, т. е. задается выражением оптимального фильтра в задаче без неопределенности (при известной ковариации $\sigma_t \sigma_t'$), в котором эта неизвестная ковариация заменена максимально возможной Σ_t. На самом деле, даже если бы удалось доказать такое утверждение, это никак не приблизило бы к практическому решению задачи фильтрации, поскольку переход от записи оптимального фильтра Вонэма к его типовой численной реализации по схеме Эйлера–Маруямы [14] приводит к неустойчивой процедуре [15], и, соответственно, гораздо важнее предложить пусть неоптимальные в минимаксном смысле, но работоспособные устойчивые алгоритмы. Такими алгоритмами служат дискретизованные фильтры, предложенные в [16, 17] и подробно исследованные в [18]. В настоящей статье результаты дополнены анализом свойств этих фильтров в постановке с неопределенностью интенсивности шума в наблюдениях.

3 Некоторые сведения об оптимальных фильтрах

Для понимания дальнейшего обратим внимание на следующие известные для рассматриваемой задачи факты. Во-первых, при условии $\sigma_t \sigma_t' = \Sigma_t$, т. е. при отсутствии неопределенности, известное оптимальное решение $\hat{y}_t^W$ определяется фильтром Вонэма [2, 13]:

$$d\hat{y}_t^W = \Lambda_t' \hat{y}_t^W\, dt + \left(\operatorname{diag}\left(\hat{y}_t^W\right) - \right.$$
$$\left. - \hat{y}_t^W \left(\hat{y}_t^W\right)'\right) a_t' \Sigma_t^{-1} \left(dz_t - b_t z_t\, dt - a_t \hat{y}_t^W\, dt\right). \quad (3)$$

Во-вторых, это общее уравнение оптимальной фильтрации в терминах обновляющих процессов [3], записанное с учетом модели системы наблюдения (1), (2):

$$d\hat{y}_t^{\mathrm{Opt}} = \Lambda_t' \hat{y}_t^{\mathrm{Opt}} dt + \mathbb{E}\left\{ \left(y_t - \hat{y}_t^{\mathrm{Opt}}\right) y_t' | \mathcal{F}_t^z, \sigma_t \right\} \times$$
$$\times\, a_t' \Sigma_t^{-1} \left(dz_t - b_t z_t\, dt - a_t \hat{y}_t^{\mathrm{Opt}}\, dt\right). \quad (4)$$

И, наконец, уравнение фильтра Калмана–Бьюси [18, 19] для модели (1), (2), записанное в предположении, что вместо цепи y_t наблюдается гауссовский процесс, т. е. мартингал Λ_t^y предполагается винеровским процессом:

$$d\hat{y}_t^K = \\ = \Lambda_t' \hat{y}_t^K dt + P_t^K a_t' \Sigma_t^{-1} \left(dz_t - b_t z_t\, dt - a_t \hat{y}_t^K dt \right), \quad (5)$$

где матрица усиления P_t^K не зависит от $\mathcal{F}_t^z$ и определяется обыкновенным дифференциальным уравнением Риккати.

Сравнивая (3) и (4) и учитывая несмещенность оценок, нетрудно увидеть, что

$$\mathbb{E}\left\{ \left(y_t - \hat{y}_t^W\right) \left(y_t - \hat{y}_t^W\right)' | \mathcal{F}_t^z, \sigma_t \right\} = \\ = \operatorname{diag}\left(\hat{y}_t^W\right) - \hat{y}_t^W \left(\hat{y}_t^W\right)'$$

для традиционной постановки фильтрации Вонэма. Аналогично, сравнивая (4) и (5), нетрудно увидеть, что

$$\mathbb{E}\left\{ \left(y_t - \hat{y}_t^K\right) \left(y_t - \hat{y}_t^K\right)' | \mathcal{F}_t^z, \sigma_t \right\} = P_t^K$$

в традиционной постановке фильтрации Калмана. Поскольку здесь P_t^K не зависит от наблюдений, то и общее качество калмановской фильтрации

$$\mathbb{E}\left\{ \left(y_t - \hat{y}_t^K\right) \left(y_t - \hat{y}_t^K\right)' \right\} = P_t^K .$$

Это обстоятельство вместе с линейностью оценки $\hat{y}_t^K$ и становится тем движущим свойством, которое обеспечивает фильтру Калмана еще и минимаксность в линейно-гауссовской модели с неопределенностью $\sigma_t \sigma_t' \le \Sigma_t$, а именно:

$$\hat{y}_t^K(\Sigma_t) = \arg\min_{\hat{y}_t} \max_{\sigma_t : \sigma_t \sigma_t' \le \Sigma_t} \mathbb{E}\left\{ \left| \hat{y}_t^K(\sigma_t) - y_t \right|^2 \right\}. \quad (6)$$

Вычислить безусловную ковариацию $\mathbb{E}\left\{ \left(y_t - \hat{y}_t^W\right) \left(y_t - \hat{y}_t^W\right)' \right\}$ в модели фильтрации Вонэма, к сожалению, не удается. Кроме того, сама оценка (3) существенно нелинейна, поэтому получить аналогичное (6) свойство минимаксности для фильтра Вонэма пока не удалось. В качестве фундаментального результата это был бы большой успех. В практических целях особого значения это не имеет из-за неустойчивости численных схем фильтра Вонэма, и минимаксная постановка ничего здесь не изменит.

4 Дискретизованные фильтры для модели Вонэма

Устойчивые аппроксимации фильтра Вонэма [16, 17] реализованы для модели, которая получается из (1), (2) дискретизацией на некотором интервале $t \in [0, T]$ с заданным временны́м шагом δ:

$$0 = t_0;\ t_1 = t_0 + \delta;\ \ldots;\ t_i = t_{i-1} + \delta;\ \ldots \\ \ldots;\ t_{T/\delta - 1} + \delta = t_{T/\delta} = T .$$

Кроме того, предполагается, что $a_t \equiv const$, $b_t \equiv const$ и $\sigma_t \equiv const$ на интервалах дискретизации $[t_{i-1}, t_i]$. Это предположение легко реализуется заменой функций a_t, b_t и σ_t на их кусочно-постоянные аппроксимации. Чтобы привести наблюдения (2) к каноническому виду [20], можно рассмотреть процесс z_t^0, представляющий собой неупреждающее преобразование исходного наблюдаемого выхода z_t (т. е. выполнено равенство $\hat{y}_t = \\ = \mathbb{E}\{y_t | \mathcal{F}_t^z\} = \mathbb{E}\left\{ y_t | \mathcal{F}_t^{z^0} \right\}$):

$$z_t^0 = \int_0^t (dz_\tau - b_\tau z_\tau\, d\tau) = \int_0^t (a_\tau y_\tau\, d\tau + \sigma_\tau\, dw_\tau). \quad (7)$$

Далее введем новые наблюдения, дискретизованные по времени с шагом δ:

$$\Delta z_{t_i}^0 = \int_{t_{i-1}}^{t_i} (a_\tau y_\tau\, d\tau + \sigma_\tau\, dw_\tau).$$

Это приращения z_t^0 на интервалах дискретизации, и они порождают σ-алгебру

$$\mathcal{F}_{t_i}^{\Delta z^0} = \sigma\left\{ \Delta z_{t_j}^0, 1 \le j \le i \right\}.$$

Если обозначить через

$$\mu_i = \int_{t_{i-1}}^{t_i} y_\tau\, d\tau = \left(\mu_i^1, \ldots, \mu_i^{n_y}\right)'$$

случайный вектор, компоненты которого равны времени пребывания марковской цепи y_t в каждом из возможных состояний на интервале времени $(t_{i-1}, t_i]$, а через $\mathcal{N}(z; m, \sigma^2)$ гауссовскую плотность вероятности со средним m и дисперсией σ^2, вычисленную в точке z, то оценка $\hat{y}_{t_i} = \mathbb{E}\left\{ y_t | \mathcal{F}_{t_i}^{\Delta z^0} \right\}$ находится с помощью следующей рекуррентной процедуры [17]:

$$\left.\begin{aligned} \hat{y}_{t_i}^{\mathrm{opt}} &= \left(\mathbf{1} \hat{q}_{t_i}' \hat{y}_{t_{i-1}}^{\mathrm{opt}} \right)^{-1} \left(\hat{q}_{t_i}' \hat{y}_{t_{i-1}}^{\mathrm{opt}} \right); \\ \hat{q}_{t_i}^{k,j} &= \mathbb{E}\left\{ \mathcal{N}\left(\Delta z_{t_i}^0; a\mu_i, \delta\sigma\sigma' \right) y_{t_i}^j | y_{t_{i-1}} = e_k \right\}, \end{aligned}\right\} \quad (8)$$

где $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^{n_y}$ — вектор из единиц; начальное условие $\hat{y}_0^{\mathrm{opt}} = \pi_0$; матрица $\hat{q}_{t_i} = \\ = \left\| \hat{q}_{t_i}^{k,j} \right\|_{k,j=1}^{n_y}$. Величины $\hat{q}_{t_i}^{k,j}$ можно аппроксимировать, используя для задающих их интегралов одну из аппроксимаций [16]:

– схему «левых» прямоугольников (порядок точности $1/2$);
– схему «средних» прямоугольников (порядок точности 1);
– схему, основанную на квадратурах Гаусса (порядок точности 2).

Полностью соотношения приведены также в [18], где показано, что в большинстве экспериментов разница между оценками дискретизованных фильтров, связанная с выбором схемы интегрирования, незначительна и может не учитываться в рассматриваемом модельном примере, поэтому в расчетах все три оценки представлены первой, $\breve{y}_{t_i}^{\delta\,1/2}$.

Отдельного пояснения требует преобразование (7). Его смысл состоит в том, что в модели (2) без ограничения общности можно считать $b_t = 0$, причем как для дискретизованных фильтров, так и для фильтра Вонэма (3). Однако для численной реализации это оказывается не так. Как показали представленные далее расчеты, наблюдения, формируемые неустойчивым процессом z_t, который можно получить именно с помощью матрицы b_t, оказывают решающее влияние на устойчивость аппроксимации фильтра Вонэма.

5 Экспериментальное исследование

5.1 Модель

Основу для численных экспериментов дала модель механического привода, использованная в [18]. Отличие состоит в том, что здесь привод не управляется, поскольку решается только задача фильтрации, поэтому модель системы наблюдения имеет вид

$$\left.\begin{aligned} dx_t &= v_t\,dt, \quad t \in (0,T]\,; \\ dv_t &= ax_t\,dt + bv_t\,dt + cy_y\,dt + \sqrt{g}\,dw_t\,, \end{aligned}\right\} \tag{9}$$

где x_t — положение привода на горизонтальной оси; v_t — скорость. Физический смысл модели состоит в том, чтобы стабилизировать привод в положениях, задаваемых состояниями цепи, т. е. в точках $-c_i/a$ для $y_t = e_i$. Если y_t неизвестно, то его надо оценивать, используя v_t в качестве наблюдений.

Марковская цепь y_t в разных расчетах будет иметь три или четыре состояния. Для $n_y = 3$ постоянная матрица интенсивностей $\Lambda_t = \Lambda$ отвечает модели простого процесса рождения-гибели:

$$\Lambda = \begin{pmatrix} -0{,}5 & 0{,}5 & 0 \\ 0{,}5 & -1 & 0{,}5 \\ 0 & 0{,}5 & -0{,}5 \end{pmatrix} \text{ или } \Lambda = \begin{pmatrix} -5 & 5 & 0 \\ 5 & -1 & 5 \\ 0 & 5 & -5 \end{pmatrix}.$$

Для $n_y = 4$ также

$$\Lambda_t = \Lambda = \begin{pmatrix} -0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0 & 0 & 0{,}5 & -0{,}5 \end{pmatrix}.$$

Начальное распределение π во всех расчетах задает $y_0 = Y = e_1$

Скаляры a, b и g — известные постоянные и меняются в разных расчетах; строки c известны и равны соответственно $(c_1, c_2, c_3) = (-1, 0, 1)$ и $(c_1, c_2, c_3, c_4) = (-1{,}5; -0{,}5; 0{,}5; 1{,}5)$; w_t — стандартный винеровский процесс.

Наблюдения (9), как видно, сами представляют собой систему. Эта линейная система устойчива, если $b < 0$ и $b^2 + 4a < 0$, поскольку b и $b^2 + 4a$ — собственные числа матрицы системы $b_t = \begin{pmatrix} 0 & 1 \\ a & b \end{pmatrix}$, и неустойчива иначе.

Интегрирование во всех расчетах как для системы (9), так и для фильтров (3) и (8) выполнялось методом Эйлера с шагом $\delta = 0{,}001$. Дискретная цепь, аппроксимирующая y_t, моделировалась независимыми экспоненциальными величинами для каждого из 100 интервалов интегрирования длины δ, т. е. использовалась выборка из распределения $E(0{,}00001)$. Для моделирования системы наблюдения, т. е. переменных x_t и v_t, шаг интегрирования δ также разбивался на 100 интервалов длины $\delta/100$.

В [18] есть подробные пояснения и примеры расходимости аппроксимаций фильтра Вонэма, полученных в этих расчетах с помощью схемы Эйлера–Маруямы. В описываемых здесь расчетах также в качестве признака расходимости оценки $\hat{y}_t^W$ использовалось условие $\left|\left(\hat{y}_t^W\right)_k\right| > 1$ хотя бы для одного k, и при выполнении этого условия оценка фильтрации $\hat{y}_t^W$ возвращалась в предельное состояние:

$$\hat{y}_\tau^W = \pi_\infty = \begin{cases} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)' & \text{при } n_y = 3\,; \\ \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)' & \text{при } n_y = 4 \end{cases}$$

для момента времени τ, в который выполнилось это условие. Так получается $\hat{y}_t^{\lim}$. Второй вариант предотвратить расходимость — это замена текущей оценки на оценку предыдущего шага, т. е. $\hat{y}_\tau^W = \hat{y}_{\tau-\delta}^W$. Так получается $\hat{y}_t^{\mathrm{del}}$. Именно эти оценки участвуют далее в расчетах и анализируются на предмет робастности в отношении сформулированной задачи.

В каждом расчете моделировалось по 1000 траекторий системы (9) и оценок фильтров. Для характеризации качества оценок фильтрации $\hat{y}_t^{\lim}$, $\hat{y}_t^{\mathrm{del}}$

и $\breve{y}_{t_i}^{\delta^{1/2}}$ рассчитывались интегральные квадратичные ошибки $\hat{D}\big(\hat{y}_t^{\lim}\big)$, $\hat{D}\big(\hat{y}_t^{\mathrm{del}}\big)$ и $\hat{D}\Big(\breve{y}_{t_i}^{\delta^{1/2}}\Big)$ для

$$\hat{D}\left(\tilde{y}_{t_i}\right) = \hat{\mathbb{E}}\left\{\frac{\delta}{T}\sum_{i=1}^{T/\delta}\left(cy_{t_i} - c\tilde{y}_{t_i}\right)^2\right\},$$

где $\hat{\mathbb{E}}$ обозначает усреднение по пучку 1000 траекторий.

Далее, в каждом из расчетов выбирался «базовый» сценарий без неопределенности с известной интенсивностью шума $\sqrt{g} = 0{,}1$ Расчеты для анализа свойства робастности выполнялись в каждом случае для этого значения g, но в предположении, что известна лишь граница Σ_t, для которой выбирались два варианта: $\Sigma_t = 3g$ и $10g$. Наконец, для оценки верхней границы точности в задаче с неопределенностью расчет повторялся в предположении, что интенсивности известны, т. е. для $g = 0{,}03$ и $0{,}1$.

5.2 Устойчивые наблюдения

Устойчивый вариант системы (9) описывается значениями $a = -1$ и $b = -0{,}5$ Для этого случая эксперименты выполнены для цепей размерности $n_y = 3$ и 4 и для всех трех матриц интенсивностей, всего три эксперимента, в каждом по пять расчетов. Для удобства параметры каждой модели повторены в табл. 1–3, представляющих результаты.

Рассмотрим табл. 1. Первая строка, соответствующая базовому сценарию, повторяет основные выводы, сделанные в [18]. Среди смоделированных траекторий не слишком часто, но срабатывали условия $|(\hat{y}_t^W)_k| > 1$, т. е. фильтр Вонэма «не справлялся». На них по-разному реагировали аппроксимации $\hat{y}_t^{\lim}$ и $\hat{y}_t^{\mathrm{del}}$ (об этом свидетельствует их разная точность). Дискретизованные фильтры дали результат лучше. Но в тех двух расчетах, что должны дать ответ на основной вопрос о наличии робастных свойств (строки выделены полужирным), этого уже нет. В остальных расчетах расходящихся траекторий не было (точность $\hat{y}_t^{\lim}$ и $\hat{y}_t^{\mathrm{del}}$ одинаковая). Дискретизованные фильтры дали результат хуже.

Таблица 1 Фильтрация по устойчивым наблюдениям. Низкая интенсивность переключений

g	Σ_t	$\hat{D}\left(\hat{y}_t^{\lim}\right)$	$\hat{D}\left(\hat{y}_t^{\mathrm{del}}\right)$	$\hat{D}\left(\breve{y}_{t_i}^{\delta^{1/2}}\right)$
0,01	0,01	0,0540	0,0495	0,0486
	0,3	**0,0612**	**0,0612**	**0,0729**
	0,1	**0,0911**	**0,0911**	**0,1225**
0,03	0,03	0,0987	0,0987	**0,0986**
0,1	0,1	0,1837	0,1837	**0,1837**

Параметры: $a = -1$; $b = -0{,}5$; $(c_1, c_2, c_3) = (-1, 0, 1)$; $\Lambda = \begin{pmatrix} -0{,}5 & 0{,}5 & 0 \\ 0{,}5 & -1 & 0{,}5 \\ 0 & 0{,}5 & -0{,}5 \end{pmatrix}$.

Таблица 2 Фильтрация по устойчивым наблюдениям. Высокая интенсивность переключений

g	Σ_t	$\hat{D}\left(\hat{y}_t^{\lim}\right)$	$\hat{D}\left(\hat{y}_t^{\mathrm{del}}\right)$	$\hat{D}\left(\breve{y}_{t_i}^{\delta^{1/2}}\right)$
0,01	0,01	0,1959	0,1956	0,1948
	0,03	**0,2164**	**0,2164**	**0,2286**
	0,1	**0,3247**	**0,3247**	**0,3250**
0,03	0,03	0,3055	0,3055	**0,3064**
0,1	0,1	0,4477	0,4477	**0,4485**

Параметры: $a = -1$; $b = -0{,}5$; $(c_1, c_2, c_3) = (-1, 0, 1)$; $\Lambda = \begin{pmatrix} -5 & 5 & 0 \\ 5 & -1 & 5 \\ 0 & 5 & 5 \end{pmatrix}$.

Таблица 3 Фильтрация по устойчивым наблюдениям. Цепь размерности 4

g	Σ_t	$\hat{D}\left(\hat{y}_t^{\lim}\right)$	$\hat{D}\left(\hat{y}_t^{\mathrm{del}}\right)$	$\hat{D}\left(\breve{y}_{t_i}^{\delta^{1/2}}\right)$
0,01	0,01	0,3615	0,3562	0,0534
	0,03	**0,3583**	**0,3583**	**0,0811**
	0,1	**0,3751**	**0,3751**	**0,1384**
0,03	0,03	0,3957	0,3957	**0,1086**
0,1	0,1	0,4767	0,4767	**0,2058**

Параметры: $a = -1$; $b = -0{,}5$; $(c_1, c_2, c_3, c_4) = (-1{,}5; -0{,}5; 0{,}5; 1{,}5)$; $\Lambda = \begin{pmatrix} -0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0 & 0 & 0{,}5 & -0{,}5 \end{pmatrix}$.

Надо отметить, что в этом эксперименте параметры оказались такими, что обеспечили идеальные условия для фильтра Вонэма. Ситуация, когда расходящихся траекторий нет, как показано в [18], возникает не слишком часто. Вместе с тем робастностью дискретизованные фильтры обладают, хотя и дают худший результат, чем оценки $\hat{y}_t^{\lim}$ и $\hat{y}_t^{\mathrm{del}}$.

Обратимся к табл. 2. Этот расчет повторил те же выводы, что и предыдущий: робастные свойства есть у всех фильтров, но аппроксимации фильтра Вонэма выигрывают, хотя и менее заметно.

Перейдем к табл. 3. Параметры этого эксперимента принципиально отличаются от предыдущих тем, что уже в базовом сценарии аппроксимации фильтра Вонэма проигрывают очень сильно. Модель с неопределенностью «добавляет» устойчивости оценкам $\hat{y}_t^{\lim}$ и $\hat{y}_t^{\mathrm{del}}$ (их качество совпадает, расходящихся траекторий нет), но на превосходство дискретизованных фильтров это не влияет.

5.3 Неустойчивые наблюдения

Еще один эксперимент был выполнен для формально неустойчивой системы (9) со значениями

$a = 0$ и $b = 0$. В реальности такая система ведет себя довольно инертно и обнаружить в моделируемом пучке траектории системы, которые успевают разойтись за время $T = 10$, не удается. Расчеты с теми же параметрами, что и в табл. 1, привели к тем же в точности результатам, т. е. в табл. 1 можно считать $a = 0$ и $b = 0$. Это отвечает замене (7) и эквивалентной записи (2) с $b_t = 0$. Но следующий эксперимент показывает, что на практике так сделать можно не всегда. Единственное отличие последнего обсуждаемого эксперимента состоит в том, что для неустойчивой системы (9) использованы значения $a = 1$ и $b = 0{,}5$. Результаты приведены в табл. 4.

Таблица 4 Фильтрация по неустойчивым наблюдениям

g	Σ_t	$\hat{D}\left(\hat{y}_t^{\lim}\right)$	$\hat{D}\left(\hat{y}_t^{\mathrm{del}}\right)$	$\hat{D}\left(\breve{y}_{t_i}^{\delta^{1/2}}\right)$
0,01	0,01	0,4860	0,6432	0,0486
	0,03	**0,5491**	**0,6407**	**0,0729**
	0,1	**0,6107**	**0,6398**	**0,1225**
0,03	0,03	0,5660	0,6569	**0,0986**
0,1	0,1	0,6496	0,6796	**0,1837**

Параметры: $a = 1$; $b = 0{,}5$; $(c_1, c_2, c_3) = (-1, 0, 1)$; $\Lambda = \begin{pmatrix} -0{,}5 & 0{,}5 & 0 \\ 0{,}5 & -1 & 0{,}5 \\ 0 & 0{,}5 & -0{,}5 \end{pmatrix}$.

Тенденция ухудшения качества аппроксимаций фильтра Вонэма, наметившаяся в эксперименте из табл. 3, здесь продолжилась вплоть до того, что оценки $\hat{y}_t^{\lim}$ и $\hat{y}_t^{\mathrm{del}}$ перестали иметь смысл, потому что установившееся значение $\mathbb{E}\{|y_t|^2\} = 2/3$, а значит, тривиальная оценка $\mathbb{E}\{y_t\} = 0$ становится лучше этих оценок. При этом с дискретизованными фильтрами не произошло ничего — качество осталось ровно таким же, как и в устойчивом случае. Это в равной степени касается и их робастных свойств в отношении неопределенности границы Σ_t.

Интересно понять, что именно происходит с аппроксимациями фильтра Вонэма. Представление об этом дают рис. 1 и 2, на которых приведены

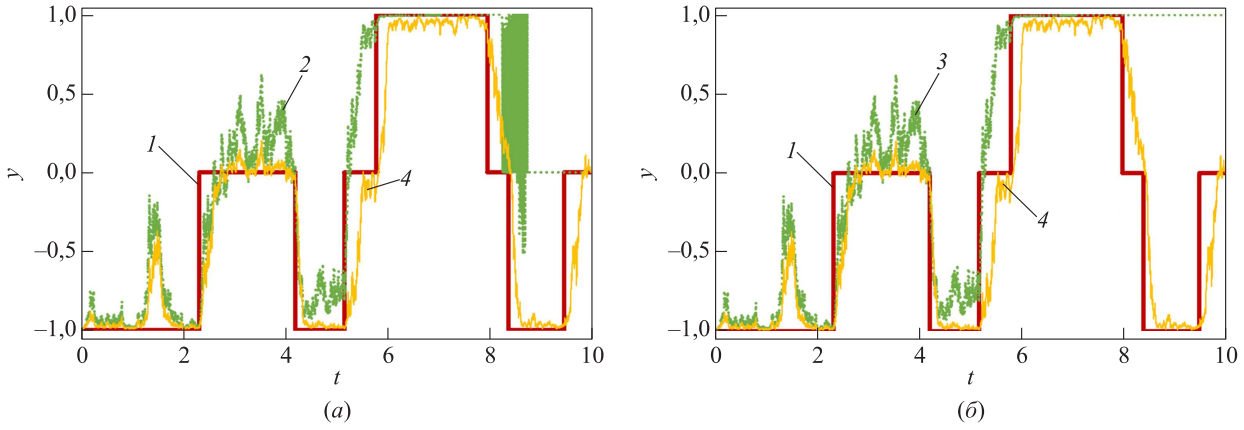


Рис. 1 Характерные траектории для $g = 0{,}01$ и $\Sigma_t = 0{,}03$: *1* — цепь y_t; *2* — оценка $\hat{y}_t^{\lim}$; *3* — оценка $\hat{y}_t^{\mathrm{del}}$; *4* — оценка $\breve{y}_{t_i}^{\delta^{1/2}}$

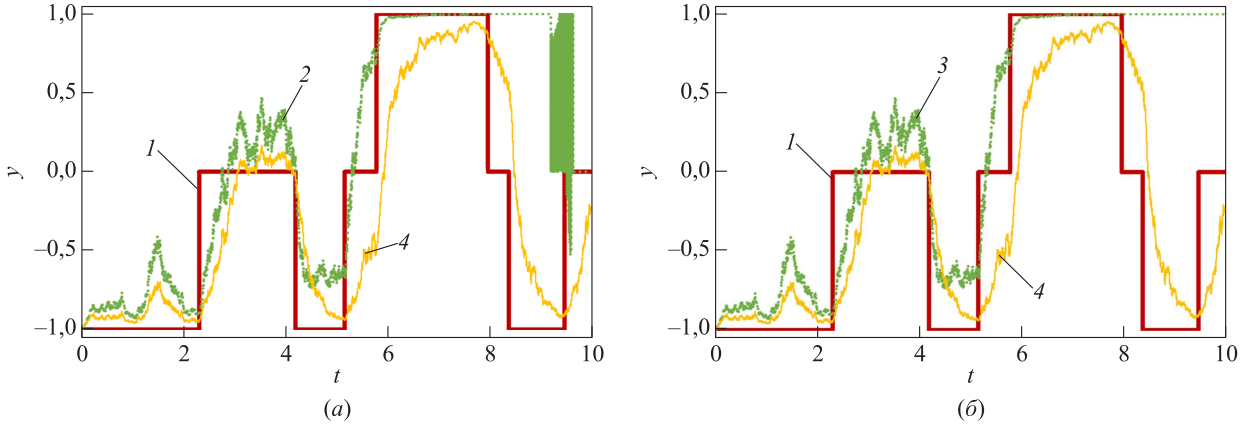


Рис. 2 Характерные траектории для $g = 0{,}01$ и $\Sigma_t = 0{,}1$: *1* — цепь y_t; *2* — оценка $\hat{y}_t^{\lim}$; *3* — оценка $\hat{y}_t^{\mathrm{del}}$; *4* — оценка $\breve{y}_{t_i}^{\delta^{1/2}}$

примеры одной и той же траектории цепи и всех оценок в двух расчетах, посвященных анализу робастности.

Видно, что оценка дискретизованного фильтра остается содержательной на всем интервале. При этом траектория $\breve{y}_{t_i}^{\delta^{1/2}}$, видимо, хуже отслеживает y_t для большего значения Σ_t, но фильтр работает. Обе аппроксимации фильтра Вонэма начинают проигрывать с самого начала, а в итоге «разваливаются», причем делают это по-разному: оценке $\hat{y}_t^{\lim}$ свойственны резкие многократные колебания, а оценка $\hat{y}_t^{\mathrm{del}}$ в итоге «зависает» и перестает работать. Надо отметить, что такое поведение имеют все смоделированные траектории. Окончательные итоги подведены в заключении.

6 Заключение

Статья продолжает исследования свойств и приложений дискретизованных фильтров — эффективных дискретных аппроксимаций фильтра Вонэма, обеспечивающих устойчивость численной реализации. Рассмотренная постановка включает неопределенность интенсивности шума в наблюдениях, типичную для минимаксных задач, успешно решенных для линейных систем. В отсутствие подходящих аналитических результатов анализ алгоритмов выполнен на объемном численном эксперименте.

Результаты подтвердили ожидаемую робастность оценок всех фильтров — самого фильтра Вонэма, его эмпирических устойчивых аппроксимаций и дискретизованных фильтров. При этом отсутствие устойчивости схемы Эйлера–Маруямы в постановке с неопределенностью остается такой же проблемой для фильтра Вонэма, как и в задаче с полной информацией. Преимущество дискретизованных фильтров, т. е. их численная устойчивость для любых комбинаций параметров системы наблюдения, принимает абсолютный характер в задаче с неопределенностью, когда система наблюдения сама становится неустойчивой.

Литература

1. *Kalman R. E., Bucy R. S.* New results in linear filtering and prediction theory // J. Basic Eng. — T. ASME, 1961. Vol. 83. P. 95–108. doi: 10.1115/1.3658902.
2. *Wonham W. M.* Some applications of stochastic differential equations to optimal nonlinear filtering // SIAM J. Control, 1965. Vol. 2. No. 3. P. 347–369. doi: 10.1137/030202.
3. *Липцер Р. Ш., Ширяев А. Н.* Статистика случайных процессов (нелинейная фильтрация и смежные вопросы). — М.: Наука, 2001. 696 с.
4. *D'Appolito J. A., Hutchinson C. E.* A minimax approach to the design of low sensitivity state estimators // Automatica, 1972. Vol. 8. No. 5. P. 599–608. doi: 10.1016/0005-1098(72)90031-3.
5. *Morris J.* The Kalman filter: A robust estimator for some classes of linear quadratic problems // IEEE T. Inform. Theory, 1976. Vol. 22. No. 5. P. 526–534. doi: 10.1109/TIT.1976.1055611.
6. *Vandelinde V.* Robust properties of solutions to linear-quadratic estimation and control problems // IEEE T. Automat. Contr., 1977. Vol. 22. No. 1. P. 138–139. doi: 10.1109/TAC.1977.1101433.
7. *Poor V., Looze D.* Minimax state estimation for linear stochastic systems with noise uncertainty // IEEE T. Automat. Contr., 1981. Vol. 26. No. 4. P. 902–906. doi: 10.1109/TAC.1981.1102756.
8. *Verdu S., Poor H.* Minimax linear observers and regulators for stochastic systems with uncertain second-order statistics // IEEE T. Automat. Contr., 1984. Vol. 29. No. 6. P. 499–511. doi: 10.1109/TAC.1984.1103576.
9. *Verdu S., Poor H.* On minimax robustness: A general approach and applications // IEEE T. Inform. Theory, 1984. Vol. 30. No. 2. P. 328–340. doi: 10.1109/TIT.1984.1056876.
10. *Панков А. Р.* Стратегии управления в линейной стохастической системе с негауссовскими возмущениями // Автомат. и телемех., 1994. № 6. С. 74–83.
11. *Миллер Г. Б., Панков А. Р.* Минимаксная фильтрация в линейных неопределенно-стохастических дискретно-непрерывных системах // Автомат. и телемех., 2006. № 3. С. 77–93.
12. *Миллер Г. Б., Панков А. Р.* Минимаксное управление процессом в линейной неопределенно-стохастической системе с неполными данными // Автомат. и телемех., 2007. № 11. С. 164–177.
13. *Elliott R. J., Aggoun L., Moore J. B.* Hidden Markov models: Estimation and control. — New York, NY, USA: Springer-Verlag, 1995. 396 p.
14. *Kloeden P. E., Platen E.* Numerical solution of stochastic differential equations. — Berlin: Springer, 1992. 636 p.
15. *Yin G., Zhang Q., Liu Y.* Discrete-time approximation of Wonham filters // J. Control Theory Applications, 2004. Vol. 2. P. 1–10. doi: 10.1007/s11768-004-0017-7.
16. *Борисов А. В.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям II: случай аддитивных шумов // Информатика и её применения, 2020. Т. 14. Вып. 1. С. 17–23. doi: 10.14357/19922264200103.
17. *Борисов А. В.* L1-оптимальная фильтрация марковских скачкообразных процессов II: численный ана-

лиз конкретных схем // Автомат. и телемех., 2020. № 12. С. 24–49.

18. *Борисов А. В., Босов А. В.* Практическая реализация решения задачи стабилизации линейной системы со скачкообразным случайным дрейфом по косвенным наблюдениям // Автомат. и телемех., 2022. № 9. С. 109–127.
19. *Davis M. H. A.* Linear estimation and stochastic control. — London: Chapman and Hall, 1977. 224 p.
20. *Борисов А. В.* Фильтрация состояний марковских скачкообразных процессов по дискретизованным наблюдениям // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 115–121. doi: 10.14357/19922264180316.

Поступила в редакцию 15.03.23

ROBUSTNESS INVESTIGATION OF THE NUMERICAL APPROXIMATION OF THE WONHAM FILTER

A. V. Bosov

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The properties of the optimal continuous Markov chain state filtering problem decision given be the linear observations noisy Wiener process, assuming incomplete information about its intensity, are investigated. The uncertainty of the observation system is set by the upper bound of the noise intensity. Numerical implementation of the optimal solution in the statement with complete information provided by the Wonham filter does not guarantee stability. It is shown that the Wonham filter in the statement with uncertainty is robust with respect to the noise intensity if the model parameters do not lead to its divergence. In the general case, the instability of the Euler–Maruyama numerical scheme of the Wonham filter is preserved. Simple heuristic techniques that provide stable approximations of the Wonham filter show the workability for a wider set of parameters. However, in the statement with uncertainty, it is possible to give examples when such heuristic filters show unacceptably low quality. The best solution is provided by discretized filters, approximations of the Wonham filter implemented for a specific model approximating the initial continuous observation system. A series of numerical experiments has shown that these filters have robustness for all sets of parameters. If there are no divergent trajectories among the modeled trajectories of the Wonham filter in the calculations, then the discretized filters lose a little. If there are divergent trajectories, then the gain of discretized filters is absolute.

Keywords: Markov jump process; stochastic filtering; robust estimation; Wonham filter; Euler–Maruyama numerical scheme; discretized filters

DOI: 10.14357/19922264230206 **EDN:** BGILKR

Acknowledgments

This work was supported by the Russian Science Foundation, project No. 22-28-00588. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Kalman, R. E., and R. S. Bucy. 1961. New results in linear filtering and prediction theory. *J. Basic Eng. — T. ASME* 83(1):95–108. doi: 10.1115/1.3658902.
2. Wonham, W. M. 1965. Some application of stochastic differential equations to optimal nonlinear filtering. *SIAM J. Control* 2(3):347–369. doi: 10.1137/ 030202.
3. Liptser, R. S., and A. N. Shiryaev. 2001. *Statistics of random processes II. Applications.* Berlin: Springer-Verlag. 402 p.
4. D'Appolito, J. A., and C. E. Hutchinson. 1972. A minimax approach to the design of low sensitivity state estimators. *Automatica* 8(5):599–608. doi: 10.1016/0005-1098(72)90031-3.
5. Morris, J. 1976. The Kalman filter: A robust estimator for some classes of linear quadratic problems. *IEEE T. Inform. Theory* 22(5):526–534. doi: 10.1109/TIT.1976.1055611.
6. Vandelinde, V. 1977. Robust properties of solutions to linear-quadratic estimation and control problems. *IEEE T. Automat. Contr.* 22(1):138–139. doi: 10.1109/TAC.1977.1101433.
7. Poor, V., and D. Looze. 1981. Minimax state estimation for linear stochastic systems with noise uncertainty. *IEEE T. Automat. Contr.* 26(4):902–906. doi: 10.1109/TAC.1981.1102756.

