СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Научный журнал Российской академии наук
(издается под руководством Отделения нанотехнологий и информационных технологий РАН)

Издается с 1989 года
Журнал выходит ежеквартально

Учредитель:
Федеральный исследовательский центр «Информатика и управление» Российской академии наук

РЕДАКЦИОННЫЙ СОВЕТ

академик РАН И.А. Соколов — председатель Редакционного совета
академик РАН Г.И. Савин
академик РАН А.Л. Стемпковский
член-корреспондент РАН Ю.Б. Зубарев
профессор Ш. Долев (S. Dolev, Beer-Sheva, Israel)
профессор Ю. Кабанов (Yu. Kabanov, Besancon, France)
профессор В. Ротарь (V. Rotar, San-Diego, USA)
профессор М. Финкельштейн (M. Finkelstein, Bloemfontein, South Africa)

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

академик РАН И.А. Соколов — главный редактор
профессор, д.ф.-м.н. С.Я. Шоргин — заместитель главного редактора
д.т.н. В.Н. Захаров
проф., д.ф.-м.н. А.И. Зейфман
проф., д.т.н. В.Д. Ильин
проф., д.т.н. К.К. Колин
проф., д.ф.-м.н. В.Ю. Королев
к.ф.-м.н. Р.В. Разумчик
д.ф.-м.н. В.И. Синицын
проф., д.т.н. И. Н. Синицын
проф., д.ф.-м.н. В.Г. Ушаков
к.ф.-м.н. А.К. Горшенин — отв. секретарь
к.ф.-м.н. С.А. Христочевский

Редакция

к.ф.-м.н. Е. Н. Арутюнов
к.ф.-м.н. Р. В. Разумчик
С. Н. Стригина

© Федеральный исследовательский центр «Информатика и управление» Российской академии наук, 2021

Журнал включен в базу данных Russian Science Citation Index (RSCI), интегрированную с Web of Science
Журнал входит в систему Российского индекса научного цитирования (РИНЦ)
Журнал включен в базу данных CrossRef (систему DOI — Digital Object Identifier), в базу данных Ulrich's periodicals directory и в информационную систему «Общероссийский математический портал Math-Net.Ru»

Журнал реферируется в «Реферативном журнале» ВИНИТИ и в системе Google Scholar

Журнал включен в сформированный Минобрнауки России Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук

http://www.ipiran.ru/journal/collected

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Том 31 № 1 Год 2021

СОДЕРЖАНИЕ

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Том 31 № 1 Год 2021

СОДЕРЖАНИЕ

МОДЕЛИ СОГЛАСОВАНИЯ СКРЫТОГО ПРОСТРАНСТВА В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ*

***Ф. Р. Яушев*[1], *Р. В. Исаченко*[2], *В. В. Стрижов*[3]**

Аннотация: Исследуется задача прогнозирования сложной целевой переменной. Под сложностью подразумевается наличие зависимостей, линейных или нелинейных. Предполагается, что исходные данные гетерогенны. Это значит, что пространства независимой и целевой переменных имеют разную природу. Предлагается построить предсказательную модель, которая учитывает зависимость в исходном пространстве независимой переменной, а также в пространстве целевой переменной. Согласование моделей предлагается проводить в низкоразмерном пространстве. В качестве базового алгоритма используется метод проекции в скрытое пространство (PLS — projection to latent space). В работе проводится сравнение линейного PLS и предложенных нелинейных моделей. Сравнение проводится на гетерогенных данных в пространствах высокой размерности.

Ключевые слова: прогнозирование; модель частичных наименьших квадратов; задача восстановления; согласование скрытого пространства

DOI: 10.14357/08696527210101

1 Введение

В данной работе решается задача прогнозирования целевой переменной с наличием зависимостей. Трудность задачи в том, что исходные данные имеют высокую размерность и в пространствах целевой и независимой переменных есть скрытые зависимости. Чрезмерно высокая размерность пространств и наблюдаемая множественная корреляция приводят к неустойчивости прогностической модели. Для решения задачи предлагается построить модель, которая бы учитывала обе эти зависимости. Она переводит данные в низкоразмерные пространства, и согласование данных происходит в полученном скрытом пространстве.

*Статья содержит результаты проекта «Математические методы интеллектуального анализа больших данных», выполняемого в рамках реализации программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по договору МГУ им. М. В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018. Работа выполнена при поддержке РФФИ (проекты 19-07-01155 и 19-07-00885).

[1]Московский физико-технический институт, fyaush@mail.ru

[2]Московский физико-технический институт, roman.isachenko@phystech.edu

[3]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук; Московский физико-технический институт, strijov@phystech.edu

Метод проекции в скрытое пространство (PLS) [1, 2] восстанавливает зависимости между двумя наборами данных. Он применяется в биоинформатике, медицине, социальных науках [3–6]. Алгоритм PLS строит матрицу совместного описания признаков и целевой переменной. Полученное пространство является низкоразмерным. Это позволяет получить простую, точную и устойчивую прогностическую модель. Наряду с PLS используется метод канонического анализа корреляций (Canonical Correlation Analysis, CCA) [7]. Метод CCA применяется для поиска зависимостей между двумя наборами данных и получения их низкоразмерного представления [8, 9]. Метод CCA максимизирует корреляции, а метод PLS — ковариации. Обзор и сравнение CCA и PLS приводится в [1]. Линейные методы PLS и CCA игнорируют сложные нелинейные зависимости.

Задачи, в которых между данными существует нелинейная зависимость, описаны в работе [2]. Аппроксимация этой зависимости линейной моделью PLS приводит к неудовлетворительным результатам. Разработаны нелинейные модификации PLS [10–12] и CCA [13, 14]. Например, модель Deep CCA [14] преобразует исходные данные с помощью нейронной сети таким образом, что результирующее представление становится согласованным. Метод Deep CCA используется для генерации текстового описания по изображениям в работе [15].

В данной работе исследуется сложность моделей для данных со сложноорганизованной целевой переменной. Для учета зависимостей в целевом пространстве используются проекции в скрытое пространство с помощью моделей PLS и CCA. В случае наличия существенно нелинейных зависимостей между независимой и целевой переменными сложность линейной модели оказывается недостаточной. В работе предлагаются методы согласования проекций для нелинейных моделей.

В работе проведены два эксперимента. Первый направлен на сравнение эффективности Deep CCA и CCA на задаче классификации зашумленных цифровых изображений MNIST [16]. Во втором эксперименте используется набор данных, полученный делением каждого изображения из MNIST на левую и правую части. На задаче регрессии правой части изображения по левой проводится сравнение нелинейных моделей с применением автоэнкодеров, моделей без преобразования данных и линейного PLS. На основании полученных результатов сделан вывод о точности и сложности нелинейных алгоритмов и о целесообразности использования той или иной модели.

2 Постановка задачи

Задана выборка $(\mathbf{X}, \mathbf{Y})$, где $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times m}$ — матрица независимых переменных; $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times k}$ — матрица целевых переменных. Предполагается, что между $\mathbf{X}$ и $\mathbf{Y}$ существует зависимость

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon}, \tag{1}$$

где $f : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times k}$ — функция регрессионной зависимости; $\boldsymbol{\varepsilon}$ — матрица регрессионных ошибок. Необходимо восстановить зависимость f по заданной выборке.

2.1 Линейная регрессия

Предположим, что зависимость (1) линейна. Требуется найти эту зависимость:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} = \mathbf{X}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\varepsilon},$$

где $\mathbf{W} \in \mathbb{R}^{k\times m}$ – матрица параметров модели.

Оптимальные параметры определяются минимизацией функции потерь. Используется квадратичная функция потерь

$$\mathcal{L}(\mathbf{W}|\mathbf{X},\mathbf{Y}) = \left\| \underset{n\times k}{\mathbf{Y}} - \underset{n\times m}{\mathbf{X}} \cdot \underset{m\times k}{\mathbf{W}}^{\mathrm{T}} \right\|_2^2 \to \min_{\mathbf{W}}. \tag{2}$$

Решение задачи (2) имеет вид:

$$\mathbf{W} = \mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

Линейная зависимость столбцов матрицы $\mathbf{X}$ приводит к неустойчивости решения задачи минимизации (2), так как в этом случае матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ является плохо обусловленной. Для борьбы с линейной зависимостью используются методы снижения размерности путем перехода в низкоразмерное латентное пространство.

Определение 2.1. Параметрическая функция $\varphi_1 : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times p}$, переводящая исходные данные в латентное пространство, называется **функцией кодирования**.

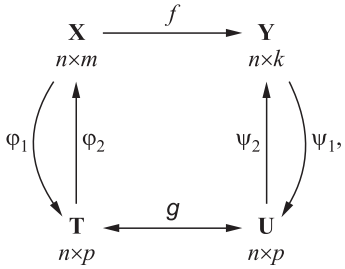
Определение 2.2. Функция $\varphi_2 : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times m}$, переводящая данные из латентного пространства в исходное, называется **функцией восстановления**.

Определение 2.3. Функция $g : \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} \to \mathbb{R}$, связывающая закономерности в низкоразмерных латентных представлениях, называется **функцией согласования**.

Определение 2.4. Согласование — алгоритмическая процедура максимизации функции согласования.

2.2 Снижение размерности

Коммутативная диаграмма процедуры выбора прогностической модели имеет вид



где $\varphi_1 : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times p}$ — функция кодирования независимых переменных; $\psi_1 : \mathbb{R}^{n\times k} \to \mathbb{R}^{n\times p}$ — функция кодирования целевых переменных; $\varphi_2 : \mathbb{R}^{n\times p} \to \to \mathbb{R}^{n\times m}$ — функция восстановления независимых переменных; $\psi_2 : \mathbb{R}^{n\times p} \to \to \mathbb{R}^{n\times k}$ — функция восстановления целевых переменных; $g : \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} \to \to \mathbb{R}$ — функция согласования. Матрицы $\mathbf{T} = \varphi_1(\mathbf{X}) \in \mathbb{R}^{n\times p}$ и $\mathbf{U} = \psi_1(\mathbf{Y}) \in \in \mathbb{R}^{n\times p}$ являются матрицами представлений данных в латентном пространстве низкой размерности.

Оптимальные параметры $\theta^*_{\varphi_1}$ и $\theta^*_{\psi_1}$ для функций кодирования φ_1 и ψ_1 находятся из следующей задачи параметрической оптимизации:

$$\left(\theta^*_{\varphi_1}, \theta^*_{\psi_1}\right) = \underset{(\theta_{\varphi_1}, \theta_{\psi_1})}{\arg\max}\, g\left(\varphi_1(\mathbf{X}; \theta_{\varphi_1}), \psi_1(\mathbf{Y}; \theta_{\psi_1})\right). \tag{3}$$

Так как параметры функции кодирования подбираются из условия максимизации функции согласования (3), то после перехода в латентное пространство между $\mathbf{T}$ и $\mathbf{U}$ существует зависимость:

$$\mathbf{U} = h(\mathbf{T}) + \boldsymbol{\eta},$$

где $h : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ — функция регрессионной зависимости; $\boldsymbol{\eta}$ — матрица регрессионных ошибок. Оптимальная функция h выбирается минимизацией функции ошибки. Используется квадратичная функция потерь $\mathcal{L}$ на $\mathbf{T}$ и $\mathbf{U}$:

$$\mathcal{L}(h|\mathbf{T}, \mathbf{U}) = \left\| \underset{n\times p}{\mathbf{U}} - h(\underset{m\times p}{\mathbf{T}}) \right\|_2^2 \to \min_h.$$

Финальная прогностическая модель имеет вид

$$\widehat{\mathbf{y}} = \psi_2\left(h(\varphi_1(\mathbf{x}))\right),$$

т. е.

$$f = \psi_2 \circ h \circ \varphi_1.$$

2.3 Метод главных компонент

Метод главных компонент (PCA — principal component analysis) снижает размерность данных и сохраняет максимальную дисперсию. Линейная модель PCA представляет собой ортогональное линейное преобразование исходного признакового пространства в новое пространство меньшей размерности. Первый базисный вектор строится так, чтобы выборочная дисперсия столбцов проекций матрицы $\mathbf{X}$ была максимальной:

$$\mathbf{p} = \arg\max_{\|\mathbf{p}\|_2=1} [\mathbf{var}(\mathbf{Xp})],$$

где $\mathbf{var}(\mathbf{Xp}) = (1/n)(\mathbf{Xp})^{\mathrm{T}}\mathbf{Xp}$ обозначает выборочную дисперсию. Последующие базисные векторы находятся итеративно после вычитания проекции на все найденные ранее.

Функция кодирования $\varphi_1 : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times p}$ имеет вид:

$$\varphi_1(\mathbf{X}) = \underset{n\times m}{\mathbf{X}} \cdot \underset{m\times p}{\mathbf{P}}^{\mathrm{T}},$$

где $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_p]$. Метод PCA не согласует независимые переменные и целевые переменные. Из-за этого зависимости в обоих пространствах не учитываются.

2.4 Метод частичных наименьших квадратов

Метод частичных наименьших квадратов восстанавливает связь между двумя наборами данных $\mathbf{X}$ и $\mathbf{Y}$. Алгоритм проецирует $\mathbf{X}$ и $\mathbf{Y}$ на латентное пространство $\mathbb{R}^p$ меньшей размерности. Метод PLS находит матрицы исходных данных $\mathbf{X}$ и $\mathbf{Y}$ в латентном пространстве $\mathbf{T}$ и $\mathbf{U}$ соответственно. Матрица объектов $\mathbf{X}$ и целевая матрица $\mathbf{Y}$ проецируются на латентное пространство следующим образом:

$$\underset{n\times m}{\mathbf{X}} = \underset{n\times p}{\mathbf{T}} \cdot \underset{p\times m}{\mathbf{P}}^{\mathrm{T}} + \underset{n\times m}{\mathbf{F}};$$

$$\underset{n\times k}{\mathbf{Y}} = \underset{n\times p}{\mathbf{U}} \cdot \underset{p\times k}{\mathbf{Q}}^{\mathrm{T}} + \underset{n\times k}{\mathbf{E}},$$

где $\mathbf{T}$ и $\mathbf{U}$ — матрицы описания объектов и исходов в латентном пространстве; $\mathbf{P}$ и $\mathbf{Q}$ — матрицы перехода из латентного пространства в исходное; $\mathbf{F}$ и $\mathbf{E}$ — матрицы остатков.

В методе PLS функции кодирования имеют вид:

$$\varphi_1(\mathbf{X}) = \mathbf{XW_x}; \quad \psi_1(\mathbf{Y}) = \mathbf{YW_y},$$

где матрицы весов $\mathbf{W_x} \in \mathbb{R}^{m\times p}$ и $\mathbf{W_y} \in \mathbb{R}^{k\times p}$ находятся путем максимизации функции согласования $g(\mathbf{XW_x}, \mathbf{YW_y}) = \mathbf{Cov}(\mathbf{XW_x}, \mathbf{YW_y})^2$:

$$(\mathbf{W_x}, \mathbf{W_y}) = \arg\max_{\mathbf{W_y},\mathbf{W_y}} [\mathbf{Cov}(\mathbf{XW_x}, \mathbf{YW_y})^2],$$

где $\mathbf{Cov}(\mathbf{XW_x}, \mathbf{YW_y})$ — выборочная ковариация.

Функции восстановления принимают вид:

$$\varphi_2(\mathbf{T}) = \mathbf{T}\mathbf{P}^{\mathrm{T}}; \quad \psi_2(\mathbf{U}) = \mathbf{U}\mathbf{Q}^{\mathrm{T}}.$$

2.5 Канонический анализ корреляций

Канонический анализ корреляций находит два набора базисных векторов $\{\mathbf{w}_{\mathbf{x}_i}\}_{i=1}^{p}$, $\mathbf{w}_{\mathbf{x}} \in \mathbb{R}^m$, и $\{\mathbf{w}_{\mathbf{y}_i}\}_{i=1}^{p}$, $\mathbf{w}_{\mathbf{y}} \in \mathbb{R}^k$, — один для матрицы $\mathbf{X}$, другой для матрицы $\mathbf{Y}$, так чтобы коэффициент корреляция между проекциями переменных на эти базисные векторы был максимальным. Функция согласования для CCA имеет вид:

$$g\left(\mathbf{X}\mathbf{W}_{\mathbf{x}}, \mathbf{Y}\mathbf{W}_{\mathbf{y}}\right) = \mathbf{corr}\left(\mathbf{X}\mathbf{W}_{\mathbf{x}}, \mathbf{Y}\mathbf{W}_{\mathbf{y}}\right),$$

где $\mathbf{corr}(\mathbf{X}\mathbf{w}_{\mathbf{x}}, \mathbf{Y}\mathbf{w}_{\mathbf{y}})$ — коэффициент корреляции между векторами.

Таким образом, функции кодирования имеют вид:

$$\varphi_1(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathbf{x}}; \quad \psi_1(\mathbf{Y}) = \mathbf{Y}\mathbf{W}_{\mathbf{y}},$$

где первые столбцы матриц весов находятся как векторы, максимизирующие функцию согласования g. Далее ищутся векторы, максимизирующие g, но с ограничением, что они не коррелируют с первой парой векторов. Процедура продолжается до тех пор, пока число векторов не станет равным p.

2.6 Нелинейный канонический анализ корреляций

Нелинейный канонический анализ корреляций — нелинейная модификация CCA. Метод Deep CCA преобразует исходные данные с помощью нейронной сети таким образом, что результирующее представление становится согласованным. В данной работе рассматриваются следующие нелинейные функции кодирования и восстановления:

$$\begin{aligned}
\mathbf{T} &= \varphi_1(\mathbf{X}) = \mathbf{W}_{\mathbf{x}}^{L}\sigma(\cdots\sigma(\mathbf{W}_{\mathbf{x}}^{2}\sigma(\mathbf{X}\mathbf{W}_{\mathbf{x}}^{1}))\cdots);\\
\mathbf{U} &= \psi_1(\mathbf{Y}) = \mathbf{W}_{\mathbf{y}}^{L}\sigma(\cdots\sigma(\mathbf{W}_{\mathbf{y}}^{2}\sigma(\mathbf{Y}\mathbf{W}_{\mathbf{y}}^{1}))\cdots);\\
\mathbf{X} &= \varphi_2(\mathbf{X}) = \mathbf{W}_{\mathbf{t}}^{L}\sigma(\cdots\sigma(\mathbf{W}_{\mathbf{t}}^{2}\sigma(\mathbf{T}\mathbf{W}_{\mathbf{t}}^{1}))\cdots);\\
\mathbf{Y} &= \psi_2(\mathbf{Y}) = \mathbf{W}_{\mathbf{u}}^{L}\sigma(\cdots\sigma(\mathbf{W}_{\mathbf{u}}^{2}\sigma(\mathbf{U}\mathbf{W}_{\mathbf{u}}^{1}))\cdots).
\end{aligned}$$

Каждая функция представляет нейронную сеть с L скрытыми слоями.

Требуется найти такие параметры, при которых функция согласования g достигает своего максимума:

$$g(\mathbf{T}, \mathbf{U}) \to \max_{\mathbf{W}}, \tag{4}$$

где $\mathbf{W} = \left\{\left\{\mathbf{W}_{\mathbf{x}}^{i}\right\}_{i=1}^{L}, \left\{\mathbf{W}_{\mathbf{y}}^{i}\right\}_{i=1}^{L}, \left\{\mathbf{W}_{\mathbf{t}}^{i}\right\}_{i=1}^{L}, \left\{\mathbf{W}_{\mathbf{u}}^{i}\right\}_{i=1}^{L}\right\}$.

3 Вычислительный эксперимент

Цель вычислительного эксперимента — сравнительный анализ рассматриваемых моделей. Рассматриваются данные, для которых сложность класса линейных методов неадекватно низка. Нелинейные модели позволяют получить точный прогноз при адекватной сложности. В рамках вычислительного эксперимента написан программный комплекс для решения поставленных задач [17].

3.1 Анализ нелинейных зависимостей в задаче фильтрации шума

Проведем сравнение качества Deep CCA и CCA на задаче классификации зашумленных цифровых изображений, представленных на рис. 1. Для этого используется набор данных MNIST [16], который состоит из 70 000 цифровых изображений 28×28 образцов рукописного написания цифр. Предлагается получить два новых набора данных $\mathbf{X}$ и $\mathbf{Y}$ следующим образом. Первый набор получается поворотом исходных изображений на угол в диапазоне $[-\pi/4, \pi/4]$. Для получения второго набора данных для каждой картинки из первого набора данных ставится в соответствие случайным образом картинка с той же цифрой, но с добавлением независимого случайного шума, распределенного равномерно на отрезке $[0, 1]$.

Применив к двум новым наборам данных DeepCCA или CCA, получаем новое низкоразмерное признаковое пространство, которое игнорирует шумы в исходных данных. Таким образом, получаем функции кодирования φ_1 и ψ_1 для исходных наборов данных. На новых признаках, полученных разными моделями

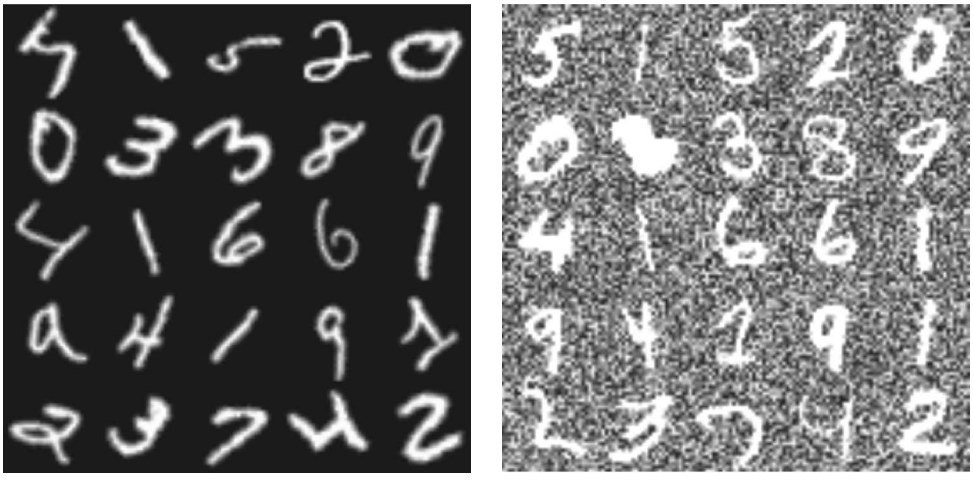


Рис. 1 Зашумленные изображения из набора данных MNIST

Таблица 1 Точность классификации линейного SVM для алгоритмов Deep CCA и CCA

Скользящий контроль	Deep CCA ($L = 3$)	CCA
Валидация	92,74%	76,21%
Тест	92,14%	76,07%

(DeepCCA и CCA), для первого набора данных, т. е. на данных после применения функции кодирования φ_1 к первому набору исходных данных, обучим линейный SVM-классификатор (SVM — support vector machine). Показателем эффективности будет точность классификации линейного SVM на тестовых данных. В случае построения адекватного скрытого пространства полученные образы объектов будут линейно разделимы. Результаты эксперимента приведены в табл. 1. Модель Deep CCA представляет собой нейронную сеть с $L = 3$ скрытыми слоями. Точность классификации нелинейной модели существенно выше линейного алгоритма CCA.

3.2 Анализ нелинейных моделей для восстановления изображений

Для анализа процедуры согласования проведен вычислительный эксперимент с предложенными нелинейными моделями. Для снижения размерности пространства используются нейросетевые модели автокодировщика с согласованием скрытого пространства (4). В качестве базовых моделей используются модель автокодировщика без согласования скрытых пространств, а также линейный PLS. В качестве исходного набора данных используется набор данных MNIST [16]. Каждое изображение поделено на левую и правую части, как показано на рис. 2. Модель по левому изображению восстанавливает правое изображение.

Модель EncNet1 — нейронная сеть с нелинейными функциями активации, которая обучается на данных после преобразования их автоэнкодером. Модель LinNet1 — нейронная сеть с одним линейным слоем, которая также обучается на преобразованных данных. Для EncNet1 и LinNet1 автоэнкодеры для объектов

Рис. 2 Набор данных MNIST, в котором каждое изображение разделено пополам

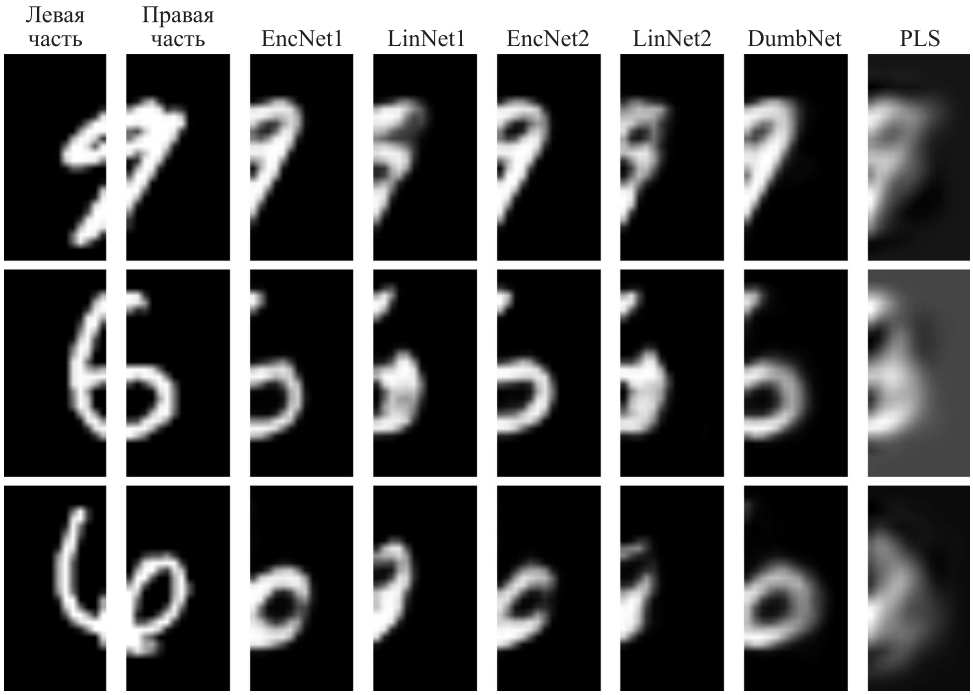


Рис. 3 Пример реконструкции правой части изображения по левой для рассматриваемых моделей

и ответов используют совместную функцию потерь, которая связывает выходы энкодеров. Модели EncNet2 и LinNet2 устроены аналогично EncNet1 и LinNet1 соответственно, но в автоэнкодерах нет совместной функции потерь. Модель DumbNet — нейронная сеть, которая обучается на исходных данных и имеет такую же структуру, что и EncNet, т. е. имеет такое же число слоев и в каждом слое такое же количество нейронов, что и у EncNet.

Для оценки качества моделей вычислялась среднеквадратичная ошибка. Примеры восстановленных изображений показаны на рис. 3. Качество моделей, а также их сложность представлены в табл. 2. На рис. 3 продемонстрировано, что предложенные модели EncNet и LinNet позволяют получить более четкие и различимые изображения в отличие от базовой нелинейной модели DumbNet и линейной модели PLS. Несмотря на заметное улучшение визуального качества изображений, ошибка предложенных моделей выше, чем у модели DumbNet. Авторы предполагают, что это связано с тем, что среднеквадратичная ошибка оказалась неадекватной метрикой в пространстве изображений. Нахождение оптимальной метрики для оценки качества предложенных алгоритмов может быть одним из возможных направлений развития текущей работы.

Таблица 2 Квадратичная ошибка для нелинейных моделей в задаче восстановления правой части изображения по левой

Метод	Число параметров	Ошибка на тесте
EncNet1	283 тыс.	0,147
LinNet1	239 тыс.	0,235
EncNet2	283 тыс.	0,149
LinNet2	239 тыс.	0,236
DumbNet	283 тыс.	0,128
PLS	—	0,188

4 Заключение

В работе рассмотрена задача восстановления для сложноорганизованной целевой переменной. Рассмотрены линейные модели согласования образов объектов в скрытом пространстве. При наличии сложных нелинейных зависимостей между независимой и целевой переменной сложности линейной модели оказывается недостаточно. Для построения точного прогноза приводятся нелинейные обобщения рассматриваемых линейных методов. В экспериментах на реальных данных изображений рукописных цифр показана адекватность рассматриваемых нелинейных моделей, а также проведен анализ различных способов согласования.

Литература

1. *Rosipal R., Kramer N., Graves A.* Overview and recent advances in partial least squares // Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop / Eds. C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor. — Lecture notes in computer science ser. — Springer, 2005. Vol. 3940. P. 34–51.
2. *Rosipal R.* Nonlinear partial least squares: An overview // Chemoinformatics advanced machine learning perspectives: Complex computational methods and collaborative techniques / Eds. H. Lodhi, Y. Yamanishi. — IGI Global, 2011. P. 169–189.
3. *Worsley K. J.* An overview and some new developments in the statistical analysis of pet and fmri data // Hum. Brain Mapp., 1997. Vol. 5. P. 254–258.
4. *Hulland J. S.* Use of partial least squares (pls) in strategic management research: A review of four recent studies // Strateg. Manage. J., 1999. Vol. 20. P. 195–204.
5. *Shalamu Abudu P. E., Pagano T. C.* Application of partial least-squares regression in seasonalstreamflow forecasting // J. Hydrol. Eng., 2010. Vol. 15. P. 612–623.
6. *Nguyen D. V., Rocke D. M.* Tumor classification by partial least squares using microarray gene expression data // Bioinformatics, 2012. Vol. 18. P. 39–50.
7. *Szedmak S. R., Hardoon D. R., Shawe-Taylor J. R.* Canonical correlation analysis: An overview with application to learning methods // Neural Comput., 2004. Vol. 16. P. 2639–2664.

8. *Schechner Y. Y., Kidron E., Elad M.* Pixels that sound // Computer Vision and Pattern Recognition Conference. — IEEE Computer Society, 2005. Vol. 1. P. 88–95.
9. *Sun S., Ji L., Ye J.* A least squares formulation for canonical correlation analysis // 25th Conference (International) on Machine Learning Proceedings. — ACM, 2008. P. 1024–1031.
10. *Qin S. J., McAvoy T. J.* Nonlinear pls modeling using neural networks // Comput. Chem. Eng., 1992. Vol. 16. P. 379–391.
11. *Hiden M., McKay B., Montague G.* Non-linear partial least squares using genetic programming // Genetic programming. — San Francisco, CA, USA: Morgan Kaufmann, 1998. P. 128–133.
12. *Chen D. Z., Yan X. F., Hu S. X.* Chaos-genetic algorithms for optimizing the operating conditions based on rbf-pls model // Comput. Chem. Eng., 2003. Vol. 27. P. 1393–1404.
13. *Lai P. L., Fyfe C.* Kernel and nonlinear canonical correlation analysis // Int. J. Neural Systems, 2000. Vol. 10. P. 365–377.
14. *Galen A., Arora R., Bilmes J., Livensen K.* Deep canonical correlation analysis // PMLR, 2013. Vol. 28. No. 3. P. 1247–1255.
15. *Yan F., Mikolajczyk K.* Deep correlation for matching images and text // Proc. CVPR IEEE, 2015. Vol. 4. P. 3441–3450.
16. *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/index.html.
17. *Yaushev F. Yu., Isachenko R. V.* Модели согласования скрытого пространства в задаче прогнозирования, 2020. https://github.com/Fyaushev/2020-Project-72.

Поступила в редакцию 15.12.20

CONCORDANT MODELS FOR LATENT SPACE PROJECTIONS IN FORECASTING

***F. Yu. Yaushev*[1], *R. V. Isachenko*[1], *and V. V. Strijov*[1,2]**

[1]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation
[2]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

Abstract: The paper examines the problem of predicting a complex structured target variable. Complexity refers to the presence of dependencies, whether linear or nonlinear. The source data are assumed to be heterogeneous. This means that the spaces of the independent and target variables are of different nature. It is proposed to build a predictive model that takes into account the dependence in the input space of the independent variable as well as in the space of the

target variable. It is proposed to make a model agreement procedure in a low-dimensional latent space. The projection to the latent space method is used as the basic algorithm. The paper compares the linear and proposed nonlinear models. The comparison is performed on heterogeneous data in high-dimensional spaces.

Keywords: prediction; partial least squares; model concordance; nonlinear projection to latent space

DOI: 10.14357/08696527210101

Acknowledgments

The paper contains results of the project "Mathematical methods for intelligent big data analysis" which is carried out within the framework of the Program "Center of Big Data Storage and Analysis" of the National Technology Initiative Competence Center. It is supported by the Ministry of Science and Higher Education of the Russian Federation according to the agreement between the M. V. Lomonosov Moscow State University and the Foundation of Project Support of the National Technology Initiative from 11.12.2018, No. 13/1251/2018. This research was supported by RFBR (projects 19-07-01155 and 19-07-00885).

References

1. Rosipal, R., N. Kramer, and A. Graves. 2005. Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop*. Eds. C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor. Lecture notes in computer science ser. Springer. 3940:34–51.
2. Rosipal, R. 2011. Nonlinear partial least squares: An overview. *Chemoinformatics and advanced machine learning perspectives: Complex computational methods and collaborative techniques*. Eds. H. Lodhi and Y. Yamanishi. IGI Global. 169–189.
3. Worsley, K. J. 1997. An overview and some new developments in the statistical analysis of pet and fmri data. *Hum. Brain Mapp.* 5:254–258.
4. Hulland, J. S. 1999. Use of partial least squares (pls) in strategic management research: A review of four recent studies. *Strateg. Manage. J.* 20:195–204.
5. Shalamu Abudu, P. E., and T. C. Pagano. 2010. Application of partial least-squares regression in seasonalstreamflow forecasting. *J. Hydrol. Eng.* 15:612–623.
6. Nguyen, D. V., and D. M. Rocke. 2012. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18:39–50.
7. Szedmak, S. R., D. R. Hardoon, and J. R. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* 16:2639–2664.
8. Schechner, Y. Y., E. Kidron, and M. Elad. 2005. Pixels that sound. *Computer Vision and Pattern Recognition Conference*. IEEE Computer Society. 1:88–95.
9. Sun, S., L. Ji, and J. Ye. 2008. A least squares formulation for canonical correlation analysis. *25th Conference (International) on Machine Learning Proceedings*. ACM. 1024–1031.

10. Qin, S. J., and T. J. McAvoy. 1992. Nonlinear pls modeling using neural networks. *Comput. Chem. Eng.* 16:379–391.
11. Hiden, M., B. McKay, and G. Montague. 1998. Non-linear partial least squares using genetic programming. *Genetic programming*. San Francisco, CA: Morgan Kaufmann. 128–133.
12. Chen, D. Z., X. F. Yan, and S. X. Hu. 2003. Chaos-genetic algorithms for optimizing the operating conditions based on rbf-pls model. *Comput. Chem. Eng.* 27:1393–1404.
13. Lai, P. L., and C. Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Systems* 10:365–377.
14. Galen, A., R. Arora, J. Bilmes, and K. Livensen. 2013. Deep canonical correlation analysis. *PMLR* 28(3):1247–1255.
15. Yan, F., and K. Mikolajczyk. 2015. Deep correlation for matching images and text. *Proc. CVPR IEEE* 4:3441–3450.
16. LeCun, Y., C. Cortes, and C. Burges. 1998. The MNIST dataset of handwritten digits. Available at: http://yann. lecun.com/exdb/mnist/index.html (accessed February 24, 2021).
17. Yaushev, F. Yu, and R. V. Isachenko. 2020. Modeli soglasovaniya skrytogo prostranstva v zadache prognozirovaniya [Concordant models for latent space projections in the complex structured prediction tasks]. Available at: https://github.com/Fyaushev/2020-Project-72 (accessed February 24, 2021).

Received December 15, 2020

Contributors

Yashev Faruh Yu. (b. 1999) — student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; fyaush@mail.ru

Isachenko Roman V. (b. 1994) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; roman.isachenko@phystech.edu

Strijov Vadim V. (b. 1967) — Doctor of Science in physics and mathematics, leading 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; professor, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; strijov@ccas.ru

ИССЛЕДОВАНИЕ МОДЕЛИ ТИПА $M_t/M_t/1$ С ДВУМЯ РАЗЛИЧНЫМИ КЛАССАМИ ТРЕБОВАНИЙ*

Я. А. Сатин[1]

Аннотация: Исследуется нестационарная модель массового обслуживания $M_t/M_t/1$ с одним сервером и двумя классами требований. Для этой модели получен одномерный процесс рождения и гибели (ПРГ), описывающий число требований в исходной системе. С помощью стандартного метода логарифмической нормы линейной операторной функции получены соответствующие оценки скорости сходимости и условия эргодичности, а также построен численный пример, показывающий применение изучаемого подхода. Приведены графические иллюстрации, построенные на основе общего алгоритма, разработанного в предыдущих работах и связанного с решением задачи Коши для прямой системы Колмогорова на соответствующем временном интервале.

Keywords: системы массового обслуживания; нестационарная модель массового обслуживания; одномерный процесс рождения и гибели; скорость сходимости; оценки эргодичности; логарифмическая норма; модель $M_t/M_t/1$ с двумя классами требований

DOI: 10.14357/08696527210102

1 Введение

В работе исследуется нестационарная модель системы массового обслуживания с одним сервером и двумя классами требований, а также одномерный ПРГ, получающийся из нее и описывающий число требований в исходной системе.

Для оценки скорости сходимости используется обобщенное понятие логарифмической нормы, введенное в [1]. Для удобства напомним, как оно вводится. Рассматривается система дифференциальных уравнений вида

$$\frac{d\mathbf{y}}{dt} = H\mathbf{y}(t) \tag{1}$$

в пространстве l_1 в предположении существования и единственности решения задачи Коши для любого начального условия $\mathbf{y}(0)$.

Тогда верна следующая оценка:

$$\frac{d\|y\|}{dt} = \sum_i \frac{d|y_i|}{dt} \le \sum_i \left(h_{ii}|y_i| + \sum_{j\ne i} |h_{ij}||y_j| \right) \le \beta^*\|y\|, \tag{2}$$

*Исследование выполнено за счет гранта Российского научного фонда (проект 19-11-00020.)

[1]Вологодский государственный университет, yacovi@mail.ru

где

$$\beta^* = \sup_i \left(h_{ii} + \sum_{j \neq i} |h_{ji}| \right).$$

Отметим, что элементы h_{ij} матрицы $H = (h_{ij})$ могут зависеть как от времени t, так и от $\mathbf{y}(t)$.

Из (2) следует

$$\|y(t)\| \leq e^{\int_0^t \beta^* \, du} \|y(0)\|.$$

При этом, если H является ограниченным при всех $t \geq 0$ линейным оператором из l_1 в себя, $\beta^*(t) = \gamma(H(t))$ совпадает с логарифмической нормой [2, 3]. Таким образом, $\beta^*(t)$ представляет собой обобщение логарифмической нормы для нелинейной системы уравнений.

В разд. 2 рассмотрено описание модели, в разд. 3 и 4 изучен процесс, описывающий число требований в системе, а затем связанный с ним одномерный процесс. В разд. 5 рассмотрен пример, иллюстрирующий изучаемый подход.

2 Описание модели

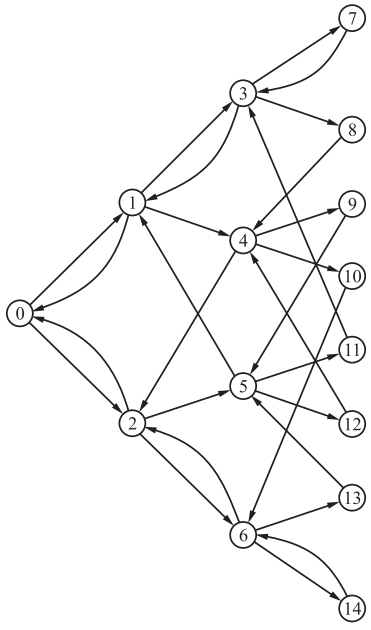


Рис. 1 Граф процесса $X(t)$

Рассмотрим систему обслуживания с двумя классами требований, которые поступают с интенсивностями $\lambda_1(t)$ и $\lambda_2(t)$ соответственно, а обслуживаются с интенсивностями $\mu_1(t)$ и $\mu_2(t)$ соответственно (см. общее описание такого типа и более общих, но стационарных моделей в [4]). Требования из очереди обслуживаются в порядке их поступления в систему.

Стандартное описание проводится с помощью двумерного ПРГ $Z(t) = (Z_1(t), Z_2(t))$, где каждая координата описывает число требований соответствующего класса, а интенсивности рождения и гибели для них — λ_1, μ_1, λ_2 и μ_2.

В настоящей работе рассмотрен другой подход, а именно: вводится одномерный процесс $X(t)$, в котором учтено как число требований каждого класса, так и то, на каком месте в очереди находятся требования первого и второго класса.

Для наглядности приведен граф с начальными состояниями процесса $X(t)$ и нумерацией состояний (рис. 1).

Вершина 0 соответствует ситуации, когда в системе нет требований. Вершина 1 (2) соответствует тому, что требование первого (второго) класса находится на обслуживании, а очередь за ним пуста. Вершины с третьей по шестую соответствуют наличию одного обслуживаемого требования и одного требования за ним в очереди (например, вершина 6 соответствует тому, что обслуживается требование второго класса и в очереди тоже требование второго класса) и т. д. Стрелки показывают возможные переходы с соответствующими интенсивностями: вправо и вверх — λ_1; вправо и вниз — λ_2; влево и вниз — μ_1; влево и вверх — μ_2.

При стандартных условиях вектор вероятностей состояний можно описать прямой системой Колмогорова

$$\frac{d\mathbf{p}}{dt} = A(t)\mathbf{p}, \tag{3}$$

где матрица A имеет вид:

$$A = \left(\begin{array}{c|cc|cccc|cccccccc|c}
 & \mu_1 & \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\hline
\lambda_1 & & 0 & \mu_1 & 0 & \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\lambda_2 & 0 & & 0 & \mu_1 & 0 & \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\hline
0 & \lambda_1 & 0 & & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & \mu_2 & 0 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & 0 & & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & \mu_2 & 0 & 0 & \cdots \\
0 & 0 & \lambda_1 & 0 & 0 & & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & \mu_2 & 0 & \cdots \\
0 & 0 & \lambda_2 & 0 & 0 & 0 & & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & \mu_2 & \cdots \\
\hline
0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \cdots \\
\hline
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots &
\end{array}\right). \tag{4}$$

Главная диагональ заполнена так, что суммы по столбцам равны нулю. В общем случае интенсивности являются функциями времени t.

3 Оценки для процесса $X(t)$

Возьмем два разных неотрицательных решения системы (3) $\mathbf{p}^*(t)$ и $\mathbf{p}^{**}(t)$ с единичной нормой. Тогда

$$\sum_i (p_i^*(t) - p_i^{**}(t)) = 0\,.$$

Это означает, что на решение системы не будет влиять изменение целой строки в матрице A на одну и ту же функцию. Следовательно, оценку скорости сходимости можно свести к изучению скорости сходимости системы

$$\frac{d(\mathbf{p}^* - \mathbf{p}^{**})}{dt} = (A - C)\,(\mathbf{p}^* - \mathbf{p}^{**})\,.$$

Здесь матрица C состоит из одних нулей за исключением самой верхней строки, каждый элемент которой равен неотрицательной c.

Для использования логарифмической нормы введем диагональную матрицу

$$D = \operatorname{diag}(1, k_1, k_1, k_1k_2, k_1k_2, k_1k_2, k_1k_2, k_1k_2k_3, \ldots)\,.$$

На главной диагонали произведения $k_1 \cdots k_s$ встречаются 2^s раз.

Найдем условия нуль-эргодичности.

Для этого возьмем в D все $k_i < 1$ и $c = 0$.

Если взять все k_i одинаковыми и равными k, получим:

$$\alpha(t) = \min \alpha_i(t) = \min\left(-(k-1)(\lambda_1 + \lambda_2) - \left(\frac{1}{k} - 1\right)\max(\mu_1, \mu_2)\right).$$

Очевидно, что если $\max(\mu_1, \mu_2) < \lambda_1 + \lambda_2$, то всегда можно подобрать $k < 1$, при которых α положительно. Отсюда получаем теорему.

Теорема 1. *Если* $\max(\mu_1, \mu_2) < (\lambda_1 + \lambda_2)$*, то процесс нуль-эргодичен.*

Найдем условия сильной эргодичности.

Для этого возьмем в D все $k_i > 1$.

Тогда матрица $D(A - C)D^{-1}$ примет вид, показанный на рис. 2.

Сумма по нулевому столбцу с противоположным знаком (вне главной диагонали элементы берутся по модулю)

$$\alpha_0 = c - (k_1 - 1)\,(\lambda_1 + \lambda_2)\,.$$

Суммы по остальным столбцам можем аналогично ограничить (i относится сразу к выделенной на рис. 2 группе столбцов):

$$\alpha_i = -\frac{c}{k_1 \cdots k_i} - (k_{i+1} - 1)\,(\lambda_1 + \lambda_2) - \left(\frac{1}{k_i} - 1\right)\min(\mu_1, \mu_2)\,.$$

Если взять все k_i одинаковыми и равными k,

$$\alpha = \min \alpha_i(t) =$$
$$= \min\left(-\frac{c}{k} - (k-1)(\lambda_1 + \lambda_2) - \left(\frac{1}{k} - 1\right)\min(\mu_1, \mu_2), c - (k-1)(\lambda_1 + \lambda_2)\right).$$

	$\frac{\mu_1-c}{k_1}$	$\frac{\mu_2-c}{k_1}$	$-c$ k_1k_2	$-c$ k_1k_2	$-c$ k_1k_2	$-c$ k_1k_2	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$-c$ $k_1k_2k_3$	$\dots$
$k_1\lambda_1$		0	$\frac{\mu_1}{k_2}$	0	$\frac{\mu_2}{k_2}$	0	0	0	0	0	0	0	0	$\dots$
$k_1\lambda_2$	0		0	$\frac{\mu_1}{k_2}$	0	$\frac{\mu_2}{k_2}$	0	0	0	0	0	0	0	$\dots$
0	$k_2\lambda_1$	0		0	0	0	$\frac{\mu_1}{k_3}$	0	0	$\frac{\mu_2}{k_3}$	0	0	0	$\dots$
0	$k_2\lambda_2$	0	0		0	0	0	$\frac{\mu_1}{k_3}$	0	0	$\frac{\mu_2}{k_3}$	0	0	$\dots$
0	0	$k_2\lambda_1$	0	0		0	0	0	$\frac{\mu_1}{k_3}$	0	0	$\frac{\mu_2}{k_3}$	0	$\dots$
0	0	$k_2\lambda_2$	0	0	0		0	0	0	$\frac{\mu_1}{k_3}$	0	0	$\frac{\mu_2}{k_3}$	$\dots$
0	0	0	$k_3\lambda_1$	0	$\frac{\mu_2}{k_2}$	0	0	0	0	0	0	0	0	$\dots$
0	0	0	$k_3\lambda_2$	0	0	$\frac{\mu_2}{k_2}$	0	0	0	0	0	0	0	$\dots$
0	0	0	0	$k_3\lambda_1$	0	0	$\frac{\mu_1}{k_3}$	0	0	0	0	0	0	$\dots$
0	0	0	0	$k_3\lambda_2$	0	0	0	$\frac{\mu_1}{k_3}$	0	0	0	0	0	$\dots$
0	0	0	0	0	$k_3\lambda_1$	0	0	0	$\frac{\mu_1}{k_3}$	0	0	0	0	$\dots$
0	0	0	0	0	$k_3\lambda_2$	0	0	0	0	0	0	0	0	$\dots$
0	0	0	0	0	0	$k_3\lambda_1$	0	0	0	0	0	0	0	$\dots$
0	0	0	0	0	0	$k_3\lambda_2$	0	0	0	0	0	0	0	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	

Рис. 2 Матрица $D(A-C)D^{-1}$

Взяв

$$c = \frac{(k-1)\min(\mu_1, \mu_2)}{k+1},$$

получим:

$$\alpha = \frac{(1-k^2)(\lambda_1+\lambda_2) + (k-1)\min(\mu_1,\mu_2)}{k+1}. \tag{5}$$

Очевидно, что если $\min(\mu_1, \mu_2) > \lambda_1 + \lambda_2$, то всегда можно подобрать $k > 1$ при которых α положительно. Отсюда получаем теорему.

Теорема 2. *Если* $\min(\mu_1, \mu_2) > (\lambda_1 + \lambda_2)$*, то процесс сильно эргодичен и верна оценка*

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \le e^{-\alpha t}\|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\|_{1D}. \tag{6}$$

Замечание 1. Теоремы 1 и 2 можно легко обобщить на случай 1-периодических интенсивностей.

Теорема 3. *Если интенсивности 1-периодичны и* $\int_0^1 \max(\mu_1(s), \mu_2(s))\,ds < \int_0^1 (\lambda_1(s) + \lambda_2(s))\,ds$*, то процесс нуль-эргодичен.*

Теорема 4. *Если интенсивности 1-периодичны и* $\int_0^1 \min(\mu_1(s), \mu_2(s))\,ds > \int_0^1 (\lambda_1(s) + \lambda_2(s))\,ds$*, то процесс слабо эргодичен и верна оценка*

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \le e^{-\int_0^t \alpha(u)\,du}\|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\|_{1D}.$$

4 Сведение к одномерному процессу рождения и гибели

Рассмотрим теперь новый процесс $\tilde{X}(t)$, определяющий число требований в системе. В отличие от $X(t)$ в проекции теряется информация о порядке требований в очереди и учитывается только число требований в системе. Для исследования будет применяться подход из статьи [1], где было показано, что при фиксированном начальном условии интенсивности проекции для многомерного ПРГ при больших значениях времени можно рассматривать как соответствующие характеристики для одномерных ПРГ. Там же получены условия нуль-эргодичности, слабой эргодичности и оценки сходимости для полученных одномерных процессов. Теоремы, приведенные в [1], справедливы и в рассматриваемой ситуации для одномерного процесса, получающегося для модели $M_t/M_t/1$ с двумя классами требований.

Из соответствующих уравнений прямой системы Колмогорова получаем

$$\frac{d(p_0)}{dt} = -(\lambda_1 + \lambda_2)\,p_0 + \mu_1 p_1 + \mu_2 p_2,$$

откуда

$$\frac{d(p_0)}{dt} = -(\lambda_1+\lambda_2)\,p_0 + \frac{\mu_1 p_1+\mu_2 p_2}{p_1+p_2}\,(p_1+p_2)\,.$$

Далее аналогично:

$$\frac{d(p_1+p_2)}{dt} = (\lambda_1+\lambda_2)\,p_0 - (\lambda_1+\lambda_2+\mu_1)\,p_1 - (\lambda_1+\lambda_2+\mu_2)\,p_2 + \\ + \mu_1 p_3 + \mu_1 p_4 + \mu_2 p_5 + \mu_2 p_6;$$

$$\frac{d(p_1+p_2)}{dt} = (\lambda_1+\lambda_2)\,p_0 - \frac{(\lambda_1+\lambda_2+\mu_1)p_1 + (\lambda_1+\lambda_2+\mu_2)p_2}{p_1+p_2}(p_1+p_2) + \\ + \frac{\mu_1 p_3+\mu_1 p_4+\mu_2 p_5+\mu_2 p_6}{p_3+\cdots+p_6}\,(p_3+\cdots+p_6)\,.$$

И так далее складываем по 2^s уравнений.

Обозначим

$$\tilde{\mu}_1 = \frac{\mu_1 p_1+\mu_2 p_2}{p_1+p_2}\,;$$
$$\tilde{\mu}_2 = \frac{\mu_1 p_3+\mu_1 p_4+\mu_2 p_5+\mu_2 p_6}{p_3+\cdots+p_6}\,.$$

Аналогично определяем остальные $\tilde{\mu}_k$. Обозначим $\tilde{\lambda}_k = \lambda_1+\lambda_2$ и

$$x_0 = p_0;\;\; x_1 = p_2+p_3;\;\; x_2 = p_4+p_5+p_6+p_7;\;\; \ldots$$

Получим систему

$$\frac{d\mathbf{x}}{dt} = \tilde{A}\mathbf{x}(t),$$

где

$$\tilde{\lambda}_i = \lambda_1+\lambda_2\,;\quad \tilde{\mu}_i = \frac{\mu_1\sum_{k=2^i-1}^{2^i+2^{i-1}-2} p_i + \mu_2\sum_{k=2^i+2^{i-1}-1}^{2^{i+1}-2} p_i}{\sum_{k=2^i-1}^{2^{i+1}-2} p_i}\,;$$

соответствующий процесс обозначим $\tilde{X}(t)$.

Положим

$$l \le \tilde{\lambda}_k \le L\,,\quad m \le \tilde{\mu}_k \le M$$

для любых k, t и любого начального условия.

Получим теоремы 5 и 6, аналогичные которым доказаны в [1].

1. Пусть $\tilde{X}(t)$ имеет счетное число состояний и

$$M < l\,. \tag{7}$$

Пусть $\sigma = \sqrt{M/l} < 1$, $\delta_n = \sigma^n$, $n \geq 0$, $\tilde{x}_n = \delta_n x_n$ и $\tilde{\mathbf{x}} = (\tilde{x}_0, \tilde{x}_1, \dots)$. Пусть $\Lambda = \operatorname{diag}(\delta_0, \delta_1, \dots)$.

Теорема 5. *Пусть верно* (7). *Тогда* $\tilde{X}(t)$ *нуль-эргодичен и верны оценки*

$$\|\tilde{\mathbf{x}}(t)\| \leq e^{-\alpha^* t}\|\tilde{\mathbf{x}}(0)\|\,;$$
$$\Pr\left(\tilde{X}(t) \leq n/\tilde{X}_j(0) = k\right) \leq \sigma^{k-n}e^{-\alpha^* t}.$$

Здесь

$$\Pr\left(\tilde{X}(t) > n/\tilde{X}(0) = k\right) > 1 - \sigma^{k-n}e^{-\alpha^* t}$$

и $\Pr\left(\tilde{X}(t) > n/\tilde{X}(0) = k\right) \to 1$ *при* $t \to \infty$ *для любых* n *и* k.

2. Пусть

$$L < m\,, \quad \alpha_* = l + m - 2\sqrt{LM} > 0\,. \tag{8}$$

Теорема 6. *Пусть верно* (8). *Тогда* $\tilde{X}(t)$ *слабо эргодичен и верна оценка*

$$\|D\mathbf{w}(t)\| \leq e^{-\alpha_* t}\|D\mathbf{w}(0)\| \tag{9}$$

для любых $t \geq 0$ *и любого соответствующего начального условия*.

Замечание 2. Аналогичные утверждения, как и в предыдущем разделе, справедливы и в случае 1-периодических интенсивностей.

Замечание 3. Условия на нуль-эргодичность в теоремах 1 и 5 и слабую эргодичность в 2 и 6, а также оценки (6) и (9) формально выглядят одинаково. Однако в них использованы разные нормы. В (1), (2) и (6) норма раскрывается привычным способом:

$$\|\mathbf{p}\| = \sum_{i=0}^{\infty} |p_i|\,,$$

в то время как в (5), (6) и (9) оценивается норма

$$\|\mathbf{x}\| = \sum_{i=0}^{\infty} |x_i| = \sum_{i=0}^{\infty} \left| \sum_{2^i-1}^{2^{i+1}-2} p_i \right|$$

(сумма модулей суммы вероятностей состояний по столбцам в графе).

Замечание 4. Рассмотренный в этом разделе процесс можно описать как сумму координат $Z_1 + Z_2$ соответствующего двумерного ПРГ.

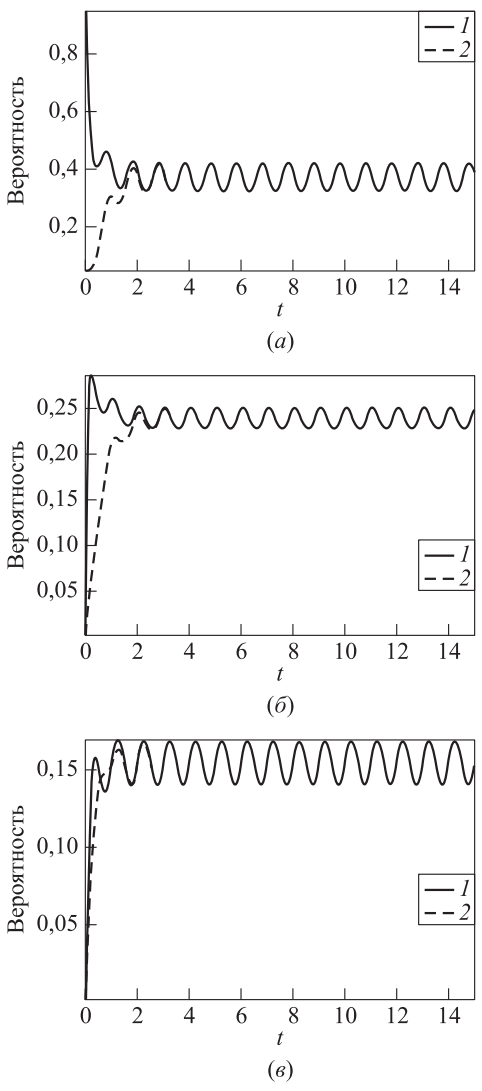


Рис. 3 Вероятности отсутствия требований (*а*), наличия ровно одного требования первого или второго класса (*б*) и наличия ровно двух каких-либо требований в системе (*в*): *1* — $p_0 = 0{,}95$; *2* — $p_0 = 0{,}05$

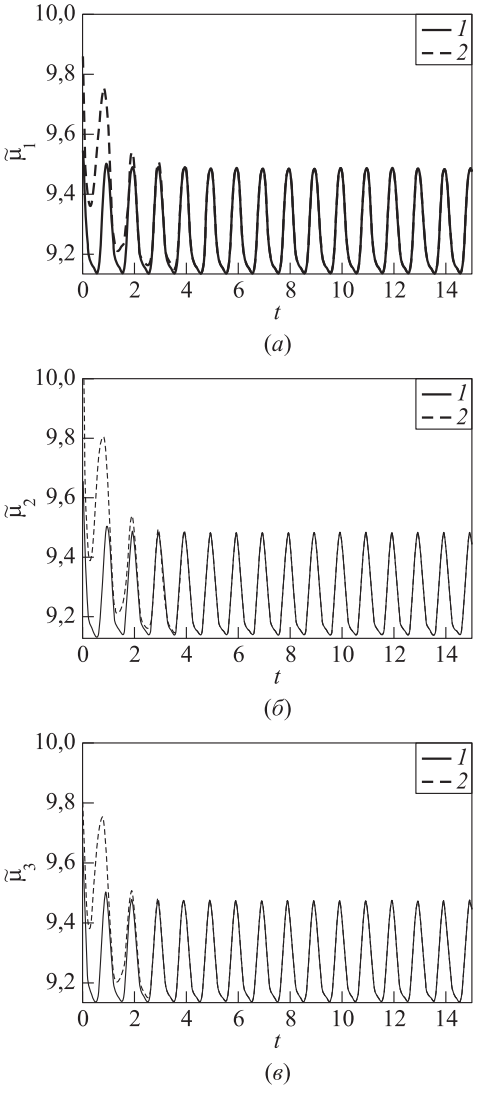


Рис. 4 Интенсивности $\tilde{\mu}_1$ (*а*), $\tilde{\mu}_2$ (*б*) и $\tilde{\mu}_3$ (*в*): *1* — $p_0 = 0{,}95$; *2* — $p_0 = 0{,}05$

5 Численный пример

Пусть интенсивности имеют вид:

$$\lambda_1(t) = 1 + \cos(2\pi t);\ \lambda_2(t) = 5 + \sin(2\pi t);\ \mu_1(t) = 11 + \sin(2\pi t);\ \mu_2 = 9\,.$$

Применить теорему 2 для получения оценок скорости сходимости затруднительно, так как нужно подбирать k, очень близкое к единице. Это связано с тем, что $\lambda_1 + \lambda_2$ не сильно отличается от $\min(\mu_1, \mu_2)$, при этом по формуле (5) получается $\alpha(t)$, почти равная нулю.