8. Verdu, S., and H. Poor. 1984. Minimax linear observers and regulators for stochastic systems with uncertain second-order statistics. *IEEE T. Automat. Contr.* 29(6):499–511. doi: 10.1109/TAC.1984.1103576.
9. Verdu, S., and H. Poor. 1984. On minimax robustness: A general approach and applications. *IEEE T. Inform. Theory* 30(2):328–340. doi: 10.1109/TIT.1984.1056876.
10. Pankov, A. R. 1994. Control strategies in a linear stochastic system with non-Gaussian perturbations. *Automat. Rem. Contr.* 55(6):832–840.
11. Miller, G. B., and A. R. Pankov. 2006. Minimax filtering in linear stochastic uncertain discrete-continuous systems. *Automat. Rem. Contr.* 67(3):413–427.
12. Miller, G. B., and A. R. Pankov. 2007. Minimax control of a process in a linear uncertain-stochastic system with incomplete data. *Automat. Rem. Contr.* 68(11):2042–2055.
13. Elliott, R. J., L. Aggoun, and J. B. Moore. 1995. *Hidden Markov models: Estimation and control*. New York, NY: Springer-Verlag. 396 p.
14. Kloden, P. E., and E. Platen. 1992. *Numerical solution of stochastic differential equations*. Berlin: Springer. 636 p.
15. Yin, G., Q. Zhang, and Y. Liu. 2004. Discrete-time approximation of Wonham filters. *J. Control Theory Applications* 2:1–10. doi: 10.1007/s11768-004-0017-7.
16. Borisov, A. V. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam II: Sluchay additivnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations II: Additive noise case]. *Informatika i ee Primeneniya — Inform. Appl.* 14(1):17–23. doi: 10.14357/19922264200103.
17. Borisov, A. V. 2020. L1-optimal filtering of Markov jump processes. II. Numerical analysis of particular realizations schemes. *Automat. Rem. Contr.* 81(12):2160–2180.
18. Borisov, A. V., and A. V. Bosov. 2022. Practical implementation of the stabilization problem solution for a linear system with discontinuous random drift by indirect observations. *Automat. Rem. Contr.* 83(9):1417–1432. doi: 10.31857/S0005231022090069.
19. Davis, M. H. A. 1977. *Linear estimation and stochastic control*. London: Chapman and Hall. 224 p.
20. Borisov, A. V. 2018. Fil'tratsiya sostoyaniy markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam [Filtering of Markov jump processes by discretized observations]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):115–121. doi: 10.14357/19922264180316.

Received March 15, 2023

Contributor

Bosov Alexey V. (b. 1969) — Doctor of Science in technology, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AVBosov@ipiran.ru

КРИТЕРИИ ВЫБОРА РАЗМЕРНОСТИ МОДЕЛИ ФАКТОРИЗАЦИИ

М. П. Кривенко[1]

Аннотация: Работа посвящена выбору размерности модели факторизации матрицы с пропущенными элементами. Задача оценивания параметров принятой модели данных решается путем многомерной оптимизации квадратичной целевой функции. Оценивание значения сниженной размерности — типичный пример задачи выбора модели, когда в ходе анализа данных возникает альтернатива, а выбор означает либо выяснение предпочтений отдельных вариантов, либо выделение «лучшего» представителя. Обычно применяемые критерии выбора основываются на функции правдоподобия, для чего требуются вероятностные предположения относительно данных. Но при оценивании параметров рассматриваемой факторной модели они не задаются, а вводить их нецелесообразно, ибо можно нарушить общность сформулированной задачи снижения размерности. Поэтому была предпринята попытка обратиться к идее использовать имеющиеся данные для целей статистического вывода повторно. Ни один из существующих подходов (бутстреп, складного ножа, перепроверки, а также перестановочные тесты) не подходит, поэтому был предложен оригинальный метод формирования новых данных путем дополнительных пропусков элементов исходной матрицы. Для обработки сформированных выборок предлагается использовать комбинацию модели смеси нормальных распределений совместно с ядерным сглаживанием. Предложенные решения позволяют корректно проводить процедуру обоснования размерности принятой модели факторизации. Изложение иллюстрируется примером обработки синтетических данных.

Ключевые слова: понижающая ранг аппроксимация матрицы; пропущенные данные; критерии выбора модели; методы повторной выборки; ядерное сглаживание

DOI: 10.14357/19922264230207 **EDN:** NQXYDC

1 Введение

Факторизация матриц данных хорошо зарекомендовала себя как метод снижения размерности в таких областях, как разведочный анализ данных, сжатие передаваемой информации, визуализация, распознавание образов и прогнозирование временны́х рядов. При этом все бóльшую востребованность стали приобретать задачи с наличием пропусков в данных. С общей характеристикой этой проблемы можно ознакомиться, например, в [1]. В [2] обращено внимание на построение алгоритмов, учитывающих в ходе оценивания параметров факторизации специфические особенности обрабатываемых матриц.

Модель факторизации $(m \times n)$-матрицы наблюдений $\mathbf{Y}$ — это представление ее в виде $\tilde{\mathbf{Y}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$, где $\mathbf{U}$ — $(m \times r)$-матрица; $\mathbf{V}$ — $(n \times r)$-матрица; r — размерность модели (Factorization Model Dimension, FMD). Наличие пропусков в данных отражается в $(m \times n)$-матрице $\mathbf{H}$:

$$h_{ij} = \begin{cases} 1\,, & y_{ij} \text{ присутствует;} \\ 0\,, & y_{ij} \text{ пропущено.} \end{cases}$$

Для суммы всех ее элементов примем обозначение p (число присутствующих элементов матрицы наблюдений $\mathbf{Y}$).

Задача построения $\tilde{\mathbf{Y}}$ формулируется как минимизация целевой функции:

$$\varphi(\mathbf{U}, \mathbf{V}) = \left\| \mathbf{H} \odot (\mathbf{Y} - \tilde{\mathbf{Y}}) \right\|_F^2 \to \min_{\mathbf{U},\mathbf{V}}.$$

Перепишем целевую функцию в более удобном для аналитических преобразований виде, исключив $\mathbf{H}$. Построчная запись матрицы $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]^{\mathrm{T}}$, где $\boldsymbol{u}_i$ суть r-векторы, даст mr-вектор $\boldsymbol{u} = [\boldsymbol{u}_1^{\mathrm{T}}, \ldots, \boldsymbol{u}_m^{\mathrm{T}}]^{\mathrm{T}}$. Также для матрицы $\mathbf{V}$ определим nr-вектор $\boldsymbol{v}$. В результате $\varphi(\mathbf{U}, \mathbf{V})$ можно переписать как

$$\varphi(\mathbf{U}, \mathbf{V}) \equiv \varphi(\boldsymbol{u}, \boldsymbol{v}) = |\mathbf{F}\boldsymbol{u} - \boldsymbol{y}|^2 = |\mathbf{G}\boldsymbol{v} - \boldsymbol{y}|^2\,,$$

где p-вектор $\boldsymbol{y}$ формируется из соответствующих элементов $\mathbf{Y}$, а $(p \times mr)$-матрица $\mathbf{F}$ — из векторов $\boldsymbol{v}_i$ и $(p \times nr)$-матрица $\mathbf{G}$ — из векторов $\boldsymbol{u}_i$ согласно только значениям $h_{ij} = 1$.

Для нахождения минимума $\hat{\varphi}(\mathbf{U}, \mathbf{V})$ целевой функции принят альтернирующий алгоритм наименьших квадратов (Alternating Least Squares, ALS). Задание начального $\boldsymbol{v}$ осуществляется с помощью методов обработки неполных данных (в данной работе — это заполнение пропусков средними значениями наблюденных значений с последующим нахождением сингулярного разложения теперь уже

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, mkrivenko@ipiran.ru

полной матрицы данных и формирование начального значения с помощью правых сингулярных векторов). Поблочная обработка возникающих в ходе итерационного процесса матриц позволяет в полной мере воспользоваться возможностями методов многократной повторной выборки [3].

Сравнение моделей для разных r осуществляется с помощью нормированной формы целевой функции $\hat{\varphi}(r) = \hat{\varphi}(\mathbf{U}, \mathbf{V})/p$, которая позволяет сопоставлять результаты оценивания параметров модели для различных значений интенсивности пропусков.

Оценивание FMD — типичный пример задачи выбора модели, когда в ходе анализа данных возникает альтернатива, а выбор означает либо выяснение предпочтений отдельных вариантов, либо выделение «лучшего» представителя. Как правило, альтернативные модели имеют вероятностный характер и включают разное число параметров, причем чем больше параметров используется, тем лучше подгонка, которой можно достичь. Критерий выбора модели (Model-Selection Criterion, MSC) должен одновременно учитывать как пригодность модели, так и число параметров, обеспечивающих ее использование. Фактически он дает ответ на вопрос: какая степень соответствия должна быть достигнута, чтобы оправдать включение дополнительных параметров в модель? Примерами близких по содержанию и достаточно проработанных задач могут служить оценивание количества элементов смеси распределений или выбор числа регрессоров.

Для выбора вероятностной модели привлекаются две основные группы методов [4].

Критерии, основанные на функции правдоподобия со штрафами. Они используют предельные свойства максимального правдоподобия и в качестве целевой рассматривают функцию вида

$$\mathrm{MSC} = \\ = \begin{bmatrix} \text{качество} \\ \text{подгонки} \end{bmatrix} + \begin{bmatrix} \text{штраф, включающий число} \\ \text{используемых параметров} \end{bmatrix}.$$

Вариантов подобных критериев множество, в частности сюда входят информационный критерий Акаике (Akaike's Information Criterion, AIC) и байесовский информационный критерий Шварца (Bayesian Information Criterion, BIC). Заметим, что как обобщение при оценивании качества подгонки можно рассматривать любую целевую функцию (например, квадратичное отклонение).

Байесовский подход к проблеме выбора модели заключается не в выборе какой-то одной модели, а в том, чтобы указать подходящее достоверное распределение для полностью исчерпывающего набора моделей, обновить эту информацию на основе данных и использовать в последующем анализе все модели с ненулевыми апостериорными вероятностями. Если подобный вывод не отвечает потребностям практики из-за своей нечеткости, можно ограничиться одной моделью с максимальной апостериорной вероятностью (Maximum *a Posteriori* Probability, MAP).

Между этими подходами существует достаточно тесная связь. Дело не только в нормальном распределении, методах максимального правдоподобия и наименьших квадратов. За счет подбора априорного распределения можно добиться, чтобы AIC или BIC, применяемые к вложенным вероятностным моделям, имели тот же эффект, что и оценка MAP, несмотря на различия в происхождении подходов [4].

Наряду с двумя представленными базовыми подходами на практике используются как их расширения (в частности, максимизация «коэффициента эффективности» — правдоподобия для каждой модели в качестве меры точности, деленного на вычислительные затраты), так и разной степени строгости эвристические приемы (в частности, последовательный перебор от простых к более сложным альтернативам вкупе с бутстреп-методами).

При любом из базовых способов выбора факторной модели требуются вероятностные предположения относительно данных, которые при оценивании параметров рассматриваемой факторной модели не задавались. Вводить их напрямую нецелесообразно, ибо можно нарушить общность сформулированной задачи снижения размерности. Поэтому объяснима попытка обратиться к процедурам, которые для целей статистического вывода используют имеющиеся данные повторно. При этом обычно рассматриваются следующие конкретные методы [5]: бутстреп, складного ножа, перепроверка, тесты перестановки.

2 Повторная выборка путем дополнительных пропусков

Содержание задачи факторизации матрицы с пропусками и желание сохранить общность ее постановки не позволяют напрямую использовать перечисленные подходы: вместо выборки имеется только единственный структурированный элемент данных — матрица, относительно вероятностных характеристик которой ничего не предполагается. В ходе экспериментов [2] было обращено внимание на то, что значения квадратичной целевой функции при близких значениях вероятности пропуска практически не отличаются, поэтому дополнительное исключение отдельных элементов $\mathbf{Y}$ по отно-

шению к уже имеющимся не должно существенно повлиять на результаты подгонки модели. Это в совокупности со сформировавшимися принципами управления обработкой данных становится основой нового метода, когда образование выборок, назовем их RSM-выборками (ReSample by Missing), осуществляется путем повторного случайного пропуска элементов исходной матрицы данных.

Реализация дополнительных пропусков должна сохранить постановку исходной задачи, т. е. удалять элементы исходной матрицы надо так, чтобы не изменились значения m и n. Простейшим алгоритмом в этом случае оказывается достижение успеха при случайном выборе элемента матрицы среди непропущенных с отбраковкой его в случае, когда удаление приводит к уменьшению значений m и n. Он не обременителен при малых значениях m и n и небольшом числе пропусков в исходной матрице. Удобно набор кандидатов на удаление сформировать заранее, а затем использовать по усмотрению.

Сложности, в первую очередь вычислительные, возрастают при росте значений m и n и степени разреженности матрицы $\mathbf{H}$. Обозначим через p_m число возможных дополнительных удалений элементов исходной матрицы, а через N_{RSM} — требуемый размер RSM-выборки. Далее необходимо рассмотреть два случая: p_m известно (посчитано заранее) или нет.

Если p_m известно, то получаем задачу случайной выборки N_{RSM} элементов из p_m имеющихся кандидатов на удаление. Соответствующий последовательный алгоритм формирования выборки известен [6]. Применительно к рассматриваемой задаче факторизации требуется напомнить, что пропуск касается только элементов матрицы, которые можно удалять.

Если p_m неизвестно, то можно посчитать его значение и тем самым свести ситуацию к уже описанной. При этом обращение к перебору элементов исходной матрицы произойдет дважды. Но [6] содержит оригинальный прием, подготавливающий при первом переборе резервуар (авторский термин), в который входят менее чем p_m элементов — возможных кандидатов в окончательную выборку. Правда, при этом на второй стадии алгоритма для того, чтобы выявить N_{RSM} нужных элементов, понадобится хоть и частично, но отсортировать резервуар.

Многовариантность базовых алгоритмов случайной выборки дополнительно пропускаемых элементов матрицы вкупе с их различными модификациями и солидный набор параметров m, n, p и N_{RSM} с богатым спектром значений оставляют возможность выбора конкретного алгоритма только с помощью экспериментов.

Для иллюстрации постановок задач и методов их решения проводились эксперименты с искусственными данными по следующей схеме: задание объемов анализируемых данных m и n, назначение сниженной размерности r_M, генерирование случайных данных с фиксированными средним и дисперсией, получение из них матрицы данных $\mathbf{Y}$ в подпространстве сниженной размерности, генерирование матрицы $\mathbf{H}$ для определенной вероятности p_m пропуска, применение того или иного метода анализа данных. Для определенности рассматривалось ограничение $n < m$. В качестве критерия завершенности итерационного процесса использовался контроль относительного изменения значений целевой функции на последовательных шагах итераций $\mathrm{Tol} = 0{,}01$ вместе с ограничением числа шагов итераций $t_{\max} = 100$.

Построив RSM-выборки из элементов $\boldsymbol{x}_i$ размерности d, можно оценивать соответствующее распределение в виде плотности $f(\boldsymbol{t})$ и обращаться к классическим вероятностным методам выбора модели. При этом доступны параметрические и непараметрические подходы с оговоркой, что общность привлекаемых моделей не должна искажать постановку исходной задачи. Последнее в первую очередь касается параметрической точки зрения, вследствие чего внимание было обращено на описание данных с помощью смеси нормальных распределений. Она позволяет обеспечивать желаемую точность аппроксимации благодаря наращиванию числа элементов смеси, сохраняя при этом в силе преимущество относительно простых аналитических решений типовых вероятностных задач.

Использование смеси многомерных нормальных распределений с плотностью $f(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ основывается на представлении

$$f(\boldsymbol{t}) = \sum_{l=1}^{k} w_l f(\boldsymbol{t}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l).$$

Вместе с оценками для параметров смеси $\hat{w}_l$, $\hat{\boldsymbol{\mu}}_l$ и $\hat{\boldsymbol{\Sigma}}_l$, $l = 1, \ldots, k$, становится известна матрица апостериорных вероятностей принадлежности i-го наблюденного значения к j-му элементу смеси, что, в свою очередь, позволяет стратифицировать исходный набор данных и получить k подвыборок. Для каждой l-й из них объемом n_l можно построить ядерную оценку плотности в следующих предположениях [7]:

- распределение данных — $N(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l)$;
- $\hat{\hat{f}}(\boldsymbol{t}, \mathbf{H}) = (1/n_l) \sum_{i=1}^{n_l} K_{\mathbf{H}}(\boldsymbol{t} - \hat{\boldsymbol{x}}_i)$ при $K_{\mathbf{H}}(\boldsymbol{t}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{t})$ и $K(\boldsymbol{t}) = (2\pi)^{-d/2} \exp(-(1/2)\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})$, $\mathbf{H} = \mathbf{H}_{\mathrm{AMISE}} = (4/(d+2))^{2/(d+4)} \hat{\boldsymbol{\Sigma}}_l n_l^{-2/(d+4)}$.

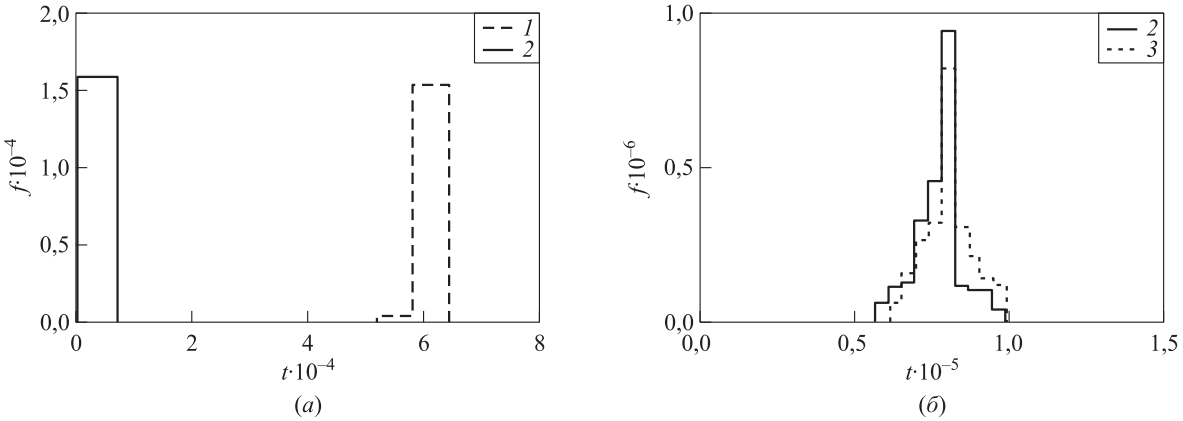


Рис. 1 Попарное сравнение гистограмм для $r = 3$ (*1*) и 4 (*2*) (*а*) и $r = 4$ (*2*) и 5 (*3*) (*б*)

Здесь AMISE — Asymptotic Mean Integrated Squared error и речь идет об асимптотически оптимальной сглаживающей матрице $\mathbf{H}_{\mathrm{AMISE}}$. В принятых обозначениях

$$f\left(\boldsymbol{t}|\boldsymbol{x}_i, \mathbf{H}_i\right) = K_{\boldsymbol{H}}(\boldsymbol{t} - \boldsymbol{x}_i).$$

Ключевым в выборе оптимального параметра сглаживания $\mathbf{H}$ становится предположение о виде оцениваемой плотности распределения, что, на первый взгляд, лишено всякого смысла: оцениваем то, что знаем. Но это не так, и причин здесь несколько: предположение о конкретном виде распределения (в данном случае речь идет о смеси нормальных) есть способ, пусть грубой, аппроксимации данных, которая, в свою очередь, позволяет относительно просто найти для параметра окна сглаживания некоторое начальное приближение для его последующего уточнения. Кроме того, полученные оценки можно использовать для формирования представления о свойствах непараметрических оценок в наилучшем случае, когда известна оцениваемая плотность (например, для иллюстрации эффекта «проклятия размерности»).

После предварительных общих замечаний перейдем к предлагаемым критериям выбора FMD.

Критерий попарного сравнения эмпирических распределений. Для каждого значения r можно получить значение $\hat{\varphi}(r)$ нормированной целевой функции и построить RSM-выборку значений $\hat{\varphi}^{\mathrm{RSM}}(r)$, которая позволяет получить эмпирическую плотность распределения $\hat{f}(t; r)$ (далее речь идет о гистограммной оценке) и представление о значимости отличий значений $\hat{\varphi}(r)$ для разных r. Если рассмотреть пару RSM-распределений $\hat{\varphi}(r)$ и $\hat{\varphi}(r+1)$, то видно, что они могут заметно различаться или нет в зависимости от того, существенно или несущественно уменьшились значения целевой функции. Такое парное сравнение можно сделать шагом последовательной процедуры от меньшего значения r к большему.

Формализовать меру различия двух распределений можно, рассмотрев простейший однопороговый критерий значимости, статистику которого будем оценивать с помощью двух RSM-выборок из условия минимизации ошибки классификации $\mathrm{Err}_{\min}$ принятого критерия. Далее, если получившаяся оценка ошибки классификации мала, то будем считать целесообразным переход от случая r к случаю $r+1$, если же нет (около 0,5), то процесс усложнения модели факторизации прекратим и в качестве оценки структурного параметра факторизации примем найденное r (не что иное, как реализация принципа «бритвы Оккама»).

Для иллюстрации рассмотрим случай $m = 50$, $n = 10$, $r_M = 4$, $p_m = 40\%$ и сравним пары одномерных гистограмм для $r = 3$ и 4 (рис. 1). Если для $r = 3$ видно, что усложнение модели до $r = 4$ оправдано (плотности распределения RSM-значений отчетливо отделимы друг от друга и $\mathrm{Err}_{\min} = 0\%$), то для $r = 4$ в нем нет необходимости ($\mathrm{Err}_{\min} = 50\%$).

Подобную процедуру можно рекомендовать только в качестве разведочного анализа данных по ряду причин:

- внешне парное сравнение очень схоже с условиями и приемами двухвыборочных критериев значимости, но на поверку последние демонстрируют неприемлемые результаты; это связано с тем, что нулевая гипотеза должна формулироваться в более общем виде, чем совпадение распределений; как результат — для сравнения приходится использовать упрощенные критерии значимости;
- требуется уточнить правило задания уровня значимости на каждом шаге, если принимается как условие задачи итоговый уровень значимости всей последовательной процедуры;

– неизвестные распределения $\hat{\varphi}(r)$, скорее всего, зависят от r, но это не принимается во внимание в описанной схеме принятия решения;

– полностью не учитывается многомерность статистики критерия выбора FMD (многократно сравнивается лишь пара «соседних» одномерных распределений).

Но перечисленные недостатки компенсируются простотой использования попарных сравнений, в особенности при больших размерах исходных матриц данных. Например, при исследовании возможности прогнозирования химического состава мочевых камней у пациентов с уролитиазом [2] без особых затрат можно показать, что желание снизить количество анализируемых метаболических показателей мочи и сыворотки крови только на основе матрицы исходных данных несостоятельно.

Критерий максимального правдоподобия позволяет полнее описать альтернативу и учесть реальную размерность при построении эмпирических распределений.

Если при статистическом выводе удастся предусмотреть фактическую размерность r_M, то это найдет отражение в обозначениях как $\varphi(r|r_M)$. Для того чтобы учесть многомерность статистики критерия выбора FMD, надо перейти к анализу $r_{\max}$-мерного набора

$$(\hat{\varphi}(1|r_M), \hat{\varphi}(2|r_M), \ldots, \hat{\varphi}(r_{\max}|r_M)) = \hat{\boldsymbol{\varphi}}(r_M),$$

где $r_{\max}$ — заранее выбранная верхняя граница рассматриваемых возможных значений размерности. Тогда станет возможным сравнивать различные варианты выбора FMD, рассчитывая при этом на эффективность предлагаемых решений. Зависимость статистического вывода от фактической размерности FMD реализуется следующим образом. Для каждого значения r_M могут быть получены оценки параметров модели $\hat{\mathbf{U}}$ и $\hat{\mathbf{V}}$, которые приводят к аппроксимации исходных данных $\hat{\mathbf{Y}} = \hat{\mathbf{U}}\hat{\mathbf{V}}^{\mathrm{T}}$ и к возможности для нее построить RSM-выборку $\hat{\boldsymbol{\varphi}}^{\mathrm{RSM}}(r_M)$. В результате эмпирическая плотность распределения, построенная на основе $\hat{\boldsymbol{\varphi}}^{\mathrm{RSM}}(r_M)$, даст оценку функции правдоподобия в точке $\hat{\varphi}(r_M)$ для принятого значения r_M.

Для рассматриваемого иллюстративного примера результативность предлагаемой схемы обработки исходных данных (рис. 2) несомненна:

– ярко выраженный экстремум функции правдоподобия точно указывает на истинное значение r;

– более тщательная аппроксимация распределений RSM-выборок, когда использование смеси из двух нормальных распределений (случай NM) дополняется ядерным сглаживанием (случай NM&K), выглядит убедительней.

Дополнительные многократные эксперименты в рамках иллюстративного примера показывают, что скудость априорной информации, относительно небольшие объемы исходных данных, а также лишь наброски идей о том, как справляться со сложностью возникающих проблем, приводят к понятному результату: в среднем оценки реального значения r оказываются несколько завышенными, но не тривиальными.

Кроме использования зависимости $\hat{\varphi}(r|r_M)$ от r_M, можно попробовать привлечь дополнительную информацию об априорном распределении размерности R_M. Но чаще всего ее нет, и чтобы както снять проблему роста набора параметров задачи факторизации, можно ввести упрощенную схему задания нужных вероятностей, используя лишь единственный параметр. Например, для этого принимается предположение, что с ростом r_M вероятности уменьшаются (т. е. более сложная модель оказывается менее вероятной) в соответствии с геометрическим распределением:

$$\Pr\{R_M = r\} = p_1(1-p_1)^{r-1}, \;\; r = 1, 2, \ldots,$$

где p_1 — вероятность выбора одномерной модели. В рассматриваемой задаче факторизации $r \le n$. Поэтому формально, если говорить о R_M и за основу брать геометрическое распределение, соответствующие вероятности надо нормировать, но для байесовского вывода это непринципиально. Когда p_1 приближается к 0, распределение R_M начинает походить на равномерное, что фактически приводит к оценке максимального правдоподобия. С ростом p_1 акцент на малых значениях

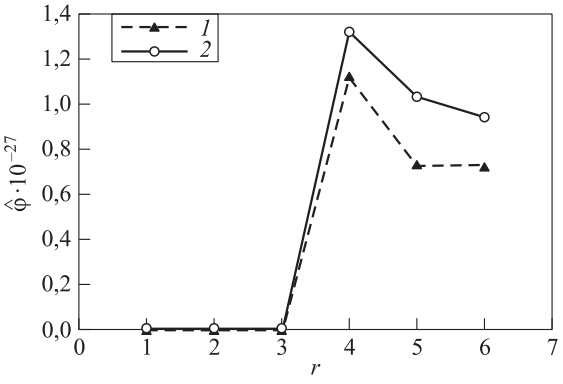


Рис. 2 Правдоподобие как функция от размерности r: *1* — случай NM, когда распределение RSM-выборки аппроксимируется смесью из двух нормальных распределений; *2* — случай NM&K, когда указанная аппроксимация дополняется ядерным сглаживанием

размерности становится отчетливей. Таким образом, параметр геометрического распределения дает полный контроль над скоростью уменьшения вероятности последовательных моделей.

3 Заключение

Сжатие матрицы данных с пропусками (например, в фиксируемом изображении потеряны его фрагменты, в ходе клинического обследования часть результатов лабораторного исследования отсутствует и т. п.) обычно осуществляется на предварительном этапе обработки информации при отсутствии обоснованных предположений о вероятностных характеристиках сведений. В связи с этим возникает вопрос: можно ли и как более-менее формально определиться со значением сниженной размерности? Если работ по оцениванию параметров факторной модели (матрицы $\mathbf{U}$ и $\mathbf{V}$) достаточно много, то по выбору числа факторов (значения сниженной размерности) нет вообще. Кроме того, в имеющихся публикациях присутствовала некоторая несогласованность деталей модели данных и методов их обработки. Поэтому пришлось обращаться к истокам темы и в [2] удалось определиться с моделью и методами оценивания ее параметров. Но стала явной проблема объемных по памяти и времени вычислений. Однократно найти оценку параметров модели возможно, но проанализировать ее с помощью многократного пересчета оказывается уже нереальным. Пришлось заниматься методами матричных вычислений: полученные в [3] результаты относительно блочно-диагонального представления матриц и правомерности операций с ними в таком представлении привели к реальным вычислительным преимуществам поблочной обработки. Кроме того, открылись новые горизонты, например возможность использования поблочной версии сингулярного разложения матриц в чистом виде без перестановочных матриц, а также привлечения иных способов факторизации матриц.

Уточнение модели факторизации матрицы данных с пропусками, выбор устойчивого варианта алгоритма оценивания параметров модели, построение его эффективной реализации позволили в данной статье предложить критерии выбора адекватной размерности модели факторизации. Ключевым моментом стала оригинальная идея получения повторных выборок путем дополнительных пропусков элементов данных. Таким образом, не вводя дополнительных ограничений на исходные данные, удалось привлечь фактор случайности и методы обработки эмпирических распределений.

При полном отсутствии знаний о вероятностных свойствах генерируемых данных продемонстрирована действенность аппроксимации распределений с помощью смеси многомерных распределений с последующим уточнением посредством ядерного сглаживания. В данной работе построение и использование модели данных RSM-выборки преследовало утилитарную цель — обосновать жизнеспособность оригинального метода формирования новых данных. При этом схема комбинированного описания выборочных распределений через смеси распределений и ядерные оценки оставляет возможность для совершенствования путем рассмотрения других критериев эффективности оценок и иных ядер, посредством учета вычислительных аспектов.

Литература

1. *Chen P.* Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix // Int. J. Comput. Vision, 2008. Vol. 80. Iss. 1. P. 125–142. doi: 10.1007/s11263-008-0135-7.
2. *Кривенко М. П.* Выбор модели при факторизации матрицы данных с пропусками // Информатика и её применения, 2022. Т. 16. Вып. 3. С. 52–58. doi: 10.14357/19922264220307.
3. *Кривенко М. П.* Эффективные вычисления при факторизации матричных данных с пропусками // Системы и средства информатики, 2023. Т. 33. Вып. 1. С. 78–89. doi: 10.14357/08696527230108.
4. *Poland W. B., Shachter R. D.* Three approaches to probability model selection // Uncertainty in artificial intelligence. — Seattle, WA, USA: Morgan Kaufmann, 1994. P. 478–483. doi: 10.1016/B978-1-55860-332-5.50065-1.
5. *Chernick M. R.* Resampling methods // WIREs Data Min. Knowl., 2012. Vol. 2. Iss. 3. P. 255–262. doi: 10.1002/widm.1054.
6. *Fan C. T., Muller M. E., Rezucha I.* Development of sampling plans by using sequential (item by item) selection techniques and digital computers // J. Am. Stat. Assoc., 1962. Vol. 57. No. 298. P. 387–402. doi: 10.1080/01621459.1962.10480667.
7. *Wand M. P.* Error analysis for general multivariate kernel estimators // J. Nonparametr. Stat., 1992. Vol. 2. Iss. 1. P. 1–15. doi: 10.1080/10485259208832538.

Поступила в редакцию 27.01.23

CRITERIA FOR CHOOSING THE FACTORIZATION MODEL DIMENSIONALITY

M. P. Krivenko

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper is devoted to the choice of model dimension of matrix factorization in the presence of missing elements. The problem of estimating the parameters of the adopted data model is solved by multidimensional optimization. Estimating the value of reduced dimensionality is a typical example of the problem of choosing a model when an alternative arises during data analysis and the choice means either finding out the preferences of individual options or highlighting the "best" representative. Typically, applied selection criteria are based on likelihood function which requires probabilistic assumptions about the data. But when evaluating the parameters of the factor model under consideration, they are not set and it is impractical to introduce them, so as not to violate the commonality of the formulated task of reducing dimensionality. Therefore, an attempt was made to turn to the idea of reusing the available data for the statistical output. None of the existing approaches (bootstrap, folding knife, rechecks, as well as permutation tests) is suitable; so, an original method for generating new data by additional omissions of elements of the original matrix was proposed. To process the formed samples, it is suggested to use a combination of the model of a mixture of normal distributions in conjunction with nuclear smoothing. The proposed solutions make it possible to correctly carry out the procedure for justifying the dimensionality of the adopted factorization model. The exposition is illustrated by an example of synthetic data processing.

Keywords: lower rank matrix approximation; missing data; criteria for model selection; resampling methods; kernel smoothing

DOI: 10.14357/19922264230207 **EDN:** NQXYDC

References

1. Chen, P. 2008. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *Int. J. Comput. Vision* 80(1):125–142. doi: 10.1007/s11263-008-0135-7.
2. Krivenko, M. P. 2022. Vybor modeli pri faktorizatsii matritsy dannykh s propuskami [Model selection for matrix factorization with missing components]. *Informatika i ee Primeneniya — Inform. Appl.* 16(3):52–58. doi: 10.14357/19922264220307.
3. Krivenko, M. P. 2023. Effektivnye vychisleniya pri faktorizatsii matrichnykh dannykh s propuskami [Efficient computations in a matrix factorization with missing components]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 33(1):78–89. doi: 10.14357/08696527230108.
4. Poland, W. B., and R. D. Shachter. 1994. Three approaches to probability model selection. *Uncertainty in artificial intelligence*. Seattle, WA: Morgan Kaufmann. 478–483. doi: 10.1016/B978-1-55860-332-5.50065-1.
5. Chernick, M. R. 2012. Resampling methods. *WIREs Data Min. Knowl.* 2(3):255–262. doi: 10.1002/widm.1054.
6. Fan, C. T., M. E. Muller, and I. Rezucha. 1962. Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *J. Am. Stat. Assoc.* 57(298):387–402. doi: 10.1080/01621459.1962.10480667.
7. Wand, M. P. 1992. Error analysis for general multivariate kernel estimators. *J. Nonparametr. Stat.* 2(1):1–15. doi: 10.1080/10485259208832538.

Received January 27, 2023

Contributor

Krivenko Michail P. (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mkrivenko@ipiran.ru

ИССЛЕДОВАНИЕ СИСТЕМ ОБСЛУЖИВАНИЯ СО СМЕШАННЫМИ ПРИОРИТЕТАМИ*

А. К. Берговин[1], В. Г. Ушаков[2]

Аннотация: Изучена однолинейная система массового обслуживания с бесконечным числом мест для ожидания, произвольным распределением времени обслуживания и пуассоновскими входящими потоками требований. Рассматриваются две модели смешанных приоритетов. В первой модели между требованиями разных потоков действует либо дисциплина абсолютного приоритета с обслуживанием заново прерванного требования, либо дисциплина относительного приоритета. Во второй модели — дисциплина абсолютного приоритета либо с потерей, либо с обслуживанием заново прерванного требования. Методом дополнительных компонент исследуется многомерный случайный процесс, компоненты которого — число требований каждого приоритета в системе и время, прошедшее с начала обслуживания требования, находящегося на приборе в момент времени t. Найдено распределение указанного процесса в нестационарном режиме работы системы.

Ключевые слова: относительный приоритет; абсолютный приоритет с обслуживанием заново; абсолютный приоритет с потерей; одноканальная система; длина очереди

DOI: 10.14357/19922264230208 **EDN:** JULPWS

1 Введение

Системы обслуживания с несколькими типами требований и различными приоритетными дисциплинами часто рассматриваются в качестве математических моделей различных инфокоммуникационных систем. Применение приоритетных дисциплин позволяет как учесть различную степень важности требований разных типов, так и обеспечить наиболее эффективную работу системы. К настоящему времени достаточно полно развита теория однолинейных систем обслуживания с приоритетами [1–3]. Практически во всех работах предполагается, что взаимоотношение требований разных типов регулируется одной и той же приоритетной дисциплиной. Возможность выбора разновидности приоритетного правила для каждой пары потоков позволяет, с одной стороны, адекватно описать работу значительно более широкого класса реальных систем и, с другой стороны, организовать более эффективную работу системы в целом без значительного ухудшения качества обслуживания наиболее приоритетных требований.

2 Обозначения и определения

Рассматривается система обслуживания типа $M_r|G_r|1|\infty$, в которую поступают $r \geqslant 2$ пуассоновских потоков требований с интенсивностями $a_1, \ldots, a_r$. Длительности обслуживания — независимые в совокупности случайные величины с функциями распределения $B_1(x), \ldots, B_r(x)$ и плотностями распределения $b_1(x), \ldots, b_r(x)$. Требования i-го потока (приоритета i) имеют приоритет перед требованиями j-го потока при $i < j$. Рассматриваются две модели:

(1) приоритет может быть или относительным, или абсолютным с обслуживанием заново прерванного требования;

(2) приоритет является абсолютным с потерей или обслуживанием заново прерванного требования.

Положим в первой модели I_i и J_i — соответственно множества номеров потоков, которые имеют абсолютный приоритет и перед которыми имеют абсолютный приоритет требования i-го потока ($I_1 = J_r = \emptyset$), а K_i и M_i во второй модели — соответственно множества номеров потоков, которые имеют абсолютный приоритет с потерей и перед которыми имеют абсолютный приоритет с потерей требования i-го потока.

Пусть $L_i(t)$ — число требований i-го потока в системе в момент времени t; $i(t)$ и $x(t)$ — соответственно номер потока, требование из которого

*Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации, грант № 075-15-2019-1621.

[1]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова, alexey.bergovin@gmail.com

[2]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, vgushakov@mail.ru

обслуживается в момент времени t, и время, прошедшее с начала его обслуживания (если в момент t система свободна, то $i(t)$ и $x(t)$ можно доопределить произвольным образом, например положить $i(t) = x(t) = 0$), $\mathbf{L}(t) = (L_1(t), \dots, L_r(t))$;

$$\beta_i(s) = \int_0^\infty e^{-sx} dB(x); \quad \beta_{ij} = \int_0^\infty x^j dB_i(x);$$

$$\eta_i(x) = \frac{b_i(x)}{1 - B_i(x)};$$

$$\sigma_k = a_1 + \dots + a_k,\ k = 1, \dots, r,$$

$$\sigma_0 = 0,\ \sigma = \sigma_r,\ \mathbf{0} = (0, \dots, 0);$$

$\mathbf{1}_i = (0, \dots, 0, 1, 0, \dots, 0)$, где 1 стоит на i-м месте;

$$\mathsf{P}(\mathbf{n}, t) = \mathbf{P}(\mathbf{L}(t) = \mathbf{n});$$

$$p(\mathbf{z}, s) = \int_0^\infty e^{-st} \sum_{n_1=0}^\infty \cdots \sum_{n_r=0}^\infty z_1^{n_1} \cdots z_r^{n_r} \mathsf{P}(\mathbf{n}, t)\, dt;$$

$$\mathsf{P}_i(\mathbf{n}, x, t) = \frac{\partial}{\partial x} \mathbf{P}(\mathbf{L}(t) = \mathbf{n}, i(t) = i, x(t) < x),$$

$$\mathbf{n} = (n_1, \dots, n_r);$$

$$p_i(\mathbf{z}, x, s) = \int_0^\infty e^{-st} \sum_{n_1=0}^\infty \cdots \sum_{n_r=0}^\infty z_1^{n_1} \cdots z_r^{n_r} \mathsf{P}_i(\mathbf{n}, x, t)\, dt,$$

$$\mathbf{z} = (z_1, \dots, z_r);$$

$$d_i(\mathbf{z}, s) = 1 - z_i^{-1} \beta_i \left(s + \sigma - \sum_{j \notin I_i} a_j z_j \right) -$$

$$- \sum_{j \in I_i} a_j z_j \frac{1 - \beta_i \left(s + \sigma - \sum_{j \notin I_i} a_j z_j \right)}{s + \sigma - \sum_{j \notin I_i} a_j z_j};$$

$$c_i(\mathbf{z}, s) = 1 - z_i^{-1} \beta_i \left(s + \sigma - \sum_{j=i}^r a_j z_j \right) -$$

$$- \left(\sum_{j \in K_i} a_j z_j z_i^{-1} + \sum_{j=1, j \notin K_i}^{i-1} a_j z_j \right) \times$$

$$\times \frac{1 - \beta_i \left(s + \sigma - \sum_{j=i}^r a_j z_j \right)}{s + \sigma - \sum_{j=i}^r a_j z_j}.$$

Будем предполагать, что в начальный момент времени $t = 0$ система свободна от требований.

3 Распределение длины очереди

В первых двух теоремах содержатся результаты для первой модели, а в 3-й и 4-й — для второй.

Теорема 1. *При каждом $k = 1, \dots, r$ система уравнений*

$$d_i(\mathbf{z}, s) = 0,\ i = 1, \dots, k,$$

имеет единственное решение $z_i = \pi_{ik}(z_{k+1}, \dots, z_r, s)$, аналитическое в области $|z_{k+1}| < 1, \dots, |z_r| < 1$, $\operatorname{Re} s > 0$, в которой $|\pi_{ik}(z_{k+1}, \dots, z_r, s)| < 1$, $i = 1, \dots, k$, $k = 1, \dots, r$.

Доказательство теоремы аналогично доказательству леммы 3 в [2, с. 122].

Теорема 2. *Функция $p(\mathbf{z}, s)$ определяется по формуле*:

$$p(\mathbf{z}, s) = p_0(s) +$$

$$+ \sum_{i=1}^r \frac{1 - \beta_i \left(s + \sigma - \sum_{j \notin I_i} a_j z_j \right)}{s + \sigma - \sum_{j \notin I_i} a_j z_j} p_i(\mathbf{z}, 0, s), \quad (1)$$

где

$$p_0(s) = \left(s + \sigma - \sum_{j=1}^r a_j \pi_{jr}(s) \right)^{-1}, \quad (2)$$

а $p_i(\mathbf{z}, 0, s)$ определяются из рекуррентных соотношений

$$\sum_{i=k+1}^r d_i \left(\pi_{1k}(z_{k+1}, \dots, z_r, s), \dots \right.$$

$$\left. \dots, \pi_{kk}(z_{k+1}, \dots, z_r, s), z_{k+1}, \dots, z_r, s \right) p_i(\mathbf{z}, 0, s) =$$

$$= 1 - \left(s + \sigma - \sum_{j=1}^k a_j \pi_{jk}(z_{k+1}, \dots, z_r, s) - \right.$$

$$\left. - \sum_{j=k+1}^r a_j z_j \right) p_0(s),\ \ k = 0, \dots, r-1. \quad (3)$$

Доказательство. Рассматривая изменения состояний процесса $(\mathbf{L}(t), x(t), i(t))$ в интервале времени $(t, t + \Delta)$ и устремляя $\Delta \to 0$, имеем

$$\frac{\partial \mathsf{P}_i(\mathbf{n}, x, t)}{\partial t} + \frac{\partial \mathsf{P}_i(\mathbf{n}, x, t)}{\partial x} = -(\sigma + \eta_i(x)) \mathsf{P}_i(\mathbf{n}, x, t) +$$

$$+ \sum_{j \notin I_i,\ j \neq i} \left(1 - \delta_{n_j, 0}\right) a_j \mathsf{P}_i(\mathbf{n} - \mathbf{1}_j, x, t) +$$

$$+ (1 - \delta_{n_i, 1})\, a_i \mathsf{P}_i(\mathbf{n} - \mathbf{1}_i, x, t). \quad (4)$$

Переходя в (4) к производящим функциям и преобразованиям Лапласа по t, получаем

$$\frac{\partial p_i(\mathbf{z}, x, s)}{\partial x} = -(s + \sigma + \eta_i(x)) p_i(\mathbf{z}, x, s) +$$

$$+ \sum_{j \notin I_i} a_j z_j p_i(\mathbf{z}, x, s).$$

Для вероятности свободного состояния $\mathsf{P}_0(t)$ имеем

$$\frac{\partial \mathsf{P}_0(t)}{\partial t} = -\sigma \mathsf{P}_0(t) + \sum_{j=1}^{r} \int_0^\infty \mathsf{P}_j(\mathbf{1}_j, x, t)\eta_j(x)\,dx$$

и

$$(s+\sigma)p_0(s) - 1 = \\ = \int_0^\infty e^{-st} \sum_{j=1}^{r} \int_0^\infty \mathsf{P}_j(\mathbf{1}_j, x, t)\eta_j(x)\,dx\,dt. \quad (5)$$

Найдем теперь краевые условия в точке $x = 0$. Имеем:

$$\mathsf{P}_i(\mathbf{n}, 0, t) = 0, \text{ если } n_1 + \cdots + n_{i-1} \neq 0 \text{ или } n_i = 0\,.$$

Для остальных $n_1, \ldots, n_r$ справедливы равенства:

$$\mathsf{P}_i(\mathbf{n}, 0, t) = \sum_{j=1, j\notin J_i}^{r} \int_0^\infty \mathsf{P}_j(\mathbf{n} + \mathbf{1}_j, x, t)\eta_j(x)\,dx + \\ + \delta_{n_i,1} \sum_{j\in J_i} \int_0^\infty \mathsf{P}_j(\mathbf{n} - \mathbf{1}_i, x, t)\,dx\,a_i + \\ + \delta_{n_i,1} \prod_{j\neq i} \delta_{n_j,0} a_i \mathsf{P}_0(t). \quad (6)$$

Из (5) и (6) получаем

$$\sum_{i=1}^{r} p_i(\mathbf{z}, 0, s) = \sum_{i=1}^{r} \int_0^\infty p_i(\mathbf{z}, x, s)\eta_i(x)\,dx + \\ + \sum_{i=1}^{r} \left(\sum_{j\in I_i} a_j z_j \right) \int_0^\infty p_i(\mathbf{z}, x, s)\,dx + 1 - \\ - \left(s + \sigma - \sum_{j=1}^{r} a_j z_j \right) p_0(s). \quad (7)$$

Решение системы уравнений (5) имеет вид:

$$p_i(\mathbf{z}, x, s) = \\ = (1 - B_i(x)) e^{-\left(s+\sigma-\sum_{j\notin I_i} a_j z_j\right)x} p_i(\mathbf{z}, 0, s).$$

Подставляя его в (7), имеем

$$\sum_{i=1}^{r} d_i(\mathbf{z}, s) p_i(\mathbf{z}, 0, s) = 1 - \left(s + \sigma - \sum_{j=1}^{r} a_j z_j \right) p_0(s).$$

Так как $\mathsf{P}_i(\mathbf{n}, 0, t) = 0$ при $n_1 + \cdots + n_{i-1} \neq 0$, то $p_i(\mathbf{z}, 0, s)$ не зависит от $z_1, \ldots, z_{i-1}$. Отсюда и из теоремы 1 следуют (2) и (3), а из равенства

$$p(\mathbf{z}, s) = p_0(s) + \sum_{i=1}^{r} \int_0^\infty p_i(\mathbf{z}, x, s)\,dx$$

получаем (1).

Теорема 3. *При каждом $k = 1, \ldots, r$ система уравнений*

$$c_i(\mathbf{z}, s) = 0, \;\; i = 1, \ldots, k,$$

имеет единственное решение $z_i = \tau_{ik}(z_{k+1}, \ldots, z_r, s)$, аналитическое в области $|z_{k+1}| < 1, \ldots, |z_r| < 1$, $\operatorname{Re} s > 0$, в которой $|\tau_{ik}(z_{k+1}, \ldots, z_r, s)| < 1$, $i = 1, \ldots, k$, $k = 1, \ldots, r$.

Теорема 4. *Функция $p(\mathbf{z}, s)$ определяется по формуле*

$$p(\mathbf{z}, s) = p_0(s) + \\ + \sum_{i=1}^{r} \frac{1 - \beta_i \left(s + \sigma - \sum_{j=i}^{r} a_j z_j \right)}{s + \sigma - \sum_{j=i}^{r} a_j z_j} p_i(\mathbf{z}, 0, s),$$

где

$$p_0(s) = \left(s + \sigma - \sum_{j=1}^{r} a_j \tau_{jr}(s) \right)^{-1},$$

а $p_i(\mathbf{z}, 0, s)$ определяются из рекуррентных соотношений

$$\sum_{i=k+1}^{r} c_i \left(\tau_{1k}(z_{k+1}, \ldots, z_r, s), \ldots \right. \\ \left. \ldots, \tau_{kk}(z_{k+1}, \ldots, z_r, s), z_{k+1}, \ldots, z_r, s \right) p_i(\mathbf{z}, 0, s) = \\ = 1 - \left(s + \sigma - \sum_{j=1}^{k} a_j \tau_{jk}(z_{k+1}, \ldots, z_r, s) - \right. \\ \left. - \sum_{j=k+1}^{r} a_j z_j \right) p_0(s), \; k = 0, \ldots, r-1.$$

Доказательство. Для второй модели функции $\mathsf{P}_i(\mathbf{n}, x, t)$ удовлетворяют системе дифференциальных уравнений

$$\frac{\partial \mathsf{P}_i(\mathbf{n}, x, t)}{\partial t} + \frac{\partial \mathsf{P}_i(\mathbf{n}, x, t)}{\partial x} = -(\sigma + \eta_i(x))\mathsf{P}_i(\mathbf{n}, x, t) + \\ + \sum_{j=i+1}^{r} \left(1 - \delta_{n_j,0}\right) a_j \mathsf{P}_i(\mathbf{n} - \mathbf{1}_j, x, t) + \\ + \left(1 - \delta_{n_i,1}\right) a_i \mathsf{P}_i\left(\mathbf{n} - \mathbf{1}_i, x, t\right).$$

Отсюда

$$\frac{\partial p_i(\mathbf{z}, x, s)}{\partial x} = \\ = -\left(s + \sigma - \sum_{j=i}^{r} a_j z_j + \eta_i(x) \right) p_i(\mathbf{z}, x, s). \quad (8)$$

Для вероятности свободного состояния $\mathsf{P}_0(t)$ и ее преобразования Лапласа $p_0(s)$ справедливы те же соотношения, что и для модели 1. При $x = 0$ имеем:

$$\mathsf{P}_i(\mathbf{n}, 0, t) = 0, \text{ если } n_1 + \cdots + n_{i-1} \neq 0 \text{ или } n_i = 0.$$

Для остальных $n_1, \ldots, n_r$ справедливы равенства:

$$\mathsf{P}_i(\mathbf{n}, 0, t) = \sum_{j=1}^{i} \int_0^\infty \mathsf{P}_j(\mathbf{n} + \mathbf{1}_j, x, t)\eta_j(x)\,dx + \\ + \delta_{n_i,1} \prod_{j \neq i} \delta_{n_j,0} a_i \mathsf{P}_0(t) + \\ + \delta_{n_i,1} a_i \left(\sum_{j=i+1, j \in M_i}^{r} \int_0^\infty \mathsf{P}_j(\mathbf{n} + \mathbf{1}_j - \mathbf{1}_i, x, t)\,dx + \right. \\ \left. + \sum_{j=i+1, j \notin M_i}^{r} \int_0^\infty \mathsf{P}_j(\mathbf{n} - \mathbf{1}_i, x, t)\,dx \right). \quad (9)$$

Из (8) и (9) находим

$$p_i(\mathbf{z}, x, s) = (1 - B_i(x)) e^{-\left(s + \sigma - \sum_{j=i}^{r} a_j z_j\right) x} p_i(\mathbf{z}, 0, s);$$

$$\sum_{i=1}^{r} c_i(\mathbf{z}, s) p_i(\mathbf{z}, 0, s) = 1 - \left(s + \sigma - \sum_{j=1}^{r} a_j z_j \right) p_0(s).$$

Окончание доказательства повторяет рассуждения при доказательстве теоремы 2.

Литература

1. *Jaiswal N. K.* Priority queues. — New York; London: Academic press, 1968. 240 p.
2. *Матвеев В. Ф., Ушаков В. Г.* Системы массового обслуживания. — М.: Изд-во Московского ун-та, 1984. 240 с.
3. *Takagi H.* Queueing analysis: A foundation of performance evaluation. — Amsterdam: North-Holland Elsevier, 1991. Vol. 1. Part 1. 487 p.

Поступила в редакцию 28.03.22

ANALYSIS OF THE QUEUEING SYSTEMS WITH MIXED PRIORITIES

A. K. Bergovin[1] and V. G. Ushakov[1,2]

[1]M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2]Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: A one-line queuing system with an infinite number of waiting places, an arbitrary distribution of service time, and Poisson incoming flows of customers is studied. Two models of mixed priorities are considered. In the first model, there is either preemptive repeat priority discipline between the customers of different flows or a head of the line priority discipline. In the second model — preemptive priority discipline either with loss or with repeat of a newly interrupted customer. By the method of additional components, a multidimensional random process is investigated, the components of which are the number of customers of each priority in the system and the time elapsed since the start of servicing the customer located on the device at time t. The distribution of the specified process in the nonstationary mode of the system is found.

Keywords: head of the line; preemptive repeat; preemptive loss; one-line; queue length

DOI: 10.14357/19922264230208 **EDN:** JULPWS

Acknowledgments

The research was supported by the Ministry of Science and Higher Education of the Russian Federation, project No. 075-15-2019-1621.

References

1. Jaiswal, N. K. 1968. *Priority queues*. New York; London: Academic press. 240 p.
2. Matveev, V. F., and V. G. Ushakov. 1984. *Sistemy massovogo obsluzhivaniya* [Queueing systems]. Moscow: M. V. Lomonosov Moscow State University Publs. 240 p.
3. Takagi, H. 1991. *Queueing analysis: A foundation of performance evaluation*. Amsterdam: North-Holland Elsevier. Vol. 1. Part 1. 487 p.

Received March 28, 2022

Contributors

Bergovin Alexey K. (b. 1995) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; alexey.bergovin@gmail.com

Ushakov Vladimir G. (b. 1952) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; vgushakov@mail.ru

ВЕРОЯТНОСТНАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ ОСНОВНЫХ ХАРАКТЕРИСТИК ПРОИЗВОДИТЕЛЬНОСТИ МАРКОВСКОЙ МОДЕЛИ СУПЕРКОМПЬЮТЕРА*

Р. В. Разумчик[1], А. С. Румянцев[2], Р. М. Гаримелла[3]

Аннотация: Рассматривается известная модель суперкомпьютера в виде двухлинейной экспоненциальной неконсервативной системы массового обслуживания (СМО) с очередью неограниченной емкости и с одновременным занятием и одновременным освобождением заявкой случайного числа приборов. Впервые доказано, что ее основные вероятностные характеристики обслуживания могут быть получены из принципиально другой модели, представляющей собой однолинейную СМО неограниченной емкости с потерями поступающих заявок, но без искусственных простоев прибора. С привлечением развитого аналитического аппарата анализа обобщенных процессов размножения и гибели (ПРГ) показано, что в важных частных случаях совместное нестационарное распределение вероятностей ее состояний представляется в матрично-геометрической форме и может быть найдено (в терминах преобразования Лапласа (ПЛ)) с помощью метода на основе информации о пересечении уровней. Приведены примеры численных расчетов, иллюстрирующие некоторые характеристики установленной связи между двумя моделями.

Ключевые слова: суперкомпьютер; система массового обслуживания (СМО); неконсервативность дисциплины обслуживания; нестационарный режим

DOI: 10.14357/19922264230209 **EDN:** KXYHPO

1 Введение

В области вычислительных и телекоммуникационных систем широко применяются системы параллельного обслуживания, такие как многопроцессорные вычислительные устройства общего назначения, суперкомпьютеры, системы передачи данных и системы хранения. Для исследования характеристик производительности таких систем вследствие случайного характера поступающей нагрузки и зачастую наличия очередей применяются методы теории массового обслуживания и многолинейные СМО. В этом классе СМО важную роль играет подкласс систем с одновременным занятием и одновременным освобождением заявкой случайного числа приборов на одно и то же случайное время. Сейчас в научном сообществе по исследованию операций он активно применяется для моделирования и анализа характеристик производительности современных суперкомпьютеров и центров обработки данных [1–5]. Однако ранее СМО этого класса применялись также к исследованию систем социального обслуживания [6, 7]. Особенность этих систем заключается в неконсервативности процесса нагрузки[4]: приборы могут простаивать при непустой очереди. Это обстоятельство существенно затрудняет анализ даже систем малого размера [9, 10], и, как следствие, многие вопросы здесь остаются либо открытыми, либо недостаточно разработанными (см. более подробно [11]).

Эта статья посвящена, главным образом, модели суперкомпьютера, представляющей собой двухлинейную неконсервативную марковскую СМО с двумя типами заявок и дисциплиной обслуживания в порядке поступления (см., например, [12] и подробное описание ниже, в начале разд. 3). Хотя эта модель в научной литературе исследована достаточно подробно [2–7, 9, 10, 12–17], здесь, по-видимому, впервые показано, что ее основные вероятностные характеристики обслуживания могут быть получены из принципиально другой модели, представляющей собой СМО, в которой, во-первых,

*Исследование выполнено за счет гранта Российского научного фонда № 21-71-10135. Работа проводилась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, rrazumchik@frccsc.ru

[2]Институт прикладных математических исследований Карельского научного центра Российской академии наук, ar0@krc.karelia.ru

[3]Университет Махиндра, Индия, rama.murthy@mahindrauniversity.edu.in

[4]Или, другими словами, неконсервативности дисциплины обслуживания. Напомним (см., например, [8, с. 49]), что свойство консервативности дисциплины обслуживания означает, что длительность обслуживания заявки не зависит от дисциплины обслуживания и нет искусственных простоев прибора.

нет искусственных простоев прибора и, во-вторых, имеют место потери заявок при неограниченной емкости очереди. Для наглядности ее описание (см. разд. 2) дано в рамках дискретной системы поллинга (о них см., например, [18]), в которой анализу подвергается только очередь высшего приоритета. В отличие от исходной модели, в которой процесс обслуживания имеет конкретное физическое наполнение (заключающееся в выполнении одного или одновременно нескольких суперкомпьютерных заданий), в новой модели каждый акт обслуживания является виртуальным и однозначного физического наполнения не имеет.

Таким образом, новая система представляет собой отвлеченную математическую модель, результаты анализа которой требуют интерпретации. Для распределения очереди она дана в разд. 3 (см. соотношение (4)). Для временны́х характеристик вопрос остается открытым, но вычислительные эксперименты указывают на наличие здесь некоторых закономерностей (см. разд. 5). Судя по публикациям в открытой печати, внимание подробному исследованию предложенной в разд. 2 СМО не уделялось[1]. Поэтому в разд. 4 внимание уделено стандартному вопросу расчета совместного нестационарного распределения вероятностей ее состояний (и, ввиду (4), состояний связанной с ней суперкомпьютерной модели) и показано, что в некоторых частных случаях оно может быть найдено (в терминах ПЛ) известным, но нестандартным методом на основе информации о пересечении уровней [21].

Далее используются следующие обозначения: $\mathbf{I}$ — тождественная матрица; $\mathbf{0}$ — нулевая матрица; $\vec{1}$ — вектор, состоящий из единиц. Там, где размерность векторов и матриц не ясна из контекста, она определяется нижним индексом.

2 Описание модели и ее вероятностные характеристики

Пусть имеется некоторая дискретная система поллинга с одним обслуживающим прибором, $N \geq 1$ очередями и приоритетным порядком их обхода. Переключения прибора между очередями происходят мгновенно. Будем рассматривать только очередь, имеющую наивысший приоритет. Предположим, что число мест для ожидания в ней неограниченно и в нее поступает пуассоновский поток групп заявок интенсивности λ. Число заявок, приходящих в каждой группе, равно двум. Если в момент поступления группы в очереди находится нечетное число заявок, то поступающая группа целиком помещается в нее, иначе только одна заявка из группы (другая считается потерянной). Внутри очереди заявки обслуживаются в порядке поступления. Дисциплина обслуживания очереди прибором — исчерпывающая, т. е. очередь обслуживается до тех пор, пока не опустеет.

Процесс обслуживания относится к марковскому типу и определяется следующим образом. Если в очереди находятся k, $k \geq 0$, заявок, то процесс обслуживания может находиться в одном из l_k, $1 \leq l_k < \infty$, состояний (фаз обслуживания), причем $l_0 = 1$. Тогда если в некоторый момент обслуживания в очереди находятся $k \geq 1$ заявок и фаза обслуживания равна i, $1 \leq i \leq l_k$, то за «малое» время Δ с вероятностью $\lambda_{ij}^{(k)}\Delta + o(\Delta)$ ни одна заявка не покинет очередь и фаза обслуживания изменится на j, $1 \leq j \leq l_k$, а с вероятностью $n_{ij}^{(k)}\Delta + o(\Delta)$ обслуживание закончится, причем если k — нечетное число, то очередь мгновенно покинет одна заявка, иначе две, и (в обоих случаях) фаза обслуживания изменится на j, $1 \leq j \leq l_{k-1-(k-1 \bmod 2)}$. Матрицы из элементов $\lambda_{ij}^{(k)}$ и $n_{ij}^{(k)}$ будем обозначать соответственно $\mathbf{\Lambda}^{(k)}$ и $\mathbf{N}^{(k)}$, $k \geq 1$. Кроме того, будем предполагать, что начиная с некоторого номера k все l_k равны, $\mathbf{\Lambda}^{(k)} = \mathbf{L}$, $\mathbf{N}^{(k)} = \mathbf{N}$, матрица $\mathbf{\Lambda} + \mathbf{N}$ неразложимая, а матрица $\mathbf{N}$ ненулевая. Наконец, будем предполагать, что если в момент поступления в очереди имеются $k \geq 0$ других заявок и фаза обслуживания равна i, $1 \leq i \leq l_k$, то после поступления она с вероятностью $\omega_{ij}^{(k)}$ изменится на j, $1 \leq j \leq l_{k+1+(k \bmod 2)}$. Матрицы из элементов $\omega_{ij}^{(k)}$ будем обозначать $\mathbf{\Omega}^{(k)}$, $k \geq 0$, и считать, что $\mathbf{\Omega}^{(k)} = \mathbf{\Omega}$ начиная с некоторого номера k.

Обозначим через $\hat{X}(t)$ число заявок в очереди, а через $\hat{Y}(t)$ — фазу обслуживания в момент t. Положим

$$\hat{Z}(t) = \left(\hat{X}(t), \hat{Y}(t)\right).$$

При сделанных предположениях относительно входящего потока и процесса обслуживания случайный процесс $\{\hat{Z}(t), t \geq 0\}$ является марковским. Множество его состояний имеет вид $\{(k, j), k \geq 0,\ 1 \leq j \leq l_k\}$, где индексы k и j указывают соответственно число заявок в очереди и фазу обслуживания. Следуя терминологии обобщенных ПРГ, назовем уровнем процесса $\{\hat{Z}(t),\ t \geq 0\}$ число заявок в очереди, а его фазой — фазу обслуживания. Из описания модели следует, что, если рассматриваемый процесс находится на уровне k, возможны переходы либо с текущего уровня на

[1] И связано это, во-видимому, с тем, что с точки зрения вероятностных характеристик она представляет собой лишь частный случай хорошо известной СМО $\mathrm{SM}/\mathrm{MSP}/n/r$ (см., например, [19, 20]).

один или на два уровня выше, либо внутри текущего уровня, либо с текущего уровня на один или два уровня ниже. Так, если k, $k \geq 1$, есть число нечетное, то строки матрицы интенсивностей переходов[2] $\hat{\mathbf{Q}}$, соответствующие k-му и $(k+1)$-му уровням процесса $\{\hat{Z}(t), t \geq 0\}$, имеют вид:

$$\hat{\mathbf{Q}} = \begin{pmatrix} \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underbrace{\mathbf{0}\cdots\mathbf{0}}_{k-1} & \mathbf{N}^{(k)} & \mathbf{M}^{(k)} & \mathbf{0} & \lambda\mathbf{\Omega}^{(k)} & \mathbf{0} & \cdots \\ \underbrace{\mathbf{0}\cdots\mathbf{0}}_{k-2} & \mathbf{N}^{(k+1)} & \mathbf{0} & \mathbf{M}^{(k+1)} & \lambda\mathbf{\Omega}^{(k+1)} & \mathbf{0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}, \quad (1)$$

где матрицы $\mathbf{M}^{(k)}$ вычисляются по формуле

$$\mathbf{M}^{(k)} = \mathbf{\Lambda}^{(k)} - \operatorname{diag}\left(\lambda\mathbf{\Omega}^{(k)}\vec{1}\right).$$

Введем обозначение

$$\vec{\hat{p}}_k(t) = (\hat{p}_{k,1}(t), \ldots\ldots, \hat{p}_{k,l_k}(t)),$$

где $\hat{p}_{(k,j)}(t) = \mathsf{P}(\hat{Z}(t) = (k,j))$ — вероятность того, что процесс $\{\hat{Z}(t), t \geq 0\}$ находится в состоянии (k,j) в момент t. Принципиально вопрос расчета совместного распределения $\vec{\hat{p}}(t) = (\hat{p}_0(t), \vec{\hat{p}}_1(t), \ldots)$ решается путем путем группировки элементов (1) в блоки вида

$$\begin{pmatrix} \mathbf{0} & \mathbf{N}^{(k)} \\ \mathbf{0} & \mathbf{N}^{(k+1)} \end{pmatrix}; \quad \begin{pmatrix} \mathbf{M}^{(k)} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{(k+1)} \end{pmatrix}; \quad \begin{pmatrix} \lambda\mathbf{\Omega}^{(k)} & \mathbf{0} \\ \lambda\mathbf{\Omega}^{(k+1)} & \mathbf{0} \end{pmatrix},$$

которые с некоторого номера k становятся регулярными, и далее применением одного из множества известных методов расчета распределений очередей в СМО, описываемых обобщенными ПРГ (для стационарного случая см., например, [22, теорема 2], [23, с. 242]; для нестационарного, например, [24, 25]).

Умея вычислять совместное распределение по стандартным алгоритмам, нетрудно рассчитать (хотя бы приближенно) и основные вероятностные характеристики обслуживания. Так, вероятность $\hat{q}(t)$ потери заявки[2] в момент t равна

$$\hat{q}(t) = \sum\nolimits_{k=0}^{\infty} \vec{\hat{p}}_{2k}(t)\vec{1}.$$

Из классической формулы Литтла можно получить стационарное[3] среднее время $\hat{V}$ пребывания заявки в очереди, равное $\hat{V} = \sum_{k=1}^{\infty} k\vec{\hat{p}}_k(\infty)\vec{1}/(\lambda(2-\hat{q}(\infty)))$. Другие временны́е характеристики требуют дополнительных построений, и их нахождение, по крайней мере в явном виде, остается открытым вопросом.

3 Связь с моделью суперкомпьютера

Рассмотрим модель суперкомпьютера (см., например, [2, 12]) в виде СМО с двумя идентичными приборами (единичной скорости), обслуживающими заявки (в порядке поступления) из единственной очереди неограниченной емкости. Заявки поступают в очередь по пуассоновскому закону с параметром λ. Для выполнения каждой заявке с вероятностью $p_1 \in (0,1)$ требуется один прибор, а с дополнительной вероятностью $p_2 = 1 - p_1$ — два; конкретное число становится известным в момент поступления заявки в систему и остается фиксированным для данной заявки. Времена обслуживания заявок являются независимыми случайными величинами и имеют экспоненциальное распределение с параметром μ_i, где i — число требуемых заявкой приборов (тип заявки). Предполагается, что занятие (и освобождение) приборов заявкой второго типа происходит одновременно. Заявка поступает на обслуживание, во-первых, когда подошла ее очередь и, во-вторых, когда требуемое ей число приборов свободно.

Обозначим через $X(t)$ общее число заявок в системе в момент t, через $Y(t)$ — комбинацию типов двух старейших заявок в системе в момент t, подразумевая, что

$$Y(t) = \begin{cases} 1, & \text{если в момент } t \text{ обе старейшие заявки первого типа;} \\ 2, & \text{если в момент } t \text{ старейшая заявка первого типа, а вторая старейшая — второго типа (ожидает на первой позиции в очереди);} \\ 3 & \text{в остальных случаях.} \end{cases}$$

Заметим, что если $X(t) = 1$, то $Y(t)$ — тип единственной заявки в системе в момент t. Положим[1] $Z(t) = (X(t), Y(t))$. Напомним (см., напри-

[2] Первая строка матрицы $\hat{\mathbf{Q}}$, соответствующая нулевому уровню, имеет вид $(-\lambda\ \lambda\mathbf{\Omega}_0\ \mathbf{0}\ \cdots)$.

[2] Считая в покидающей очередь группе одну из двух заявок необслуженной, можно ввести такие характеристики, как (виртуальная) вероятность (не)обслуживания. Их расчет более сложен и требует отдельного анализа.

[3] Критерий существования которого может быть выписан в явном виде, так как после указанной выше группировки матрица $\hat{\mathbf{Q}}$ оказывается трехдиагональной с повторяющимися блоками на «верхних уровнях».

[1] При $X(t) = 0$ вторая компонента процесса опускается.

мер, [26]), что матрица интенсивностей $\mathbf{Q}$ переходов процесса $\{Z(t), t \geq 0\}$ имеет вид:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}^{0,0} & \mathbf{A}^{0,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}^{1,0} & \mathbf{A}^{1,1} & \mathbf{A}^{1,2} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}^{2,1} & \mathbf{A}^{(0)} & \mathbf{A}^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{(-1)} & \mathbf{A}^{(0)} & \mathbf{A}^{(1)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}, \quad (2)$$

и, значит, $\{Z(t), t \geq 0\}$ — обобщенный ПРГ, причем $X(t)$ — уровень процесса; $Y(t)$ — его фаза. Ненулевые блоки в (2) имеют следующую структуру:

$$\mathbf{A}^{0,0} = -\lambda; \quad \mathbf{A}^{0,1} = \lambda(p_1 \ p_2);$$

$$\mathbf{A}^{1,0} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \quad \mathbf{A}^{1,1} = -\operatorname{diag}\left(\mathbf{A}^{1,0} + \mathbf{A}^{1,2}\vec{1}\right);$$

$$\mathbf{A}^{1,2} = \begin{pmatrix} \lambda p_1 & \lambda p_2 & 0 \\ 0 & 0 & \lambda \end{pmatrix}; \quad \mathbf{A}^{2,1} = \begin{pmatrix} 2\mu_1 & 0 \\ 0 & \mu_1 \\ p_1\mu_2 & p_2\mu_2 \end{pmatrix};$$

$$\mathbf{A}^{(0)} = -\operatorname{diag}\left(\mathbf{A}^{2,1} + \mathbf{A}^{(1)}\vec{1}\right); \quad \mathbf{A}^{(1)} = \lambda \mathbf{I}_3;$$

$$\mathbf{A}^{(-1)} = \begin{pmatrix} 2\mu_1 p_1 & 2\mu_1 p_2 & 0 \\ 0 & 0 & \mu_1 \\ p_1^2\mu_2 & p_1 p_2 \mu_2 & \mu_2 p_2 \end{pmatrix}.$$

Введем обозначения:

$$\vec{p}_1(t) = (p_{1,1}(t), p_{1,2}(t));$$
$$\vec{p}_k(t) = (p_{k,1}(t), p_{k,2}(t), p_{k,3}(t)), \ k \geq 2;$$
$$\vec{p}(t) = (p_0(t), \vec{p}_1(t), \dots),$$

где $p_{(k,j)}(t) = \mathsf{P}(Z(t) = (k,j))$ — вероятность того, что процесс $\{Z(t), t \geq 0\}$ находится в состоянии (k,j) в момент t.

Рассмотрим бесконечную матрицу $\mathbf{T}$ вида:

$$\mathbf{T} = \operatorname{diag}\left(1, \vec{1}_2 \otimes \mathbf{I}_2, \vec{1}_2 \otimes \mathbf{I}_3, \vec{1}_2 \otimes \mathbf{I}_3, \dots\right),$$

где $\otimes$ — кронекерово произведение. Отметим, что $\mathbf{T}$ — это матрица Теплица [27, с. 86]. Как можно убедиться непосредственной проверкой, если элементы матрицы $\hat{\mathbf{Q}}$ выбраны таким образом, что

$$\mathbf{N}^{(1)} = \mathbf{N}^{(2)} = \mathbf{A}^{1,0}; \quad \mathbf{N}^{(3)} = \mathbf{N}^{(4)} = \mathbf{A}^{2,1};$$
$$\mathbf{N}^{(k)} = \mathbf{A}^{(-1)}, \ k \geq 5;$$
$$\mathbf{\Omega}^{(0)} = \mathbf{A}^{0,1}; \quad \mathbf{\Omega}^{(1)} = \mathbf{\Omega}_2 = \mathbf{A}^{1,2};$$
$$\mathbf{\Omega}^{(k)} = \mathbf{A}^{(1)}, \ k \geq 3, \quad (3)$$

то имеет место тождество $\hat{\mathbf{Q}}\mathbf{T} = \mathbf{T}\mathbf{Q}$. Поскольку вероятности $\vec{\hat{p}}(t)\mathbf{T}$ удовлетворяют системе дифференциальных уравнений $(d/dt)\vec{\hat{p}}(t)\mathbf{T} = \vec{\hat{p}}(t)\hat{\mathbf{Q}}\mathbf{T}$, то с учетом цепочки равенств $\vec{\hat{p}}(t)(\hat{\mathbf{Q}}\mathbf{T}) = \vec{\hat{p}}(t)(\mathbf{T}\mathbf{Q}) = (\vec{\hat{p}}(t)\mathbf{T})\mathbf{Q}$, ассоциативность умножения в которой гарантируется конечным числом ненулевых элементов в каждом столбце матрицы $\mathbf{T}$, покомпонентно имеем:

$$\vec{p}(t) = \vec{\hat{p}}(t)\mathbf{T}, \ \ t \geq 0\,. \quad (4)$$

Таким образом, вероятности состояний рассматриваемой суперкомпьютерной модели совпадают с вероятностями укрупненных (с помощью матрицы $\mathbf{T}$) состояний частного случая модели из разд. 2.