Применим теорему 6 для получения оценок скорости сходимости «проекции». Подбираем l_1, L_1, m_1 и M_1, при которых верно $l_1 \le \tilde{\lambda} \le L_1$ и $m_1 \le \tilde{\mu} \le M_1$. Положив $\beta = \sqrt{M_1/L_1} = \sqrt{6}$, получаем

$$\alpha_* = l_1 + m_1 - 2\sqrt{L_1 M_1} \ge 10 - 4\sqrt{6}\,.$$

Получаем оценку

$$\|D\mathbf{w}(t)\| \le e^{-\alpha_* t}\|D\mathbf{w}(0)\|$$

для любых $t \ge 0$ и любого соответствующего начального условия. Здесь D — треугольная матрица с $d_{k+1} = \beta^k$, $k \ge 0$.

На рис. 3 и 4 приведены графики для усеченной системы с состояниями от 0 до $2^7 - 2$. Графики строились для двух разных начальных условий. В одном случае $p_0 = 0{,}95$, остальные вероятности равны между собой. В другом случае $p_0 = 0{,}05$, остальные вероятности также равны между собой. «Проекция» описывает число требований в системе, и получается ПРГ с периодическими интенсивностями и семью состояниями. В частности, $\tilde{\mu}_1$, $\tilde{\mu}_2$ и $\tilde{\mu}_3$ приведены соответственно на рис. 4. Можно заметить, что изначально интенсивности сильно отличаются от периодических, но затем практически совпадают друг с другом и мало отличаются от периодических, причем они заключены между μ_1 и μ_2.

Литература

1. *Zeifman A., Satin Y., Kiseleva K., Korolev V.* On the rate of convergence for a characteristic of multidimensional birth–death process // Mathematics, 2019. Vol. 7. Iss. 5. Art. No. 477. 10 p. doi: 10.3390/math7050477.
2. *Zeifman A. I.* On the estimation of probabilities for birth and death processes // J. Appl. Probab., 1995. Vol. 32. No. 3. P. 623–634.
3. *Granovsky B., Zeifman A.* Nonstationary queues: Estimation of the rate of convergence // Queueing Syst., 2004. Vol. 46. P. 363–388.
4. *Adan I., Hathaway B., Kulkarni V. G.* On first-come, first-served queues with two classes of impatient customers // Queueing Syst., 2019. Vol. 91. P. 113–142. doi: 10.1007/s11134-018-9592-z.

Поступила в редакцию 22.01.21

ON THE BOUNDS OF THE RATE OF CONVERGENCE FOR $M_t/M_t/1$ MODEL WITH TWO DIFFERENT TYPES OF REQUESTS

Ya. A. Satin

Department of Applied Mathematics, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation

Abstract: The author deals with a nonstationary queuing model $M_t/M_t/1$ with one server and two different types of requests. For this model, the author obtains a one-dimensional birth and death process that describes the number of requirements in the original system. By applying the standard method of the logarithmic norm of the operator of a linear function, corresponding estimates for the rate of convergence and ergodicity are obtained. A numerical example with exact given values of intensities showing the application of the studied approach is constructed and corresponding graphic illustrations are provided. The author uses the general algorithm to build graphs, it is associated with solving the Cauchy problem for the forward Kolmogorov system on the corresponding interval which has already been used by the authors in previous papers.

Keywords: queuing systems; nonstationary queuing model; one-dimensional birth–death process; rate of convergence; ergodicity bounds; logarithmic norm; $M_t/M_t/1$ queue

DOI: 10.14357/08696527210102

Acknowledgments

This work was financially supported by the Russian Science Foundation (grant No. 19-11-00020).

References

1. Zeifman, A., Y. Satin, K. Kiseleva, and V. Korolev. 2019. On the rate of convergence for a characteristic of multidimensional birth-death process. *Mathematics* 7(5):477. 10 p. doi: 10.3390/math7050477.
2. Zeifman, A. I. 1995. On the estimation of probabilities for birth and death processes. *J. Appl. Probab.* 32(3):623–634.
3. Granovsky, B., and A. Zeifman. 2004. Nonstationary queues: Estimation of the rate of convergence. *Queueing Syst.* 46:363–388.
4. Adan, I., B. Hathaway, and V. G. Kulkarni. 2019. On first-come, first-served queues with two classes of impatient customers. *Queueing Syst.* 91:113–142. doi: 10.1007/s11134-018-9592-z.

Received January 22, 2021

Contributor

Satin Yacov A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation; yacovi@mail.ru

ОБ АППРОКСИМАЦИИ С ПОМОЩЬЮ УСЕЧЕНИЙ ДЛЯ ОДНОЙ НЕСТАЦИОНАРНОЙ МОДЕЛИ МАССОВОГО ОБСЛУЖИВАНИЯ*

Я. А. Сатин[1]

Аннотация: Исследуется нестационарная модель массового обслуживания типа $M_t/M_t/1$ с одним сервером. Предполагается, что требования поступают с интенсивностью $\lambda(t)$, а обслуживаются парами, т. е. в данном случае $\mu(t)$ — это интенсивность обслуживания группы из двух требований. Для рассматриваемой модели построены предельные характеристики с помощью метода усечений пространства состояний системы. Приведены численный пример, а также графические иллюстрации, построенные на основе общего алгоритма, разработанного в предыдущих работах автора и связанного с решением задачи Коши для прямой системы Колмогорова на соответствующем временном интервале.

Ключевые слова: системы массового обслуживания; модель типа $M_t/M_t/1$; нестационарная модель массового обслуживания; аппроксимация; предельные характеристики; усечение пространства состояний

DOI: 10.14357/08696527210103

1 Введение

В работе изучается модель типа $M_t/M_t/1$ с особенностями в обслуживании, а именно предполагается, что требования поступают на обслуживание по одному с интенсивностью $\lambda(t)$, а обслуживаются *только парами*, т. е. с интенсивностью $\mu(t)$ обслуживается группа из двух требований.

В статье [1] получены оценки скорости сходимости к предельному режиму. В настоящей работе исследуется вопрос о построении аппроксимаций предельных характеристик с помощью усечений пространства состояний системы.

Для усечения будет применяться подход, разработанный в [2]. Отметим, что имеется много работ, в которых строятся те же (по существу, предельные) характеристики, в которых применяются другие методы аппроксимации, не использующие оценки скорости сходимости и из-за этого приводящие к существенно большему объему вычислений (см., например, [3, 4]).

На рис. 1 приведен граф процесса, связанный с числом требований в системе обслуживания. Соответствующая прямая система Колмогорова имеет вид:

$$\frac{d}{dt}\mathbf{p}(t) = A(t)\mathbf{p}(t), \tag{1}$$

*Исследование выполнено за счет гранта Российского научного фонда (проект 19-11-00020.)

[1]Вологодский государственный университет, yacovi@mail.ru

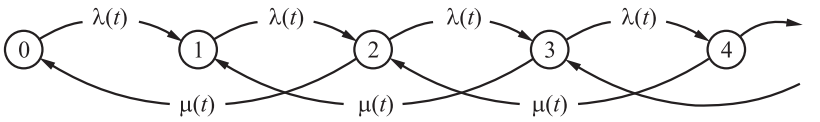


Рис. 1 Граф процесса $X(t)$

где

$$A(t) = \begin{pmatrix} -\lambda(t) & 0 & \mu(t) & 0 & 0 & \cdots \\ \lambda(t) & -\lambda(t) & 0 & \mu(t) & 0 & \cdots \\ 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \mu(t) & \cdots \\ 0 & 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Заменяя $p_0(t) = 1 - \sum_{i\geq 1} p_i(t)$, получаем из (1) систему

$$\frac{d}{dt}\,\mathbf{z}(t) = B(t)\mathbf{z}(t) + \mathbf{f}(t),$$

где

$$\mathbf{f}(t) = (\lambda(t), 0, 0, \dots)^{\mathrm{T}}; \qquad \mathbf{z}(t) = (p_1(t), p_2(t), \dots)^{\mathrm{T}};$$

$$B(t) = \begin{pmatrix} -2\lambda(t) & -\lambda(t) & \mu(t)-\lambda(t) & -\lambda(t) & \cdots \\ \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \mu(t) & \cdots \\ 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \end{pmatrix}.$$

Скорость сходимости процесса $X(t)$ может быть найдена из системы

$$\frac{d}{dt}\,\mathbf{y}(t) = B(t)\mathbf{y}(t)\,.$$

Пусть T — треугольная матрица:

$$t_{ij} = \begin{cases} 1 & \text{при } j \geq i\,; \\ 0 & \text{в противном случае.} \end{cases}$$

Положим $\mathbf{u}(t) = T\mathbf{y}(t)$. Тогда

$$\frac{d}{dt}\,\mathbf{u}(t) = B^*(t)\mathbf{u}(t),$$

где

$$B^*(t) =$$

$$= \begin{pmatrix} -\lambda(t) & -\mu(t) & \mu(t) & 0 & 0 & 0 & \cdots \\ \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \mu(t) & 0 & 0 & \cdots \\ 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \mu(t) & 0 & \cdots \\ 0 & 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \mu(t) & \cdots \\ 0 & 0 & 0 & \lambda(t) & -(\lambda(t)+\mu(t)) & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Возьмем $\{d_i, i \geq 0\}$ так, что $\inf_{i\geq 0}|d_i| = d > 0$. Пусть $D = \operatorname{diag}(d_0, d_1, d_2, \dots)$. Делаем подстановку $\mathbf{w}(t) = D\mathbf{u}(t)$. Получаем

$$\frac{d}{dt}\,\mathbf{w}(t) = B^{**}(t)\mathbf{w}(t),$$

где $B^{**}(t) = (b^{**}(t))_{i,j=1}^{\infty} = DB^*(t)D^{-1}$ имеет вид:

$$B^{**}(t) = \begin{pmatrix} -\lambda(t) & -\mu(t)\frac{d_1}{d_2} & \mu(t)\frac{d_1}{d_3} & 0 & 0 & \cdots \\ \lambda(t)\frac{d_2}{d_1} & -(\lambda(t)+\mu(t)) & 0 & \mu(t)\frac{d_2}{d_4} & 0 & \cdots \\ 0 & \lambda(t)\frac{d_3}{d_2} & -(\lambda(t)+\mu(t)) & 0 & \mu(t)\frac{d_3}{d_5} & \cdots \\ 0 & 0 & \lambda(t)\frac{d_4}{d_3} & -(\lambda(t)+\mu(t)) & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

В работе [1] показано, что сходимость оценивается следующим образом. Пусть

$$\alpha^*(t) \geq \min\left(\lambda(t)\left(1-\delta^{-1}\right), \mu(t)(1+\delta) - \lambda(t)\left(\delta^2-1\right), \right.$$
$$\left. \mu(t)\left(1-\delta^{-1}\right) - \lambda(t)(\delta-1)\right). \quad (2)$$

Тогда верно неравенство:

$$\|\mathbf{w}(t)\| \le e^{-\int_0^t \alpha^*(\tau)\,d\tau}\|\mathbf{w}(0)\|, \tag{3}$$

Пусть числа δ_i, где $i = 1, 2$, больше единицы. Обозначим

$$D_{M_i} = \operatorname{diag}\left(1, \delta_i, \delta_i^2, \delta_i^3, \ldots\right); \quad D_{m_i} = \operatorname{diag}\left(1, \frac{1}{\delta_i}, \delta_i, \delta_i^2, \delta_i^3, \ldots\right).$$

Тогда (3) можно переписать в виде:

$$\|\mathbf{w}(t)\|_{D_m} \le e^{-\int_0^t \alpha^*(\tau)\,d\tau}\|\mathbf{w}(0)\|_{D_M}, \tag{4}$$

или в виде:

$$\|D_m T\mathbf{y}(t)\| \le e^{-\int_0^t \alpha^*(\tau)\,d\tau}\|D_M T\mathbf{y}(0)\|.$$

Замечание 1. Можно показать, что для 1-периодической $\alpha^*(t)$ можно подобрать такие M и a, что будет верно неравенство $e^{-\int_0^t \alpha^*(\tau)\,d\tau} \le Me^{-at}$. Например, можно взять $M = \exp\left(\sup_{|t-s|\le 1}\int_s^t \alpha(s)\,ds\right)$, $a = \int_0^1 \alpha(s)\,ds$. Если найти целое число A, такое что $\alpha(t) > A$, то можно положить $a = A$, $M = 1$.

Замечание 2. Если интенсивности постоянны, то $\alpha^*(t)$ постоянна, и, взяв $M = 1$ $a = \alpha(s)$, получаем неравенство $e^{-\int_0^t \alpha^*(\tau)\,d\tau} \le Me^{-at}$.

2 Аппроксимация усечениями

Зададим усеченный процесс, заменив на нули столбцы с $N+1$ и строки с $N+1$ в матрице $A(t)$. Заменим еще $a_{N,N}$ на $-\mu(t)$.

Далее будем считать, что в начальный момент времени $t = 0$ исходный процесс и его усеченный аналог находятся в нулевом состоянии, т. е. в системе нет требований.

Будем считать далее, что при всяком t верно $\lambda(t) \le L$.

Замечание 3. Столбцы в матрице $A(t)$ нумеруются с нуля.

Тогда усеченный процесс описывается системой

$$\frac{d}{dt}\,\widetilde{\mathbf{p}}(t) = A_N(t)\widetilde{\mathbf{p}}(t)$$

и соответствующей преобразованной системой

$$\frac{d}{dt}\,\widetilde{\mathbf{z}}(t) = B_N(t)\widetilde{\mathbf{z}}(t) + \widetilde{\mathbf{f}}(t), \tag{5}$$

где

$$\widetilde{\mathbf{f}}(t) = (\lambda(t), 0, 0, \dots)^{\mathrm{T}}; \quad \widetilde{\mathbf{z}}(t) = (\widetilde{z}_1(t), \widetilde{z}_2(t), \dots)^{\mathrm{T}} = (\widetilde{p}_1(t), \widetilde{p}_2(t), \dots)^{\mathrm{T}}.$$

Далее возьмем два разных числа $1 < \delta_2 < \delta_1$, для которых $\int_0^1 \alpha_i^*(s)\,ds \geq 0$ (см. (2)). Затем подберем M_i и a_i, для которых верно $e^{-\int_0^t \alpha_i^*(\tau)\,d\tau} \leq M_i e^{-a_i t}$ (см. замечание 1).

Для начала оценим p_N.

Из (5) получаем

$$\widetilde{z}_N(t) \sum_{k=0}^{N-1} \delta_1^k \leq$$

$$\leq \|T\widetilde{\mathbf{z}}(t)\|_{D_{m_1}} \leq e^{-\int_0^t \alpha_1^*(\tau)\,d\tau} \|T\widetilde{\mathbf{z}}(0)\|_{D_{M_1}} + \int_0^t e^{-\int_s^t \alpha_1^*(\tau)} \|T\mathbf{f}(t)\|_{D_{M_1}}.$$

Так как усеченный аналог исходного процесса в начальный момент времени находится в нулевом состоянии, положим $\widetilde{\mathbf{z}}(0) = \mathbf{0}$. Получаем

$$\widetilde{z}_N(t) \leq \frac{LM_1}{a_1 \sum_{k=0}^{N-1} \delta_1^k}.$$

Получим теперь оценку усечения.

Найдем разность

$$\frac{d}{dt}(\mathbf{z} - \widetilde{\mathbf{z}}(t)) = B(t)\mathbf{z}(t) - B_N(t)\widetilde{\mathbf{z}}(t).$$

Перепишем ее в виде:

$$\frac{d}{dt}\left(\mathbf{z}(t) - \widetilde{\mathbf{z}}(t)\right) = B(t)\left(\mathbf{z}(t) - \widetilde{\mathbf{z}}(t)\right) + \left(B(t) - B_N(t)\right)\widetilde{\mathbf{z}}(t). \tag{6}$$

Так как усеченный аналог исходного процесса и сам исходный процесс в начальный момент времени находятся в нулевом состоянии, то $\widetilde{\mathbf{z}}(0) = \mathbf{z}(0) = \mathbf{0}$. С учетом (4) из (6) получаем:

$$\|T(\mathbf{z}(t) - \widetilde{\mathbf{z}}(t))\|_{D_{m_2}} \leq e^{-\int_0^t \alpha_2^*(\tau)\,d\tau} \left\|T\left(\mathbf{z}(0) - \widetilde{\mathbf{z}}(0)\right)\right\|_{D_{M_2}} +$$

$$+ \int_0^t e^{-\int_s^t \alpha_2^*(\tau)} \left\|T\left(B(t) - B_N(t)\right)\widetilde{\mathbf{z}}(t)\right\|_{D_{M_2}} ds.$$

Далее

$$T(B(t)-B_N(t))\widetilde{\mathbf{z}}(t)=(0,\dots,0,-\lambda(t)\widetilde{z}_N(t),\lambda(t)\widetilde{z}_N(t),0,0,0,\dots)^{\mathrm{T}}.$$

Значит,

$$\|T(B(t)-B_N(t))\widetilde{\mathbf{z}}_N(t)\|_{D_{M_2}}\le 2L\frac{LM_1\delta_2^N}{a_1\sum_{k=0}^{N-1}\delta_1^k}.$$

Отсюда

$$\|T(\mathbf{z}(t)-\widetilde{\mathbf{z}}(t))\|_{D_{m_2}}\le\frac{2L^2M_1M_2\delta^N}{a_2a_1\sum_{k=0}^{N-1}\delta_1^k}.$$

Перепишем оценку в других нормах.
С учетом

$$\|T\mathbf{u}(t)\|_{D_{m_2}}=|u_1|+\left(1+\frac{1}{\delta_2}\right)|u_2|+\left(1+\frac{1}{\delta_2}+\delta_2\right)|u_3|+$$
$$+\left(1+\frac{1}{\delta_2}+\delta_2+\delta_2^2\right)|u_4|+\left(1+\frac{1}{\delta_2}+\delta_2+\delta_2^2+\delta_2^3\right)|u_5|+\dots\ge\|\mathbf{u}(t)\|$$

получаем

$$\|\mathbf{p}(t)-\widetilde{\mathbf{p}}(t)\|\le 2\|\mathbf{z}(t)-\widetilde{\mathbf{z}}(t)\|\le 2\|T(\mathbf{z}(t)-\widetilde{\mathbf{z}}(t))\|_{D_{m_2}}\le 4\frac{L^2M_1M_2\delta_2^N}{a_1a_2\sum_{k=0}^{N-1}\delta_1^k}.$$

С учетом

$$\|T\mathbf{u}(t)\|_{D_{m_2}}=|u_1|+\frac{1}{2}\left(1+\frac{1}{\delta_2}\right)2|u_2|+\frac{1}{3}\left(1+\frac{1}{\delta_2}+\delta_2\right)3|u_3|+$$
$$+\frac{1}{4}\left(1+\frac{1}{\delta_2}+\delta_2+\delta_2^2\right)4|u_4|+\frac{1}{5}\left(1+\frac{1}{\delta_2}+\delta_2+\delta_2^2+\delta_2^3\right)5|u_5|+\dots\ge$$
$$\ge W\|\mathbf{u}(t)\|_{1E},$$

где $W=\min_{k\ge1}(\delta_2^{k-2}/k)$, получаем

$$|\phi(t)-\tilde{\phi}(t)|=\|\mathbf{z}(t)-\widetilde{\mathbf{z}}(t)\|_{1E}\le\frac{2}{W}\,\frac{L^2M_1M_2\delta_2^N}{a_1a_2\sum_{k=0}^{N-1}\delta_1^k},$$

где $\phi(t)=\sum_{i=0}^{\infty}ip_i$; $\tilde{\phi}(t)=\sum_{i=0}^{N}i\widetilde{p}_i$.

Теорема 1. *Пусть найдутся числа* $1 < \delta_2 < \delta_1$, *такие что при* $i = 1, 2$

$$\int_0^1 \alpha_i^*(t)\,dt \geq \int_0^1 \min\left(\lambda(t)\left(1-\delta_i^{-1}\right), \mu(t)\left(1+\delta_i\right) - \right.$$
$$\left. - \lambda(t)\left(\delta_i^2-1\right), \mu(t)\left(1-\delta_i^{-1}\right) - \lambda(t)\left(\delta_i-1\right)\right)\,dt > 0\,.$$

Тогда верны оценки

$$\|\mathbf{p}(t) - \widetilde{\mathbf{p}}(t)\| \leq 4\frac{L^2 M_1 M_2 \delta_2^N}{a_1 a_2 \sum_{k=0}^{N-1} \delta_1^k}\,;$$

$$|\phi(t) - \tilde{\phi}(t)| \leq \frac{2}{W}\,\frac{L^2 M_1 M_2 \delta_2^N}{a_1 a_2 \sum_{k=0}^{N-1} \delta_1^k}\,,$$

где

$$W = \min_{k\geq 1} \frac{\delta_2^{k-2}}{k}; \quad e^{-\int_0^t \alpha_i^*(\tau)\,d\tau} \leq M_i e^{-a_i t}; \quad \phi(t) = \sum_{i=0}^{\infty} i p_i; \quad \tilde{\phi}(t) = \sum_{i=0}^{N} i\widetilde{p}_i.$$

Замечание 4. В статье [1] показано, что в качестве одной из δ_i можно выбрать число $\sqrt{\mu}/\sqrt{\lambda}$. В этом случае получается

$$\alpha_i^* = \min\left((\sqrt{\mu}-\sqrt{\lambda})^2, \lambda\left(1-\sqrt{\frac{\mu}{\lambda}}\right)\right),$$

и можно взять $a_i = \alpha_i^*$, $M_i = 1$.

3 Численный пример

Пусть $\lambda(t) = 2 + \sin(2\pi t)$, $\mu(t) = 4 - \cos(2\pi t)$.

Положим $\delta_1 = 11/10$. Тогда $\alpha_1^* \geq 1/22$, и можно взять $a_1 = 1/22$, $M_1 = 1$.

Положим $\delta_2 = 101/100$. Тогда $\alpha_2^* \geq 0,005$, и можно взять $a_2 = 0{,}005$, $M_2 = 1$, $W > 0{,}02$.

С учетом оценки скорости сходимости отсюда получаем, что при $N = 300$ справедливы неравенства:

$$\|\mathbf{p}(t) - \widetilde{\mathbf{p}}(t)\| \leq 10^{-5}\,;$$

$$|\phi(t) - \tilde{\phi}(t)| \leq 10^{-4}.$$

Соответствующие графики приведены на рис. 2 и 3.

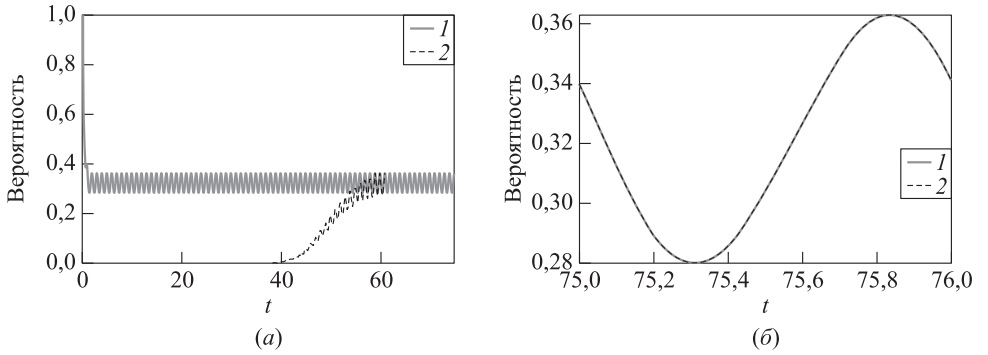


Рис. 2 Вероятность отсутствия требований в системе $p_0(t)$ на отрезке $[0, 75]$ (*а*) и «в предельном режиме» на отрезке $[75, 76]$ (*б*) при начальных условиях $X(0) = 0$ (*1*) и 300 (*2*)

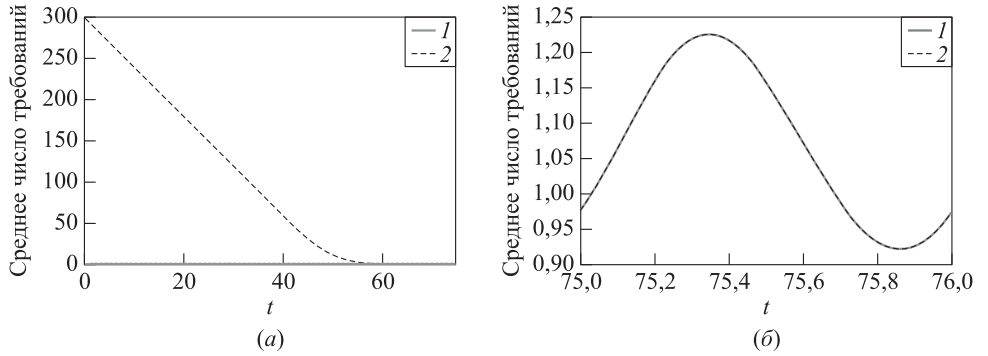


Рис. 3 Среднее число требований в системе $E(t, 0)$ (*1*) и $E(t, 300)$ (*2*) на отрезке $[0, 75]$ (*а*) и «в предельном режиме» на отрезке $[75, 76]$ (*б*)

Литература

1. *Satin Y., Zeifman A., Kryukova A.* On the rate of convergence and limiting characteristics for a nonstationary queueing model // Mathematics, 2019. Vol. 7. Iss. 8. Art. No. 678. 11 p.
2. *Зейфман А. И., Коротышева А. В., Королев В. Ю., Сатин Я. А.* Оценки погрешности аппроксимаций неоднородных марковских цепей с непрерывным временем // Теория вероятностей и ее применения, 2016. Т. 61. № 3. С. 563–569.
3. *Arns M., Buchholz P., Panchenko A.* On the numerical analysis of inhomogeneous continuous-time Markov chains // Informs J. Comput., 2010. Vol. 22. P. 416–432.
4. *Andreychenko A., Sandmann W., Wolf V.* Approximate adaptive uniformization of continuous-time Markov chains // Appl. Math. Model., 2018. Vol. 61. P. 561–576.

Поступила в редакцию 22.01.21

ON APPROXIMATION WITH TRUNCATIONS FOR THE NONSTATIONARY QUEUING MODEL

Ya. A. Satin

Department of Applied Mathematics, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation

Abstract: The author deals with a nonstationary queuing model $M_t/M_t/1$ with one server. It is assumed here that the customers arrive with the intensity $\lambda(t)$ but are served in pairs (that is, in this case, $\mu(t)$ is the service rate of a group of two customers). For the considered model, the limiting characteristics are constructed using the method of truncating the state space of the system. A numerical example with exact given values of intensities showing the application of the studied approach is constructed and corresponding graphic illustrations are provided. The author uses the general algorithm to build graphs, it is associated with solving the Cauchy problem for the forward Kolmogorov system on the corresponding interval which has already been used by the author in previous papers.

Keywords: queuing systems; $M_t/M_t/1$ queue; nonstationary queuing model; approximation; limiting characteristics; rate of convergence; truncation of the state space

DOI: 10.14357/08696527210103

Acknowledgments

This work was financially supported by the Russian Science Foundation (grant No. 19-11-00020).

References

1. Satin, Y., A. Zeifman, and A. Kryukova. 2019. On the rate of convergence and limiting characteristics for a nonstationary queueing model. *Mathematics* 7(8):678. 11 p.
2. Zeifman, A. I., A. V. Korotysheva, V. Yu. Korolev, and Ya. A. Satin. 2016. Truncation bounds for approximations of inhomogeneous continuous-time Markov chains. *Theor. Probab. Appl.* 61(3):513–520.
3. Arns, M., P. Buchholz, and A. Panchenko. 2010. On the numerical analysis of inhomogeneous continuous-time Markov chains. *Informs J. Comput.* 22:416–432.
4. Andreychenko, A., W. Sandmann, and V. Wolf. 2018. Approximate adaptive uniformization of continuous-time Markov chains. *Appl. Math. Model.* 61:561–576.

Received January 22, 2021

Contributor

Satin Yacov A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation; yacovi@mail.ru

АНАЛИТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И ФИЛЬТРАЦИЯ ПРОЦЕССОВ В ИНТЕГРОДИФФЕРЕНЦИАЛЬНЫХ СТОХАСТИЧЕСКИХ СИСТЕМАХ, НЕ РАЗРЕШЕННЫХ ОТНОСИТЕЛЬНО ПРОИЗВОДНЫХ

И. Н. Синицын[1]

Аннотация: Для нелинейных интегродифференциальных стохастических систем (ИДСтС), не разрешенных относительно производных и приводимых к дифференциальным методом сингулярных ядер, разработаны алгоритмы аналитического моделирования нормальных стохастических процессов (СтП), при этом нелинейность под интегралом может быть разрывной, а также синтеза нормальных субоптимальных фильтров (НСОФ) для онлайн-обработки информации в ИДСтС. Подробно рассмотрен тестовый пример с разрывной нелинейностью под знаком интеграла. Предложены алгоритмы оценки качества НСОФ на основе теории чувствительности.

Ключевые слова: метод аналитического моделирования (МАМ); метод нормальной аппроксимации (МНА); метод статистической линеаризации (МСЛ); нормальный субоптимальный фильтр (НСОФ); стохастическая система (СтС); стохастические системы, не разрешенные относительно производных; формирующий фильтр (ФФ)

DOI: 10.14357/08696527210104

1 Введение

Интегродифференциальные стохастические системы, разрешенные относительно производных, служат подходящими моделями для эредитарных стохастических систем (ЭСтС). В случаях, когда эредитарные ядра асимптотически устойчивы, кроме того вырождены (или удовлетворяют обыкновенным дифференциальным уравнениям), уравнения ИДСтС приводятся к дифференциальным стохастическим системам (СтС) [1, 2]. Для таких ЭСтС вопросы аналитического и статистического моделирования распределений рассмотрены в [3–10]. Работы [11, 12] посвящены синтезу суб- и условно-оптимальных фильтров (СОФ и УОФ) для ЭСтС. Для ИДСтС, не разрешенных относительно производных, вопросы аналитического моделирования рассмотрены в [13, 14]. Теория НСОФ для дифференциальных СтС разработана в [15].

Рассмотрим развитие [13–15] на случай субоптимальных (по среднеквадратичному критерию) нормальных (гауссовских) процессов и фильтров.

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sinitsin@dol.ru

Раздел 2 посвящен методам аналитического моделирования (МАМ) нормальных СтП в негауссовских ИДСтС, не разрешенных относительно производных и приводимых к дифференциальным. Приведены алгоритмы синтеза НСОФ в дифференциальных СтС. Вопросам синтеза НСОФ для гауссовских и негауссовских ИДСтС, не разрешенных относительно производных, посвящен разд. 3. В разд. 4 приведен тестовый пример. Заключение содержит выводы и возможные обобщения.

2 Нормальный субоптимальный фильтр для дифференциальных стохастических систем, не разрешенных относительно производных

Следуя [13, 14], рассмотрим дифференциальную СтС с нелинейностями, описываемыми гладкими функциями

$$\varphi = \varphi\left(t, \Theta, X_t, \dot{X}_t, \ldots, X_t^{(k)}, U_t\right) = 0,$$
$$X(t_0) = X_0, \;\; \dot{X}(t_0) = \dot{X}_0, \ldots, X^{(k)}(t_0) = X_0^{(k)}. \quad (1)$$

Уравнение нелинейного формирующего фильтра (ФФ) возьмем в виде, разрешенном относительно возмущений U_t:

$$\dot{U}_t = a^U(t, \Theta, U_t) + b^U(t, \Theta, U_t)\, V_t^U, \;\; U(t_0) = U_0. \quad (2)$$

Здесь $a^U = a^U(t, \Theta, U_t)$ и $b^U(t, \Theta, U_t)$ — $(n^U \times 1)$- и $(n^U \times n^V)$-мерные функции; V_t^U — белый шум в строгом смысле [1, 2], допускающий представление в виде суммы гауссовской и пуассоновской составляющих

$$V_t^U = \dot{W}_t^U, \;\; W_t^U = W_0^U(t, \Theta) + \int\limits_{R_0^q} c^U(\Theta, \rho) P^0(t, \Theta, d\rho),$$

где ν_t — его интенсивность:

$$\nu_t = \nu_t^W = \nu_t^{W_0} + \int\limits_{R_0^q} c^U(\Theta, \rho) \left[c^U(\Theta, \rho)\right]^{\mathrm{T}} \nu_P(t, \Theta, \rho) d\rho;$$

$c^U = c^U(\Theta, \rho)$ — известная векторная функция той же размерности, что и W_t^0, а интеграл при любом $t \geq t_0$ представляет собой стохастический интеграл по центрированной пуассоновской мере $P^0(t, \Theta, \mathcal{A})$, независимой от W_0^U и имеющей независимые значения на попарно непересекающихся множествах; $\mathcal{A}$ — борелевское множество пространства R_0^q с выколотым началом; ν_t^W, $\nu_t^{W_0}$ и ν_P —

интенсивности СтП W_t^U, W_0^U и P^0. Уравнение (2) понимается в смысле Ито и имеет единственное среднеквадратичное решение [1, 2].

В случае гладких функций в (1), допускающих стохастические производные Ито до h-го порядка и статистическую линеаризацию по Казакову [1–4], выполним следующие преобразования. Будем дифференцировать сполна по t левые части уравнений (1) по обобщенной формуле Ито [1, 2] до тех пор, пока не появятся производные белого шума. В результате получим следующие системы нелинейных дифференциальных уравнений:

$$\varphi = 0, \;\; \dot{\varphi} = 0, \; \ldots \, , \; \varphi^{(h)} = 0 \, , \tag{3}$$

где

$$\begin{aligned} \varphi^{(i)} \left(t, \Theta, X_t^{(i)}, U_t \right) = \varphi_{0X_{1t}}^{(i)} \left(t, \Theta, m_t^{X^i}, K_t^{X^i}, m_t^U, K_t^U \right) + \\ + k_{1X_{1t}}^{(i)} \left(t, \Theta, m_t^{X^i}, K_t^{X^i}, m_t^U, K_t^U, K_t^{X^iU} \right) X_t^{i0} + \\ + k_{1U}^{(i)} \left(t, \Theta, m_t^{X^i}, K_t^{X^i}, m_t^U, K_t^U, K_t^{X^iU} \right) U_t^0 = 0 \quad (i = 1, 2, \ldots, h). \end{aligned} \tag{4}$$

Здесь $X_{1t} = \left[X_t^{\mathrm{T}} \dot{X}_t^{\mathrm{T}} \cdots X_t^{(k-1)\mathrm{T}} \right]^{\mathrm{T}}$; $m_t^{X_1}$ и $K_t^{X_1}$ — вектор математического ожидания и ковариационная матрица; $k_{1X_1}^{(i)} = k_{1X_1}^{(i)}(t, \Theta, m_t^{X^i}, K_t^{X^i}, m_t^U, K_t^U, K_t^{X^iU})$ и $k_{1U}^{(i)} = k_{1U}^{(i)}(t, \Theta, m_t^{X^i}, K_t^{X^i}, m_t^U, K_t^U)$ — матричные коэффициенты статистической линеаризации функций (4).

Далее введем вектор $\bar{X}_t = \left[X_{1t}^{\mathrm{T}} X_{2t}^{\mathrm{T}} \right]^{\mathrm{T}}$, составленный из $X_{1t} = \left[X_t^{\mathrm{T}} \dot{X}_t^{\mathrm{T}} \cdots X_t^{(h-1)\mathrm{T}} \right]^{\mathrm{T}}$ и вспомогательного вектора $X_{2t} = [X_t^{(i)}]_{i=\overline{1,h}}$ на основе (4). В результате придем к уравнениям, разрешенным относительно дифференциалов, следующего вида:

$$d\bar{X}_t = a^{\bar{X}} dt + b^{\bar{X}} dW_0 + \int\limits_{R_0^q} c^{\bar{X}} P^0(t, \Theta, du), \tag{5}$$

где

$$a^{\bar{X}} = a^{\bar{X}} \left(t, \Theta, \bar{X}_t \right); \;\; b^{\bar{X}} = b^{\bar{X}} \left(t, \Theta, \bar{X}_t \right); \;\; c^{\bar{X}} = c^{\bar{X}} \left(t, \Theta, \bar{X}_t, u \right).$$

Таким образом, имеем следующее утверждение [13, 14].

Теорема 1. *Пусть нелинейная негауссовская СтС* (1), (2), *не разрешенная относительно производных k-го порядка, удовлетворяет условиям*:

1^0 *функции* (1) *допускают обобщенные стохастические дифференциалы Ито вплоть до* h*-го порядка включительно и статистическую линеаризацию по Казакову*;

2^0 *возмущения* U_t *негауссовские, причем уравнение ФФ* (2) *разрешено относительно возмущений* U_t *и имеет единственное среднеквадратичное решение*.

3^0 *СтП* X_{1t} *более гладкий, чем возмущение* U_t.

Тогда система (1), (2) *приводима к дифференциальной системе, разрешенной относительно производных* (5).

Теперь перейдем к фильтрации процессов в дифференциальных СтС. Предположим, что СтС удовлетворяет условиям теоремы 1, а также полностью наблюдаема [16]. При этом уравнения наблюдения описываются уравнениями (5) при следующих условиях: во-первых, уравнения наблюдения не содержат пуассоновского шума ($c_1 \equiv 0$), а во-вторых, коэффициент при винеровском шуме b_1^Y не зависит от состояния $\bar{X}_t(b_1(\bar{X}^t, Y_t)) = b_1(Y_t, t)$. Размерность вектора Y_t примем равной n^Y, а $\bar{X}_t$ — $n^{\bar{X}}$. В дальнейшем для краткости будем принимать $\bar{X}_t = X_t$, а уравнения «система плюс наблюдения» записывать в следующем виде [16]:

$$dX_t = $$

$$= a\left(X_t, Y_t, t, \Theta\right) dt + b\left(X_t, Y_t, t, \Theta\right) dW_0 + \int\limits_{R_0^q} c\left(X_t, Y_t, t, \Theta, u\right) P^0(dt, du, \Theta), \quad (6)$$

$$dY_t = a_1\left(X_t, Y_t, t, \Theta\right) dt + b_1\left(Y_t, t, \Theta\right) dW_0. \quad (7)$$

Для СтС (6), (7) с аддитивными шумами имеют место следующие условия:

$$b\left(X_t, Y_t, t, \Theta\right) = b_0(t, \Theta); \;\; c\left(X_t, Y_t, t, \Theta, v\right) = 0; \;\; b_1\left(Y_t, t, \Theta\right) = b_{10}(t, \Theta). \quad (8)$$

Условия (8) будут выполняться, если в последнем уравнении (3) с номером h появился белый шум, а коэффициент при нем зависит только от t и Θ.

Как известно из теории субоптимальной фильтрации [16] для гауссовской СтС, так как гауссовское (нормальное) распределение, аппроксимирующее апостериорное распределение вектора X_t, полностью определяется апостериорными математическим ожиданием $\hat{X}_t$ и ковариационной матрицей R_t вектора X_t, то при аппроксимации апостериорного распределения вектора X_t нормальным распределением все математические ожидания для $d\hat{X}_t$ и dR_t будут определенными функциями $\hat{X}_t$, R_t и t, т. е. будут представлять собой стохастические дифференциальные уравнения, определяющие $\hat{X}_t$ и R_t:

$$d\hat{X}_t = B\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) = f\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) dt +$$

$$+ h\left(\hat{X}_t, Y_t, R_t, t, , \Theta\right) dt \left[dY_t - f^{(1)}\left(\hat{X}_t, Y_t, R_t, t\right) dt\right]; \quad (9)$$

$$dR_t = \left\{ f^{(2)}\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) - \right.$$
$$\left. - h\left(\hat{X}_t, Y_t, R_t, t, \Theta\right)\left(b_1 \nu b_1^{\mathrm{T}}\right)(Y_t, t, \Theta)\, h\left(\hat{X}_t, Y_t, R_t, t, \Theta\right)^{\mathrm{T}} \right\} dt +$$
$$+ \sum_{r=1}^{n_y} \rho_r(\hat{X}_t, Y_t, R_t, t, \Theta)\left[dY_r - f_r^{(1)}(\hat{X}_t, Y_t, R_t, t, \Theta)\, dt\right],$$

где $\nu = \nu(t)$ — интенсивность белого шума;

$$f(\hat{X}_t, Y_t, R_t, t, \Theta) =$$
$$= [(2\pi)^n\ |R_t|\,]^{-1/2} \int\limits_{-\infty}^{\infty} a\,(Y_t, x, t, \Theta) \exp\left\{ -\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) R_t^{-1} \frac{x - \hat{X}_t}{2} \right\} dx, \quad (10)$$

$$f^{(1)}(\hat{X}_t, Y_t, R_t, t, \Theta) = \left\{ f_r^{(1)}\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) \right\} =$$
$$= [(2\pi)^{n_x}\ |R_t|\,]^{-1/2} \int\limits_{-\infty}^{\infty} a_1(Y_t, x, t) \exp\left\{ -\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) R_t^{-1} \frac{x - \hat{X}_t}{2} \right\} dx\,; \quad (11)$$

$$h\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) =$$
$$= \left\{ [(2\pi)^{n_x}\ |R_t|\,]^{-1/2} \int\limits_{-\infty}^{\infty} \left[x a_1(Y_t, x, t)^{\mathrm{T}} + b\nu b_1^{\mathrm{T}}\,(Y_t, x, t, \Theta)\right] \times \right.$$
$$\left. \times \exp\left\{ -(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}) R_t^{-1} \frac{x - \hat{X}_t}{2} \right\} dx - \hat{X}_t f^{(1)}\left(\hat{X}_t, Y_t, R_t, t, \Theta\right)^{\mathrm{T}} \right\} \times$$
$$\times \left(b_1 \nu b_1^{\mathrm{T}}\right)^{-1}(Y_t, t)\,; \quad (12)$$

$$f^{(2)}\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) = [(2\pi)^{n_x}\ |R_t|\,]^{-1/2} \int\limits_{-\infty}^{\infty} \left\{ (x - \hat{X}_t) a(Y_t, x, t, \Theta)^{\mathrm{T}} + \right.$$
$$\left. + a\,(Y_t, x, t, \Theta)\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) + b\nu b_1^{\mathrm{T}}(Y_t, x, t, \Theta) \right\} \times$$
$$\times \exp\left\{ -\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) R_t^{-1} \frac{x - \hat{X}_t}{2} \right\} dx; \quad (13)$$

$$\rho_r\left(\hat{X}_t, Y_t, R_t, t, \Theta\right) =$$

$$= [(2\pi)^{n_x}\ |R_t|\,]^{-1/2} \int\limits_{-\infty}^{\infty} \left\{\left(x - \hat{X}_t\right)\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right)\alpha_r\left(Y_t, x, t, \Theta\right) + \right.$$

$$\left. + \left(x - \hat{X}_t\right)\beta_r\left(Y_t, x, t, \Theta\right)^{\mathrm{T}}\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) + \beta_r\left(Y_t, x, t, \Theta\right)\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right)\right\} \times$$

$$\times \exp\left\{-\left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) R_t^{-1} \frac{x - \hat{X}_t}{2}\right\} dx \quad (r = 1, \ldots, n_y). \quad (14)$$

Здесь α_r — r-й элемент матрицы-строки $(a_1 - \hat{a}_1^{\mathrm{T}})(b_1 \nu b_1^{\mathrm{T}})^{-1}$; β_{kr} — элемент k-й строки и r-го столбца матрицы $b\nu b_1^{\mathrm{T}}(b_1 \nu b_1^{\mathrm{T}})^{-1}$, при этом $\beta_r = [\beta_{1r} \cdots \beta_{pr}]^{\mathrm{T}}$ $(r = 1, \ldots, n_X)$.

За начальные значения $\hat{X}_t$ и R_t при интегрировании уравнений (8) и (9), естественно, следует принять условные математическое ожидание и ковариационную матрицу величины X_0 относительно Y_0:

$$\hat{X}_0 = \mathsf{M}\left[X_0 \mid Y_0\right]; \quad R_0 = \mathsf{M}\left[\left(X_0 - \hat{X}_0\right)\left(X_0^{\mathrm{T}} - \hat{X}_0^{\mathrm{T}}\right) \mid Y_0\right].$$

Если нет информации об условном распределении X_0 относительно Y_0, то начальные условия можно взять в виде:

$$\hat{X}_0 = \mathsf{M}\, X_0; \quad R_0 = \mathsf{M}\left(X_0 - \mathsf{M}\, X_0\right)\left(X_0^{\mathrm{T}} - \mathsf{M}\, X_0^{\mathrm{T}}\right).$$

Если же и об этих величинах нет никакой информации, то начальные значения $\hat{X}_t$ и R_t приходится задавать произвольно.

Таким образом, справедливо следующее утверждение.

Теорема 2. *Пусть уравнения нелинейной гауссовской дифференциальной СтС* (6) *и* (7) *допускают применение* МНА. *Тогда в основе синтеза НСОФ лежат уравнения* (10)–(14) *при соответствующих начальных условиях*.

Число уравнений МНА одномерного апостериорного распределения определяется по формуле:

$$Q_{\text{МНА}} = n_x + \frac{n_x(n_x + 1)}{2} = \frac{n_x(n_x + 3)}{2}.$$

Для негауссовских СтС, не разрешенных относительно производных, как показано в [13, 14], требуется ограниченность функций f, $f^{(1)}$, h, ρ_r и функции

$$\bar{f}^{(2)} = f^{(20} + \mathsf{M}^N \left[\int\limits_{R_0^q} cc^{\mathrm{T}} \nu_P(t, \Theta)\, dv\right]$$

(**теорема 3**).

Пусть для дифференциальной СтС с аддитивными шумами (8) функции a, b, a_1 и b_1 в (6), (7) удовлетворяют условиям:

$$\left.\begin{array}{ll} a = a\left(X_t, Y_t, t, \Theta\right) = a\left(X_t, t, \Theta\right); & a_1 = a_1\left(X_t, Y_t, t, \Theta\right) = a_1\left(X_t, t, \Theta\right); \\ b\left(X_t, Y_t, t, \Theta\right)\, dW = b(t, \Theta)\, dW_1; & b_1\left(X_t, t, \Theta\right)\, dW = dW_2. \end{array}\right\} \quad (15)$$

Здесь W_1 и W_2 — независимые винеровские процессы размерности $n_{w_1} = n_x$ и $n_{w_2} = n_y$. Тогда после перехода от дифференциалов к производным будем иметь

$$\dot{X}_t = a\left(X_t, t, \Theta\right) + b(t, \Theta) V_1; \quad Z_t = \dot{Y}_t = a_1\left(X_t, t, \Theta\right) + V_2, \quad (16)$$

где $V = [V_1 V_2]^{\mathrm{T}}$ — нормальный белый шум интенсивности $\nu = \operatorname{diag}(\nu_1, \nu_2)$.

Заменим (16) статистически линеаризованной системой, нелинейной относительно математических ожиданий m_t^x и m_t^z и линейной относительно центрированных составляющих $X_t^0 = X_t - m_t^x$ и $\hat{X}_t^{\,0} = \hat{X}_t - \hat{m}_t^x$:

$$\dot{m}_t^x = a_{00}\left(m_t^x, K_t^x, t, \Theta\right); \quad (17)$$

$$m_t^z = a_{10}\left(m_t^x, K_t^x, t, \Theta\right); \quad (18)$$

$$\dot{X}_t^0 = a_{01}\left(m_t^x, K_t^x, t, \Theta\right) X_t^0 + b(t, \Theta) V_1; \quad (19)$$

$$Z_t^0 = a_{11}\left(m_t^x, K_t^x, t, \Theta\right) X_t^0 + V_2, \quad (20)$$

где $a_{00} = a_{00}\left(m_t^x, K_t^x, t, \Theta\right)$, $a_{10} = a_{10}\left(m_t^x, K_t^x, t, \Theta\right)$, $a_{01}\left(m_t^x, K_t^x, t, \Theta\right) = \partial a_0 / \partial m_t^x$ и $a_{11}(m_t^x, K_t^x, t, \Theta) = \partial a_{10} / \partial m_t^x$ — коэффициенты статистической линеаризации нелинейных функций a и a_1, вычисляемые для нормального распределения $\mathcal{N}(m_t^x, K_t^x)$. При этом в силу (19) ковариационная матрица K_t^x будет определяться уравнением:

$$\dot{K}_t^x = a_{11}\left(m_t^x, K_t^x, t, \Theta\right) K_t^x + K_t^x a_{11}\left(m_t^x, K_t^x, t, \Theta\right)^{\mathrm{T}} + \\ + b\left(t, \Theta\right) \nu_1(t, \Theta) b(t, \Theta)^{\mathrm{T}}. \quad (21)$$

Применяя к модели (19), (20) уравнения линейного фильтра Калмана–Бьюси [16], получим искомые уравнения фильтра на основе МСЛ:

$$\dot{\hat{X}}_t = a_{00}\left(m_t^x, K_t^x, t, \Theta\right) - a_{01}\left(m_t^x, K_t^x, t, \Theta\right) m_t^x + \\ + a_{01}\left(m_t^x, K_t^x, t, \Theta\right) \hat{X}_t + R_t a_{11}\left(m_t^x, K_t^x, t, \Theta\right)^{\mathrm{T}} \nu_2(t, \Theta)^{-1} \times \\ \times \left[Z_t - a_{11}\left(m_t^x, K_t^x, t, \Theta\right) \hat{X}_t - a_{10}\left(m_t^x, K_t^x, t, \Theta\right) + a_{11}\left(m_t^x, K_t^x, t, \Theta\right) m_t^x\right], \quad (22)$$

где $\hat{X}_0 = \mathsf{M}_{\mathcal{N}} X_0$;

$$\dot{R}_t = a_{01}\left(m_t^x, K_t^x, t, \Theta\right) R_t + R_t a_{01}\left(m_t^x, K_t^x, t, \Theta\right)^{\mathrm{T}} -$$
$$- R_t a_{11}\left(m_t^x, K_t^x, t, \Theta\right)^{\mathrm{T}} \nu_2(t,\Theta)^{-1} a_{11}\left(m_t^x, K_t^x, t, \Theta\right) R_t +$$
$$+ b(t,\Theta)\nu_1(t,\Theta)b(t,\Theta)^{\mathrm{T}}, \;\; R_0 = \mathsf{M}_{\mathcal{N}}\left[\left(X_0 - \hat{X}_0\right)\left(X_0 - \hat{X}_0\right)^{\mathrm{T}}\right]. \quad (23)$$

Входящие сюда m_t^x и K_t^x определяются из уравнений (17) и (18).

Теорема 4. *Пусть в условиях аддитивных шумов* (15) *и теоремы 1 уравнения нелинейной гауссовской дифференциальной СтС* (16) *допускают применение МСЛ, а линеаризованные уравнения удовлетворяют условиям стохастической наблюдаемости. Тогда НСОФ согласно МСЛ определяется уравнениями* (22), (23) *при условиях* (17), (18) *и* (21) *и соответствующих начальных условиях*.

При чисто линейных наблюдениях, когда $a_1\left(X_t, t, \Theta\right) = b_1(t,\Theta)X_t + b_0(t,\Theta)$, уравнения (6) и (7) упрощаются, поскольку $a_{10}(m_t^x, K_t^x, t, \Theta) = b_0(t,\Theta)$, $a_{11}(m_t^x, K_t^x, t, \Theta) = b_1(t,\Theta)$ и $\beta_t = R_t b_1(t,\Theta)^{\mathrm{T}}\nu_2(t,\Theta)^{-1}$, и принимают вид:

$$\dot{\hat{X}}_t = a_{00}\left(m_t^x, K_t^x, t, \Theta\right) - a_{01}\left(m_t^x, K_t^x, t, \Theta\right) m_t^x + a_{01}\left(m_t^x, K_t^x, t, \Theta\right)\hat{X}_t +$$
$$+ \beta_t\left[Z_t - b_1(t,\Theta)\hat{X}_t - b_0(t,\Theta) + b_1(t,\Theta)m_t^x\right], \;\; \hat{X}_0 = \mathsf{M}_{\mathcal{N}}X_0; \quad (24)$$

$$\dot{R}_t = a_{01}\left(m_t^x, K_t^x, t, \Theta\right) R_t + R_t a_{01}\left(m_t^x, K_t^x, t, \Theta\right)^{\mathrm{T}} - \beta_t b_1(t,\Theta)R_t +$$
$$+ b(t,\Theta)\nu_1(t,\Theta)b(t,\Theta)^{\mathrm{T}}, \;\; R_0 = \mathsf{M}_{\mathcal{N}}\left[\left(X_0 - \hat{X}_0\right)\left(X_0 - \hat{X}_0\right)^{\mathrm{T}}\right]. \quad (25)$$

Коэффициенты статистической линеаризации a_{00}, a_{01}, a_{10} и a_{11} и вспомогательная (инструментальная) матрица R_t размерности $n_z \times n_z$ не содержат результаты наблюдений Z_t и могут быть определены отдельно (до получения результатов наблюдений). Тогда линеаризованные уравнения (17) и (19) в силу их простоты могут быть проинтегрированы в реальном масштабе времени в течение наблюдений системы. При этом возможна априорная оценка точности фильтра [16].

Алгоритмы синтеза НСОФ зависят от инструментальных параметров Θ. Они представляют собой случайные величины или медленно меняющиеся функции времени. В [3, 5, 15] для оценки качества НСОФ приведены соответствующие алгоритмы по методу теории чувствительности для условной функции потерь $\rho = R(\Theta, t)$, допускающей квадратичную аппроксимацию.

3 Нормальный фильтр для интегродифференциальных стохастических систем, не разрешенных относительно производных

Обобщая [15], рассмотрим ИДСтС, описываемую следующими векторными уравнениями:

$$\bar{\psi}_0 = \psi_0(t, \Theta, X_t, \dot{X}_t, \dots, X_t^{(k)}, U_t) + \\ + \int_{t_0}^{t} H(\Theta, t, \tau)\psi_1(\Theta, t, \tau, X_\tau, \dot{X}_\tau, \dots, X_\tau^{(l)}, U_\tau)\, d\tau = 0\,,$$

$$X_{t_0} = X_0,\ \ \dot{X}_{t_0} = \dot{X}_0, \dots X_{t_0}^{(k)} = X_0^{(k)}; \quad (26)$$

$$\dot{U}_t = a^U(t, \Theta) + a^U(t, \Theta)U_t + b^U(t, \Theta)V_t^U,\ \ U_{t_0} = U_0\,, \quad (27)$$

при условиях физической реализуемости и затухающей памяти:

$$H(t, \tau, \Theta) = [H_{ij}(t, \tau, \Theta)] = 0,\ \ \forall \tau > t,\ \ \int_{-\infty}^{\infty} |H_{ij}(t, \tau, \Theta)|\, d\tau < \infty\,. \quad (28)$$

Следуя [1, 2], в случае вырожденного ядра $H(t, \tau, \Theta)$, когда

$$H_{ij}(t, \tau, \Theta) = H_{ij}^{+}(t, \Theta)H_{ij}^{-}(\tau, \Theta),$$

положим

$$H^{+}Y_t = \int_{t_0}^{t} H(t, \tau, \Theta)\psi_1(t, \tau, \Theta, X_\tau, \dot{X}_\tau, \dots, X_\tau^{(k-1)}, U_\tau, \tau)\, d\tau\,;$$

$$\dot{Y}_t = H^{-}Y_t,\ \ Y_{t_0} = 0\,.$$

Тогда (26) примет следующий вид:

$$\bar{\psi}_0 = \psi_0(t, \Theta, X_{1t}, U_t) + H^{+}Y_t = 0,$$

где

$$X_{1t} = \left[X_t^{\mathrm{T}} \dot{X}_t^{\mathrm{T}} \cdots X_t^{(k-1)\mathrm{T}}\right]^{\mathrm{T}}.$$

Далее проведем статистическую линеаризацию по Казакову [1, 2] функций $\bar{\psi}_0$ и ψ_1 в (26):

$$\bar{\psi}_0 \approx m_0^{\bar{\psi}_0} + k_{1X_1}^{\bar{\psi}_0} X_{1t}^0 + k_{1U}^{\bar{\psi}_0} U_t^0 + k_{1Y}^{\bar{\psi}_0} Y_t^0;$$
$$\psi_1 \approx m_0^{\psi_1} + k_1^{\psi_1} X_{1t}^0 + k_{1U}^{\psi_1} U_t^0.$$

Здесь коэффициенты статистической линеаризации зависят от $m_t^{X_1}$, m_t^U, $K_t^{X_1}$, K_t^U и K_t^{XU}. В результате получим систему уравнений для математических ожиданий:

$$
\begin{aligned}
m_0^{\psi_0} + H^+ m_t^Y &= 0\,; \qquad (29)\\
H^+ m_t^Y &= \int\limits_{t_0}^{t} H(t,\tau,\Theta) m_{0\tau}^{\psi_1}\, d\tau,\\
\dot{m}^Y &= H^- m^Y;\\
\dot{m}^U &= a^U m^U
\end{aligned}
$$

и центрированных векторов $X_{1t}^0 = X_{1t} - m_t^{X_1}$, $U_t^0 = U_t - m_t^U$ и $Y_t^0 = Y_t - m_t^Y$ в виде:

$$\tilde{\varphi} = k_{X_1}^{\psi_0} X_{1t}^0 + k_{U_t}^{\psi_0} U_t^0 + H^+ Y_t^0 = 0\,; \qquad (30)$$

$$H^+ Y_t^0 = \int\limits_0^t H(t,\tau,\Theta)\left[k_{X_1\tau}^{\psi_1} X_{1\tau}^0 + k_{U\tau}^{\psi_1} U_\tau^0\right] d\tau\,; \qquad (31)$$

$$\dot{Y}_t^0 = H^- Y_t^0\,; \qquad (32)$$

$$\dot{U}_t^0 = a^U U_t^0 + b^U V_t^0\,. \qquad (33)$$

Продифференцируем уравнение (29) один раз. В результате получим:

$$\dot{m}_0^{\psi_0} + H^+ \dot{m}_t^Y + \dot{H}_t m_t^Y = 0. \qquad (34)$$

Аналогично продифференцируем один раз (34), а (30) — h раз до появления белого шума. Тогда будем иметь дополнительную систему уравнений:

$$\left.\begin{aligned}
&\dot{\tilde{\varphi}} = \dot{k}_{X_1,t}^{\psi_0} X_{1t}^0 + k_{X_1,t}^{\psi_0} \dot{X}_{1t}^0 + \dot{k}_{Ut}^{\psi_0} U_t^0 + K_{Ut}^{\psi_0} \dot{U}_t^0 + \dot{H}_t^+ Y_t^0 + H^+ \dot{Y}_t^0;\\
&\quad \ldots;\\
&\tilde{\varphi}^{(h)} = 0.
\end{aligned}\right\} \qquad (35)$$

Теперь на основе системы уравнений (35) образуем вектор центрированных переменных

$$X_{2t}^0 = \left[X_{1t}^{0\mathrm{T}} \dot{X}_{1t}^{0\mathrm{T}} \cdots X_{1t}^{0(h)\mathrm{T}}\right]^{\mathrm{T}}. \qquad (36)$$

Таким образом, составной вектор

$$Z_t = \left[X_{1t}^{0\mathrm{T}} X_{2t}^{0\mathrm{T}} Y_t^{0\mathrm{T}} U_t^{\mathrm{T}}\right]^{\mathrm{T}} \qquad (37)$$

для ИДСтС (26) с вырожденным ядром $H(t,\tau)$ при условиях (28) будет определяться векторным дифференциальным уравнением вида (5):

$$dZ_t = a^Z dt + b^Z dW_0 + \int\limits_{R_0^q} c^Z P^0(t,\Theta,du). \tag{38}$$

Здесь функции a^Z, b^Z и c^Z определяются (30)–(37).