4 Нестационарное распределение очереди

Пусть элементы $\hat{\mathbf{Q}}$ заданы согласно (3). Нетрудно видеть, что матрица $\hat{\mathbf{Q}} - s\mathbf{I}$ допускает представление

$$\begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{C} & \mathbf{B} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{C} & \mathbf{B} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}, \quad (5)$$

в котором (не все квадратные) блоки $\mathbf{B}_0$, $\mathbf{B}_1$, $\mathbf{B}$ и $\mathbf{C}$ зависят от s и имеют вид:

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}^{(-1)} & \mathbf{A}^{(0)} - s\mathbf{I} \\ \mathbf{A}^{(-1)} & \mathbf{0} \end{pmatrix};$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{A}^{(1)} \\ \mathbf{A}^{(0)} - s\mathbf{I} & \mathbf{A}^{(1)} \end{pmatrix}; \quad \mathbf{B}_1 = \begin{pmatrix} \mathbf{0} \\ \mathbf{B} \end{pmatrix};$$

$$\mathbf{B}_0 = \begin{pmatrix} \mathbf{A}^{0,0} - s\mathbf{I} & \mathbf{A}^{0,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{A}^{1,0} & \mathbf{A}^{1,1} - s\mathbf{I} & \mathbf{0} & \mathbf{A}^{1,2} \\ \mathbf{A}^{1,0} & \mathbf{0} & \mathbf{A}^{1,1} - s\mathbf{I} & \mathbf{A}^{1,2} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{2,1} & \mathbf{A}^{(0)} - s\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{2,1} & \mathbf{0} \end{pmatrix}.$$

Обозначим распределение вероятностей состояний обобщенного ПРГ с матрицей интенсивностей (при $s = 0$) переходов (5) через $\vec{\check{p}}(t) = (\vec{\check{p}}_0(t), \vec{\check{p}}_1(t), \dots)$, а его покомпонентное преобразование Лапласа в точке s, $s \geq 0$, — через $\vec{\tilde{\pi}}(s) = (\vec{\tilde{\pi}}_0(s), \vec{\tilde{\pi}}_1(s), \dots)$. Далее без ограничения общности можно считать, что в начальный момент процесс находится в нулевом состоянии, т. е. $\check{p}_0(0) = 1$. Поскольку $\vec{\tilde{\pi}}(s)$ удовлетворяет системе линейных алгебраических уравнений

$$\vec{\tilde{\pi}}(s)(\hat{\mathbf{Q}} - s\mathbf{I}) = -\vec{\check{p}}(0),$$

то, если матрица $\mathbf{C}$ не вырождена, расчет $\vec{\tilde{\pi}}_k(s)$, $k \geq 2$, можно осуществить по матрично-рекуррентной формуле. Однако $\mathbf{C}$ обратимой не является (определитель $\mathbf{A}^{(-1)}$ равен нулю). Выполним, следуя [21], операции над столбцами матрицы (5), не

изменяющие решения. Выделим в (5) все группы столбцов, содержащие целиком матрицу $\mathbf{C}$, домножим в каждой группе второй столбец на p_2/p_1 и вычтем из первого столбца. Столбцы матрицы $\mathbf{B}$ автоматически претерпят аналогичные изменения. В результате преобразований первый столбец каждой матрицы $\mathbf{C}$ будет содержать только нули.

Перегруппируем теперь элементы матрицы (5) с сохранением блочной структуры и обозначений следующим образом. Добавим в матрицу $\mathbf{B}_0$ (справа) первый столбец $\mathbf{B}_1$. Новую матрицу $\mathbf{B}_1$ «начнем» со второго столбца прежней и дополним ее справа нулевым столбцом.

В качестве новой матрицы $\mathbf{C}$ возьмем прежнюю, но начиная со второго столбца, и дополним ее (справа) первым столбцом прежней матрицы $\mathbf{B}$. Новую матрицу $\mathbf{B}$ определим аналогичным образом, начиная со второго столбца прежней, дополнив нулевым столбцом (справа). Теперь определитель матрицы $\mathbf{C}$, равный $\mu_1\mu_2p_1p_2(\lambda+s)(\lambda+\mu_2+s)(\lambda+\mu_1+s)(\lambda+2\mu_1+s)$, всегда отличен от нуля и, таким образом, ПЛ распределения вероятностей состояний $\vec{\pi}(s)$ представляется в матрично-геометрической форме

$$\vec{\pi}_{k+1}(s) = \vec{\pi}_1(s)(-\mathbf{B}\mathbf{C}^{-1})^k \quad k \geq 1\,,$$

а векторы $\vec{\pi}_0(s)$ и $\vec{\pi}_1(s)$ определяются единственным образом из системы уравнений

$$\vec{\pi}_0(s)\mathbf{B}_0 = -(1,0,\ldots,0); \quad \vec{\pi}_0(s)\mathbf{B}_1 = -\vec{\pi}_1(s)\mathbf{C}.$$

По построению число уравнений в этой системе на два меньше числа неизвестных. Как следует из [21, теорема 4.1], недостающие два уравнения имеют вид:

$$\vec{\pi}_1(s)\vec{\psi} = 0\,; \quad \vec{\pi}_1(s)\vec{\phi} = 0,$$

где $\vec{\psi}$ и $\vec{\phi}$ — правые собственный векторы, соответствующие тем (двум из шести) собственным значе-

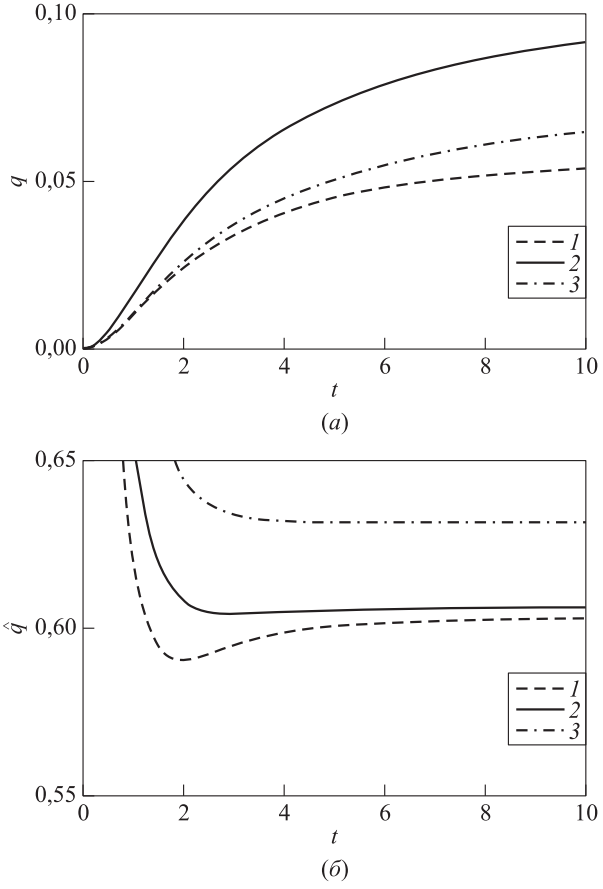


Рис. 1 Условные вероятности вынужденного простоя прибора (*а*) и вероятности потери заявки (*б*) как функции времени. Среднее время обслуживания заявок разных типов одинаковое: *1* — $p_1 = 8/9$; *2* — $2/3$; *3* — $p_1 = 1/3$

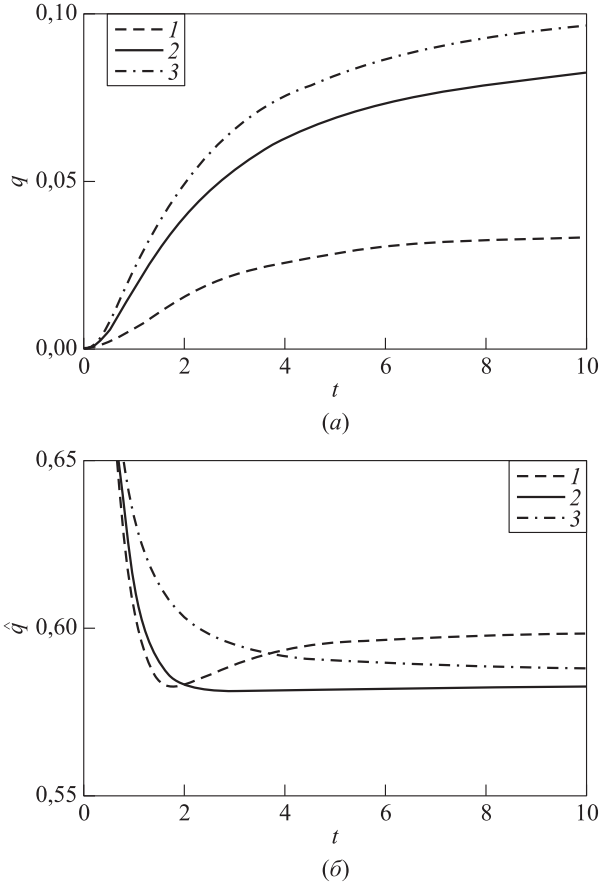


Рис. 2 Условные вероятности вынужденного простоя прибора (*а*) и вероятности потери заявки (*б*) как функции времени. Среднее время обслуживания заявок разных типов различное: *1* — $p_1 = 8/9$; *2* — $2/3$; *3* — $p_1 = 1/3$

ниям матрицы $-\mathbf{BC}^{-1}$, которые лежат на границе или вне единичного круга.

Хотя полученный результат и позволяет находить (в терминах ПЛ) нестационарные вероятностные характеристики (обеих моделей, рассмотренных в разд. 2 и 3), окончательные формулы не являются замкнутыми в полном смысле этого слова и требуют применения численных методов. Не останавливаясь на обсуждении достоинств и недостатков описанного решения[1] по сравнению с традиционными, проиллюстрируем сказанное двумя примерами. Верхняя группа кривых на рис. 1 и 2 показывает поведение условной вероятности $q(t)$ вынужденного простоя прибора[2] с ростом t при трех соотношениях между типами заявок, когда приборы загружены в среднем наполовину[3]. Рисунок 1 соответствует однородному случаю, когда $\mu_1 = \mu_2 = 1$, рис. 2 — неоднородному, когда $\mu_2 = 3\mu_1 = 3$. Нижняя группа кривых на каждом рисунке показывает соответствующие значения вероятности потерь $\hat{q}(t)$ в модели из разд. 2, параметры которой выбраны согласно (3).

Отметим немонотонный характер изменения вероятности $q(t)$ (при каждом t) как функции от p_1: в одном случае доля времени, в течение которого прибор вынужденно простаивает, достигает наибольшего значения, когда существенно преобладает один из двух типов заявок, в другом случае — когда такого преобладания нет. Аналогичное по характеру замечание (но, как видно из рис. 2, уже с учетом значения t) справедливо и для вероятности $\hat{q}(t)$.

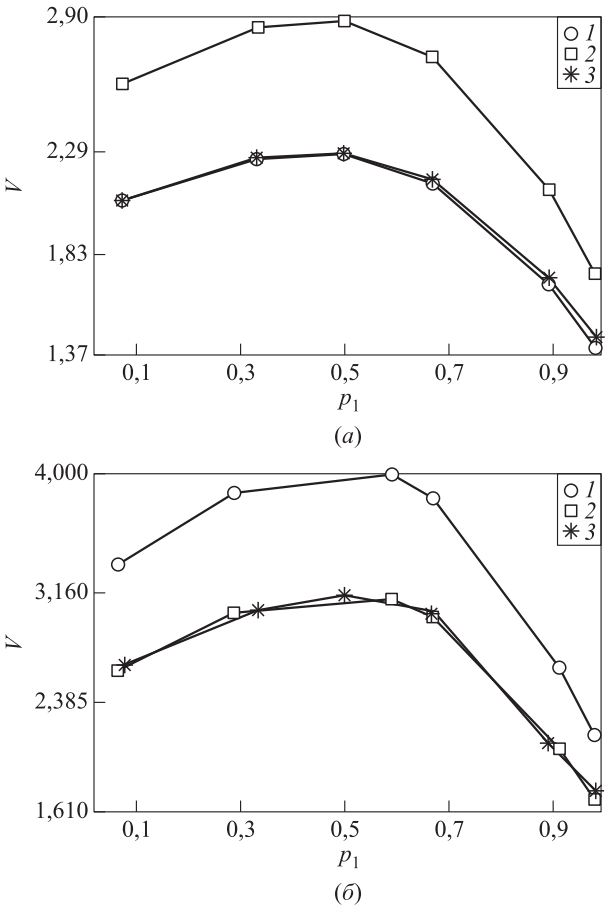


Рис. 3 Стационарное среднее время V (*1*) ($\hat{V}$ (*2*) и $\hat{V}^*$ (*3*)) пребывания заявки в модели из разд. 3 (разд. 2) при стационарной средней загрузке приборов $\mathsf{E}\,U = 0{,}5$ (*а*) и 0,6 (*б*) и различных значениях p_1. Среднее время обслуживания заявок разных типов одинаковое и равно 1

5 Заключение

В этой статье впервые установлена связь (см. соотношение (4)), существующая между хорошо исследованной в литературе моделью суперкомпьютера в виде двухлинейной неконсервативной СМО неограниченной емкости и отвлеченной математической моделью, представляющей собой однолинейную СМО с групповым поступлением, механизмом активного управления очередью (приводящего к потерям поступающих заявок) и групповым обслуживанием, зависящим от состояния очереди. Характер этой связи требует дальнейшего прояснения. В частности, неясно, существует ли для условной вероятности вынужденного простоя прибора в первой СМО эквивалентная характеристика во второй: как видно из рис. 1 и 2, вероятность $\hat{q}(t)$ потери поступающей заявки таковой не является. Как показывают вычислительные эксперименты, с точки зрения такой характеристики, как стационарное среднее время пребывания заявки в системе, вторая СМО (при дисциплине FIFO — first in, first out) мажорирует первую сверху (рис. 3), и эти оценки лучше, по крайней мере известных из литературы, «наивных» (соответствующих случаю $p_1 = 0$). Вместе с тем вопрос о стохастической упорядоченности соответствующих случайных величин остается открытым.

Что касается описанной в разд. 2 модели, то, как уже отмечалось выше, она обобщается на случай полумарковского входящего потока и теоретический[4] интерес здесь представляет получение

[1] Которое можно назвать спектральным, поэтому см. [28, Ch. 13].

[2] Вычисляемой по формуле $q(t) = \sum_{k=2}^{\infty} p_{k,2}(t)/(2p_0(t) + p_{1,1}(t) + \sum_{k=2}^{\infty} p_{k,2}(t))$.

[3] То есть их стационарная средняя загрузка $\mathsf{E}U = 1 - p_0(\infty) - 0{,}5p_{1,1}(\infty) - 0{,}5\sum_{k=2}^{\infty} p_{k,2}(\infty) = 0{,}5$.

[4] А с учетом установленной в этой статье связи с суперкомпьютерными моделями и практический, поскольку для последних проблема анализа и оптимизации временны́х характеристик не исчерпана (см., например, [7]).

временны́х характеристик при различных[1] правилах учета и выбора заявок из очереди на обслуживание. Стандартные результаты для обобщенных ПРГ (см., например, [29]) оказываются здесь малопригодными.

Литература

1. *Морозов Е. В., Румянцев А. С.* Модели многосерверных систем для анализа вычислительного кластера // Труды Карельского научного центра РАН, 2011. № 5. С. 75–85.
2. *Rumyantsev A., Morozov E.* Stability criterion of a multiserver model with simultaneous service // Ann. Oper. Res., 2017. Vol. 252. P. 29–39. doi: 10.1007/s10479-015-1917-2.
3. *Hong Y., Wang W.* Sharp waiting-time bounds for multiserver jobs // 23rd Symposium (International) on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing Proceedings. — ACM, 2022. P. 161–170. doi: 10.1145/3492866.3549717.
4. *Grosof I.* Optimal scheduling in the multiserver-job model under heavy traffic // Proceedings ACM Measurement Analysis Computing Systems, 2022. Vol. 3. No. 6. Art. 51. doi: 10.1145/3570612.
5. *Wang W., Xie Q., Harchol-Balter M.* Zero queueing for multi-server jobs // ACM Sigmetrics Performance Evaluation Review, 2022. Vol. 1. No. 49. P. 13–14. doi: 10.1145/3543516.3453924.
6. *Kim S.* $M/M/s$ queueing system where customers demand multiple server use: PhD Diss. — Dallas, TX, USA: Southern Methodist University, 1979. 104 p.
7. *Brill P. H., Green L.* Queues in which customers receive simultaneous service from a random number of servers: A system point approach // Manage. Sci., 1984. Vol. 30. No. 1. P. 51–68.
8. *Яшков С. Ф.* Анализ очередей в ЭВМ. — М.: Радио и связь, 1989. 216 с.
9. *Filippopoulos D., Karatza H.* An $M/M/2$ parallel system model with pure space sharing among rigid jobs // Math. Comput. Model., 2007. Vol. 45. P. 491–530.
10. *Chakravarthy S. R., Karatza H. D.* Two-server parallel system with pure space sharing and Markovian arrivals // Comput. Oper. Res., 2013. Vol. 40. No. 1. P. 510–519.
11. *Harchol-Balter M.* Open problems in queueing theory inspired by datacenter computing // Queueing Syst., 2021. Vol. 97. P. 3–37.
12. *Rumyantsev A., Basmadjian R., Golovin A., Astafiev S.* A three-level modelling approach for asynchronous speed scaling in high-performance data centres // 12th Conference (International) on Future Energy Systems e-Energy Proceedings. — ACM, 2021. P. 417–423.
13. *Unwin A. R.* Results for dual resource queues // Modelling and performance evaluation methodology / Eds. F. Baccelli, G. Fayolle. — Lecture notes in control and information sciences ser. — Berlin/Heidelberg: Springer-Verlag, 1984. Vol. 60. P. 351–370. doi: 10.1007/BFb0005182.
14. *Melikov A. Z.* Computation and optimization methods for multiresource queues // Cybern. Syst. Anal., 1996. Vol. 32. No. 6. P. 821–836. doi: 10.1007/BF02366862.
15. *Afanaseva L., Bashtova E., Grishunina S.* Stability analysis of a multi-server model with simultaneous service and a regenerative input flow // Methodol. Comput. Appl., 2020. Vol. 22. P. 1439–1455. doi: 10.1007/s11009-019-09721-9.
16. *Grosof I., Harchol-Balter M., Scheller-Wolf A.* Stability for two-class multiserver-job systems. — Cornell University, 2020. 29 p. arXiv:2010.00631 [cs.PF].
17. *Harchol-Balter M.* The multiserver job queueing model // Queueing Syst., 2022. Vol. 100. No. 3-4. P. 201–203. doi: 10.1007/s11134-022-09762-x.
18. *Вишневский В. М., Семенова О. В.* Математические методы исследования систем поллинга // Автомат. и телемех., 2006. Вып. 2. С. 3–56.
19. *Печинкин А. В., Чаплыгин В. В.* Стационарные характеристики системы массового обслуживания $\mathrm{SM/MSP}/n/r$ // Автомат. и телемех., 2004. Вып. 9. С. 85–100.
20. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The theory of queuing systems with correlated flows. — Cham, Switzerland: Springer, 2020. 410 p.
21. *Zhang J., Coyle E. J.* Transient analysis of quasi-birth-death processes // Commun. Stat. Stochastic Models, 1989. Vol. 5. No. 3. P. 459–496. doi: 10.1080/15326348908807119.
22. *Вишневский В. М., Дудин А. Н.* Системы массового обслуживания с коррелированными входными потоками и их применение для моделирования телекоммуникационных сетей // Автомат. и телемех., 2017. Вып. 8. С. 3–59.
23. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. — М.: РУДН, 1995. 529 с.
24. *Le Ny L. M., Sericola B.* Transient analysis of the $\mathrm{BMAP/PH}/1$ queue // Int. J. Simulation Systems Science Technology, 2002. Vol. 3. No. 3. P. 4–14.
25. *Hiroyuki M., Takine T.* Algorithmic computation of the time-dependent solution of structured Markov chains and its application to queues // Stoch. Models, 2005. Vol. 21. P. 885–912.

[1] Обратимся к примерам на рис. 3. Если при удалении из очереди двух заявок считать обслуженной ту, что стоит на последнем месте, то стационарное среднее время пребывания уменьшается с $\hat{V}$ до $\hat{V}^*$.

26. *Rumyantsev A., Basmadjian R., Astafiev S., Golovin A.* Three-level modeling of a speed-scaling supercomputer // Ann. Oper. Res., 2022. doi: 10.1007/s10479-022-04830-0.

27. *Гелбаум Б., Олмстед Д.* Контрпримеры в анализе / Пер. с англ. — М.: Мир, 1967. 251 с. (*Gelbaum B. R., Olmsted J. M. H.* Counterexamples in analysis. — 1st ed. — Dover books on matematics ser. — Dover Publicatrions, 1964. 194 p.)

28. *Dshalalow J. H.* Advances in queueing: Theory, methods and open problem. — London: CRC Press, 1995. 528 p.

29. *Ozawa T.* Sojourn time distributions in the queue defined by a general QBD process // Queueing Syst., 2006. Vol. 53. No. 4. P. 203–211. doi: 10.1007/s11134-006-7651-3.

Поступила в редакцию 15.04.23

A QUEUEING SYSTEM FOR PERFORMANCE EVALUATION OF A MARKOVIAN SUPERCOMPUTER MODEL

R. V. Razumchik[1], A. S. Rumyantsev[2], and R. M. Garimella[3]

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[2]Institute of Applied Mathematical Research of the Karelian Research Center of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Russian Federation
[3]Mahindra University, 62/1A Bahadurpally Jeedimetla, Hyderabad 500043, India

Abstract: Consideration is given to the well-known supercomputer model in the form of a Markovian nonwork-conserving two-server queueing system with unlimited queue capacity, in which customers are served by a random number of servers simultaneously. For the first time, it is shown that its basic probabilistic characteristics can be calculated from an unrelated single-server queueing system with infinite capacity, work conserving scheduling, and forced customers' losses. Based on the known matrix-analytic techniques for quasi-birth-and-death processes, it is shown that in certain special cases, the transient queue-size distribution can be found (in terms of Laplace transform) using the Level Crossing Information method and has a matrix-geometric form. Numerical examples which illustrate some properties of the established connection between the two queueing systems are provided.

Keywords: supercomputer model; queueing system; nonwork-conserving scheduling; transient regime

DOI: 10.14357/19922264230209 **EDN:** KXYHPO

Acknowledgments

The research was funded by the Russian Science Foundation, project No. 21-71-10135. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Morozov, E. V., and A. S. Rumyantsev. 2011. Modeli mnogoservernykh sistem dlya analiza vychislitel'nogo klastera [Multi-server models to analyze high performance cluster]. *Transactions of the Karelian Research Centre of the Russian Academy of Sciences* 5:75–85.

2. Rumyantsev, A., and E. Morozov. 2017. Stability criterion of a multiserver model with simultaneous service. *Ann. Oper. Res.* 252:29–39. doi: 10.1007/s10479-015-1917-2.

3. Hong, Y., and W. Wang. 2022. Sharp waiting-time bounds for multiserver jobs. *23rd Symposium (International) on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing Proceedings*. ACM. 161–170. doi: 10.1145/3492866.3549717.

4. Grosof, I. 2022. Optimal scheduling in the multiserver-job model under heavy traffic. *Proceedings ACM Measurement Analysis Computing Systems* 3(6):51. 32 p. doi: 10.1145/3570612.

5. Wang, W., Q. Xie, and M. Harchol-Balter. 2022. Zero queueing for multi-server jobs. *ACM Sigmetrics Performance Evaluation Review* 1(49):13–14. doi: 10.1145/3543516.3453924.

6. Kim, S. 1979. $M/M/s$ queueing system where customers demand multiple server use. Dallas, TX: Southern Methodist University, 1979. PhD Diss. 104 p.

7. Brill, P. H., and L. Green. 1984. Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Manage. Sci.* 30(1): 51–68.

8. Yashkov, S. F. 1989. *Analiz ocheredey v EVM* [Analysis of queues in computer systems]. Moscow: Radio i svyaz'. 216 p.

9. Filippopoulos, D., and H. Karatza. 2007. An $M/M/2$ parallel system model with pure space sharing among rigid jobs. *Math. Comput. Model.* 45:491–530.
10. Chakravarthy, S. R., and H. D. Karatza. 2013. Two-server parallel system with pure space sharing and Markovian arrivals. *Comput. Oper. Res.* 40(1):510–519.
11. Harchol-Balter, M. 2021. Open problems in queueing theory inspired by datacenter computing. *Queueing Syst.* 97:3–37.
12. Rumyantsev, A., R. Basmadjian, A. Golovin, and S. Astafiev. 2021. A three-level modelling approach for asynchronous speed scaling in high-performance data centres. *12th Conference (International) on Future Energy Systems e-Energy Proceedings*. ACM. 417–423.
13. Unwin, A. R. 1984. Results for dual resource queues. *Modelling and performance evaluation methodology*. Eds. F. Baccelli and G. Fayolle. Lecture notes in control and information sciences ser. Berlin/Heidelberg: Springer-Verlag. 60:351-370. doi: 10.1007/BFb0005182.
14. Melikov, A. Z. 1996. Computation and optimization methods for multiresource queues. *Cybern. Syst. Anal.* 32(6):821–836. doi: 10.1007/BF02366862.
15. Afanaseva, L., E. Bashtova, and S. Grishunina. 2020. Stability analysis of a multi-server model with simultaneous service and a regenerative input flow. *Methodol. Comput. Appl.* 22:1439–1455. doi: 10.1007/s11009-019-09721-9.
16. Grosof, I., M. Harchol-Balter, and A. Scheller-Wolf. 2020. Stability for two-class multiserver-job systems. *arXiv.org*. 29 p. Available at: https://arxiv.org/abs/2010.00631 (accessed April 30, 2023).
17. Harchol-Balter, M. 2022. The multiserver job queueing model. *Queueing Syst.* 100(3-4):201–203. doi: 10.1007/s11134-022-09762-x.
18. Vishnevskiy, V. M., and O. V. Semenova. 2006. Mathematical methods to study the polling systems. *Automat. Rem. Contr.* 67(2):173–220.
19. Pechinkin, A. V., and V. V. Chaplygin. 2004. Stationary characteristics of the $\mathrm{SM}/\mathrm{MSP}/n/r$ queuing system. *Automat. Rem. Contr.* 65(9):1429–1443.
20. Dudin, A. N., V. I. Klimenok, and V. M. Vishnevsky. 2020. *The theory of queuing systems with correlated flows*. Cham, Switzerland: Springer. 410 p.
21. Zhang, J., and E. J. Coyle. 1989. Transient analysis of quasi-birth-death processes. *Commun. Stat. Stochastic Models* 5(3):459–496. doi: 10.1080/15326348908807119.
22. Vishnevskii, V. M., and A. N. Dudin. 2017. Queueing systems with correlated arrival flows and their applications to modeling telecommunication network. *Automat. Rem. Contr.* 78:1361–1403.
23. Bocharov, P. P., and A. V. Pechinkin. 1995. *Teoriya massovogo obsluzhivaniya* [Queueing theory]. Moscow: RUDN. 529 p.
24. Le Ny, L. M., and B. Sericola. 2002. Transient analysis of the $\mathrm{BMAP}/\mathrm{PH}/1$ queue. *Int. J. Simulation Systems Science Technology* 3(3):4–14.
25. Hiroyuki, M., and T. Takine. 2005. Algorithmic computation of the time-dependent solution of structured Markov chains and its application to queues. *Stoch. Models* 21:885–912.
26. Rumyantsev, A., R. Basmadjian, S. Astafiev, and A. Golovin. 2022. Three-level modeling of a speed-scaling supercomputer. *Ann. Oper. Res.* doi: 10.1007/s10479-022-04830-0.
27. Gelbaum, B. R., and J. M. H. Olmsted. 2003. *Counterexamples in analysis*. Courier Corporation. 195 p.
28. Dshalalow, J. H. 1995. *Advances in queueing: Theory, methods and open problem*. London: CRC Press. 528 p.
29. Ozawa, T. 2006. Sojourn time distributions in the queue defined by a general QBD process. *Queueing Syst.* 53(4):203–211. doi: 10.1007/s11134-006-7651-3.

Received April 15, 2023

Contributors

Razumchik Rostislav V. (b. 1984) — Doctor of Science in physics and mathematics, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; rrazumchik@ipiran.ru

Rumyantsev Alexander S. (b. 1986) — Doctor of Science in physics and mathematics, senior scientist, Institute of Applied Mathematical Research of the "Karelian Research Center of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Russian Federation; ar0@krc.karelia.ru

Garimella Rama Murthy (b. 1962) — PhD in computer engineering, professor, Department of Computer Science and Engineering, Mahindra University, 62/1A Bahadurpally Jeedimetla, Hyderabad 500043, India; rama.murthy@mahindrauniversity.edu.in

К МОДЕЛИРОВАНИЮ ЭФФЕКТОВ ОБСЛУЖИВАНИЯ МНОГОАДРЕСНОГО ТРАФИКА В СЕТЯХ 5G NR*

А. К. Самуйлов[1], А. А. Платонова[2], В. С. Шоргин[3], Ю. В. Гайдамака[4]

Аннотация: Многоадресная передача данных в сетях беспроводного доступа позволяет эффективно предоставлять услугу группе абонентов и оказывается полезной для сокращения ресурса, необходимого для обслуживания пользователей, запрашивающих одни и те же данные. Поддержка этой функции в современной технологии 5G New Radio (NR) и будущих субтерагерцевых системах 6G сталкивается с особенностями, связанными с использованием фазированных антенных решеток (ФАР), формирующих направленные лучи. Представленная модель обслуживания многоадресного и одноадресного трафика позволяет исследовать области значений параметров сети связи 5G/6G для снижения плотности размещения базовых станций (БС) при поддержании качества предоставления услуг абонентам.

Ключевые слова: 5G; 6G; многоадресная передача; миллиметровые волны; терагерцевые частоты; фазированные антенные решетки; технологии радиодоступа; математическое моделирование

DOI: 10.14357/19922264230210 **EDN:** SLMGZU

1 Введение

Приложения недалекого будущего, такие как дополненная (AR, augmented reality) и виртуальная реальность (VR, virtual reality), телеприсутствие, видео 8/16 K, требуют резкого увеличения скорости передачи по радиоинтерфейсу [1]. Значительное увеличение скорости до 20 Гбит/с на БС ожидается в системах 5G NR за счет использования миллиметрового диапазона длин волн (mmWave) (с частотами 24–52,6 ГГц), а для сотовых систем 6G планируется использовать нижнюю часть диапазона миллиметровых волн (с частотами 52,6–100 ГГц) и даже субтерагерцевый диапазон частот (100–300 ГГц) [2], потенциально повышая скорость доступа до 100 Гбит/с на одну БС.

Многоадресная передача является важной возможностью сетей беспроводного доступа, повышающей эффективность использования ресурсов при обслуживании абонентов сети, запрашивающих одни и те же данные [3]. Если для предоставления услуги одноадресной передачи необходимо организовать отдельную сессию для каждого абонента, то для одновременного предоставления услуги многоадресной передачи нескольким абонентам из группы многоадресной доставки информации достаточно организовать одну многоадресную сессию. Под сессией понимается процесс непрерывной передачи данных от БС, инициированный запросом абонента на предоставление соответствующей услуги и завершающийся освобождением выделенного на БС радиоресурса [4]. По сравнению с проводными сетями в беспроводных системах одновременная передача данных нескольким абонентским устройствам (АУ) затруднена не только из-за отличающихся условий распространения радиосигнала до разных АУ, но и вследствие использования ФАР, формирующих несколько высоконаправленных лучей, что приводит к невозможности обслуживания всех АУ из группы многоадресной доставки одним лучом. Преимущество высоконаправленных лучей при этом заключается в увеличении радиуса покрываемого лучом сектора, что позволяет размещать БС на большем расстоянии друг от друга и снижает капительные затраты оператора сети. Таким образом, возникает задача исследования особенностей обслуживания многоадресного трафика в сетях 5G NR с точки зрения баланса для оператора сети между эффективностью использования радиоресурсов, качеством предоставления услуг и затратами на развертывание сети.

2 Формализация модели

Рассмотрим сценарий с одной БС, предоставляющей $M = |\mathcal{M}|$ услуг, требующих многоадресную, и $K = |\mathcal{K}|$ услуг, требующих одноадресную

*Исследование выполнено за счет Российского научного фонда (грант № 21-79-00142).

[1]Российский университет дружбы народов, samuylov-ak@rudn.ru

[2]Российский университет дружбы народов, platonova-aa@rudn.ru

[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, vshorgin@ipiran.ru

[4]Российский университет дружбы народов; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, gaydamaka-yuv@rudn.ru

доставку информации. Абонент формирует запрос на услугу многоадресной передачи с вероятностью p_M и на услугу одноадресной передачи с вероятностью p_U, $p_M + p_U = 1$. Назовем (I, m)-сессией многоадресную сессию класса m, соответствующую услуге m, и будем считать, что поступивший от абонента запрос с вероятностью $p_{M,m}$ представляет собой запрос на предоставление услуги m, $m = 1, 2, \ldots, M$. Аналогично (II, k)-сессия — одноадресная сессия класса k, соответствующая услуге k и вероятности $p_{U,k}$, $k = 1, 2, \ldots, K$. При этом $\sum_{m=1}^{M} p_{M,m} = p_M$, $\sum_{k=1}^{K} p_{U,k} = p_U$.

Пусть λ_{UE} — интенсивность поступления запросов на установление сессии от абонента, запрашивающего одну из M многоадресных или одну из K одноадресных услуг. Тогда $\Lambda = \lambda_{UE} N_{UE}$ — общая интенсивность поступления запросов от N_{UE} абонентов в зоне покрытия БС. Для оценки числа N_{UE} при заданной плотности абонентов λ_B делаем предположение о равномерном распределении АУ в зоне покрытия антенны БС, обслуживающей сектор 120° с радиусом r [5]. В этом случае $N_{UE} = \lambda_B \pi r^2/3$, а интенсивности поступления на БС запросов на предоставление услуги класса m и класса k равны $\lambda_m = p_{M,m}\Lambda$ и $\nu_k = p_{U,k}\Lambda$ соответственно. Кроме этого, для всех классов заданы средние длительности предоставления услуги абоненту μ_m^{-1} и κ_k^{-1}, а также объем требуемого ресурса b_m и d_k в первичных ресурсных блоках (РБ), причем два последних параметра, а также общее число V доступных РБ на БС зависят от размера блока, спектральной эффективности, полосы пропускания БС и могут быть вычислены, как показано в [6].

Описанный сценарий моделируется мультисервисной системой массового обслуживания (СМО) с потерями [7], схема которой представлена на рис. 1. Система позволяет моделировать предоставление двух типов услуг многоадресной передачи: услуг передачи хранимых данных (например, мультимедийное оповещение для целей общественной безопасности, массовые обновления программного обеспечения и прошивки в интернете вещей) и услуг передачи данных в реальном времени (например, средства массовой информации, развлечения, включая вещание в дополненной и виртуальной реальности, создание профессионального контента в беспроводной студии с несколькими камерами, иммерсивное видеопроизводство в реальном времени) [8–10]. Ключевое отличие заключается в длительности сессии для предоставления многоадресной услуги на стороне БС, т. е. интервала занятости ресурса, выделенного на БС для непрерывного предоставления услуги абонентам из группы многоадресной доставки информации. Продолжительность интервала занятости отражает особенности многоадресной передачи хранимых данных и данных в реальном времени и моделируется в СМО двумя дисциплинами «прозрачного» обслуживания [11] — Tr_1 для многоадресной передачи хранимых данных и Tr_2 для многоадресной передачи данных в реальном времени. Изменение числа абонентов $\xi^{(1)}(t)$ для Tr_1 и $\xi^{(2)}(t)$ для Tr_2, обслуживаемых в течение одной многоадресной сессии, показано на рис. 2.

Для обеих дисциплин запрос абонента на предоставление услуги класса (I, m) принимается к обслуживанию, когда при поступлении он не находит на обслуживании запросов такого же класса и свободны не менее b_m РБ. В этом случае (I, m)-запрос инициирует интервал занятости ресурса запросами класса (I, m) и занимает b_m РБ в течение случайного интервала времени, не зависящего от моментов поступления и длительностей обслуживания многоадресных и одноадресных запросов других классов. Все (I, m)-запросы, поступившие в течение этого интервала занятости, т. е. в период, когда на обслуживании находится один или несколько запросов класса (I, m), принимают-

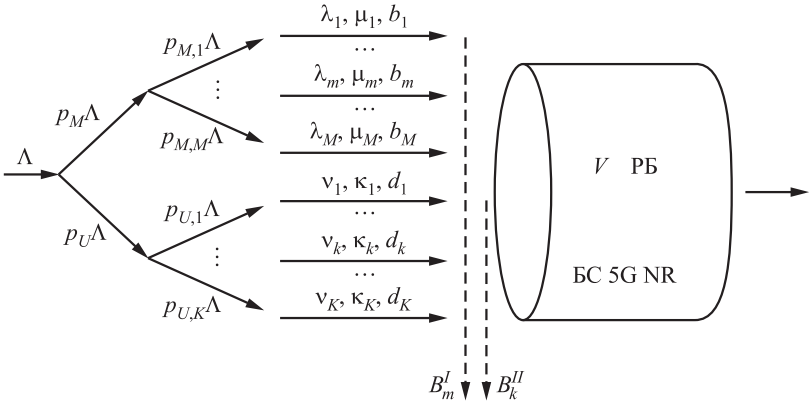


Рис. 1 Схема СМО для модели совместного обслуживания многоадресного и одноадресного трафика в сетях 5G/6G

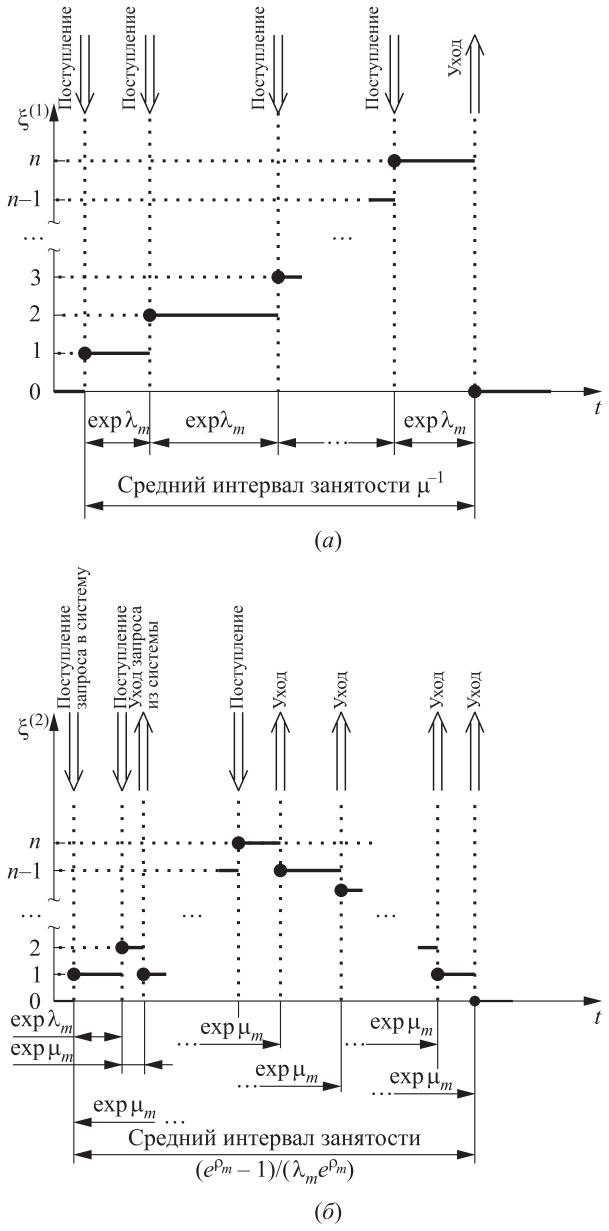


Рис. 2 Интервалы занятости при передаче хранимых данных (Tr_1) (*а*) и при передаче данных в реальном времени (Tr_2) (*б*)

ся в систему и получают обслуживание без выделения дополнительных РБ. Для обеих дисциплин доступ в систему (I,m)-запроса блокируется, если в системе нет запросов этого класса, а свободного ресурса недостаточно для начала этой сессии.

Существует принципиальное различие в моделировании момента окончания интервала занятости для дисциплин Tr_1 и Tr_2. Для дисциплины Tr_1 все обслуживающиеся (I,m)-запросы покидают систему одновременно в момент окончания обслуживания первого запроса, инициировавшего период занятости [12]. Для дисциплины Tr_2 обслуживающиеся (I,m)-запросы покидают систему в произвольные моменты, а интервал занятости заканчивается, когда все выделенные РБ освобождаются при окончании обслуживания последнего (I,m)-запроса, не оставившего после себя в системе запросов того же класса [13]. Под нижними горизонтальными стрелками на схемах рис. 2 указаны средние значения длительности интервала занятости для обеих дисциплин. Здесь $\rho_m = (1+\rho_{UE,m})^{p_{M,m}N_{UE}} - 1$ соответствует предлагаемой нагрузке запросами класса (I,m) от всех АУ в секторе покрытия антенны, при этом $\rho_{UE,m} = p_{M,m}\lambda_{UE}/\mu_m$ — предлагаемая нагрузка запросами класса (I,m) от одного абонента, $m \in \mathcal{M}$.

Обслуживание запросов абонентов на предоставление одноадресной услуги класса (II,k) моделируется классической СМО с потерями [14], для которой задана интенсивность ν_k входящего пуассоновского потока, экспоненциальное случайное время обслуживания запроса κ_k^{-1} и требование к ресурсу для обслуживания запроса d_k РБ. Предлагаемая нагрузка класса (II,k) обозначена $a_k = \nu_k/\kappa_k$, $k \in \mathcal{K}$.

3 Метрики качества предоставления услуг

Функционирование СМО описывается марковским процессом (МП)

$$\mathbf{Z}(t) = (Y_1(t),\ldots,Y_M(t),N_1(t),\ldots,N_K(t))\,,\ \ t \ge 0\,.$$

Здесь $Y_m(t)$, $t \ge 0$, $m \in \mathcal{M}$, — индикатор наличия в системе (I,m)-запросов в момент t:

$$Y_m(t) = \begin{cases} 1, & \text{если в момент } t \text{ обслуживается} \\ & \text{хотя бы один } (I,m)\text{-запрос;} \\ 0 & \text{иначе.} \end{cases}$$

Марковский процесс $N_k(t)$, $t \ge 0$, отражает число (II,k)-запросов в момент t, $N_k(t) \in \{0,1,2,\ldots,\lfloor V/d_k \rfloor\}$. Состояние МП $\mathbf{Z}(t)$ имеет вид $(\mathbf{y},\mathbf{n})$, а пространство состояний определяется как

$$\mathcal{Z} = \Bigg\{ (\mathbf{y},\mathbf{n}) \in \{0,1\}^M \times \{0,1,2,\ldots\}^K : \sum_{m=1}^{M} b_m y_m + \sum_{k=1}^{K} d_k n_k \le V \Bigg\}.$$

В [15] показано, что стационарное распределение МП $\mathbf{Z}(t)$ имеет мультипликативное представление

$$p(\mathbf{y},\mathbf{n}) = G^{-1}(\mathcal{Z}) \prod_{m=1}^{M} \gamma_m^{y_m} \prod_{k=1}^{k} \frac{a_k^{n_k}}{n_k!}, (\mathbf{y},\mathbf{n}) \in \mathcal{Z}, \quad (1)$$

с нормировочной константой $G(\mathcal{Z})$, представленной в виде

$$G(\mathcal{Z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{m=1}^{M} \gamma_m^{y_m} \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}.$$

Здесь γ_m имеет смысл загрузки БС многоадресными сессиями, которая различается для услуг передачи хранимых данных и услуг передачи данных в реальном времени и определяется как

$$\gamma_m = \begin{cases} \rho_m & \text{для } \mathrm{Tr}_1; \\ e^{\rho_m} - 1 & \text{для } \mathrm{Tr}_2. \end{cases}$$

Основные характеристики качества предоставления услуг можно найти суммированием стационарных вероятностей (1) по соответствующим подмножествам Ω пространства состояний $\mathcal{Z}$:

$$\mathsf{P}\{(\mathbf{y},\mathbf{n}) \in \Omega\} = \sum_{(\mathbf{y},\mathbf{n})\in\Omega} p(\mathbf{y},\mathbf{n}) = \frac{G(\Omega)}{G(\mathcal{Z})}, \quad \Omega \subseteq \mathcal{Z}. \quad (2)$$

Так, множество состояний блокировки доступа в систему для многоадресного запроса класса (I,m) определяется выражением

$$\mathcal{B}_m^I = \left\{ (\mathbf{y},\mathbf{n}) \in \mathcal{Z} : \sum_{m=1}^{M} b_m y_m + \sum_{k=1}^{K} d_k n_k + b_m > V, \; y_m = 0 \right\}, \quad m \in \mathcal{M},$$

а для одноадресного запроса класса (II,k) — выражением

$$\mathcal{B}_k^{II} = \left\{ (\mathbf{y},\mathbf{n}) \in \mathcal{Z} : \sum_{m=1}^{M} b_m y_m + \sum_{k=1}^{K} d_k n_k + d_k > V \right\}, \quad k \in \mathcal{K}.$$

Соответствующие вероятности $B_m^I = \mathsf{P}\{(\mathbf{y},\mathbf{n}) \in \mathcal{B}_m^I\}$ и $B_k^{II} = \mathsf{P}\{(\mathbf{y},\mathbf{n}) \in \mathcal{B}_k^{II}\}$ можно найти с помощью (2).

4 Результаты численного эксперимента

Для иллюстрации эффектов обслуживания многоадресного трафика при использовании частот миллиметрового диапазона длин волн и субтерагерцевого диапазона частот рассмотрена сота сети 5G NR с техническими параметрами из [16], радиус которой в зависимости от ФАР (от 4×4 до 32×4) может варьироваться от 107 до 288 м. Абоненты получают одну одноадресную услугу и одну услугу многоадресной передачи данных в реальном времени, при этом для последней высокая направленность лучей в технологии 5G NR может привести к необходимости поддерживать одновременно несколько многоадресных сессий, число которых ограничено числом антенных элементов ФАР.

На рис. 3 приведены графики вероятности блокировки доступа в систему для многоадресного и одноадресного трафика при $p_M = 0{,}5$ в зависимости от плотности λ_B абонентов в зоне покрытия БС для 4 значений расстояния между БС (ISD, inter-site distance).

Заметим, что увеличение интенсивности поступления запросов от абонентов из зоны покрытия сначала приводит к увеличению вероятности блокировки доступа в систему запросов обоих классов. Однако начиная с определенной интенсивности вероятность блокировки многоадресных запросов начинает уменьшаться, как показано на рис. 3, *а*. Объяснение заключается в том, что возрастает вероятность обслуживания в системе

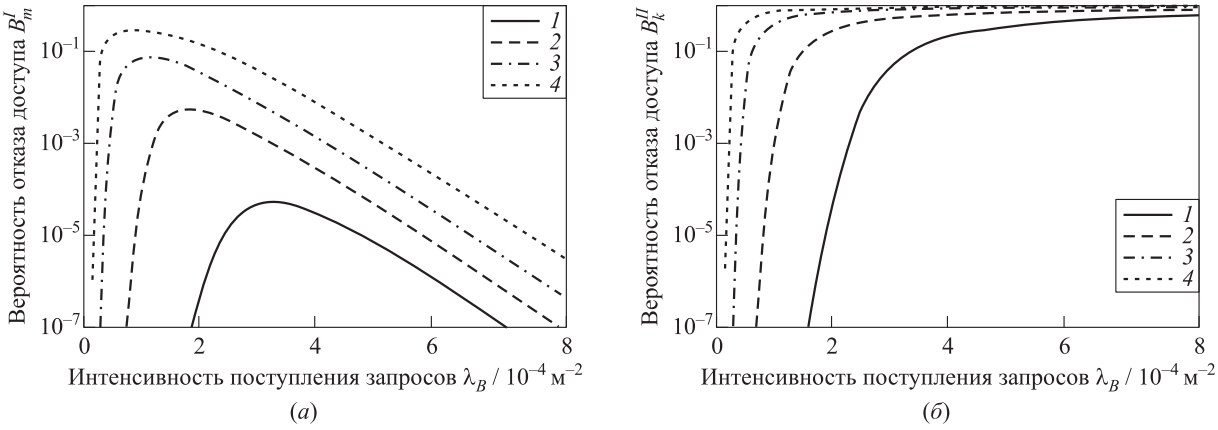


Рис. 3 Вероятность блокировки доступа в систему многоадресных (*а*) и одноадресных (*б*) запросов в зависимости от ISD: *1* — ISD = 321; *2* — 446; *3* — 620; *4* — ISD = 863

многоадресной сессии, которая занимает ресурс, в том числе для обслуживания в рамках текущей сессии всех следующих поступающих запросов на эту многоадресную услугу. Вышеупомянутый эффект приводит к резкому увеличению вероятности блокировки доступа для одноадресных запросов, как показано на рис. 3, *б*, вплоть до момента, когда система начинает обслуживать почти исключительно многоадресные запросы. Подобный эффект обычно наблюдается, когда предлагаемая нагрузка многоадресных запросов увеличивается или когда падает предлагаемая нагрузка одноадресных запросов, и усиливается для ФАР с большим числом элементов, соответствующих большим значениям расстояния между БС соседних сот сети.

Проиллюстрированная неявная приоритизация многоадресных сессий не всегда может быть предпочтительной для оператора сети, поскольку она может блокировать запросы на одноадресные услуги с более высоким приоритетом. Однако фактический баланс между вероятностями блокировки доступа одноадресных и многоадресных запросов существенно зависит от функции полезности сетевого оператора. Последний может обеспечить соблюдение необходимого баланса, введя, например, приоритеты на этапе приема запросов на обслуживание.

5 Заключение

В работе предложена математическая модель обслуживания многоадресного и одноадресного трафика на БС в системах 5G/6G при использовании высоконаправленных антенных решеток, характерных для сетей 5G NR. Показано, что совместная передача многоадресного и одноадресного трафика на радиоинтерфейсе приводит к ряду эффектов, связанных с использованием ресурсов этими классами трафика. Предложенная модель позволяет анализировать области значений параметров технической системы для эффективного применения ФАР при обслуживании многоадресного трафика, в том числе оценивать ограничения на расстояние между БС в таких сетях.

Литература

1. *David K., Berndt H.* 6G vision and requirements: Is there any need for beyond 5G? // IEEE Veh. Technol. Mag., 2018. Vol. 13. Iss. 3. P. 72–80. doi: 10.1109/MVT.2018.2848498.
2. *Petrov V., Kurner T., Hosako I.* IEEE 802.15. 3d: First standardization efforts for sub-terahertz band communications toward 6G // IEEE Commun. Mag., 2020. Vol. 58. Iss. 11. P. 28–33. doi: 10.1109/MCOM.001.2000273.
3. *Kompella V. P., Pasquale J. C., Polyzos G. C.* Multicasting for multimedia applications // Conference on Computer Communications. — IEEE, 1992. P. 2078–2085. doi: 10.1109/INFCOM.1992.263480.
4. Multimedia Broadcast/Multicast Service (MBMS); Stage 1 (Release 16): Technical Specification 22.146 V16.0.0. — 3GPP, 2020. https://www.3gpp.org/ftp/Specs/archive/22_series/22.146/22146-g00.zip.
5. *Moltchanov D.* Distance distributions in random networks // Ad Hoc Netw., 2012. Vol. 10. Iss. 6. P. 1146–1166. doi: 10.1016/j.adhoc.2012.02.005.
6. *Kovalchukov R., Moltchanov D., Gaidamaka Y., Bobrikova E.* An accurate approximation of resource request distributions in millimeter wave 3GPP New Radio systems // Internet of things, smart spaces, and next generation networks and systems / Eds. O. Galinina, S. Andreev, S. Balandin, Y. Koucheryavy. — Lecture notes in computer science ser. — Springer, 2019. Vol. 11660. P. 572–585. doi: 10.1007/978-3-030-30859-9_50.
7. *Basharin G., Gaidamaka Y., Samouylov K.* Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks // Autom. Control Comp. S., 2013. Vol. 47. P. 62–69. doi: 10.3103/S0146411613020028.
8. *Araniti G., Condoluci M., Scopelliti P., Molinaro A., Iera A.* Multicasting over emerging 5G Networks: Challenges and perspectives // IEEE Network, 2017. Vol. 31. No. 2. P. 80–89. doi: 10.1109/MNET.2017.1600067NM.
9. Study on architectural enhancements for 5G multicast-broadcast services (Release 17): Technical Report 23.757 V1.2.0. — 3GPP, 2020. https://www.3gpp.org/ftp/Specs/archive/23_series/23.757/23757-120.zip.
10. *Tran T., Navrátil D., Sanders P., Hart J., Odarchenko R., Barjau C., Altman B., Burdinat C., Gomez-Barquero D.* Enabling multicast and broadcast in the 5G core for converged fixed and mobile networks // IEEE T. Broadcast., 2020. Vol. 66. No. 2. P. 428–439. doi: 10.1109/TBC.2020.2991548.
11. *Рыков В. В., Самуйлов К. Е.* К анализу вероятностей блокировок ресурсов сети с динамическими многоадресными соединениями // Электросвязь, 2000. № 10. С. 27–30.
12. *Karvo J., Martikainen O., Virtamo J., Aalto S.* Blocking of dynamic multicast connections // Telecommun. Syst., 2001. Vol. 16. P. 467–481. doi: 10.1023/A:1016631431617.
13. *Boussetta K., Belyot A.-L.* Multirate resource sharing for unicast and multicast connections // Broadband communications / Eds. D. H. K. Tsang, P. J. Kühn. — Boston, MA, USA: Springer, 2000. Vol. 30. P. 561–570. doi: 10.1007/978-0-387-35579-5_47.

14. *Kelly F. P.* Loss networks // Ann. Appl. Probab., 1991. Vol. 1. Iss. 3. P. 319– 378. doi: 10.1214/aoap/1177005872.
15. *Naumov V., Gaidamaka Y., Yarkina N., Samouylov K.* Matrix and analytical methods for performance analysis of telecommunication systems. — Springer Nature, 2021. 308 p.
16. *Samuylov A., Moltchanov D., Kovalchukov R., Pirmagomedov R., Gaidamaka Y., Andreev S., Koucheryavy Y., Samouylov K.* Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic // IEEE T. Wirel. Commun., 2020. Vol. 19. No. 5. P. 3421–3434. doi: 10.1109/TWC.2020.2973375.

Поступила в редакцию 15.04.23

ON MODELING THE EFFECTS OF MULTICAST TRAFFIC SERVICING IN 5G NR NETWORKS

A. K. Samouylov[1], A. A. Platonova[1], V. S. Shorgin[2], and Yu. V. Gaidamaka[1,2]

[1]RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation
[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: Multicasting in wireless access networks allows efficient provision of a service to a group of subscribers and is useful for reducing the resource required to serve user equipments requesting the same data. The support of this feature in current 5G New Radio (NR) technology and future subterahertz (sub-THz) 6G systems faces challenges associated with the use of the directional beamforming phased array antennas. The presented multicast and unicast traffic service model allows one to explore the range of 5G/6G network parameters to reduce the density of base stations while maintaining the quality of services provided to subscribers.

Keywords: 5G; 6G; multicasting; millimeter wave; terahertz; multibeam antennas; multi-RAT; numerical simulation

DOI: 10.14357/19922264230210 **EDN:** SLMGZU

Acknowledgments

The reported study was funded by the Russian Science Foundation, project No. 21-79-00142.

References

1. David, K., and H. Berndt. 2018. 6G vision and requirements: Is there any need for beyond 5G? *IEEE Veh. Technol. Mag.* 13(3):72–80. doi: 10.1109/MVT.2018.2848498.
2. Petrov, V., T. Kurner, and I. Hosako. 2020. IEEE 802.15. 3d: First standardization efforts for sub-terahertz band communications toward 6G. *IEEE Commun. Mag.* 58(11):28–33. doi: 10.1109/MCOM.001.2000273.
3. Kompella, V. P., J. C. Pasquale, and G. C. Polyzos. 1992. Multicasting for multimedia applications. *Conference on Computer Communications*. IEEE. 2078–2085. doi: 10.1109/INFCOM.1992.263480.
4. Multimedia broadcast/multicast service (MBMS); Stage 1 (Release 16): Technical specification 22.146 V16.0.0. 3GPP. Available at: https://www.3gpp.org/ftp/Specs/archive/22_series/22.146/22146-g00.zip (accessed May 20, 2023).
5. Moltchanov, D. 2012. Distance distributions in random networks. *AD Hoc Netw.* 10(6):1146–1166. doi: 10.1016/j.adhoc.2012.02.005.
6. Kovalchukov, R., D. Moltchanov, Y. Gaidamaka, and E. Bobrikova. 2019. An accurate approximation of resource request distributions in millimeter wave 3GPP New Radio systems. *Internet of things, smart spaces, and next generation networks and systems*. Eds. O. Galinina, S. Andreev, S. Balandin, and Y. Koucheryavy. Lectures notes in computer science ser. Springer. 11660:572–585. doi: 10.1007/978-3-030-30859-9_50.
7. Basharin, G., Y. Gaidamaka, and K. Samouylov. 2013. Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks. *Autom. Control Comp. S.* 47:62–69. doi: 10.3103/S0146411613020028.
8. Araniti, G., M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera. 2017. Multicasting over emerging 5G networks: Challenges and perspectives. *IEEE Network* 31(2):80–89. doi: 10.1109/MNET.2017.1600067NM.
9. Study on architectural enhancements for 5G multicast-broadcast services (Release 17): Technical Report 23.757 V1.2.0. 3GPP. Available at: https://www.3gpp.org/ftp/Specs/archive/23_series/23.757/23757-120.zip (accessed May 20, 2023).
10. Tran, T., D. Navátil, P. Sanders, J. Hart, R. Odarchenko, C. Barjau, B. Altman, C. Burdinat, and D. Gomez-Barquero. 2020. Enabling multicast and broadcast in

the 5G core for converged fixed and mobile networks. *IEEE T. Broadcast.* 66(2):428–439. doi: 10.1109/TBC.2020.2991548.

11. Rykov, V. V., and K. E. Samuylov. 2000. K analizu veroyatnostey blokirovok resursov seti s dinamicheskimi mnogoadresnymi soedineniyami [To the analysis of blocking probabilities in a network with dynamic multicast connections]. *Elektrosvyaz'* [Electrosvyaz Magazine] 10: 27–30.
12. Karvo, J., O. Martikainen, J. Virtamo, and S. Aalto. 2001. Blocking of dynamic multicast connections. *Telecommun. Syst.* 16:467–481. doi: 10.1023/A:1016631431617.
13. Boussetta, K., and A.-L. Belyot. 2000. Multirate resource sharing for unicast and multicast connections. *Broadband communications*. Eds. D. H. K. Tsang and P. J. Kühn. Boston, MA: Springer. 30:561–570. doi: 10.1007/978-0-387-35579-5_47.
14. Kelly, F. P. 1991. Loss networks. *Ann. Appl. Probab.* 1(3):319–378. doi: 10.1214/aoap/1177005872.
15. Naumov, V., Y. Gaidamaka, N. Yarkina, and K. Samouylov. 2021. *Matrix and analytical methods for performance analysis of telecommunication systems*. Springer Nature. 308 p.
16. Samuylov, A., D. Moltchanov, R. Kovalchukov, R. Pirmagomedov, Y. Gaidamaka, S. Andreev, Y. Koucheryavy, and K. Samouylov. 2020. Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic. *IEEE T. Wirel. Commun.* 19(5):3421–3434. doi: 10.1109/TWC.2020.2973375.

Received April 15, 2023

Contributors

Samuylov Andrey K. (b. 1988) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Applied Probability and Informatics, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; samuylov-ak@rudn.ru

Platonova Anna A. (b. 1996) — PhD student, Department of Applied Probability and Informatics, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; platonova-aa@rudn.ru

Shorgin Vsevolod S. (b. 1978) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; vshorgin@ipiran.ru

Gaidamaka Yuliya V. (b. 1971) — Doctor of Science in physics and mathematics, professor, Department of Applied Probability and Informatics, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gaydamaka-yuv@rudn.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2023. Т. 17. Вып. 2. С. 78–83

САМООБУЧЕНИЕ АВТОНОМНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ РОБОТОВ В ПРОЦЕССЕ ПОИСКОВО-ИССЛЕДОВАТЕЛЬСКОЙ ДЕЯТЕЛЬНОСТИ*

В. Б. Мелехин[1], В. М. Хачумов[2], М. В. Хачумов[3]

Аннотация: Рассматривается один из эффективных подходов к организации целесообразного поведения автономных интегральных роботов (АИР) в процессе поисково-исследовательской деятельности в априори неописанных условиях проблемной среды (ПС). Предлагается в основе целесообразного поведения роботов использовать процедуры наглядно-действенного мышления, основанные на формализации рефлекторного поведения высокоорганизованных живых систем. Разработан алгоритм самообучения в условиях с высоким уровнем неопределенности, позволяющий автоматически формировать условные программы целесообразного поведения, обеспечивающие АИР возможность достигать заданной цели поведения в процессе поисково-исследовательской деятельности. Найдены граничные оценки функциональной сложности предложенного алгоритма самообучения в условиях неопределенности, показывающие возможность его реализации на бортовой ЭВМ автономных интегральных роботов, имеющих, как правило, ограниченные вычислительные ресурсы. Проведено имитационное моделирование процесса самообучения АИР в априори неописанной ПС, подтвердившее эффективность применения предложенного подхода для организации планирования целесообразного поведения в априори неописанных ПС.

Ключевые слова: автономный интегральный робот; алгоритм самообучения; условия неопределенности; проблемная среда; условные сигналы

DOI: 10.14357/19922264230211 **EDN:** SOFDKW

1 Введение

Разработка информационных технологий, связанных с построением интеллектуального решателя задач АИР, способных целесообразно (рационально) функционировать в априори неописанных ПС, — актуальная и сложная проблема искусственного интеллекта. К одному из эффективных подходов решения данной проблемы следует отнести разработку когнитивных инструментов наглядно-действенного мышления интеллектуальных систем различного назначения [1]. В общем случае наглядно-действенное мышление АИР строится на основе формализации рефлекторного поведения живых систем [2, 3] и включает следующие три основные составляющие [4].

1. Самообучение на основе выполнения пробных действий и механизмов избирательности поступающей из ПС информации, обеспечивающих возможность поиска заданных объектов в априори неописанных условиях функционирования.

Организовать самообучение АИР в априори неописанных условиях ПС можно на основе, например, алгоритмов роевого поведения [5, 6] или генетических алгоритмов [7, 8]. Однако непосредственная отработка пробных действий может привести к негативным изменениям, не связанным с достижением заданной цели. Обойти этот недостаток можно на основе алгоритмов самообучения, которые имитируют выполнение пробных действий на формальном описании текущей ситуации ПС.

Процесс самообучения АИР в априори неописанной ПС сводится к формированию и закреплению элементарных актов поведения в формируемых условных программах целесообразной деятельности (УПЦД) по новизне происходящих в ПС изменений, а для всей автоматически построенной упорядоченной последовательности действий характеризуется достижением заданного безусловного сигнала.

В формируемых в процессе самообучения УПЦД запоминаются происходящие в ПС из-

*Исследование выполнено при поддержке Российского научного фонда (проект 21-71-10056).

[1]Дагестанский государственный технический университет, pashka1602@rambler.ru

[2]Институт программных систем им. А. К. Айламазяна Российской академии наук; Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Российский университет дружбы народов, vmh48@mail.ru

[3]Институт программных систем им. А. К. Айламазяна Российской академии наук; Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Российский университет дружбы народов, khmike@inbox.ru

менения в форме сигналов, которые возникают в результате отрабатываемых АИР действий. Различные сигналы ПС в процессе самообучения приобретают роль условных сигналов — знаков, вызывающих у АИР определенные реакции, связанные с отработкой закрепленных в УПЦД действий. Таким образом, условные сигналы после закрепления в УПЦД приобретают роль ориентиров или предвестников, появление которых в ПС сигнализирует АИР о возможности достижения в ней соответствующего безусловного сигнала.

2. Целесообразное поведение АИР, связанное с отработкой в текущих условиях функционирования действий, ранее сформированных УПЦП для достижения соответствующих им безусловных сигналов при восприятии в ПС закрепленных в этих программах условных сигналов.
3. Отработка безусловных реакций для достижения заданной цели при появлении в ПС соответствующих им безусловных сигналов.

В настоящей статье предлагаются процедуры самообучения в процессе поисково-исследовательской деятельности, позволяющие АИР организовать целесообразное поведение в априори неописанной ПС с препятствиями для поиска заданных объектов. Например, при выполнении различных спасательных работ в труднодоступных для человека условиях функционирования.

2 Постановка задачи

Рассмотрим АИР, оснащенный техническим зрением, манипулятором и моторной системой, позволяющей ему перемещаться в ПС. Проблемная среда представляет собой пересеченную местность с расположенными на ее территории препятствиями и различными объектами $O = \{o_{i_1}(X_{i_1})\}$, $i_1 = \overline{1, n_1}$, где X_{i_1} — множество характеристик, по которым робот способен идентифицировать воспринимаемые в ПС препятствия и объекты.

Проблемную среду можно охарактеризовать множеством условных сигналов $A = \{a_{i_2}\}$, $i_2 = \overline{1, n_2}$, каждый из которых представляет собой проходимый между препятствиями участок ПС.

В общем случае АИР способен отрабатывать множество действий $B = \{b_{i_3}\}$, $i_3 = \overline{1, n_3}$, и распознавать проходимые участки местности, а также найденные в ПС объекты $o_{i_1}(X_{i_1}) \in O$. (Следует отметить, что проблема распознавания проходимых участков и объектов ПС является самостоятельной задачей и в настоящей статье не рассматривается.)

Требуется разработать алгоритм самообучения АИР в процессе поисково-исследовательской деятельности, позволяющий автоматически формировать в априори неописанной проблемной среде УПЦП следующего вида:

$$a^1_{i_2}\&b^1_{i_3} \to a^2_{i_2}\&b^2_{i_3} \to \cdots \to a^k_{i_2}\&b^k_{i_3} \to a^P_{i_2}\,, \qquad (1)$$

где $a^1_{i_2}\&b^1_{i_3} \to a^2_{i_2}$ — элементарный акт поведения, означающий, что если АИР воспринимает в ПС условный сигнал $a^1_{i_2}$, то отрабатываемое им действие $b^1_{i_3}$ приводит к появлению условного сигнала $a^2_{i_2}$; $a^1_{i_2}$ и $a^2_{i_2}$ — условные сигналы ПС, определяющиеся проходимыми для АИР участками ПС; $a^P_{i_2}$ — безусловный сигнал, вызывающий у АИР соответствующие безусловные реакции, связанные с выполнением определенных действий над найденным объектом.

Требуется разработать алгоритм самообучения, позволяющий АИР в процессе поисково-исследовательской деятельности формировать УПЦД в виде простой цепи (1) в априори неописанных ПС.

3 Синтез алгоритма самообучения автономных интегральных роботов

В общем случае алгоритм самообучения АИР в процессе поисково-исследовательской деятельности опирается на закрепление в формируемой УПЦД элементарных актов поведения по новизне условных сигналов, воспринимаемых в ПС. Вся же полученная таким образом цепь действий закрепляется достижением в ПС заданного безусловного сигнала. Роль безусловного сигнала в этом случае играет заданный объект, воспринятый в ПС после выхода АИР за пределы последнего закрепленного в УПЦД условного сигнала. Данный алгоритм самообучения АИР имеет следующее структурированное описание.

Исходные условия: заданный АИР объект ПС $o_{i_1}(X_{i_1}) \in O$, выполняющий роль безусловного сигнала $a^P_{i_2}$; множество действий B, которые способен отрабатывать АИР.

Входные переменные: условные сигналы $a_{i_2} \in A$ и заданный $a^P_{i_2}$ безусловный сигнал ПС.

Выходные переменные: формируемые УПЦД в виде простой цепи.

Начало.

1. Установить $j_1 = 1$. Определить в качестве исходного сигнала $a^{j_1}_{i_2}$ в формируемой УПЦД непосредственно воспринимаемое роботом в ПС препятствие.

2. Принять в качестве подцели поведения на текущем шаге самообучения появление в ПС нового условного сигнала $a_{i_2}^{j_1+1}$, определяемого воспринятым после отработки пробного действия новым проходимым участком.
3. Определить на текущем шаге самообучения согласно равномерному закону распределения вероятностей выбора пробное действие $b_{i_3}^{j_1} \in B$. Выполнить выбранное действие $b_{i_3}^{j_1}$ и сформировать по результатам его отработки элементарный акт поведения $a_{i_2}^{j_1} \& b_{i_3}^{j_1} \to a_{i_2}^{j_1+1}$.
4. Проверить условие «условный сигнал $a_{i_2}^{j_1+1}$ был ранее закреплен в формируемой УПЦД»: если условие выполняется, то перейти к п. 5; в противном случае перейти к п. 8.
5. Удалить все элементарные акты поведения, закрепленные в УПЦД после первого восприятия АИР в ПС условного сигнала $a_{i_2}^{j_1+1}$.
6. Исключить выбранное пробное действие $b_{i_3}^{j_1}$ из множества B как нерезультативное на текущем шаге самообучения.
7. Проверить условие «множество B является пустым»: если условие выполняется, то перейти к п. 11; в противном случае перейти к п. 3.
8. Сохранить элементарный акт поведения $a_{i_2}^{j_1} \& b_{i_3}^{j_1} \to a_{i_2}^{j_1+1}$ в формируемой УПЦД.
9. Проверить условие «после выхода за пределы проходимого участка, определяемого условным сигналом $a_{i_2}^{j_1+1}$, АИР воспринимает в ПС заданный безусловный сигнал $a_{i_2}^{P}$»: если условие выполняется, перейти к п. 12; в противном случае перейти к п. 10.
10. Восстановить все исключенные действия из заданного множества B, $j_1 = j_1 + 1$, перейти к п. 2.
11. Сформировать требуемую УПЦД в текущих условиях ПС не представляется возможным, перейти к п. 13.
12. Требуемая УПЦД сформирована; выполнить для достижения заданной цели безусловные реакции.

Конец.

Введем понятие функциональной сложности β алгоритма самообучения АИР, зависящей от общего числа действий $b_{i_3} \in B$, апробируемых роботом в процессе формирования УПЦД. Тогда для данного алгоритма можно доказать следующее утверждение.

Утверждение. *Функциональная сложность β алгоритма самообучения АИР определяется следующими граничными оценками:*

$$n_{10} \le \beta \le n_1 n_{10},$$

где n_{10} — общее число выполненных АИР шагов самообучения; n_1 — общее число различного вида действий, которые робот отрабатывает в процессе самообучения.

Доказательство. Справедливость сформулированного утверждения вытекает из следующих соображений.