Наконец, применим к уравнениям (38) стандартные уравнения МНА [1, 2]:

$$\left.\begin{aligned} \dot{m}_t^X &= \Phi_t^m\left(t,\Theta,m_t^X,K_t^X\right), & m_0^X &= m_{t_0}^X;\\ \dot{K}_t^X &= \Phi_t^K\left(t,\Theta,m_t^X,K_t^X\right), & K_0^X &= K_{t_0}^X;\\ \frac{\partial K^X(t_1,t_2)}{\partial t_2} &= \Phi_{t_1,t_2}^K\left(t_1,t_2,\Theta,m_{t_2}^X,K_{t_2}^X\right), & K^X(t_1,t_2) &= K_{t_1}^X, \end{aligned}\right\} \tag{39}$$

где введены обозначения:

$$\Phi_t^m(t,\Theta,m_t^X,K_t^X) = \mathsf{M}_{\mathcal{N}}\left[a^X\right]; \tag{40}$$

$$\begin{aligned} \Phi_t^K\left(t,\Theta,m_t^X,K_t^X\right) &= \mathsf{M}_{\mathcal{N}}\Big[\Phi_{1t}\left(t,\Theta,m_t^X,K_t^X\right) + \Phi_{1t}^{\mathrm{T}}\left(t,\Theta,m_t^X,K_t^X\right) + \\ &\quad + \Phi_{2t}\left(t,\Theta,m_t^X,K_t^X\right)\Big], \\ \Phi_{1t}\left(t,\Theta,m_t^X,K_t^X\right) &= \mathsf{M}_{\mathcal{N}}\left[a^X\left(X_t - m_t^X\right)^{\mathrm{T}}\right], \\ \Phi_{2t}\left(t,\Theta,m_t^X,K_t^X\right) &= \mathsf{M}_{\mathcal{N}}\left[\sigma\left(t,\Theta,X_t\right)\right]; \end{aligned} \tag{41}$$

$$\sigma\left(t,\Theta,X_t\right) = \sigma_0\left(t,\Theta,X_t\right) + \int\limits_{R_0^q} c^X\left(t,\Theta,X_t,u\right)\left[c^X\left(t,\Theta,X_t,u\right)^{\mathrm{T}}\right]\nu_P(t,\Theta,du),$$

$$\sigma_0\left(t,\Theta,X_t\right) = b^X(t,\Theta,X_t)\nu_0(t,\Theta)b^X(t,\Theta,X_t)^{\mathrm{T}}; \tag{42}$$

$$\Phi_{t_1,t_2}^k\left(t_1,t_2,\Theta,m_{t_2}^X,K_{t_2}^X\right) = K^X\left(t_1,t_2\right)\left(K_{t_2}^Z\right)^{-1}\Phi_{1t}^{\mathrm{T}}. \tag{43}$$

В стационарном случае уравнения МНА имеют следующий вид:

$$\left.\begin{aligned} \Phi_*^m\left(\Theta,m_*^Z,K_*^Z\right) &= 0, \quad \Phi_*^K\left(\Theta,m_*^Z,K_*^Z\right) = 0;\\ \frac{dk^Z(\tau)}{dt} &= \Phi_\tau^k\left(k^Z(\tau),\Theta,m_*^Z,K_*^Z\right), \quad k^Z(0) = K_*^Z, \end{aligned}\right\} \tag{44}$$

где принято

$$\Phi_*^m\left(\Theta,m_*^Z,K_*^Z\right) = \mathsf{M}_{\mathcal{N}}\left[a^Z\right]; \tag{45}$$

$$\Phi_*^K\left(\Theta, m_*^Z, K_*^Z\right) = \Phi_{1*}\left(\Theta, m_*^Z, K_*^Z\right) + \Phi_{1*}\left(\Theta, m_*^Z, K_*^Z\right)^{\mathrm{T}} + \Phi_{2*}\left(\Theta, m_*^Z, K_*^Z\right),$$
$$\Phi_{1*}\left(\Theta, m_*^Z, K_*^Z\right) = \mathrm{M}_{\mathcal{N}}\left[a^Z\left(Z_t - m_*^Z\right)\right],$$
$$\Phi_{2*}\left(\Theta, m_*^Z, K_*^Z\right) = \mathrm{M}_{\mathcal{N}}\left[\sigma_*\left(\Theta, Z_t\right)\right]; \quad (46)$$

$$\sigma_*\left(\Theta, Z_t\right) = \sigma_{0*}\left(\Theta, Z_t\right) + \int\limits_{R_0^q} c^Z\left(\Theta, Z_t, u\right) c^Z\left(\Theta, Z_t, u\right)^{\mathrm{T}} \nu_P(t, \Theta, du),$$
$$\sigma_{0*}\left(\Theta, Z_t\right) = b^Z\left(\Theta, Z_t\right) \nu_{0*}(\Theta) b^Z\left(\Theta, Z_t\right)^{\mathrm{T}}, \quad (47)$$

$$\Phi_\tau^k\left(k^Z(\tau), \Theta, m_*^Z, K_*^Z\right) = k^Z(\tau)\left(K_*^Z\right)^{-1}\Phi_{1*}^{\mathrm{T}}, \ \ \tau = t_1 - t_2. \quad (48)$$

Таким образом, имеем следующие результаты.

Теорема 5. *В условиях теоремы* 1 *уравнения МАМ для ИДСтС, приведенной к дифференциальной системе* (38) *на основе МНА, имеют вид* (39) *при условиях* (40)–(43). *В стационарном случае имеем уравнения* (44)–(48).

Для ИДСтС с аддитивными шумами уравнения МАМ по МНА переходят в уравнения по МСЛ следующего вида:

$$\dot{Z}_t = a_0^э + a^э Z_t + b^э V. \quad (49)$$

Здесь вектор $a_0^э$ и матрица $a^э$ зависят от коэффициентов статистической линеаризации, следовательно, от $m_t^{X_1}$, m_t^U, K^{X_1}, K^U и K^{X_1U}, а матрица $b^э$ может зависеть от времени. В таком случае в силу (49) уравнения МАМ по МСЛ примут вид (**теорема 6**):

$$\dot{m}_t^Z = a_0^э, \qquad m_{t_0}^э = m_0^э;$$
$$\dot{K}_t^Z = a^э K_t^Z + K_t^Z\left(a^э\right)^{\mathrm{T}} + b^э \nu \left(b^э\right)^{\mathrm{T}}, \qquad K_{t_0}^Z = K_0^Z;$$
$$\frac{\partial K^Z\left(t_1, t_2\right)}{\partial t_2} = K^Z\left(t_1, t_2\right)\left(a_{t_2}^э\right)^{\mathrm{T}}, \qquad K^Z\left(t_1, t_1\right) = K_{t_1}^Z.$$

Наконец, для наблюдений (7), основываясь на теоремах разд. 2 для ИДСтС (26) и (27), приведенной к дифференциальной методом вырожденных ядер, придем к следующим результатам.

Теорема 7. *Пусть гауссовская ИДСтС описывается уравнениями* (6), (7), *причем ее ядро* $H = H(t, \tau, \Theta)$ *является вырожденным. Тогда алгоритм синтеза НСОФ определяется теоремой* 2.

Теорема 8. *Пусть гауссовская ИДСтС описывается уравнениями* (6), (7) *с аддитивными гауссовскими возмущениями, а ядро вырожденно. Тогда алгоритм синтеза НСОФ определяется теоремой* 4.

Теорема 9. *Пусть негауссовская ИДСтС описывается уравнениями* (6), (7) *и ядро вырожденно. Тогда алгоритм синтеза НСОФ определяется теоремой* 3.

4 Пример

Рассмотрим скалярную гауссовскую СтС, не разрешенную относительно производной, вида

$$\int_{t_0}^{t} e^{-\alpha(t-\tau)} \operatorname{sgn} \dot{X}_\tau \, d\tau = a_0 + aX + bU \,; \tag{50}$$

$$\dot{U} = \gamma U + V_1. \tag{51}$$

Здесь a_0, a, b и γ — постоянные, а V_1 — гауссовский белый шум интенсивности ν_1. Получим уравнения МАМ.

Выполним с ИДСтС (50) следующие преобразования:

1. Введем переменные $X_1 = X$, $X_2 = \dot{X}_1$ и $X_3 = U$. Заменим нелинейную функцию $\varphi = \operatorname{sgn} \dot{X} = \operatorname{sgn} X_2$ статистически линеаризованной по Казакову:

$$\varphi(X_2) \approx \tilde{\varphi}_0 + k_1^{\varphi} X_2.$$

Здесь

$$\dot{\tilde{\varphi}}_0 = \varphi_0 - k_1^{\varphi} m^{X_2},$$

где

$$\varphi_0 = 2\Phi(\zeta); \quad \zeta = \frac{m^{X_2}}{\sqrt{D^{X_2}}} \,;$$

$$k_1^{\varphi} = \frac{\partial \varphi_0}{\partial m^{X_2}} = \frac{2}{\sqrt{2\pi D^{X_2}}} \exp\left[-\frac{(m^{X_2})^2}{2D^{X_2}}\right]; \quad \Phi(\zeta) = \frac{1}{\sqrt{2\pi}} \int_0^{\zeta} e^{-(t^2/2)} dt \,.$$

2. Введем инструментальную переменную X_4 согласно уравнению

$$\frac{1}{\alpha} X_4 = \int_{t_0}^{t} e^{-\alpha(t-\tau)} \left(\tilde{\varphi}_{0\tau} + k_{1\tau}^{\varphi} X_{2\tau}\right) d\tau \,. \tag{52}$$

Пользуясь формулой дифференцирования интеграла, зависящего от параметра

$$\frac{d}{d\lambda} \int_{u(\lambda)}^{v(\lambda)} \frac{\partial}{\partial \lambda} f(X, \lambda) \, dX + f(v(\lambda), \lambda) \frac{dv}{d\lambda} - f(u(\lambda), \lambda) \frac{du}{d\lambda} \,,$$

заключаем, что X_4 удовлетворяет дифференциальному уравнению

$$\dot{X}_4 = -\alpha \left(X_4 - \tilde{\varphi}_0 - k_1^{\varphi} X_2\right). \tag{53}$$

3. Перепишем (50) и (51) с учетом (52) в виде:

$$a_0 + aX_1 + bX_3 - \frac{1}{\alpha} X_4 = 0\,; \tag{54}$$

$$\dot{X}_3 = \gamma X_3 + V_1 \equiv a_{33} X_3 + V_1. \tag{55}$$

4. Продифференцируем (54) один раз:

$$a\dot{X}_1 + b\dot{X}_3 - \frac{1}{\alpha}\,\dot{X}_4 = 0\,,$$

а затем разрешим уравнение относительно $X_2 = \dot{X}_1$:

$$X_2 = \frac{1}{a + (1+\alpha b) k_{16}^{\varphi}} \left[-(1+\alpha b)\tilde{\varphi}_0 + \alpha b X_3 + X_4\right] \equiv$$

$$\equiv A_0^2 + A_0^2 X_3' + A_0^2 X_4. \tag{56}$$

В результате ИДСтС (50) и (51) будет заменена на эквивалентную статистически линеаризованную систему уравнений, состоящую из уравнений

$$\dot{X}_1 = X_2,$$

а также уравнений (53)–(55), (56).

Упростим уравнения (53) с учетом (56). Тогда получим

$$\dot{X}_4 = -\alpha X_4 + \alpha\tilde{\varphi}_0 + \frac{\alpha k_1^{\varphi}}{a + (1+\alpha b) k_1^{\varphi}} \left[-(1+\alpha b)\tilde{\varphi}_0 + \alpha b X_3 + X_4\right] \equiv$$

$$\equiv a_{40} + a_{43} X_3 + a_{44} X_4. \tag{57}$$

Уравнения (55) и (57) отделяются от уравнений для X_1 и X_2 в силу (55) и (56). Пользуясь формулами теории линейных СтС [1, 2], получим искомые уравнения для математических ожиданий и ковариационной матрицы вектора $X^{34} = [X_3 X_4]^{\mathrm{T}}$:

$$\dot{m}^{34} = a_0^{34} + a^{34} m^{34};$$
$$\dot{K}^{34} = a^{34} K^{34} + K^{34} (a^{34})^{\mathrm{T}} + b^{34} \nu_1 (b^{34})^{\mathrm{T}}.$$

Здесь введены обозначения:

$$m^{34} = \begin{bmatrix} m_3 \\ m_4 \end{bmatrix};\ \ K^{34} = \begin{bmatrix} D_3 & K_{34} \\ K_{34} & D_4 \end{bmatrix};\ \ a^{34} = \begin{bmatrix} a_{33} & 0 \\ a_{43} & a_{44} \end{bmatrix};\ \ b^{34} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Для переменных X_1 и X_2 в силу (54), если учесть соотношение

$$X_1 = \frac{1}{a}\left(-a_0 - bX_3 + \frac{1}{\alpha}X_4\right) = A_0^1 + A_3^1 X_3 + A_4^1 X_4, \tag{58}$$

получим уравнения для математических ожиданий и ковариационной матрицы вектора $X^{12} = [X_1 X_2]^{\mathrm{T}}$ следующие уравнения:

$$\dot{m}^{12} = a_0^{12} + a^{12} m^{12} = 0\,;$$
$$\dot{K}^{12} = a^{12} K^{12} + K^{12} (a^{12})^{\mathrm{T}},$$

где

$$a_0^{12} = \begin{bmatrix} A_0^1 \\ A_0^2 \end{bmatrix};\ a^{12} = \begin{bmatrix} A_{13}^1 & A_{14}^1 \\ A_{13}^2 & A_{14}^2 \end{bmatrix}.$$

В силу зависимости φ_0 и k_1^{φ} от m^{X_2} и D^{X_2} полученные системы уравнений МАМ взаимосвязаны и решаются численно стандартными методами численного анализа.

Теперь, пользуясь уравнениями (24) и (25), в случае линейного несмешанного измерения инструментальной переменной X_4

$$Z_4 = \dot{Y}_4 = X_4 + V_2$$

в силу (56) и (58) имеем соотношения:

$$\dot{\hat{X}}^{12} = a_0^{12} + a_1^{12} \hat{X}^{34}.$$

При этом алгоритм синтеза НСОФ в силу (24) и (25) будет определяться следующими уравнениями:

$$\dot{\hat{X}}^{34} = a_{06}^{34} m^{34} + a_{01} \hat{X}^{34} + \beta^{34}\left(Z^{34} - \hat{X}^{34} + m^{34}\right);$$
$$\dot{k}^{34} = a_{01}^{34} k^{34} + k^{34}\left(a_{01}^{34}\right)^{\mathrm{T}} - \beta k^{34} + b^{34} \nu_1 (b^{34})^{\mathrm{T}},$$

где $\beta^{34} = k^{34} \nu_2^{-1}$. Входящие сюда матрицы a^{34} и b^{34} определены в (13) и зависят от m^{34} и K^{34}. Заметим, что НСОФ позволяет проводить фильтрацию в масштабе реального времени, поскольку вычисление β и R проводится заранее и не требует текущих измерений.

5 Заключение

Для нелинейных ИДСтС, не разрешенных относительно производных, приводимых к дифференциальным методом сингулярных (вырожденных) ядер, разработаны методы и алгоритмы:

– аналитического моделирования нормальных СтП, при этом нелинейность под интегралом может быть разрывной;

– синтеза НСОФ для онлайн-обработки информации в ИДСтС.

Результаты допускают обобщение на случай дифференцируемых нелинейностей методом сведения ядер к дифференциальным, если воспользоваться [1, 2].

Предложены алгоритмы оценки качества НСОФ на основе теории чувствительности.

Подробно рассмотрен пример с разрывной нелинейностью под знаком интеграла.

В настоящее время идет разработка учебного экспериментального программного обеспечения по тематике [16–20] для случаев, когда в уравнениях состояния можно пренебречь постоянными времени высокого порядка.

В качестве дальнейших обобщений могут рассматриваться следующие задачи:

– аналитическое моделирование на основе уравнений теории марковских процессов и параметризации распределений методами моментов, семиинвариантов, ортогональных разложений, эллипсоидальной аппроксимации и др.;

– синтез среднеквадратичных УОФ, экстраполяторов и интерполяторов на основе параметризации распределений;

– в том случае, когда вектор возмущений и/или нелинейные функции заданы каноническими представлениями случайных функций, полученные результаты допускают обобщения на более общие классы ИДСтС, если воспользоваться [2, 4, 19].

Теоретический и практический интерес представляют задачи, когда функции (1) являются стохастическими функциями отмеченных переменных.

Литература

1. *Пугачев В. С., Синицын И. Н.* Стохастические дифференциальные системы. Анализ и фильтрация. — М.: Наука, 1990. 632 с.
2. *Пугачев В. С., Синицын И. Н.* Теория стохастических систем. — М.: Логос, 2000; 2004. 1000 с.
3. *Синицын И. Н.* Анализ и моделирование распределений в эредитарных стохастических системах // Информатика и её применения, 2014. Т. 8. Вып. 1. С. 2–11.
4. *Синицын И. Н., Сергеев И. В., Синицын В. И., Корепанов Э. Р., Белоусов В. В.* Математическое обеспечение параметрического моделирования распределений в интегродифференциальных системах // Системы и средства информатики, 2014. Т. 24. № 1. С. 4-45.
5. *Синицын И. Н.* Аналитическое моделирование распределений с инвариантной мерой в негауссовских дифференциальных и приводимых к ним эредитарных стохастических системах // Информатика и её применения, 2014. Т. 8. Вып. 2. С. 2–14.

6. *Синицын И. Н., Синицын В. И.* Нормальные и эллипсоидальные распределения в интегродифференциальных системах // Системы компьютерной математики и их приложения: Мат-лы XV Междунар. конф. — Смоленск: СмолГУ, 2014. Вып. 15. С. 196–197.
7. *Синицын И. Н., Синицын В. И., Корепанов Э. Р.* Аналитическое моделирование эредитарных стохастических систем со сложными нелинейностями // Современные проблемы прикладной математики, информатики автоматизации и управления: Мат-лы IV Междунар. научн.-технич. семинара. — М.: ИПИ РАН, 2014. CD-R. С. 29–34.
8. *Синицын И. Н., Сергеев И. В., Синицын В. И., Корепанов Э. Р., Белоусов В. В.* Аналитическое компьютерное моделирование в эредитарных стохастических системах с автокоррелированными шумами // Кибернетика и высокие технологии XXI века: Сб. докл. XV Междунар. научн.-технич. конф. — Воронеж: САКВОЕЕ, 2014. CD-R. С. 543–551.
9. *Синицын И. Н., Корепанов Э. Р., Белоусов В. В.* Развитие математического обеспечения аналитического и статистического моделирования эредитарных стохастических систем // Идентификация систем и задачи управления: Тр. X Междунар. конф. — М.: ИПУ РАН, 2015. С. 1275–1297.
10. *Sinitsyn I. N.* Analytical modeling of “linear” and circular control stochastic systems // Advances in robotics and automatic control: Reviews. — Book ser. — IFSA Publishing, 2021 (in press). Vol. 2.
11. *Синицын И. Н., Синицын В. И., Сергеев И. В., Корепанов Э. Р., Белоусов В. В.* Условно оптимальная фильтрация нормальных процессов в эредитарных стохастических системах // Современные проблемы прикладной математики, информатики, автоматизации и управления: Тр. 5-го юбилейного семинара / Под ред. И. А. Соколова, В. И. Кошкина. — Севастополь: СевГУ, 2015. С. 23–33.
12. *Синицын И. Н., Сергеев И. В., Синицын В. И., Корепанов Э. Р., Белоусов В. В., Шоргин В. С.* Математическое обеспечение синтеза дискретных фильтров Пугачёва для обработки нормальных процессов в эредитарных стохастических системах // Системы и средства информатики, 2015. Т. 25. № 2. С. 62–101.
13. *Синицын И. Н.* Аналитическое моделирование широкополосных процессов в стохастических системах, не разрешенных относительно производных // Информатика и её применения, 2017. Т. 11. Вып. 1. С. 3–10.
14. *Синицын И. Н.* Параметрическое аналитическое моделирование процессов в стохастических системах, не разрешенных относительно производных // Системы и средства информатики, 2017. Т. 27. № 1. С. 21–45.
15. *Синицын И. Н.* Нормальные субоптимальные фильтры для дифференциальных стохастических систем, не разрешенных относительно производных // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 3–10.
16. *Синицын И. Н.* Фильтры Калмана и Пугачева. — 2-е изд. — М.: Логос, 2007. 776 с.
17. ГОСТ 23743–88. Изделия авиационной техники. Номенклатура показателей безопасности полета, надежности, контролепригодности, эксплуатационной и ремонтной технологичности. — М.: Изд-во стандартов, 1993. 84 с.
18. *Александровская Л. Н., Аронов И. З., Круглов В. И. и др.* Безопасность и надежность технических систем. — М.: Университетская книга, Логос, 2008. 376 с.

19. *Синицын И. Н.* Канонические представления случайных функций и их применение в задачах компьютерной поддержки научных исследований. — М.: ТОРУС ПРЕСС, 2009. 768 с.
20. *Синицын И. Н.* Лекции по теории систем интегрированной логистической поддержки. — М.: ТОРУС ПРЕСС, 2019. 1072 с.

Поступила в редакцию 08.07.20

ANALYTICAL MODELING AND FILTERING FOR INTEGRODIFFERENTIAL SYSTEMS WITH UNSOLVED DERIVATIVES

I. N. Sinitsyn

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

Abstract: For nonlinear integrodifferential stochastic systems (IDStS) with unsolved derivatives reducible to differential stochastic systems (StS) by means of singular kernels, the following methods and algorithms are proposed: analytical modeling of normal (Gaussian) stochastic processes and analytical synthesis of normal suboptimal filters for information processing in IDStS. Both Gaussian and non-Gaussian StS white noises are considered. Quality estimation methods based on the sensitivity theory are suggested. An example with discontinuous nonlinearity is considered in details. Directions for future investigations are given.

Keywords: integrodifferential stochastic system (IDStS); method of analytical modeling (MAM); method of normal approximation (MNA); method of statistical linearization (MSL); normal suboptimal filter (NSOF); stochastic system (StS); stochastic systems with unsolved derivatives

DOI: 10.14357/08696527210104

References

1. Pugachev, V. S., and I. N. Sinitsyn. 1987. *Stochastic differential systems. Analysis and filtering*. Chichester, NY: J. Wiley & Sons. 549 p.
2. Pugachev, V. S., and I. N. Sinitsyn. 2001. *Stochastic systems. Theory and applications*. Singapore: World Scientific. 908 p.
3. Sinitsyn, I. N. 2014. Analiz i modelirovanie raspredeleniy v ereditarnykh stokhasticheskikh sistemakh [Analysis and modeling of distributions in hereditary stochastic systems]. *Informatika i ee Primeneniya — Inform. Appl.* 8(1):2–11.

4. Sinitsyn, I. N., I. V. Sergeev, V. I. Sinitsyn, E. R. Korepanov, and V. V. Belousov. 2014. Matematicheskoe obespechenie parametricheskogo modelirovaniya raspredeleniy v integrodifferentsial'nykh stokhasticheskikh sistemakh [Mathematical software for parametric modeling of distributions in integrodifferential stochastic systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 24(1):4–45.
5. Sinitsyn, I. N. 2014. Analiticheskoe modelirovanie raspredeleniy s invariantnoy meroy v negaussovskikh differentsial'nykh i privodimykh k nim ereditarnykh stokhasticheskikh sistemakh [Analytical modeling of distributions with invariant measure in non-Gaussian differential and reducible to differential hereditary stochastic systems]. *Informatika i ee Primeneniya — Inform. Appl.* 8(2):2–14.
6. Sinitsyn, I. N., and V. I. Sinitsyn. 2014. Normal'nye i ellipsoidal'nye raspredeleniya v integrodifferentsial'nykh sistemakh [Normal and ellipsoidal distributions in integrodifferential systems]. *Sistemy komp'yuternoy matematiki i ikh prilozheniya: Matly XV Mezhdunar. konf.* [15th Conference (International) "Systems of Computer Mathematics and Their Applications" Proceedings]. Smolensk: SmolGU. 15:196–197.
7. Sinitsyn, I. N., V. I. Sinitsyn, and E. R. Korepanov. 2014. Analiticheskoe modelirovanie ereditarnykh stokhasticheskikh sistem so slozhnymi nelineynostyami [Analytical modeling of hereditary stochastic systems with complex nonlinearities]. *Sovremennye problemy prikladnoy matematiki, informatiki, avtomatizatsii, upravleniya: Mat-ly 4-go Mezhdunar. nauchn.-tekhnich. seminara* [4th Science and Technology Seminar (International) "Recent Developments in Applied Mathematics, Computer Science, Automation, and Control" Proceedings]. Moscow: IPI RAN. CD-R. 29–34.
8. Sinitsyn, I. N., I. V. Sergeev, V. I. Sinitsyn, E. R. Korepanov, and V. V. Belousov. 2014. Analiticheskoe komp'yuternoe modelirovanie v ereditarnykh stokhasticheskikh sistemakh s avtokorrelirovannymi shumami [Analytical computer modeling in hereditary stochastic systems with autocorrelated noise]. *Kibernetika i vysokie tekhnologii XXI veka: Sb. dokl. XV Mezhdunar. nauchn.-tekhnich. konf.* [15th Science and Technology Conference (International) "Cybernetics and High Technology of the XXI Century" Proceedings]. Voronezh: SAKVOEE. CD-R. 543–551.
9. Sinitsyn, I. N., E. R. Korepanov, and V. V. Belousov. 2015. Razvitie matematicheskogo obespecheniya analiticheskogo i statisticheskogo modelirovaniya ereditarnykh stokhasticheskikh sistem [Development of analytical and statistical modeling for gereditary stochastic systems]. *Identifikatsiya sistem i zadachi upravleniya: Tr. X Mezhdunar. konf.* [10th Conference (International) on System Identification and Control Problems Proceedings]. Moscow: IPU RAN. 1275–1297.
10. Sinitsyn, I. N. 2021 (in press.) Analytical modeling of "linear" and circular control stochastic systems. *Advances in robotics and automatic control: Reviews*. Book ser. IFSA Publishing. 2.
11. Sinitsyn, I. N., V. I. Sinitsyn, I. V. Sergeev, E. R. Korepanov, and V. V. Belousov. 2015. Uslovno optimal'naya fil'tratsiya normal'nykh protsessov v ereditarnykh stokhasticheskikh sistemakh [Conditionally optimal filtering of normal processes in hereditary stochastic systems]. *Sovremennye problemy prikladnoy matematiki, informatiki, avtomatizatsii i upravleniya: Tr. 5-go yubileynogo seminara* ["Modern Problems of Applied Mathematics, Informatics, Automation and Control": 5th Anniversary Seminar Proceedings]. Sevastopol: SevSU. 23–33.
12. Sinitsyn, I. N., I. V. Sergeev, V. I. Sinitsyn, E. R. Korepanov, V. V. Belousov, and V. S. Shorgin. 2015. Matematicheskoe obespechenie sinteza diskretnykh fil'trov

Pugacheva dlya obrabotki normal'nykh protsessov v ereditarnykh stokhasticheskikh sistemakh [Software for synthesis of discrete Pugachev filters for normal processes in hereditary stochastic systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(2):62–101.

13. Sinitsyn, I. N. 2017. Analiticheskoe modelirovanie shirokopolosnykh protsessov v stokhasticheskikh sistemakh, ne razreshennykh otnositel'no proizvodnykh [Analytical modeling of wide band processes in stochastic systems with unsolved derivatives]. *Informatika i ee Primeneniya — Inform. Appl.* 11(1):3–10.
14. Sinitsyn, I. N. 2017. Parametricheskoe analiticheskoe modelirovanie protsessov v stokhasticheskikh sistemakh, ne razreshennykh otnositel'no proizvodnykh [Parametric analytical modeling of processes in stochastic systems that are not allowed with respect to derivatives]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 27(1):21–45.
15. Sinitsyn, I. N. 2021. Normal'nye suboptimal'nye fil'try dlya differentsial'nykh stokhasticheskikh sistem, ne razreshennykh otnositel'no proizvodnykh [Normal suboptimal filtering for differential stochastic systems with unsolved derivatives]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):3–10.
16. Sinitsyn, I. N. 2007. *Fil'try Kalmana i Pugacheva* [Kalman and Pugachev filters]. 2nd ed. Moscow: Logos. 776 p.
17. GOST 23743-88. 1993. Izdeliya aviatsionnoy tekhniki. Nomenklatura pokazateley bezopasnosti poleta, nadezhnosti, kontroleprigodnosti, ekspluatatsionnoy i remontnoy tekhnologichnosti [Aircraft products. Nomenclature of flight safety, reliability, testability, and operational and repair manufacturability]. Moscow: Standards Publs. 84 p.
18. Aleksandrovskaya, L. N., I. Z. Aronov, V. I. Kruglov, *et al.* 2008. *Bezopasnost' i nadezhnost' tekhnicheskikh system* [Security and reliability of technical systems]. Moscow: Universitetskaya kniga, Logos. 376 p.
19. Sinitsyn, I. N. 2009. *Kanonicheskie predstavleniya sluchaynykh funktsiy i ikh primenenie v zadachakh komp'yuternoy podderzhki nauchnykh issledovaniy* [Canonical expansions of random functions and their applications in computer-aided support]. Moscow: TORUS PRESS. 768 p.
20. Sinitsyn, I. N. 2019. *Lektsii po teorii sistem integrirovannoy logisticheskoy podderzhki* [Lectures on theory of integrated logistic support systems]. Moscow: TORUS PRESS. 1072 p.

Received July 8, 2020

Contributor

Sinitsyn Igor N. (b. 1940) — Doctor of Science in technology, professor, Honored scientist of RF, 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; sinitsin@dol.ru

СТРАТЕГИЯ ИССЛЕДОВАНИЙ И РАЗРАБОТОК В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА I: ОСНОВНЫЕ ПОНЯТИЯ И КРАТКАЯ ХРОНОЛОГИЯ

А. В. Борисов[1]***, А. В. Босов***[2]***, Д. В. Жуков***[3]

Аннотация: Статья начинает цикл работ, представляющих результаты выполненного исследования влияния государственного управления на эффективность проведения исследований и разработок в области искусственного интеллекта (Artificial Intelligence Research and Development, AI R&D), ставших стратегически важной отраслью любой технологически развитой страны. Первая часть посвящена обсуждению разных трактовок термина «искусственный интеллект» и смежных с ним понятий. Представлено современное деление мировой истории развития AI R&D на эпохи. Так как практические результаты в данной области неотъемлемы от развития аппаратной платформы, в исторической ретроспективе представлена параллельная хронология отечественных достижений в области создания средств вычислительной техники и решений государственных органов, стимулирующих развитие данной области.

Ключевые слова: искусственный интеллект; смежные понятия и технологии искусственного интеллекта; временная хронология

DOI: 10.14357/08696527210105

1 Введение

В настоящее время опубликовано множество обзоров, касающихся различных аспектов *исследований и разработок в области искусственного интеллекта* [1–4]. Из-за обширности объекта исследований данные работы акцентируются на различных частных аспектах: начиная с анализа результатов внедрения AI в различные сферы жизни общества, интеграции математических знаний, алгоритмического и информационного обеспечения решения некоторого класса задач, связанных с AI, и заканчивая формированием портфелей футуристических проектов, нацеленных на далекую перспективу.

Цель данного цикла статей — освещение и анализ государственного участия в планировании и финансировании AI R&D на примере ведущих экономически развитых государств — локомотивов развития AI: США и КНР, а также РФ.

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ABorisov@ipiran.ru

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, AVBosov@ipiran.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, DZhukov@ipiran.ru

Будут представлены стратегические решения властей указанных государств по AI R&D, а также сравнение текущих результатов в этой области по комплексу стандартных наукометрических показателей. Данная статья является первой частью цикла и имеет следующую структуру. В разд. 2 обсуждается толкование термина «искусственный интеллект» и смежных с ним понятий, необходимых для дальнейшего изложения материала. Раздел 3 содержит этапность развития AI, начиная с 1940-х гг. по настоящее время. В разд. 4 представлен очерк государственного управления AI R&D в СССР и Российской Федерации. Предварительный итог подведен в заключении.

2 Феномен «искусственный интеллект» и смежные понятия

Для начала изложения основ важно понимать, что «*искусственный интеллект*» — это изначально название научного направления, которое существует с конца 1950-х гг. Это ни в коем случае не название некоторого устройства, или программы, или программно-технической системы. Целью этого направления ставится создание интеллектуальных систем, т. е. систем, способных к обучению, разумным рассуждениям и целенаправленному поведению. Такие системы сейчас существуют и решают много полезных задач. Характерным образцом современной ошибочной терминологии в области AI является, например, отождествление искусственного интеллекта с нейронными сетями — нейронные сети можно рассматривать лишь как одну из технологий AI. Мифологический оттенок носят дискуссии на тему «Может ли искусственный интеллект заменить человека?» — научное направление в принципе никого заменить не может, и уж тем более не может нести цивилизационную угрозу, о чем тоже ведутся сомнительные дискуссии.

Исторически тематику AI следует отсчитывать с 1944 г., когда в США был создан первый программируемый компьютер Mark-1, использовавшийся для сложных баллистических расчетов. А уже в 1956 г. на Конференции в колледже Дартмута (Нью-Гэмпшир) выдающиеся специалисты в области информатики Дж. МакКартни, М. Минский, Н. Рочестер и К. Шеннон [5] спрогнозировали появление новых вычислительных систем, которые впоследствии были обозначены термином «искусственный интеллект». Они представили *мир «с машинами, использующими человеческий язык, формирующими обобщения и умозаключения, решающие задачи, ранее предназначенные только для человека, и способные к самосовершенствованию»*. Эта историческая встреча заложила основу для десятилетий правительственных и отраслевых исследований в области AI, приведших к прогрессу в построении моделей восприятия, автоматизированных рассуждений/планирования, созданию когнитивных систем и систем машинного обучения, средств обработки естественного языка, робототехники и смежных областей. Сегодня эти достижения привели к появлению новых секторов экономики, влияющих на повседневную жизнь: от геоинформационных технологий до смартфонов, управляемых голосом, от распознавания рукописного текста до финансовой торговли, интеллектуальной логистики, фильтрации спама и многого

другого. Деятельность в последние 25 лет по большей части была направлена на адаптацию AI к нуждам статистических и вероятностных методов, обеспечение доступности больших массивов информации и развитие вычислительных мощностей. За последнее десятилетие особенно быстрый рост демонстрирует машинное обучение — методология AI, позволяющая компьютерам обучаться на опытах или примерах. В то время как основное внимание уделялось применению статистических подходов типа глубокого обучения, серьезные достижения были получены и в других областях, таких как построение умозаключений, обработка естественных языков, формальная логика, представление знаний, робототехника, теория управления, архитектура когнитивных систем, технологии поиска и оптимизации и др.

Вот определение AI, приведенное в электронной версии энциклопедии Britannica [6]: *искусственный интеллект* — способность цифрового компьютера или управляемого компьютером робота решать задачи, обычно ассоциирующиеся с разумными существами. Этот термин часто применяют к проектируемым системам, наделенным интеллектуальными возможностями, характерными для разумных существ, такими как способность к рассуждениям, пониманию (раскрытию) смысла, обобщению и обучению на прошлом опыте.

Теперь приведем определение AI из ключевого российского документа [7]. *Искусственный интеллект* — комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе то, в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

Как видно, это два принципиально разных подхода к исследованию, развитию и внедрению AI: фундаментальный и прикладной. Последний достаточно утилитарно представляет AI как набор готовых «блоков» («технологических решений»), поэтому важно представить определение технологий AI.

Нейротехнологии [7] — технологии, которые используют или помогают понять работу мозга, мыслительные процессы, высшую нервную деятельность, в том числе технологии по усилению, улучшению работы мозга и психической деятельности. Далее даны определения основных технологий из [7].

Компьютерное зрение — система решений, которые находят, отслеживают и классифицируют объекты.

Обработка естественного языка — система решений, направленных на понимание языка и генерацию грамотного текста, а также создание более удобной формы взаимодействия компьютера и человека.

Распознавание и синтез речи — система решений, позволяющих осуществлять перевод речевого запроса в текстовый вид, в том числе анализ тембра и тональности голоса, распознавание эмоций.

Рекомендательные системы и интеллектуальные системы поддержки принятия решений — система решений, посредством которых процесс выполняется без участия человека, поддержка в выборе решения, а также предсказание объектов, которые будут интересны пользователю по информации его профиля.

Перспективные методы и технологии в AI — методы и технологии, развитие которых влияет на все текущие технологии, а также на создание новых технологий в области AI.

Кроме того, в [7] к технологиям (точный термин — субтехнологиям) отнесены *нейропротезирование, нейроинтерфейсы, нейростимуляция и нейросенсинг*, что отражает субъективное представление прикладной идеологии AI.

3 Этапы мирового развития искусственного интеллекта

С самого начала исследований в области AI можно выделить три технологические волны [8]. Исследования *первой волны* (1980-е гг.) были четко сфокусированы на рукотворном знании и заключались в создании базирующихся на правилах экспертных систем в хорошо изученных областях. Знания в этих системах были собраны от экспертов и выражены в виде правил-импликаций, а затем реализованы в программно-аппаратной форме. Такие ориентированные системы рассуждений/умозаключений были успешно применены к узко определенным группам проблем, но у них не было возможности учиться или справляться с неопределенностью. Тем не менее они привели к важным решениям и разработке методов, которые все еще активно используются.

Вторая волна исследований AI — с 2000-х гг. по настоящее время — характеризуется развитием машинного обучения. Доступность значительно бóльших объемов цифровых данных, относительно недорогие массивно-параллельные вычислительные средства и улучшенные методы обучения привели к значительным достижениям применительно к таким задачам, как распознавание изображений и письма, понимание речи и перевод с человеческого языка. Ключом к ряду из этих успехов было развитие глубокого обучения.

Несмотря на прогресс в развитии, системы AI все еще имеют некоторые ограничения. Успех сопутствует AI при решении узкоспециализированных задач. В «общем AI» достигнут лишь некоторый прогресс, связанный с когнитивными технологиями. Системы AI распознавания изображений базируются на значительных человеческих усилиях, связанных с разметкой многотысячных обучающих банков, в то время как большинство людей способны «за один раз» учиться сразу на нескольких примерах. Понимание сцены, легкое для человека, все еще трудно для машины.

Далее, считается, что область AI сейчас находится в начальной стадии *третьей волны*, которая фокусируется на объяснительных и общих технологиях. Цели этих подходов состоят в том, чтобы улучшить изученные модели с помощью интерфейса объяснения и исправления, прояснить основу и надежность выходных данных, работать с высокой степенью прозрачности и выйти за пределы узкого AI

до возможностей, которые можно обобщать для более широкой задачи. В случае успеха инженеры смогут создавать системы, генерирующие объяснительные модели для классов явлений реального мира и вступающие в естественное общение с людьми. Предполагается, что они будут учиться и рассуждать, сталкиваясь с новыми задачами и ситуациями, и решать новые проблемы, обобщая прошлый опыт. Пояснительные модели для этих систем AI могут быть построены автоматически с помощью передовых методов. Эти модели могут обеспечить быстрое обучение в системах AI. Они могут придавать «смысл» или «понимание» системе AI, что может затем позволить системам AI достичь более широких возможностей.

4 Государственное участие в AI R&D в СССР и Российской Федерации: краткий экскурс

Развитие AI R&D в СССР и РФ, как и во всем мире, неразрывно связано с развитием вычислительной техники. В таблице в хронологическом порядке приводятся даты принятия органами государственной власти соответствующих директивных документов и события, связанные с развитием вычислительной техники в СССР и РФ [9–11].

Согласно [12], в настоящий момент времени в РФ сложились и функционируют следующие научные школы в области теоретического AI:

- Ю. Д. Апресяна — модель «Смысл → Текст»;
- С. Н. Васильева — логические методы в теории управления;
- С. Ю. Желтова — обработка информации в сложных системах управления;
- Ю. И. Журавлёва — теория распознавания образов;
- В. А. Лекторского — гуманитарные аспекты искусственного интеллекта;
- В. Л. Макарова — моделирование экономических процессов;
- Д. А. Поспелова — ситуационное управление, экспертные системы, нечеткие системы, моделирование рассуждений;
- К. В. Рудакова — анализ данных;
- К. В. Анохина — исследование мозга;
- Б. М. Величковского — исследование сознания и внимания в психологии;
- Н. В. Вапника — теория машинного обучения;
- Г. А. Золотовой — модель коммутативной грамматики в лингвистике;
- Г. С. Осипова — интеллектуальные динамические системы, анализ естественного языка;
- В. К. Финна — правдоподобный вывод, ДСМ-метод (названный так в честь Джона Стюарта Миля).

Основные вехи развития вычислительной техники в СССР и РФ

Дата	Событие
1950	Создание первого в СССР программируемого цифрового компьютера МЭСМ С. Лебедевым в Институте электротехники АН УССР (Киев)
1954	Создание ВЦ-1 МО СССР
Середина 1950-х	Создание ЭВМ «Минск» и «Урал» (Министерство радиопромышленности СССР). Использование для научных, космических и военных расчетов
1956	Создание ЭВМ серии «Стрела» (СКБ-245 (НИЭМ, НИЦЭВТ) Министерства радиопромышленности СССР), использование его для расчетов в МО СССР, ВЦ АН СССР, министерстве среднего машиностроения. Главный конструктор — Ю. Я. Базилевский
1958	Создание ЭВМ «Сетунь» с троичной логикой (МГУ, создатели — С. Соболев, Н. Брусенцов)
1959	Постановление Совета Министров СССР о создании при Госплане СССР Вычислительного центра для обеспечения компьютерных плановых расчетов
1959	Начало работ над проектом Общегосударственной автоматизированной системы учета и обработки информации ОГАС (авторы — А. Китов, В. Глушков)
1962	Строящийся город-спутник Москвы Зеленоград переориентирован на проектирование и производство электроники и микроэлектроники
1963	Постановление ЦК КПСС и СМ СССР от 21 мая 1963 г. № 564 «Об улучшении руководства внедрением вычислительной техники и автоматизированных систем управления в народное хозяйство»
1963	Создание ЦЭМИ АН СССР
1965	Создание Министерства электронной промышленности СССР
1965	Создание в Зеленограде специализированного вуза МИЭТ (Постановление Совета Министров СССР от 26 ноября 1965 г. № 1006)
1965–1969	Создание серии миникомпьютеров второго поколения (полупроводниковых) «Мир» (Киевский институт кибернетики, главный конструктор — В. Глушков)
1967	Постановлением ЦК КПСС и Совета Министров СССР от 30 декабря 1967 г. № 1180/420 определен комплекс мер, направленных на разработку, производство и эффективное использование средств вычислительной техники в народном хозяйстве и оборонном комплексе страны. Разделение полномочий по выпуску ЭВМ между Министерством радиоэлектронной промышленности (универсальные и специальные ЭВМ) и Министерством приборостроения (управляющие ЭВМ). Преобразование НИЭМ в НИЦЭВТ
1967	Создание космического корабля «Союз 7К-Л1» с первым бортовым компьютером «Аргон-11С»

Продолжение таблицы на с. 63

Основные вехи развития вычислительной техники в СССР и РФ (*продолжение*)

Дата	Событие
1967	Создание БЭСМ (ИТМиВТ, главный конструктор — С. А. Лебедев)
1969	Создание Института проблем управления АН СССР
1970	Создание факультета вычислительной математики и кибернетики МГУ имени М. В. Ломоносова
1971	Начало выпуска ЭВМ серии ЕС (аналог IBM 360/370)
1972	В АН СССР создан Научный совет по проблеме искусственного интеллекта (ИВТА АН СССР, Г. С. Поспелов)
1973	Начало выпуска ЭВМ серии «Эльбрус»
1974	Создание НПО «Центр программных систем» (подчинено Министерству приборостроения) как единого центра дистрибуции программного обеспечения
1971–1975	В план 9-й пятилетки заложено увеличение числа установленных компьютеров в СССР до 25 000 (в 2,6 раза), а также работы по реализации урезанного проекта ОГАС — Единой государственной сети вычислительных центров (ЕГСВЦ)
1982	Начало запуска спутников системы ГЛОНАСС
1983	Создание ИПИ АН СССР и ИПК АН СССР
1983	В АН СССР создано Отделение информатики, вычислительной техники и автоматизации.
1984	В СССР около 300 000 программистов
1985	Покупка 10 000 ЭВМ Nippon Gakki
1985	Постановление ЦК КПСС и Совета Министров СССР от 28 марта 1985 г. № 271 «О мерах по обеспечению компьютерной грамотности учащихся средних учебных заведений и широкого внедрения электронно-вычислительной техники в учебный процесс»
1985	Создание ИПС АН СССР
1986	Постановление ЦК КПСС и Совета Министров СССР от 23 января 1986 г. «О создании и развитии производства в СССР персональных ЭВМ»
1986	Создание ИАП АН СССР и Научного центра по фундаментальным проблемам вычислительной техники и систем управления
1986–1988	Поставка в школы около 90 000 персональных компьютеров
1988	Введена в строй первая в СССР ВОЛС «Ленинград – Сосновый Бор»
1989	Создание НИИСИ
1989	В СССР около 200 000 ЭВМ
1990	Создание сети RELCOM, создание домена .su
1990	Создание ИВМ АН СССР и ИММ АН СССР
1990	Создание сети RELCOM компании «Demos»
1990	Постановление Совета Министров СССР от 13 января 1990 г. № 49 «Об образовании межотраслевого государственного объединения по разработке, производству и обслуживанию персональных ЭВМ»
1992	Создание ИСА
1993	Создание ИСОИ РАН

Окончание таблицы на с. 64

Основные вехи развития вычислительной техники в СССР и РФ (*окончание*)

Дата	Событие
1994	Создание ИВВС РАН и ИСП РАН
2001	В Межведомственном суперкомпьютерном центре установлен отечественный суперкомпьютер МВС 1000М
2002	Письмо Президента РФ от 30 марта 2002 г. № Пр-576 «Основы политики Российской Федерации в области развития науки и технологий на период до 2010 года и дальнейшую перспективу»
2006	Утверждена «Стратегия развития науки и инноваций в Российской Федерации на период до 2015 года»
2011	Распоряжение Правительства РФ от 8 декабря 2011 г. № 2227-р «Об утверждении Стратегии инновационного развития Российской Федерации на период до 2020 года»
2013–2018	Федеральный закон от 27.09.2013 № 253-ФЗ «О Российской академии наук, реорганизации государственных академий наук и внесении изменений в отдельные законодательные акты Российской Федерации», реформа РАН
2016	Указом Президента РФ от 1 декабря 2016 г. № 642 утверждена «Стратегия научно-технического развития России», одним из направлений содержащая «Переход к цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создание систем обработки больших данных, машинного обучения и искусственного интеллекта»
2018	Создание Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации (Минцифры России) в Правительстве РФ (переименование Минкомсвязи)
2018	Утверждение национальной программы «Цифровая экономика Российской Федерации»
2019	Указ Президента РФ от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»)

Согласно той же ссылке, среди значимых российских достижений можно отметить создание методов:

– машинного обучения при решении задач обработки изображений и распознавания образов;

– автономной координации и управления в коалициях интеллектуальных агентов;

– коллективного взаимодействия роботов при решении групповых задач;

– когнитивных компьютерных моделей с пониманием естественного языка, систем поддержки научных исследований;

– искусственного интеллекта для обеспечения информационной безопасности;

– автоматизации рассуждений;

– планирования и управления поведением в сложных непрогнозируемых средах.

5 Заключение

Широко используемый в настоящее время термин «искусственный интеллект» понимается исследователями и практиками по-разному. Современная трактовка его как комплекса технологических решений представляется не вполне удачной, так как сужает для него горизонт фундаментального научного развития современным уровнем понимания и достижений. Словосочетание «технологическое решение» в самом себе несет ограничение: имеющееся «решение» не может породить новую задачу. Поэтому в следующих частях предполагается использовать этот термин в смысле, близком к [6].

Приведенная хронология проектирования и создания средств вычислительной техники как аппаратной базы AI в СССР и РФ, а также научных направлений в этой области позволяет сделать следующие выводы.

Во-первых, в период 1950–1990-х гг. советское государство уделяло много внимания и средств развитию вычислительной техники и программным системам — основе AI. Несмотря на сложившееся отставание, предпринимались значительные и разнообразные усилия по его исключению: создавались оригинальная собственная элементная база, аппаратные средства и программное обеспечение, несмотря на жесткие санкционные ограничения, имелся положительный опыт воспроизведения и модернизации передовых зарубежных систем. Параллельно с этим была разработана и реализована программа обучения собственных специалистов в этой области. Притом что передовые образцы отечественных вычислительных средств и программного, в особенности алгоритмического, обеспечения были на достаточно современном мировом уровне, общий показатель надежности и оснащенности был несравнимо ниже, чем у ведущих держав. Отставание «в среднем», наличие бюрократических препон внутри, санкционных запретов извне и разрушение в 1991 г. существовавшей на тот момент экономической системы в СССР предопределили серьезнейшее отставание Российской Федерации в части аппаратно-программной платформы AI.

Во-вторых, после периода стагнации конца 1990-х – начала 2000-х гг. государство вернулось к стратегическому управлению наукой в связи с провозглашенным курсом на научно-техническую модернизацию РФ. Принятием «Национальной стратегии развития искусственного интеллекта на период до 2030 года» Россия сделала тот же шаг, что и остальные государства, стремящиеся получить или сохранить статус экономически и промышленно развитых держав. Следует отметить, что к таким странам относятся не только США и КНР, но и страны ЕС, Индия, Корея, также недавно принявшие аналогичные стратегические документы.

В-третьих, стратегия [7] опосредованно предполагает не только внедрение «технологических решений», но и проведение полноценных фундаментальных научных исследований в области AI. В данной части цикла было упомянуто, что в РФ имеется научное сообщество, занимающееся AI R&D. Сравнительному анализу наукометрических показателей научных публикаций российских ученых в этой области будет посвящена следующая статья цикла.

Литература

1. *Bender E. A.* Mathematical methods in artificial intelligence. — Washington, D.C., USA: IEEE Computer Society Press, 1996. 656 p.
2. *Nilsson N. J.* The quest for artificial intelligence. — Cambridge: Cambridge University Press, 2010. 707 p.
3. *Haenlein M., Kaplan A. A.* Brief history of artificial intelligence: On the past, present, and future of artificial intelligence // Calif. Manage. Rev., 2019. Vol. 61. No. 4. P. 5–14.
4. *Russell S., Norvig P.* Artificial intelligence: A modern approach. — New York, NY, USA: Pearson, 2020. 1136 p.
5. *McCarthy J., Minsky M. L., Rochester N., Shannon C. E.* A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955 // AI Mag., 2006. Vol. 27. No. 4. P. 12–14. doi: 10.1609/aimag.v27i4.1904.
6. *Copeland B. J.* Artificial intelligence // Britannica, 2020. https://www.britannica.com/technology/artificial-intelligence.
7. О развитии искусственного интеллекта в Российской Федерации: Указ Президента РФ от 10 октября 2019 г. № 490. http://static.kremlin.ru/media/events/?les/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf.
8. The National Artificial Intelligence Research and Development Strategic Plan. — Washington, D.C., USA: National Science and Technology Council, Networking and Information Research and Development Subcommittee, 2016. https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf.
9. *Golubev K. M.* Overview of AI research history in USSR and Ukraine: Up-to-date just-in-time knowledge concept // Artificial intelligence for knowledge management / Eds. E. Mercier-Laurent, D. Boulanger. — IFIP advances in information and communication technology ser. — Springer, 2014. Vol. 422. P. 1–18.
10. Библиотека нормативно-правовых актов Союза Советских Социалистических Республик. http://www.libussr.ru.
11. Виртуальный компьютерный музей: календарь событий. https://www.computer-museum.ru/calendar.
12. *Соколов И. А.* Теория и практика применения методов искусственного интеллекта // Вестник РАН, 2019. Т. 89. № 4. С. 365–370.

Поступила в редакцию 30.12.20

RESEARCH AND DEVELOPMENT STRATEGY IN THE FIELD OF ARTIFICIAL INTELLIGENCE I: BASIC CONCEPTS AND BRIEF CHRONOLOGY

A. V. Borisov, A. V. Bosov, and D. V. Zhukov

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

Abstract: The paper begins a series of works presenting the results of the study of the impact of public administration on the effectiveness in the field of artificial intelligence research and development (AI R&D), which has become a strategically important industry in any technologically developed country. This first part is devoted to a discussion of different interpretations of the term "artificial intelligence" and related concepts. The modern division of the world history of AI R&D into eras is presented. Since the practical results in this area are integral to the development of the hardware platform, a historical retrospective presents a parallel chronology of domestic achievements in the field of creating computer technology and decisions of state bodies that stimulate the development of this area.

Keywords: artificial intelligence; related concepts and technologies of artificial intelligence; time chronology

DOI: 10.14357/08696527210105

References

1. Bender, E. A. 1996. *Mathematical methods in artificial intelligence*. Washington, D.C.: IEEE Computer Society Press. 656 p.
2. Nilsson, N. J. 2010. *The quest for artificial intelligence*. Cambridge: Cambridge University Press. 707 p.
3. Haenlein, M., and A. A. Kaplan. 2019. Brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *Calif. Manage. Rev.* 61(4):5–14.
4. Russell, S., and P. Norvig. 2020. *Artificial intelligence: A modern approach*. New York, NY: Pearson. 1136 p.
5. McCarthy, J., M. L. Minsky, N. Rochester, and C. E. Shannon. 2006. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag.* 27(4):12–14. doi: 10.1609/aimag.v27i4.1904.
6. Copeland, B. J. Artificial intelligence. Available at: https://www.britannica.com/technology/artificial-intelligence (accessed February 8, 2021).
7. O razvitii iskusstvennogo intellekta v Rossiyskoy Federatsii: Ukaz Prezidenta ot 10.10.2019 No. 490 [About strategy of scientific and technological development of the Russian Federation. Presidential Decree No. 490 dated 10.10.2019]. Available at: http://static.kremlin.ru/media/events/?les/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf (accessed February 8, 2021).

8. The National Artificial Intelligence Research and Development Strategic Plan. 2016. Washington, D.C.: National Science and Technology Council, Networking and Information Research and Development Subcommittee. Available at: https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf (accessed February 8, 2021).
9. Golubev, K. M. 2014. Overview of AI research history in USSR and Ukraine: Up-to-date just-in-time knowledge concept. *Artificial intelligence for knowledge management*. Eds. E. Mercier-Laurent and D. Boulanger. IFIP advances in information and communication technology ser. Springer. 422:1–18.
10. Biblioteka normativno-pravovykh aktov Soyuza Sovetskikh Sotsialisticheskikh Respublik [Library of normative legal acts of the Union of Soviet Socialist Republics]. Available at: http://www.libussr.ru (accessed February 8, 2021).
11. Virtual'nyy komp'yuternyy muzey: kalendar' sobytiy [Virtual Computer Museum: Events Calendar]. Available at: https://www.computer-museum.ru/calendar/ (accessed February 8, 2021)
12. Sokolov, I. A. 2019. Theory and practice of application of artificial intelligence methods. *Her. Russ. Acad. Sci.* 89:115–119.

Received December 30, 2020

Contributors

Borisov Andrey V. (b. 1965) — Doctor of Science in physics and mathematics, 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; ABorosov@ipiran.ru

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

Zhukov Denis V. (b. 1979) — principal specialist, 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; DZhukov@ipiran.ru

ПОДДЕРЖКА РЕШЕНИЯ ЗАДАЧ ДИАГНОСТИЧЕСКОГО ТИПА*

М. И. Забежайло[1], *А. А. Грушо*[2], *Н. А. Грушо*[3], *Е. Е. Тимонина*[4]

Аннотация: Обсуждаются некоторые значимые особенности математических методов анализа данных (АД) и поддержки принятия решений (ППР) в задачах диагностического типа. Рассмотрены наиболее существенные характеристические особенности, позволяющие выделить задачи обсуждаемого типа в особый класс. Этот класс требует одновременной разработки решений ряда взаимосвязанных задач, которые без учета таких взаимосвязей практически бесполезны. Опыт работы с задачами диагностического типа позволил сформулировать рекомендации по направлениям разработки результативных подходов и методов интеллектуального анализа данных (ИАД) для решения таких прикладных задач.

Ключевые слова: искусственный интеллект; интеллектуальный анализ данных; математические методы; диагностика

DOI: 10.14357/08696527210106

1 Введение

Цель данной работы — аргументировать потребность в развитии и использовании проблемно-ориентированного математического инструментария АД и ППР в задачах диагностического типа. Требуется, чтобы такой инструментарий позволял:

- реализовать причинный анализ;
- решать задачи в условиях динамически пополняемых эмпирических данных;
- строить представления знаний;
- формализовать рассуждения, включая подходы и методы оценивания приемлемости (достаточности) основания для принятия получаемых заключений;
- обеспечивать эффективную разрешимость соответствующих математических задач;

*Работа частично поддержана РФФИ (проект 18-29-03081).

[1]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, m.zabezhailo@yandex.ru

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, grusho@yandex.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, info@itake.ru

[4]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, eltimon@yandex.ru

– обеспечивать прозрачную интеграцию (интерактивный режим взаимодействия эксперта и компьютерной системы АД и ППР в процессе ИАД).

Надежная реализация представленных требований даст основание доверять полученным в процессе результатам ИАД.

2 Неформальное представление проблемы

Среди современных приложений математических моделей, алгоритмов и программных систем компьютерного анализа данных все более значимую роль играют задачи диагностического характера. Спектр востребованных практикой областей применения подобного «инструментария» достаточно широк.