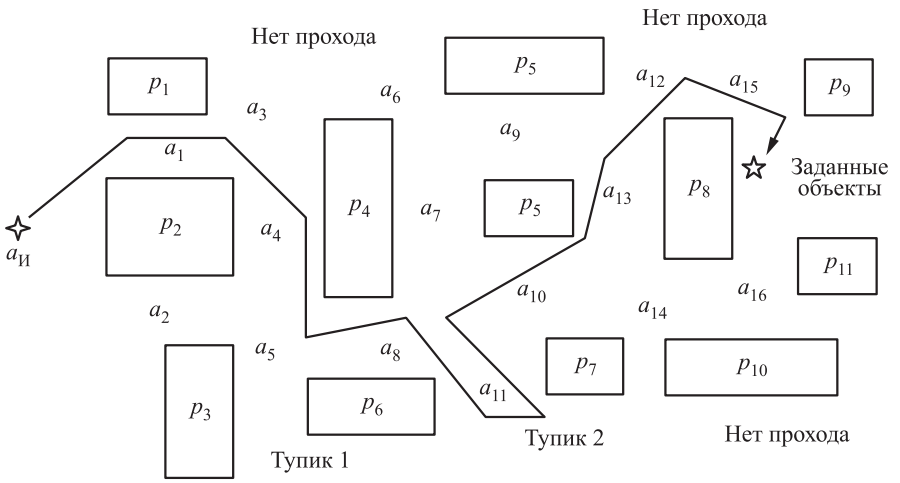
1. Согласно пп. 3–8 *алгоритма самообучения*, вполне вероятно, что в лучшем случае на каждом j_1-м шаге самообучения АИР первым случайным образом выбирает результативное действие $b_{i_3}^{j_1}$. Следовательно, нижнее граничное значение оценки сложности β_1 в этом случае определяется величиной, равной n_{10}.
2. В худшем случае результативное действие $b_{i_3}^{j_1}$ на каждом j_1-м шаге самообучения АИР может быть выбрано случайным образом в последнюю очередь. Отсюда следует, что на каждом шаге самообучения АИР апробирует отработку не более n_1 действий $b_k(j_1) \in B$. Таким образом, число выполнений пробных действий в процессе самообучения АИР не может превышать величины, равной $n_1 n_{10}$.
3. Из пп. 1 и 2 проведенного доказательства с очевидностью следует справедливость сформулированного утверждения.

Рассмотрим гипотетический пример, связанный с использованием АИР алгоритма самообучения для решения целевой задачи, когда у робота отсутствует формальное описание карты местности, а известны только границы участка ПС, на котором требуется найти заданные объекты.

4 Пример решения задачи в процессе поисково-исследовательской деятельности автономных интеллектуальных роботов

Пусть АИР требуется найти заданный объект в априори неописанной ПС, представляющей собой местность с расположенными на ней препятствиями, структура которой приведена на рисунке.

Таким образом, ПС характеризуется 11 расположенными в ней препятствиями $P = \{p_{i_6}\}$, $i_6 = \overline{1,11}$, и 16 проходимыми между препятствиями зонами, обозначенными сигналами $A = \{a_{i_7}\}$, $i_7 = \overline{1,16}$. В этой среде АИР требуется найти объекты,



Структура ПС с препятствиями и обозначенными исходными местоположениями АИР и заданных объектов

Закономерности перехода АИР из текущего положения в ПС к смежной проходимой зоне

	$a_{\text{И}}$	a_1	a_2	a_3	a_4	a_5	a_1^*	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_2^*	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
b_1	a_1	a_3	a_4	—	a_3	a_8	—	—	a_9	a_{10}	a_{12}	a_{13}	a_{10}	—	—	a_{12}	a_{16}	—	$a_{\text{Ц}}$
b_2	a_2	a_4	a_5	a_4	a_5	a_1^*	—	a_9	a_{10}	a_{11}	a_{13}	a_{14}	a_2^*	—	a_{15}	a_{14}	—	$a_{\text{Ц}}$	—
b_3	—	—	—	—	—	—	a_5	—	—	—	—	—	—	a_{11}	—	—	—	—	—

расположенные за препятствием $p_8 \in P$, при его исходном местоположении напротив препятствия $p_2 \in P$. Восприятие данного объекта в ПС соответствует достижению роботом заданной цели. Пусть АИР для решения поставленной перед ним задачи способен отрабатывать следующие три действия: b_1 — поворот влево и движение вперед до выхода за наблюдаемую в результате этого проходимую зону; b_2 — поворот вправо и движение вперед до выхода за наблюдаемую в результате этого проходимую зону; b_3 — безусловные реакции «разворот и выход из тупика в предыдущее исходное текущее местоположение».

При этом система технического зрения АИР способна распознавать и отличать друг от друга проходимые участки $a_{i_7} \in A$ ПС и заданный ему объект. Для имитации процесса поиска АИР цели в заданной ПС строится конечный автомат со случайными реакциями [4], в память которого занесена таблица команд, отражающая закономерности перехода АИР от одного проходимого участка ПС к другому такому участку (см. таблицу).

В таблице использованы следующие обозначения: $a_{\text{И}}$ — исходное местоположение АИР; a_1^* и a_2^* — соответственно первый и второй тупик; $a_{\text{Ц}}$ — местоположение заданных объектов; прочерк означает отсутствие прохода или выхода за пределы заданного участка местности.

По итогам проведенного на ПЭВМ эксперимента были получены следующие результаты. Автономный интеллектуальный робот, выполнив 12 пробных действий, прошел по следующему маршруту в процессе поиска заданных объектов (см. рисунок):

$$a_{\text{И}}\&b_1 \to a_1\&b_2 \to a_4\&b_2 \to a_5\&b_1 \to a_8\&b_2 \to$$
$$\to a_{11}\&b_2 \to a_2^*(\text{тупик 2})\&b_3 \to a_{11}\&b_1 \to$$
$$\to a_{10}\&b_1 \to a_{13}\&b_1 \to a_{12}\&b_2 \to a_{15}\&b_2 \to a_{\text{Ц}}.$$

При этом у АИР после обнуления в процессе самообучения цикла сформировалась следующая УПЦД:

$$a_{\text{И}}\&b_1 \to a_1\&b_2 \to a_4\&b_2 \to a_5\&b_1 \to a_8\&b_1 \to$$
$$\to a_{10}\&b_1 \to a_{13}\&b_1 \to a_{12}\&b_2 \to a_{15}\&b_2 \to a_{\text{Ц}}\,.$$

Данную УПЦД интеллектуальный робот может использовать, например, для перевозки большого числа заданных объектов на участок ПС, определяющий заданное их местоположение.

5 Заключение

1. Предложенный алгоритм самообучения позволяет организовать целесообразное поведение АИР в процессе поисково-исследовательской

деятельности в априори неописанных труднодоступных для человека ПС.

2. Найденные граничные оценки и результаты имитационного моделирования алгоритма самообучения показали эффективность его использования для проведения АИР поисково-исследовательской деятельности в априори неописанной ПС с препятствиями с целью поиска заданных объектов, например при выполнении различных спасательных работ в труднодоступных для человека условиях функционирования.

Литература

1. *Мелехин В. Б., Хачумов М. В.* Формы мышления автономных интеллектуальных агентов: особенности и проблемы их организации // Морские интеллектуальные технологии, 2020. № 4-1. С. 224–230. doi: 10.37220/MIT.2020.50.4.031.
2. *Брайнес С. Н., Напалков А. Н., Свечинский В. Б.* Нейрокибернетика. — М.: Госмедиздат, 1962. 172 с.
3. *Шингаров Г. Х.* Условные рефлексы и проблема знака и значения. — М.: Наука, 1986. 200 с.
4. *Мелехин В. Б., Хачумов М. В.* Инструментальные средства управления целесообразным поведением самоорганизующихся автономных интеллектуальных агентов // Мехатроника, автоматизация, управление, 2021. Т. 22. № 4. С. 171–180. doi: 10.17587/mau.22.171-180.
5. *Карпов В. Э., Карпова И. П., Кулинич А. А.* Социальные сообщества роботов. — М.: Ленанд, 2019. 352 с.
6. *Guan B., Xu T., Zhao Y., Li Y., Dong X.* A random grouping-based self-regulating artificial bee colony algorithm for interactive feature detection // Knowl.-Based Syst., 2021. Vol. 243. P. 1–12. doi: 10.1016/j.knosys.2022.108434.
7. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы. — М.: Горячая линия – Телеком, 2008. 452 с.
8. *Саймон Д.* Алгоритмы эволюционной оптимизации / Пер. с англ. — М.: ДМК Пресс, 2020. 940 с. (*Simon D.* Evolutionary optimization algorithms. — 1st ed. — New York, NY, USA: Wiley, 2013. 784 p.)

Поступила в редакцию 02.11.22

SELF-LEARNING OF AUTONOMOUS INTELLIGENT ROBOTS IN THE PROCESS OF SEARCH AND EXPLORE ACTIVITIES

V. B. Melekhin[1], V. M. Khachumov[2,3,4], and M. V. Khachumov[2,3,4]

[1]Dagestan State Technical University, 70A Imam Shamil Ave., Makhachkala 367015, Republic of Dagestan

[2]Ailamazyan Program Systems Institute of the Russian Academy of Sciences, 4A Petra Pervogo Str., Veskovo 152024, Yaroslavl Region, Russian Federation

[3]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[4]RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation

Abstract: One of the effective approaches to organizing the goal-seeking behavior of autonomous integral robots in the process of search and explore activities in an a priori undescribed conditions of a problematic environment is considered. It is proposed to use the procedures of visual-effective thinking based on the formalization of the reflex behavior of highly organized living systems as the basis for the goal-seeking behavior of robots. A self-learning algorithm has been developed for the conditions with a high level of uncertainty which allows automatically generating conditional programs of expedient behavior that provide autonomous integral robots with the ability to achieve a given behavioral goal in the process of search and explore activities. The boundary estimates of the functional complexity of the proposed self-learning algorithm under uncertainty are found showing the possibility of its implementation on the onboard computer of autonomous integral robots which have, as a rule, limited computing resources. A modeling of self-learning process for an autonomous integral robot in an a priori undescribed and problematic environment was carried out which confirmed the effectiveness of the proposed approach for organizing the planning of goal-seeking behavior in an a priori undescribed and problematic environments.

Keywords: autonomous integral robot; self-learning algorithm; uncertainty conditions; problematic environment; conditional signals

DOI: 10.14357/19922264230211 **EDN:** SOFDKW

Acknowledgments

This work was supported by the Russian Science Foundation, project No. 21-71-10056.

References

1. Melekhin, V. B., and M. V. Khachumov. 2020. Formy myshleniya avtonomnykh intellektual'nykh agentov: osobennosti i problemy ikh organizatsii [Forms of thinking of autonomous intelligent agents: Features and problems of their organization]. *Morskie intellektual'nye tekhnologii* [Marine Intelligent Technologies] 4-1:224–230. doi: 10.37220/MIT.2020.50.4.031.
2. Braynes, S. N., A. N. Napalkov, and V. B. Svechinskiy. 1962. *Neyrokibernetika* [Neurocybernetics]. Moscow: Gosmedizdat. 172 p.
3. Shingarov, G. Kh. 1986. *Uslovnye refleksy i problema znaka i znacheniya* [Conditioned reflexes and the problem of sign and meaning]. Moscow: Nauka. 200 p.
4. Melekhin, V. B., and M. V. Khachumov. 2021. Instrumental'nye sredstva upravleniya tselesoobraznym povedeniem samoorganizuyushchikhsya avtonomnykh intellektual'nykh agentov [Instrumental means for managing the rational behavior of self-organizing autonomous intelligent agents]. *Mekhatronika, avtomatizatsiya, upravlenie* [Mechanatronics, Automation, and Control] 4:171–180. doi: 10.17587/mau.22.171-180.
5. Karpov, V. E., I. P. Karpova, and A. A. Kulinich. 2019. *Sotsial'nye soobshchestva robotov* [Social communities of robots]. Moscow: LENAND. 352 p.
6. Guan, B., T. Xu, Y. Zhao, Y. Li, and X. Dong. 2021. A random grouping-based self- regulating artificial bee colony algorithm for interactive feature detection. *Knowl.-Based Syst.* 243:1–12. doi: 10.1016/j.knosys.2022.108434.
7. Rutkovskaya, D., M. Pilin'skiy, and L. Rutkovskiy. 2008. *Neyronnye seti, geneticheskie algoritmy i nechetkie sistemy* [Neural networks, genetic algorithms, and fuzzy systems]. Moscow: Goryachaya Liniya – Telekom. 452 p.
8. Simon, D. 2013. *Evolutionary optimization algorithms*. 1st ed. New York, NY: Wiley. 784 p.

Received November 2, 2022

Contributors

Melekhin Vladimir B. (b. 1954) — Doctor of Science in technology, professor, Department of Software for Computers and Automated Systems, Dagestan State Technical University, 70A Imam Shamil Ave., Makhachkala 367015, Republic of Dagestan; pashka1602@rambler.ru

Khachumov Vyacheslav M. (b. 1948) — Doctor of Science in technology, head of laboratory, Intelligent Control Laboratory, Ailamazyan Program Systems Institute of the Russian Academy of Sciences, 4A Petra Pervogo Str., Veskovo 152024, Yaroslavl Region, Russian Federation; principal scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Department of Information Technology, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; vmh48@mail.ru

Khachumov Mikhail V. (b. 1986) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Intelligent Control Laboratory, Ailamazyan Program Systems Institute of the Russian Academy of Sciences, 4A Petra Pervogo Atr., Veskovo 152024, Yaroslavl Region, Russian Federation; senior scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Department of Information Technology, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; khmike@inbox.ru

СЛОЖНЫЕ ПРИЧИННО-СЛЕДСТВЕННЫЕ СВЯЗИ

А. А. Грушо[1], Н. А. Грушо[2], М. И. Забежайло[3], Е. Е. Тимонина[4], С. Я. Шоргин[5]

Аннотация: Рассматривается задача построения логического вывода конкретного свойства из данных, выбираемых из множества наборов исходных данных. При решении задачи учитываются как возможности нарушения причинно-следственной схемы из-за шума, так и возможности недостижимости получения необходимого вывода. Построена причинно-следственная схема приближенного вывода, состоящая из объектов покрытий причин и следствий, которая начинается из исходных предпосылок и заканчивается выводом объекта, содержащего требуемое свойство. Для описания процесса порождения логического вывода объекта с интересующим свойством из исходных данных введено понятие активации объектов. Это свойство позволяет представить схему вывода в форме DAG (Directed Acyclic Graph). Построен простой алгоритм конструирования причинно-следственной схемы из исходных данных к объекту, содержащему искомое свойство. Этот алгоритм также определяет условия существования возможностей вывода из исходных условий объекта с требуемым свойством.

Ключевые слова: причинно-следственные связи; приближенный логический вывод; вероятность правильного вывода в условиях шума

DOI: 10.14357/19922264230212 **EDN:** TGXQIW

1 Введение

Исследованию причинно-следственных связей посвящено много научных работ [1–3]. С помощью причинно-следственных связей решаются задачи поиска первопричины сбоев (Root Cause Analyses — RCA) и аномалий [4–6] в компьютерных системах и сетях. Методы искусственного интеллекта в медицине [3, 7] основаны на поиске причин заболевания. Причинно-следственные связи используются при анализе атак на компьютерные системы [8] и др.

Причинно-следственные связи обычно моделируются графами. Причины и следствия — узлы графа, если A — причина, а B — следствие этой причины, то графически это изображается ориентированным ребром (дугой) от узла A к узлу B. Сложные причинно-следственные связи часто представляются ориентированными ациклическими графами (DAG) [9, 10].

В причинно-следственные (каузальные) модели [3, 11] часто вносятся элементы случайности. В [11] рассматривается функциональная зависимость причина $\rightarrow$ следствие, в которую в качестве дополнительного аргумента внесена случайная величина. Такой подход значительно затрудняет исследование каузальных отношений, так как требует внесения дополнительных ограничений в модель.

Схема внесения случайности в [11] может быть изменена, как это было сделано в работе [12], а именно: причинно-следственная связь рассматривается как детерминированное следствие причины, но в следствие вносится случайный шум, который может не позволить получить («увидеть») данное следствие. В такой модели допущение о независимом шуме для каждого следствия рассматриваемой причины позволяет получать точные оценки вероятностных распределений возможности правильного распознавания наличия причины по наблюдаемым следствиям [13]. Более того, используя детерминированность отношения причина $\rightarrow$ следствие в такой модели, удалось получить оценки вероятностей достижимости правильного результата классификации для сложных схем причинно-следственных связей.

В данной работе рассматривается задача построения логического вывода конкретного свойства из данных, выбираемых из множества наборов исходных данных. При решении задачи учитываются как возможности нарушения причинно-следственной схемы из-за шума, так и возможности недостижимости получения необходимого вывода. Построена причинно-следственная схема приближенного вывода, состоящая из объектов покрытий причин и следствий, которая начинается из исходных предпосылок и заканчивается выводом объекта, содержащего требуемое свойство.

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, grusho@yandex.ru
[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, info@itake.ru
[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, m.zabezhailo@yandex.ru
[4]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, eltimon@yandex.ru
[5]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, sshorgin@ipiran.ru

2 Математическая модель причинно-следственных связей в задачах классификации

В данной работе будем опираться на модель из публикации [14]. Пусть задано некоторое пространство характеристик $U = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ и множество объектов $\boldsymbol{O} = \{O_1, O_2, \ldots, O_m\}$. Каждый объект из множества $\boldsymbol{O}$ есть подмножество пространства U, но не всякое подмножество пространства U представляет собой объект. Будем считать, что множества характеристик в объектах неизвестны. Положим по определению, что $A \subseteq U$ служит причиной появления следствия $B \subseteq U$, если характеристики множеств A и B могут взаимодействовать между собой, и в этом случае при появлении причины A детерминированно возникает следствие B. При этом будем считать, что появившееся следствие B выступает носителем свойства B.

Для удобства можно представить механизм появления следствия B следующим образом. Появление причины A в каком-то смысле означает активацию элементов множества A. Тогда, используя все существующие взаимодействия с потенциальными следствиями, A рассылает «сигнал» от своих характеристик по этим взаимодействиям. Если B является следствием A, то принятые «сигналы» активируют элементы множества B, которое может в свою очередь рассылать «сигналы» о своей активации.

Положим, что взаимодействие A может происходить, не обязательно «сигнализируя» о своей активации другому подмножеству пространства U, но также некоторому взаимодействующему с A множеству характеристик в другом пространстве характеристик U^*.

Более того, активация B происходит только тогда, когда приходят «сигналы» от всех характеристик множества A, т. е. причина A является минимальным множеством, порождающим следствие B. Следует обратить внимание на то, что множества A и B не обязательно представляют собой объекты, хотя понятие активации характеристик объекта может относиться не только к характеристикам A. В этом случае в пространстве U^* селективно выбираются «сигналы» от характеристик из A и активируются все объекты, определенные на пространстве U^* и содержащие B.

Далее будем считать, что понятие активации в множестве объектов $\boldsymbol{O}$ относится только к объектам. Активация причины A позволяет активировать все следствия A при наличии взаимодействия с соответствующими пространствами характеристик.

Будем называть любой объект, содержащий причину A, *покрытием причины*, а любой объект, содержащий следствие B причины A, *покрытием следствия* B. Будем предполагать, что в отношении «причина $A \to$ следствие B» существует единственная причина A у данного следствия B.

Расширим понятие покрытия причины или следствия следующим образом. Пусть часть характеристик множества A содержится в объекте O_1, а остальные характеристики A содержатся в объекте O_2. Тогда при активации обоих объектов O_1 и O_2 и наличии взаимодействия обоих объектов с пространством U^* возникает множество «сигналов» для активации следствия B, т. е. активируется объект, содержащий B. В этом случае причина A распределена между объектами O_1 и O_2.

Если причина A, состоящая из характеристик пространства U, порождает следствие B в пространстве характеристик U^*, а B как причина порождает следствие C в пространстве характеристик U^{**}, то из условия детерминированности отношения причина $\to$ следствие следует транзитивность причинно-следственных связей. Это означает, что A служит *транзитивной причиной* появления свойства C. При этом введение понятия транзитивной причины не нарушает свойство единственности причины, так как у C есть единственная причина — это B из пространства характеристик U^*, а у B есть единственная причина — это A из пространства характеристик U.

Если $A \to B$ и A может быть покрыто распределенной системой объектов, то B должно быть покрыто хотя бы одним объектом $O(B)$. Это свойство на графе причинно-следственных отношений будет соответствовать совокупности дуг, входящих в $O(B)$ и выходящих из объектов распределенного покрытия A. Заметим, что активация нужна для выделения объекта, содержащего следствие. Поскольку возможно участие $O(B)$ в транзитивной причине, то активация сохраняется.

Предположим, что нас интересует свойство C, которое является следствием B, но неизвестно ни одного его покрытия. Пусть известно, что B может быть активировано из строгих подмножеств множества A, которые образуют причину B. Если возможно активировать объекты O_1 и O_2, содержащие подмножества множества A и образующие причину B, то по свойству транзитивной причины будет активировано следствие C вместе с некоторым объектом $O(C)$, содержащим C. Объекты O_1 и O_2 могут быть активированы исходными (входными) данными или другими активированными объектами.

Нетрудно видеть, что в построенном по этим правилам графе с любым числом узлов не могут

образоваться ориентированные циклы. Ориентированный цикл требует повторной активации хотя бы одного объекта, который ранее был активирован. Но повторная активация невозможна по соглашению выше. Тогда цикл не может быть замкнут.

Если в объекте, предполагающем зацикливание, нужны другие характеристики, не использованные раньше, то (по соглашению) активируются все объекты, содержащие активированное следствие. Отсюда следует, что для получения в качестве следствия нужного свойства граф многих причинно-следственных связей объектов, содержащих используемые причинно-следственные связи, должен быть ориентированным связным графом без ориентированных циклов, т. е. DAG. При этом, чтобы не оставлять висячие вершины, среди всех объектов, содержащих нужное следствие, можно выбрать одно для дальнейшего построения причинно-следственного графа, но можно оставить и другие объекты для последующего контроля правильности порожденного графа. Отметим, что по соглашению выше наличие свойства или существование нужного следствия в объекте известно, но может быть неизвестен состав его характеристик.

3 Вероятностная модель шума

Построим модель шума, который может влиять на характеристики причинно-следственных связей. Пусть дано причинно-следственное отношение $A \to B$ и в пространстве характеристик U^* следствия B любая его характеристика независимо от других может из-за воздействия шума перестать распознаваться с вероятностью p. Одинаковые вероятности нераспознавания характеристик взяты для удобства, а свойство независимости допустимо, так как дискретный шум в условиях неупорядоченности характеристик в множестве B трудно описать иначе.

Так как следствие, если в нем присутствуют все характеристики, валидно, то нераспознавание хотя бы одной из них означает непоявление всего следствия и всех связанных со следствием активаций. Таким образом, граф причинно-следственных связей (причинно-следственная схема), состоящий из M узлов, расположенных в пространствах характеристик $U_1, U_2, \ldots, U_M$, является ориентированным, связным и ациклическим, т. е. DAG. В условиях наличия шума могут получаться частичные подграфы, которые заведомо не порождают интересующее свойство.

Для простоты предположим, что все причины и следствия имеют одинаковое число характеристик s. Тогда вероятность непоявления данного следствия B (далее — событие $\overline{B}$) равна

$$\mathsf{P}(\overline{B} = 1 - (1-p)^s.$$

Вероятность того, что причинно-следственная схема от исходных данных до порождения требуемого следствия позволяет правильно решать задачу логического вывода, равна

$$\mathsf{P} = (1-p)^{sM},$$

а вероятность того, что причинно-следственная схема не позволяет правильно решать данную задачу (сбой), соответственно, равна $1 - \mathsf{P}$.

Для контроля решения задачи построения логического вывода конкретного свойства из исходных данных в схеме причинно-следственных связей можно оставлять контрольные объекты, содержащие соответствующие следствия. Отсутствие контрольного объекта означает сбой. Это позволяет удаленно контролировать работоспособность схемы до сбоя.

Отметим, что для логического вывода определенного свойства могут существовать несколько DAG. Это возможно, например, за счет различного представления распределенной причины и построения других вариантов исходного графа причинно-следственной схемы.

4 Алгоритмы построения DAG причинно-следственных связей

В данном разделе построим алгоритм получения искомого свойства C с помощью схемы причинно-следственных связей. Пусть по-прежнему $\boldsymbol{O}$ — ограниченное множество объектов в семействе различных пространств характеристик $U_1, U_2, \ldots, U_M$, которые можно использовать при построении причинно-следственной схемы.

По определению свойство C принадлежит объекту $O(C)$ в одном из рассматриваемых пространств характеристик. В противном случае свойство C недостижимо. Свойство C должно быть получено из исходных данных, которые активируют $O(C)$. Если такие исходные данные сразу активируют $O(C)$, то задача решена. Если это не так, то надо построить причинно-следственную схему, в которой последовательность активаций порождает в качестве одного из следствий $O(C)$. Для этого выберем из $U_1, U_2, \ldots, U_M$ те пространства, которые взаимодействуют с пространством, содержащим $O(C)$. Последовательным перебором опробуем все объекты этих пространств на предмет поиска объекта, содержащего причину следствия C. Если такого объекта не найдено, то опробуем все пары

объектов в объединении этих пространств. Если не найдена распределенная причина для следствия C, то перейдем к опробованию троек объектов, и т. д.

Ограничениями этого алгоритма могут стать допустимая сложность вычислений или ограниченность множества возможных наборов объектов. Если найдена прямая или распределенная причина для C, тогда проверяется возможность активации соответствующих объектов из исходных данных. Если найдены несколько распределенных причин следствия C, то далее независимо строятся несколько допустимых DAG для каждой распределенной причины. Если найден хотя бы один объект, который может быть активирован из исходных данных, то в дальнейших шагах алгоритма построения данного DAG этот объект не участвует.

Если в найденной прямой или распределенной причине следствия C участвует причина или часть причины B, то алгоритм работает с B так же, как происходила работа с C.

Алгоритм прекращает свою работу, когда все объекты в порожденной причинно-следственной схеме могут быть активированы из исходных данных. При этом совсем не обязательно, что все исходные данные должны быть использованы. Следует также учитывать, что на каждом цикле алгоритма при нахождении искомой части необходимой причины нужно активировать все объекты соответствующего пространства характеристик, содержащих искомую часть причины.

Если при выполнении перечисленных выше ограничений невозможно вывести или получить из исходных данных все узлы графа, то необходимо перейти к следующему допустимому DAG.

Если все возможности исчерпаны без положительного результата, то считаем, что C не может быть получено из данного набора исходных данных.

При построении шагов алгоритма для объектов, порождающих распределенные причины, возникает необходимость порождения причины объекта с неизвестным свойством в отличие от прямых причин, которые могут быть идентифицированы. В этом случае используется содержащий эту часть причины объект и для него ищется другой объект, активирующий данный.

Несмотря на сложность переборного процесса, алгоритм корректно определен. В самом деле, отсутствие прямой или распределенной причины на каждом шаге алгоритма определяется ограничением сложности или исчерпанностью перебора либо возможностью остановки данной ветви алгоритма за счет исходных данных. Как было показано ранее, ориентированные циклы получить невозможно, других противоречий объектов в узлах в каждом DAG не существуют. Наборы характеристик в причинах и следствиях неизвестны, но активация происходит на объектах, содержащих причину или следствие. Поэтому один объект может участвовать несколько раз с разными причинами или следствиями.

Можно предусмотреть некоторые сокращения времени перебора за счет распараллеливания. Укажем точки возможного выхода на параллельные пути выполнения алгоритма. Прежде всего, построение продолжений DAG из различных узлов их возможного возникновения. Другой вариант распараллеливания состоит в том, что шаг алгоритма, состоящий в поиске причины для данного следствия, универсален и может быть масштабирован в произвольном числе экземпляров при использовании на различных ветках поиска промежуточных причин.

5 Заключение

Идея покрытия причин и следствий объектами из определенного множества лежит в основе приближенного логического вывода. Например, объектами могут быть программно-технические устройства в реализуемой информационной технологии, а искомым свойством служить сбой в этой технологии.

В работе построена логическая схема приближенного вывода, состоящая из объектов покрытий причин и следствий, которая начинается из исходных предпосылок и заканчивается выводом объекта, содержащего требуемое свойство. В условиях появления случайного шума оценена вероятность сбоя и правильного срабатывания логического вывода объекта, содержащего требуемое свойство. В отличие от большинства схем причинно-следственного вывода шум может присутствовать только в следствиях простейших схем причина → следствие.

Для описания процесса порождения логического вывода объекта с интересующим свойством из исходных данных введено понятие активации объектов. Это свойство позволяет представить схему вывода в форме DAG. Построен простой алгоритм конструирования причинно-следственной схемы из исходных данных к объекту, содержащему искомое свойство. Этот алгоритм также определяет условия существования возможностей вывода из исходных условий объекта с требуемым свойством.

Литература

1. *Halpern J. Y., Pearl J.* Causes and explanations: A structural-model approach. Part I: Causes // Brit. J. Philos. Sci., 2005. Vol. 56. No. 4. P. 843–887.
2. *Pearl J.* Causal inference // Causality: Objectives and assessment / Eds. I. Guyon, D. Janzing, B. Scholkopf. — Proceedings of machine learning research ser. — Whistler, Canada, 2010. Vol. 6. P. 39–58.
3. *Pearl J.* The mathematics of causal inference // Joint Statistical Meetings Proceedings. — ASA, 2013. P. 2515–2529.
4. *Jurn J., Kim T., Kim H.* A survey of automated root cause analysis of software vulnerability // Innovative mobile and internet services in ubiquitous computing / Eds. L. Barolli, F. Xhafa, N. Javaid, T. Enokido. — Advances in intelligent systems and computing ser. — Cham: Springer, 2019. Vol. 773. P. 756–761. doi: 10.1007/978-3-319-93554-6_74.
5. *Grusho A., Grusho N., Zabezhailo M., Timonina E., Senchilo V.* Metadata for root cause analysis // Communications ECMS, 2021. Vol. 35. Iss. 1. P. 267–271. doi: 10.7148/2021-0267.
6. *Grusho A. A., Grusho N. A., Zabezhailo M. I., Timonina E. E.* Localization of the root cause of the anomaly // Autom. Control Comp. S., 2021. Vol. 55. No. 8. P. 978–983. doi: 10.3103/s0146411621080137.
7. *Грушо А. А., Грушо Н. А., Забежайло М. И., Тимонина Е. Е.* Поддержка решения задач диагностического типа // Системы и средства информатики, 2021. Т. 31. № 1. С. 69–81. doi: 10.14357/08696527210106.
8. *Грушо А. А., Забежайло М. И., Зацаринный А. А., Тимонина Е. Е.* О некоторых возможностях управления ресурсами при организации проактивного противодействия компьютерным атакам // Информатика и её применения, 2018. Т. 12. Вып. 1. С. 62–70. doi: 10.14357/19922264180108.
9. *Williams T. C., Bach C. C., Matthiesen N. B., Henriksen T. B., Gagliardi L.* Directed acyclic graphs: a tool for causal studies in pediatrics // Pediatr. Res., 2018. Vol. 84. P. 487–493. doi: 10.1038/s41390-018-0071-3.
10. *Grusho A., Grusho N., Zabezhailo M., Timonina E.* Generation of metadata for network control // Distributed computer and communication networks / Eds. V. M. Vishnevskiy, K. E. Samouylov, D. V. Kozyrev. — Lecture notes in computer science ser. — Cham: Springer, 2020. Vol. 12563. P. 723–735. doi: 10.1007/978-3-030-66471-8_55.
11. *Schölkopf B.* Causality for machine learning. — Cornell University, 2019. arXiv:1911.10500v2 [cs.LG]. 20 p.
12. *Грушо А. А., Грушо Н. А., Забежайло М. И., Кульченков В. В., Тимонина Е. Е., Шоргин С. Я.* Причинно-следственные связи в задачах классификации // Информатика и её применения, 2023. Т. 17. Вып. 1. С. 43–49. doi: 10.14357/19922264230106.
13. *Грушо А. А., Забежайло М. И., Кульченков В. В., Смирнов Д. В., Тимонина Е. Е., Шоргин С. Я.* Причинно-следственные связи в задачах анализа ненаблюдаемых свойств процессов // Системы и средства информатики, 2023. Т. 33. № 2. С. 71–78.
14. *Аншаков О. М.* Об одной интерпретации ДСМ-метода автоматического порождения гипотез // Автоматическое порождение гипотез в интеллектуальных системах / Под ред. В. К. Финна. — М.: Либроком, 2009. С. 81–95.

Поступила в редакцию 11.04.23

COMPLEX CAUSE-AND-EFFECT RELATIONSHIPS

A. A. Grusho, N. A. Grusho, M. I. Zabezhailo, E. E. Timonina, and S. Ya. Shorgin

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: The paper discusses the task of constructing a logical inference of the specific property from data selected from a plurality of sets of source data. When solving the problem, it should be taken into account both the possibility of violating the cause-and-effect scheme due to noise and the possibility of not achieving the necessary conclusion. The cause-and-effect scheme of approximate inference has been built, consisting of objects of covering causes and consequences, which begins from the initial prerequisites and ends with the output of the object containing the required property. To describe the process of generating logical inference of the object with the property of interest from the source data, the concept of activating objects is introduced. This property allows one to represent the inference scheme in the form of a DAG (Directed Acyclic Graph). The simple algorithm for constructing the cause-and-effect scheme was built from the source data up to the object containing the desired property. This algorithm also determines the conditions for the existence of the ability to inference from the original conditions up to the object with the required property.

Keywords: cause-and-effect relationships; approximate logical inference; probability of correct inference under noise conditions

DOI: 10.14357/19922264230212 **EDN:** TGXQIW

References

1. Halpern, J. Y., and J. Pearl. 2005. Causes and explanations: A structural-model approach. Part I: Causes. *Brit. J. Philos. Sci.* 56(4):843–887.
2. Pearl, J. 2010. Causal inference. *Causality: Objectives and assessment*. Eds. I. Guyon, D. Janzing, and B. Scholkopf. Proceedings of machine learning research ser. Whistler, Canada. 6:39–58.
3. Pearl, J. 2013. The mathematics of causal inference. *Joint Statistical Meetings Proceedings*. ASA. 2515–2529.
4. Jurn J., T. Kim, and H. Kim. 2019. A survey of automated root cause analysis of software vulnerability. *Innovative mobile and internet services in ubiquitous computing*. Eds. L. Barolli, F. Xhafa, N. Javaid, and T. Enokido. Advances in intelligent systems and computing ser. Cham: Springer. 773:756–761. doi: 10.1007/978-3-319-93554-6_74.
5. Grusho, A., N. Grusho, M. Zabezhailo, E. Timonina, and V. Senchilo. 2021. Metadata for root cause analysis. *Communications ECMS* 35(1):267–271. doi: 10.7148/2021-0267.
6. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2021. Localization of the root cause of the anomaly. *Autom. Control Comp. S.* 55(8):978–983. doi: 10.3103/s0146411621080137.
7. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2021. Podderzhka resheniya zadach diagnosticheskogo tipa [Support for solving diagnostic type problems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 31(1):69–81. doi: 10.14357/08696527210106.
8. Grusho, A. A., M. I. Zabezhailo, A. A. Zatsarinny, and E. E. Timonina. 2018. O nekotorykh vozmozhnostyakh upravleniya resursami pri organizatsii proaktivnogo protivodeystviya komp'yuternym atakam [On some possibilities of resource management for organizing active counteraction to computer attacks]. *Informatika i ee Primeneniya — Inform. Appl.* 12(1):62–70. doi: 10.14357/19922264180108.
9. Williams, T. C., C. C. Bach, N. B. Matthiesen, T. B. Henriksen, and L. Gagliardi. 2018. Directed acyclic graphs: A tool for causal studies in pediatrics. *Pediatr. Res.* 84(4):487–493. doi: 10.1038/s41390-018-0071-3.
10. Grusho, A., N. Grusho, M. Zabezhailo, and E. Timonina. 2020. Generation of metadata for network control. *Distributed computer and communication networks*. Eds. V. M. Vishnevskiy, K. E. Samouylov, and D. V. Kozyrev. Lecture notes in computer science ser. Cham: Springer. 12563:723–735. doi: 10.1007/978-3-030-66471-8_55.
11. Schölkopf, B. 2019. Causality for machine learning. Cornell University. *arXiv.org*. 20 p. Available at: https://arxiv.org/abs/1911.10500v2 (accessed June 12, 2023).
12. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, V. V. Kulchenkov, E. E. Timonina, and S. Ya. Shorgin. 2023. Prichinno-sledstvennye svyazi v zadachakh klassifikatsii [Causal relationships in classification problems]. *Informatika i ee Primeneniya — Inform. Appl.* 17(1):43–49. doi: 10.14357/19922264230106.
13. Grusho, A. A., M. I. Zabezhailo, V. V. Kul'chenkov, D. V. Smirnov, E. E. Timonina, and S. Ya. Shorgin. 2023. Prichinno-sledstvennye svyazi v zadachakh analiza nenablyudaemykh svoystv protsessov [Cause-and-effect relationships in analysis of unobservable process properties]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 33(2):71–78.
14. Anshakov, O. M. 2009. Ob odnoy interpretatsii DSM-metoda avtomaticheskogo porozhdeniya gipotez [On one interpretation of the JSM-method of automatic generation of hypotheses]. *Avtomaticheskoe porozhdenie gipotez v intellektual'nykh sistemakh* [Automatic hypotheses generation in intelligent systems]. Ed. V. K. Finn. Moscow: Librokom. 81–95.

Received April 11, 2023

Contributors

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, professor, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; grusho@yandex.ru

Grusho Nikolai A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; info@itake.ru

Zabezhailo Michael I. (b. 1956) — Doctor of Science in physics and mathematics, professor, principal scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; m.zabezhailo@yandex.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; eltimon@yandex.ru

Shorgin Sergey Ya. (b. 1952) — Doctor of Science in physics and mathematics, professor, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; sshorgin@ipiran.ru

МЕТОДОЛОГИЯ КОРПУСНО-ОРИЕНТИРОВАННОГО ИССЛЕДОВАНИЯ В ОБЛАСТИ КОНТРАСТИВНОЙ ПУНКТУАЦИИ*

В. А. Нуриев[1], В. И. Карпов[2]

Аннотация: Уточняется подход к современным исследованиям в области контрастивной пунктуации с точки зрения методологии. С учетом новейших достижений информатики, компьютерной лингвистики и теории перевода такие исследования очевидным образом должны иметь корпусно-ориентированный характер. В данной статье представлена методологическая схема подобного исследования, направленного на выявление межъязыковой пунктуационной асимметрии посредством сравнения функционального диапазона одного и того же знака препинания в разных языках. Показываются основные методологические тенденции, характерные для этой научной области. Внимание уделяется особенностям корпусной методологии при контрастивном изучении пунктуации. В качестве одного из современных методологических инструментов предлагаются надкорпусные базы данных (НБД), разрабатываемые в ФИЦ ИУ РАН.

Ключевые слова: контрастивная пунктуация; перевод; корпусное переводоведение; корпусно-ориентированное исследование; параллельный корпус; надкорпусная база данных; межъязыковая асимметрия; методология

DOI: 10.14357/19922264230213 **EDN:** VBOZAO

1 Введение

Важность и необходимость исследований в области контрастивной пунктуации в научной литературе отмечалась неоднократно (см., например, [1–7]). Обычно эта необходимость выводится из нужд переводческой практики, которая предполагает при обработке письменного текста обязательную речемыслительную программу, связанную с исходным пунктуационным компонентом и его переносом в систему переводящего языка. Так, Ньюмарк в своем «Учебнике перевода» пишет, что «пунктуация может быть мощнейшим инструментом, но ее настолько легко упустить из виду, что я советую переводчикам: специально сравнивайте, где у вас расставлены знаки препинания, а где они стоят в оригинале» [1, с. 58]. В работе «Переводчик в тексте: о чтении русской литературы по-английски» значение пунктуации отмечает Мей, критикуя англоязычных переводчиков за недостаточное внимание к межъязыковой пунктуационной асимметрии — за «игнорирование отличительных особенностей, присущих знакам препинания» [2, с. 121]. О пунктуации в переводе говорит Юдейл, выделяя три аспекта: (1) «знаки препинания — важная часть перевода, но, концентрируясь на общем смысле переводимого, ее часто не замечают»; (2) «изменения в пунктуации при переводе могут значительно повлиять на выразительность текста, его связанность и ритм»; (3) «часто возникает впечатление, что литературные переводчики наделили себя правом менять границы исходного предложения и пунктуационные знаки, как им заблагорассудится» [6, с. 121]. Гораздо реже в специализированной литературе подчеркивается роль, которую исследования в области контрастивной пунктуации играют при обучении иностранным языкам, в частности при обучении иноязычной письменной речи [5].

Признавая безусловную значимость данного научного направления и его дальнейшего развития, необходимо предметно разрабатывать методологию исследования в области контрастивной пунктуации, которая учитывала бы новейшие достижения информатики, компьютерной лингвистики и корпусного переводоведения. Представляется, что такая методология должна основываться на использовании современных информационных корпусных инструментов, позволяющих автоматизированным образом обрабатывать представительные массивы текстовых данных, и, следовательно, носить корпусно-ориентированный характер (о корпусных данных при контрастивном изучении пунктуации см. [7]).

*Работа выполнена за счет гранта Российского научного фонда (проект 23-28-00548) с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, nurieff.v@gmail.com

[2]Институт языкознания Российской академии наук; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, wi.karpow@gmail.com

2 Методологические модели корпусно-ориентированного исследования контрастивной пунктуации

В мае 2019 г. в Регенсбурге (Германия) прошла научная конференция под названием «Punctuation Seen Internationally. System–Norm–Practice» («Пунктуация в мировом масштабе: система–норма–практика») — первая конференция, полностью посвященная проблемам контрастивной пунктуации. Оргкомитет, собирая заявки на участие, справедливо отмечал, что до настоящего времени пунктуации едва ли уделялось внимание в рамках типологии, контрастивной лингвистики, прагмалингвистики, а также в исследованиях индивидуальной языковой манеры на фоне языкового стандарта. Сейчас появляются отдельные работы, где проводится сопоставительное изучение пунктуации, однако по-прежнему ощущается острая необходимость в исследованиях по контрастивной пунктуации, которые бы учитывали типологические (системные), социолингвистические (нормативные) и прагматические (речевые) аспекты.

Итогом конференции стала коллективная монография [8], состоящая из шестнадцати статей, которые представляют собой пионерские исследования, направленные на формирование целостной парадигмы контрастивного изучения пунктуации и борьбу с маргинализацией важной научной отрасли. Все статьи условно можно разделить на 4 категории, первые две из которых имеют в большей степени теоретический характер и связаны с системой и нормой, а вторые — более практической направленности — с узусом и освоением пунктуационных навыков. В представленных работах доминируют два подхода к исследованию контрастивной пунктуации:

(1) интралингвистический (контрастивный анализ знаков препинания и конкурирующих с ними маркеров синтаксических отношений в рамках одного языка) [9, с. 110];

(2) интерлингвистический (контрастивный анализ знаков препинания и конкурирующих с ними средств в разных языках, контрастивная пунктуация рассматривается в том числе как часть методики обучения неродному языку, например при интеграции трудовых мигрантов в иноязычную среду) [10, с. 57–73].

Интралингвистический подход часто носит смешанный характер: если речь идет об эволюции пунктуационной системы отдельно взятого языка на фоне развития аналогичных систем других языков, контрастивный анализ сопровождается историко-этимологическим [11, с. 187–206]. В рамках этого подхода в указанной монографии имеются психолингвистические исследования с нетривиальным корпусным материалом. Так, в статье [12, с. 163–186] корпусные данные привлекаются для контрастивного анализа пунктуационных предпочтений двух групп испытуемых. Автор использует корпус CoPaDocs (Corpus of Patient Documents), основу которого составили письма и другие личные документы бывших пациентов психиатрических учреждений Германии на рубеже XIX–XX вв. Корпус позволяет установить, зависит ли языковое оформление письма от личности адресата — происходит ли переключение регистров сознательно. Данный корпус создан с целью разработки интегративной методики анализа языковой вариативности, в том числе и в области пунктуации. Изучив специфику расстановки 12 знаков препинания, Эбер-Хаммерль приходит к выводу, что пациенты, чей род деятельности прежде не был связан с письменной сферой, использовали больше пунктуационных маркеров (но с меньшей вариативностью), чем представители второй опытной группы — канцелярские служащие. В личной переписке участники обеих групп к знакам препинания прибегали гораздо реже, чем в документах, адресованных официальным лицам.

В статье [10, с. 57–73] представлено контрастивное исследование, выполненное в интерлигвистическом ключе. Сопоставление пунктуации в итальянском и немецком языках здесь проводится на основе комплексной методологии, включающей приемы дескриптивного, просодического, синтаксического и коммуникативно-текстового анализа. Примеры приводятся из различных источников, причем к корпусным данным в статье отсылают не напрямую, а опосредованно — через более раннюю работу [13]. По мнению авторов, пунктуирование в этих языках организовано по-разному, что объясняется резкими различиями в пунктуационном узусе: если в итальянском знаки препинания коммуникативно нагружены, то в немецком они подчинены формально-синтаксическому принципу. Иначе говоря, итальянская пунктуация выполняет не формальную функцию, а сигнализирует о тонких смысловых нюансах, которых нельзя достичь другими языковыми средствами (аргументативный конфликт, полифонические эффекты, метатекстовые комментарии). В этом же духе выполнена и другая интерлингвистическая работа [14], посвященная контрастивному исследованию многоточия и тире в итальянском и английском языках и продуктивно использующая

корпусный метод сбора и обработки эмпирических данных.

Объединенные в коллективную монографию работы позволяют вывести обобщенную методологическую схему контрастивного изучения пунктуации. Она имеет трехфазную структуру. Первая фаза включает теоретическое описание пунктуации в изучаемом языке с привлечением исторических и современных нормативных грамматик и справочников. Вторая фаза направлена на описание трансформаций в других языках, оказавших существенное влияние на статус и место пунктуации в системе конкретного языка. Обе фазы нацелены на создание такого исследовательского поля, которое позволит выявить значение пунктуации для языковой культуры. Это, в свою очередь, должно стать задачей третьей фазы. Вторая и третья фазы предполагают межъязыковое сравнение как функционального диапазона отдельно взятых знаков препинания, так и пунктуационного репертуара в целом. На этих стадиях применяется корпусный метод. Контрастивный анализ в зависимости от поставленных целей и задач наряду со знаками препинания может охватывать и конкурирующие с ними языковые средства. Направление контрастивного исследования пунктуации может быть и синхронным, и диахроническим.

3 Методологические особенности корпусно-ориентированного исследования в области контрастивной пунктуации

Особенности методологии при корпусном контрастивном изучении пунктуации, как, впрочем, и при любом корпусно-ориентированном исследовании, связаны прежде всего со стремлением получить непротиворечивые, валидные и надежные данные. Электронный корпус, будучи методологически новаторским инструментом для получения научного знания, позволяет, с одной стороны, автоматическим образом обрабатывать большие массивы данных и тем самым серьезно сокращает временные издержки на поиск эмпирического материала. С другой стороны, электронные корпусные ресурсы имеют свои особенности, и без их надлежащего учета пользователь рискует получить искаженные результаты.

Например, в указанной выше работе [14, с. 291] авторы, описывая методологию своего исследования, отмечают, что итальянские примеры заимствованы из корпуса, хранящегося в Базельском университете и состоящего из двух частей — 33 современных романа-бестселлера (1 млн словоупотреблений) и нехудожественных текста разной направленности (1 млн 40 тыс. словоупотреблений), в то время как англоязычные примеры извлечены из подкорпуса «Книги и периодические издания» Британского национального корпуса (80 млн словоупотреблений). Итальянский материал, по словам авторов, был проанализирован весь, а для английского из-за гораздо большего объема ограничились анализом случайной выборки, объем которой сопоставим с выборкой из итальянского корпуса. Очевидным образом валидность выводов по результатам анализа англоязычного материала здесь может оказаться под вопросом в силу методологически неоднородных установок применительно к процедуре обработки данных, полученных по двум языкам. Примечательно к тому же, что базельский корпус, в отличие от британского, закрыт для общественного пользования.

О подобных ограничениях рассуждает Надворникова в своей работе, посвященной корпусной методологии контрастивного изучения пунктуации [15], где анализируется частотность употребления шести знаков препинания (запятой, точки, двоеточия, точки с запятой, вопросительного и восклицательного знака) в английском, французском и чешском языках. Для сбора данных используются сопоставимые веб-корпусы, моноязычные общие (референтные) и параллельные корпусы. Цель автора — определить, какой из трех типов корпусных ресурсов наиболее подходит для исследований в области контрастивной пунктуации.

Полученные данные показывают, что при изучении пунктуации показатели частотности проявляют высокую чувствительность к типу текста; следовательно, веб-корпусы, которые, как правило, отличают стихийное наполнение, неупорядоченность и низкая степень структурированности, не могут служить источником достоверной информации об употреблении знаков препинания в том или ином языке. Моноязычный общий корпус, наоборот, содержит специальную разметку (морфологическую, синтаксическую и т. д.) и позволяет гибко настраивать поиск (в том числе выбирать соответствующий тип текста) в зависимости от конкретных исследовательских задач. Такие корпусы располагают большими массивами данных, поскольку призваны представить язык во всей его полноте и многообразии, что, казалось бы, обеспечивает надежность и валидность полученных результатов. Между тем этот тип корпусов имеет существенный недостаток — ограниченную межъязыковую сопоставимость. Как правило, моноязычные общие корпусы разных языков разительно отличаются по объему данных и их составу и поэтому не подходят

в качестве основного инструмента контрастивного исследования, а могут служить лишь референтным (проверочным) источником для дополнительной верификации результирующих данных. Кроме того, сопоставительный анализ относительной частотности употребления знаков препинания в разных языках на основе данных, извлеченных из корпусов этого типа, также имеет свои ограничения. Он не применим для изучения пунктуации в языках разного строя, которым для кодирования информации требуется количественно больше (аналитические языки типа французского) или меньше слов (синтетические языки типа русского). Таким образом, лучше всего для контрастивного изучения пунктуации подходят параллельные корпусы, которые, несмотря на свой сравнительно небольшой объем, представляют существенно больше возможностей для качественного анализа употребления знаков препинания и непосредственного сопоставления их абсолютной частотности в параллельных текстах — оригинале и переводе. Однако и этот тип информационного ресурса не может служить универсальным исследовательским инструментом. При его использовании необходимо учитывать, что пунктуационные расхождения в исходном и переводном тексте могут быть не результатом системных дифференциаций, а возникнуть под влиянием переводческих предпочтений. Следовательно, чтобы избежать искажения результирующих данных, надо следовать некоторым методологическим принципам:

(1) данные собираются в обоих переводных направлениях;

(2) выявленные тенденции проходят обязательную проверку с помощью референтного моноязычного корпуса;

(3) контрастивное изучение пунктуации с применением параллельных корпусов требует системного подхода в том смысле, что в функциональном диапазоне разных знаков препинания могут быть общие зоны, указывающие на их потенциальную внутриязыковую и межъязыковую конкуренцию.

4 Заключение

В статье представлена обобщенная методологическая схема корпусно-ориентированного исследования в области контрастивной пунктуации — отрасли научного знания, интенсивно развивающейся и привлекающей внимание специалистов самого широкого профиля. Несмотря на то что появляются работы, где описываются сопоставительные исследования пунктуации на примере одного произведения или литературного наследия отдельно взятого писателя (см., например, [16, 17]), очевидно, что для каких-либо существенных, крупномасштабных обобщений относительно межъязыковой пунктуационной асимметрии и специфики функционирования знаков препинания в разных языках требуется привлечение корпусного материала.

Дальнейшее изучение контрастивной пунктуации видится в нескольких направлениях. Необходимо качественное углубление сопоставительного анализа, чтобы его тонкая нюансировка позволила установить, в какой мере совпадает и разнится функциональный диапазон того или иного знака препинания в контактирующих языках в зависимости от жанровой принадлежности текста. Этот анализ целесообразно проводить комплексно, охватывая всю совокупность синтаксических изменений, которые влекут за собой отказ от исходного пунктуирования при переводе с одного языка на другой. Такая комплексность поможет выявить и с большей полнотой описать существующие межъязыковые структурные различия, что необходимо и для переводческой практики, и для обучения иностранным языкам. Требует дальнейшего уточнения вопрос, как на пунктуационные преференции переводчика влияет родная языковая культура, пунктуационные установки которой могут меняться со временем. По мере наращивания опыта и мастерства могут меняться пунктуационные предпочтения и самого переводчика, и это также представляет определенный научный интерес.

В заключение следует отметить, что одним из современных информационных инструментов корпусного исследования в области контрастивной пунктуации могут быть НБД, разрабатываемые в отделе 54 Федерального исследовательского центра «Информатика и управление» Российской академии наук (о возможностях НБД см. [7]). В данный момент этот методологический инструмент проходит апробацию в контрастивном исследовании двоеточия и многоточия в трех языках — русском, французском и немецком.

Литература

1. *Newmark P.* A textbook of translation. — New York, London, Toronto, Sydney, Tokyo: Prentice Hall, 1988. 402 p.
2. *May R.* The translator in the text: On reading Russian literature in English. — Evanston, IL, USA: Northwestern University Press, 1994. 209 p.
3. *Munday J.* Systems in translation: A systemic model for descriptive translation studies // Crosscultural transgressions: Research models in translation studies II — histori-

cal and ideological issues / Ed. T. Hermans. — Manchester, U.K.: St. Jerome, 2002. P. 76–92.

4. *Baker M.* In other words. — 2nd ed. — London, New York: Routledge, 2011. 352 p.
5. *Сигал К. Я.* Контрастивная пунктуация в начале XXI века // Язык. Текст. Дискурс: Научный альманах Ставропольского отделения РАЛК. — Ставрополь: СКФУ, 2019. Вып. 17. С. 69–78.
6. *Youdale R.* Using computers in the translation of literary style: Challenges and opportunities. — London, New York: Routledge, 2020. 242 p.
7. *Нуриев В. А., Кружков М. Г.* Корпусные данные при контрастивном изучении пунктуации // Системы и средства информатики, 2023. Т. 33. № 1. С. 14–23. doi: 10.14357/08696527230102.
8. Vergleichende Interpunktion — comparative punctuation / Eds. P. Rössler, P. Besl, A. Saller. — Berlin, Boston: De Gruyter, 2021. 454 p.
9. *Rinas K.* Vom genormten Satzbau zur genormten Interpunktion. Zur Funktion der Zeichensetzung in älterer und neuerer Zeit // Vergleichende Interpunktion — comparative punctuation / Eds. P. Rössler, P. Besl, A. Saller. — Berlin, Boston: De Gruyter, 2021. P. 109–136. doi: 10.1515/9783110756319-006.
10. *Ferrari A., Stojmenova Weber R.* Das Komma in kontrastiver Perspektive Italienisch-Deutsch // Vergleichende Interpunktion — comparative punctuation / Eds. P. Rössler, P. Besl, A. Saller. — Berlin, Boston: De Gruyter, 2021. P. 57–73. doi: 10.1515/9783110756319-003.
11. *Besch W.* Zur Entwicklung der deutschen Interpunktion seit dem späten Mittelalter // Interpretation und Edition deutscher Texte des Mittelalters. Festschrift für John Asher zum 60. Geburtstag / Eds. K. Smits, W. Besch, V. Lange. — Berlin: Erich Schmidt, 1981. P. 187–206.
12. *Eber-Hammerl F.* Interpunktion in historischen Patientenbriefen // Vergleichende Interpunktion — comparative punctuation / Eds. P. Rössler, P. Besl, A. Saller. — Berlin, Boston: De Gruyter, 2021. P. 163–186.
13. *Ferrari A.* Leggere la virgola. Una prima ricognizione // Chimera Romance Corpora Linguistic Studies, 2017. Vol. 4. Iss. 2. P. 145–162. doi: 10.15366/chimera2017.4.2.001.
14. *Pecorari F., Longo F.* The ellipsis and the dash in Italian and English: A contrastive perspective // Vergleichende Interpunktion — comparative punctuation / Eds. P. Rössler, P. Besl, A. Saller. — Berlin, Boston: De Gruyter, 2021. P. 289–314. doi: 10.1515/9783110756319-013.
15. *Nádvorníková O.* The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: A question of methodology // Linguist. Prag., 2020. Vol. 30. Iss. 2. P. 30–50. doi: 10.14712/18059635.2020.1.2.
16. *Сигал К. Я.* Пунктуация как средство создания эмоционального подтекста (на материале рассказа М. А. Шолохова «Судьба человека» и его переводов на английский язык) // Известия РАН. Серия литературы и языка, 2014. Т. 73. № 6. С. 38–50.
17. *Богданов К. А.* Пунктуация как мотив: многоточие и тире // НЛО, 2022. № 2(174). С. 241–253. doi: 0.53953/ 08696365_2022_174_2_241.

Поступила в редакцию 15.04.23

METHODOLOGY OF THE CORPUS-BASED STUDIES IN THE FIELD OF CONTRASTIVE PUNCTUATION

V. A. Nuriev[1] and V. I. Karpov[1,2]

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[2]Institute of Linguistics of the Russian Academy of Sciences, 1 bld. 1 Bolshoy Kislovsky Lane, Moscow 125009, Russian Federation

Abstract: The paper refines the methodological approach to the contrastive studies of punctuation. Given the recent achievements of information science, computer linguistics, and translation theory, such studies are most likely to be corpus-based. The paper presents a methodological model of research into interlingual punctuation asymmetry, the aim of which is to shed light on the functional scope of the same punctuation marks in different languages. It shows what methodological trends are characteristic of this research area. The focus is also on the specificities of corpus methodology in the contrastive study of punctuation. It is argued that one of the methodological tools, tailored specifically to the needs of contrastive punctuation research, may be the supracorpora databases developed at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

Keywords: contrastive punctuation; translation; corpus-based translation studies; corpus-based studies; parallel corpus; supracorpora database; asymmetry between languages; methodology

DOI: 10.14357/19922264230213 **EDN:** VBOZAO

Acknowledgments

The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow). The research was supported by the Russian Science Foundation (project No. 23-28-00548).

References

1. Newmark, P. 1988. *A textbook of translation*. New York, London, Toronto, Sydney, Tokyo: Prentice Hall. 402 p.
2. May, R. 1994. *The translator in the text: On reading Russian literature in English*. Evanston, IL: Northwestern University Press. 209 p.
3. Munday, J. 2002. Systems in translation: A systemic model for descriptive translation studies. *Crosscultural transgressions: Research models in translation studies II — historical and ideological issues*. Ed. T. Hermans. Manchester, U.K.: St. Jerome. 76–92.
4. Baker, M. 2011. *In other words*. 2nd ed. London, New York: Routledge. 352 p.
5. Seagal, K. Ya. 2019. Kontrastivnaya punktuatsiya v nachale XXI veka [Contrastive punctuation at the beginning of the XXI century]. *Yazyk. Tekst. Diskurs: Nauchnyy al'manakh Stavropol'skogo otdeleniya RALK* [Language. Text. Discourse: Scientific almanac of Stavropol Branch of the Russian Cognitive Linguists Association]. Stavropol': SKFU. 17:69–78.
6. Youdale, R. 2020. *Using computers in the translation of literary style: Challenges and opportunities*. London, New York: Routledge. 242 p.
7. Nuriev, V. A., and M. G. Kruzhkov. 2023. Korpusnye dannye pri kontrastivnom izuchenii punktuatsii [The parallel corpora perspective on studying contrastive punctuation]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 33(1):14–23. doi: 10.14357/08696527230102.
8. Rössler, P., P. Besl, and A. Saller, eds. 2021. *Vergleichende Interpunktion — comparative punctuation*. Berlin, Boston: De Gruyter. 454 p.
9. Rinas, K. 2021. Vom genormten satzbau zur genormten interpunktion. Zur funktion der zeichensetzung in älterer und neuerer zeit. *Vergleichende Interpunktion — comparative punctuation*. Eds. P. Rössler, P. Besl, and A. Saller. Berlin, Boston: De Gruyter. 109–136. doi: 10.1515/9783110756319-006.
10. Ferrari, A., and R. Stojmenova. 2021. Weber das komma in kontrastiver perspektive Italienisch-Deutsch. *Vergleichende Interpunktion — comparative punctuation*. Eds. P. Rössler, P. Besl, and A. Saller. Berlin, Boston: De Gruyter. 57–73. doi: 10.1515/9783110756319-003.
11. Besch, W. 1981. Zur entwicklung der deutschen interpunktion seit dem späten mittelalter. *Interpretation und Edition deutscher Texte des Mittelalters. Festschrift für John Asher zum 60. Geburtstag*. Eds. K. Smits, W. Besch, and V. Lange. Berlin: Erich Schmidt. 187–206.
12. Eber-Hammerl, F. 2021. Interpunktion in historischen Patientenbriefen. *Vergleichende Interpunktion — comparative punctuation*. Eds. P. Rössler, P. Besl, and A. Saller. Berlin, Boston: De Gruyter. 163–186.
13. Ferrari, A. 2017. Leggere la virgola. Una prima ricognizione. *Chimera Romance Corpora Linguistic Studies* 4(2):145–162. doi: 10.15366/chimera2017.4.2.001.
14. Pecorari, F., and F. Longo. 2021. The ellipsis and the dash in Italian and English: A contrastive perspective. *Vergleichende Interpunktion — comparative punctuation*. Eds. P. Rössler, P. Besl, and A. Saller. Berlin, Boston: De Gruyter. 289–314. doi: 10.1515/9783110756319-013.
15. Nádvorníková, O. 2020. The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: A question of methodology. *Linguist. Prag.* 30(2):30–50. doi: 10.14712/18059635.2020.1.2.
16. Seagal, K. Ya. 2014. Punktuatsiya kak sredstvo sozdaniya emotsional'nogo podteksta (na materiale rasskaza M. A. Sholokhova "Sud'ba cheloveka" i ego perevodov na angliyskiy yazyk) [Punctuation as a means of revealing the emotional subtext (the case of Mikhail Sholokhov's short story "The Fate of a Man" and its translations into English)]. *Izvestiya RAN. Seriya literatury i yazyka* [The Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 73(6):38–50.
17. Bogdanov, K. A. 2022. Punktuatsiya kak motiv: mnogotochie i tire [Punctuation as a motive: The ellipsis and the dash]. *NLO* [New Literary Observer] 2(174):241–253. doi: 0.53953/08696365_2022_174_2_241.

Received April 15, 2023

Contributors

Nuriev Vitaly A. (b. 1980) — Doctor of Science in philology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; nurieff.v@gmail.com

Karpov Vladimir I. (b. 1971) — Candidate of Science (PhD) in philology, leading scientist, Institute of Linguistics of the Russian Academy of Sciences, 1 bld. 1 Bolshoy Kislovsky lane, Moscow 125009, Russian Federation; scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; wi.karpow@gmail.com

ПОДХОДЫ К ПОДБОРУ СПЕЦИАЛИСТОВ ПРИ ОРГАНИЗАЦИИ КОЛЛЕКТИВНОГО РЕШЕНИЯ ПРОБЛЕМ*

С. Б. Румовская[1]

Аннотация: Исследование малых групп (коллективов, команд), их особенностей, проблем, динамики и особенностей подбора специалистов стоит на стыке психологии в управлении персоналом и социальной психологии. Особое место в широком спектре направлений современной науки занимает моделирование взаимодействия людей в малых коллективах специалистов, в частности в рамках многоагентного подхода. При этом, разрабатывая интеллектуальные системы (ИС) (искусственные гетерогенные коллективы) для решения практических проблем, сейчас требуется объединять в составе системы модели специалистов (агентов), созданных различными командами разработчиков, имеющих несовместимые цели и модели предметной области. Отбор специалистов в естественные и моделей специалистов в искусственные гетерогенные коллективы — важная задача, результаты решения которой влияют на дальнейший процесс принятия решений. Представлен анализ методов и подходов к подбору специалистов и комплектования малых групп (коллективов, команд), измерительные инструменты которых должны подвергаться оценке качества.

Ключевые слова: группа; малый коллектив специалистов; команда; методы подбора специалистов и формирования малых групп; командообразование

DOI: 10.14357/19922264230214 **EDN:** VJWNOE

1 Введение

Взаимосвязь психологии в управлении персоналом и социальной психологии состоит в том, что объект исследования первой — как отдельно взятая личность, так и малые группы — один из сложнейших феноменов социальной психологии [1]. «Малая группа» [2] — элементарное звено структуры социальных отношений, обретающее через непосредственные межличностные контакты структурные, динамические и феноменологические характеристики, отражающие признаки группы как целостной системы социальных и психологических отношений. В организациях руководители получают значительный эффект, создавая малые группы с учетом групповой сплоченности, единства и других социально-психологических феноменов.

Отбор специалистов — важная задача, результаты выполнения которой влияют на дальнейшую работу группы (коллектива, команды), решающей в различных сферах проблемы, осложненные слабой формализацией, комплексным строением, сетевым характером условий и целей, неопределенностью, субъективностью и динамичностью. Подобный коллектив, который называют естественным гетерогенным коллективным интеллектом поддержки принятия решений [3], — это малая группа экспертов (специалистов), которой присущи неоднородность, разнообразие, сотрудничество, дополнительность и относительность знаний.

Основная форма организации малых коллективов — совещания, построенные по принципу круглого стола. Особое место в современной науке занимает моделирование взаимодействия людей в таких коллективах, в частности методами многоагентных систем. Сейчас для решения практических проблем требуется объединять в составе ИС модели специалистов (агентов), созданных различными командами разработчиков и имеющих несовместимые цели и модели предметной области. В этой связи для учета субъективности и динамического характера проблем предполагается разработать новый класс ИС — рефлексивно-активные системы искусственных гетерогенных интеллектуальных агентов (РАСИГИА), в которых агенты будут взаимно моделировать рефлексивные позиции друг друга, динамически вырабатывать стратегии своего поведения, по мере необходимости в процессе решения проблем привлекать новых агентов из пула доступных агентов от различных разработчиков и исключать существующих.

Результат работы РАСИГИА зависит от состава выбираемых для включения агентов, а значит, методы подбора специалистов, разрабатываемые в области психологии управления персоналом, должны

*Исследование выполнено за счет гранта РНФ № 23-21-00218.

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, sophiyabr@gmail.com

быть проанализированы на предмет возможности адаптации для РАСИГИА.

В настоящей работе проанализированы методы подбора специалистов и комплектования малых групп (коллективов и команд), разработанные в области психологии управления персоналом.

2 Малые высокоорганизованные группы (коллективы, команды)

В работах [4–7] о малых группах специалистов, рассматриваемых в данном исследовании как естественный гетерогенный коллектив, говорят как о коллективах — высокоразвитой форме организации групповой деятельности, при которой связи и отношения между индивидами опосредованы общественно значимыми целями. Коллективы в отечественной литературе рассматриваются как высший уровень развития группы, характеризуемый высокой степенью сплоченности, единством, ценностно-нормативной ориентации, глубокой идентификацией индивида с группой и ответственностью за результаты совместной групповой деятельности [5]. В работах [8–10] говорится о стадиях зрелости коллектива (команды) в рамках более широкого процесса жизни малой группы, которые можно свести к таким, как (1) притирка и формирование; (2) конфликтная (образуются подгруппы, появляются разногласия); (3) экспериментирование в методах и средствах; (4) появление сплоченности (границы подгрупп стираются, успешное решение задач, творчество); (5) высокий уровень сплоченности (формируются прочные связи, роли и полномочия динамично согласовываются, личные разногласия быстро устраняются). Катценбах и Смит [11] определяют команду как малое число людей (от 2 до 25 человек, но обычно не более 10) с взаимодополняющими умениями, связанных единым замыслом, стремящихся к общим целям и ответственных за их достижение. Команде присущи постоянство состава, жесткое распределение ролей, ясная и формальная цель, а члены команды сыгранны и действуют одинаково по отношению к окружению [12]. В [4] отмечается, что в современной трактовке команды много общего с описанием коллектива в работах отечественных авторов прошлых лет. Команда как группа высокого уровня развития, сравнительно с пониманием коллектива, более реалистична, прагматична, лишена идеологических ярлыков. Таким образом, понятия «малый коллектив специалистов» и «команда» идентичны, поэтому актуально исследование формирования и коллективов, и команд.

3 Методы и подходы к отбору специалистов

Методы комплектования малых групп и коллективов [13]. Определяются оптимальные количественные соотношения между работниками в малых группах и коллективах. Выделяют два взаимодополняющих принципа отбора: сработанность и совместимость [14]. Сработанность характеризуется высокой согласованностью у членов группы [15] и ее продуктивностью и базируется на профессионально-квалификационной дополняемости. Совместимость — оптимальное сочетание свойств участников, обеспечивающее их эффективное существование и способность оптимизировать свои взаимоотношения и согласовывать свои действия [15]. Выделяются три уровня совместимости.

1. Согласованность функционально-ролевых ожиданий — выделяют своеобразные роли, которые в совместной деятельности дополнительно к основным играют люди, например выделяют целевые и поддерживающие [16].
2. Ценностно-ориентационное единство (ЦОЕ) по А. В. Петровскому [17] — сходство мнений, позиций членов группы по отношению к объектам, наиболее значимым для группы в целом. Есть ряд процедур оценки ЦОЕ.
3. Психофизиологическая совместимость:
   - первичное комплектование малых групп [18]. Применяют метод изучения характерных особенностей индивидуальной ориентации человека по отношению к другим людям, диагностируемых опросником межличностных отношений (ОМО) В. Шультца, который определяет межличностную совместимость как отношения между двумя или более индивидами, при которых достигается та или иная степень взаимного удовлетворения межличностных потребностей. Также применяется соционический метод, например Е. С. Филатовой и Ю. В. Иванова [19, 20], базирующийся на том, что отношения между сотрудниками группы можно представить в виде парных взаимодействий, а психологический тип человека проявляется во взаимодействии с другими людьми;
   - перекомплектование — предполагает, что члены обследуемого коллектива достаточно хорошо знают друг друга по совместной деятельности. Тогда можно использовать социометрический тест Дж. Морено [21],

представляющий собой процедуру перекрестного опроса членов группы друг о друге по вопросам или критериям, которые направлены на выявление особенностей их взаимоотношений, взаимных оценок тех или иных качеств личности и поведения. Данные ответов кодируются в специальные матрицы и анализируются вероятностно-статистическими методами. В [22–24] рассмотрен ряд приемов экспертизы психологической совместимости в сложившихся малых группах.

Профессиональный отбор [7] предполагает выбор по критериям профессиональной подготовленности и опыта, уровню и профилю образования.

1. Формирование профиля должности [6]. Каждая должностная позиция, в том числе и в составе коллектива, решающего проблему, предъявляет специалисту ряд требований, информацию о которых структурируют и сводят в единую систему в профиле должности. Используют несколько видов критериев [15]:
   - квалификационные;
   - объективные, констатирующие соответствие реальных достижений оцениваемых субъектов некоторым количественным и качественным показателям;
   - внешние, характеризующие наличие качеств, позволяющих добиваться высоких результатов;
   - психологические, разрабатываемые на основе профессиограммы соответствующего вида деятельности, которая представляет собой систему признаков, описывающих профессию, перечень норм и требований [7];
   - тестовые по индивидуально-психологическим характеристикам.

   В профиль включают факторы приоритетов при принятии решений, основные мотивации и др. В профиле должны быть сформулированы предельно конкретно со шкалами измерений компетенции, которые необходимы, желательны или безразличны, в частности для роли специалиста в коллективе, формирующемся для выработки решения по некоторой проблеме [6, 25].
2. Первичный отбор [7] начинается с анализа списка кандидатов с точки зрения их соответствия общим требованиям посредством анкетирования, тестирования или испытания, графологического анализа (экспертиза почерка и стиля изложения), морфологического анализа и анализа по фотографии.
3. На втором этапе отбора в основном используют [26, 27]:
   - комплексные исследования в центрах оценки персонала — оценки одних и тех же критериев в разных ситуациях и различными методами, а также ролевые и имитационные деловые игры и анализ конкретных ситуаций (моделируются существенные моменты деятельности и оцениваются реальные достижения испытуемых или демонстрируемое поведение) [27];
   - тесты:
     (а) на профпригодность — оценка психофизиологических качеств человека, умений выполнять определенную деятельность;
     (б) общие — оценка общего уровня развития и особенностей мышления, внимания, памяти и других высших психических функций;
     (в) биографические тесты и изучение биографии — анализируются семейные отношения, характер образования, физическое развитие, главные потребности и интересы, особенности интеллекта, общительность;
     (г) личностные тесты — психодиагностические тесты на оценку уровня отдельных качеств и предрасположенность к определенному типу поведения;
   - интервью — беседа, направленная на сбор информации об опыте и уровне знаний и на оценку профессионально важных качеств претендента;
   - также анализируют рекомендации (их источники и оформление) и используют: полиграф, медицинские тесты, психоанализ.

Методика подбора персонала О. С. Насташевской с использованием психограмм и профессиограмм [28]. Методика базируется на выделении типов личности по уровню отклонений от теоретической психограммы через систему отбора. Сначала разрабатывается психограмма на основе профессиограммы специалиста, на которого объявлен отбор, — перечисляются психологические профессионально важные качества специалиста [7]. Затем выбираются диагностические методики и проводится собеседование-тестирование, по результатам которого определяют степень соответствия кандидатов требованиям психограммы и по ним выделяют типы личности (кластерный анализ, k-средних). Кластеризация предполагает неаддитивную модель учета

соответствия психограмме, что повышает адекватность получаемого решения. В итоге кандидаты, тип личности которых с максимальной степенью соответствует требованиям, обсуждаются руководством.