Общее представление о задачах такого типа удобно описать, используя аналогию с врачебной практикой. Для разработки соответствующего «инструментария» АД и ППР изучаются действия врача, цель которого — идентификация и противодействие состоянию болезни (ПАТОЛОГИИ) у конкретного пациента. При этом в организации своих действий врач опирается на опыт (как позитивный, так и негативный) ранее выполненных противодействий «аналогичным» болезненным состояниям у других (ранее уже лечившихся у него) пациентов.

Мониторинг поведения объекта анализа предполагает в первую очередь аргументированную идентификацию состояния НОРМЫ или же, наоборот, отклонения от нее — аномалии / ПАТОЛОГИИ.

При последующей организации противодействия возможным аномалиям, а также их последствиям естественно сначала выделить факторы влияния, результатами которого становится идентифицированная аномалия. Затем организовать результативное противодействие вредоносному влиянию. Следует учитывать, что в некоторых случаях предполагаемые меры противодействия могут носить более радикальный характер. Так, при боевом применении «умного» оружия идентифицированный вредоносный объект (живая сила или же военная техника противника) просто переводится в статус цели для последующего уничтожения. Однако такая ситуация в данной статье не рассматривается.

Ключевым элементом обсуждаемой схемы оказывается проблема результативности организуемого противодействия. Предполагается, что, воздействуя именно на факторы влияния, можно гарантированно обеспечить желаемый результат. Именно на этом основании при разработке эффективных «инструментов» АД и ППР естественно ориентироваться на анализ *причинности* — выделение «*каузальных оснований*» анализируемых эффектов.

Таким образом, задачи диагностического характера — это задачи АД, ППР и организации выполнения целенаправленных действий, предполагающие:

(1) мониторинг текущего состояния объекта анализа;

(2) идентификацию анализируемого эффекта (аномалии в поведении объекта мониторинга — смены состояния НОРМА на состояние ПАТОЛОГИЯ);

(3) идентификацию факторов, влиянием которых обусловлено появление идентифицированной аномалии;

(4) организацию противодействия влиянию выявленных вредоносных факторов, обеспечивающего результативное сопротивление развитию ПАТОЛОГИИ и возвращение к состоянию НОРМА.

Приведем некоторые примеры задач обсуждаемого класса.

– медицинская диагностика, в том числе высокотехнологичная, где необходимо анализировать большие объемы эмпирических данных, фиксируемых объективными средствами, и последующие терапевтические воздействия;

– техническая диагностика и восстановление штатного режима работы: идентификация отказов техники, выявление причин отказа, организация действий по возврату к нормальному режиму функционирования;

– обеспечение информационной безопасности компьютерных систем, в частности идентификация целенаправленных вредоносных воздействий на компьютерные системы и организация адекватного противодействия их вредоносному влиянию;

– борьба с мошенничеством в финансовой сфере, т. е. выявление фактов мошенничества, подготовка и реализация мер по эффективному противодействию мошенническим активностям и др.

3 Возможные подходы к решению диагностических проблем

Обращаясь к известным подходам, позволяющим рассчитывать на результативность противодействия развитию идентифицируемых аномальных эффектов, естественно начать с опыта классической физики, а именно: формирования так называемых «*панфизических*» моделей изучаемых явлений. Здесь анализ причинности и последующее использование основанных на результатах такого анализа управляющих воздействий можно представить следующей процедурной схемой:

– фиксируется анализируемый целевой эффект;

– выделяются факторы влияния на этот эффект;

– выясняется наличие взаимосвязей между такими факторами;

– выделенные факторы и взаимосвязи между ними сводятся в подходящую систему «балансовых» соотношений (отражающих, например, те или иные физические законы сохранения);

– обеспечивается разрешимость (аналитическая или численная) сформированной системы соотношений;

– в явной форме получаются причинные зависимости вида

[**факторы влияния**] ⇒ **целевой эффект**.

К сожалению, прямой реализуемости подобного подхода в рассматриваемых предметных областях препятствует по крайней мере одно весьма существенное обстоятельство. Для классической физики характерны *малофакторные* (по числу факторов влияния) причинные обусловленности, например «простые» (*малопричинные*) зависимости определяют основные законы механики. В отличие от физики для рассматриваемых областей приложений диагностического типа характерна *многопричинность*, например в медицинской и технической диагностике. В этом случае критически значимой оказывается проблема полноты учета факторов влияния, а также связей между ними.

Не менее существенной оказывается и проблема практической (с привлечением «разумных» вычислительных ресурсов) разрешимости формируемых математических моделей «балансового» типа.

Реально работающей альтернативой «*панфизическим*» подходам стали *интерполяционно-экстраполяционные* модели АД и ППД, в рамках которых сначала на уже имеющихся эмпирических данных формируются интерполяционные зависимости того или иного вида: статистические, алгебраические, логические и т. д. Затем задача принятия решения в новой ситуации формализуется как проверка экстраполируемости найденных зависимостей на вновь анализируемый случай. Обширный перечень реализаций подобного подхода можно найти в статистическом анализе данных, машинном обучении и др.

Однако в общем случае и здесь приходится иметь дело с серьезной проблемой. Интерполяционные решения, вообще говоря, не обязательно отражают причинность возникновения идентифицируемого целевого эффекта. Не менее существенной оказывается наследуемость (сохраняемость) уже порожденных эмпирических зависимостей при расширении исходных данных описаниями новых прецедентов, что, вообще говоря, непросто обеспечить при построении стандартных интерполяционных моделей. Как следствие, можно говорить о необходимости сужения множества математических «инструментов» АД, релевантных уже представленной выше «природе» задач диагностического характера таким образом, чтобы обеспечить:

– анализ *причинности* возникновения исследуемого целевого эффекта;

– сохраняемость в *динамике* эмпирических закономерностей при расширениях обучающей выборки;

– *достаточность оснований* для принятия формируемых результатов АД (выявления причинных факторов);

– *неопровергаемость* (*неоспариваемость*) полученных результатов по крайней мере на имеющихся для анализа эмпирических данных.

4 Методы интеллектуального анализа данных в решении задач диагностического типа

Понятно, что основополагающее для computer science моделирование явлений окружающего мира компьютерными средствами реализует представление *семантики*, т. е. смысловых особенностей исследуемых объектов *синтаксическими* средствами, например кодированием выражениями в алфавите {0, 1} и их последующими проблемно-ориентированными преобразованиями. В общем случае следствием этого становится *приближенный* характер подобных представлений, а также их *множественность*, которая определяется выбором того или иного конкретного языка описания (представления) знаний об исследуемых эффектах и явлениях.

При построении компьютерно-ориентированных формализаций однозначное соответствие между синтаксисом и семантикой требует выполнимости специальных условий типа теорем о полноте, хорошо известных в математической логике (см., например, [1, 2]). Так, в частности, дедуктивное доказательство в формальной системе, для которой доказана теорема о полноте, — это объект, синтаксическая корректность которого гарантирует его семантическую корректность, т. е. выводимость формулы в такой формальной системе гарантирует ее истинность.

Таким образом, выбирая тот или иной язык представления знаний (ЯПЗ) для использования в схеме эмпирических закономерностей (ЭЗ) компьютерного АД, следует учитывать, что соответствующие объемы (сложность) вычислений при поиске решений могут стать своего рода «платой» за детальность представления семантики конкретным ЯПЗ.

Как следствие, важным и, в значительной мере, определяющим качество формируемых результатов компонентом рассматриваемого процесса порождения ЭЗ оказывается выбор того или иного конкретного способа описания (формализованного представления) прецедентов исходной обучающей выборки.

Для примера рассмотрим задачу медицинской диагностики. «Стандартные» для текущей медицинской практики характеристики физиологических состояний пациента могут быть дополнены, например, результатами исследования:

- метаболома (данными о метаболитах, задействованных в патологических биохимических процессах) плазмы крови пациента;
- протеома (данными о белковом составе анализируемых проб);
- генома, в том числе данными о структуре мутаций определенных групп генов и др.

Таким образом, варьируя «выразительные» возможности используемого способа описания данных (знаний) о релевантных цели (формированию соответствующего диагноза) явлениях (эффектах), можно выявлять те или иные совокупности действующих факторов, характеризующих причинность возникновения изучаемых эффектов. И далее в той или иной мере управлять этими эффектами.

При этом естественно искать такие ЭЗ, которые сохраняли бы свои «экстраполяционные» возможности и при варьировании используемого языка описания прецедентов. Другими словами, не только востребованной, но и результативной при решении трудных прикладных задач оказывается традиционная для исследований в области искусственного интеллекта проблематика методов и средств представления знаний (см., например, [3]). Собственно, именно такой смысл естественно вкладывается в часто используемый термин «управление знаниями» (см., например, [4, 5]).

5 Представление знаний и наследуемость результатов диагностики при варьировании исходных данных

Критически значимую роль при поиске решений задач рассматриваемого типа играет адекватность представления знаний о природе причинности возникновения изучаемых эффектов (явлений). Особого внимания здесь заслуживают эффекты устойчивости (наследуемости) формируемых результатов ИАД при варьировании используемых средств представления знаний (ЯПЗ о природе исследуемой причинности).

С «процедурной» точки зрения в приоритетном порядке следует обратить внимание на два аспекта:

(1) необходимо разрабатывать приближенные представления о природе причинности в конкретном ЯПЗ, т. е. необходимо решать проблему представления семантических «нюансов» синтаксическими «средствами». Или находить надежный ответ на вполне естественный вопрос: «Насколько конкретный ЯПЗ адекватен в конкретном случае ИАД?» Ответ на этот вопрос требует оценки адекватности (приемлемости) порождаемых результатов ИАД, в частности идентификации *границ* «области надежности» и приемлемости таких результатов;

(2) с учетом варьирования как используемых средств представления знаний, так и «инструментов» АД и формирования новых знаний об объекте исследования иметь возможность проследить подобласти «наследуемости» диагностических заключений «вдоль» областей варьирования, выбираемых ЯПЗ и, соответственно, инструментальных средств ИАД. Сравнимость формируемых таким образом диагностических заключений требует как сопоставимости средств представления данных и знаний, так и использования сопоставимых процедурных средств ИАД (в первую очередь в части реализованной в них единой «логики» АД).

Дополнительно можно заметить, что при расширении обучающей выборки прецедентов новыми объектами традиционные интерполяционно-экстраполяционные процедурные схемы ИАД в общем случае не гарантируют устойчивости (наследуемости) интерполяций, формируемых их средствами. Разумеется, при варьировании привлекаемых средств представления знаний наследуемость

«вдоль» области варьирования соответствующих интерполяционных ЭЗ становится еще более проблематичной.

6 Формализация рассуждений

Интеллектуальный анализ данных как процесс перехода от исходных данных к конечному результату позволяет выделить наряду с собственно «вычислительной» составляющей еще два принципиально значимых этапа «познавательного» цикла:

(1) отбор исходных данных, релевантных цели выполняемого ИАД;

(2) оценку «доверия» результатам, полученным в ходе «вычислений».

Эксперт, решая трудные задачи рассматриваемого типа, в случае возникающих каких-либо сомнений, способен:

– возвратиться к началу представленного «познавательного» цикла;

– изменить, если это потребуется, состав исходных данных и/или набор «инструментов» их компьютерного анализа с целью добиться порождения приемлемого, характеризуемого достаточными основаниями для его принятия результата либо аргументированным образом отказаться от последующих попыток его получения.

Этапы процедурной схемы ИАД

Релевантные «посылки» (входные данные для ИАД)	$\Leftrightarrow$	**Способ вычислений.** Его корректность (отсутствие ошибочных действий)	$\langle = \rangle$	**Оценка приемлемости результата** (достаточности оснований для принятия результата)

Таким образом, объединив все три этапа (см. таблицу) рассматриваемой процедурной схемы ИАД в единый комплекс, ориентированный на перенос его далее в компьютерную систему АД, можно говорить о формализации рассуждений (см., например, [6]). Следовательно, значимым является не только способ «вычислений», но и аргументируемая приемлемость всей конструкции (цепочки рассуждений), при этом оценка приемлемости результатов ИАД становится общей составляющей для представления знаний и формализации рассуждений.

7 Эффективная разрешимость математических задач

Опыт попыток преодоления проблем надежной разрешимости систем балансовых соотношений в «*панфизических*» (см. разд. 3) моделях учит уделять особое

внимание не только потенциальной, но и эффективной разрешимости возникающих в рамках ИАД математических задач. Критически значимым аспектом является эффективная вычислимость при порождении результатов ИАД, т. е. возможность формировать диагностические заключения полиномиально-сложными вычислениями. Фактически речь ведется о прямых аналогиях с проблемой аналитической, численной и т. д. разрешимости «*панфизических*» моделей при анализе причинности с учетом особенностей схем АД и ППР.

Примерами весьма чувствительных барьеров, с которыми приходится сталкиваться в практически значимых приложениях, могут служить:

- приближенный, не обеспечивающий однозначного отображения всех аспектов семантики анализируемого эффекта (явления) характер формализованного описания [7];
- идентификация в процессе формализации процесса АД и ППР доказуемо трудноразрешимых комбинаторных задач;
- необходимость в целом ряде случаев «укладываться» в жесткие ресурсные ограничения, так называемое процессно-реальное время;
- учитывать производительность доступных вычислительных установок и др.

Для преодоления названных ограничений существенным является:

- уже накопленный эмпирический опыт, например: проблемно-ориентированные подходы и методы в части идентификации быстроразрешимых подклассов доказуемо трудноразрешимых переборных задач, приближенные методы оптимизации и управления перебором вариантов при поиске решений, использование специализированных аппаратно-программных платформ и т. п.;
- «гибкость мысли» исследователей, позволяющая порождать новые эффективные эвристики, при возможности выделяя далее области доказуемой надежности (корректности) их применения.

8 Интерактивный режим взаимодействия эксперта и компьютерной системы

В работе У. Р. Эшби [8] обращено внимание на продуктивность интерактивного режима взаимодействия человека и компьютерной системы искусственного интеллекта (ИИ). Роль интеллектуальных компьютерных систем — *ассистентов* специалиста, в том числе «*акселераторов*» его познавательных возможностей, в различных областях знаний и технологий хорошо известна в настоящее время. Однако опыт использования компьютерных систем АД и ППР в критически значимых областях обозначил принципиально важный аспект — *ответственность* за конечный результат и последствия принятых решений, в том числе и соответствующих управляющих воздействий. Оказалось, что определяющую роль здесь в организации взаимодействия человека и компьютерной системы

ИАД играет понимание экспертом того, как именно система ИАД сформировала соответствующее заключение, почему именно эксперт должен принять на себя ответственность за предлагаемое решение. Стало понятно, что необходимым условием решения проблем в таком «тандеме» человека и системы ИИ становятся средства согласования способа рассуждения эксперта, используемого им при АД, со «способом» (формализованной моделью) рассуждений, реализованным в компьютерной системе — «акселераторе» его ИАД-возможностей. Естественным выводом здесь оказалось вполне очевидное требование к «архитектуре» взаимодействия эксперта и интеллектуальной компьютерной системы и, как следствие, к архитектуре собственно систем ИАД, реализующих «бесшовную» интеграцию «способа» рассуждений эксперта и компьютерной системы.

Итак, в задачах рассматриваемого класса, предполагающих ответственность за управляющие воздействия, доверие к порождаемым результатам определяется возможностями математического инструментария АД с учетом следующих аспектов:

- **представления знаний** о причинности изучаемых эффектов, в частности *интерпретируемости наследования* (*ненаследования*) результатов при варьировании используемых ЯПЗ;
- **причинности** в задаче, изучаемой средствами ИАД, позволяющей достигнуть интерпретируемости результатов и достаточной для принятия экспертом ответственности за предлагаемое решение;
- **динамики** изменения имеющихся эмпирических данных, т. е. работы с постоянно расширяемым множеством прецедентов в анализируемой выборке;
- **достаточности оснований для принятия результатов** ИАД, т. е. *неоспариваемости* полученных результатов на имеющихся эмпирических данных (отсутствие противоречий);
- **формализации рассуждений**, средствами которых ведется ИАД, включая подбор релевантных «посылок» (исходных данных) вместе с однородностью «расчетных» процедур для иерархий ЯПЗ и оценкой адекватности формируемых результатов;
- **эффективной разрешимости** возникающих в процессе ИАД комбинаторных (переборных) математических задач с учетом необходимости обеспечить процессно-реальное время при порождении результатов ИАД;
- **«бесшовной» интеграции** «инструментов» ИАД, используемых человеком и компьютерной системой.

9 Заключение

В работе сформулированы принципы построения математического аппарата, необходимого для формирования адекватных решений задач диагностического

типа. Речь идет о математических методах, эффективная разработка и использование которых позволит осуществить «прорыв» в решении задач диагностического типа.

На сегодняшний день можно уверенно указать на ряд результативных заделов, позволяющих рассчитывать на успешное решение поставленной в работе проблемы. Так, уже предложены:

- проблемно-ориентированные средства анализа причинности, позволяющие формировать бесспорные на имеющейся обучающей выборке заключения (см., например, [9, 10]);
- методы отслеживания устойчивости (наследуемости) таких ЭЗ при расширении исходной обучающей выборки описаниями новых прецедентов (см., например, [11, 12]);
- теоретический базис подхода к формированию машинно-ориентированных формализаций эмпирической индукции [6].

В частности, разработанный математический аппарат открывает возможности формировать «однородные» по «логике рассуждений» семейства инструментальных средств ИАД, которые способны оперировать различными представлениями данных об анализируемых явлениях (булевы векторы, нумерованные множества признаков, порядковые шкалы, графы и др.).

Тем не менее не вызывает сомнений, что значительная часть ответов на обозначенные в данной работе вопросы и вызовы еще ждет своего исследования и своих решений.

Литература

1. *Клини С.* Математическая логика / Пер. с англ. — М.: Мир, 1973. 480 с. (*Kleene S. C.* Mathematical logic. — New York, NY, USA: Wiley, 1967. 398 p.)
2. *Мендельсон Э.* Введение в математическую логику / Пер. с англ. — М.: Наука, 1976. 320 с. (*Mendelson E.* Introduction to mathematical logic. — Van Nostrand, 1964. 310 p.)
3. *Davis R., Shrobe H., Szolovits P.* What is a knowledge representation? // AI Mag., 1993. Vol. 14. Iss. 1. P. 17–33.
4. Handbook of knowledge representation / Eds. F. van Harmelen, V. A. Lifschitz, B. Porter. — Amsterdam: Elsevier, 2008. 1005 p. http://dai.fmph.uniba.sk/~sefranek/kri/handbook/handbook_of_kr.pdf.
5. Mathematical knowledge management. NIST Digital Library. https://www.nist.gov/mathematical-knowledge-management.
6. *Финн В. К.* Индуктивные методы Д. С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений, 2010. № 3. С. 3–21 (Ч. I); № 4. С. 14–40 (Ч. II).
7. *Грушо А. А., Забежайло М. И., Зацаринный А. А., Николаев А. В., Писковский В. О., Тимонина Е. Е.* Классификация ошибочных состояний в распределенных вычислительных системах и источники их возникновения // Системы и средства информатики, 2017. Т. 27. № 2. С. 29–40.

8. *Эшби У. Р.* Введение в кибернетику / Пер. с англ. — М.: ИЛ, 1959. 432 с. http://pcp.vub.ac.be/ASHBBOOK.html. (*Ashby W. R.* An introduction to cybernetics. — New York, NY, USA: Wiley, 1956. 316 p.)
9. *Грушо А. А., Забежайло М. И., Тимонина Е. Е.* О каузальной репрезентативности обучающих выборок прецедентов в задачах диагностического типа // Информатика и её применения, 2020. Т. 14. Вып. 1. С. 80–86.
10. *Грушо Н. А., Грушо А. А., Забежайло М. И., Тимонина Е. Е.* Методы нахождения причин сбоев в информационных технологиях с помощью метаданных // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 33–39.
11. *Забежайло М. И.* О некоторых оценках сложности вычислений при прогнозировании свойств новых объектов средствами характеристических функций // Научно-техническая информация. Сер. 2, 2020. № 12. С. 1–12.
12. *Забежайло М. И., Трунин Ю. Ю.* К проблеме надежности медицинского диагноза, формируемого на основе эмпирических данных // Искусственный интеллект и принятие решений, 2020. № 4. С. 3–13.

Поступила в редакцию 09.02.21

SUPPORT FOR SOLVING DIAGNOSTIC TYPE PROBLEMS

***M. I. Zabezhailo*[1], *A. A. Grusho*[2], *N. A. Grusho*[2], *and E. E. Timonina*[2]**

[1]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

[2]Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: Some significant features of mathematical methods of data analysis and decision support in diagnostic-type problems are discussed. The most significant characteristic features are considered allowing to distinguish the tasks of the discussed type into a special class. This class requires the simultaneous development of solutions to a number of interrelated problems which, without taking into account such relationships, are practically useless. Using the experience with diagnostic-type tasks, recommendations are made on the areas of development of effective approaches and methods of data mining for solving such applications.

Keywords: artificial intelligence; intelligent data analysis; mathematical methods; diagnostics

DOI: 10.14357/08696527210106

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-29-03081).

References

1. Kleene, S. C. 1967. *Mathematical logic*. New York, NY: Wiley. 398 p.
2. Mendelson, E. 2015. *Introduction to mathematical logic*. 6th ed. New York, NY: CRC Press. 499 p.
3. Davis, R., H. Shrobe, and P. Szolovits. 1993.What is a knowledge representation? *AI Mag.* 14(1):17–33.
4. Van Harmelen, F., V. A. Lifschitz, and B. Porter, eds. 2008. *Handbook of knowledge representation*. Amsterdam: Elsevier. 1005 p. Available at: http://dai.fmph.uniba.sk/~sefranek/kri/handbook/handbook_of_kr.pdf (accessed March 1, 2021).
5. NIST Digital Library. Mathematical knowledge management. Available at: https://www.nist.gov/mathematical-knowledge-management (accessed February 4, 2021).
6. Finn, V. K. 2011–2012. J. S. Mill's inductive methods in artificial intelligence systems. *Scientific and Technical Information Processing*. Part I (2011). 38(6):385–402. Part II (2012). 39(5):241–260.
7. Grusho, A. A., M. I. Zabezhailo, A. A. Zatsarinny, A. V. Nikolaev, V. O. Piskovski, V. V. Senchilo, and E. E. Timonina. 2017. Klassifikatsiya oshibochnykh sostoyaniy v raspredelennykh vychislitel'nykh sistemakh i istochniki ikh vozniknoveniya [Erroneous states classifications in dictributed computing systems and sources of their occurences]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 27(2):29–40.
8. Ashby, W. R. 1956. *An introduction to cybernetics*. New York, NY: Wiley. 316 p.
9. Grusho, A. A., M. I. Zabezhailo, and E. E. Timonina. 2020. O kauzal'noy reprezentativnosti obuchayushchikh vyborok pretsedentov v zadachakh diagnosticheskogo tipa [On causal representativeness of training samples of precedents in diagnostic type tasks]. *Informatika i ee Primeneniya — Inform. Appl.* 14(1):80–86.
10. Grusho, N. A., A. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2020. Metody nakhozhdeniya prichin sboev v informatsionnykh tekhnologiyakh s pomoshch'yu metadannykh [Methods of finding the causes of information technology failures by means of meta data]. *Informatika i ee Primeneniya — Inform. Appl.* 14(2):33–39.
11. Zabezhailo, M. I. 2020. Some estimates of computational complexity when predicting the properties of new objects using characteristic functions. *Autom. Doc. Math. Linguist.* 54(6):298–305. doi: 10.3103/S0005105520060072.
12. Zabezhailo, M. I., and Y. Y. Trunin. 2020. K probleme nadezhnosti meditsinskogo diagnoza, formiruyemogo na osnove empiricheskikh dannykh [To the reliability of medical diagnosis based on empirical data]. *Iskusstvennyy intellekt i prinyatie resheniy* [Scientific and Technical Information Processing] 4:3–13.

Received February 9, 2021

Contributors

Zabezhailo Michael I. (b. 1956) — Doctor of Science in physics and mathematics, 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

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, professor, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, 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 Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; grusho@yandex.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; eltimon@yandex.ru

НЕЙРОСЕТЕВОЙ ПОДХОД К ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ ПОДДЕРЖКЕ ПРОЦЕССОВ КОНТРОЛЯ И ОХРАНЫ ВОДНЫХ БИОЛОГИЧЕСКИХ РЕСУРСОВ*

А. А. Зацаринный[1], *А. М. Растрелин*[2], *А. П. Сучков*[3]

Аннотация: Рассмотрены вопросы использования искусственных нейронных сетей (ИНС) для решения части задач информационно-аналитического обеспечения процессов целеполагания и ситуационного управления в системе контроля и охраны водных биологических ресурсов (ВБР). Анализ данной предметной области позволяет выделить ряд наукоемких прикладных задач информационно-аналитического обеспечения контроля и охраны ВБР, связанных прежде всего с целеполаганием, расчетом сил и средств, а также с их ситуационным управлением. Осуществлена классификация задач целеполагания, планирования и ситуационного управления в данной области. Обоснованы структура, состав исходных входных данных и выходов ИНС двух типов — классификации и прогнозирования. Обсуждены вопросы обучения и тестирования нейросистем.

Ключевые слова: искусственные нейронные сети; водные биологические ресурсы; информационно-аналитическая поддержка

DOI: 10.14357/08696527210107

1 Введение

Расширение возможностей цифровой науки должно происходить не только путем создания системы поддержки научных исследований, но и, прежде всего, разработки прикладных научных сервисов для различных отраслей народного хозяйства. Одна из таких наукоемких задач — охрана природных и биологических ресурсов.

Для каждого рыбохозяйственного бассейна Минсельхозом России устанавливаются правила рыболовства, регламентирующие добычу (вылов) ВБР в соответствии с Федеральным законом № 166-ФЗ [1]. Такие правила распространяются на промысел во внутренних морских водах РФ, в территориальном море, на континентальном шельфе РФ и в исключительной экономической зоне РФ.

*Работа выполнена при частичной поддержке РФФИ (проект 18-29-03091).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, AZatsarinny@ipiran.ru

[2]Научно-исследовательский институт систем автоматизации, Rastrelin@niisa.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ASuchkov@ipiran.ru

Контроль выполнения правил рыболовства осуществляется пограничными органами с целью выявления и пресечения их нарушения в ходе промысла ВБР. Для реализации этих целей на пограничные органы возложены задачи организации контроля и охраны ВБР путем рационального применения имеющихся сил и средств с учетом предшествующего опыта и складывающейся текущей промысловой обстановки.

Анализ данной предметной области позволяет выделить ряд наукоемких прикладных задач информационно-аналитического обеспечения контроля и охраны ВБР, связанных прежде всего с целеполаганием, расчетом сил и средств, а также с их ситуационным управлением. С точки зрения теории управления целеполагание осуществляется на долгосрочную перспективу (стратегическое планирование), среднесрочную перспективу (планирование на ближайший период времени) и краткосрочную перспективу (динамическое ситуационное целеполагание). При этом необходимым условием корректного целеполагания является установление измеримых, ресурсообеспеченных и привязанных ко времени целевых показателей [2–5]. В данном случае в качестве целевых показателей могут служить причиненный ущерб (минимизация) и предотвращенный ущерб (максимизация).

Цель статьи — рассмотрение методических вопросов использования ИНС для информационно-аналитического обеспечения процессов целеполагания и ситуационного управления в системе контроля и охраны ВБР.

2 Постановка задач

Рассмотрим основные задачи, связанные с двумя последними видами обеспечения целесообразности системы управления процессами контроля и охраны ВБР.

Планирование указанной деятельности пограничным органом осуществляется или на предстоящий период времени (полугодие, год) на основании имеющихся ретроспективных статистических данных, или на текущее время с целью выявления и пресечения нарушений правил промысла ВБР с учетом складывающейся реальной промысловой обстановки.

Для информационно-аналитического обеспечения процессов планирования на предстоящий период времени можно выделить следующие задачи.

А1 Прогнозирование на предстоящую путину числа промысловых судов (ПС), осуществляющих лов ВБР во всех районах промысла зон ответственности пограничных органов в разрешенное для промысла время.

А2 Определение числа возможных судов – нарушителей правил рыболовства среди всего множества судов в рыбопромысловом районе и зоне ответственности пограничного органа в планируемый период.

А3 Расчет необходимого количества сил и средств для осуществления контроля промысла ВБР в конкретном районе промысла в разрешенный период лова в рассматриваемый период времени.

Текущее ситуационное управление по складывающейся промысловой обстановке предполагает решение следующих задач.

Б1 Выявление среди ПС, вышедших из порта и следующих к районам промысла, предрасположенных к нарушению правил рыболовства, и организация отслеживания трасс их движения к району промысла на фоне контролируемой надводной обстановки.

Б2 Выявление среди судов, ведущих вылов ВБР в районах промысла, судов, нарушивших правила рыболовства.

Б3 Выявление среди судов, завершивших промысел и следующих в порт сдачи добытого ВБР, судов, пытающихся уклониться от досмотра в морском контрольном пункте.

Б4 Расчет сил и средств пограничного органа для осуществления текущего контроля промысла ВБР и пресечения нарушений в конкретном районе промысла в разрешенный период лова ВБР в рассматриваемый период времени.

Рассматриваемая в статье научная проблема состоит в обосновании выбора численных методов анализа для обеспечения решения перечисленных задач в условиях больших объемов и номенклатуры данных и обусловленной этим сложности выявляемых зависимостей. Обширность данных предметной области определяется [6]:

- огромной площадью всех акваторий России, в которых возможен промысел ВБР, — около 24,7 млн км^2, в том числе исключительная экономическая зона — 8,2 млн км^2;
- большой номенклатурой контролируемых ВБР — 250 добываемых видов ВБР общим объемом вылова более 4,2 млн т в год;
- численностью ПС, которая превышает 2,5 тыс. единиц;
- сложностью общей судовой обстановки; так, в Дальневосточном регионе она составляет более 4 тыс. целей ежемоментно;
- большим объемом текущих рыночных показателей (данные лицензирования и квотирования, стоимость добываемого валютоемкого вида ВБР, мировые статистические данные по рыболовству и т. п.);
- учетом статистики погодных условий.

В связи с этим в статье в качестве типовых задач для применения ИНС выделены задачи классификации (в данном случае — задачи А2, Б1, Б2 и Б3) и прогнозирования (задачи А1 и А2). Задачи А3 и Б4 относятся к классу задач по многокритериальному выбору в условиях ограничения ресурсов с учетом установленных целевых показателей и решаются по существующим методикам. Рассмотрены методические подходы к использованию ИНС с целью решения части задач информационно-аналитического обеспечения процессов целеполагания и ситуационного управления в системе контроля и охраны ВБР.

3 Задачи прогнозирования

Обсудим использование ИНС в целях информационно-аналитической поддержки решения задачи по разработке плана на предстоящий период времени. Для прогнозирования числа ПС, осуществляющих лов ВБР, и числа возможных нарушителей в определенном районе промысла в разрешенное для промысла время необходимо осуществить:

- отбор входных статистических ретроспективных данных, влияющих на результат;
- подготовку и преобразование исходных данных с учетом характера проблемы;
- проектирование структуры ИНС, выбор функции активации;
- обучение сети на специально подготовленных обучающих выборках;
- проверку адекватности обучения (тестирование сети);
- при необходимости вербализацию сети для удобства использования.

Правилами рыболовства запрещается одному ПС вести промысел в двух и более рыбопромысловых зонах (районах) в течение одного рейса (выхода в море). Поэтому при проектировании ИНС в качестве основной структурной единицы предлагается принять образ района промысла ВБР, который характеризуется двумя показателями: численностью ПС, ведущих лов конкретного вида ВБР, и численностью возможных нарушителей.

Для всех районов промысла, входящих в зону ответственности пограничного органа, прогноз можно получить либо последовательно, используя ИНС, разработанную для одного типового района, с поочередной заменой входных данных для каждого района, либо разработать ИНС с расширенной структурой, учитывающей все районы промысла в зоне ответственности пограничного органа.

Статистика показывает, что наибольший интерес для нарушителей правил рыболовства представляют такие валютоемкие виды ВБР, как краб, морской еж, лососевые, минтай. Исходя из этого, имеет смысл определять численность ПС, ведущих в конкретном районе промысел только валютоемких видов ВБР, причем с учетом самого вида ВБР, так как орудия лова на борту ПС ориентированы на промысел одного конкретного вида ВБР.

3.1 Структура входных данных

В качестве входных данных для ИНС предлагается использовать значения следующих факторов, влияющих на интерес к ведению промысла в конкретном районе:

- рыночная стоимость добываемого валютоемкого вида ВБР;
- количество вида ВБР, добытого в предыдущие путины в конкретном промысловом районе;

– погодные условия, имевшие место в районе промысла (количество дней в путину, когда волнение моря не позволяло вести промысел);
– разрешенный объем вылова определенного вида ВБР в районе промысла.

Необходимо добавить также временной фактор, отражающий ежегодные тенденции по увеличению вылова ВБР и уменьшению нарушений, обусловленных укреплением контроля и охраны ВБР. Можно также рассмотреть временны́е сезонные факторы, но это можно будет использовать для улучшения прогностических свойств нейросети в случае неудовлетворительных результатов ее тестирования.

В качестве выходных данных сети предлагается прогнозируемое число судов и судов-нарушителей на предстоящую путину.

Сбор данных должен осуществляться в пограничном органе (пограничном управлении (ПУ)) по результатам годовой деятельности (отчеты, справки) и по результатам промысла ВБР, отражаемых в данных отраслевой системы мониторинга (ОСМ) Росрыболовства, в данных Всероссийского научно-исследовательского института рыбного хозяйства и океанографии и в отчетах Счетной палаты РФ.

Для использования исходных данных при обучении сети удобно их представить в виде, приведенном в таблице, где $T_j + \Delta T_i$ — год и период времени в месяцах, разрешенный для промысла ВБР в этом году; T^P — прогнозируемый период; входные данные:

X_1 — стоимость добываемого вида ВБР;

X_2 — объем добытого ВБР;

X_3 — разрешаемый объем вылова вида ВБР;

X_4 — погода в районе промысла;

выходные данные:

Y^r — число промысловых судов, реально участвовавших в промысле; Y^p — прогнозируемое число промысловых судов; Z^r — число реальных судов-нарушителей; Z^p — прогнозируемое число судов-нарушителей.

Статистические данные по промыслу ВБР

СТАТИСТИЧЕСКИЕ ДАННЫЕ ПО ПРОМЫСЛУ ВБР В ЗОНЕ ОТВЕТСТВЕННОСТИ ПУ (*название*)							
РЕЗУЛЬТАТЫ ПРОМЫСЛА (*вид валютоемкого ВБР*) в зоне ответственности ПУ							
№	Район промысла (*название*)	Входные данные				Выходные данные	
п/п	и периоды промысла	X_1	X_2	X_3	X_4	Y^r	Z^r
1	$T_1 + \Delta T_1$	X_{11}	X_{12}	X_{13}	X_{14}	Y_1^r	Z_1^r
2	$T_2 + \Delta T_2$	X_{21}	X_{22}	X_{23}	X_{24}	Y_2^r	Z_2^r
3	$T_3 + \Delta T_3$	X_{31}	X_{32}	X_{33}	X_{34}	Y_3^r	Z_3^r
...	$T_j + \Delta T_j$	X_{j1}	X_{j2}	X_{j3}	X_{j4}	Y_j^r	Z_j^r
	T^P					Y^p	Z^p

В течение года может быть несколько периодов времени, в которые разрешен (или запрещен) промысел ВБР. Например, правилами рыболовства в подзоне Приморье разрешен промысел краба в периоды ΔT_1 с 1 по 31 мая и ΔT_2 с 1 августа по 31 октября текущего года. Поэтому подбор статистических данных и прогнозирование численности ПС, ведущих промысел в конкретном районе, должны вестись по каждому разрешенному периоду лова.

3.2 Структура и настройка нейросети

Количество сил и средств, необходимое для контроля промысла непосредственно в районе промысла на определенном интервале времени, будет зависеть от числа ПС, ведущих промысел, и продолжительности проверки ПС государственным инспектором. Поэтому предлагается с помощью ИНС прогнозировать число ПС, ведущих промысел, и ПС-нарушителей в конкретном районе для каждого вида промышляемых ВБР. Структура ИНС в этом случае будет иметь вид, приведенный на рис. 1, где Y — предполагаемое число ПС в районе; Z — предполагаемое число ПС-нарушителей в районе; X_1 — усредненная рыночная стоимость определенного вида ВБР в период промысла (усреднение проводится по стоимости на рынках стран-потребителей браконьерского ВБР); X_2 — объем добытого определенного вида ВБР в данном районе промысла за предыдущий период промысла; X_3 — разрешенный объем вылова определенного вида ВБР (по установленным квотам на рассматриваемый район и период промысла); X_4 — погода в районе промысла в рассматриваемый период промысла; T — год промысла; w_{ij} — веса соответствующих синапсов; F^Y и F^Z — функции активации (функция может быть линейной, сигмоидом (логистической) и гиперболическим тангенсом).

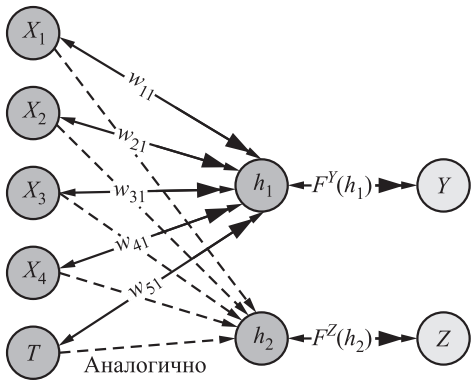


Рис. 1 Структура нейросети прогнозирования

Все входящие данные должны быть нормализованы (обычно приведены к интервалу $[0, 1]$).

Искусственная нейронная сеть прямого распространения может быть представлена в виде системы уравнений, где прогнозируемое число ПС на промысле валютоемких видов ВБР в рассматриваемом районе в интересующий интервал промысла составит:

$$Y = F^Y\left(Tw_{15} + \sum_{i=1}^{4} w_{1i}x_i\right); \quad Z = F^Z\left(Tw_{25} + \sum_{i=1}^{4} w_{2i}x_i\right).$$

Обучение сети предполагается проводить с учителем путем подбора значений весов синапсов с использованием метода среднеквадратичной ошибки и, при необходимости, путем изменения топологии сети за счет включения в скрытый слой дополнительных нейронов. При обучении с учителем для каждого элемента выборки (X_1, X_2, X_3, X_4, T) — реальных входных данных (статистика за предыдущие путины) — вычисляется «экспериментальная» оценка Y и Z .

Процесс обучения заключается в пошаговом вычислении Y и Z для элементов выборки (X_1, X_2, X_3, X_4, T) с изменением на каждом шаге значения веса синапсов и сравнении полученных значений Y^p и Z^p с имевшими место реальными значениями Y^r и Z^r. Вес синапса должен меняться так, чтобы ответы сети стремились к сближению с реальными значениями. Для выяснения, насколько хорошо сеть справляется с поставленной задачей, используется среднеквадратичная ошибка. Для целенаправленного уменьшения ошибки применяется метод обратного распространения, который использует алгоритм градиентного спуска. В качестве начальных значений весов синапсов целесообразно выбрать вариант с равными вкладами в целевое прогнозируемое значение.

Для проверки качества обучения сети (тестирования сети) должны использоваться элементы выборки (X_1, X_2, X_3, X_4, T) и соответствующие им реальные значения Y^r и Z^r, не вошедшие в обучающую выборку. Технология тестирования сети аналогична технологии обучения.

4 Задачи классификации

4.1 Отбор входных данных, влияющих на результат

План ситуационного управления подразумевает оперативное выявление судов, нарушающих правила промысла, и организацию их проверки и пресечения нарушений. Исходные данные для решения этой задачи получаются от систем слежения за промысловой обстановкой реального времени КИИС МоРе, АИС, ОСМ Росрыболовства, автоматизированной системой технического контроля (АСТК) надводной обстановки в зоне ответственности пограничного органа [7].

Предполагается, что комплекс программ классификационной ИНС входит в состав программного обеспечения мониторинга промысловой обстановки и моменты подключения ИНС к оценке ПС определяются моментом первичного появления судна в надводной обстановке или темпом обновления надводной обстановки (данных о каждом судне) по данным от систем КИИС МоРе, АИС, ОСМ, АСТК. С использованием ИНС при каждом обновлении данных о судне осуществляется оценка принадлежности каждого промыслового судна к одному из трех классов:

класс 1: суда-нарушители правил рыболовства;

класс 2: суда — возможные нарушители правил рыболовства;

класс 3: суда, не являющиеся нарушителями.

Для упрощения структуры ИНС весь процесс добычи ВБР условно представим в виде трех этапов:

(1) следование рыбопромыслового судна из порта в район промысла;

(2) ведение вылова (добычи) ВБР в районе промысла;

(3) следование ПС из района промысла в порт сдачи продукции или к плавучему рыбопромысловому заводу (к этому же этапу относится следование судна-перевозчика ВБР или плавучего завода в порт сдачи продукции).

Таким образом, первоначально на основании статистических данных о результатах промысла судном ВБР в предыдущие путины можно определить расположенность судна к нарушениям в текущей путине. Выявление таких судов позволит организовать отслеживание трасс их движения на фоне надводной обстановки к району промысла и при промысле. На втором этапе требуется выявить суда, нарушающие правила промысла ВБР, а на третьем этапе — суда, нарушающие правила транспортировки ВБР в порты сдачи продукции.

В этом случае обобщенную структуру ИНС можно представить в виде трех взаимосвязанных ИНС, обеспечивающих классификацию возможных судов-нарушителей на каждом этапе промысловой деятельности (рис. 2).

Входные данные ИНС-1 для выявления потенциально возможных нарушителей на предварительной стадии оценки промысловой обстановки, т. е. на стадии следования ПС по акватории зоны ответственности пограничного органа от входа в нее до захода в район промысла, могут быть косвенно связаны с возможностью нарушения.

Эти характеристики должны выявляться по результатам анализа деятельности ПС в предшествующие путины. К таким входным данным отнесем:

x_1 — ПС (название и номера IMO (International Maritime Organization), MMSI (Maritime Mobile Service Identify)) совершало нарушение правил промысла в предыдущих путинах;

x_2 — капитан судна ранее был капитаном ПС, совершавшего нарушения правил промысла;

x_3 — владельцу данного судна принадлежат суда, совершавшие ранее нарушения правил рыболовства;

x_4 — имеет место замена названия ПС (при прежних номерах IMO, MMSI);

x_5 — имеет место замена флага (или флага и названия) ПС;

x_6 — ПС уклонялось от досмотра в морском контрольном пункте (МКП) в предыдущие путины;

x_7 — объем выловленного в предыдущие путины и сданного в порту ВБР не превышает установленного порогового значения выбора квоты.

Входными данными ИНС-2 для выявления ПС-нарушителей правил промысла в процессе добычи ВБР должны служить характеристики, устанавливающие ограничения на порядок промысла. К таким входным данным отнесем:

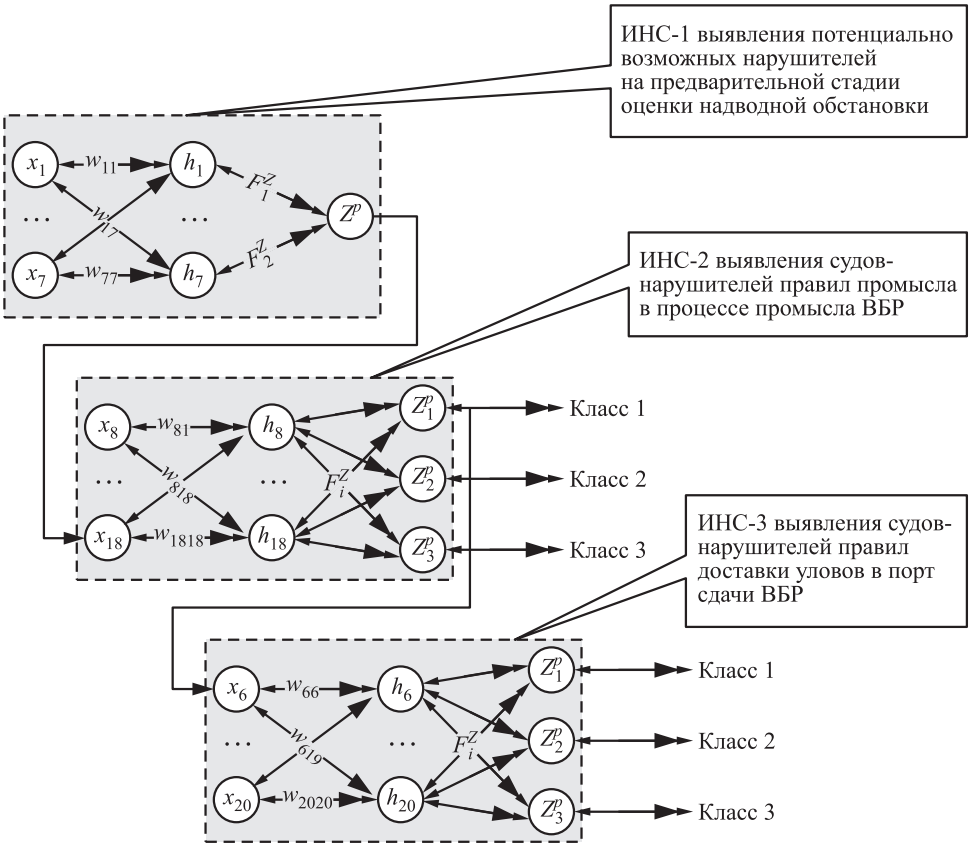


Рис. 2 Структура ИНС классификации

x_8 — отсутствие у ПС документа «Разрешение на добычу (вылов) водных биологических ресурсов» (далее — Разрешение), выдаваемого этому судну Росрыболовством;

x_9 — нахождение в не разрешенном для плавания районе;

x_{10} — ведение промысла в районе, не указанном в Разрешении;

x_{11} — ведение промысла во время, запрещенное для промысла;

x_{12} — ведение промысла ВБР, не указанного в Разрешении на промысел данного ПС;

x_{13} — ведение промысла орудиями лова, не указанными в Разрешении на промысел данного ПС;

x_{14} — наличие сведений об отключении технических средств контроля (ТСК) на время, превышающее допустимое;

x_{15} — наличие несоответствия в позициях судна по данным ТСК ОСМ и данным из судового суточного донесения (ССД) капитана судна;

x_{16} — объем выловленного ВБР превышает установленный в Разрешении на промысел;

x_{17} — использование ПС тактики лова, свойственной браконьерскому промыслу;

x_{18} — результаты работы ИНС-1 по данному судну: «1», если ПС — предполагаемый нарушитель, и «0», если расположенности к нарушениям правил рыболовства не выявлено.

Входными данными ИНС-3 для выявления ПС-нарушителей правил прохождения МКП при следовании в порт сдачи продукции может служить расхождение трассы движения в иностранный порт ПС с местом дежурства пограничного сторожевого корабля (ПСКР) в МКП:

x_6 — ПС уклонялось от досмотра в МКП в предыдущие путины;

x_{19} — результаты исследования судна с использованием ИНС-3: «1», если ПС — нарушитель, и «0», если судно — не нарушитель;

x_{20} — прохождение ПС вне зоны дежурства ПСКР в МКП.

Входной сигнал x_{20} принимает значение «1» при уклонении судна от досмотра в МКП и следовании в иностранный порт, «0» — при завершении промысла и выходе из района промысла.

Сети связаны между собой объектом исследования (нейрон — это ПС) и через результаты работы: выход ИНС-1 является одним из входных сигналов ИНС-2, выход ИНС-2 является одним из входных сигналов ИНС-3. Все сети — прямого распространения, синхронные, с обучением учителем.

С использованием перечисленных ИНС осуществляется анализ каждого судна, находящегося в зоне ответственности пограничного органа, на основе данных о нем, предоставляемых системами мониторинга надводной и промысловой обстановки АСТК и ОСМ Росрыболовства. В зависимости от места нахождения судна программным комплексом мониторинга обстановки АСТК подключаются для исследования судна либо сеть ИНС-1, либо ИНС-2, либо ИНС-3.

Подготовка исходных данных для ИНС осуществляется с использованием системы мониторинга АСТК надводной и промысловой обстановки и хранит в своих базах данных всю необходимую информацию и дополняет ее текущими результатами мониторинга надводной и промысловой обстановки. Наряду с данными о текущей путине хранятся статистические данные по ПС за предыдущие путины.

Функция F_i^z служит функцией активации нейрона сети и является выходом нейрона. Для предлагаемой ИНС в качестве функции активации выберем логистическую функцию:

$$F(x) = \frac{1}{1+e^{-x}}, \quad 0 < F(x) \le 1.$$

Система уравнений, отражающая зависимость выходов ИНС-1, ИНС-2 и ИНС-3 в алгебраической форме, следующая:

$$\begin{aligned}
Z^P &= F^Z\left(\sum w_i x_i + b_i\right), \qquad i = 1,\dots,7;\\
Z_1^P &= F_1^Z\left(\sum w_{1i}^p x_i + b_{i1}^p\right), \quad i = 8,\dots,18;\\
Z_2^P &= F_3^Z\left(\sum w_{i2}^p x_i + b_{i2}^p\right), \quad i = 8,\dots,18;\\
Z_3^P &= F_3^Z\left(\sum w_{i3}^p x_i + b_{i3}^p\right), \quad i = 8,\dots,18;\\
Z_1^R &= F_1^Z\left(\sum w_{i1}^R x_i + b_{i1}^R\right), \quad i = 6,19,20;\\
Z_2^R &= F_2^Z\left(\sum w_{i2}^R x_i + b_{i2}^R\right), \quad i = 6,19,20;\\
Z_3^R &= F_3^Z\left(\sum w_{i3}^R x_i + b_{i3}^R\right), \quad i = 6,19,20.
\end{aligned}$$

4.2 Обучение системы

В рассматриваемом случае при обучении с учителем на вход сети ИНС-1 подается сигнал $\langle X_1,\dots,X_7\rangle$, ИНС-2 — сигнал $\langle X_8,\dots,X_{18}\rangle$, ИНС-3 — сигнал $\langle X_6, X_{19}, X_{20}\rangle$. Так как входы нейросетей принимают значения 0 и 1, то источником обучающих входных выборок могут быть генераторы случайных чисел, выходных — экспертная оценка принадлежности ПС тому или иному классу. Задаются начальные значения весовых коэффициентов w_{ij} (обычно это набор равных значений) и генерируется первая последовательность входных сигналов. Ответ сети сравнивается с экспертной оценкой. Затем с помощью специальных алгоритмов изменяются веса синапсов нейронной сети и снова подается входной сигнал. После сравнения и верификации ответа сети этот процесс повторяется до тех пор, пока сеть не начнет отвечать с приемлемой точностью.

Все алгоритмы обучения нейронных сетей, применяемые в настоящее время, основываются на оценочной функции. Методики обучения базируются на принципе достаточности, в котором ошибка системы не может превышать определенного показателя. Для вычисления очередных поправок весов синапсов обычно применяется метод градиентного спуска.

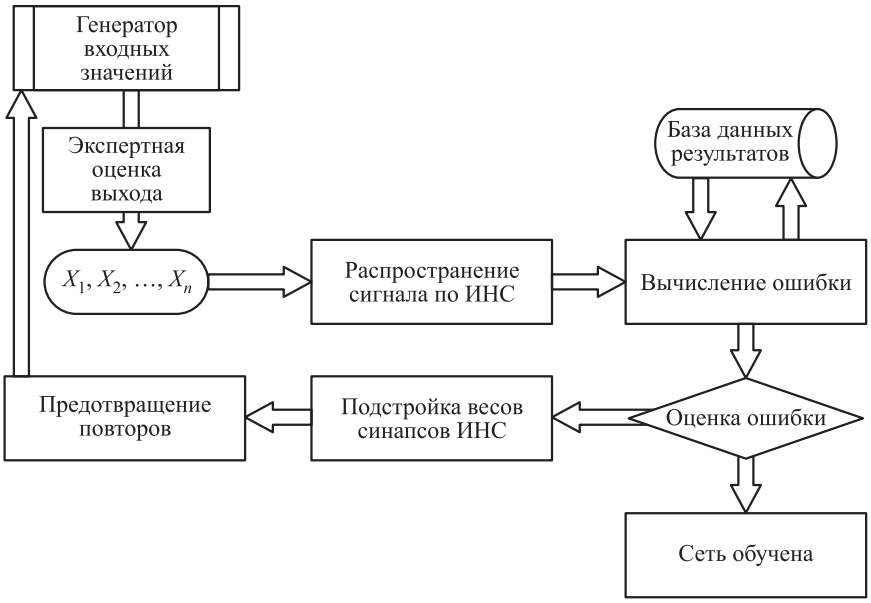


Рис. 3 Схема обучения ИНС классификации

Общая схема обучения ИНС представлена на рис. 3.

Следует отметить, что процесс обучения и тестирования может завершиться неудачно, т. е. не обеспечить требуемую точность классификации. В этом случае необходимо либо пересмотреть состав и объем обучающих выборок, либо модифицировать структуру ИНС по следующим возможным направлениям:

- пересмотр номенклатуры входных данных;
- увеличение числа внутренних слоев нейронов;
- изменение функции активации нейронов;
- подбор других методов поиска экстремума.

5 Заключение

1. Представлен и обоснован возможный методический подход к использованию ИНС для информационно-аналитической поддержки деятельности пограничных органов по организации охраны ВБР и государственному контролю в этой сфере.
2. Следующим шагом должно стать создание опытного образца ИНС, опирающегося на реальные результаты мониторинга промысловой обстановки за

несколько предыдущих лет с целью уточнения косвенных признаков возможных нарушений и конкретизации входных данных.

3. Для эффективного решения рассмотренных задач необходима постановка комплекса НИОКР с привлечением ведущих научных и промышленных организаций, а также специалистов, участвовавших непосредственно в контроле и охране ВБР в морских акваториях России, и специалистов в области создания ИНС.
4. Дальнейшее развитие данного методического подхода может осуществляться в направлении создания системы типовых научных сервисов, доступных для их целевого использования в любом промысловом районе.

Литература

1. О рыболовстве и сохранении водных биологических ресурсов: Федеральный закон от 20 декабря 2004 года № 166-ФЗ.
2. *Зацаринный А. А., Сучков А. П., Босов А. В.* Ситуационные центры в современных информационно-телекоммуникационных системах специального назначения // ВКСС Connect! (Ведомственные корпоративные сети и системы), 2007. № 5(44). С. 64–75.
3. *Сучков А. П.* Ситуационный подход и информационная модель предметной области в правоохранительной сфере // Методы построения и технологии функционирования ситуационных центров. — М.: ИПИ РАН, 2011. С. 76–88.
4. *Сучков А. П.* Формирование системы целей для ситуационного управления // Системы и средства информатики, 2013. Т. 23. № 2. С. 171–182.
5. *Зацаринный А. А., Сучков А. П.* Системотехнические подходы к созданию системы поддержки принятия решений на основе ситуационного анализа // Информатика и её применения, 2016. Т. 10. Вып. 4. С. 105–113.
6. *Сухаренко А. Н., Туровец А. Е., Жерновой М. В., Хренков О. В.* Незаконный оборот водных биоресурсов на Дальнем Востоке как угроза экономической безопасности России. — Владивосток: Экономическая газета, 2014. 66 с.
7. *Неилко О. Б., Айдаров Е. Ю., Панченко В. В., Растрелин А. М.* Информационные технологии мониторинга состояния на основе сбора информации от технических средств наблюдения // Методы построения и технологии функционирования ситуационных центров. — М.: ИПИ РАН, 2011. С. 124–135.

Поступила в редакцию 26.10.20

NEURAL NETWORK APPROACH FOR INFORMATION AND ANALYTICAL SUPPORT OF CONTROL AND PROTECTION OF AQUATIC BIOLOGICAL RESOURCES

***A. A. Zatsarinny*[1], *A. M. Rastrelin*[2], *and A. P. Suchkov*[1]**

[1]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

[2]Institute of Science and Technology of Automated Systems, 2 Volgogradsky Prosp., Moscow 109316, Russian Federation

Abstract: The article deals with the use of artificial neural networks (ANN) to solve some of the information and analytical support problems of goal-setting and situational management processes in the control and protection of aquatic biological resources (ABR) system. The analysis of this subject area allows one to identify a number of high-tech applied tasks of information and analytical support for the control and protection of ABR, primarily related to goal setting, calculation of forces and means, as well as their situational management. The classification of goal-setting, planning, and situational management tasks in this area is carried out. The structure and composition of the initial input data and outputs of two types of ANN — classification and forecasting — are justified. Issues of training and testing of neural systems are discussed.

Keywords: artificial neural networks; aquatic biological resources; information and analytical support

DOI: 10.14357/08696527210107

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-29-03091).

References

1. 166-FZ. December 20, 2004. O rybolovstve i sokhranenii vodnykh biologicheskikh resursov: Federal'nyy zakon [On fishing and conservation of aquatic biological resources: Federal law No. 166-FZ dated December 20, 2004].
2. Zatsarinnyj, A. A., A. V. Suchkov, and A. V. Bosov. 2007. Situatsionnye tsentry v sovremennykh informatsionno-telekommunikatsionnykh sistemakh spetsial'nogo naznacheniya [Situational centers in modern information-telecommunicational network of special purposes]. *VKSS Connect! (Vedomstvennye korporativnye seti i sistemy)* [VKSS Connect! (Departmental Corporate Networks and Systems] 5(44):64–76.
3. Suchkov, A. P. 2011. *Situatsionnyy podkhod i informatsionnaya model' predmetnoy oblasti v pravookhranitel'noy sfere* [Situational approach and information model of a subject domain in the law enforcement sphere]. Metody postroeniya i tehnologii

funktsionirovaniya situatsionnykh tsentrov [Methods of construction and technology of operation of situational centers]. Moscow: IPI RAN. 76–88.

4. Suchkov, A. P. 2013. Formirovanie sistemy tseley dlya situatsionnogo upravleniya [The formation of the objective system to situational management]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(2):171–182.
5. Zatsarinny, A. A., and A. P. Suchkov. 2016. Sistemotekhnicheskie podkhody k sozdaniyu sistemy podderzhki prinyatiya resheniy na osnove situatsionnogo analiza [Systems engineering approaches to a decision support system based on situational analysis]. *Informatika i ee Primeneniya — Inform. Appl.* 10(4): 105–113.
6. Sukharenko, A. N., A. E. Turovets, M. V. Zhernovoy, and O. V. Khrenkov. 2014. *Nezakonnyy oborot vodnykh bioresursov na Dal'nem Vostoke kak ugroza ekonomicheskoy bezopasnosti Rossii* [Illegal turnover of aquatic bioresources in the Far East as a threat to Russia's economic security]. Vladivostok: Ekonomicheskaya gazeta. 66 p.
7. Neilko, O. B., E. Yu. Aydarov, V. V. Panchenko, and A. M. Rastrelin. 2011. *Informatsionnye tekhnologii monitoringa sostoyaniya na osnove sbora informatsii ot tekhnicheskikh sredstv nablyudeniya* [Information technologies for monitoring the state based on collecting information from technical monitoring tools]. Metody postroeniya i tekhnologii funktsionirovaniya situatsionnykh tsentrov [Methods of construction and technology of operation of situational centers]. Moscow: IPI RAN. 124–135.

Received October 26, 2020

Contributors

Zatsarinny Alexander A. (b. 1951) — Doctor of Science in technology, professor, Deputy Director, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS); principal scientist, Institute of Informatics Problems, FRC CSC RAS; 44-2 Vavilov Str., Moscow 119333, Russian Federation; AZatsarinny@ipiran.ru

Rastrelin Anatolij M. (b. 1937) — Doctor of Science in technology, professor, Director on Research and Development, Institute of Science and Technology of Automated Systems, 2 Volgogradsky Prosp., Moscow 109316, Russian Federation; rastrelin@niisa.ru

Suchkov Alexander P. (b. 1954) — Doctor of Science in technology, 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; ASuchkov@frccsc.ru

ОЦЕНКА ВЛИЯНИЯ ПОРЯДКА РАСПРЕДЕЛЕНИЯ ПРОЦЕССОВ И ПОТОКОВ В ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ IBM POWER НА ЭФФЕКТИВНОСТЬ ВЫПОЛНЕНИЯ ПАРАЛЛЕЛЬНЫХ ПРИЛОЖЕНИЙ*

***С. И. Мальковский*[1], *А. А. Сорокин*[2], *Г. И. Цой*[3], *В. Ю. Черных*[4], *К. И. Волович*[5]**

Аннотация: Статья посвящена исследованию производительности вычислительных систем на базе современных процессоров IBM POWER при выполнении параллельных приложений. Изучено влияние различных способов распределения вычислительных процессов и потоков по ядрам центральных процессоров на эффективность выполнения программ, разработанных с использованием технологий MPI (Message Passing Interface) и OpenMP (Open Multi-Processing). Полученные результаты могут быть использованы при оценке эффективности алгоритмов распределения вычислительных ресурсов и организации вычислительного процесса в распределенных гетерогенных вычислительных системах.

Ключевые слова: гибридная вычислительная система; IBM POWER8; IBM POWER9; одновременная многопоточность; распределение потоков; распределение процессов; OpenMP; MPI; NAS Parallel Benchmark

DOI: 10.14357/08696527210108

1 Введение

Одной из основных тенденций современного этапа развития высокопроизводительных вычислительных систем является активное использование гибридных компьютерных архитектур, предусматривающих в дополнение к центральным процессорам установку различных «ускорителей вычислений» или сопроцессоров (на ноябрь 2020 г. число таких систем в списке 500 самых производительных

*Работа выполнена при частичной поддержке РФФИ (проект 18-29-03100).

[1]Вычислительный центр Дальневосточного отделения Российской академии наук, sergey.malkovsky@ccfebras.ru

[2]Вычислительный центр Дальневосточного отделения Российской академии наук, alsor@febras.net

[3]Вычислительный центр Дальневосточного отделения Российской академии наук, tsoy.dv@mail.ru

[4]Вычислительный центр Дальневосточного отделения Российской академии наук, syler1983.9@gmail.com

[5]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, KVolovich@frccsc.ru

суперкомпьютеров мира составило 29%). Это связано со снижением темпов роста производительности новых поколений процессоров, вызванным трудностями в разработке новых технологических процессов производства микросхем.

В качестве основы самых производительных из таких гибридных систем совместно с графическими сопроцессорами компании NVIDIA активно применяются центральные процессоры с архитектурой POWER [1]. Так, 2 из 10 самых быстрых суперкомпьютеров мира на ноябрь 2020 г. созданы на ее основе. Одна из причин востребованности новых вычислительных систем на базе процессоров IBM POWER — наличие в них шин NVLink и CAPI (Coherent Accelerator Processor Interface), используемых для связи с сопроцессорами и периферийными устройствами. Они позволяют значительно сократить время на передачу данных между различными устройствами и, соответственно, ускорить их обработку.

Несмотря на возрастающую актуальность суперкомпьютеров, использующих современные процессоры IBM семейства POWER, относительно низкая доля их распространения в отрасли не позволяет в полной мере оценить вычислительные возможности и эффективность применения данной архитектуры в решении различных научных задач, поэтому вопрос оценки производительности вычислительных систем, построенных на ее основе, приобретает несомненную важность и актуальность.

Настоящая статья продолжает исследования [2], связанные с изучением производительности отдельных подсистем вычислительных систем, основанных на процессорах IBM POWER8 [3] и IBM POWER9 [4]. В ней с использованием теста NAS Parallel Benchmark (NPB) [5] дана оценка влияния различных способов распределения вычислительных процессов и потоков по ядрам центральных процессоров на эффективность выполнения программ, использующих различные технологии параллельного программирования. При этом отдельное внимание уделено режимам функционирования центральных процессоров. Полученные результаты могут быть применены при оценке эффективности алгоритмов распределения вычислительных ресурсов и организации вычислительного процесса в распределенных гетерогенных вычислительных системах.

2 Описание вычислительных платформ и используемого системного программного обеспечения

В работе исследовалась производительность двух вычислительных систем: IBM Power System S822LC 8335-GTB (далее — система IBM POWER8) и IBM Power System AC922 8335-GTG (далее — система IBM POWER9). Первая из них создана на базе двух процессоров IBM POWER8, двух сопроцессоров NVIDIA Tesla P100 и оснащена 256 ГБ оперативной памяти. Вторая система основана на двух процессорах IBM POWER9 и четырех сопроцессорах NVIDIA Tesla V100. Объем оперативной памяти этой системы составляет 1 ТБ. Рассмотрим характеристики используемых в системах процессоров.

Десятиядерный процессор IBM POWER8 имеет максимальную частоту 4,023 ГГц и пиковую производительность 0,32 ТФлопс. Он поддерживает технологию одновременной многопоточности (simultaneous multithreading, SMT) [6], позволяющую выполнять до восьми аппаратных потоков на ядро. Каждое вычислительное ядро содержит 32 КБ кеша первого уровня (L1) инструкций, 64 КБ кеша L1 данных и 512 КБ кеша второго уровня (L2). Также на каждое ядро приходится 8 МБ eDRAM (embedded dynamic random access memory) кеша третьего уровня (L3), разделяемого между всеми вычислительными ядрами процессора. Дополнительный кеш четвертого уровня (L4) объемом 64 МБ, реализованный по технологии eDRAM, находится вне процессора на микросхемах Centaur [7], обеспечивающих планирование и управление обменами данными с памятью.

Процессор IBM POWER9 (версия Scale-out) с максимальной частотой 3,5 ГГц и пиковой производительностью 0,56 ТФлопс имеет 20 вычислительных ядер, поддерживающих технологию SMT для четырех потоков. Вычислительные ядра содержат 32 КБ кеша L1 инструкций и 32 КБ кеша L1 данных. На каждую пару ядер IBM POWER9 приходится 512 КБ кеша L2 и 10 МБ eDRAM кеша L3. Будем называть такие два ядра центрального процессора с едиными кешами «модулем».

При проведении исследований использовалось программное обеспечение, рекомендованное производителем вычислительных систем. В качестве операционной системы применялся дистрибутив GNU/Linux CentOS версии 7.6. Компиляция NPB осуществлялась компиляторами IBM XL C/C++ и Fortran 16.1.1 (описание процедуры сборки тестов приведено в работе [2]). При сборке MPI-версий тестов использовалась библиотека IBM Spectrum MPI 10.3.

3 Оценка влияния порядка распределения процессов и потоков на эффективность выполнения параллельных вычислений

В процессе исследований рассматривались вопросы эффективности аппаратного обеспечения при выполнении параллельных приложений. В частности, с применением тестов EP, LU, MG, CG, FT и IS (класс сложности C) из состава NPB 3.4 проводилось исследование достигаемой производительности вычислительных систем от количества вычислительных потоков (здесь и далее применительно к MPI под потоками будут подразумеваться процессы) при различных способах их распределения по ядрам центральных процессоров в различных SMT-режимах (ST, SMT2, SMT4 и SMT8 — 1, 2, 4 и 8 потоков на ядро соответственно) при выполнении параллельных вычислений с использованием технологий OpenMP и MPI. Число потоков, которое при этом использовалось, представлено в таблице. Из нее видно, что при выполнении первой группы тестов были задействованы все процессорные ядра вычислительных систем, а при выполнении второй — только 80%. Это объясняется тем, что тесты MG, CG, FT и IS могут выполняться лишь на числе потоков, кратном степени двойки.

Число потоков, использованных при запуске тестов

Тесты	Система IBM POWER8	Система IBM POWER9
EP, LU	1, 2, 4, 8, 16 и 20 в режиме ST; 40, 80 и 160 — в режимах SMT2, SMT4 и SMT8 соответственно	1, 2, 4, 8, 16, 32 и 40 в режиме ST; 80 и 160 — в режимах SMT2 и SMT4 соответственно
MG, CG, FT, IS	1, 2, 4, 8 и 16 в режиме ST; 32, 64 и 128 — в режимах SMT2, SMT4 и SMT8 соответственно	1, 2, 4, 8, 16 и 32 в режиме ST; 64 и 128 — в режимах SMT2 и SMT4 соответственно

При тестировании вычислительные потоки приложений распределялись между ядрами центральных процессоров вычислительных систем тремя различными способами. Первый способ заключался в равномерном распределении потоков между процессорами. На системе IBM POWER8 при запуске двух процессов задействовалось первое ядро каждого из процессоров, при запуске четырех — первые два ядра и т. д. На системе IBM POWER9 при числе используемых ядер, меньшем либо равном 16, вычислительные потоки запускались на отдельных «модулях» процессоров, включающих по 2 ядра и объединенные кеши L2 и L3. Для примера, при числе вычислительных потоков, равном 4, на первых ядрах первых двух «модулей» каждого из процессоров выполнялось по одному потоку. При числе потоков, равном 32 и 40, на каждое второе ядро задействованных «модулей» добавлялось по одному дополнительному потоку. При использовании второго, компактного способа распределения потоков происходило последовательное заполнение вычислительных ядер сначала первого, а затем второго процессора вычислительной системы. Стоит отметить, что в случае первого и второго способов распределения при активации режимов SMT на каждое задействованное процессорное ядро добавлялось соответствующее число дополнительных вычислительных потоков. Третий способ распределения вычислительных потоков между ядрами центральных процессоров заключался в использовании режима по умолчанию, т. е. какое-либо специальное распределение потоков и их привязка не осуществлялись.

В качестве метода передачи сообщений для Spectrum MPI был применен PAMI (Parallel Active Messaging Interface) [8]. Далее представлены результаты проведенных экспериментов, причем их порядок выстроен в соответствии с уровнем нагрузки на сеть передачи данных [9]. Показатели производительности усреднялись по результатам 5 испытаний. Во избежание получения ошибочных результатов все численные расчеты проводились в монопольном режиме. Динамическое управление частотой центральных процессоров отключалось. При этом устанавливалась максимальная частота процессоров, допустимая при одновременной максимальной загрузке всех вычислительных ядер (см. разд. 1).

Тест EP служит для оценки производительности в расчетах с плавающей точкой при отсутствии заметных межпроцессорных взаимодействий. Он включает генерацию псевдослучайных нормально распределенных чисел. На рис. 1

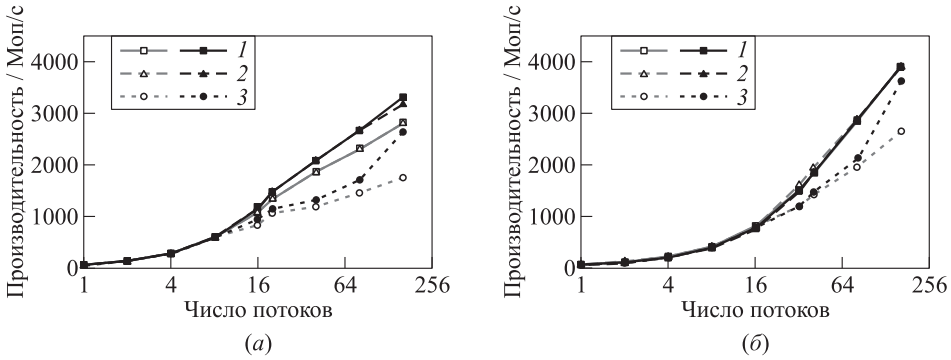


Рис. 1 Зависимость производительности в тесте EP от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

приведены графики зависимости производительности от числа вычислительных потоков для технологий OpenMP и MPI при различных способах их распределения по процессорным ядрам.

Из результатов, приведенных на рис. 1, видно, что на системе IBM POWER8 технология MPI показывает в 1,2 раза бóльшую максимальную производительность, чем технология OpenMP. При этом на системе IBM POWER9 подобная разница между указанными технологиями отсутствует.

Проведенные эксперименты не позволили обнаружить значимой разницы в достигаемой производительности при равномерном и компактном распределении потоков между ядрами центральных процессоров. Лишь на системе IBM POWER8 равномерное распределение потоков позволяет повысить производительность в 1,04 раза для MPI-версии теста. При этом стоит отметить, что без задания фиксированной привязки вычислительных потоков к процессорным ядрам достигаемая максимальная производительность теста EP снижается на всех системах. Так, на системе IBM POWER8 она начинает снижаться с 8 потоков, а на системе IBM POWER9 — с 16 потоков.

Наибольшую эффективность в тесте EP демонстрирует система IBM POWER9. Ее производительность в 1,2 раза выше производительности системы IBM POWER8.

В тесте LU проводится LU-разложение. На рис. 2 показаны полученные результаты оценки производительности. Из него видно, что OpenMP-версия этого теста достигает большей производительности, чем его MPI-версия. Так, на системе IBM POWER8 первая из них более чем в 1,2 раза быстрее второй. При этом на системе IBM POWER9 OpenMP-версия теста быстрее MPI-версии практически в 1,1 раза. Что касается способов распределения вычислительных

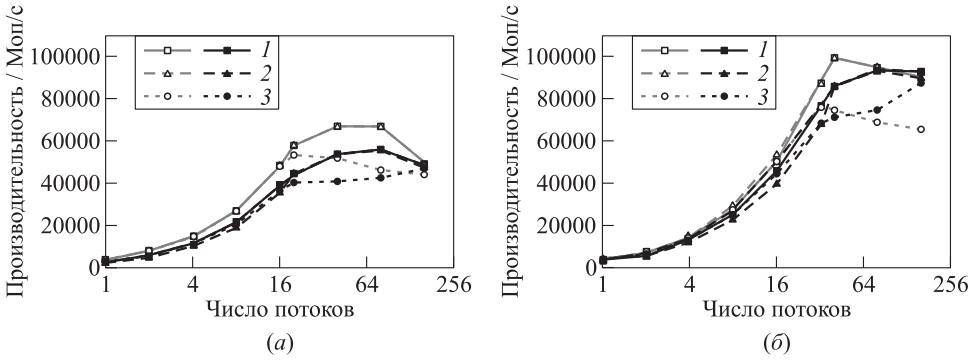


Рис. 2 Зависимость производительности в тесте LU от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

потоков между ядрами центральных процессоров, то на системе IBM POWER8 разницы в достигаемой производительности между ними не наблюдается. На системе IBM POWER9 при числе потоков, меньшем 32, для OpenMP-версии теста небольшое преимущество имеет компактное распределение потоков. Равномерное распределение потоков обладает преимуществом над компактным для MPI-версии при числе потоков, меньшем 40. При отсутствии привязки достигаемая системами производительность снижается для всех версий теста начиная с 16 потоков.

Максимальную производительность в тесте LU имеет вычислительная система IBM POWER9. Она быстрее вычислительной системы IBM POWER8 на 47,9%.

В тесте MG с использованием многосеточного алгоритма находится приближенное решение трехмерного уравнения Пуассона с периодическими граничными условиями. На рис. 3 показаны результаты проведенных экспериментов. Из них видно, что на системе IBM POWER8 максимальная производительность достигается MPI-версией теста, которая оказывается на 20% быстрее OpenMP-версии. При этом на системе IBM POWER9 OpenMP-версия производительнее MPI-версии на 2,5%. В обоих случаях достичь наибольшей эффективности позволяет равномерное распределение потоков по ядрам центральных процессоров вычислительных систем. Выводы относительно производительности вычислительных систем при отсутствии привязки потоков к вычислительным ядрам аналогичны выводам, сделанным для тестов EP и LU.

Что касается общей эффективности вычислительных систем при выполнении теста MG, то система IBM POWER9 оказывается в 1,2 раза производительнее системы IBM POWER8.

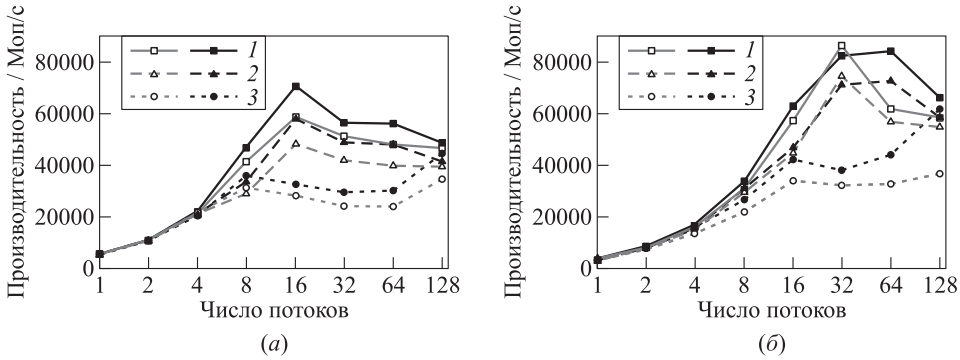


Рис. 3 Зависимость производительности в тесте MG от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

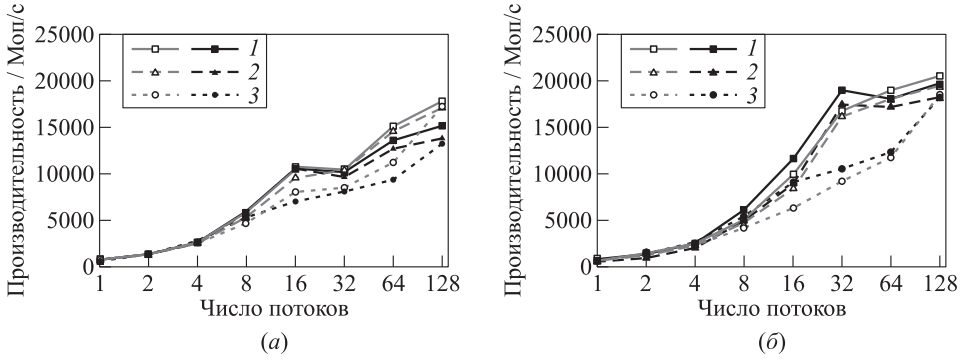


Рис. 4 Зависимость производительности в тесте CG от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

В тесте CG решается система линейных алгебраических уравнений с разреженной произвольной матрицей методом сопряженных градиентов. Коммутации в MPI-реализации организованы с помощью неблокирующих двухточечных взаимодействий. Из рис. 4 видно, что при числе потоков, меньшем 32, преимущество в производительности демонстрирует MPI-версия теста. Однако максимальная производительность достигается его OpenMP-версией при числе потоков, равном 128. Стоит отметить, что в обоих случаях наиболее эффективным оказывается равномерное распределение потоков. При этом обращает

на себя внимание сверхлинейное ускорение выполнения MPI-версии теста при числе процессов от двух до шестнадцати на системе IBM POWER9: в 2,2, 4,03, 9,5 и 18,3 раза соответственно. Оно не наблюдается в том случае, когда процессы равномерно распределяются не между вычислительными «модулями», а между отдельными ядрами системы, что было показано в дополнительных экспериментах. Это связано с архитектурой процессора IBM POWER9, в котором кеш второго и третьего уровня разделяется между парами вычислительных ядер. Поэтому при равномерном распределении процессов по «модулям» возрастает число процессов, единолично использующих эти кеши, что приводит к росту доступной приложению кеш-памяти и, соответственно, дополнительному росту его производительности. Такое поведение не отмечено на системе IBM POWER8, в которой кеш L2 не разделяется между ядрами процессора, а кеш L3 общий.

При отсутствии привязки вычислительных потоков к ядрам центральных процессоров производительность теста CG снижается так же, как и для других тестов NPB.

При выполнении теста CG вычислительная система IBM POWER9 показывает в 1,1 раза большую производительность, чем система IBM POWER8.

В тесте FT решается трехмерная задача с использованием дискретного преобразования Фурье. Взаимодействие между процессами MPI-версии осуществляется с помощью следующих коллективных операций: MPI_Reduce, MPI_Barrier, MPI_Bcast, MPI_Alltoall. Данные, приведенные на рис. 5, свидетельствуют о том, что технология OpenMP показывает на системе IBM POWER8 на 41,6% бóльшую производительность, чем технология MPI. Та же разница в производительности на системе IBM POWER9 достигает 60,7%. При этом на обеих системах оптимальным оказывается равномерное распределение

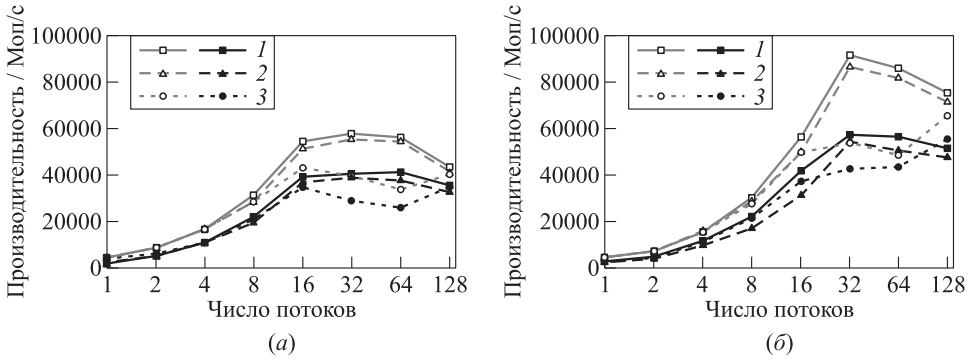


Рис. 5 Зависимость производительности в тесте FT от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

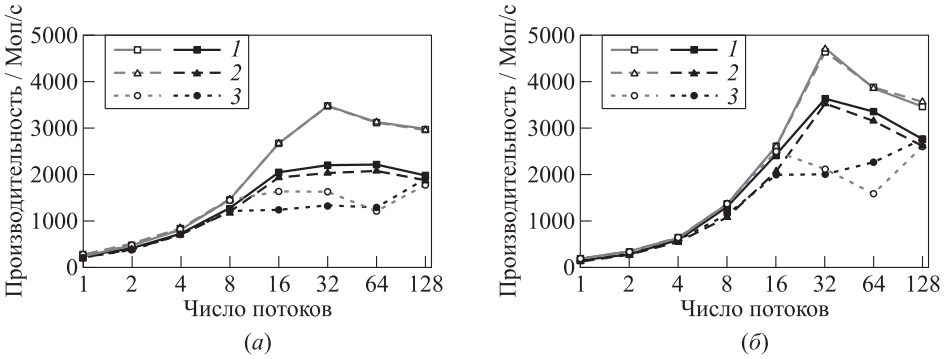


Рис. 6 Зависимость производительности в тесте IS от числа потоков для системы IBM POWER8 (*а*) и IBM POWER9 (*б*): пустые значки и серые кривые — OpenMP; залитые значки и черные кривые — MPI; *1* — равномерное распределение потоков; *2* — компактное распределение потоков; *3* — без привязки

потоков по ядрам центральных процессоров, а в отсутствие привязки наблюдается снижение производительности.

Вычислительная система IBM POWER8 показывает в тесте FT на 36,7% меньшую производительность, чем вычислительная система IBM POWER9.

В тесте IS осуществляется параллельная сортировка большого массива целых чисел (рис. 6). Он используется для оценки производительности целочисленных вычислений при интенсивном межпроцессорном взаимодействии. Передача сообщений между процессами MPI-версии теста осуществляется с помощью операций MPI_Alltoall и MPI_Allreduce. Результаты экспериментов показывают, что технология MPI демонстрирует меньшую производительность, чем OpenMP: на вычислительной системе IBM POWER8 первая из них оказывается на 36,3%, а на вычислительной системе IBM POWER9 — на 24,3% менее производительной, чем вторая. Значимой разницы в компактном и равномерном распределении потоков для технологии OpenMP не наблюдается. Для технологии MPI оптимальным оказывается равномерное распределение процессов по центральным процессорам. При отсутствии привязки потоков для обеих технологий параллельного программирования наблюдается снижение производительности.

При выполнении теста IS вычислительная система IBM POWER9 оказывается в 1,4 раза производительней вычислительной системы IBM POWER8.

4 Анализ полученных результатов

Оценивая результаты, полученные в ходе экспериментальных расчетов, можно сделать вывод о том, что процессор IBM POWER9 имеет в 1,1–1,6 раза

большую производительность в NPB, чем процессор IBM POWER8. При этом наименьшая разница в производительности между процессорами наблюдается в тесте CG, а наибольшая — в тесте FT.

Приложения, разработанные с использованием технологии OpenMP, в большинстве случаев демонстрируют несколько лучший уровень производительности, чем приложения, разработанные с использованием технологии MPI. Лишь MPI-версии теста EP (на обеих системах) и MG (на системе IBM POWER8) оказались быстрее OpenMP-версий. Стоит отметить, что наибольшее преимущество технология OpenMP имеет в тех случаях, когда требуется реализация интенсивного взаимодействия между вычислительными потоками (тесты FT и IS).

Наилучшую масштабируемость при запуске в режиме ST демонстрирует тест EP. Так, на вычислительной системе IBM POWER8 при его выполнении с использованием всех ядер центральных процессоров достигается максимальное ускорение в 19,3 раза, а на вычислительной системе IBM POWER9 — в 31,1 раза. Хуже масштабируются тесты LU, CG, MG и FT (максимальное ускорение в режиме ST — в 15,6, 15,1, 12,5 и 12,3 раза на системе IBM POWER8 и в 25,2, 25,8, 21,4 и 21,6 раза на системе IBM POWER9 соответственно). Наихудшую масштабируемость на системе IBM POWER8 показывает тест IS. При выполнении на 16 ядрах в режиме ST обеспечивается ускорение расчетов в 11,1 раза. При этом на системе IBM POWER9 тест IS демонстрирует масштабируемость, лишь немногим уступающую масштабируемости теста EP. Его выполнение при запуске на 32 ядрах в режиме ST ускоряется в 26,9 раза.

Технология SMT наиболее эффективна на процессорах IBM POWER8. На оснащенной ими системе она позволяет улучшить масштабируемость всех исследованных тестов, за исключением теста MG. Значительно хуже данная технология работает на процессорах IBM POWER9. С ее помощью удается ускорить лишь выполнение тестов EP и CG. При этом наибольший выигрыш от использования SMT на обеих системах демонстрирует тест EP. Данная технология позволяет ускорить его выполнение в 2 раза.

Результаты исследования показали, что для всех вычислительных систем и рассматриваемых технологий параллельного программирования задание фиксированной привязки вычислительных процессов и потоков к процессорным ядрам является важным условием для получения максимальной производительности. Что касается характера распределения процессов и потоков по ядрам процессоров, то для большинства случаев оптимальным оказывается равномерное распределение, позволяющее максимизировать доступную пропускную способность подсистемы памяти. При этом необходимо учитывать архитектурные особенности используемой системы. Так, на системе IBM POWER9 при задействовании небольшого числа процессорных ядер может оказаться полезным разнести потоки таким образом, чтобы каждый из них использовал отдельные кеши L2 и L3. Это не только увеличит объем доступной программе кеш-памяти, но и снизит число потенциальных кеш-конфликтов, что важно для приложений, чувствительных к производительности кешей процессора.

5 Заключение

В статье представлены результаты комплексных исследований по оценке влияния порядка распределения процессов и потоков на производительность вычислительных систем, построенных с использованием современных центральных процессоров IBM POWER. Они показали, что равномерное распределение и привязка вычислительных процессов и потоков к процессорным ядрам позволяют повысить производительность параллельных приложений, а отсутствие привязки снижает ее на всех системах. При задании характера их распределения важно учитывать архитектурные особенности используемой вычислительной системы.

Среди рассмотренных технологий параллельного программирования наибольшую производительность в большинстве тестов показала технология OpenMP.

При выполнении тестов из состава NPB процессор IBM POWER9 оказывается в среднем в 1,3 раза производительнее процессора IBM POWER8 при разнице в пиковой производительности между ними в 1,75 раза. Такая разница между ростом пиковой и реальной производительности может быть объяснена, в том числе, тем, что на процессоре IBM POWER9 технология SMT оказалась менее эффективной, чем на процессоре IBM POWER8.

Полученные научные результаты могут быть использованы при разработке программных средств, оценке эффективности алгоритмов распределения вычислительных ресурсов, а также при определении оптимальной конфигурации и организации вычислительного процесса высокопроизводительных вычислительных систем, построенных на рассматриваемых процессорных архитектурах.

При проведении численных расчетов было использовано оборудование ЦКП «Центр данных ДВО РАН» (ВЦ ДВО РАН, г. Хабаровск) [10] и ЦКП «Информатика» Федерального исследовательского центра «Информатика и управление РАН» (г. Москва) [11].

Литература

1. *Karkhanis T. S., Moreira J. E.* IBM Power architecture // Encyclopedia of parallel computing / Ed. D. Padua. — Boston – New York: Springer, 2011. P. 900–907.
2. *Sorokin A., Malkovsky S., Tsoy G., et al.* Comparative performance evaluation of modern heterogeneous high-performance computing systems CPUs // Electronics, 2020. Vol. 9. Iss. 6. Art. ID: 1035. P. 1–13. doi: 10.3390/electronics9061035.
3. *Sinharoy B., Van Norstrand J. A., Eickemeyer R. J., et al.* IBM POWER8 processor core microarchitecture // IBM J. Res. Dev., 2015. Vol. 59. Iss. 1. P. 2:1–2:21. doi: 10.1147/JRD.2014.2376112.
4. *Sadasivam S. K., Thompto B. W., Kalla R., et al.* IBM Power9 processor architecture // IEEE Micro, 2017. Vol. 37. Iss. 2. P. 40–51. doi: 10.1109/MM.2017.40.

5. *Bailey D., Barszcz E., Barton J., et al.* The NAS parallel benchmarks. RNR Technical Report RNR-94-007, 1994. https://www.nas.nasa.gov/assets/pdf/techreports/1994/rnr-94-007.pdf.
6. *Eggers S. J., Emer J. S., Levy H. M., et al.* Simultaneous multithreading: A platform for next-generation processors // IEEE Micro, 1997. Vol. 17. Iss. 5. P. 12–19. doi: 10.1109/40.621209.
7. *Starke W. J., Stuecheli J., Daly D. M., et al.* The cache and memory subsystems of the IBM POWER8 processor // IBM J. Res. Dev., 2015. Vol. 59. Iss. 1. P. 3:1–3:13. doi: 10.1147/JRD.2014.2376131.
8. *Kumar S., Mamidala A. R., Faraj D. A., et al.* PAMI: A parallel active message interface for the Blue Gene/Q supercomputer // IEEE 26th Parallel and Distributed Processing Symposium (International) Proceedings. — Piscataway, NJ, USA: IEEE, 2012. P. 763–773. doi: 10.1109/IPDPS.2012.73.
9. *Takouna I., Dawoud W., Meinel C.* Analysis and simulation of HPC applications in virtualized data centers // IEEE Conference (International) on Green Computing and Communications Proceedings. — Piscataway, NJ, USA: IEEE, 2012. P. 498–507. doi: 10.1109/ GreenCom.2012.80.
10. *Сорокин А. А., Макогонов С. В., Королев С. П.* Информационная инфраструктура для коллективной работы ученых Дальнего Востока России // Научно-техническая информация. Сер. 1: Организация и методика информационной работы, 2017. № 12. С. 14–16. doi: 10.3103/S0147688217040153.
11. Положение о ЦКП «Информатика». http://www.frccsc.ru/ckp.

Поступила в редакцию 12.08.20

ASSESSMENT OF THE EFFECT OF PROCESSES AND THREADS AFFINITY IN IBM POWER COMPUTING SYSTEMS ON THE PARALLEL APPLICATIONS PERFORMANCE

***S. I. Malkovsky*[1], *A. A. Sorokin*[1], *G. I. Tsoy*[1], *V. Y. Chernykh*[1], *and K. I. Volovich*[2]**

[1]Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim U Chen Str., Khabarovsk 680000, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: The article is devoted to the study of the performance of computing systems based on modern IBM POWER processors when running parallel applications. The effect of different methods of distributing computational processes and threads among central processors' cores on the efficiency of executing programs developed using MPI (Message Passing Interface) and OpenMP (Open Multi-Processing) technologies is studied. The results obtained can

be useful for evaluating the efficiency of algorithms for the distribution of computing resources and the organization of the computing process in distributed heterogeneous computing systems.

Keywords: heterogeneous computing; IBM POWER8; IBM POWER9; simultaneous multithreading; threads affinity; processes affinity; OpenMP; MPI; NAS Parallel Benchmark

DOI: 10.14357/08696527210108

Acknowledgments

The research was partially supported by the Russian Foundation for Basic Research (project 18-29-03100).

References

1. Karkhanis, T. S., and J. E. Moreira. 2011. IBM power architecture. *Encyclopedia of parallel computing.* Ed. D. Padua. Boston–New York: Springer. 900–907.
2. Sorokin, A., S. Malkovsky, G. Tsoy, *et al.* 2020. Comparative performance evaluation of modern heterogeneous high-performance computing systems CPUs. *Electronics* 9(6):1035. 13 p. doi: 10.3390/electronics9061035.
3. Sinharoy, B., J. A. Van Norstrand, R. J. Eickemeyer, *et al.* 2015. IBM POWER8 processor core microarchitecture. *IBM J. Res. Dev.* 59(1):2.1–2.21. doi: 10.1147/JRD.2014.2376112.
4. Sadasivam, S. K., B. W. Thompto, R. Kalla, *et al.* 2017. IBM Power9 processor architecture. *IEEE Micro* 37(2):40–51. doi: 10.1109/MM.2017.40.
5. Bailey, D., E. Barszcz, J. Barton, *et al.* 1994. The NAS parallel benchmarks. RNR Technical Report RNR-94-007. Available at: https://www.nas.nasa.gov/assets/pdf/techreports/1994/rnr-94-007.pdf (accessed March 5, 2021).
6. Eggers, S. J., J. S. Emer, H. M. Levy, *et al.* 1997. Simultaneous multithreading: A platform for next-generation processors. *IEEE Micro* 17(5):12–19. doi: 10.1109/40.621209.
7. Starke, W. J., J. Stuecheli, D. M. Daly, *et al.* 2015. The cache and memory subsystems of the IBM POWER8 processor. *IBM J. Res. Dev.* 59(1):3.1–3.13. doi: 10.1147/JRD.2014.2376131.
8. Kumar, S., A. R. Mamidala, D. A. Faraj, *et al.* 2012. PAMI: A parallel active message interface for the Blue Gene/Q supercomputer. *26th Parallel and Distributed Processing Symposium (International) Proceedings*. Piscataway, NJ: IEEE. 763–773. doi: 10.1109/IPDPS.2012.73.
9. Takouna, I., W. Dawoud, and C. Meinel. 2012. Analysis and simulation of HPC applications in virtualized data centers. *Conference (International) on Green Computing and Communications Proceedings*. Piscataway, NJ: IEEE. 498–507. doi: 10.1109/GreenCom.2012.80.
10. Sorokin, A. A., S. V. Makogonov, and S. P. Korolev. 2017. The information infrastructure for collective scientific work in the Far East of Russia. *Sci. Tech. Inf. Proc.* 44(4):302–304. doi: 10.3103/S0147688217040153.

11. Polozheniye o TsKP «Informatika» ["Informatics" core facility statute]. Available at: http://www.frccsc.ru/ckp (accessed March 5, 2021).

Received August 12, 2020

Contributors

Malkovsky Sergey I. (b. 1983) — scientist, Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim U Chen Str., Khabarovsk 680000, Russian Federation; sergey.malkovsky@ccfebras.ru

Sorokin Aleksei A. (b. 1980) — Candidate of Science (PhD) in technology, leading scientist, Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim U Chen Str., Khabarovsk 680000, Russian Federation; alsor@febras.net

Tsoy Georgiy I. (b. 1992) — Candidate of Science (PhD) in physics and mathematics, scientist, Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim U Chen Str., Khabarovsk 680000, Russian Federation; tsoy.dv@mail.ru

Chernykh Vladimir Y. (b. 1995) — senior programmer, Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim U Chen Str., Khabarovsk 680000, Russian Federation; syler1983.9@gmail.com

Volovich Konstantin I. (b. 1970) — 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; KVolovich@frccsc.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2021. Т. 31. № 1. С. 111–121

ЭВОЛЮЦИЯ СЕТЕВЫХ ПРОЦЕССОРОВ

В. Б. Егоров[1]

Аннотация: Сетевые процессоры (СП) прошли долгий путь развития от универсальных компьютеров, оснащенных сетевыми интерфейсными картами, до высокоинтегрированных систем на кристалле, содержащих десятки и сотни программируемых процессоров, аппаратных ускорителей и сетевых интерфейсов. Высокая производительность СП достигалась по двум разным направлениям: старожилы рынка микроэлектроники наращивали в них число процессорных ядер с традиционной архитектурой, а небольшие молодые компании делали ставку на специализированные архитектуры. Второе направление обеспечивало лучшие удельные характеристики СП, но его приверженцам было трудно выдерживать конкуренцию из-за сложности программирования в нетрадиционных архитектурах и ограниченности ресурсов для создания полноценного программного обеспечения (ПО) их СП. К настоящему времени практически все они покинули рынок СП, будучи так или иначе поглощенными крупными компаниями. Последние же по большей части использовали приобретенные ноу-хау не для выхода на рынок СП, а применяли опыт агрессивной многоядерности для завоевания рынка высокопроизводительных серверных процессоров, бурно растущего в эпоху «облачных» технологий. Таким образом, СП постепенно прекращают существование как самостоятельные продукты и превращаются во внутрифирменные инструменты интеллектуализации сетевых устройств.

Ключевые слова: сетевой процессор; интегрированный сетевой процессор; архитектура процессорного ядра; производительный многоядерный сервер

DOI: 10.14357/08696527210109

1 Введение. Предпосылки появления сетевых процессоров

Понятие сетевого процессора возникло практически одновременно с появлением пакетных сетей передачи данных около полувека тому назад. Сетевой процессор прошлого столетия — это обычный универсальный компьютер, дополненный интерфейсными картами для подключения его к сетям. Все сетевые функции, в том числе коммутация (soft switching) пакетов, реализовывались программно. Такое универсальное решение позволяло легко создавать и совершенствовать устройства сетевой инфраструктуры, а также по мере необходимости переходить на новые сетевые технологии и методы коммутации, причем не важно, шла речь о пакетах или виртуальных соединениях. Практические ограничения этого подхода стали проявляться лишь с повышением скоростей передачи данных

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, VEgorov@ipiran.ru

в сетях и на технологиях, специально ориентированных на быструю коммутацию. Неслучайно в коммутаторах ATM (asynchronous transfer mode) возобладали аппаратные решения по продвижению (hard forwarding) пакетов (в случае ATM — ячеек) в узлах сети.

Однако эра интернета с доминированием протокольного стека TCP/IP (transport protocol control / internet protocol) инициировала разворот сетевой инфраструктуры вновь в сторону программной коммутации. В IP-сетях базовыми узлами стали маршрутизаторы (routers), принципиально обязанные поддерживать ряд протоколов маршрутизации. Кроме того, в сетях масштаба MAN (metropolitan-area networks) и WAN (wide-area networks) актуальными стали вопросы сетевой безопасности, что также потребовало от сетевых устройств поддержки целого комплекса соответствующих протоколов, а также возможности инспекции содержимого пакетов «на лету». Расширение круга задач, растущее число поддерживаемых протоколов и постоянная их обновляемость практически исключали аппаратную реализацию устройств инфраструктуры IP-сетей. В то же время растущие скорости передачи данных предъявляли все более жесткие требования к задержкам коммутации пакетов. Это объективное противоречие заставило разработчиков сетевого оборудования искать решения, способные обеспечить компромисс между гибкостью программной коммутации и скоростью аппаратного продвижения пакетов.

По счастливому совпадению, ко времени возникновения указанного противоречия, где-то на рубеже тысячелетий, микроэлектронная промышленность преодолела свой «рубеж» в миллиард транзисторов на кристалле кремния и искала сферы широкой устойчивой применимости новых СБИС-технологий. Одной из таких сфер стали интегрированные СП. Новые возможности микроэлектроники позволили объединить на кристалле кремния множество процессоров, в том числе многоядерных (multicore) и многопоточных (multithreaded), с разнообразными специализированными средствами обработки сетевого трафика и сетевыми интерфейсами [1].

Сетевые процессоры в качестве основы сетевых устройств предоставляли разработчикам последних функциональную гибкость за счет программируемости при высокой производительности и пропускной способности благодаря большому числу разнообразных обрабатывающих устройств [2]. В свою очередь, микроэлектронным фирмам интегрированные СП обещали устойчивый высокий спрос их продукции за счет быстрого роста рынка сетевого оборудования вследствие расширяющегося охвата всех сторон жизни современного общества сетями от локальных до глобальных.

2 Эволюция архитектур сетевых процессоров

Вторая половина первого и первая половина второго десятилетий XXI в. ознаменовались бурным расцветом СП. На многообещающий рынок вышли со своими продуктами как гранды микроэлектроники, такие как Intel и Motorola,

так и ряд небольших фирм. Соответственно обозначились два направления развития СП, различающиеся степенью универсальности программируемых процессоров. Крупные компании, имевшие опыт разработки и выпуска универсальных процессорных ядер, стремились максимально использовать имеющиеся заделы для захвата рынка СП. Мелкие фирмы старались составить им конкуренцию за счет нетривиальных, более эффективных архитектурных решений на основе фирменных ноу-хау.

Первый подход получил свое наиболее яркое обобщенное выражение в макроархитектуре Layerscape, анонсированной компанией Freescale Semiconductor в 2011 г. Макроархитектура определяет три функциональных слоя обобщенной архитектуры интегрированного СП:

(1) универсальной обработки GPPL (general-purpose processing layer);

(2) ускоренной пакетной обработки APPL (accelerated packet processing layer);

(3) быстрого пакетного ввода-вывода EPIL (express packet input-output layer).

Макроархитектура не определяет систему инструкций (instruction set architecture, ISA) программируемых процессоров слоя GPPL, конкретный набор функций в слое APPL и интерфейсы слоя EPIL. Она дает самое общее представление о функциональности компонентов СП, но не конкретизирует их работу и взаимодействие друг с другом. С этой точки зрения интересен предложенный компанией LSI Corp. метод организации работы СП в виде «виртуального конвейера», предложенный в 2010 г. для семейства СП AXXIA. Суть метода в том, что входящему в СП пакету после первичной классификации (в слое EPIL в категориях Layerscape) назначается определенный маршрут обработки, по которому он движется, как по конвейеру, переходя от одного блока СП к другому, причем этими блоками могут быть как программируемые процессоры, так и аппаратные ускорители (т. е. соответственно компоненты слоев GPPL и APPL в Layerscape). Из блока в блок пакет попадает через иерархические очереди в соответствии с требованиями качества обслуживания для данного маршрута [3, 4].

Наряду с компаниями Freescale и LSI крупным поставщиком СП с традиционной ISA процессорных ядер была компания Cavium Networks. Ее СП были буквально нашпигованы десятками программируемых процессоров с ISA MIPS или ARM, всевозможными сопроцессорами и аппаратными ускорителями, а также разнообразными сетевыми интерфейсами [5].

Не обладая возможностями конкурировать с крупными компаниями в количественном отношении, небольшие фирмы делали ставку на качественно отличные, инновационные архитектурные подходы и фирменные ноу-хау, которые должны были существенно улучшить удельные характеристики СП и тем самым нивелировать исходные конкурентные преимущества старожилов рынка микроэлектроники. Как правило, ставка делалась на интеграцию большого числа предельно упрощенных процессорных ядер со специализированными ISA, узко ориентированными на типовые сетевые функции плоскости данных (data plane),

оставляя задачи плоскости управления (control plane) подключаемому универсальному host-процессору.

Характерен в этом плане пример компании Netronome. Ее СП семейства NFP могли включать до 96 ядер, специализированных под обработку пакетов, и до 120 ядер для обработки их потоков [6]. В СП также интегрировались функциональные ускорители табличного поиска, криптографии, управления очередями, балансировки нагрузки и др. Все это дополнялось современными высокоскоростными сетевыми интерфейсами. Для подключения host-процессоров предусматривались интерфейсы PCIe (peripheral component interconnect express).

Другой пример аналогичного подхода — семейство СП TILE-Gx компании Tilera Corp. В основу семейства заложена ортогональная матрица (8×9 в старшей модели TILE-Gx72) «плиток» (tiles) — процессорных ядер с оригинальной ISA, — объединяемых фирменной сетью iMesh, которая служит главным отличием данной архитектуры [7]. Эта конфигурируемая сеть позволяет фрагментировать участки матрицы и организовывать параллельную или конвейерную работу «плиток» и их фрагментов. В СП включены ускорители шифрования. Интегрированные сетевые интерфейсы подключаются через специализированный пакетный процессор. Для связи с host-процессором также используются интерфейсы PCIe.

Наконец, к этой же категории СП, вероятно, относятся высокоинтегрированные приборы семейства nPower одного из лидеров рынка сетевых устройств, компании Cisco Systems. Кроме рекордного числа программируемых процессоров — до 672 ядер с фирменной ISA — и наличия аппаратных ускорителей, прочие подробности архитектуры компанией не разглашаются [8]. В этом и нет прямой необходимости, так как компания использует свои СП сугубо в собственных продуктах. (Недавно были анонсированы пробные поставки СП nPower для Facebook и Microsoft [9].)

Рассмотренные выше архитектуры со специализированными ISA реализуют параллелизм вычислений тривиальным тиражированием одинаковых упрощенных процессорных ядер. Они, как правило, предполагают всю обработку пакета от начала до конца на одном процессоре с привлечением при необходимости интегрированных аппаратных ускорителей, причем механизмы распределения пакетов между процессорами и вовлечения ускорителей обычно не раскрываются. Однако в противовес такой прямолинейной реализации параллелизма предлагались и более изощренные подходы.

Один из них продемонстрировала компания EZchip Semiconductor [10]. Ее архитектура «ориентированных на задачи процессоров» (task-oriented processors, TOP) базируется на коротком конвейере обработки пакетов из четырех функционально специализированных ступеней и, соответственно, четырех типах процессорных ядер с различными ISA. Каждая ступень допускает масштабирование «вширь» и, как следствие, теоретически неограниченное наращивание производительности СП. Очевидный недостаток архитектуры — необходимость

для каждого типа обработки четырех программ, написанных на языках низкого уровня в четырех разных ISA. Вследствие оригинальности архитектуры вся информация о ней и все средства разработки — компиляторы, библиотеки, симуляторы — безальтернативно доступны только от владельца ноу-хау, вследствие чего разработка ПО для СП не только усложнена нетривиальностью архитектуры, но и принципиально ограничена по числу возможных участников, т. е. данный подход в значительной степени лишает СП главного преимущества — гибкости за счет удобной и доступной пользователям программируемости.

Попытка преодолеть этот недостаток была сделана в архитектуре СП другой небольшой фирмы — Xelerated [10]. Здесь вместо короткого конвейера с функционально ориентированными ступенями предлагается длинный конвейер с концептуально идентичными ступенями, на которых последовательно выполняются инструкции программы: i-я инструкция на i-й ступени. Таким образом, число ступеней в конвейере определяется числом инструкций в самой длинной программе. Хотя писать программу по-прежнему приходится на фирменном языке низкого уровня и отлаживать на фирменном симуляторе, вместо четырех программ пишется только одна для некоего «виртуального процессора», при этом программист не озабочен параллелизмом, который автоматически реализуется во время исполнения программы на конвейере, обрабатывающем одновременно столько пакетов, сколько в нем ступеней. Особенности подхода: одинаковое время обработки любого пакета, пропорциональное длине конвейера и не зависящее от типа обработки, и постоянная гарантированная пропускная способность, определяемая только частотой работы конвейера. Основной недостаток архитектуры кроется в больших аппаратных затратах, так как в исходной концепции все ступени должны быть функционально эквивалентны «виртуальному процессору». При длине программы в сотни инструкций степень интеграции СП оказывается эквивалентной сотням «виртуальных процессоров». Эту проблему Xelerated частично решает оптимизацией программ и, нарушая изначальную концепцию, функциональной специализацией ступеней конвейера.

3 Современные тенденции в развитии сетевых процессоров

Сетевые процессоры со стандартными ISA эволюционировали в основном количественно. Прогресс технологии позволял размещать на кристалле все больше различных компонентов СП. Характерным трендом в СП этой категории стало их «ARMирование» [11]. Как видно из рис. 1, на ISA ARM в своих СП перешли не только приверженцы других стандартных ISA, но и некоторые апологеты специализированных архитектур: полностью на ARM мигрировала компания AMCC, перевод своих «плиток» на ARM планировала (но не успела выполнить) Tilera. Превалирование архитектуры ARM стало следствием как ее энергоэффективности, так и гибкой лицензионной политики патентообладателя. В итоге ARM победила, как видно на примере компании Cavium, даже архитектуру MIPS, хотя с чисто технической точки зрения последняя более эффективна по энерго-

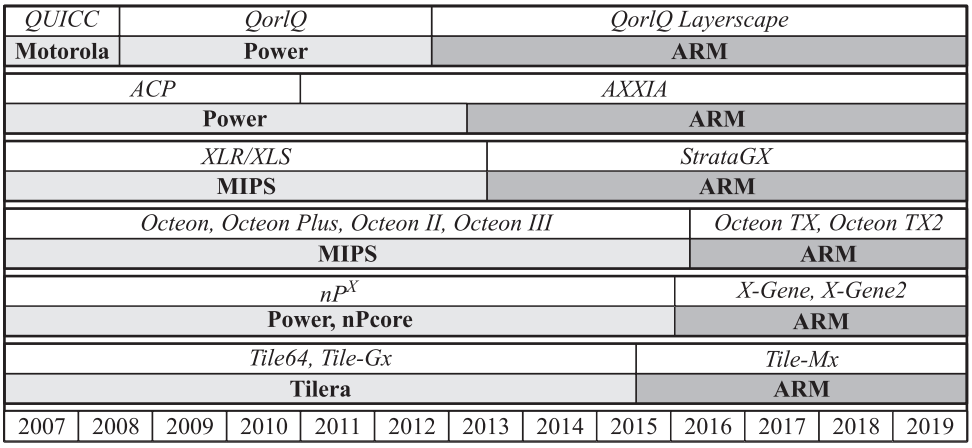


Рис. 1 «ARMирование» СП

потреблению и занимаемой на кристалле площади [12]. В свою очередь, массовое «ARMирование» СП повлекло за собой практически полный отказ в них от многопоточности в русле технической политики лицензиара архитектуры [13].

Параллельно эволюции СП со стандартными ISA программистское сообщество нарабатывало под эти ISA все бóльшие объемы ПО как системного, так и прикладного, в том числе общедоступного с открытыми кодами. Многое из этих наработок могло быть использовано в СП. В результате, хотя по удельным параметрам СП со стандартными ISA уступали своим специализированным конкурентам, потребители СП готовы были немного переплатить за кремний, чтобы гораздо больше выиграть на стоимости разработки ПО для своих продуктов и сокращении времени их вывода на рынок.

Поскольку СП с нетрадиционной архитектурой и нестандартными ISA разрабатывались и продвигались на рынок в основном небольшими фирмами, их общая «ахиллесова пята» — сложность программирования на специальных языках низкого уровня — усугублялась недостатком фирменных ресурсов и ограниченностью круга разработчиков ПО. В 2010-х гг. эти фирмы одна за другой прекращают самостоятельную деятельность, либо будучи поглощенными мощными компаниями, либо добровольно влившись в них, что иллюстрирует рис. 2. Но, что характерно, новыми владельцами ноу-хау неудачников часто оказывались отнюдь не их конкуренты из лагеря СП со стандартными ISA, а сторонние игроки. Будучи более практичными, новые обладатели ноу-хау предпочитали использовать приобретения для внутрифирменных нужд, следуя примеру компании Cisco. Обладая бóльшими возможностями, они могли себе позволить содержать штат программистов, создававших и поддерживавших специфическое встроенное ПО (firmware) для ограниченного круга интеллектуальных фирменных продуктов.

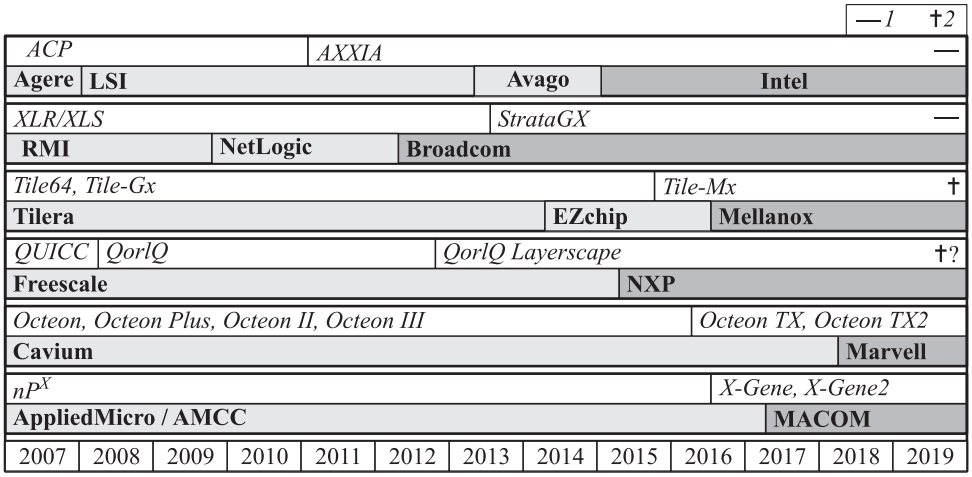


Рис. 2 Смена правообладателей архитектур СП: *1* — только для внутрифирменного использования; *2* — выпуск СП прекращен

В середине 2010-х гг. определенный кризис СП проявился отмеченной выше чередой поглощений связанных с этим бизнесом компаний. Число поставщиков СП сократилось, спектр приборов заметно обеднел. Если раньше Freescale регулярно (раз в пару лет) выпускала новое семейство СП с несколькими моделями, то после слияния ее в 2015 г. с NXP объединенная компания анонсировала лишь один-единственный СП QorIQ LX2160A, разработка которого была проведена еще до слияния. Компания Cavium, разработавшая целый ряд семейств СП Octeon, с 2014 г. переносит акцент своих усилий на универсальные многоядерные серверные процессоры ThunderX. Это направление становится у Cavium доминирующим, а после поглощения ее в 2018 г. компанией Marvell, похоже, единственным. Тот же путь прошла компания AMCC. Выпускавшая СП довольно популярной серии nP^x, в 2013 г. она анонсирует серверные процессоры X-Gene, а после ее поглощения в 2017 г. компанией MACOM выпуск СП окончательно прекращается. Множащиеся примеры переориентации прежних поставщиков СП на многоядерные серверные процессоры заставляют говорить о тенденции.

Даже на пике популярности СП аппаратные решения по продвижению пакетов сохранялись в технологиях, ориентированных на быструю коммутацию, таких как ATM или MPLS (multiprotocol label switching). Поэтому, хотя формально в узлах сетей MPLS предполагаются маршрутизаторы, реально непосредственно коммутация по меткам в них осуществляется аппаратно согласно простой таблице, которую формирует маршрутизирующий процессор, поддерживающий требуемые протоколы маршрутизации, распределения меток и пр.

Такое явное физическое разделение функций плоскостей данных и управления нашло наиболее полное выражение в программно определяемой организации сетей (software-defined networking, SDN). Пожалуй, именно SDN, вкупе с виртуализацией сетевых функций (network functions virtualization, NFV), нанесла по СП самый болезненный удар.

В исходной концепции SDN все функции плоскости управления сосредоточены в логически едином контроллере сети, а плоскость данных обслуживается простейшими коммутаторами, для которых правила коммутации задаются контроллером сети. В качестве такого контроллера предполагается типовой сервер (или виртуальная машина на нем) со стандартным системным и специальным прикладным ПО, реализующим все функции управления сетью. В такой рафинированной концепции SDN для СП просто нет места. Но для SDN-контроллеров потенциально нужны высокопроизводительные серверы, построенные на основе универсальных процессоров с большим числом ядер, и это, вероятно, объясняет отмеченную выше тенденцию к миграции бывших поставщиков многоядерных СП в область универсальных серверных процессоров. По крайней мере, практически все производители таких «перелицованных» СП обозначают SDN в качестве одной из основных областей применения своих продуктов.

Стоит также заметить, что в их рекламе рядом с SDN практически всегда фигурирует NFV. Таким образом, если изначально СП программно реализовывали типовые сетевые функции, то теперь, «перелицованные» в универсальные многоядерные процессоры, они, словно цепляясь за уплывающий от них сетевой рынок, пытаются реализовывать все те же привычные сетевые функции, но уже в виртуальном пространстве.

4 Заключение. Перспективы сетевых процессоров

За время существования СП сетевые технологии совершили несколько витков развития. Программная реализация сетевых функций то дополнялась аппаратным продвижением пакетов, то вновь возвращалась к преимущественно программной коммутации. Постоянно расширялся спектр сетевых протоколов, росла значимость мер обеспечения сетевой безопасности, соответственно менялись роль и значение СП. Технология СБИС позволила объединить в одном микроэлектронном приборе — интегрированном СП — множество обрабатывающих устройств и сетевых интерфейсов. «Золотой век» интегрированных СП пришелся на начало нынешнего тысячелетия, когда массово появлялись не только новые приборы, но и инновационные архитектуры со специализированными ISA. Однако последние в итоге проиграли в противостоянии со стандартными, в частности с ISA ARM.

В настоящее время наметилась отчетливая тенденция к сокращению числа фирм-поставщиков СП вследствие их слияний и поглощений, а также к сужению области применимости самих приборов. Новые обладатели ноу-хау массово

переориентируются с СП на универсальные серверные процессоры, перенося туда опыт агрессивной многоядерности СП.

В наступившую эру «облачных» технологий потребность в витающих в «облаках» сверхвысокопроизводительных серверах будет продолжать расти быстрыми темпами. Открытым, однако, остается вопрос, смогут ли «перелицованные» из СП серверные процессоры, в абсолютном большинстве своем «ARMированные», конкурировать в качестве основы таких серверов с давно захватившими этот сегмент рынка процессорами архитектуры x86 от Intel и AMD. Но, вероятно, независимо от исхода этой схватки СП будут в ней не более чем наблюдателями с обочины мейнстрима в качестве всего лишь внутрифирменных инструментов интеллектуализации ограниченного числа продуктов в узком кругу крупных микроэлектронных фирм.