Психологическая оценка персонала при выдвижении в кадровый резерв [29]. Кадровый резерв составляет группа сотрудников, которая должна быть обучена внутренними и внешними экспертами. Создается для обеспечения гибкости в замещении сотрудников. Для выдвижения в кадровый резерв проводятся: оценка профессиональных компетенций, психологическое тестирование и экспертное оценивание в ходе деловых игр (ассессмент-центр). Принцип включения в кадровый резерв базируется на двух основных измерениях: профессиональном и управленческом потенциале, а также эффективности в деятельности. Оценивается уровень следующих базовых психологических характеристик: интеллектуальный уровень (тест структуры интеллекта Р. Амтхауэра, методика оценки социального интеллекта Дж. Гилфорда); лидерский потенциал (калифорнийский личностный опросник (КЛО); стандартизированный метод исследования личности Л. Н. Собчик (СМИЛ)); коммуникабельность (КЛО, СМИЛ); психологическая устойчивость (опросник эмоционального интеллекта Д. В. Люсина, СМИЛ); этичность, порядочность во взаимодействии с коллегами, подчиненными, руководством (экспертная оценка и сбор информации). Сотрудника можно зачислять в кадровый резерв уже при одновременном наличии средних показателей по эффективности и потенциалу.

Формирование команд [9]. Выделяют:

(1) динамический подход, направленный на развитие социоэмоциональных и инструментальных отношений в команде посредством различных тренингов — базируются на стадиях развития групп (M. Kelly, W. Wellins, B. W. Tuckman, Т. Ю. Базаров) — модели здесь дескриптивные;

(2) специально организованные социально-психологические технологии формирования коллективного субъекта деятельности — команды, а именно: командные испытания для развития эмоциональных отношений; тренинги навыков командной работы; командный коучинг; деловые игры, тренинги по разработке общего видения.

При комплектовании/перекомплектовании команд превалирует ориентация на формирование гетерогенных групп [30]:

(1) по полу, возрасту и профессии — не вызывает трудностей;

(2) по интеллекту и личностным чертам — используются тесты оценки IQ, методики диагностики когнитивного стиля и креативности мышления, учитывающие специфику деятельности формирующейся команды (например, тесты Фланагана и Векслера), а также методики, основанные на типологическом подходе К. Г. Юнга (модели Майерс–Бриггс и Кейрси — включают ситуационно-поведенческое тестирование и/или глубинное биографическое интервью), и методики, построенные на базе концепции командных ролей Р. М. Белбина — используют сопоставление результатов применения опросника и внешних оценок, полученных от коллег, — 360-градусная обратная связь.

Диагностика ценностной ориентации осуществляется, например, с помощью проективных методик, обращающихся непосредственно к символическому содержанию, в котором объединены и образ, и отношение к организации, и личный опыт, и ценности, и переживания.

За рубежом выделяют четыре основных подхода к образованию команд [31]:

(1) основанный на развитии и согласовании целей команды (E. A. Locke, E. Weldon и др.) — развитие способности группы людей достигать своих целей;

(2) интерперсональный, ориентированный на анализ процессов и улучшение межличностных отношений (C. Argyris, W. Schutz и др.);

(3) ролевой подход — улучшение работы команды за счет увеличения ясности ролей и, как следствие, увеличения организационной эффективности. Предварительно сотрудников тестируют, определяют типы их поведения, а далее соединяют в команды по принципу взаимодополняемости (модели Р. М. Белбина и Т. Ю. Базарова; взаимодополняющая команда по И. А. Адизес) [3, 32];

(4) проблемно-ориентированный подход (W. G. Dyer, R. Kilman, I. Kilman и др.) — более общий, может включать все предыдущие. Считается, что команда становится более эффективной в результате совместного решения проблем.

Результаты анализа подходов из разд. 3 сведены в таблицу.

В таблице курсивом отмечены подходы, которые могут быть в адаптированном виде использованы при моделировании коллективного принятия

Анализ отечественных и зарубежных подходов к отбору специалистов

Основные подходы к отбору специалистов	Качественные	Количественные	Индивидуальные	Групповые	Возможность моделирования в рамках РАСИГИА
Согласованность функционально-ролевых ожиданий [15, 16]	+			+	±
Ценностно-ориентационное единство по А. В. Петровскому [17]		+		+	±
Комплектование групп с использованием ОМО [18]		+		+	
Соционический метод [19, 20]		+		+	
Социометрическая теория Я. Морено [21]		+	+	+	
Метод формирования профиля должности [7]		+	+		±
Анкетирование и тестирование [7]	+	+	+		±
Графологический и морфологический анализ [7]	+		+		
Комплексные исследования в центрах оценки персонала [27]	+	+	+		±
Биографические тесты и изучение биографии [27]	+		+		
Интервью [23, 25–27]	+	+	+		
Ролевые и имитационные деловые игры [27]	+	+	+	+	
Методика подбора персонала О. С. Насташевской [28]		+			±
Психологическая оценка персонала для кадрового резерва [29]	+	+	+		
Динамический подход к командообразованию [9]	+	+		+	±
Специальные организованные социально-психологические технологии формирования команд [9]	+			+	
Тесты и методики оценки когнитивной сферы для формирования гетерогенных групп [30]	+	+	+		
Проективные методики [30]	+	+	+		
Развитие и согласование целей команды [31]	+	+		+	±
Интерперсональный подход к командообразованию [31]	+	+		+	
Ролевой подход объединения в команды [30, 32]	+	+		+	±
Проблемно-ориентированный подход к формированию команд	+	+			±

Обозначения: ± — в рамках соответствующего подхода есть ограничения и/или противоречия по выделенному критерию.

решения в РАСИГИА для отбора моделей специалистов (агентов).

4 Качество измерительного инструмента отбора специалистов

Одна из проблем отбора — повышение прогностичности, надежности процедуры выделения таких характеристик человека, которые укажут на его последующую успешную профессиональную деятельность и соответствие запросам организации, т. е. необходимо обоснование качества выбранного измерительного инструмента (анкеты, теста и т. д.) [14, 28]. При этом оцениваются:

– эмпирическая валидность инструмента, т. е. соответствие его результатов характеристике, для измерения которой он разработан, — проводится пилотное исследование, в ходе которого респонденты оценивают исследуемый объект при помощи альтернативных анкет. Связь между результатами измерения определяется расчетом коэффициента ранговой корреляции Спирмена [14];
– надежность инструмента:
  (1) с позиции согласованности, т. е. степени однородности состава вопросов (заданий) с точки зрения измеряемой характеристики — определяется связью каждого конкретного элемента инструмента с общим результатом [33];
  (2) с позиции устойчивости — проводится несколько измерений с некоторым промежутком времени одним и тем же инструментом. Инструмент устойчив, если имеет место статистически значимое значение коэффициента корреляции между данными измерений и статистически незначимые различия в средних значениях, полученных при измерениях.

5 Заключение

В работе представлены результаты исследования, которое по материалам открытой печати вы-

явило большое разнообразие подходов к отбору специалистов на должность и в группу (команду, коллектив) — групповые, индивидуальные, качественные, количественные и комбинированные. При этом инструменты оценки кандидатов должны подвергаться проверке на валидность и надежность. На практике часто комбинируют несколько методов, чтобы повысить качество отбора. По результатам анализа также были выделены несколько подходов (при условии их адаптации), которые могут быть в той или иной степени использованы для отбора моделей специалистов (агентов) в искусственный гетерогенный коллектив РАСИГИА из всего пула интеллектуальных агентов: согласованность функционально-ролевых ожиданий; ценностно-ориентационное единство; формирование профиля должности; анкетирование и тестирование; принципы комплексных исследований в центрах; методика О. С. Насташевской; развитие и согласование целей команды; динамический, ролевой и проблемно-ориентированный подходы.

Литература

1. *Куроедова Е. О.* Интернет-курс по дисциплине «Психология в управлении персоналом». http://www.e-biblio.ru/book/bib/04_pravo/psiholog_v_uprav_personalom/sg_online.html.
2. *Кричевский Р. Л., Дубовская Е. М.* Социальная психология малой группы. — М.: Аспект Пресс, 2001. 318 с.
3. *Колесников А. В.* Гетерогенные естественные и искусственные системы // Интегрированные модели и мягкие вычисления в искусственном интеллекте. — М.: Физматлит, 2013. Т. 1. С. 86–103.
4. *Андреева Г. М.* Социальная психология. — М.: Аспект Пресс, 2009. 393 с.
5. *Меньшиков А. А.* Основы интегрированных коммуникаций. — Комсомольск-на-Амуре: КнАГУ, 2012. 101 с.
6. *Коноваленко В. А., Коноваленко М. Ю., Соломатин А. А.* Психология управления персоналом. — М.: Юрайт, 2014. 477 с.
7. Психология управления персоналом / Под ред. Е. И. Рогова. — М.: Юрайт, 2023. 350 с.
8. Социальная психология в современном мире / Под ред. Г. М. Андреевой, А. И. Донцова. — М.: Аспект Пресс, 2002. 335 с.
9. *Короткина Е. Д.* Современные технологии создания команды в организации // Вестник Санкт-Петербургского университета. Сер. 12. Психология. Социология. Педагогика, 2009. Т. 3. № 2. С. 46–53.
10. *Нургалиева А. М., Ахметшина А. Р., Сайфудинова Н. З.* Современные методики формирования эффективной команды в организации // СКИФ. Вопросы студенческой науки, 2018. Вып. 11(27). С. 221–230.
11. *Katzenbach J. R., Smith D. K.* The discipline of teams // Harward Business Review, 1993. Vol. 71. Iss. 2. P. 111–120.
12. *Картушина Т. Н.* Командообразование как потребность в современном процессе управления персоналом // Социально-экономические явления и процессы, 2013. № 5(051). С. 99–102.
13. *Жаглин А. В., Ульянов А. Д.* Основы управления и делопроизводства в органах внутренних дел: Альбом схем. — М.: Юнити-Дана, 2014. 191 с.
14. *Горленко О. А., Ерохин Д. В., Можаева Т. П.* Управление персоналом. — М.: Юрайт, 2023. 217 с.
15. *Кабаченко Т. С.* Психология в управлении человеческими ресурсами. — СПб.: Питер, 2003. 400 с.
16. *Mescon M. H., Albert M., Khedouri F.* Management. — New York, NY, USA: Harper & Row Publs., 1988. 777 p.
17. Психологическая теория коллектива / Под ред. А. В. Петровского. — М.: Педагогика, 1979. 240 с.
18. Рабочая книга практического психолога / Под ред. А. А. Бодалева, А. А. Деркача, Л. Г. Лаптева. — М.: Изд-во Института психотерапии, 2001. 640 с.
19. *Иванов Ю. В.* Деловая соционика. — М.: Топ-персонал, 2004. 200 с.
20. *Филатова Е. С.* Соционика в портретах и примерах. — М.: Черная белка, 2009. 443 с.
21. *Миронова Е. Е.* Сборник психологических тестов. Часть I: Пособие. — Мн.: Женский институт ЭНВИЛА, 2005. 155 с.
22. Лучшие психологические тесты для профотбора и профориентации / Отв. ред. А. Ф. Кудряшов. — Петрозаводск: Петроком, 1992. 318 с.
23. Психологические тесты / Под ред. А. А. Карелина: в 2 т. — М.: Владос, 2002. Т. 1. 312 с. Т. 2. 246 с.
24. *Елисеев О. П.* Практикум по психологии личности. — М.: Юрайт, 2023. 390 с.
25. *Иванова С. В.* Искусство подбора персонала: как оценить человека за час. — М.: Альпина Паблишер, 2012. 269 с.
26. *Мякушкин Д. Е.* Отбор и подбор персонала. — Челябинск: ЮУрГУ, 2006. 26 с.
27. *Базаров Т. Ю.* Технология центров оценки персонала: процессы и результаты. Практическое пособие. — М.: КноРус, 2021. 301 с.
28. *Насташевская О. С.* Психологические аспекты технологии подбора персонала для торговой организации // Вестник Самарской гуманитарной академии. Сер. Психология, 2015. № 1(17). С. 11–29.
29. *Васильева И. В.* Психотехники и психодиагностика в управлении персоналом: Практическое пособие. — М.: Юрайт, 2023. 122 с.
30. *Жуков Ю. М., Журавлев А. В., Павлова Е. Н.* Технологии командообразования. — М.: Аспект Пресс, 2008. 320 с.

31. *Безрукова Е. Ю.* Информационно-методическое обеспечение процесса командообразования: Дисс. . . . канд. псих. наук. — М., 1998. 289 с.
32. *Семина А. П.* Анализ моделей и подходов в формировании команды компании // Вестник Алтайской академии экономики и права, 2020. № 12-2. С. 399–404. doi: 10.17513/vaael.1526.
33. *Яхонтова Е. С.* Стратегическое управление персоналом. — М.: Дело, 2013. 378 с.

Поступила в редакцию 05.04.23

SELECTION OF SPECIALISTS IN THE ORGANIZATION OF COLLECTIVE SOLVING PROBLEMS

S. B. Rumovskaya

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The study of small groups (collectives, teams), their characteristics, problems, dynamics, and features of selection of specialists stands at the intersection of psychology of personnel management and social psychology. A special place in a wide range of areas of modern science is occupied by modeling the interaction of people in small collectives of specialists, in particular, within the framework of a multiagent approach. At the same time, when developing intelligent systems (artificial heterogeneous collectives) to solve practical problems, it is now required to combine in the system the models of specialists (agents) with incompatible goals and domain models. These agents are created by different development teams. The selection of specialists in natural and models of specialists in artificial heterogeneous teams is an important task, the results of which influence the further decision-making process. The paper presents an analysis of methods and approaches to the selection of specialists and the acquisition of small groups (collectives, teams) whose measuring tools should be exposed to quality assessment.

Keywords: group; small collective of specialists; team; methods of selecting specialists and forming small groups; teambuilding

DOI: 10.14357/19922264230214 **EDN:** VJWNOE

Acknowledgments

The research was supported by the Russian Science Foundation (project No. 23-21-00218).

References

1. Kuroedova, E. O. Internet-kurs po distsipline "Psikhologiya v upravlenii personalom" [Online course on the discipline "Psychology in personnel management"]. Available at: http://www.e-biblio.ru/book/bib/04_pravo/psiholog_v_uprav_personalom/sg_online.html (accessed May 11, 2023).
2. Krichevskiy, R. L., and E. M. Dubovskaya. 2001. *Sotsial'naya psikhologiya maloy gruppy* [Social psychology of a small group]. Moscow: Aspect Press. 318 p.
3. Kolesnikov, A. V. 2013. Geterogennye estestvennye i iskusstvennye sistemy [Natural and artificial heterogeneous systems]. *Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte* [Integrated models and oft computing in artificial intelligence]. Moscow: Fizmatlit. 1:86–103.
4. Andreeva, G. M. 2009. *Sotsial'naya psikhologiya* [Social psychology]. Moscow: Aspect Press. 393 p.
5. Men'shikov, A. A. 2012. *Osnovy integrirovannykh kommunikatsiy* [Fundamentals of integrated communications]. Komsomolsk-on-Amur: KnAGU. 101 p.
6. Konovalenko, V. A., M. Yu. Konovalenko, and A. A. Solomatin. 2014. *Psikhologiya upravleniya personalom* [Psychology of human resources management]. Moscow: Yurayt. 477 p.
7. Rogov, E. I., ed. 2023. *Psikhologiya upravleniya personalom* [Psychology of human resources management]. Moscow: Yurayt. 350 p.
8. Andreeva, G. M., and A. I. Dontsov, eds. 2002. *Sotsial'naya psikhologiya v sovremennom mire* [Social psychology in the modern world]. Moscow: Aspect Press. 335 p.
9. Korotkina, E. D. 2009. Sovremennye tekhnologii sozdaniya komandy v organizatsii [Modern approaches to teambuilding in organization]. *Vestnik Sankt-Peterburgskogo universiteta. Ser. 12. Psikhologiya. Sotsiologiya. Pedagogika* [Vestnik of Saint Petersburg University. Ser. 12. Psychology. Sociology. Pedagogy] 3(2):46–53.
10. Nurgalieva, A. M., A. R. Akhmetshina, and N. Z. Sayfudinova. 2018. Sovremennye metodiki formirovaniya effektivnoy komandy v organizatsii [Modern methods of forming an effective team in the organization]. *Skif. Vo-*

prosy studencheskoy nauki [Skif. Issues of Student Science] 11(27):221–230.

11. Katzenbach, J. R., and D. K. Smith. 1993. The discipline of teams. *Harward Business Review* 71(2):111–120.
12. Kartushina, T. N. 2013. Komandoobrazovanie kak potrebnost' v sovremennom protsesse upravleniya personalom [Teambuilding as need in modern HR management]. *Sotsial'no-ekonomicheskie yavleniya i protsessy* [Social-Economic Phenomena and Processes] 5(051):99–102.
13. Zhaglin, A. V., and A. D. Ul'yanov. 2014. *Osnovy upravleniya i deloproizvodstva v organakh vnutrennikh del: Al'bom skhem* [Fundamentals of management and office work in the internal affairs bodies: Album of schemes]. Moscow: Unity-Dana. 191 p.
14. Gorlenko, O. A., D. V. Erokhin, and T. P. Mozhaeva. 2023. *Upravlenie personalom* [Human resource management]. Moscow: Yurayt. 217 p.
15. Kabachenko, T. S. 2003. *Psikhologiya v upravlenii chelovecheskimi resursami* [Psychology in human resource management]. Saint Petersburg: Piter Publishing House. 400 p.
16. Mescon, M. H., M. Albert, and F. Khedouri. 1988. *Management*. New York, NY: Harper & Row Publs. 777 p.
17. Petrovskiy, A. V., ed. 1979. *Psikhologicheskaya teoriya kollektiva* [Psychological theory of the team]. Moscow: Pedagogika. 240 p.
18. Bodalev, A. A., A. A. Derkach, and L. G. Laptev, eds. 2001. *Rabochaya kniga prakticheskogo psikhologa* [Practical psychologist's workbook]. Moscow: Publishing house of the Institute of Psychotherapy Publs. 640 p.
19. Ivanov, Yu. V. 2004. *Delovaya sotsionika* [Business socionics]. Moscow: Top-personal. 200 p.
20. Filatova, E. S. 2009. *Sotsionika v portretakh i primerakh* [Socionics in portraits and examples]. Moscow: Chernaya belka. 443 p.
21. Mironova, E. E. 2005. *Sbornik psikhologicheskikh testov. Chast' I: Posobie* [Collection of psychological tests. Part I: Manual]. Minsk: Zhenskiy Institut ENVILA [ENVIL Women's Institute]. 155 p.
22. Kudryashov, A. F., ed. 1992. *Luchshie psikhologicheskie testy dlya profotbora i proforientatsii* [The best psychological tests for vocational selection and vocational guidance]. Petrozavodsk: Petrokom. 318 p.
23. Karelin, A. A., ed. 2002. *Psikhologicheskie testy*: v 2 tomakh [Psychological tests in two volumes]. Moscow: Vlados. Vol. 1. 312 p. Vol.2. 246 p.
24. Eliseev, O. P. 2023. *Praktikum po psikhologii lichnosti* [Practical work on personality psychology]. Moscow: Yurayt. 390 p.
25. Ivanova, S. V. 2012. *Iskusstvo podbora personala: kak otsenit' cheloveka za chas* [The art of recruiting: How to evaluate a person in an hour]. Moscow: Alpina Publisher. 269 p.
26. Myakushkin, D. E. 2006. *Otbor i podbor personala* [Selection and recruitment of personnel]. Chelyabinsk: Publishing Center of South Ural State University. 26 p.
27. Bazarov, T. Yu. 2021. *Tekhnologiya tsentrov otsenki personala: protsessy i rezul'taty. Prakticheskoe posobie* [Technology of personnel assessment centers: Processes and results. Practical guide]. Moscow: KnoRus. 301 p.
28. Nastashevskaya, O. S. 2015. Psikhologicheskie aspekty tekhnologii podbora personala dlya torgovoy organizatsii [Psychological aspects of technology recruitment for a trade organization]. *Vestnik Samarskoy gumanitarnoy akademii. Ser. Psikhologiya* [Bulletin of Samara Academy for the Humanities. Ser. Psychology] 1(17):11–29.
29. Vasil'eva, I. V. 2023. *Psikhotekhniki i psikhodiagnostika v upravlenii personalom: Prakticheskoe posobie* [Psychotechnics and psychodiagnostics in personnel management: A practical guide]. Moscow: Yurayt. 122 p.
30. Zhukov, Yu. M., A. V. Zhuravlev, and E. N. Pavlova. 2008. *Tekhnologii komandoobrazovaniya* [Teambuilding technologies]. Moscow: Aspect Press. 320 p.
31. Bezrukova, E. Yu. 1998. Informatsionno-metodicheskoe obespechenie protsessa komandoobrazovaniya [Information and methodological support of the teambuilding process]. Moscow. PhD Diss. 289 p.
32. Semina, A. P. 2020. Analiz modeley i podkhodov v formirovanii komandy kompanii [Analysis of models and approaches in the formation of team in company]. *Vestnik Altayskoy akademii ekonomiki i prava* [Bulletin of the Altai Academy of Economics and Law] 12-2:399–404. doi: 10.17513/vaael.1526.
33. Yakhontova, E. S. 2013. *Strategicheskoe upravlenie personalom* [Strategic human resources management]. Moscow: Delo. 378 p.

Received April 5, 2023

Contributor

Rumovskaya Sophiya B. (b. 1985) — Candidate of Science (PhD) in technology, senior scientist, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; sophiyabr@gmail.com

ОБ АВТОРАХ

Берговин Алексей Константинович (р. 1995) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова

Борисов Андрей Владимирович (р. 1965) — доктор физико-математических наук, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Васильев Николай Семенович (р. 1952) — доктор физико-математических наук, профессор Московского государственного технического университета имени Н. Э. Баумана

Воронцов Михаил Олегович (р. 1996) — аспирант факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; математик Московского центра фундаментальной и прикладной математики

Гайдамака Юлия Васильевна (р. 1971) — доктор физико-математических наук, профессор, профессор кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Гаримелла Рама Мурти (р. 1962) — PhD в области вычислительной техники, профессор Университета Махиндра (Индия, Хайдерабад)

Грушо Александр Александрович (р. 1946) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Грушо Николай Александрович (р. 1982) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Забежайло Михаил Иванович (р. 1956) — доктор физико-математических наук, профессор, главный научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Карпов Владимир Ильич (р. 1971) — кандидат филологических наук, ведущий научный сотрудник Института языкознания Российской академии наук; научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ковалёв Сергей Протасович (р. 1972) — доктор физико-математических наук, ведущий научный сотрудник Института проблем управления им. В. А. Трапезникова Российской академии наук

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Мелехин Владимир Борисович (р. 1954) — доктор технических наук, профессор кафедры программного обеспечения вычислительной техники и автоматизированных систем Дагестанского государственного технического университета

Нуриев Виталий Александрович (р. 1980) — доктор филологических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Платонова Анна Алексеевна (р. 1996) — аспирант кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Разумчик Ростислав Валерьевич (р. 1984) — доктор физико-математических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Румовская София Борисовна (р. 1985) — кандидат технических наук, старший научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Румянцев Александр Сергеевич (р. 1986) — доктор физико-математических наук, старший научный сотрудник Института прикладных математических исследований Федерального исследовательского центра «Карельский научный центр Российской академии наук»

Самуйлов Андрей Константинович (р. 1988) — кандидат физико-математических наук, доцент кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Тимонина Елена Евгеньевна (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Торшин Иван Юрьевич (р. 1972) — кандидат физико-математических наук, кандидат химических наук, старший научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ушаков Владимир Георгиевич (р. 1952) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Хачумов Вячеслав Михайлович (р. 1948) — доктор технических наук, заведующий лабораторией интеллектуального управления Института программных систем им. А. К. Айламазяна Российской академии наук; главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук; профессор кафедры информационных технологий Российского университета дружбы народов

Хачумов Михаил Вячеславович (р. 1986) — кандидат физико-математических наук, старший научный сотрудник лаборатории интеллектуального управления Института программных систем им. А. К. Айламазяна Российской академии наук; старший научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук; доцент кафедры информационных технологий Российского университета дружбы народов

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; ведущий научный сотрудник Московского центра фундаментальной и прикладной математики

Шоргин Всеволод Сергеевич (р. 1978) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шоргин Сергей Яковлевич (р. 1952) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Правила подготовки рукописей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются статьи, содержащие результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях.

   Публикация предоставленной автором(ами) рукописи не должна нарушать положений глав 69, 70 раздела VII части IV Гражданского кодекса, которые определяют права на результаты интеллектуальной деятельности и средства индивидуализации, в том числе авторские права, в РФ.

   Ответственность за нарушение авторских прав, в случае предъявления претензий к редакции журнала, несут авторы статей.

   Направляя рукопись в редакцию, авторы сохраняют свои права на данную рукопись и при этом передают учредителям и редколлегии журнала неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на перевод ее на английский язык, а также на ее распространение в России и за рубежом. Каждый автор должен представить в редакцию подписанный с его стороны «Лицензионный договор о передаче неисключительных прав на использование произведения», текст которого размещен по адресу http://www.ipiran.ru/publications/licence.doc. Этот договор может быть представлен в бумажном (в 2-х экз.) или в электронном виде (отсканированная копия заполненного и подписанного документа).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности публикации представленной статьи в открытой печати.

2. К статье прилагаются данные автора (авторов) (см. п. 8). При наличии нескольких авторов указывается фамилия автора, ответственного за переписку с редакцией.

3. Редакция журнала осуществляет экспертизу присланных статей в соответствии с принятой в журнале процедурой рецензирования.

   Возвращение рукописи на доработку не означает ее принятия к печати.

   Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.

4. Решение редколлегии о публикации статьи или ее отклонении сообщается авторам.

   Редколлегия может также направить авторам текст рецензии на их статью. Дискуссия по поводу отклоненных статей не ведется.

5. Редактура статей высылается авторам для просмотра. Замечания к редактуре должны быть присланы авторами в кратчайшие сроки.

6. Рукопись предоставляется в электронном виде в форматах MS WORD (.doc или .docx) или LaTeX (.tex), дополнительно — в формате .pdf, на дискете, лазерном диске или электронной почтой. Предоставление бумажной рукописи необязательно.

7. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки.

   Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3.

   Основной текст: стиль — «Обычный», шрифт — Times New Roman, размер — 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине.

Рекомендуемый объем рукописи — не свыше 10 страниц указанного формата. При превышении указанного объема редколлегия вправе потребовать от автора сокращения объема рукописи.

Сокращения слов, помимо стандартных, не допускаются. Допускается минимальное количество аббревиатур.

Все страницы рукописи нумеруются.

Шаблоны примеров оформления представлены в Интернете: http://www.ipiran.ru/journal/template.doc

8. Статья должна содержать следующую информацию на ***русском и английском языках***:
   - название статьи;
   - Ф.И.О. авторов, на английском можно только имя и фамилию;
   - место работы, с указанием почтового адреса организации и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах: http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp и http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   - аннотация (не менее 100 слов на каждом из языков). Аннотация — это краткое резюме работы, которое может публиковаться отдельно. Она является основным источником информации в информационных системах и базах данных. Английская аннотация должна быть оригинальной, может не быть дословным переводом русского текста и должна быть написана хорошим английским языком. В аннотации не должно быть ссылок на литературу и, по возможности, формул;
   - ключевые слова — желательно из принятых в мировой научно-технической литературе тематических тезаурусов. Предложения не могут быть ключевыми словами;
   - источники финансирования работы (ссылки на гранты, проекты, поддерживающие организации и т. п.).

9. Требования к спискам литературы.

   Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в каждом из списков литературы в порядке первых упоминаний. Если источник имеет DOI и/или EDN, то их необходимо указывать.

   Списки литературы представляются в двух вариантах:

   (1) **Список литературы к русскоязычной части**. Русские и английские работы — на языке и в алфавите оригинала;
   (2) **References**. Русские работы и работы на других языках — в латинской транслитерации с переводом на английский язык; английские работы и работы на других языках — на языке оригинала.

   Необходимо для составления списка “References” пользоваться размещенной на сайте http://www.translit.net/ru/bgn/ бесплатной программой транслитерации русского текста в латиницу.

   Список литературы “References” приводится полностью отдельным блоком, повторяя все позиции из списка литературы к русскоязычной части, независимо от того, имеются или нет в нем иностранные источники. Если в списке литературы к русскоязычной части есть ссылки на иностранные публикации, набранные латиницей, они полностью повторяются в списке “References”.

   Ниже приведены примеры ссылок на различные виды публикаций в списке “References”.

   Описание статьи из журнала:

   Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe hozyaystvo* [*Oil Industry*] 11:54–57.

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Russ. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Описание статьи из электронного журнала:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Описание статьи из продолжающегося издания (сборника трудов):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental’noe issledovanie prochnosti soedineniy “stal’–kompozit” [Experimental study of the strength of joints “steel–composite”]. *Trudy MGTU “Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem”* [*Bauman MSTU “Mathematical Modeling of Complex Technical Systems” Proceedings*]. 593:125–130.

Описание материалов конференций:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Описание книги (монографии, сборники):

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Latyshev, V. N. 2009. *Tribologiya rezaniya. Kn. 1: Friktsionnye protsessy pri rezanii metallov* [*Tribology of cutting. Vol. 1: Frictional processes in metal cutting*]. Ivanovo: Ivanovskii State Univ. 108 p.

Описание переводной книги (в списке литературы к русскоязычной части необходимо указать: / Пер. с англ. — после названия книги, а в конце ссылки указать оригинал книги в круглых скобках):

1. В русскоязычной части:
   Тимошенко С. П., Янг Д. Х., Уивер У. Колебания в инженерном деле / Пер. с англ. — М.: Машиностроение, 1985. 472 с. (*Timoshenko S. P., Young D. H., Weaver W.* Vibration problems in engineering. — 4th ed. — New York, NY, USA: Wiley, 1974. 521 p.)
2. В англоязычной части:
   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. New York: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004 (unpubl.). Geologiya i dobycha (NGT GiD) [Geology and production (NGT GiD)]. Certificate on official registration of the computer program No. 2004611198.

Описание интернет-ресурса:

Pravila tsitirovaniya istochnikov [Rules for the citing of sources]. Available at: http://www.scribd.com/doc/1034528/ (accessed February 7, 2011).

Описание диссертации или автореферата диссертации:

Semenov, V. I. 2003. Matematicheskoe modelirovanie plazmy v sisteme kompaktnyy tor [Mathematical modeling of the plasma in the compact torus]. Moscow. D.Sc. Diss. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. Moscow: IPI RAN. PhD Thesis. 23 p.

Описание ГОСТа:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. Moscow: Standardinform Publs. 10 p.

Описание патента:

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

10. Присланные в редакцию материалы авторам не возвращаются.
11. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
12. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Информатика и её применения»:
Москва 119333, ул. Вавилова, д. 44, корп. 2, ФИЦ ИУ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05
e-mail: iiep@frccsc.ru (Стригина Светлана Николаевна)
http://www.ipiran.ru/journal/issues/

Requirements for manuscripts submitted to Journal "Informatics and Applications"

Journal "Informatics and Applications" (Inform. Appl.) publishes theoretical, review, and discussion articles on the research and development in the field of informatics and its applications.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

The topics covered include the following areas:

- theoretical fundamentals of informatics;
- mathematical methods for studying complex systems and processes;
- information systems and networks;
- information technologies; and
- architecture and software of computational complexes and networks.

1. The Journal publishes original articles which have not been published before and are not intended for simultaneous publication in other editions. An article submitted to the Journal must not violate the Copyright law. Sending the manuscript to the Editorial Board, the authors retain all rights of the owners of the manuscript and transfer the nonexclusive rights to publish the article in Russian (or the language of the article, if not Russian) and its distribution in Russia and abroad to the Founders and the Editorial Board. Authors should submit a letter to the Editorial Board in the following form:

   Agreement on the transfer of rights to publish:

   "*We, the undersigned authors of the manuscript ". . . ", pass to the Founder and the Editorial Board of the Journal "Informatics and Applications" the nonexclusive right to publish the manuscript of the article in Russian (or in English) in both print and electronic versions of the Journal. We affirm that this publication does not violate the Copyright of other persons or organizations.*

   Author(s) signature(s): (name(s), address(es), date).

   This agreement should be submitted in paper form or in the form of a scanned copy (signed by the authors).

2. A submitted article should be attached with **the data on the author(s)** (see item 8). If there are several authors, the contact person should be indicated who is responsible for correspondence with the Editorial Board and other authors about revisions and final approval of the proofs.
3. The Editorial Board of the Journal examines the article according to the established reviewing procedure. If the authors receive their article for correction after reviewing, it does not mean that the article is approved for publication. The corrected article should be sent to the Editorial Board for the subsequent review and approval.
4. The decision on the article publication or its rejection is communicated to the authors. The Editorial Board may also send the reviews on the submitted articles to the authors. Any discussion upon the rejected articles is not possible.
5. The edited articles will be sent to the authors for proofread. The comments of the authors to the edited text of the article should be sent to the Editorial Board as soon as possible.
6. The manuscript of the article should be presented electronically in the MS WORD (.doc or .docx) or LaTeX (.tex) formats, and additionally in the .pdf format. All documents may be sent by e-mail or provided on a CD or diskette. A hard copy submission is not necessary.
7. The recommended typesetting instructions for manuscript.

   Pages parameters: format A4, portrait orientation, document margins (cm): left — 2.5, right — 1.5, above — 2.0, below — 2.0, footer 1.3.

   Text: font —Times New Roman, font size — 14, paragraph indent — 0.5, line spacing — 1.5, justified alignment.

   The recommended manuscript size: not more than 10 pages of the specified format. If the specified size exceeded, the editorial board is entitled to require the author to reduce the manuscript.

   Use only standard abbreviations. Avoid abbreviations in the title and abstract. The full term for which an abbreviation stands should precede its first use in the text unless it is a standard unit of measurement.

   All pages of the manuscript should be numbered.

   The templates for the manuscript typesetting are presented on site: http://www.ipiran.ru/journal/template.doc.
8. The articles should enclose data both in **Russian and English**:

   - title;
   - author's name and surname;
   - affiliation — organization, its address with ZIP code, city, country, and official e-mail address;
   - data on authors according to the format (see site):
     http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp and
     http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;

- abstract (not less than 100 words) both in Russian and in English. Abstract is a short summary of the article that can be published separately. The abstract is the main source of information on the article and it could be included in leading information systems and data bases. The abstract in English has to be an original text and should not be an exact translation of the Russian one. Good English is required. In abstracts, avoid references and formulae;
- indexing is performed on the basis of keywords. The use of keywords from the internationally accepted thematic Thesauri is recommended.

  Important! Keywords must not be sentences;
- Acknowledgments.

9. References. Russian references have to be presented both in English translation and Latin transliteration (refer http://www.translit.net/ru/bgn/).

   Please take into account the following examples of Russian references appearance:

   Article in journal:

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Russ. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Journal article in electronic format:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Article from the continuing publication (collection of works, proceedings):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental'noe issledovanie prochnosti soedineniy "stal'–kompozit" [Experimental study of the strength of joints "steel–composite"]. *Trudy MGTU "Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem"* [*Bauman MSTU "Mathematical Modeling of Complex Technical Systems" Proceedings*]. 593:125–130.

   Conference proceedings:

   Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

   Books and other monographs:

   Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

   Dissertation and Thesis:

   Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. Moscow: IPI RAN. PhD Thesis. 23 p.

   State standards and patents:

   GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. M.: Standardinform Publs. 10 p.

   Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

   References in Latin transcription are presented in the original language.

   References in the text are numbered according to the order of their first appearance; the number is placed in square brackets. All items from the reference list should be cited.

10. Manuscripts and additional materials are not returned to Authors by the Editorial Board.
11. Submissions of files by e-mail must include:
    - the journal title and author's name in the "Subject" field;
    - an article and additional materials have to be attached using the "attach" function;
    - an electronic version of the article should contain the file with the text and a separate file with figures.
12. "Informatics and Applications" journal is not a profit publication. There are no charges for the authors as well as there are no royalties.

Editorial Board address:
FRC CSC RAS, 44, block 2, Vavilov Str., Moscow 119333, Russia
Ph.: +7 (499) 135 86 92, Fax: +7 (495) 930 45 05
e-mail: iiep@frccsc.ru (to Svetlana Strigina)
http://www.ipiran.ru/english/journal.asp