Литература

1. *Serpanos D., Wolf T.* Specialized hardware components // Architecture of network systems. — The Morgan Kaufmann ser. in computer architecture and design. — Elsevier Inc., 2011. P. 211–227.
2. The basics of network processors, 2004. https://www.embedded.com/the-basics-of-network-processors.
3. *Егоров В. Б.* Современные тенденции в развитии архитектур интегрированных сетевых процессоров // Системы и средства информатики, 2014. Т. 24. № 3. С. 78–90.
4. *Егоров В. Б.* Способ организации обработки пакетов в интегрированных сетевых процессорах // Системы и средства информатики, 2017. Т. 27. № 1. С. 108–121.
5. OCTEON III CN7XXX Multi-Core MIPS64 Processors, 2020. https://www.marvell.com/products/infrastructure-processors/multi-core-processors/octeon-multi-core-mips64-processors/octeon-iii-cn7xxx.html.
6. Netronome NFP-6000 Flow Processor, 2018. https://www.netronome.com/m/documents/PB_NFP-6000-7-20.pdf.
7. TILE-Gx72 processor // Mellanox Technologies, 2015–2016. https://www.mellanox.com/sites/default/files/related-docs/prod_multi_core/PB_TILE-Gx72.pdf.
8. *Markevitch J., Malladi S.* A 400 Gbps multi-core network processor, 2017. https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.80-Architectdure-Pub/HC29,22,810-400gbs-NPU-Markevitch-Cisco.pdf.
9. *Clarke P.* Cisco enters chip market with network processor // eeNews Analog, 2019. https://www.eenewsanalog.com/news/cisco-enters-chip-market-network-processor#.
10. *Егоров В. Б.* Современные интегрированные сетевые процессоры: архитектура, возможности, средства разработки. — М.: ФИЦ ИУ РАН, 2016. 103 с.
11. *Егоров В. Б.* Тенденция к ARMированию многоядерных интегрированных сетевых процессоров // Электронные компоненты, 2018. № 6. С. 40–44.
12. *Егоров В. Б.* Обзор и сравнение процессорных ядер ARM Cortex и MIPS Aptive // Электронные компоненты, 2013. № 6. С. 64–69.

13. *Егоров В. Б.* Особенности многоядерности и многопоточности в сетевых процессорах // Системы и средства информатики, 2020. Т. 30. № 1. С. 82–92.

Поступила в редакцию 31.08.20

EVOLUTION OF NETWORK PROCESSORS

V. B. Egorov

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

Abstract: Network processors (NP) have passed a long evolution way from general-purpose computers equipped with network interfacing cards to highly integrated on-chip systems comprising dozens of programmable processors, hardware accelerators, and network interfaces. The high performance was reached in the NPs on two differing directions: old microelectronic firms scaled up the number of processing cores with a conventional architecture, while small young companies invented specialized approaches. The second direction provided better relative characteristics, but the small firms were not competitive due to complexities of programming in their nonconventional architectures and paucity of resources to create comprehensive NP software. Ultimately, almost all of them left recently the NP market, having been absorbed, in one way or another, by some larger companies. Herewith, the latter utilized the acquired technologies mostly not for expansion into the NP market so much as for applying the multicore experience to conquer the high-performance server processors market which rapidly grows in the nowadays' era of cloud technologies. Thus, the NPs are actually ceasing to exist as separate products and turning into an intrafirm tool for intellectualization of a restricted circle of some specific network devices.

Keywords: network processor; integrated network processor; processor core architecture; high-performance multicore server

DOI: 10.14357/08696527210109

References

1. Serpanos, D., and T. Wolf. 2011. Specialized hardware components. *Architecture of network systems*. The Morgan Kaufmann ser. in computer architecture and design. Elsevier Inc. 211–227.
2. The basics of network processors. 2004. Available at: https://www.embedded.com/the-basics-of-network-processors/ (accessed February 17, 2021).
3. Egorov, V. B. 2014. Sovremennye tendentsii v razvitii arkhitektur integrirovannykh setevykh protsessorov [Modern trends in evolution of integrated network processor architectures]. *Sistemy i sredstva informatiki — Systems and Means of Informatics* 24(3):78–90.

4. Egorov, V. B. 2017. Sposob organizatsii obrabotki paketov v integrirovannykh setevykh protsessorakh [A method of packet processing in integrated network processors]. *Sistemy i sredstva informatiki — Systems and Means of Informatics* 27(1):108–121.
5. OCTEON III CN7XXX Multi-Core MIPS64 Processors. 2020. Available at: https://www.marvell.com/products/infrastructure-processors/multi-core-processors/octeon-multi-core-mips64-processors/octeon-iii-cn7xxx.html (accessed February 17, 2021).
6. Netronome NFP-6000 Flow Processor. 2018. Available at: https://www.netronome.com/m/documents/PB_NFP-6000-7-20.pdf (accessed February 17, 2021).
7. TILE-Gx72 processor. 2015–2016. Available at: https://www.mellanox.com/sites/default/files/related-docs/prod_multi_core/PB_TILE-Gx72.pdf (accessed February 17, 2021).
8. Markevitch, J., and S. Malladi. 2017. A 400 Gbps multi-core network processor. Available at: https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.80-Architectdure-Pub/HC29,22,810-400gbs-NPU-Markevitch-Cisco.pdf (accessed February 17, 2021).
9. Clarke, P. 2019. Cisco enters chip market with network processor. *eeNews Analog*. Available at: https://www.eenewsanalog.com/news/cisco-enters-chip-market-network-processor# (accessed February 17, 2021).
10. Egorov, V. 2016. *Sovremennye integrirovannye setevye protsessory: Arkhitektura, vozmozhnosti, sredstva razrabotki* [Modern integrated network processors: Architecture, facilities, and design tools]. Moscow: FRC CSC RAS. 103 p.
11. Egorov, V. 2018. Tendentsiya k ARMirovaniyu mnogoyadernykh integrirovannykh setevykh protsessorov [A trend to ARMing of multicore integrated network processors]. *Elektronnye komponenty* [Electronic Components] 6:40–44.
12. Egorov, V. B. 2013. Obzor i sravnenie protsessornykh yader ARM Cortex i MIPS Aptive [A review and comparison of the processor cores ARM Cortex and MIPS Aptive]. *Elektronnye Komponenty* [Electronic Components] 6:64–69.
13. Egorov, V. B. 2020. Osobennosti mnogoyadernosti i mnogopotochnosti v setevykh protsessorakh [Multicore and multithreading peculiarities in network processors]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 30(1):82–92.

Received August 31, 2020

Contributor

Egorov Vladimir B. (b. 1948) — Candidate of Science (PhD) in technology, 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; VEgorov@ipiran.ru

МЕТОДЫ СРАВНЕНИЯ КОНКУРИРУЮЩИХ ГИПОТЕЗ В ГИПОТЕЗООРИЕНТИРОВАННЫХ СИСТЕМАХ*

***Е. М. Тириков*[1], *Д. Ю. Ковалев*[2]**

Аннотация: С появлением класса систем управления виртуальными экспериментами использование гипотез и моделей в явном виде становится все более распространенным. В рамках подобных систем используются как порожденные из данных гипотезы, так и теоретические гипотезы. При этом критически важным становится сравнение между собой нескольких конкурирующих гипотез разной природы. В работе рассмотрены различные подходы к сравнению конкурирующих гипотез и реализующих их вычислительных моделей между собой. Рассмотренные подходы реализованы в виде программного компонента, являющегося частью системы управления виртуальными экспериментами. Применение компонента проиллюстрировано на задаче поиска различий функциональной связности областей головного мозга в состоянии покоя у мужчин и женщин.

Ключевые слова: системы управления виртуальными экспериментами; конкурирующие гипотезы; сравнение гипотез

DOI: 10.14357/08696527210110

1 Введение

Исследования в областях с интенсивным использованием данных породили новый класс систем по управлению виртуальными экспериментами [1]. В подобных системах в явном виде используется набор взаимодействующих гипотез и реализующих их моделей. Для объяснения одного и того же природного явления часто могут быть предложены несколько конкурирующих между собой гипотез. Важным становится использование корректных механизмов сравнения гипотез между собой, а также обеспечение ранжирования гипотез по выбранной экспертом метрике, вычисленной на данных наблюдений.

Сложность сравнения гипотез между собой заключается в широте класса моделей, реализующих представленные гипотезы. Часть моделей относится к регрессионным и строится непосредственно на данных наблюдений без добавления теоретических знаний из предметной области. Зависимости в моделях могут быть как линейными, так и нелинейными, а сами уравнения могут содержать частные

* Работа выполнена при финансовой поддержке РФФИ (проекты 18-07-01434 и 18-29-22096).

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, em.tirikov@gmail.com

[2] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, dkovalev@ipiran.ru

производные или интегральную часть [2]. Построенные из данных регрессионные модели, в отличие от теоретических, могут иметь большое число параметров и быть плохо интерпретируемыми. Часть моделей относится к теоретическим, обычно представленным только системами уравнений и выделенным из обзоров литературы предметной области. Идет активное развитие гибридных моделей, сочетающих элементы как теоретических моделей, так и моделей, построенных непосредственно на данных наблюдений.

Данная работа выполняется в рамках проекта, направленного на разработку прототипа гипотезоориентированной системы управления виртуальными экспериментами [1]. Виртуальный эксперимент содержит описание гипотез, моделей, потока работ, а также онтологии предметной области. Сравнение гипотез является важной частью системы. В работе рассмотрены различные подходы к сравнению гипотез и реализующих их моделей между собой. Приведено описание программного компонента системы, реализующего сравнение, позволяющее не только выделить лучшую из гипотез, но и ранжировать модели, реализующие гипотезы, по качеству предсказания на выбранном наборе данных с целью предоставления нескольких лучших гипотез эксперту для последующей более тщательной проверки. Таким образом, использование описываемого компонента при анализе виртуальных экспериментов позволяет сократить множество рассматриваемых гипотез и реализующих их моделей в несколько раз.

Примером области с интенсивным использованием данных, где требуется применение подходов к сравнению гипотез, служит нейроинформатика. Практическое применение компонента демонстрируется на примере задачи поиска различий функциональной связности областей головного мозга мужчин и женщин на основании данных функциональной магнитно-резонансной томографии в состоянии покоя [3].

Работа структурирована следующим образом. В разд. 2 приводится обзор существующих подходов к сравнению гипотез. В разд. 3 представлено описание компонента сравнения гипотез и соответствующих моделей. В разд. 4 представлена демонстрация подходов к сравнению гипотез в задаче поиска различий в функциональной связности областей головного мозга у мужчин и женщин в состоянии покоя.

2 Методы сравнения конкурирующих гипотез

В данном разделе рассмотрены подходы к сравнению гипотез и соответствующих им моделей, использованные при реализации компонента системы по управлению виртуальными экспериментами. Описание компонента приведено в следующем разделе. Сравнение конкурирующих гипотез выполняется на основе сравнения между собой реализующих их моделей.

Существует множество подходов к сравнению различных вычислительных моделей. Один из самых простых подходов [4, 5] — *упорядочение моделей по выбранной метрике* на тестовых данных, например среднеквадратичной ошибке или

коэффициенту детерминации R^2. Преимущество данного подхода заключается в его простоте, интерпретируемости и возможности сравнения моделей различных типов. Этот подход в основном применяется для отсечения худших моделей с целью уменьшения в дальнейшем числа попарных сравнений моделей, а также для ранжирования моделей по выбранной метрике.

При подходе с *использованием статистической проверки гипотез* формулируются нулевая и альтернативная гипотезы, экспертом выбирается уровень значимости, задается статистика с известным распределением при верности нулевой гипотезы. В зависимости от проверяемой гипотезы экспертом выбирается статистический тест и вычисляется достигаемый уровень значимости. По результатам сравнения достигаемого уровня значимости с изначально выбранным значением отвергается или не отвергается нулевая гипотеза. Данный способ сравнения наиболее распространен для моделей, линейно зависящих от своих параметров [4, 6]. В работе [7] для сравнения моделей на схожесть используются критерий Уилкоксона и критерий знаков, проверяющие гипотезу о равенстве нулю медианы разности предсказаний двух рассматриваемых моделей. В работе [3] критерий знаков используется для попарного сравнения двух множеств нелинейных моделей.

Подход к сравнению различных вычислительных моделей *с использованием методов теории информации* дает возможность оценить не только качество предсказания модели, но также и степень ее переобучения. Одним из самых распространенных методов является информационный критерий Акаике (AIC, Akaike's Information Criterion) [8, 9]. Критерий подходит для сравнения не только моделей, построенных на экспериментальных данных, но и теоретических моделей. Лучшей из нескольких моделей считается та, которая имеет наименьшее значение AIC. Недостаток данного метода кроется в плохой применимости на выборках малого размера. Преодолеть данный недостаток позволяет использование модифицированного информационного критерия Акаике, являющегося менее общим и применимого при условии, что модель линейно зависима от своих параметров. Другой информационный критерий носит имя Байеса (BIC, Bayesian Information Criterion) [10]. Его основное отличие от критерия Акаике состоит в том, что накладывается более строгое ограничение на число настраиваемых параметров. Для некоторых моделей, например построенных при помощи генетического программирования, рекомендуется использовать модифицированные критерии Акаике и Байеса, учитывающие особенности применяемого алгоритма [11]. Подход с использованием информационных критериев применяется для сравнения моделей разной природы, а также для отсечения моделей большого размера.

Идея *байесовского подхода* [12] к сравнению вычислительных моделей заключается в переходе от априорных знаний к апостериорным с учетом наблюдаемых данных. Вначале вычисляется разность предсказаний двух моделей. Затем проверяется гипотеза о том, что среднее вычисленной разности равно нулю. Преимущество байесовского подхода заключается в возможности определения

степени уверенности в проверке гипотезы (доказательная сила неубедительная; слабое, среднее, сильное доказательство).

Подход к сравнению вычислительных моделей на основе *поиска схожести графового представления моделей* [13] предусматривает сравнение сходства двух ориентированных ацикличных графов, построенных на основе систем уравнений, задающих модели. Строится разностный ориентированный ацикличный граф. Данный метод отличается от других тем, что с его помощью можно сравнивать множества гипотез в совокупности. Его преимущество заключается в возможности аппроксимации разностного графа без явного построения изначальных графов. Недостатком данного подхода является экспоненциальная вычислительная сложность построения разностного графа и невозможность учета параметризующих переменных. Подход применяется, когда модели представимы в виде ацикличных графов.

3 Компонент сравнения гипотез и соответствующих им моделей

В данном разделе приведено описание компонента[1] для сравнения гипотез и соответствующих им моделей. Компонент представляет собой совокупность функций на языке Python, реализующих подходы, рассмотренные в разд. 2.

Функция range_models(models, dataset, metrics, threshold) предназначена для ранжирования моделей по выбранной метрике на выбранном наборе данных. На вход функция принимает набор моделей, которые требуется ранжировать, набор данных, на котором строится выбранная метрика, метрику предсказаний, порог значений для отсечения плохих моделей. В качестве метрик предсказаний могут быть использованы метрики из библиотеки Sklearn, например среднеквадратичная ошибка или коэффициент детерминации, а также информационные критерии, определенные в виде функций ниже. На выходе функция возвращает ранжированный список моделей с вычисленными метриками.

Фукции get_AIC(model, dataset) и get_BIC(model, dataset) осуществляют вычисление информационного критерия Акаике и байесовского информационного критерия соответственно для выбранной модели и набора данных. На вход функции принимают модель и набор данных, на выходе возвращают значение информационных критериев. Функции get_AIC_nonlinear(model, dataset) и get_BIC_nonlinear(model, dataset) осуществляют вычисление критериев для нелинейных моделей.

Функция update_bayesian_probability(model, dataset, prior_probability) осуществляет вычисление байесовской вероятности для выбранной модели на выбранном наборе данных с учетом заданной априорной оценки вероятности. Функция compute_bayesian_hypothesis_score(model, dataset) осуществляет вычисление байесовской оценки соответствия предсказаний модели наблюдениям.

[1] https://github.com/dmkovalev/hypothesis_platform/blob/main/core/comparison/compare_models.py.

Функция compare_models(models, dataset, stat_test) осуществляет сравнение двух моделей между собой на одном наборе данных. На вход функция принимает две модели, набор данных, а также выбранный экспертом статистический тест. На выходе метод возвращает истину, если нельзя отбросить гипотезу о различии предсказаний моделей, иначе — ложь. Функция compare_preds_on_different_datasets(models, dataset_1, dataset_2, stat_test) осуществляет аналогичное сравнение двух моделей между собой, но уже на двух наборах данных. На вход функция принимает модели, два набора данных, а также выбранный экспертом статистический тест. На выходе функция возвращает истину, если результат сравнения моделей одинаков для обоих наборов данных.

Функция compute_diff_dag(models, dataset) предназначена для вычисления графа разности двух ориентированных ацикличных графов, построенных по данным входных моделей. На входные модели налагается дополнительное ограничение о линейности. На выходе функция возвращает вычисленный граф разности моделей. Если множество ребер графа пустое, то модели считаются схожими.

4 Сравнение конкурирующих гипотез для задачи поиска различий в функциональной связности областей головного мозга у мужчин и женщин

В данном разделе приводятся результаты сравнения различных гипотез и соответствующих им моделей для задачи поиска различий в функциональной связности областей головного мозга у мужчин и женщин в состоянии покоя. Постановка и описание задачи приведены в работе [3]. Для поставленной задачи были сформулированы гипотезы и реализующие их модели (рис. 1). Слева приведен граф зависимости нескольких гипотез между собой. Вершиной графа служит набор конкурирующих между собой гипотез. Конкурирующими являются гипотезы, описывающие один и тот же естественный феномен и реализованные разными вычислительными моделями. Приведенные методы сравнения работают для конкурирующих гипотез, для зависимых гипотез они не применимы.

Для данной задачи выделены три множества конкурирующих гипотез: выбор атласов и выделение регионов головного мозга человека, построение функциональной связности регионов головного мозга, поиск значимых различий в функциональной связности у мужчин и женщин в состоянии покоя.

Для первого множества нет единого способа выбрать ту или иную модель. В работе [14] сравнивается множество различных методов для выделения регионов. Отмечается, что модели не полностью сравнимы, поэтому окончательный выбор остается за экспертом. В данной работе используется атлас Harvard–Oxford [15], с помощью которого выделяются 48 регионов кортикальной части головного мозга.

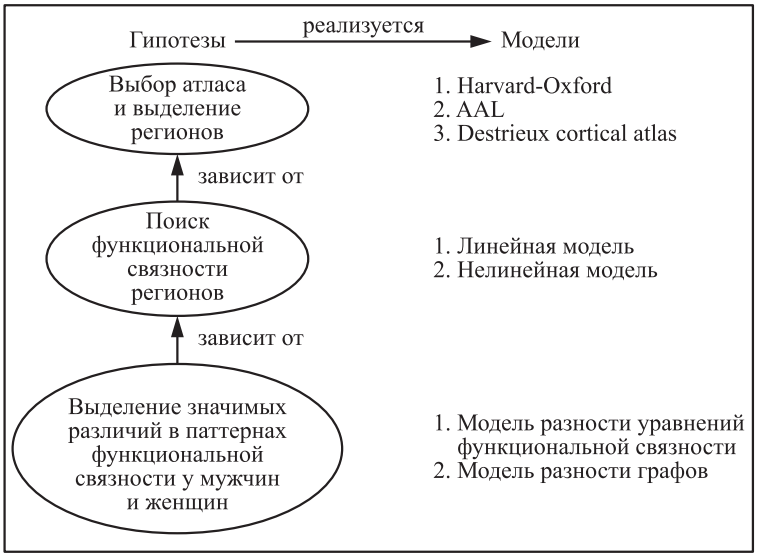


Рис. 1 Гипотезы и модели для задачи поиска различий в функциональной связности регионов головного мозга у мужчин и женщин в состоянии покоя

Для второго множества конкурирующих гипотез о функциональной связности регионов головного мозга используются несколько методов. Первым шагом для данного сравнения служит отсечение моделей по пороговому значению коэффициента детерминации. Это делается для сокращения пространства сравниваемых моделей. Из 48 регионов для 23 построены модели с оценкой выше порогового коэффициента детерминации, равного 0,7. Вторым шагом является сравнение линейных и нелинейных моделей, где нелинейные модели были получены методом генетического программирования. Для сравнения моделей используются критерии Акаике и Байеса.

Результаты сравнения с использованием информационных критериев Акаике и Байеса приведены в табл. 1. Здесь и далее результаты приводятся только для регионов Precentual Gyrus, Postcentral Gyrus и Planum Polare. Для критерия

Таблица 1 Информационные критерии

Регион	Критерий Акаике		Критерий Байеса	
	Линейная модель	ГП	Линейная модель	ГП
Precentual Gyrus	−53 058	**−58 638**	**−54 987**	−37 821
Postcentral Gyrus	−46 048	**−52 315**	**−42 874**	−34 508
Planum Polare	−7 109	**−10 658**	**−6 925**	−3 215

Акаике лучшие результаты показала нелинейная модель, в то время как для критерия Байеса лучшей оказалась линейная модель.

Несоответствие между двумя информационными критериями вызвано тем, что в критерии Байеса учитываются не только точность предсказания, но также и сложность проверяемой модели. В модели генетического программирования (ГП) число параметров значительно больше, чем в обычной линейной модели, поэтому у критерия Байеса оценка модели ГП выше, чем у линейной. Если требуется не только точность предсказаний, но и простота модели, то следует использовать критерий Байеса, иначе — критерий Акаике.

Третье множество конкурирующих гипотез о поиске значимых различий в функциональной связности регионов головного мозга у мужчин и женщин в состоянии покоя состоит из двух конкурирующих гипотез (моделей). Первая модель представляет собой разность между предсказаниями для мужчин и женщин в выбранном регионе. Вторая модель строится вычитанием графов, построенных для мужчин и женщин, каждый из которых описывает взаимодействие между регионами головного мозга. Данные гипотезы не являются сравнимыми напрямую, а скорее являются взаимодополняющими. Если полученные регионы со значимыми различиями не совпадают, то эксперту рекомендуется дополнительно проверить данный регион, например используя другой набор данных.

Для первой модели проверяется равенство медианы разности нулю с использованием статистического теста — критерия знаков [16], а также байесовского подхода. Результаты для трех регионов представлены в табл. 2.

Уровень значимости выбран равным 0,05. Регионы Precentual Gyrus и Postcentral Gyrus оказались разными для мужчин и женщин. Для региона Planum Polare нулевая гипотеза о схожести регионов не отвергается. Важное замечание: хотя нулевая гипотеза не отвергается, но полученное значение близко к порогу 0,05. Если выбрать значение порога, например 0,1, то нулевая гипотеза о схожести региона будет отвергнута и для Planum Polare.

Результаты сравнения с применением байесовского подхода представлены в табл. 3. Результаты совпадают с аналогичными результатами для статистического теста, при этом появляется возможность оценить степень уверенности в результате.

Для регионов Precentual Gyrus и Postcentral Gyrus нулевая гипотеза схожести этих регионов у мужчин и женщин отвергается со средней степенью уверенности.

Таблица 2 Результаты сравнения двух моделей ГП, обученных на разных выборках

Регион	Достигаемый уровень значимости
Precentual Gyrus	0,0045 (отвергается)
Postcentral Gyrus	0,0003 (отвергается)
Planum Polare	0,08 (не отвергается)

Таблица 3 Результаты сравнения байесовским подходом

Регион	η	Доказательная сила
Precentual Gyrus	29 : 1 (29)	Среднее доказательство
Postcentral Gyrus	54 : 1 (54)	Среднее доказательство
Planum Polare	7 : 1 (0,14)	Слабое доказательство

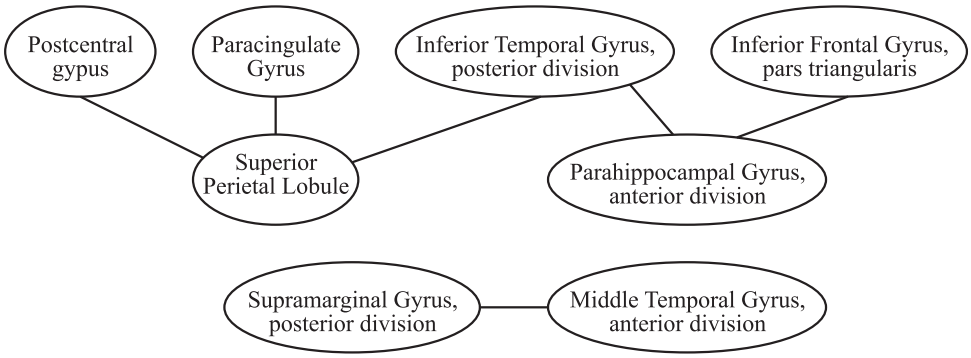


Рис. 2 Граф разности для мужчин и женщин

При этом для Planum Polare нулевая гипотеза не отвергается со слабой степенью уверенности, что говорит о необходимости дополнительного исследования данного региона. В целом результаты статистического и байесовского подхода схожи, но иногда они могут давать разные результаты.

Для второй модели строится разность графов связности для мужчин и женщин (рис. 2). Вершинами построенного графа разности служат регионы мозга, а ребрами — связи между регионами. Если функциональная связность региона различна у мужчин и женщин, то в графе разности присутствует хотя бы одно ребро для данного региона. Для региона Postcentral Gyrus в графе есть такое ребро. Также для этого региона нулевая гипотеза об отсутствии различий отвергается по обоим критериям. Для региона Planum Polare нет ребер, и гипотеза об отсутствии различий не отвергается по обоим критериям. Для региона Precentual Gyrus по статистическим критериям гипотеза отвергается, но ребра в графе отсутствуют, таким образом, результат сравнения с использованием разных методов отличается, при этом эксперту рекомендуется провести дополнительные исследования данного региона.

5 Заключение

Развитие систем поддержки виртуальных экспериментов привело к необходимости явного сравнения и ранжирования гипотез и соответствующих им моделей.

В статье рассмотрены подходы к сравнению гипотез и реализующих их вычислительных моделей. Подходы реализованы в компоненте сравнения гипотез, который является одной из составных частей системы управления виртуальными экспериментами. Сравнение моделей демонстрируется на решении задачи из области нейрофизиологии.

Литература

1. *Kovalev D., Tarasov E.* Virtual experiments in data intensive research // Информатика и её применения, 2019. Т. 13. Вып. 2. С. 117–125.
2. *Kiebel S. J., Klöppel S., Weiskopf N., Friston K. J.* Dynamic causal modeling: A generative model of slice timing in fMRI // Neuroimage, 2007. Vol. 34. Iss. 4. P. 1487–1496.
3. *Ковалев Д. Ю., Сергеев Д. И., Тириков Е. М., Пономарева Н. В.* Методы и средства анализа сигналов головного мозга человека на данных функциональной магнитно-резонансной томографии // Data analytics and management in data intensive domains / Eds. B. K. Thalheim, A. V. Sychev, S. D. Makhortov. — CEUR, 2020. Vol. 2790. P. 214–229.
4. *Pham H.* System software reliability. — London: Springer-Verlag, 2006. 441 p.
5. *Pham H.* A new criterion for model selection // Mathematics, 2019. Vol. 7. Iss. 12. Art. ID: 1215. 12 p.
6. *Rencher A. C., Schaalje G. B.* Linear models in statistics. — New York, NY, USA: John Wiley & Sons, 2008. 672 p.
7. *Mahmoudi M. R., Maleki M., Pak A.* Testing the equality of two independent regression models // Commun. Stat. A — Theor., 2018. Vol. 47. Iss. 12. P. 2919–2926.
8. *Akaike H.* A new look at the statistical model identification // IEEE T. Automat. Contr., 1974. Vol. 19. Iss. 6. P. 716–723.
9. *Liddle A. R.* Information criteria for astrophysical model selection // Mon. Not. R. Astron. Soc. Lett., 2007. Vol. 377. Iss. 1. P. L74–L78.
10. *Giraud C.* Introduction to high-dimensional statistics. — Boca Raton, FL, USA: CRC Press, 2014. 270 p.
11. *Borges C. E., Alonso C. L., Montaña J. L.* Model selection in genetic programming // 12th Annual Conference on Genetic and Evolutionary Computation Proceedings. — New York, NY, USA: ACM, 2010. P. 985–986. doi: 10.1145/1830483.1830662.
12. *Тарасов Е. А., Ковалев Д. Ю.* Оценка качества научных гипотез в виртуальных экспериментах в областях с интенсивным использованием данных // Data analytics and management in data intensive domains / Eds. L. Kalinichenko, Ya. Manolopoulos, N. Skvortsov, V. Sukhomlin. — CEUR, 2017. Vol. 2022. P. 281–292.
13. *Wang Y., Squires C., Belyaeva A., Uhler C.* Direct estimation of differences in causal graphs // Adv. Neur. In., 2018. Vol. 31. P. 3770–3781.
14. *Arslan S., Ktena S. I., Makropoulos A., Robinson E. C., Rueckert D., Parisot S.* Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex // Neuroimage, 2018. Vol. 170. P. 5–30.

15. *Desikan R. S., Ségonne F., Fischl B., et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest // Neuroimage, 2006. Vol. 31. Iss. 3. P. 968–980.
16. *Dixon W. J., Mood A. M.* The statistical sign test // J. Am. Stat. Assoc., 1946. Vol. 41. Iss. 236. P. 557–566.

Поступила в редакцию 27.12.20

METHODS FOR COMPARING COMPETING HYPOTHESES IN HYPOTHESIS-ORIENTED SYSTEMS

E. M. Tirikov and D. Y. Kovalev

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

Abstract: With the advent of a new class of virtual experiments management systems, the use of hypotheses and models in an explicit form becomes more and more widespread. Such systems apply both hypotheses generated from the data and theoretical hypotheses. It becomes critically important to compare several competing hypotheses of different origin with each other. The paper considers various approaches to comparing competing hypotheses and computational models implementing them. The considered approaches are implemented as a software component that is a part of a virtual experiment management system. The component is applied for problem solving in neurophysiology.

Keywords: virtual experiments management systems; competing hypotheses; comparison of hypotheses

DOI: 10.14357/08696527210110

Acknowledgments

The research was partially supported by the Russian Foundation for Basic Research (projects 18-07-01434 and 18-29-22096).

References

1. Kovalev, D., and E. Tarasov. 2019. Virtual experiments in data intensive research. *Informatika i ee Primeneniya — Inform. Appl.* 13(2):117–125.
2. Kiebel, J., S. Klöppel, N. Weiskopf, and K. J. Friston. 2007. Dynamic causal modeling: A generative model of slice timing in fMRI. *Neuroimage* 34(4):1487–1496.

3. Kovalev, D., D. Sergeev, E. Tirikov, and N. Ponomareva. 2020. Metody i sredstva analiza signalov golovnogo mozga cheloveka na dannykh funktsional'noy magnitno-rezonansnoy tomografii [Methods and tools for analyzing human brain signals based on functional magnetic resonance imaging data]. *Data analytics and management in data intensive domains*. Eds. B. K. Thalheim, A. V. Sychev, and S. D. Makhortov. CEUR. 2790:214–229.
4. Pham, H. 2006. *System software reliability*. London: Springer-Verlag. 441 p.
5. Pham, H. 2019. A new criterion for model selection. *Mathematics* 7(12):1215. 12 p.
6. Rencher, A. C., and G. B. Schaalje. 2008. *Linear models in statistics*. New York, NY: John Wiley & Sons. 672 p.
7. Mahmoudi, M. R., M. Maleki, and A. Pak. 2018. Testing the equality of two independent regression models. *Commun. Stat. A — Theor.* 47(12):2919–2926.
8. Akaike, H. 1974. A new look at the statistical model identification. *IEEE T. Automat. Contr.* 19(6):716–723.
9. Liddle, A. R. 2007. Information criteria for astrophysical model selection. *Mon. Not. R. Astron. Soc. Lett.* 377(1):L74–L78.
10. Giraud, C. 2014. *Introduction to high-dimensional statistics*. Boca Raton, FL, USA: CRC Press. 270 p.
11. Borges, C. E., C. L. Alonso, and J. L. Montaña. 2010. Model selection in genetic programming. *12th Annual Conference on Genetic and Evolutionary Computation Proceedings*. New York, NY: ACM. 985–986. doi: 10.1145/1830483.1830662.
12. Tarasov, E., and D. Kovalev. 2017. Otsenka kachestva nauchnykh gipotez v virtual'nykh eksperimentakh v oblastyakh s intensivnym ispol'zovaniem dannykh [Estimation of scientific hypotheses quality in virtual experiments in data intensive domains]. *Data analytics and management in data intensive domains*. Eds. L. Kalinichenko, Ya. Manolopoulos, N. Skvortsov, and V. Sukhomlin. CEUR. 2022:281–292.
13. Wang, Y., C. Squires, A. Belyaeva, and C. Uhler. 2018. Direct estimation of differences in causal graphs. *Adv. Neur. Inf.* 31:3770–3781.
14. Arslan, S., S. I. Ktena, A. Makropoulos, E. C. Robinson, D. Rueckert, and S. Parisot. 2018. Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *Neuroimage* 170:5–30.
15. Desikan, R. S., F. Ségonne, B. Fischl, *et al.* 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31(3):968–980.
16. Dixon, W. J., and A. M. Mood. 1946. The statistical sign test. *J. Am. Stat. Assoc.* 41(236):557–566.

Received December 27, 2020

Contributors

Kovalev Dmitry Y. (b. 1988) — 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; dkovalev@ipiran.ru

Tirikov Egor M. (b. 1996) — PhD student, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; em.tirikov@gmail.com

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ В МАТЕМАТИЧЕСКОМ ОБЕСПЕЧЕНИИ ЦИФРОВЫХ ДВОЙНИКОВ ЭЛЕКТРОЭНЕРГЕТИЧЕСКИХ СИСТЕМ

С. П. Ковалёв[1]

Аннотация: Рассматривается проблематика разработки цифровых двойников современных активных распределительных электроэнергетических систем (РЭЭС). Выделены подходы к применению искусственных нейронных сетей глубокого обучения в интеллектуальном управлении указанными системами на базе цифровых двойников. Дан краткий обзор релевантных архитектур нейросетей. Приведены примеры нейросетевых средств для решения ряда ключевых задач интеллектуального управления, включая прогнозирование нагрузки, прогнозирование цены электроэнергии, оптимизацию распределения нагрузки между доступным генерирующим оборудованием, оценку и прогнозирование технического состояния энергетического оборудования, диагностику отказов и катастроф. Сформулированы рекомендации по альтернативным вариантам применения рассмотренных нейросетевых средств, таким как включение в состав базового математического обеспечения цифрового двойника либо поставка в виде дополнительных приложений для определенных категорий пользователей.

Ключевые слова: цифровой двойник; распределительная электроэнергетическая система; искусственная нейронная сеть глубокого обучения; прогнозирование; диагностика отказов

DOI: 10.14357/08696527210111

1 Введение

Современные активные РЭЭС представляют собой сложные высокотехнологичные объекты, соединяющие разнообразные энергоприемники (в том числе с управляемым потреблением), локальное генерирующее оборудование (в том числе на возобновляемых источниках энергии — ВИЭ) и системы накопления электроэнергии. Главным принципом организации управления такими объектами, в соответствии с подходом модельно-ориентированной системной инженерии (Model-Based Systems Engineering) и парадигмой четвертой промышленной революции (Industrie 4.0), является формирование цифрового двойника (Digital Twin) — виртуальной копии объекта, достоверно воспроизводящей и задающей структуру, состояние и поведение оригинала в реальном времени [1]. Цифровой

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

двойник, будучи интеллектуальной надстройкой над средой интернета вещей и информационной моделью объекта, становится ключевым базовым элементом системы управления объектом.

Однако в настоящее время многие вопросы формирования и применения цифровых двойников в энергетике не решены [2]. Традиционные физико-математические модели плохо справляются с высокой динамикой и скрытыми поведенческими закономерностями, присущими РЭЭС, поэтому математическое обеспечение цифрового двойника требует применения средств искусственного интеллекта, в частности на базе нейронных сетей глубокого обучения [3]. Анализу таких средств, включая рекомендации по характеру их применения, посвящена настоящая работа.

2 Возможности искусственных нейронных сетей

Искусственной нейронной сетью (кратко — нейросетью) называется многокомпонентный информационно-вычислительный блок, способный к самообучению на примерах аналогично мозгу живого организма [4]. Нейросети развиваются начиная с 1950-х гг. и в настоящее время переживают настоящий бум в качестве самой перспективной технологии искусственного интеллекта — автоматического «угадывания» заранее не известных зависимостей. Еще в 1980-х гг. было установлено, что достаточно большая нейросеть может быть обучена с любой точностью аппроксимировать любую функцию многих переменных F, не имея никакой априорной информации о ней [5]. Примерами для обучения аппроксимирующей нейросети служат пары вида $(\boldsymbol{Y}, F(\boldsymbol{Y}))$ с разными наборами аргументов $\boldsymbol{Y}$, в совокупности в той или иной степени покрывающими область определения функции F. Обученная нейросеть, получив на вход произвольное значение набора аргументов $\boldsymbol{X}$, быстро выдает значение, близкое к $F(\boldsymbol{X})$.

Широко известны примеры применения нейросетей в задачах, где F задает: принадлежность входного набора тому или иному классу (распознавание образов различной модальности — изображений, фрагментов речи, массивов численных данных и др.); очередное значение в некотором ряду, заданному входным набором (прогнозирование); перевод входных значений в иную систему обозначений (кодирование) и т. д. Существуют также задачи обучения нейросети «без учителя», когда от нее требуется вычислить не аппроксимацию некоторой функции, а статистические характеристики распределения входных наборов данных. В инженерных прикладных задачах необходимо тщательно подбирать структуру набора аргументов так, чтобы он: (1) включал все параметры, существенно влияющие на неизвестную аппроксимируемую зависимость F; (2) не был зашумлен незначащими параметрами; (3) позволял сравнительно легко получить много достоверных обучающих примеров, более или менее равномерно распределенных по области определения функции F. Но даже при очень хорошем наборе следует иметь в виду, что качество аппроксимации гаранти-

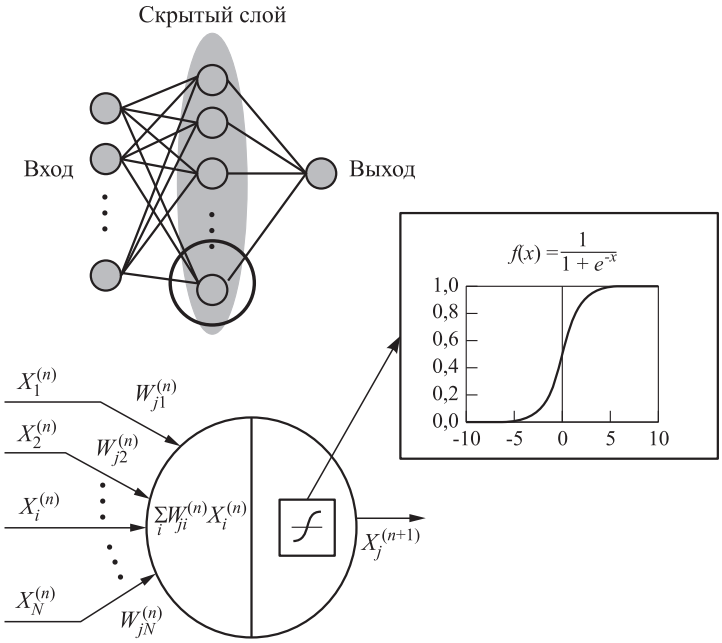


Рис. 1 Искусственная нейронная сеть

руется лишь в среднем: в общем случае могут существовать значения набора аргументов, на которых выход обученной нейросети будет сколь угодно сильно отличаться от правильного значения функции F. Известны вычислительные эксперименты в распознавании изображений, где небольшие, незаметные человеческому глазу искажения, вычисленные по «белым пятнам» хорошей обучающей выборки, вызывали у нейросетевых распознавателей фатальную ошибку — например, искаженное изображение автобуса распознавалось как страус [6]. Поэтому современные нейросети могут значительно обогатить, но не заменить собой другие средства математического обеспечения интеллектуального управления.

Элементарным строительным блоком нейросети, как нетрудно догадаться, служит несложный вычислительный модуль, грубо имитирующий поведение клетки-нейрона. Этот модуль имеет несколько входов и один выход, на который при поступлении входных сигналов выдает результат действия функции возбуждения на их взвешенную сумму (рис. 1). Функция возбуждения имеет вид сигмоида, как на рис. 1, или другой непрерывной аппроксимации пороговой зависимости. Наборы $\boldsymbol{W}$ весовых коэффициентов входов всех нейронов и служат теми параметрами нейросети, которые подбираются в ходе обучения в целях минимизации функции ошибки.

Для нейросетей традиционной является многослойная топология: сигнал последовательно проходит сквозь слои в направлении от входного (input) через промежуточные или скрытые (hidden) к выходному (output). Нейросеть с базовой многослойной архитектурой называется перцептроном (perceptron). В современных перцептронах число промежуточных слоев доходит до сотен, поэтому их называют глубокими (deep). Считается, что по мере прохождения слоев входной сигнал преобразуется в представления все более высокоуровневых признаков (features), характеризующих его подлинный «смысл». Для задач типа распознавания образов, которые сводятся именно к выделению признаков, можно оптимизировать сеть путем уменьшения числа связей между нейронами соседних слоев, соответственно уменьшая количество коэффициентов, подлежащих подбору. Оптимизированная таким способом глубокая сеть называется сверточной (convolutional neural network, CNN) [7].

Существуют также рекуррентные архитектуры сетей (recurrent neural network, RNN), включающие обратные связи от нейронов последующих слоев к предыдущим. Благодаря обратным связям появляется возможность учитывать результаты предыдущих актов прохождения сигналов в последующих, т. е. частично запоминать историю функционирования сети. Эта возможность особенно удобна в задачах прогнозирования временны́х рядов. Чтобы учитывать «медленные» корреляции, заметные только на достаточно длинных фрагментах ряда, в нейроны такой сети в явном виде встраивают элементы памяти и управляющие их наполнением внутренние коммутаторы (гейты). Наиболее широко применяются два типа нейронов такого рода: Long Short-Term Memory (LSTM) и Gated Recurrent Unit (GRU) [8].

В завершение краткого обзора рассмотрим архитектуры нейросетей, применяемые в обучении без учителя. Одним из подходов здесь является построение сети, в которой входной и выходной слой состоят из одинакового числа нейронов, а скрытые слои — из существенно меньшего количества. Такую нейросеть тренируют на произвольном массиве предметных данных в целях минимизации отличия выхода от входа, вынуждая скрытый слой формировать набор признаков, статистически достоверно классифицирующий входной массив. Так устроены автокодировщики (autoencoders, AE), известные своим применением в распознавании речи: автокодировщик берет на себя трудоемкую функцию разметки потока речи на смысловые единицы [9]. Другой подход к обучению без учителя состоит в организации взаимосвязанных нейронов в ансамбль в физико-статистическом смысле. По мере прохождения наборов входных данных сквозь такой ансамбль он стремится к равновесному состоянию (минимуму функционала «энергии»), в котором воспроизводится наиболее «похожее на правду» вероятностное распределение входных данных. Такой подход реализован в ограниченных машинах Больцмана (restricted Boltzmann machine, RBM), из которых собирают глубокие сети доверия (deep belief network, DBN).

В электроэнергетике естественным образом выделяются несколько областей применения нейросетей [10], каждая из которых кратко рассмотрена далее.

3 Области применения нейросетей в энергетике

3.1 Прогнозирование нагрузки (load forecasting)

Оценка объема потребления некоторого объекта энергетики в некоторый будущий период времени при известных фактических исторических значениях потребления объекта относится к классическим задачам управления в электроэнергетике. Точное прогнозирование позволяет правильно рассчитать энергетический баланс и выбрать режимные параметры РЭЭС. Типовая задача в условиях современного рынка состоит в построении прогнозного профиля потребления актива одного субъекта на очередные календарные сутки с дискретностью 1 ч, но встречаются и задачи с масштабом объекта прогнозирования от одного прибора до национальной энергосистемы и с горизонтом прогнозирования от 1 мин до 20 лет.

Хорошо известны такие грубые алгоритмы (формулы) прогнозирования, как усреднение фактических профилей за четыре предшествующих типовых (рабочих/выходных) дня и суммирование мощности всех энергоприемников объекта с учетом коэффициентов их использования. Существуют и более тонкие статистические методы прогнозирования, такие как регрессионные модели, в которых по историческим значениям подбираются коэффициенты заранее заданной функциональной зависимости потребления от времени (в виде Фурье-разложения с несколькими начальными гармониками) и от температуры воздуха (в виде полинома невысокой степени). В целях дальнейшего увеличения точности и скорости прогнозирования (а также отчасти из стремления отдать дань моде) привлекают нейросети глубокого обучения.

Примеров краткосрочного прогнозирования нагрузки при помощи нейросетей известно очень много. Как и в регрессионных моделях, основными входными

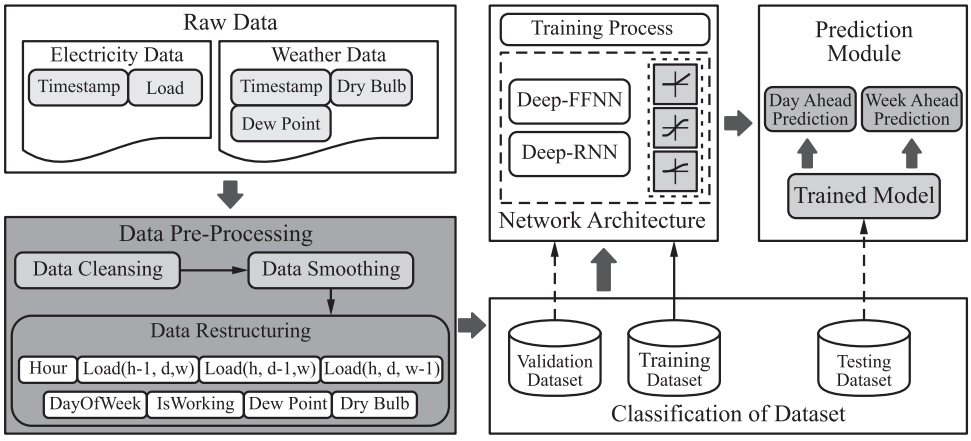


Рис. 2 Вычислительный эксперимент по прогнозированию нагрузки [11]

данными для них служат массивы исторических значений потребления за различные периоды (предыдущий час, день, неделя) и температура воздуха. В качестве дополнительных аргументов могут подаваться: номер дня недели, тип дня (рабочий/выходной), температура точки росы (для учета влажности). Известны вычислительные эксперименты по прогнозированию почасовых профилей потребления для энергосистем регионального уровня на базе как многослойных перцептронов, так и рекуррентных сетей, обученных на вышеперечисленных данных, где удалось достичь погрешности менее 1% на реальных данных [11] (рис. 2). Наименьшую погрешность продемонстрировала рекуррентная сеть, причем с гиперболическим тангенсом в качестве функции активации.

3.2 Прогнозирование цен на электроэнергию (electricity price forecasting)

Постановка задачи прогнозирования цен на энергоресурсы также вполне очевидна: требуется предсказать цену на некоторый будущий период. Благодаря наличию обширных массивов исторических данных с фактическими ценами, нетрудно реализовать «наивный» алгоритм расчета ожидаемой цены на очередные сутки, состоящий в выборе массива за предыдущие сутки с таким же типом дня недели (рабочий/выходной), с последующим внесением «фундаментальных» поправок (сезонные, общеэкономические и прочие факторы) и добавлением небольшого случайного «шума». Известны примеры, когда такой алгоритм давал ошибку примерно в 16%, превосходя по точности некоторые сложные статистические методы [12]. Однако по мере развития энергетического рынка требования к точности прогноза цены со стороны участников повышаются. Увеличиваются разброс цен и степень влияния косвенных факторов, вплоть до возникновения «эффекта бабочки». Поэтому востребованными становятся интеллектуальные методы прогнозирования цены.

В качестве перспективного метода здесь естественным образом рассматривается применение рекуррентных нейросетей глубокого обучения. При этом на вход такой сети целесообразно подавать не только накопленный ряд фактических значений цен за предыдущие периоды (отдельно по каждой действующей модели рынка), но и дополнительные переменные технико-экономической «обстановки», такие как температура воздуха или соотношение спроса и предложения. Известен пример практической реализации инструмента прогнозирования цен на электроэнергию с учетом этих переменных на базе рекуррентной сети нейронов типа GRU [12].

3.3 Оптимизация распределения нагрузки (economic dispatch)

Распределение нагрузки между имеющимися генерирующими источниками относится к числу основных задач управления РЭЭС. Актуальность этой задачи обусловлена возможностью РЭЭС передавать потребителям энергию, полученную из различных источников — как от локальных генераторов и накопителей,

так и от удаленных станций, доступных через подключение к магистральной сети. Появляется возможность принимать конкретные решения по отбору мощности от каждого доступного источника так, чтобы максимизировать интегральный экономический эффект от поставки электроэнергии при соблюдении ограничений по надежности.

При распределении известной (достоверно предсказанной) нагрузки следует минимизировать эксплуатационные расходы генерирующих мощностей. Однако зависимость расходов от объема выработки часто не известна точно и подвержена влиянию многих факторов и труднопредсказуемых физических ограничений. Кроме того, при принятии решения для каждого источника следует учитывать потери при передаче электроэнергии до нагрузки, уровень загрязнения окружающей среды вследствие работы источника и другие нефинансовые факторы. Эти условия затрудняют применение традиционных методов решения оптимизационных задач типа математического программирования и открывают поле для привлечения нейросетей.

Для оптимизации распределения нагрузки могут использоваться классические многослойные перцептроны [13]. Их обучают на примерах, сгенерированных путем имитационного моделирования. В дополнение к ним задействуются перцептроны, обученные предсказывать по метеоданным выработку ВИЭ.

3.4 Оценка и прогнозирование технического состояния энергетического оборудования (power machine health assessment and prediction)

Задача оценки технического состояния оборудования состоит в определении фактического наличия, характера, мест износа, сбоев и отказов и, соответственно, потребности в техническом обслуживании. Исходным материалом для выполнения оценки служат массивы результатов измерения непосредственно наблюдаемых характеристик состояния оборудования, таких как вибрация, температура, давление, выходное напряжение и др. Прогноз состояния сводится к оценке, но не на текущий момент времени, а на некоторый будущий период (горизонт прогнозирования). На основе прогноза рассчитываются индикаторы технического состояния, такие как оставшееся время полезной работы оборудования (remaining useful life, RUL).

Современные технологии интернета вещей позволяют автоматически формировать в реальном времени большие потоки достоверных значений первичных характеристик состояния оборудования, и для их обработки требуются вычислительные модули, способные оперативно выдавать заслуживающие доверия оценки и прогнозы. Существуют методики очень грубой оценки индикаторов технического состояния путем простых алгебраических преобразований первичных характеристик, однако их нельзя считать удовлетворительными для высокотехнологичных РЭЭС. В свою очередь, традиционные физико-математические модели подходят плохо, поскольку они в основном рассчитаны на режим постобработки (офлайн). Кроме того, практически невозможно построить физически

достоверную модель возникновения сбоев в оборудовании сложной конструкции с учетом всех влияющих факторов и условий его функционирования. Возникает естественное поле для применения машинного обучения и нейросетей.

Для оценки и прогнозирования технического состояния оборудования в последние годы были апробированы все основные архитектуры сетей глубокого обучения [14]. Были установлены и две основные трудности применения нейросетей: большое число входных сигналов и малое количество обучающих примеров (массивов значений характеристик, зарегистрированных во время сбоев и размеченных результатами их анализа). В целях уменьшения числа сигналов входной поток переводят в частотную область путем быстрого преобразования Фурье либо вейвлет-разложения. А чтобы задействовать распознавательный потенциал сверточных сетей, строится двумерная картина входного потока в частотно-временно́й области. Рекуррентные сети также находят применение для анализа временны́х рядов характеристик. А что касается недостатка обучающих примеров, то его обходят, привлекая технологии обучения без учителя (на базе автокодировщиков и машин Больцмана), позволяющие идентифицировать сбои как статистические аномалии входного потока без какой-либо априорной информации о них. Следуя условной аналогии между нейросетью и человеческим мозгом, можно сказать, что обучение оценке технического состояния без учителя наделяет компьютер хорошо развитой у опытного эксплуатирующего персонала способностью «нутром чувствовать, что оборудование работает неправильно».

3.5 Диагностика отказов и катастроф (faults and disasters diagnosis)

Диагностика отказов в РЭЭС состоит в определении места короткого замыкания, его вида и максимального отклонения напряжения от допустимого. Диагностика естественным образом сводится к анализу профилей напряжения на шинах в небольшой временно́й окрестности момента отказа с учетом конфигурации (топологии) РЭЭС. Процедуру такого анализа можно описать как задачу распознавания образов на профилях, поэтому он еще с 1990-х гг. рассматривается как естественная область применения нейросетей.

Однако реальные РЭЭС имеют такой масштаб, что нейросетевой анализ отказов на них требует очень больших вычислительных ресурсов. Поэтому осязаемые перспективы его практического применения появились лишь в последнее время благодаря развитию технологий глубокого обучения и их аппаратной базы. Известны разработки в области автоматической диагностики отказов и аварий на базе сверточных нейросетей [15] и автокодировщиков [16]. Много обучающих примеров для тренировки таких сетей можно сгенерировать путем имитационного моделирования в программах-симуляторах промышленного уровня готовности, таких как Siemens PSS/E. Благодаря высокой скорости срабатывания хорошо обученной нейросети, появляется перспектива перехода к превентивному анализу состояния РЭЭС до наступления аварий и определения мер по их предотвращению. Дополнительно при помощи нейросетей решаются задачи типа верификации

топологии РЭЭС, разложения интегральных профилей потребления на профили отдельных энергоприемников (дезагрегации) и т. п.

Еще более перспективным представляется применение нейросетей для прогнозирования структурных катастроф РЭЭС — эксцессов, основной ущерб от которых вызван не отказом какого-либо отдельного элемента, а каскадным разрушением взаимосвязей между элементами («эффект домино»). Хотя такие катастрофы немногочисленны, ущерб от них настолько велик, что любой вклад в повышение достоверности их предсказания имеет большое социальное и экономическое значение. Физико-математические модели структурных катастроф очень трудоемки в разработке и решении ввиду большого числа разноплановых влияющих факторов и негладкого/разрывного характера моделируемых зависимостей, а база наблюдений за такими катастрофами невелика, поэтому можно предположить, что для их прогнозирования целесообразно будет применять нейросети с обучением без учителя. Однако практическая реализация этого подхода требует значительного объема дальнейших исследований.

4 Заключение

Нейросети обладают значительным потенциалом применения в интеллектуальном управлении РЭЭС на базе цифровых двойников. Уже сейчас они позволяют решать ряд задач с ошибкой порядка 10%, что считается достаточным для многих практических целей. Нейросетевые средства прогнозирования нагрузки, цен и генерации целесообразно включать в базовое математическое обеспечение цифрового двойника, в то время как средства диагностики технического состояния и аварий относятся к специализированным приложениям. А в перспективе нейросетевые средства смогут занять лидирующее положение среди компонентов математического обеспечения цифрового двойника энергосистемы. Здесь возникает ряд перспективных направлений дальнейших исследований.

Литература

1. *Madni A. M., Madni C. C., Lucero S. D.* Leveraging digital twin technology in model-based systems engineering // Systems, 2019. Vol. 7. Iss. 1. Art. No. 7. 13 p.
2. *Andryushkevich S. K., Kovalyov S. P., Nefedov E.* Composition and application of power system digital twins based on ontological modeling // 17th IEEE Conference (International) on Industrial Informatics Proceedings. — Helsinki–Espoo, Finland: IEEE, 2019. P. 1536–1542.
3. *Frolov D.* How machine learning empowers models for digital twins // Benchmark, 2018. Vol. 6. P. 48–53.
4. *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. — СПб.: Питер, 2018. 480 с.
5. *Cybenko G. V.* Approximation by superpositions of a sigmoidal function // Math. Control Signal., 1989. Vol. 2. No. 4. P. 303–314.

6. *Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I. J., Fergus R.* Intriguing properties of neural networks // arXiv.org, 2013. arXiv:1312.6199 [cs.CV]. https://arxiv.org/abs/1312.6199.
7. *Srinivas S., Sarvadevabhatla R. K., Mopuri K. R., Prabhu N., Kruthiventi S. S. S., Babu R. V.* A taxonomy of deep convolutional neural nets for computer vision // arXiv.org, 2016. arXiv:1601.06615 [cs.CV]. https://arxiv.org/abs/1601.06615.
8. *Chung J., Gulcehre C., Cho K., Bengio Y.* Empirical evaluation of gated recurrent neural networks on sequence modeling // arXiv.org, 2014. arXiv:1412.3555 [cs.NE]. https://arxiv.org/abs/1412.3555.
9. *Hsu W.-N., Zhang Y., Glass J.* Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation // arXiv.org, 2017. arXiv:1707.06265 [cs.CL]. https://arxiv.org/abs/1707.06265.
10. *Панов М., Хмелев И., Смирнов А.* Нейронные сети на службе энергетиков // Открытые системы. СУБД, 2016. № 4. С. 39–41.
11. *Mohammad F., Lee K.-B., Kim Y.-C.* Short term load forecasting using deep neural networks // arXiv, 2018. arXiv:1811.03242 [cs.NE]. https://arxiv.org/abs/1811.03242.
12. *Ugurlu U., Oksuz I., Tas O.* Electricity price forecasting using recurrent neural networks // Energies, 2018. Vol. 11. Iss. 5. Art. No. 1255. 21 p. doi: 10.3390/en11051255.
13. *Bhattacharya B., Sinha A.* Intelligent subset selection of power generators for economic dispatch // arXiv.org, 2017. arXiv:1709.02513 [cs.CE]. https://arxiv.org/abs/1709.02513.
14. *Zhao R., Yan R., Chen Z., Mao K., Wang P., Gao R. X.* Deep learning and its applications to machine health monitoring // Mech. Syst. Signal Pr., 2019. Vol. 115. P. 213–237.
15. *Rudin F., Li G.-J., Wang K.* An algorithm for power system fault analysis based on convolutional deep learning neural networks // Int. J. All Research Education Scientific Methods, 2017. Vol. 5. Iss. 9. P. 11–18.
16. *Chen K., Hu J., He J.* Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder // IEEE T. Smart Grid, 2016. Vol. 9. Iss. 3. P. 1748–1758.

Поступила в редакцию 22.05.19

EMPLOYING DEEP LEARNING NEURAL NETWORKS IN MATHEMATICAL BASIS OF DIGITAL TWINS OF ELECTRICAL POWER SYSTEMS

S. P. Kovalyov

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

Abstract: Development problems for digital twins of contemporary active power distribution systems are considered. Approaches to employing deep learning

neural networks in digital twin-based intelligent control of these systems are highlighted. A brief review of relevant neural network architectures is outlined. Examples of neural network tools for solving a number of key intelligent control problems are presented, including load forecasting, electricity price forecasting, economic dispatch, power machine health assessment and prediction, and faults and disasters diagnosis. Recommendations are provided regarding alternative deployment modes of the presented neural network tools, such as inclusion into a digital twin basic mathematical software, or supply as auxiliary applications for certain categories of users.

Keywords: digital twin; electrical distribution system; deep learning neural network; forecasting; fault diagnosis

DOI: 10.14357/08696527210111

References

1. Madni, A. M., C. C. Madni, and S. D. Lucero. 2019. Leveraging digital twin technology in model-based systems engineering. *Systems* 7(1):7. 13 p
2. Andryushkevich, S. K., S. P. Kovalyov, and E. Nefedov. 2019. Composition and application of power system digital twins based on ontological modeling. *17th IEEE Conference (International) on Industrial Informatics Proceedings*. Helsinki–Espoo, Finland: IEEE. 1536–1542.
3. Frolov, D. 2018. How machine learning empowers models for digital twins. *Benchmark* 6:48–53.
4. Nikolenko, S., A. Kadurin, and E. Arkhangel'skaya. 2018. *Glubokoe obuchenie* [Deep learning]. St. Petersburg: Piter. 480 p.
5. Cybenko, G. V. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control Signal.* 2(4):303–314.
6. Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. 2013. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]. Available at: https://arxiv.org/abs/1312.6199 (accessed February 11, 2021).
7. Srinivas, S., R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu. 2016. A taxonomy of deep convolutional neural nets for computer vision. arXiv:1601.06615 [cs.CV]. Available at: https://arxiv.org/abs/1601.06615 (accessed February 11, 2021).
8. Chung, J., C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 [cs.NE]. Available at: https://arxiv.org/abs/1412.3555 (accessed February 11, 2021).
9. Hsu, W.-N., Y. Zhang, and J. Glass. 2017. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. arXiv:1707.06265 [cs.CL]. Available at: https://arxiv.org/abs/1707.06265 (accessed February 11, 2021).
10. Panov, M., I. Khmelev, and A. Smirnov. 2016. Neyronnye seti na sluzhbe energetikov [Neural networks in the service of power engineers]. *Otkrytye sistemy. SUBD* [Open Systems J.] 4:39–41.
11. Mohammad, F., K.-B. Lee, and Y.-C. Kim. 2018. Short term load forecasting using deep neural networks. arXiv:1811.03242 [cs.NE]. Available at: https://arxiv.org/abs/1811.03242 (accessed February 11, 2021).

12. Ugurlu, U., I. Oksuz, and O. Tas. 2018. Electricity price forecasting using recurrent neural networks. *Energies* 11(5):1255. 21 p. doi: 10.3390/en11051255.
13. Bhattacharya, B., and A. Sinha. 2017. Intelligent subset selection of power generators for economic dispatch. arXiv:1709.02513 [cs.CE]. Available at: https://arxiv.org/abs/1709.02513 (accessed February 11, 2021).
14. Zhao, R., R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao. 2019. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Pr.* 115:213–237.
15. Rudin, F., G.-J. Li, and K. Wang. 2017. An algorithm for power system fault analysis based on convolutional deep learning neural networks. *Int. J. All Research Education Scientific Methods* 5(9):11–18.
16. Chen, K., J. Hu, and J. He. 2016. Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder. *IEEE Trans. Smart Grid* 9(3):1748–1758.

Received May 22, 2019

Contributor

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

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2021. Т. 31. № 1. С. 145–156

МОДЕЛЬ СООБЩЕСТВА ПОЛЬЗОВАТЕЛЕЙ ТЕХНОЛОГИИ ПОДДЕРЖКИ КОНКРЕТНО-ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЙ

И. М. Адамович[1]**, *О. И. Волков***[2]

Аннотация: Статья продолжает серию работ, посвященных описанию и анализу распределенной технологии поддержки конкретно-исторических исследований (ПКИИ), основанной на принципах краудсорсинга. Данная статья посвящена описанию и обоснованию подхода к моделированию сообщества пользователей технологии (СПТ) и процессов распространения искаженной информации между ее членами с целью изучения устойчивости технологии ПКИИ к попыткам искажения истории, что становится актуальной задачей в современном обществе. Предложенный подход заключается в формировании модели структуры сообщества на базе ориентированного графа Боллобаша–Риордана. Принципы распространения информации между узлами графа определены с учетом факторов и эффектов, имеющих место в реальных социальных сетях и обусловленных как характеристиками их членов, характером их взаимодействия, так и свойствами самой социальной сети. Проверка адекватности модели была осуществлена за счет сравнения характера распространения искаженной информации в режиме отсутствия противодействия информационным атакам с логистической кривой, отражающей процесс диффузии инноваций.

Ключевые слова: виртуальное сообщество; модель; технология; искажение истории; конкретно-историческое исследование

DOI: 10.14357/08696527210112

1 Введение

Поддержка конкретно-исторических исследований становится одной из актуальных задач современности в связи с вовлечением в исследовательский процесс не только членов профессионального исторического сообщества, но и самых широких слоев непрофессионалов в связи со все увеличивающимся интересом к частной, семейной истории [1].

В [2, 3] описана разработанная в ФИЦ ИУ РАН распределенная технология ПКИИ, основанная на принципах краудсорсинга (мобилизации ресурсов широкого круга добровольцев посредством информационных технологий). Технология ПКИИ включает в себя онлайн-платформу для коммуникации и совместной

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Adam@amsd.com

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Volkov@amsd.com

деятельности исследователей, и СПТ ПКИИ обладает всеми признаками научно-профессионального сетевого сообщества [4]. Также его можно рассматривать как виртуальное деятельное сообщество (сообщество практики) — группу людей, объединенных общим интересом, профессией или хобби, существующую в сети Интернет без какого-либо стороннего финансирования или принуждения и характеризующуюся:

– приверженностью к единой предметной области (в данном случае — истории);

– осуществлением взаимодействия и информационного обмена;

– участием в практической деятельности [5].

Под устойчивостью технологии ПКИИ к фальсификации истории будем понимать способность СПТ ПКИИ противостоять попыткам сознательного искажения исторической информации, что становится актуальной задачей в современном обществе [6].

Данная статья посвящена моделированию как самой СПТ ПКИИ, так и процессов распространения искаженной информации между ее узлами с целью изучения устойчивости технологии ПКИИ к попыткам искажения истории.

2 Описание подхода к моделированию

Существует множество подходов к моделированию процесса распространения информации в социальных сетях: модель Далея–Кендалла, модель на базе клеточного автомата, пороговая модель, каскадная модель, марковская модель влияния и др. Особую популярность приобрела SIR (susceptible-infectious-removed) модель, основанная на сходстве процессов распространения информации в виртуальном сообществе и распространения инфекции в популяции [7]. Недостаток данного подхода — усреднение последствий, в то время как фундаментальным источником риска в таких явлениях выступают экстремумы, а не средние значения [8], поэтому представляется более целесообразным провести изучение СПТ ПКИИ с помощью методов имитационного моделирования. Данный метод позволит оценить близость плотности распределения последствий информационных атак к нормальному распределению и тем самым сделать вывод о целесообразности оценки рисков на базе средних значений. Для этого следует определить как принципы моделирования структуры сообщества, так и принципы распространения информации между ее узлами.

3 Принципы построения модели структуры сообщества

Для изучения генезиса социальных сетей применяются случайные графы различных типов. Особо следует выделить случайные графы с предпочтительным связыванием, которые объясняют механизм роста сетевых структур и обеспечивают их адекватное моделирование [9].

Социальные сети относятся к так называемым «малым мирам» — графам, обладающим при значительном числе вершин сравнительно небольшим диаметром. Для малых миров типичны степенные распределения вершин по числу имеющихся у них связей (безмасштабные сети) [10].

Исторически первая модель предпочтительного связывания, разработанная для моделирования сети Интернет и наиболее известная на сегодняшний день, была предложена в работе A. L. Barabasi и R. Albert [11]. Впоследствии случайные неориентированные графы, реализующие эту модель, получили название графов Барабаши–Альберт. Но эти графы, несмотря на свою популярность, — не единственные представители класса случайных графов с предпочтительным связыванием.

Предметом исследования является влияние членов социальной сети друг на друга и изменение их мнений по тем или иным вопросам истории в результате этого влияния. Модель влияния в социальной сети представляет собой взвешенный ориентированный граф, вершины которого соответствуют членам социальной группы, а дуги описывают их взаимное влияние [12]. Действительно, в сообществе, организованном вокруг некоторого общего занятия, обмен влиянием между участниками, обладающими разным уровнем профессионализма и авторитетности, не может быть симметричным. Если участник A прислушивается к мнению участника B и внимательно читает его публикации в сообществе, то участник B при этом может и не подозревать о существовании участника A.

Модель, описывающая ориентированный граф, в котором предпочтительное присоединение зависит от входящих и исходящих степеней, была предложена Боллобашем и др. в работе [13]. Именно эта модель была взята за основу при моделировании СПТ ПКИИ.

4 Принципы моделирования процесса распространения информации

При моделировании социальных сетей, взаимного влияния их членов (агентов), динамики их мнений и т. д. возникает необходимость учета факторов (эффектов), имеющих место в реальных социальных сетях. Как показано в [14], в реальных социальных сетях могут иметь место следующие эффекты и свойства, обусловленные как характеристиками и потребностями агентов (оказывающих влияние и подвергающихся влиянию), характером их взаимодействия, так и свойствами самой социальной сети:

(1) наличие собственных мнений агентов;

(2) изменение мнений под влиянием других членов социальной сети;

(3) различная значимость мнений (влиятельности, доверия) одних агентов для других агентов;

(4) различная степень подверженности агентов влиянию (конформизм, устойчивость мнений);

(5) существование «лидеров мнений» (агентов с максимальным «влиянием»);

(6) существование порога чувствительности к изменению мнения окружающих;

(7) воздействие структурных свойств социальных сетей на динамику мнений;

(8) активность (целенаправленное поведение) агентов;

(9) неполная и/или асимметричная информированность агентов.

Поэтому для корректного представления процесса распространения информации в СПТ ПКИИ модель должна поддерживать следующие характеристики агентов.

Роль: обычный/атакующий агент. Атакующий агент соответствует злоумышленнику, сознательно и активно распространяющему дезинформацию. Обычные агенты становятся объектами этой атаки. Роль есть постоянная характеристика агента.

Позиция: нормальная, искаженная, нейтральная. Относится к вопросу, мнение по которому атакующий агент пытается исказить. Нормальная позиция означает, что агент имеет сложившееся корректное, неискаженное мнение по данной теме. Искаженная позиция соответствует искаженному мнению, сложившемуся в результате информационной атаки. Нейтральная позиция означает, что у агента нет сложившегося мнения по данному вопросу. Позиция агента может меняться во времени.

Уверенность: степень уверенности агента в состоятельности своей позиции. Может меняться при воздействии мнений окружения агента. При достижении нулевого уровня уверенности агент меняет свою позицию. Ненулевой уровень уверенности агента с нейтральной позицией соответствует инертности его мышления.

Авторитетность: коэффициент, с которым воздействие агента оказывает влияние на позицию и уверенность соседа. По сути, служит коэффициентом ослабления воздействия. Хотя в принципе авторитетность может меняться с течением времени, можно без ограничения общности считать, что на протяжении одной информационной атаки данная характеристика может считаться константой. Возможны три варианта интерпретации авторитетности:

(1) независимая характеристика агента (объективная характеристика агента);

(2) независимая характеристика дуги «агент→агент» (субъективная характеристика агента);

(3) характеристика агента, зависящая от числа его исходящих связей (объективная характеристика, отражающая число участников сообщества, интересующихся мнением данного агента).

Активность: переменный коэффициент, отражающий степень воздействия агента на соседей. Исходно равен нулю, но при появлении воздействия на агента со стороны окружения становится положительным и остается положительным

до окончания информационной атаки. Растет при необходимости защищать свою позицию по мере роста уровня воздействия со стороны части окружения, имеющей позицию, отличную от позиции агента.

Воздействие: вычисляемая характеристика, оказывающее влияние на активность, уверенность и, в конечном итоге, позицию агента:

- исходящее воздействие агента пропорционально активности и имеет знак, соответствующий позиции (равно нулю при нейтральной позиции);
- входящее воздействие агента является суммой исходящих воздействий его окружения с учетом авторитетности воздействующих агентов.

Зависимость значений характеристик агента от уровня воздействия следует считать линейной и не учитывать дополнительные психологические факторы, такие как эффект «продолжительного влияния», эффект «обратного действия», эффект переполнения и склонность к избирательному потреблению информации [15] в связи с отсутствием их удовлетворительной модели. Правомерность такого допущения может быть окончательно подтверждена только при реальной эксплуатации технологии ПКИИ, но косвенным подтверждением может служить совпадение результатов моделирования в сравнимых режимах с результатами работы моделей, чья адекватность уже подтверждена.

5 Описание модели

5.1. Модель структуры сообщества формируется на базе ориентированного графа Боллобаша–Риордана [9]. В качестве начального графа G_0 выступает единственная вершина без дуг. На каждом шаге генерации в графе G появляется единственная дуга, с некоторой вероятностью на этом шаге может появиться и новая вершина. Для построения графа фиксируются некоторые неотрицательные параметры α, β, γ, δ_{in} и δ_{out} таким образом, что $\alpha + \beta + \gamma = 1$. Затем определяется граф $G(t)$, который в момент времени t имеет в точности t дуг и случайное число $n(t)$ вершин.

Правило предпочтительного связывания состоит в том, что выбор вершины v графа $G(t)$ согласно величине $d_{\text{out}} + \delta_{\text{out}}$ означает, что вероятность выбора

$$\mathsf{P}\,(v = v_i) = \frac{d_{\text{out}}(v_i) + \delta_{\text{out}}}{t + \delta_{\text{out}} n(t)}\,.$$

Аналогично выбор вершины v согласно величине $d_{\text{in}} + \delta_{\text{in}}$ означает, что вероятность

$$\mathsf{P}\,(v = v_i) = \frac{d_{\text{in}}(v_i) + \delta_{\text{in}}}{t + \delta_{\text{in}} n(t)}\,.$$

Здесь под $d_{\text{out}}(v_i)$ и $d_{\text{in}}(v_i)$ понимаются полустепень исхода и соответственно полустепень захода вершины v_i графа $G(t)$. Для $t \geq t_0$ из графа $G(t)$ формируется граф $G(t+1)$ в соответствии со следующими правилами роста.

Правило A. С вероятностью α добавляется новая вершина v совместно с дугой от вершины v к существующей вершине w, которая выбирается согласно $d_{\text{in}} + \delta_{\text{in}}$.

Правило B. С вероятностью β добавляется дуга между существующими вершинами v и w, где v и w выбираются независимо: v согласно $d_{\text{out}} + \delta_{\text{out}}$ и w согласно $d_{in} + \delta_{in}$.

Правило C. С вероятностью γ добавляется новая вершина w и дуга от существующей вершины v к вершине w, где вершина v выбирается согласно $d_{\text{out}} + \delta_{\text{out}}$.

5.2. При формировании графа для каждого узла исходно задаются следующие характеристики агентов:

- роль $r \in \{0, 1\}$: для всех устанавливается значение 0 (обычный агент), кроме случайного набора агентов A, для которых устанавливается значение 1 (атакующий). Число атакующих агентов $|A|$ — параметр модели;
- позиция $p \in \{-1, 0, 1\}$: устанавливается значение -1 (искаженная) для атакующих агентов. Для обычных агентов устанавливается значение 0 (нейтральная), за исключением случайного поднабора агентов N^+, для которых устанавливается значение 1 (нормальная). Размер поднабора $|N^+|$ — параметр модели;
- уверенность $u \in |0, +\infty|$: значение распределено логнормально для обычных агентов и равно $+\infty$ для атакующих агентов;
- авторитетность $b \in |0, 1|$: значение распределено нормально;
- активность $a = \langle a^{\min}, a^{\text{cur}} \rangle$; $a^{\min}, a^{\text{cur}} \in |0, 1|$: $a^{\min}$ — минимум активности агента после начала защиты позиции, значение которого распределено нормально, а a^{cur} — вычисляемый текущий уровень активности агента, исходно задаваемый как 0 для обычных агентов и 1 для атакующих.

Для каждой дуги d графа G задается авторитетность $p \in \{-1, 0, 1\}$: значение распределено нормально.

5.3. Работа модели осуществляется с дискретным временем. На каждом шаге используются два массива состояний модели: начальное и конечное состояние шага. Конечное состояние вычисляется на основе начального, после чего при переходе к следующему шагу конечное состояние копируется в начальное.

В зависимости от заданного как параметр модели варианта интерпретации авторитетности под u_{xy} будем понимать: (а) авторитетность u_y узла y; (б) авторитетность u_d дуги d_{yx}; (в) вычисляемую характеристику $\overline{u}$, зависящую от числа исходящих связей y:

$$\overline{u} = 1 - \frac{1}{(|O_y| + 1)^{\alpha}},$$

где α — параметр модели.

Алгоритм вычисления конечного состояния следующий.

1. Для каждого обычного узла $x \in G \backslash A$ вычисляется его окружение $O_x \subset G$: $\forall\, y \in O_x\ \exists\, d_{yx}$.
2. Для узла x вычисляется входящее воздействие:

$$v_x = \sum_{y \in O_x} p_y a_y^{\text{cur}} u_{xy} \beta\,,$$

где коэффициент β — параметр модели.

3. Если $v_x \neq 0$, то:
   - если p_x и v_x имеют один знак ($p_x v_x > 0$), то $u_x \leftarrow u_x + v_x$;
   - в противном случае $u_x \leftarrow u_x - |v_x|$. Если в результате $u_x < 0$, то $u_x \leftarrow |u_x|$, $p_x \leftarrow \operatorname{sgn}(v_x)$.
4. Для узла $x : p_x \neq 0$ вычисляется окружение с отличной от p_x позицией $O_x^- \subset O_x : y \in O_x^- \Rightarrow p_y \neq p_x$. Если $p_x = 0 \Rightarrow O_x^- = \emptyset$.
5. Для узла вычисляется входящее воздействие иного знака:

$$v_x^- = \sum_{y \in O_x^-} p_y a_y^{\text{cur}} u_{xy} \beta\,.$$

6. Актуализируется уровень активности:

$$v_x^- \neq 0 \Rightarrow a_x^{\text{cur}} = a_x^{\min} + \left(1 - a_x^{\min}\right)\left(1 - \frac{1}{(|v_x^-| + 1)^{\alpha}}\right).$$

6 Результаты моделирования

Анализ показал соответствие распределения узлов сгенерированного графа по числу их связей степенному закону с показателем степени, лежащим в диапазоне $|2, 3|$ и положительной ассортативностью (характером корреляции узлов, описываемым коэффициентом Пирсона), что типично для социальных сетей в целом и сетей сотрудничества ученых в частности [16].

Анализ структуры графа показал, что до 60% узлов сети могут быть недоступны для информационной атаки. Такое положение дел соответствуют реальным научно-профессиональным сетевым сообществам, в которых отдельный авторитетный специалист или некоторая компактная группа специалистов могут генерировать информацию, интересующую многих участников сообщества, но сами они при этом могут не интересоваться мнением менее авторитетных в данном вопросе коллег и быть не в курсе ведущихся среди них дискуссий.

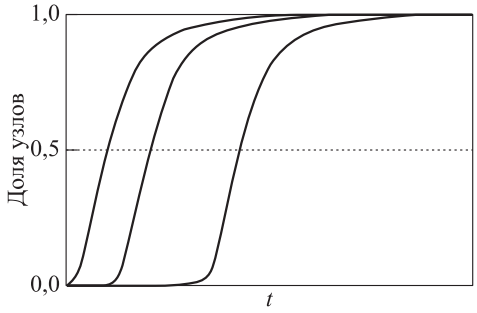


Рис. 1 Графики распространения дезинформации

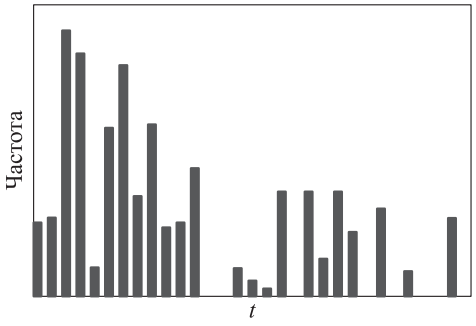


Рис. 2 Распределение последствий информационных атак

Была проведена серия расчетов модели со значением параметра $|N^+| = 0$, т. е. всем агентам, за исключением атакующего, была установлена нейтральная позиция. Это позволило сравнить результаты распространения искаженной информации в сети с результатами, полученными с помощью моделей «диффузии инноваций», адекватность которых и соответствие процессам распространения информации уже доказана [14].

На рис. 1 приведены три примера графиков распространения искаженной информации, показывающих динамику доли узлов с искаженной позицией в общем объеме уязвимых узлов. S-образная форма графика соответствует форме логистической кривой, отражающей процесс диффузии инноваций [14].

Сравнение графиков на рис. 1 демонстрирует значительный разброс результатов. На рис. 2 приведена гистограмма плотности распределения времени достижения 50%-ного порога распространения искаженной позиции среди уязвимых узлов.

Вид распределения не соответствует нормальному закону, что подтверждает нецелесообразность оценки рисков на базе средних значений.

Сравнение результатов для различных вариантов интерпретации авторитетности, описанных выше, показало, что при определенных значениях параметра α принципиальной разницы между ними нет. Это означает, что для дальнейших экспериментов может использоваться первый вариант — независимая характеристика агента.

7 Выводы

Результаты моделирования в режиме отсутствия противодействия информационным атакам соответствуют классическим моделям распространения информации в сообществах, что свидетельствует об адекватности построенной модели.

Значительный разброс результатов продемонстрировал преимущество построенной модели перед популярной SIR-моделью за счет возможности оценки

не только средних последствий, но и экстремумов, которые выступают фундаментальным источником риска.

Можно сделать вывод о перспективности модели для исследования процессов распространения искаженной информации между узлами СПТ ПКИИ с целью изучения эффективности мер противодействия и устойчивости технологии ПКИИ к попыткам искажения истории.

Литература

1. *Помникова А. Ю.* Семейная история в дискурсивном пространстве // Вестник Мининского ун-та, 2019. Т. 7. № 1. С. 9.
2. *Адамович И. М., Волков О. И.* Технология распределенного автоматизированного анализа исторических текстов // Системы и средства информатики, 2016. Т. 26. № 3. С. 148–161. doi: 10.14357/08696527160311.
3. *Адамович И. М., Волков О. И.* Единая технология поддержки конкретно-исторических исследований // Системы и средства информатики, 2019. Т. 29. № 1. С. 194–205. doi: 10.14357/08696527190116.
4. *Каменский Е. Г., Гримов О. А.* Сетевые сообщества в социальных сетях как фактор развития личностной субъектности // Вестник Нижегородского ун-та им. Н. И. Лобачевского. Сер. Социальные науки, 2014. № 2(34). С. 62–67.
5. *Wenger-Trayner E., Wenger-Trayner B.* Introduction to communities of practice: A brief overview of the concept and its uses, 2015. https://wenger-trayner.com/introduction-to-communities-of-practice.
6. *Лещев Е. Н., Харитонова Н. И.* Фальсификация истории как угроза национальной безопасности России: политический аспект // Среднерусский вестник общественных наук, 2016. Т. 11. № 6. С. 132–142.
7. *Васенин В. А., Афонин С. А., Панюшкин Д. С.* Модели распространения информации в социальных сетях // Программная инженерия, 2014. № 2. С. 33–42.
8. *Cirillo P., Taleb N.* Tail risk of contagious diseases // Nat. Phys., 2020. Vol. 16. P. 606–613.
9. *Бадрызлов В. А.* Классификация случайных графов с предпочтительным связыванием // Омский научный вестник, 2017. № 4(154). С. 124–128.
10. *Подлазов А. В., Щетинина Д. П.* Модель роста социальной сети // Препринты ИПМ им. М. В. Келдыша, 2013. № 95. С. 1–16.
11. *Barabási A.-L., Albert R.* Emergence of scaling in random networks // Science, 1999. Vol. 286. No. 5439. P. 509–512.
12. *Агиева М. Т.* Классификация моделей управления целевой аудиторией в маркетинге // Инженерный вестник Дона, 2019. № 1(52). С. 71.
13. *Bollobás B., Borgs C., Chayes T., Riordan O. M.* Directed scale-free graphs // 14th Annual ACM-SIAM Symposium on Discrete Algorithms Proceedings. — New York, NY, USA: ACM, 2003. P. 132–139.
14. *Губанов Д. А., Новиков Д. А., Чхартишвили А. Г.* Модели влияния в социальных сетях // Управление большими системами, 2009. № 27. С. 205–281.

15. *Богданова Д. А.* О дезинформации в интернет-эпоху // Наука. Информатизация. Технологии. Образование: Мат-лы XI Междунар. научн.-практич. конф., 2018. С. 317–322.
16. *Евин И. А., Хабибуллин Т. Ф.* Социальные сети // Компьютерные исследования и моделирование, 2012. Т. 4. № 2. С. 423–439.

Поступила в редакцию 01.10.20

THE MODEL OF COMMUNITY OF CONCRETE HISTORICAL INVESTIGATION SUPPORT TECHNOLOGY USERS

I. M. Adamovich and O. I. Volkov

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

Abstract: The article continues a series of works devoted to description and analysis of distributed technology of concrete historical investigation support based on the principles of crowdsourcing. This article is devoted to description and substantiation of the approach to modeling the community of technology users and the processes of spreading distorted information among its members to study the resistance of technology of concrete historical investigation support to attempts of history distortion, which is an urgent task in modern society. The proposed approach is to create a model of community structure on the basis of the Bollobas–Riordan directed graph. The principles of information spreading between graph nodes are based on factors and effects that occur in real social networks and are determined by the characteristics of their members, the nature of their interaction, and the properties of the social network. The adequacy of the model was verified by comparing the character of distorted information spread in the mode of resistance to information attacks lack with the logistic curve that shows the process of innovation diffusion.

Keywords: virtual community; model; technology; distortion of history; concrete historical investigation

DOI: 10.14357/08696527210112

References

1. Pomnikova, A. Yu. 2019. Semeynaya istoriya v diskursivnom prostranstve [Family stories in different types of discourse]. *Vestnik of Minin University B.* 7(1):9.
2. Adamovich, I. M., and O. I. Volkov. 2016. Tekhnologiya raspredelennogo avtomatizirovannogo analiza istoricheskikh tekstov [The distributed automated technology of historical texts analysis]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 26(3):148–161. doi: 10.14357/08696527160311.

3. Adamovich, I. M., and O. I. Volkov. 2019. Edinaya tekhnologiya podderzhki konkretno-istoricheskikh issledovaniy [Unified technology of concrete historical research support]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(1):194–205. doi: 10.14357/08696527190116.
4. Kamensky, E. G., and O. A. Grimov. 2014. Setevye soobshchestva v sotsial'nykh setyakh kak faktor razvitiya lichnostnoy sub"ektnosti [Network communities in social networks as a factor in the development of personal subjectivity]. *Vestnik of Lobachevsky State University of Nizhni Novgorod. Ser. Social sciences* 2(34):62–67.
5. Wenger-Trayner, E., and B. Wenger-Trayner. 2015. Introduction to communities of practice: A brief overview of the concept and its uses. Available at: https://wenger-trayner.com/introduction-to-communities-of-practice/ (accessed February 2, 2021).
6. Leshchev, E. N., and N. I. Kharitonova. 2016. Fal'sifikatsiya istorii kak ugroza natsional'noy bezopasnosti Rossii: politicheskiy aspekt [Falsification of history as threat of national security of Russia: Political aspect]. *Central Russian J. Social Sciences* 6(11):132–142.
7. Vasenin, V. A., S. A. Afonin, and D. S. Panushkin. 2014. Modeli rasprostraneniya informatsii v sotsial'nykh setyakh [Models of information dissemination in social networks]. *Programmnaya inzheneriya* [Software Engineering] 2:33–42.
8. Cirillo, P., and N. Taleb. 2020. Tail risk of contagious diseases. *Nat. Phys.* 16:606–613.
9. Badryzlov, V. A. 2017. Klassifikatsiya sluchaynykh grafov s predpochtitel'nym svyazyvaniem [Classification of random graphs with preferential attachment]. *Omskiy nauchnyy vestnik* [Omsk Scientific B.] 4(154):124–128.
10. Podlazov, A. V., and D. P. Shchetinina. 2013. Model' rosta sotsial'noy seti [The model of social network growth]. *Preprinty IPM im. M. V. Keldysha* [Keldysh Institute preprints] 95:1–16.
11. Barabási, A. L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439):509–512.
12. Agieva, M. T. 2019. Klassifikatsiya modeley upravleniya tselevoy auditoriey v marketinge [Analysis problems in the models on social networks in marketing]. *Engineering J. Don* 1(52):71.
13. Bollobás, B., C. Borgs, T. Chayes, and O. M. Riordan. 2003. Directed scale-free graphs. *14th Annual ACM-SIAM Symposium on Discrete Algorithms Proceedings*. New York, NY: ACM. 132–139.
14. Gubanov, D. A., D. A. Novikov, and A. G. Chkhartishvili. 2009. Modeli vliyaniya v sotsial'nykh setyakh [Models of influence in social networks]. *Upravlenie bol'shimi sistemami* [Large-Scale Systems Control] 27:205–281.
15. Bogdanova, D. A. 2018. O dezinformatsii v internet-epokhu [On misinformation in the Internet epoch]. *9th Scientific and Practical Conference (International) "Science. Informatization. Technologies. Education" Proceedings*. 317–322.
16. Yevin, I. A., and T. F. Khabibullin. 2012. Sotsial'nye seti [Social networks]. *Computer Research Modeling* 2(4):423–439.

Received October 1, 2020

Contributors

Adamovich Igor M. (b. 1934) — Candidate of Science (PhD) in technology, 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; Adam@amsd.com

Volkov Oleg I. (b. 1964) — leading programmer, 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; Volkov@amsd.com

ПОДХОД К СОВЕРШЕНСТВОВАНИЮ КОНЦЕПТУАЛЬНЫХ СХЕМ БАЗ ГЕОДАННЫХ ПОСРЕДСТВОМ МОДЕЛЕЙ ДЛЯ ПРОСТРАНСТВЕННО-ЛОГИЧЕСКОГО СВЯЗЫВАНИЯ ГЕООБЪЕКТОВ

Д. А. Никишин[1]

Аннотация: В контексте исследований по совершенствованию концептуальных схем представления топографической информации с целью развития процессов геоаналитики в перспективных географических информационных системах (ГИС) здесь рассмотрена одна из проблем, имеющих место в традиционных концептуальных схемах. Представлена расширенная типология видов пространственной локализации геообъектов, включающая составные геообъекты. В качестве решения предложено внедрение в структуру баз геоданных (БГД) специальных объектов данных, обеспечивающих явное пространственно-логическое связывание компонентов составных геообъектов и функциональных инфраструктур между собой.

Ключевые слова: концептуальные схемы баз геоданных; типы пространственной локализации геоданных; пространственно-логическое связывание геообъектов

DOI: 10.14357/08696527210113

1 Введение

Работа посвящена вопросам совершенствования концептуальных схем для перспективных БГД, призванных обеспечить потенциал для развития методологии анализа и отображения информации о местности в ГИС. При этом функциональность ГИС следует рассматривать как одно из важнейших свойств информационных, управляющих и телекоммуникационных систем (ИУТС). Данные исследования проводятся в рамках общего направления исследования особенностей информационных трансформаций в контексте парадигмы полиадического компьютинга [1–3].

В последние десятилетия пространственные данные все глубже проникают в различные отраслевые сферы. Тенденцией стало *возрастание роли сегмента геоаналитики* (аналитики пространственных данных), направленной в первую очередь на реализацию эффективного управления инфраструктурой. К тому же «анализ больших объемов пространственных данных с целью определения целевых ориентиров и принятия решений стал необходимым условием успеха

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, dmnikishin@mail.ru

бизнеса и научных открытий» [4]. Анализируя эти тенденции, можно ожидать, что дальнейшее развитие ГИС в целом и применяемых в ГИС моделей пространственных данных в частности со временем потребует качественных изменений (например, для задач, представленных в [5, 6]).

Как было показано в работе [7], некоторые существующие традиционные концептуальные схемы[1] оперируют в качестве модели для описания геообъектов не их аналитическим представлением, а моделями их картографического изображения — условными знаками (УЗ). Между ними есть различия, которые приводят к ряду следствий, в том числе к отсутствию в некоторых традиционных системах УЗ (СУЗ) полноценной, однозначной связи как между геообъектом и УЗ, так и между моделями отдельных геообъектов.

Примером такой концептуальной схемы может служить традиционная СУЗ топографической направленности. Ее структура и содержание фактически закреплены в нормативных документах [8–12]. Нормативный базис методологии применения этой СУЗ в контексте ГИС определяется группой стандартов [13–16].

Для первого случая примером могут служить геообъекты типа «овраг», «насыпь», «выемка», которые передаются несколькими структурными элементами (бровка, подошва), при этом явное связывание этих компонентов в более сложные геообъекты (овраг, вал и т. п.) не предусматривается, эта их связь может быть восстановлена лишь на ментальном уровне или путем сложных алгоритмов/процедур пространственного анализа.

Примером второго случая служат дороги, которые могут претерпевать разрывы в населенных пунктах, на мостах/эстакадах, насыпях и т. п. Процедуры восстановления их конфигурации также не тривиальны.

И, в-третьих, в большинстве существующих СУЗ не предусматривается связывание отдельных геообъектов компонентов в более сложные структуры (по функциональному, топологическому и другим признакам), например объединение дорог и других элементов в дорожную сеть. Это также ограничивает возможности анализа.

Соответственно, одним из направлений совершенствования применяемых в ГИС моделей пространственных данных может стать обогащение традиционных концептуальных схем механизмом явного пространственно-логического связывания компонентов моделей составных геообъектов и функциональных инфраструктур, что и является предметом рассмотрения в данной публикации.

2 Типология составных геообъектов

Модель геообъекта будем рассматривать как наличие в некоторой области пространства, заданной пространственной моделью геообъекта (его «метри-

[1] Концептуальная схема здесь понимается как система взаимосвязанных понятий, необходимая и достаточная для описания требуемого аспекта моделируемого объекта, включающая модели данных (классы геообъектов), сопряженные с ними ограничения и методы их обработки.

кой» [14]), определенной совокупности семантических свойств с конкретными их значениями. Значения метрических и семантических свойств в совокупности представляют собой *данные* модели этого геообъекта. При этом устойчиво встречающаяся совокупность этих свойств, применимая для достаточно большого множества геообъектов, может быть определена как *класс* геообъекта.

Рассмотрим модели геообъектов с точки зрения метрического аспекта модели геообъекта (описания его локализации в пространстве [13, 16]). Тогда типы метрики, используемые для моделирования геобъектов, можно разделить на две группы:

(1) элементарные (соответствующие традиционным типам метрики):

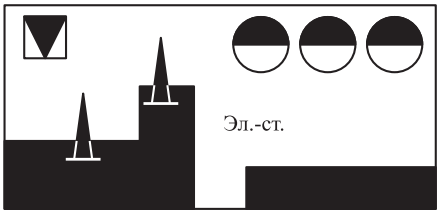
- точечный объект — означает наличие в данной точке (области пространства, не выражающейся в текущем масштабе описания) определенной совокупности семантических свойств, например малоразмерное строение, дерево и др.;
- линейный объект — наличие определенной совокупности свойств вдоль некоторой линии (полосы, ширина которой не выражается в текущем масштабе описания), например дорога, просека, ограда, линия границы и т. п.;
- площадной объект — обладание некоторой области пространства определенной совокупностью свойств, например лесной массив, сельхозугодие, административно-территориальное образование и т. п.;

(2) составной[1] геообъект (далее СГО) — состоящий из нескольких более простых геообъектов. Анализируя СУЗ, можно отметить, что во многих случаях представленные в ней типы геообъектов носят составной характер — они представляют собой совокупность нескольких компонентов более низкого уровня агрегации. Их примеры будут рассмотрены ниже, на рис. 1–3. Примером различия уровней агрегации могут служить сущности отдельного здания, квартала застройки и населенного пункта в целом.

Базой для построения составных геообъектов служат более простые объекты, которые, в свою очередь, могут представлять собой как составные структуры, так и элементарные объекты (примитивы). При этом речь идет именно о разных видах геообъектов, т. е. здесь не сопоставляются представления объектов определенного вида различными по сложности моделями метрики.

Отсутствие явного связывания как между компонентами таких составных геообъектов, так и между отдельными объектами в составе функциональных инфраструктур потенциально затрудняет их специфический анализ. Решением

[1]То, что имеется здесь в виду, соответствует определению «сложный объект» в [16], при этом термин «комплексный объект», не совсем удачно представленный в этом же смысле в [13], в контексте данной работы применен в другом качестве.



(*а*)

```
struct TRelationAggregate
{
    TGObjPtr* Components;
    TGObjPtr* RelatedSpatial;
    TGObjPtr* RelatedLogical;
};
```

(*б*)

Рис. 1 Иллюстрация составного объекта: (*а*) изображение; (*б*) структура данных для описания ПЛС

здесь может стать включение в БГД специальных структур данных, описывающих в явном виде пространственно-логические связи (ПЛС) между отдельными геообъектами.

Кроме того, на геообъект составного типа могут быть наложены особые правила формирования его структуры. На этой основе можно выделить следующие типы[1]:

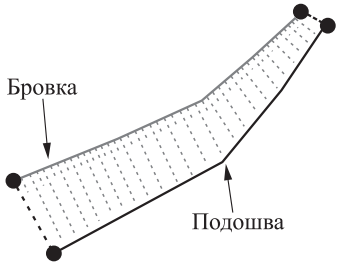
(а) объект как *совокупность* (англ. *aggregate*) компонентов, которая не накладывает каких-либо условий на количество и взаимное расположение и/или примыкание объектов-компонентов. Примерами могут служить: предприятие (рис. 1, *а*), включающее в себя контур владения (который может быть оформлен в виде ограды), а также комплекс зданий и сооружений (здания, трубы и другие сооружения). На рис. 1, *б* приведена специальная структура данных, предназначенная для всестороннего описания ПЛС данного объекта;

(б) объект как *комплекс* компонентов, конфигурация которого, в отличие от предыдущего случая, обусловлена особыми правилами организации его компонентов, например их определенным количеством и/или допустимыми вариантами их взаимного сопряжения.

Аналогичный подход используется, например, в структурах геоданных OSM (Open Street Map), где предусмотрено использование различных ролей для отдельных структурных элементов СГО: например, внутренний и внешний контуры для полигона с «островом». Другим применением является описание логических схем (например, маршрута транспорта [17] и т. п.).

Характерными примерами таких объектов-комплексов служат формы рельефа. Так, отдельные линии перегиба или перелома склона, выступающие как самостоятельные геообъекты, образуют собой отдельные элементарные склоны, которые, в свою очередь, могут быть объединены в формы рельефа различной конфигурации, вплоть до общей модели рельефа местности. Рассмотрим их подробнее.

[1] Речь идет о крайних случаях, между которыми может иметь место множество промежуточных вариантов, отличающихся по объему наложенных условий.

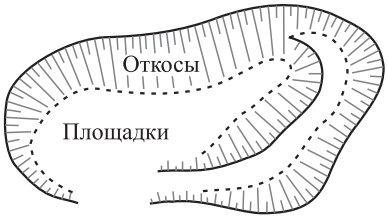


(*а*)

```
struct TRelationSlope
{
      TGObj* Edge;
      TGObj* Sole;
      TGObjPtr* RelatedSpatial;
      TGObjPtr* RelatedLogical;
};
```

(*б*)

Рис. 2 Иллюстрация комплексного геообъекта типа «склон», образованного отдельными структурными линиями: (*а*) изображение; (*б*) структура для описания ПЛС



(*а*)

```
struct TRelationCareer
{
      TGObjPtr* Slopes;
      TGObjPtr* Grounds;
      TGObjPtr* RelatedSpatial;
      TGObjPtr* RelatedLogical;
};
```

(*б*)

Рис. 3 Иллюстрация комплексного геообъекта, образованного отдельными склонами (карьер): (*а*) изображение; (*б*) структура для описания ПЛС

1. Склон, или откос[1,2], по своей конфигурации имеет сходство с полигональным типом метрики (рис. 2, *а*), но, в отличие от него, имеет две ветви контура: верхнюю, соответствующую линии бровки, и нижнюю, соответствующую линии подножия. При этом нужно отметить, что средствами СУЗ определяются только эти два несмежных участка склона, для образования полигона между ними необходимо восстановить недостающие связующие линии (показаны на рис. 2, *а* пунктиром). На рис. 2, *б* показана структура данных, связывающая эти объекты-линии в комплексный геообъект склона, а также описывающая его связи со смежными склонами — другими геообъектами.
2. Следующим уровнем агрегирования является форма рельефа, которая включает совокупность взаимосвязанных между собой элементов-склонов. Так, на рис. 3, *а* показан карьер, представляющий собой совокупность

[1] Здесь рассматриваются только склоны, имеющие выраженные линии бровки и подошвы.

[2] Вообще говоря, в отличие от искусственно спланированного откоса, склон, как правило, имеет более сложную конфигурацию: кривой (выгнутый или вогнутый) профиль, а также несколько различных «поясов» в профиле — например, осыпной склон включает обрыв и подстилающую его собственно осыпь.

откосов и пологих площадок. На рис. 3, *б* показана структура данных для описания ПЛС данного объекта, включающая множество геообъектов, описывающих склоны и специально выделяемых как геообъекты горизонтальных площадок.

3 Проблема агрегирования геообъектов

Как говорилось выше, недостаток традиционной СУЗ заключается в отсутствии возможности явного структурированного описания составных объектов в виде целостной модели геообъекта. Это препятствует полноценному описанию и анализу топографической ситуации, требования к которым в настоящее время повышаются.

Соответственно, характерные линии образуют собой не целостную модель склона, а совокупность отдельных склонов — оврага, бугра, ямы или другой формы рельефа. Анализ таких объектов как единого целого предполагает дополнительное восстановление таких объектов путем поиска и составления отдельных их элементов. Отдельные формы рельефа, в свою очередь, также могли бы быть объединены в общую модель рельефа местности, которая, по сути, стала бы топологически и логически более корректным аналогом традиционно используемых моделей рельефа типа TIN (triangular irregular network).

Характерными применениями такого подхода могут служить такие специфические типы объектов:

- мультиобъекты — не имеющие явных пространственных связей элементов, но имеющие логическую подоплеку, объединяющую их в единое целое; их можно считать частным случаем несвязной сети;
- полигон с неполной границей — когда предполагается объект типа «полигон», но его контур описывается лишь фрагментами линии: например, водоем, у которого не вся береговая линия доступна для топографической съемки.

Следующим уровнем данной проблемы является отсутствие полноценного связывания между собой отдельных взаимозависимых объектов местности. В настоящее время широко практикуется связывание объектов посредством «метрической согласованности» [16] (за счет совпадения местоположения узловых точек пространственного описания), что для задач геоаналитики становится уже неприемлемым [18, 19].

Примером такого связывания может служить последовательность отдельных участков дороги (с разным покрытием или даже разного класса) или проезжих частей многополосной дороги, образующих дорогу в целом, а эта дорога, в свою очередь, в числе прочих может быть включена в более сложный объект — дорожную сеть[1]. При этом следует иметь в виду различие связей в метрическом

[1] В контексте этого можно предложить ввести в концептуальную схему объекты, представляющие инфраструктуры в целом, например транспортную сеть, сеть связи и т. п.

(топологическом) и в семантическом (логическом, смысловом) плане; при этом возможны случаи, когда между элементами отсутствует видимая пространственная (топологическая) связь, но тем не менее логическая связь присутствует.

Несмотря на то что БГД ГИС потенциально дает возможность использования механизма явных ПЛС [13, 16], при переходе к ГИС традиционная номенклатура типов метрики моделей геообъектов практически не претерпела изменений: в основном используются все те же элементарные типы метрики — точки, линии и области. Универсальных, конвенциональных (общепринятых) методов для пространственно-логического связывания в явном виде в настоящее время нет, такое связывание применяется лишь на уровне конкретных, частных технологических решений.

Соответственно, *решением* этой проблемы и одним из направлений дальнейшего развития геомоделей представляется введение в структуру БГД специальных объектов данных, обеспечивающих явное пространственно-логическое связывание компонентов СГО и функциональных инфраструктур между собой.

Все это поможет осуществлять специфические виды геоанализа, а также проводить автоматизированный контроль целостности данных в БГД.

4 Заключение

Данная работа направлена на совершенствование функционала геоаналитики; конечным результатом исследований должна стать разработка концептуальной схемы для описания топографической ситуации, которая будет служить основой для совершенствования информационных трансформаций геоданных в контексте ИУТС.

В работе рассмотрена одна из проблем, присущая традиционным концептуальным схемам геоданных — проблема моделирования СГО. В качестве возможного подхода к ее решению и одного из направлений дальнейшего развития ГИС-технологий предложено введение специальных объектов данных, обеспечивающих явное пространственно-логическое связывание компонентов СГО и функциональных инфраструктур между собой. Такой подход, несмотря на некоторое усложнение структуры БГД, позволит обеспечить потенциал для совершенствования методов и технологий геоаналитики, а также автоматизировать контроль целостности геоданных; в настоящее время данная функциональность обеспечивается достаточно сложными методами.

Литература

1. *Rosenbloom P. S.* On computing: The fourth great scientific domain. — Cambridge, MA, USA: MIT Press, 2013. 307 p.
2. *Зацман И. М.* Методология обратимой генерализации в контексте классификации информационных трансформаций // Системы и средства информатики, 2018. Т. 28. № 2. С. 128–144.

3. *Гончаров А. А., Зацман И. М.* Информационные трансформации параллельных текстов в задачах извлечения знаний // Системы и средства информатики, 2019. Т. 29. № 1. С. 180–193.
4. *Xin Ch., Hoang V., Ablimit A., Fusheng W.* High performance integrated spatial big data analytics // 3rd ACM SIGSPATIAL Workshop (International) on Analytics for Big Geospatial Data Proceedings. — New York, NY, USA: ACM, 2014. P. 11–14. doi: 10.1145/2676536.2676538.
5. *Izham M. Y., Uznir U., Alias A. R., Ayob K.* Georeference, rainfall-runoff modeling and 3D dynamic simulation: Physical influence, integration and approaches // 1st Conference and Exhibition (International) on Computing for Geospatial Research & Application Proceedings. — New York, NY, USA: ACM, 2010. Art. No. 21. 8 p. doi: 10.1145/1823854.1823879.
6. *Jun I.* Lessons learned from data preparation for geographic information systems using Open Data // 14th Symposium (International) on Open Collaboration Proceedings. — New York, NY, USA: ACM, 2018. Art. No. 1. 5 p. doi: 10.1145/3233391.3233525.
7. *Дулин С. К., Никишин Д. А.* Подходы к интеграции прикладных концептуальных схем в составе унифицированной геоонтологии // Системы и средства информатики, 2020. Т. 30. № 2. С. 68–77.
8. Условные знаки для топографической карты масштаба 1:10 000. — М.: Недра, 1977. 143 с.
9. Условные знаки для топографических карт масштабов 1:25 000, 1:50 000, 1:100 000. — М.: ВТУ ГШ, 1983. 92 с.
10. Условные знаки для топографических карт масштабов 1:200 000 и 1:500 000. — М.: ВТУ ГШ, 1983. 56 с.
11. Условные знаки для топографических карт масштабов 1:5000, 1:2000, 1:1000, 1:500. — М.: Картгеоцентр, 2004. 286 с.
12. ГОСТ Р 52439-2005. Модели местности цифровые. Каталог объектов местности. Требования к составу. — М.: Стандартинформ, 2006. 58 с.
13. ГОСТ Р 50828-95. Геоинформационное картографирование. Пространственные данные, цифровые и электронные карты. Общие требования. — М.: Изд-во стандартов, 1996. 20 с.
14. ГОСТ Р 51605-2000. Карты цифровые топографические. Общие требования. — М.: Изд-во стандартов, 2000. 10 с.
15. ГОСТ Р 51606-2000. Карты цифровые топографические. Система классификации и кодирования цифровой картографической информации. — М.: Изд-во стандартов, 2000. 8 с.
16. ГОСТ Р 51607-2000. Карты цифровые топографические. Правила цифрового описания картографической информации. — М.: Изд-во стандартов, 2000. 10 с.
17. *Bast H., Brosi P., Storandt S.* Efficient generation of geographically accurate transit maps // ACM T. Spatial Algorithms Syst., 2019. Vol. 5. No. 4. Art. No. 25. 36 p. doi: 10.1145/3337790.
18. *Yu F., West G., Arnold L., McMeekin D., Moncrieff S.* Automatic geospatial data conflation using semantic web technologies // Australasian Computer Science Week Multiconference Proceedings. — New York, NY, USA: ACM, 2016. Art. No. 57. 10 p. doi: 10.1145/2843043.2843375.

19. *Yan M., Jing N., Zhong Z., Wu Y.* Geographical entity community mining based on spatial and semantic association // 3rd Conference (International) on Computer Science and Application Engineering Proceedings. — New York, NY, USA: ACM, 2019. Art. No. 92. 6 p. doi: 10.1145/3331453.3361652.

Поступила в редакцию 03.02.21

AN APPROACH TO IMPROVING THE CONCEPTUAL SCHEMES OF GEODATA BY MEANS OF MODELS FOR SPATIAL AND LOGICAL LINKING OF GEOOBJECTS

D. A. Nikishin

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

Abstract: In the context of research on improving the conceptual schemes for the representation of topographic information to develop the processes of geoanalytics in promising geographical information systems, one of the problems that occur in traditional conceptual schemes is considered here. An extended typology of types of spatial localization of geoobjects, including composite geoobjects, is presented. As a solution, it is proposed to introduce special data objects into the structure of the geodata databases which provide an explicit spatial-logical link between the components of composite geoobjects and functional infrastructures.

Keywords: conceptual schemes of geodata databases; types of spatial localization of geodata; spatial-logical linking of geoobjects

DOI: 10.14357/08696527210113

References

1. Rosenbloom, P. S. 2013. *On computing: The fourth great scientific domain*. Cambridge, MA: MIT Press. 307 p.
2. Zatsman, I. M. 2018. Metodologiya obratimoy generalizatsii v kontekste klassifikatsii informatsionnykh transformatsiy [Methodology of reversible generalization in the context of classification of information transformations]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(2):128–144.
3. Goncharov, A. A., and I. M. Zatsman. 2019. Informatsionnye transformatsii parallel'nykh tekstov v zadachakh izvlecheniya znaniy [Information transformations of parallel texts in problems of knowledge extraction]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(1):180–193.
4. Xin, Ch., V. Hoang, A. Ablimiti, and F. Wang. 2014. High performance integrated spatial big data analytics. *3rd ACM SIGSPATIAL Workshop (International) on Analytics for Big Geospatial Data Proceedings*. New York, NY: ACM. 11–14. doi: 10.1145/2676536.2676538.

5. Izham, M. Y., U. Uznir, A. R. Alias, and K. Ayob. 2010. Georeference, rainfall-runoff modeling and 3D dynamic simulation: Physical influence, integration and approaches. *1st Conference and Exhibition (International) on Computing for Geospatial Research & Application Proceedings*. New York, NY: ACM. Art. No. 21. 8 p. doi: 10.1145/1823854.1823879.
6. Jun, I. 2018. Lessons learned from data preparation for geographic information systems using open data. *14th Symposium (International) on Open Collaboration Proceedings*. New York, NY: ACM. Art. No. 1. 5 p. doi: 10.1145/3233391.3233525.
7. Dulin, S. K., and D. A. Nikishin. 2020. Podkhody k integratsii prikladnykh kontseptual'nykh skhem v sostave unifitsirovannoy geoontologii [Approaches to the integration of the application conceptual schemas in the unified geoontology]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 30(2):68–77.
8. Uslovnye znaki dlya topograficheskoy karty masshtaba 1:10 000 [Conditional signs for a topographic map of scale 1:10 000]. 1977. Moscow: Nedra. 143 p.
9. Uslovnye znaki dlya topograficheskikh kart masshtabov 1:25 000, 1:50 000, 1:100 000 [Symbols for topographic scale maps 1:25 000, 1:50 000, and 1:100 000]. 1983. Moscow: VTU GSh. 92 p.
10. Uslovnye znaki dlya topograficheskikh kart masshtabov 1:200 000 i 1:500 000 [Conditional signs for topographic maps of scales 1:200 000 and 1:500 000]. 1983. Moscow: VTU GSh. 56 p.
11. Uslovnye znaki dlya topograficheskikh kart masshtabov 1:5000, 1:2000, 1:1000, 1:500 [Symbols for topographic maps of scale 1:5000, 1:2000, 1:1000, and 1:500]. 2004. Moscow: Kartgeotsentr. 286 p.
12. GOST R 52439-2005. 2006. Modeli mestnosti tsifrovye. Katalog ob"ektov mestnosti. Trebovaniya k sostavu [Terrain models digital. Catalog of locality objects. Requirements for the composition]. Moscow: Standardinform Publs. 58 p.
13. GOST R 50828-95. 1996. Geoinformatsionnoe kartografirovanie. Prostranstvennye dannye, tsifrovye i elektronnye karty. Obshchie trebovaniya [Geoinformation mapping. Spatial data, digital and electronic maps. General requirements]. Moscow: Standards Publs. 20 p.
14. GOST R 51605-2000. 2000. Karty tsifrovye topograficheskie. Obshchie trebovaniya [Digital topographic maps. General requirements]. Moscow: Standards Publs. 10 p.
15. GOST R 51606-2000. 2000. Karty tsifrovye topograficheskie. Sistema klassifikatsii i kodirovaniya tsifrovoy kartograficheskoy informatsii [Digital topographic maps. System of classification and coding of digital cartographic information]. Moscow: Standards Publs. 8 p.
16. GOST R 51607-2000. 2000. Karty tsifrovye topograficheskie. Pravila tsifrovogo opisaniya kartograficheskoy informatsii [Digital topographic maps. Rules of digital description of cartographic information]. Moscow: Standards Publs. 10 p.
17. Bast, H., P. Brosi, and S. Storandt. 2019. Efficient generation of geographically accurate transit maps. *ACM T. Spatial Algorithms Syst.* 5(4):25. 36 p. doi: 10.1145/3337790.
18. Yu, F., G. West, L. Arnold, D. McMeekin, and S. Moncrieff. 2016. Automatic geospatial data conflation using semantic web technologies. *Australasian Computer Science Week Multiconference Proceedings*. New York, NY: ACM. Art. No. 57. 10 p. doi: 10.1145/2843043.2843375.

19. Yan, M., N. Jing, Z. Zhong, and Y. Wu. 2019. Geographical entity community mining based on spatial and semantic association. *3rd Conference (International) on Computer Science and Application Engineering Proceedings*. New York, NY: ACM. Art. No. 92. 6 p. doi: 10.1145/3331453.3361652.

Received February 3, 2021

Contributor

Nikishin Dmitry A. (b. 1976) — Candidate of Science (PhD) in technology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; dmnikishin@mail.ru

SIR-МОДЕЛЬ КАК ИНСТРУМЕНТ ИССЛЕДОВАНИЯ ДЕСТРУКТИВНЫХ ПРОЦЕССОВ ПРИ УСВОЕНИИ НОВОГО ЗНАНИЯ

О. М. Корчажкина[1]

Аннотация: Описывается подход к анализу способов усвоения нового знания, основанный на математическом моделировании учебной ситуации, которая представлена по образцу SIR (susceptible–infected–recovered) модели Уильяма Кермака и Андерсона МакКендрика, изначально использованной для прогнозирования процесса распространения эпидемии на обширные замкнутые группы населения с целью предотвращения губительных последствий всеобщего заражения. Модифицированная SIR-модель позволяет изучить ряд закономерностей познавательного процесса и выявить способы поведения динамической системы в виде замкнутого малочисленного ученического сообщества, когда работа с избыточной информацией может привести к когнитивной перегрузке и ошибкам в решении учебных задач. Подход реализуется как численно-графический эксперимент на фазовой плоскости, позволяющий составить целостную картину явления и проанализировать условия движения системы к состоянию устойчивого равновесия, что эквивалентно преодолению когнитивной перегрузки.

Ключевые слова: познавательный процесс; новое знание; когнитивная перегрузка; замкнутое ученическое сообщество; модель Кермака–МакКендрика; SIR-модель; фазовая плоскость/портрет; устойчивость/неустойчивость системы

DOI: 10.14357/08696527210114

1 Введение

В настоящее время проблема усвоения учащимися предметного знания рассматривается с новых позиций, учитывающих общий объем информации, который требуется для овладения предметом. Доступность многочисленных учебных материалов и увеличение времени, отводимого на самостоятельную работу, приводит и учителя, и учащихся к необходимости постоянного отбора ресурсов — источников полезного знания. Однако учащиеся, находящиеся под воздействием нарастающего информационного потока, когда у них еще не сформированы в достаточной степени умения «фильтрации» полезной и избыточной информации, ее критическая оценка и классификация по определенным критериям,

[1]Институт кибернетики и образовательной информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, olgakomax@gmail.com

могут быть дезориентированы, что приводит к потере мотивации при изучении предмета. Поэтому одной из проблем, стоящих перед учителем, становится оценка объема нового учебного материала, предъявляемого учащимся для усвоения.

Цель настоящей статьи — предложить математическую модель, с помощью которой можно оценивать условия нивелирования негативного влияния избыточной учебной информации при усвоении нового знания в замкнутом ученическом сообществе.

2 Постановка задачи

Рассмотрим подход к моделированию процесса усвоения нового знания в ученическом сообществе, когда учитель взаимодействует с группой учащихся, работающих совместно над решением единой задачи (например, проектом) или усваивающих определенную тему в рамках предметной программы. В основу модели положим процессы, к которым приводит деструктивное воздействие на учащихся избыточной (т. е. излишней, ненужной) информации.

Предлагаемая модель базируется на модели Кермака–МакКендрика, или SIR-модели, с помощью которой в медицинской статистике описывается динамика охвата замкнутой группы населения инфекционными заболеваниями с целью анализа и прогнозирования дальнейшего течения инфекции — ее угасания или распространения [1].

С развитием информационных технологий модель Кермака–МакКендрика, предложенная в 1927 г., стала применяться не только в области медицины. В настоящее время известно некоторое число исследований, привлекающих SIR-модель для предотвращения «эпидемий» в виртуальной среде, например:

- для исследования влияния вредоносного программного обеспечения (ВПО), вызванного неконтролируемым распространением вирусов по узлам сети, для оценки объемов трафика, генерируемого ВПО [2];
- для оценки структурных способов защиты компьютерных сетей от вирусных атак [3];
- для других инженерных решений [4].

В работе [5] исследуется динамическое развитие структуры замкнутой социальной группы, находящейся под негативным информационным воздействием, и делается прогноз относительно возможности «социального взрыва» в исследуемой группе, а коллектив авторов [6] привлекает данную модель для анализа работы динамических систем в биологии.

Тем не менее с помощью SIR-модели еще не рассматривались **закономерности познавательного процесса, осуществляемого в замкнутом ученическом сообществе, которые приводят к деструктивным явлениям, вызванным неумением учащихся работать с избыточной информацией**.

3 Элементы классической SIR-модели

Классическая SIR-модель описывается следующей системой уравнений[1]:

$$\frac{\partial S}{\partial t} = \dot{S} = -\beta SI\,; \tag{1}$$

$$\frac{\partial I}{\partial t} = \dot{I} = \beta SI - \gamma I\,; \tag{2}$$

$$\frac{\partial R}{\partial t} = \dot{R} = \gamma I\,. \tag{3}$$

Она описывает поведение некоей замкнутой группы испытуемых N_D, находящихся в зоне риска, т. е. в зоне распространения инфекции D. В некоторый момент времени t, находящийся в пределах отрезка наблюдения $[0,\, T]$, ее членов можно разделить на три группы:

(1) $S(t)$ — уязвимые испытуемые, которые на момент времени t здоровы, однако, находясь в группе риска, могут быть инфицированы вирусом D с высокой степенью вероятности (восприимчивая группа);

(2) $I(t)$ — инфицированные испытуемые, которые определенно являются переносчиками инфекции D (инфицированная группа);

(3) $R(t)$ — испытуемые, которые выздоровели, приобретя тем самым иммунитет к инфекции D, а также те испытуемые, которые не справились с вирусом D и скончались (группа выбывших испытуемых).

Очевидно, что процесс взаимодействия испытуемых с инфекцией D может быть выражен динамической формулой: $S \to I \to R$, что отражает процесс перехода испытуемых из одной группы в другую. В пределе (по окончании эпидемии) все испытуемые концентрируются в группе выбывших R, контингент которой состоит как из полностью излечившихся членов, так и умерших, не справившихся с инфекцией. При этом общее число испытуемых остается постоянным и равным N_D.

Для анализа процесса инфицирования оперируют несколькими производными функций $S(t)$, $I(t)$ и $R(t)$, определяющи ми скорость перехода испытуемых из одной группы в другую, или скорость инфицирования и выздоровления или смерти: $\partial S/\partial t = \dot{S}$ — скорость заражения испытуемых в группе S (работа на «выход»); $\partial I/\partial t = \dot{I}$ — скорость движения испытуемых в группе I (работа на «вход–выход»); $\partial R/\partial t = \dot{R}$ — скорость пополнения группы R за счет испытуемых из группы I (работа на «вход»).

В формулах (1)–(3) величины этих производных зависят как от начальных значений самих функций $S(0)$ и $I(0)$, так и от некоторых параметров процесса β

[1] Более подробно см. в [1; 5; 6, с. 157–177].

и γ. В SIR-модели параметр β — это эффективная частота контактов в данной популяции во время эпидемии D. Она показывает, какая часть общего числа контактов в единицу времени, обозначенная через γ, привела к заражению.

4 Специфика использования SIR-модели для исследования информационных процессов в замкнутой малочисленной ученической группе

В контексте настоящего исследования представляет интерес, как будет вести себя динамическая система в виде замкнутой малочисленной (в один–три десятка человек) ученической группы при решении учебно-познавательной задачи, когда относительный объем избыточной информации в общем потоке учебной информации может стать критическим и привести к взрывному росту когнитивной перегрузки[1] в процессе восприятия, переработки и усвоения учебного материала. Для этого сформулируем исходные данные, отвечающие SIR-модели, в которой аналогом инфекции D служит избыточная информация E.

1. Пусть N_E — общее число учащихся рассматриваемой замкнутой группы (курса, класса), находящихся в данном информационном пространстве, например в информационной образовательной среде, и подверженных воздействию непрерывно поступающей учебной информации, содержащей определенный объем избыточной информации E.
2. Каждый из N_E учащихся замкнутого ученического сообщества внутри отрезка времени $[0, T]$, где T — это период времени, отведенный на изучение материла или решение учебной задачи, может находиться в одном из трех различных состояний, образующих открытые группы $S(t)$, $I(t)$ и $R(t)$. Кроме того, при определенных обстоятельствах, вызванных особенностями учебного процесса, и в зависимости от скорости и глубины усвоения учебного материала учащиеся могут переходить из одной группы в другую, причем для каждого момента времени t выполняется очевидное условие: $N_E = S(t) + I(t) + R(t)$.
3. При $t = 0$ имеем: $S(0) > 0$, $I(0) > 0$ и $R(0) = 0$, т. е. изначально все учащиеся сконцентрированы в двух группах — S и I: $S(0) + I(0) = N_E$. А по окончании изучения темы или решения общей учебной задачи группы S и I пустеют, поскольку все учащиеся переходят в группу R: $S(T) = 0$; $I(T) = 0$; $R(T) = N_E$.

[1] Когнитивная перегрузка — психологическое состояние человека, подверженного деструктивному влиянию избыточной информации, которую он не в силах воспринять, переработать или усвоить. При когнитивной перегрузке возрастает объем недостоверного или недополученного знания за счет накопления ошибок, которые допускаются и не отрабатываются учащимися в ходе познавательной деятельности, что приводит к потере полезной информации, тормозящей процесс усвоения новых знаний.

4. По окончании изучения темы или решения учебной задачи все учащиеся перемещаются в группу $R(T)$, и по результатам работы их можно разделить на три подгруппы (две из группы S и одна из группы I):

   подгруппа а: учащиеся, которые сразу перешли из группы S в группу R, минуя группу I, поскольку изначально имели высокий интеллектуальный уровень и полученный ранее позитивный опыт работы с избыточной информацией (подобно стойкому иммунитету к инфекции); в формулах (4) и (6) ниже они представлены слагаемым αSR, где α указывает на «*неэффективное*» *число контактов* с избыточной информацией в единицу времени[1], совершенное учащимися группы S, относительно всех N учащихся; величину α можно назвать **коэффициентом потенциальной защищенности от когнитивной перегрузки**, что означает полное отсутствие у этой части учащихся когнитивной перегрузки;

   подгруппа б: учащиеся из группы S, принадлежащие к «группе риска», т. е. те, что могли быть «инфицированы» избыточной информацией и перешли сначала в группу I; затем они сумели преодолеть трудности и, проанализировав свои ошибки, перешли в группу R; в формулах (4) и (5) ниже они представлены слагаемым βSI, где величина β — это «*эффективное*» *число контактов* с избыточной информацией в единицу времени, совершенное учащимися группы I, относительно всех N учащихся; величину β назовем **коэффициентом потенциальной подверженности когнитивной перегрузке**;

   подгруппа в: суммарный контингент учащихся, включающий как тех, что изначально находились в группе I, т. е. уже были «инфицированы» избыточной информацией, так и учащихся из «группы риска», перешедших в группу I из группы S; в формулах (5) и (6) ниже они представлены слагаемым γI; где величина γ, называемая, согласно SIR-модели, *коэффициентом выздоровления*, указывает на относительное число учащихся из общего числа N, изначально подвергшихся когнитивной перегрузке в единицу времени, а затем преодолевших ее; назовем $\boldsymbol{\gamma}$ **показателем преодоления когнитивной перегрузки**[2].

Тогда система дифференциальных уравнений (1)–(3) приобретает следующий вид:

[1] В данной задаче за единицу времени следует принять академический или астрономический час.

[2] По аналогии с моделью распространения эпидемии, где γ — общее число контактов в единицу времени, а $1/\gamma$ выступает как среднее время болезни, если продолжительность болезни распределена по экспоненциальному закону [6, с. 159], можно принять, что продолжительность преодоления когнитивной перегрузки также распределена по экспоненциальному закону со средним временем $1/\gamma$.

$$\frac{\partial S}{\partial t} = \dot{S} = -\beta SI - \alpha SR\,; \tag{4}$$

$$\frac{\partial I}{\partial t} = \dot{I} - \beta SI - \gamma I\,; \tag{5}$$

$$\frac{\partial R}{\partial t} = \dot{R} = \gamma I + \alpha SR\,. \tag{6}$$

В [6, с. 160] отмечается, что в классической SIR-модели уравнение, имеющее вид (3), является избыточным. Поэтому, исключив подобное уравнение (6) из системы (4)–(6) и учитывая, что $R = N - S - I$, получим следующую результирующую систе -му обыкновенных дифференциальных уравнений, которую используем для исследования процессов, происходящих в замкнутой малочисленной ученической группе при работе с избыточной информацией:

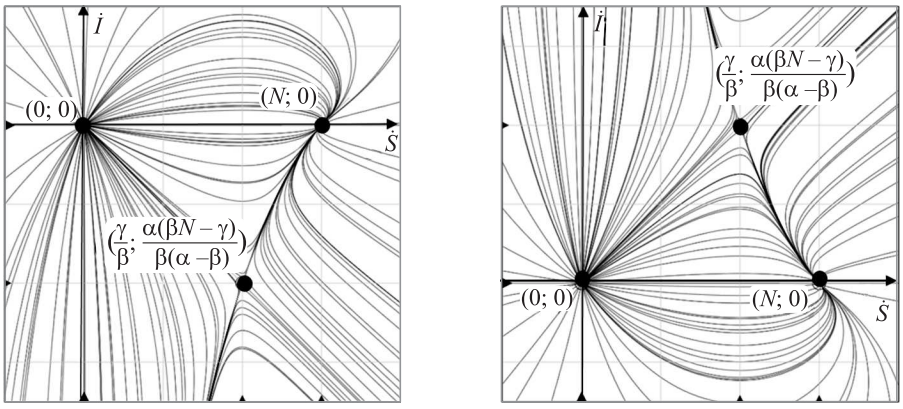
$$\dot{S} = [\alpha S + (\alpha - \beta)I - \alpha N]\, S\,; \tag{7}$$

$$\dot{I} = (\beta S - \gamma)I\,. \tag{8}$$

Процессы в замкнутых группах по SIR-модели подчиняются законам развития динамических систем, представляющих собой соответствующие реальным системам математические объекты, эволюция которых определяется начальными условиями, т. е. значениями искомых функций в начальный момент времени, или временны́х начальных условий. Для описания поведения подобных систем обычно используют их фазовые портреты (ФП), которые демонстрируют фазовые траектории — векторные пути изменения параметров системы во времени. В публикациях, посвященных биологическим и информационным процессам в больших популяциях с числом участников в десятки и сотни тысяч человек, исследование обычно осуществляется путем анализа системы дифференциальных уравнений типа (1)–(3) и построением графиков функций $S(t)$, $I(t)$ и $R(t)$ [1–3, 5, 6], тогда как для малочисленных групп испытуемых метод построения ФП представляется более наглядной иллюстрацией происходящих в них процессов.

5 Результаты исследования динамической системы методом фазовых портретов

Задача численно-графического исследования динамической системы, каковой является замкнутая малочисленная ученическая группа, подверженная воздействию избыточной учебной информации, состоит в разбиении фазовой плоскости $(\dot{S}, \dot{I})$ в зависимости от параметров α, β и γ на целевые окрестности и изучении бифуркаций, происходящих внутри этих областей, с целью определения устойчивости равновесных состояний системы.



(*а*) (*б*)

Рис. 1 Фазовые портреты SIR-модели на фазовой плоскости $(\dot{S}, \dot{I})$ при $\alpha < \beta$ (*а*) и $\alpha > \beta$ (*б*)

Расчеты и построения ФП, проведенные в веб-среде PhaPl [7, 8], показали, что система дифференциальных уравнений (7)–(8) при любых соотношениях параметров α, β и γ имеет три точки равновесия — особые точки (ОТ): ОТ_1 с координатами $(0; 0)$ — асимптотически устойчивый дикритический узел; ОТ_2 с координатами $(\gamma/\beta; \alpha(\beta N - \gamma)/(\beta(\alpha - \beta))$ — неустойчивое седло; ОТ_3 с координатами $(N; 0)$ — неустойчивый узел (рис. 1). Области ФП, обеспечивающие устойчивую работу системы вблизи ОТ_1 $(0; 0)$, располагаются между осью ординат и одной из сепаратрис ОТ_2. Это означает, что если начальные условия, отвечающие значениям $\dot{S}_0$ и $\dot{I}_0$, будут заданы в пределах этой области, то система придет в устойчивое состояние равновесия, локализуясь вблизи ОТ_1, а когнитивная перегрузка, вызванная избыточной информацией, будет преодолена всеми N учащимися за некоторый промежуток времени, определяемый временем движения от начального состояния к ОТ_1.

Примеры наиболее реальных случаев представлены на графиках рис. 2, где α — коэффициент потенциальной защищенности от когнитивной перегрузки («неэффективное» число контактов с избыточной информацией в единицу времени, совершенное всеми учащимися группы численностью N); β — коэффициент потенциальной подверженности когнитивной перегрузке («эффективное» число контактов с избыточной информацией в единицу времени, совершенное всеми учащимися группы численностью N); γ — показатель преодоления когнитивной перегрузки (число учащихся из группы численностью N, преодолевших когнитивную перегрузку в единицу времени). Следует обратить внимание, что значения параметров α, β и γ выбраны с учетом малочисленности группы ($N = 20$).

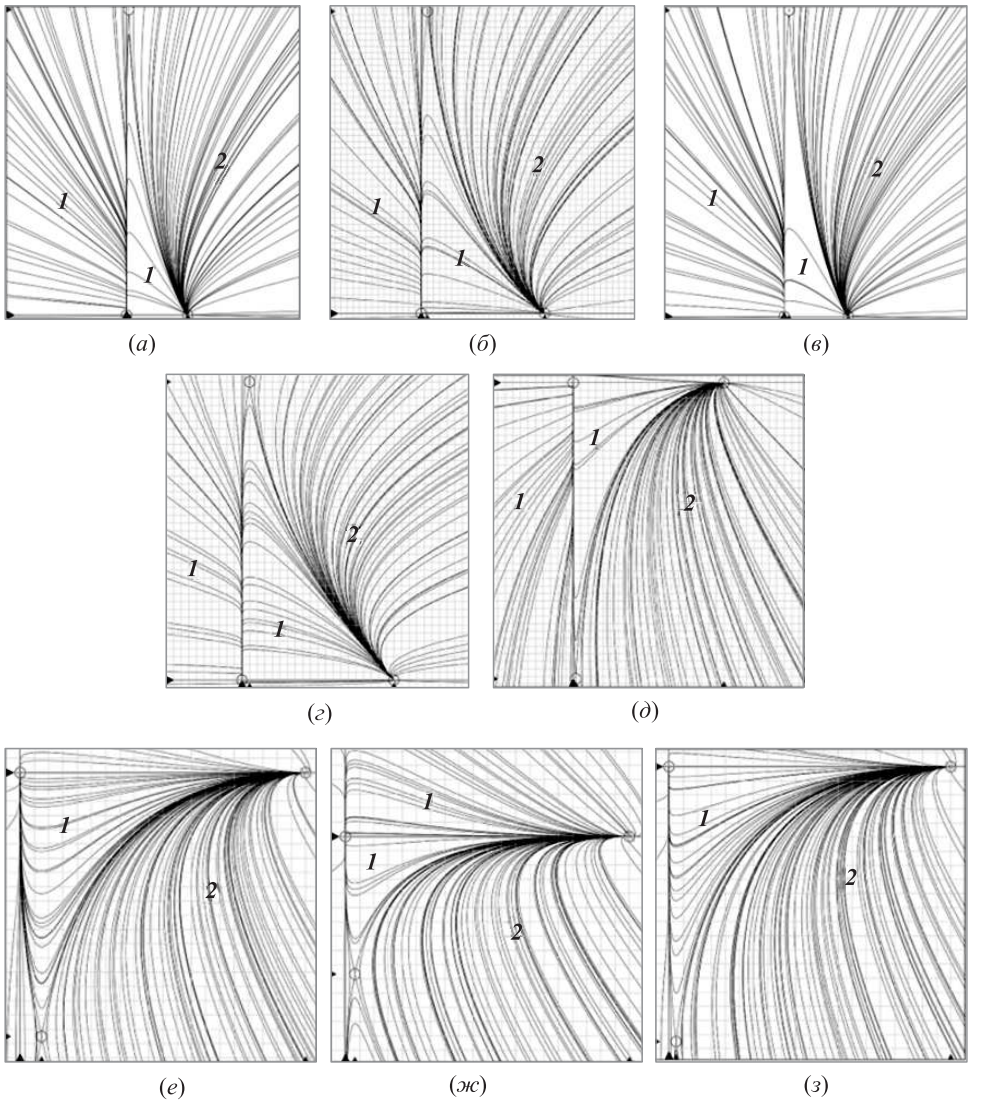
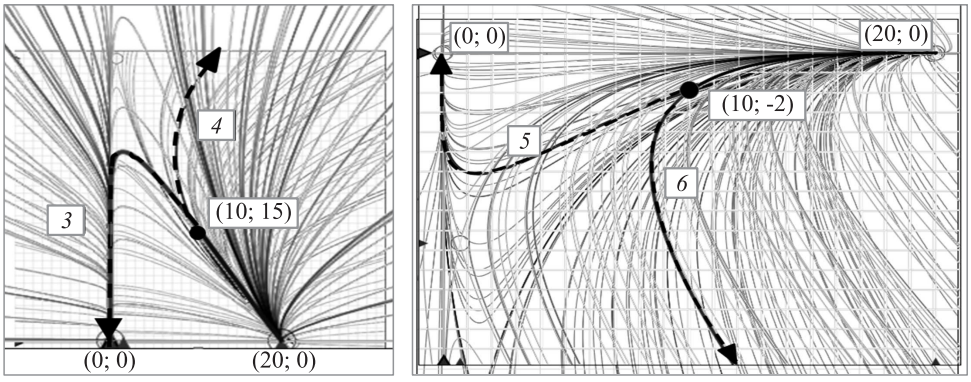


Рис. 2 Фазовые портреты динамической системы при $N = 20$: *1* — области устойчивости; *2* — области неустойчивости; (*а*) $\alpha = 5$; $\beta = 4$, $\gamma = 3$; (*б*) $\alpha = 5$; $\beta = 3$, $\gamma = 4$; (*в*) $\alpha = 5$; $\beta = 4$, $\gamma = 6$; (*г*) $\alpha = 2$; $\beta = 1$, $\gamma = 1$; (*д*) $\alpha = 2$; $\beta = 3$, $\gamma = 1$; (*е*) $\alpha = 1$; $\beta = 2$, $\gamma = 3$; (*ж*) $\alpha = 1$; $\beta = 3$, $\gamma = 2$; (*з*) $\alpha = 1$; $\beta = 2$, $\gamma = 1$

Фазовые портреты на рис. 1 и 2 демонстрируют наличие трех ОТ разного типа с шестью целевыми окрестностями, расположенными в первом и четвертом квадрантах ФП. Поэтому движение к той или иной ОТ будет зависеть от того, из какой из этих окрестностей начнется движение по соответствующей фазовой траектории в момент времени $t = 0$, т. е. от начальной точки с координатами $(\dot{S}_0, \dot{I}_0)$. В эти координаты, в свою очередь, в неявном виде заложен объем избыточной информации, который через соотношение параметров N, α, β и γ либо преодолевается системой, либо нет. Из этой начальной точки на ФП начинается движение во времени по проходящей через нее фазовой траектории. Продолжительность переходного процесса определяется, как показывают ФП на рис. 2, углом наклона сепаратрис неустойчивого седла ОТ_2, поскольку именно эта ОТ формирует основное пространство равновесных состояний системы.

Если точка $(\dot{S}_0, \dot{I}_0)$ находится в зоне захвата ОТ_1, которая является единственной устойчивой точкой, то система с определенной скоростью придет в состояние устойчивого равновесия (области 1 на рис. 2). Это будет означать, что всем учащимся группы удалось преодолеть когнитивную перегрузку. Если же начальная точка находится вне области захвата ОТ_1, то она уходит в бесконечность по одной из соответствующих траекторий (области 2 на рис. 2). Это означает, что объем избыточной учебной информации приводит к возникновению непреодолимой когнитивной перегрузки, препятствующей усвоению актуального учебного материала большинством учащихся группы.

На рис. 3 приведены примеры двух соотнесенных по осям координат и попарно наложенных друг на друга ФП исследуемой системы.



(*а*) (*б*)

Рис. 3 Движение системы в окрестностях ОТ при $(\dot{S}_0, \dot{I}_0) = (10; 15)$ (*а*) и $(\dot{S}_0, \dot{I}_0) = (10; -2)$ (*б*): *3* — $\alpha = 5$, $\beta = 4$, $\gamma = 6$ (см. рис. 2, *в*); *4* — $\alpha = 2$, $\beta = 1$, $\gamma = 1$ (см. рис. 2, *г*); *5* — $\alpha = 1$, $\beta = 2$, $\gamma = 3$ (см. рис. 2, *е*); *6* — $\alpha = 1$, $\beta = 3$, $\gamma = 2$ (см. рис. 2, *ж*)

В группах, где большинство учащихся «сильные» ($\alpha > \beta$) и высок показатель преодоления когнитивной перегрузки γ, может выполняться любое из соотношений: $\gamma > \alpha > \beta$ или $\alpha > \gamma > \beta$, т. е. число подвергшихся когнитивной перегрузке в единицу времени β меньше γ — числá ее преодолевших в ту же единицу времени. На ФП для этих групп получаем бóльшие по площади зоны устойчивого равновесия и меньшее время преодоления когнитивной перегрузки.

Для групп с преимущественно слабым составом учащихся ($\alpha < \beta$) справедливы соотношения: $\gamma < \alpha < \beta$ или $\alpha < \gamma < \beta$, т. е. число подвергшихся когнитивной перегрузке в единицу времени β больше числá ее преодолевших в ту же единицу времени γ. Тогда площадь зоны устойчивого равновесия снижается, а время преодоления когнитивной перегрузки растет.

6 Заключение

Численно-графическое исследование, проведенное на фазовой плоскости, показало, что SIR-модель, изначально имевшая целью анализ условий распространения эпидемии в обширных группах населения, при определенных допущениях применима к описанию специфики протекания деструктивных информационных процессов в замкнутых малочисленных ученических группах. С этой целью была представлена модификация системы обыкновенных дифференциальных уравнений классической SIR-модели путем введения информационных параметров: коэффициента потенциальной защищенности учащихся от когнитивной перегрузки α; коэффициента потенциальной подверженности учащихся когнитивной перегрузке β; показателя преодоления учащимися когнитивной перегрузки γ.

Исследование предложенной модели с помощью метода фазовой плоскости показало, что ФП системы однотипны и не зависят от значений информационных параметров для ученической группы с фиксированной численностью учащихся N. Каждый ФП имеет три ОТ: асимптотически устойчивый дикритический узел с координатами $(0;0)$; неустойчивое седло $(N;0)$; неустойчивый узел, координаты которого определяются информационными параметрами группы α, β и γ. Выявлены целевые окрестности каждой ОТ, которые делят правую координатную полуплоскость на шесть зон — три зоны устойчивого и три зоны неустойчивого равновесия.

В качестве примера проведен сопоставительный анализ четырех ФП при различных значениях информационных параметров системы, который подтвердил интуитивно понятную закономерность: системы с бóльшим коэффициентом потенциальной защищенности α и/или меньшим коэффициентом потенциальной подверженности когнитивной перегрузке β и/или более высоким показателем преодоления когнитивной перегрузки γ имеют более обширные зоны устойчивого равновесия, а ее члены преодолевают когнитивные перегрузки за меньшее время.

В дальнейшем предполагается изучить способы представления объема избыточной информации и информационных параметров замкнутой ученической группы через начальные условия; оценить пороговый объем избыточного учебного материала, способный привести к возрастанию когнитивной перегрузки членов группы; зависимость скорости движения системы вблизи особых точек от характеристик группы. С практической точки зрения продолжение данного исследования поможет учителю определить, какой относительный объем дополнительного учебного материала целесообразно привлекать к изучению предметной темы или решению учебной задачи в конкретных условиях обучения.

Литература

1. *Kermack W. O., McKendrick A. G.* A contribution to the mathematical theory of epidemics // P. R. Soc. Lond. A. — Conta., 1927. Vol. 115. Iss. 772. P. 700–721.
2. *Абрамов Н. А., Качалин А. И.* Выбор моделей распространения ВПО при разработке модели глобальной сети. Методы и средства обработки информации // Тр. 3-й Всеросс. научн. конф. / Под ред. Л. Н. Королева. — М.: Фак. ВМиК МГУ им. М. В. Ломоносова, 2009. С. 433–438.
3. *Бабанин Д. В.* Модели оценки структурных решений по защите компьютерных сетей от вирусных атак: Дис. . . . канд. техн. наук. — М.: МИЭМ, 2011. 131 с.
4. *Леоненко В. Н.* Математическая эпидемиология. — СПб: ИТМО, 2018. 39 с.
5. *Зиновеев И. В., Манько Н. И.-В., Спица И. А.* Построение математической модели поведения социальной группы на основе медико-биологической SIR-модели распространения эпидемии // Вісник Запорізького національного університету. Фізико-математичні науки, 2013. № 2. С. 36–41.
6. *Братусь А. С., Новожилов А. С., Платонов А. П.* Динамические системы и модели биологии. — М.: Физматлит, 2010. 400 с.
7. PhaPl: Phase Plane Helper. https://phapl.github.io.
8. *Черепанов А. А.* Веб-приложение PhaPl для автоматического построения и исследования фазовых портретов на плоскости // Вестник Самарского ун-та. Естественнонаучная серия, 2018. Т. 24. № 3. С. 41–52.

Поступила в редакцию 03.07.20

SIR-MODEL AS A TOOL TO STUDY DESTRUCTIVE PROCESSES IN NEW KNOWLEDGE ACQUISITION

O. M. Korchazhkina

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

Abstract: The article describes an approach to the analysis of mastering new knowledge with the use of mathematical modeling of a learning situation. The proposed model is based on W. Kermack and A. McKendrick's SIR (susceptible–infected–recovered) model which was originally used to predict the spread of an epidemic to large closed populations in order to prevent the disastrous consequences of global infection. The modified SIR-model allows researchers to investigate a number of regularities that a cognitive process carried out in a closed small student community has. The model also aims at identifying the ways of behavior of such a dynamic system when working with excessive information can lead to cognitive overload and errors in solving learning tasks. The approach is implemented as a numerical and graphical experiment on the phase plane, which makes it possible to compose a holistic picture of the phenomenon and to analyze the conditions under which the system moves towards the state of stable equilibrium. The latter is equal to overcoming cognitive overwork by the doers.

Keywords: cognitive process; new knowledge; cognitive overwork; closed student community; Kermack–McKendrick model; SIR-model; phase plane/portrait; system stability/instability

DOI: 10.14357/08696527210114

References

1. Kermack, W. O., and A. G. McKendrick. 1927. A contribution to the mathematical theory of epidemics. *P. R. Soc. Lond. A. — Conta.* 115(772):700–721.
2. Abramov, N. A., and A. I. Kachalin. 2009. Vybor modeley rasprostraneniya VPO pri razrabotke modeli global'noy seti. Metody i sredstva obrabotki informatsii [Selection of malware distribution models in developing a global network model. Methods and means of information processing]. *Tr. 3-y Vseross. nauchn. konf. "Metody i sredstva obrabotki informatsii"* [3rd All-Russian Scientific Conference "Methods and Means of Information Processing" Proceedings]. Moscow. 433–438.
3. Babanin, D. V. 2011. Modeli otsenki strukturnykh resheniy po zashchite komp'yuternykh setey ot virusnykh atak [Models for evaluating structural solutions to protect computer networks from virus attacks]. Moscow: MIEM HSE. PhD Diss. 131 p.
4. Leonenko, V. N. 2018. *Matematicheskaya epidemiologiya* [Mathematical epidemiology]. St. Petersburg: ITMO. 39 p.

5. Zinoveev, I. V., N. I.-V. Man'ko, and A. Spicza. 2013. Postroenie matematicheskoy modeli povedeniya sotsial'noy gruppy na osnove mediko-biologicheskoy SIR-modeli rasprostraneniya epidemii [Building a mathematical model of social group behavior based on the medical and biological SIR-model of epidemic spread]. *Vestnik ZNU. Fiziko-matematicheskie nauki* [ZNU Bull. Physics and Mathematics] 2:36–41.
6. Bratus', A. S., A. S. Novozhilov, and A. P. Platonov. 2010. *Dinamicheskie sistemy i modeli biologii* [Dynamic systems and models in biology]. Moscow: Fizmatlit. 400 p.
7. Phase Plane Helper. Available at: https://phapl.github.io/ (accessed February 25, 2021).
8. Cherepanov, A. A. 2018. Veb-prilozhenie PhaPl dlya avtomaticheskogo postroeniya i issledovaniya fazovykh portretov na ploskosti [PhaPl web application for automatic construction and research of phase portraits on the plane]. *Vestnik Samarskogo universiteta. Estestvennonauchnaya seriya* [Vestnik of Samara University. Natural science ser.] 24(3):41–52.

Received July 3, 2020

Contributor

Korchazhkina Olga M. (b. 1953) — Candidate of Science (PhD) in technology, senior scientist, Institute of Cybernetics and Educational Computing of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; olgakomax@gmail.com

МОДЕЛЬ НОРМАЛИЗОВАННОЙ ЭКОНОМИКИ И АКТУАЛЬНЫЕ ТЕХНОЛОГИИ ЦИФРОВИЗАЦИИ

В. Д. Ильин[1]

Аннотация: Представлены ключевые обновления модели нормализованной экономики и актуальные технологии цифровизации. Обновленная модель представлена кратким описанием нормализованного экономического механизма (НЭМ), реализуемого на основе онлайн-сервисов, функционирующих в среде цифровых двойников. Нормализованный экономический механизм включает ситуационно управляемые комплексы ресурсного обеспечения, производства реальных товаров и др. Банковская система НЭМ включает персональные электронные банки (ПЭБы) физических лиц, корпоративные электронные банки (КЭБы), банки-провайдеры и банк-регулятор, под управлением которого функционируют все другие банки. Нормализованные деньги (НД) предназначены для представления стоимости товаров и имущественных статусов участников экономической деятельности (эд-участников), для оплаты товаров, инвестирования и накопления богатства. Нормализованные деньги служат универсальным электронным средством количественного документирования имущественных отношений, удостоверяемых государством посредством онлайн-сервисов банка-регулятора. Управление экономическим поведением осуществляется посредством систем обязательных и ориентирующих требований к результатам решения задач в составе комплексов НЭМ.

Ключевые слова: модель нормализованной экономики; нормализованный экономический механизм; персональные электронные банки; корпоративные электронные банки; нормализованные деньги; онлайн-сервисы

DOI: 10.14357/08696527210115

1 Введение

В 2009 г. в статье, подготовленной на основе доклада, представленного Президиуму РАН, были рассмотрены модели, отражающие «*сложившиеся внутренние механизмы развития экономики как целого*» [1, с. 492]. Тогда же в первой статье о *модели нормализованной экономики* [2] была предложена концепция экономической системы, где «*установленные законом правила и государственные управляющие воздействия должны направлять экономическую деятельность населения на защиту и развитие потенциала страны*» [2, с. 118]. Была обоснована необходимость своевременной адаптации устройства экономического механизма к изменяющимся социально-экономическим отношениям и передовым техноло-

[1]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, vdilyin@yandex.ru

гиям человеко-машинной среды, в которой он функционирует (а «*экономисты продолжали держаться концепции денег как товара-посредника, не реагируя на принципиальные технологические перемены в возможностях документирования экономических сделок*» [2, с. 117]). В разд. 3.4 [*Внешний обмен товарами (экспорт/импорт)*] [2, с. 131] было предложено принять, что «*применимы деньги только систем, участвующих в сделке*».

В наши дни, подтверждая положения обновляющейся концепции *модели нормализованной экономики* [2–4], быстрыми темпами развиваются онлайн-сервисы в торговле, банковской и других областях деятельности. Интенсивно растет применение роботов в промышленности[1], добывающих компаниях[2], сельском хозяйстве[3] и других областях деятельности. Увеличивается число стран, с которыми Россия заключила соглашения о применении национальных валют в торговых сделках. Приближаются дни, когда в *нормализованной банковской системе* привычные для активных *участников экономической деятельности* смартфоны и планшеты станут выполнять роли *персональных электронных банков* [2, с. 126]. На государственном уровне России признание определяющей связи эффективности хозяйственной и других видов современной деятельности с применением информационных технологий выражено в программе «*Цифровая экономика Российской Федерации*» [5]. Формально участвуя в реализации этой программы, банки по-прежнему стремятся сохранить возможность слабо ограниченного распоряжения денежными средствами клиентов, торговли кредитами, валютной спекуляции и другой деятельности, уменьшающей товарную емкость денег. С началом реализации идеи совершенствования различных видов деятельности (не только относящихся к экономике) путем применения компьютеров и компьютерных устройств (названной *цифровизацией экономики* [6, 7]) возникли и с тех пор не утихают дискуссии о значении *цифровизации*, направлениях ее развития и последствиях [8, 9] . На современном этапе в технологическом ядре цифровизации интенсивно продвигаются *M2M-технологии* (*Machine-to-Machine*) [10, 11], технологии *облачных вычислений* (*Cloud Computing*) и электронных сервисов [12–15], *интернета вещей* (*Internet of Things*, IoT) [16], *цифровых двойников* (*Digital Twins*[4]). Успешное развитие этих и других передовых информационных технологий создает предпосылки для ускорения перехода к *нормализованному экономическому механизму* [2–4].

Запись формул и выделение фрагментов текста. Для выделения определений, замечаний и примеров используются средства языка TSM-комплекса (TSM:

[1]Применение промышленных роботов: популярные направления роботизации // MENTAMORE. https://mentamore.com/robototexnika/primenenie-promyshlennyx-robotov.html.

[2]*Бойко А.* Добывающие компании и роботизация // RoboTrends. http://robotrends.ru/robopedia/1711-dobyvayushie-kompanii-i-robotizaciya.

[3]*Бойко А.* Сельское хозяйство и роботы // RoboTrends. http://robotrends.ru/robopedia/selskoe-hozyaystvo-i-roboty.

[4]The New Age of Manufacturing: Digital Twin Technology & IIoT. https://medium.com/@lior.kitain/the-new-age-of-manufacturing-digital-twin-technology-iiot-494acee5572a.

textual symbolic modeling), разработанного для формализованного описания текстовых моделей[1].

В статье применены следующие средства выделения фрагментов текста:

□ ⟨фрагмент описания⟩ □ ≈ утверждение (определение, аксиома и др.) (здесь и далее символ ≈ заменяет слово «означает»);

◊ ⟨фрагмент описания⟩ ◊ ≈ замечание;

◯ ⟨фрагмент описания⟩ ◯ ≈ пример.

Курсивом выделены первые вхождения названий понятий и фрагменты описания, к которым автор хочет привлечь внимание.

Обсуждаемые результаты. В обзоре представлены результаты, полученные при выполнении научно-исследовательской работы «*Моделирование социальных, экономических и экологических процессов*» (№ 0063-2016-0005), выполняемой в соответствии с государственным заданием ФАНО России для Федерального исследовательского центра «Информатика и управление» РАН.

2 Нормализованный экономический механизм: общая характеристика

□ *Нормализованный экономический механизм — рыночный экономический механизм*, комплексы которого (см. таблицу) работают на основе онлайн-сервисов, функционирующих в среде цифровых двойников. Представляет собой *человеко-машинную систему, связанную отношениями координации и подчинения с государственным механизмом*. Устройство и правила функционирования НЭМ, установленные государственными законами, стимулируют вести хозяйственную деятельность, ориентированную на защиту и развитие *потенциала страны*. Делается это посредством налогов, пошлин и других экономических инструментов. □

Основные комплексы НЭМ

Управление экономической деятельностью	Ресурсное обеспечение
	Производство реальных товаров
	Торговля
	Резервирование жизненно-необходимых товаров
	Инвестирование
	Государственный бюджет, резервы, налоги, пошлины
	Региональные бюджеты и налоги
	Профессиональное образование и развитие населения
	Развитие систем организации и обеспечения жизнедеятельности
	Восстановление и развитие среды обитания
	Фонды социального обеспечения
	Документирование товарно-денежного обращения и имущественных статусов (осуществляется нормализованной банковской системой)

[1] *Ильин В. Д.* Символьное моделирование // Большая российская энциклопедия (электронная версия). http://dev.bigenc.ru/technology and technique/text/4010980.

НЭМ-комплекс *управления экономической деятельностью* определяет цели и реализующие их задачи развития экономической системы страны; координирует решение этих задач [3, 4]. НЭМ-комплексы *ресурсного обеспечения* и *производства товаров* служат ядром НЭМ. В дополнение к государственному резервированию целесообразно развивать сеть хранилищ негосударственного резерва *жизненно-необходимых товаров* (*жн-товаров*). Ядром комплекса *документирования товарно-денежного обращения и имущественных статусов эд-участников* является *система имущественных статусов* (см. разд. 3). *Специфицированные имущественные объекты* (*си-объекты*) нормализованной экономической системы — это принадлежащие физическим и юридическим лицам средства производства, реализации и резервирования товаров, средства документирования, а также предметы потребления, зарегистрированные в экономической системе.

□ *Товар* — продаваемый си-объект. □ ◊ Каждому товару поставлена в соответствие электронная унифицированная спецификация, в которой указаны наименование товара, назначение и другие сведения. ◊

3 Система имущественных статусов и нормализованные деньги

□ *Система имущественных статусов* (*ис-система*) — это реализуемая на основе онлайн-сервисов система документального представления денежной и неденежной составляющих имущественных статусов эд-участников [2–4]. Денежная составляющая выражена значениями сумм НД в разделах *уникальных поливалютных счетов эд-участников* (*ис-счетов*). Неденежная — документами, подтверждающими право собственности на недвижимость, транспорт и другое имущество, которое при необходимости может рассматриваться как залоговое. □

□ *Ис-счет* — уникальный поливалютный банковский счет, размещенный в персональном электронном банке физического лица или корпоративном электронном банке юридического лица. □

□ *Нормализованные деньги* — универсальное электронное средство количественного документирования имущественных отношений, удостоверяемых государством. Предназначено для представления стоимости товаров и имущественных статусов эд-участников, для оплаты товаров, инвестирования и накопления богатства. *Универсальность НД* выражается в применимости во всех внутренних и внешних экономических сделках, разрешенных законом государства, под юрисдикцией которого функционирует экономическая система. □

4 Нормализованная банковская система

Банковская система НЭМ (нормализованная банковская система) включает ПЭБы) *физических лиц, КЭБы*, *банки-провайдеры* и *банк-регулятор*, под управлением которого функционируют все другие банки [2–4].

□ *Банк-регулятор* — государственное учреждение, управляющее функционированием банковской системы НЭМ. Задачи банка-регулятора, реализуемые посредством онлайн-сервисов:

- выдача и отзыв лицензий на право банковской деятельности (владельцам ПЭБов, КЭБов и банков-провайдеров);
- контроль использования валютных частей ис-счетов;
- удостоверение операций над ис-счетами при совершении контролируемых сделок;
- обслуживание запросов на досрочный возврат долгов и выполнение долговой денежной эмиссии (при реализации технологии долговой торговли) [4, 17];
- контроль выполнения эд-участниками правил банковской деятельности; анализ финансовой составляющей экономической деятельности и предоставление результатов в порядке, установленном законом;
- разработка, модификация и утверждение прошедших тестирование унифицированных форм банковских документов (включая ис-счета);
- контроль эффективности размещения средств *государственного денежного резерва, фондов государственной социальной защиты*, других государственных фондов и фондов с государственным участием. □

Банк-регулятор располагает сетью серверов, размещенных на территории страны, под юрисдикцией которой находится экономическая система.

□ *Банки-провайдеры* — коммерческие предприятия, учрежденные юридическими лицами (или объединениями юридических лиц, объединениями физических лиц, объединениями юридических и физических лиц), занимающимися производством товаров и/или их реализацией. Банк-провайдер располагает объединенными в сеть серверами, предназначенными для обслуживания запросов от ПЭБов и КЭБов клиентов и взаимодействия с серверами банка-регулятора.

Онлайн-сервисы банка-провайдера предназначены:

- для обслуживания запросов владельцев ис-счетов, направляемых посредством ПЭБов и КЭБов при совершении сделок (включая запросы на удостоверение состояния ис-счетов партнеров по сделкам, направляемые с их разрешения);
- хранения и обновления зашифрованных копий ис-счетов клиентов банка-провайдера;
- анализа инвестиционных запросов клиентов (потенциальных инвесторов и получателей инвестиций), при этом банки-провайдеры могут выполнять заказы получателей инвестиций на консолидацию заявок инвесторов, чтобы комплексировать заказанную сумму инвестиции;
- регистрации заключенных договоров (с контролем допустимости сделок) и ведения баз данных таких договоров;
- юридического сопровождения реализуемых сделок. □

5 Технологии долговой электронной торговли и денежной эмиссии

Одна из актуальных технологий нормализации действующего в России экономического механизма — технология электронной *долговой торговли*, при которой отсроченная часть оплаты товара оформляется как *долг покупателя продавцу* (не банку!), *имеющему КЭБ или ПЭБ* [4, 17]. Применительно к долгам, образовавшимся в результате продажи приоритетных жн-товаров, могут действовать правила досрочного возврата долгов продавцам из средств банка-регулятора. ◊ *Досрочным возвратом долгов продавцам приоритетных товаров* банк-регулятор реализует текущую государственную программу поддержки производства и продажи таких товаров. ◊

□ *Долговая денежная эмиссия* выполняется банком-регулятором только тогда, когда сумма возвращенных покупателями долгов меньше очередной долговой суммы, запрошенной для возврата продавцу. Эмитируемая сумма равна разности запрошенной долговой суммы и суммы на счету долгового отдела. □ Долговая денежная эмиссия служит средством *регулирования общей суммы денег в экономической системе*. Таким способом осуществляется регулирование товарной емкости денег. ◊ Правила, реализуемые в технологии, исключают возможность эмиссии денежных сумм, не обеспеченных товарами. ◊

6 Технология поливалютного рынка

В концепции поливалютного рынка НЭМ технологии внешнеэкономических сделок функционируют так, что для любой пары стран на каждом отрезке времени определено некоторое множество товаров, множество применимых для расчетов валют, таможенные правила и правила внешнеторговых сборов [4]. Для каждого типа товара эд-участники имеют возможность совершать сделки в любой валюте из списка, являющегося пересечением множеств валют, разрешенных банками-регуляторами государств, к которым относятся участники сделки.

◊ На поливалютном рынке НЭМ долговая торговля служит средством, стимулирующим продажи приоритетных товаров с оплатой в национальной валюте: возможность в относительно короткие сроки получить долговую часть стоимости товаров существует у продавцов в тех случаях, когда товары оплачены национальной валютой. ◊

◊ Целесообразны отношения *координации* между странами (при выработке и реализации схем контроля исполнения указанных правил) и нецелесообразны любые глобальные регуляторы, ограничивающие свободу экономического выбора эд-участников. ◊

7 Технология назначенных платежей

□ *Назначенный платеж* — технологически гарантированная оплата заказанных товаров, правила реализации которой жестко связаны с правилами

исполнения заказа [18]. Сумма назначенного платежа переводится на счет получателя сразу после того, как заказчик подтвердил исполнение заказа. В случае неисполнения заказа назначенный платеж отменяется. □

□ *Технология назначенных платежей* — совокупность методов, средств и правил реализации назначенных платежей в среде цифровых двойников. □

8 Технология ситуационного онлайн-бюджетирования

Онлайн-бюджетирование (национальных и региональных проектов) рассматривается как *задача интервального планирования расходов* с учетом *ситуационно зависимых обязательных и ориентирующих требований* к искомому решению [19, 20].

Для числового отрезка $[a^{\min}, a^{\max}]$ ($a^{\min} \geq 0$, $a^{\max} > 0$), задающего распределяемую величину (◯ сумму денег ◯), отрезков $[b_i^{\min}, b_i^{\max}]$ ($b_i^{\min} \geq 0$, $b_i^{\max} > 0$, $i = 1, \ldots, n$), задающих запросы по расходным статьям, и весовых коэффициентов (приоритетов) расходных статей $p_i > 0$ ($i = 1, \ldots, n$) требуется найти план расходов по статьям

$$\left[x_i^{\min}, x_i^{\max}\right] : \left\{0 \leq x_i^{\min} \leq b_i^{\min},\ x_i^{\max} \leq b_i^{\max},\ \sum x_i^{\min} \leq a^{\min},\right.$$
$$\left.\sum x_i^{\max} \leq a^{\max},\ i = 1, \ldots, n\right\}. \quad (1)$$

Для совокупного вектора искомого плана $\boldsymbol{x} = (x_1^{\min}, \ldots, x_n^{\min}, x_1^{\max}, \ldots, x_n^{\max})$ может быть также задан набор требований

$$\boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{d} \leftarrow \boldsymbol{q}\,, \quad (2)$$

где $\boldsymbol{C}$ — матрица вещественных коэффициентов размера $k \times 2n$ ($k \geq 1$); $\boldsymbol{d}$ — вектор-столбец вещественных констант размера k; $\boldsymbol{q}$ — вектор-столбец весовых коэффициентов (приоритетов) требований ($0 < q_i \leq +\infty$, $i = 1, \ldots, k$).

Обязательные требования имеют приоритет $+\infty$. Приоритеты *ориентирующих требований* задаются положительными вещественными числами. ◊ Приоритеты задают эксперты-экономисты, учитывая относительную важность выполнения рассматриваемых требований. ◊

Требования (1) являются обязательными. Требования (2) могут быть как обязательными, так и ориентирующими.

Задача решается либо *методом приоритетного интервального распределения*, реализованным в действующем *интернет-сервисе планирования расходов*[1], либо *методом целевого перемещения решения* в режиме вычислительного эксперимента.

[1] *Ильин А. В.* Интернет-сервис планирования расходов. https://www.res-plan.com/services-ru.

9 Заключение

К началу 2021 г. часть положений концепции *модели нормализованной экономики* реализована (см. введение). К актуальным технологиям современного этапа цифровизации, которые целесообразно реализовать, относятся *технологии нормализованного товарно-денежного обращения*, предназначенные для формирования и реализации договорных отношений в цифровой среде, платежного и товарного кредитования [21], онлайн-банкинга на основе банков-провайдеров, КЭБов и ПЭбов [3, 4].

Литература

1. *Петров А. А., Поспелов И. Г.* Математические модели экономики России // Вестник Российской академии наук, 2009. Т. 79. № 6. С. 492–506.
2. *Ильин В. Д.* Модель нормализованной экономики (НЭк-модель): основы концепции // Управление большими системами, 2009. Вып. 25. С. 116–138.
3. *Ilyin A V., Ilyin V. D.* Towards a normalized economic mechanism based on E-services // Agris on-line Papers in Economics and Informatics, 2014. Vol. 6. Iss. 3. P. 39–49.
4. *Ilyin A. V., Ilyin V. D.* The normalized economic mechanism in the digital environment // Int. J. Open Information Technologies, 2019. Vol. 7. Iss. 12. P. 77–83.
5. Цифровая экономика Российской Федерации: Программа, утвержденная Распоряжением Правительства Российской Федерации от 28 июля 2017 г. № 1632-р. http://d-russia.ru/wpcontent/uploads/2017/07/programma-tsifrov-econ.pdf.
6. *Tapscott D.* The digital economy: Promise and peril in the age of networked intelligence. — New York, NY, USA: McGraw-Hill, 1996. 342 p.
7. *Christensen C. M.* The innovator's dilemma: When new technologies cause great firms to fail. — Boston, MA, USA: Harvard Business School Press, 1997. 288 p.
8. The New Digital Economy: How it will transform business. — Oxford Economics, 2015. 34 p. http://www.pwc.com/mt/en/publications/assets/thenew-digitaleconomy.pdf.
9. G20 digital economy development and cooperation initiative // G20 Summit, 2016. http://en.kremlin.ru/supplement/5111.
10. *Kim R. Y.* Efficient wireless communications schemes for machine to machine communications // Comm. Com. Inf. Sc., 2011. Vol. 181. No. 3. P. 313–323.
11. *Lien S. Y., Liau T. H., Kao C. Y., et al.* Cooperative access class barring for machine-to-machine communications // IEEE T. Wirel. Commun., 2012. Vol. 11. No. 1. P. 27–32.
12. *Armbrust M., Fox A., Griffith R., et al.* A view of cloud computing // Commun. ACM, 2010. Vol. 53. No. 4. P. 50–58.
13. *Wei Y., Blake M. B.* Service-oriented computing and cloud computing: Challenges and opportunities // IEEE Internet Comput., 2010. Vol. 14. No. 6. P. 72–75.
14. *Rogers O., Cliff D.* A financial brokerage model for cloud computing // J. Cloud Comput., 2012. Vol. 1. No. 1. P. 1–12.
15. *Jede A., Teuteberg F.* Understanding socio-technical impacts arising from software as-a-service usage in companies // Bus. Inf. Syst. Eng., 2016. Vol. 58. No. 3. P. 161–176.

16. *Perera C., Liu C. H., Jayawardena S.* The emerging Internet of Things marketplace from an industrial perspective: A survey // IEEE T. Emerging Topics Computing, 2015. Vol. 3. No. 4. P. 585–598.
17. *Ilyin A. V., Ilyin V. D.* E-trade with direct lending and normalized money // Agris on-line Papers in Economics and Informatics, 2015. Vol. 7. Iss. 4. P. 57–64.
18. *Ильин В. Д.* Технология назначенных платежей в среде цифровых двойников // Системы и средства информатики, 2018. Т. 28. № 3. С. 227–235.
19. *Ilyin A. V., Ilyin V. D.* Variational online budgeting taking into account the priorities of expense items // Agris on-line Papers in Economics and Informatics, 2016. Vol. 8. Iss. 3. P. 51–56.
20. *Ilyin A. V., Ilyin V. D.* Solving situationally definable linear problems of resource planning: A review of updated technology // Информационные технологии и вычислительные системы, 2019. № 3. P. 99–106.
21. *Ilyin A. V., Ilyin V. D.* The technologies of commodity-money circulation on the basis of personal and corporative e-banks // Int. J. Open Information Technologies, 2020. Vol. 8. No. 5. P. 81–84.

Поступила в редакцию 05.02.21

THE MODEL OF NORMALIZED ECONOMICS AND RELEVANT TECHNOLOGIES OF DIGITALIZATION

V. D. Ilyin

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

Abstract: The review presents key updates of the model of normalized economics and relevant technologies of digitalization. The updated model is presented by description of the normalized economic mechanism (NEM) which should be implemented on the basis of online services operating in the environment of digital twins. Normalized economic mechanism includes the situationally managed systems of resource support, production of real goods, etc. The NEM banking system includes personal electronic banks of individuals, corporate electronic banks, provider banks, and the regulation bank under which all other banks operate. The normalized money is intended to represent the value of goods and property statuses of the economic activity participants, to pay for goods, invest, and accumulate wealth. Normalized money serves as a universal electronic means of quantitative documentation of property relations to be certified by the state through the online services of the regulation bank. An economic management is to be based on the systems of mandatory and orienting requirements to the results of solving problems inside the NEM complexes.

Keywords: model of normalized economics; normalized economic mechanism; personal electronic banks; corporate electronic banks; designated payments technology; online cost planning service

DOI: 10.14357/08696527210115

References

1. Petrov, A. A., and I. G. Pospelov. 2009. Mathematical models of the Russian economy. *Her. Russ. Acad. Sci.* 79(3):205–216.
2. Ilyin, V. D. 2009. Model' normalizovannoy ekonomiki (NEk-model'): osnovy kontseptsii [The model of normalized economics (NEc-model): Basics of framework]. *Upravlenie bol'shimi sistemami* [Large-Scale Systems Control] 25:116–138.
3. Ilyin, A. V., and V. D. Ilyin. 2014. Towards a normalized economic mechanism based on E-services. *Agris on-line Papers in Economics and Informatics* 6(3):39–49.
4. Ilyin, A. V., and V. D. Ilyin. 2019. The normalized economic mechanism in the digital environment. *Int. J. Open Information Technologies* 7(12):77–83.
5. Tsifrovaya ekonomika Rossiyskoy Federatsii: Programma, utverzhdennaya Rasporyazheniem Pravitel'stva Rossiyskoy Federatsii [Digital economy of the Russian Federation: Program Approved by Order No. 1632-r dated July 28, 2017 of the Government of the Russian Federation]. Available at: http://d-russia.ru/wpcontent/uploads/2017/07/programma-tsifrov-econ.pdf (accessed February 5, 2021).
6. Tapscott, D. 1996. The digital economy: Promise and peril in the age of networked intelligence. New York, NY: McGraw-Hill. 342 p.
7. Christensen, C. M. 1997. *The innovator's dilemma: When new technologies cause great firms to fail*. Boston, MA: Harvard Business School Press. 288 p.
8. The new digital economy: How it will transform business. 2015. Oxford Economics. 34 p. Available at: http://www.pwc.com/mt/en/publications/assets/thenew-digitaleconomy.pdf (accessed February 5, 2021).
9. G20 digital economy development and cooperation initiative. 2016. G20 Summit. Available at: http://en.kremlin.ru/supplement/5111 (accessed February 5, 2021)
10. Kim, R. Y. 2011. Efficient wireless communications schemes for machine to machine communications. *Comm. Com. Inf. Sc.* 181(3):313–323.
11. Lien, S. Y., T. H. Liau, C. Y. Kao, *et al.* 2012. Cooperative access class barring for machine-to-machine communications. *IEEE T. Wirel. Commun.* 11(1):27–32.
12. Armbrust, M., A. Fox, R. Griffith, *et al.* 2010. A view of cloud computing. *Commun. ACM* 53(4):50–58.
13. Wei, Y., and M. B. Blake. 2010. Service-oriented computing and cloud computing: Challenges and opportunities. *IEEE Internet Comput.* 14:72–75.
14. Rogers, O., and D. Cliff. 2012. A financial brokerage model for cloud computing. *J. Cloud Computing* 1(1):1–12.
15. Jede, A., and F. Teuteberg. 2016. Understanding socio-technical impacts arising from software-as-a-service usage in companies. *Bus. Inf. Syst. Eng.* 58(3):161–176.
16. Perera, C., C. H. Liu, and S. Jayawardena. 2015. The emerging Internet of Things marketplace from an industrial perspective: A survey. *IEEE T. Emerging Topics Computing* 3(4):585–598.
17. Ilyin, A. V., and V. D. Ilyin. 2015. E-trade with direct lending and normalized money. *Agris on-line Papers in Economics and Informatics* 7(4):57–64.
18. Ilyin, V. D. 2018. Tekhnologiya naznachennykh platezhey v srede tsifrovykh dvoynikov [Designated payments technology in digital twins environment]. *Systemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):227–235.

19. Ilyin, A. V., and V. D. Ilyin. 2016. Variational online budgeting taking into account the priorities of expense items. *Agris on-line Papers in Economics and Informatics* 8(3):51–56.
20. Ilyin, A. V., and V. D. Ilyin. 2019. Solving situationally definable linear problems of resource planning: A review of updated technology. *Informatsionnye tekhnologii i vychislitel'nye sistemy* [J. Information Technologies Computing Systems] 3:99–106.
21. Ilyin, A. V., and V. D. Ilyin. 2020. The technologies of commodity-money circulation on the basis of personal and corporative e-banks. *Int. J. Open Information Technologies* 8(5):81–84.

Received February 5, 2021

Contributor

Ilyin Vladimir D. (b. 1937) — Doctor of Science in technology, professor, leading scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow119333, Russian Federation; vdilyin@yandex.ru

ОБ АВТОРАХ

Адамович Игорь Михайлович (р. 1934) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Борисов Андрей Владимирович (р. 1965) — доктор физико-математических наук, главный научный сотрудник Института проблем информатики 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

Волков Олег Игоревич (р. 1964) — ведущий программист Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Волович Константин Иосифович (р. 1970) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

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

Егоров Владимир Борисович (р. 1948) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Жуков Денис Владимирович (р. 1979) — главный специалист Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Забежайло Михаил Иванович (р. 1956) — доктор физико-математических наук, доцент, главный научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Зацаринный Александр Алексеевич (р. 1951) — доктор технических наук, профессор, заместитель директора Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН); главный научный сотрудник Института проблем информатики ФИЦ ИУ РАН

Ильин Владимир Дмитриевич (р. 1937) — доктор технических наук, профессор, ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Исаченко Роман Владимирович (р. 1994) — аспирант Московского физико-технического института

Ковалев Дмитрий Юрьевич (р. 1988) — 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

Корчажкина Ольга Максимовна (р. 1953) — кандидат технических наук, старший научный сотрудник Института кибернетики и образовательной информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Мальковский Сергей Иванович (р. 1983) — научный сотрудник Вычислительного центра Дальневосточного отделения Российской академии наук

Никишин Дмитрий Александрович (р. 1976) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Растрелин Анатолий Матвеевич (р. 1937) — доктор технических наук, профессор, директор по НИОКР Научно-исследовательского института систем автоматизации

Сатин Яков Александрович (р. 1978) — кандидат физико-математических наук, доцент Вологодского государственного университета

Синицын Игорь Николаевич (р. 1940) — доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Сорокин Алексей Анатольевич (р. 1980) — кандидат технических наук, ведущий научный сотрудник Вычислительного центра Дальневосточного отделения Российской академии наук

Стрижов Вадим Викторович (р. 1967) — доктор физико-математических наук, ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук; профессор Московского физико-технического института

Сучков Александр Павлович (р. 1954) — доктор технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Тимонина Елена Евгеньевна (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Тириков Егор Михайлович (р. 1996) — аспирант Федерального исследовательского центра «Информатика и управление» Российской академии наук

Цой Георгий Ильич (р. 1992) — кандидат физико-математических наук, научный сотрудник Вычислительного центра Дальневосточного отделения Российской академии наук

Черных Владимир Юрьевич (р. 1995) — старший программист Вычислительного центра Дальневосточного отделения Российской академии наук

Яушев Фарух Юрьевич (р. 1999) — студент Московского физико-технического института

Правила подготовки рукописей статей для публикации в журнале «Системы и средства информатики»

Журнал «Системы и средства информатики» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информационных технологий.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- информационно-телекоммуникационные системы и средства их построения;
- архитектура и программное обеспечение вычислительных машин, комплексов и сетей;
- методы и средства защиты информации.

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 и 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/publications/collected/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 города и страны и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах:
     http://www.ipiran.ru/journal/collected/2019_29_03_rus/authors.asp и
     http://www.ipiran.ru/journal/collected/2019_29_03_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 (ссылкм 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.

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".

Примеры ссылок на различные виды публикаций в списке "References":

Описание статьи из журнала:

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 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.

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, NY: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004. Geology and production (NGT GiD). Certificate on official registration of the computer program No. 2004611198. (In Russian, unpubl.)

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]. PhD Thesis. Moscow: IPI RAN. 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 «Системы и средства информатики» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Системы и средства информатики»:

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: ssi@frccsc.ru (Стригина Светлана Николаевна)
http://www.ipiran.ru/journal/collected

Requirements for manuscripts submitted to Journal "Systems and Means of Informatics"

Journal "Systems and Means of Informatics" publishes theoretical, review, and discussion articles on the research and development in the field of information technology.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

Topics covered include the following areas:

- information and communication systems and tools of their design;
- architecture and software of computational complexes and networks; and
- methods and tools of information protection.

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 "Systems and Means of Informatics" 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).

   The Editorial Board has the right to request from the authors an official expert conclusion that the submitted article has no classified data prohibited for publication.
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 authors receive their article for correction after reviewing, it does not mean that the article is approved to be published. 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/publication/collected/template.doc

8. 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/collected/2019_29_03_rus/authors.asp and
     http://www.ipiran.ru/journal/collected/2019_29_03_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 in 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).

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. "System and Means of Informatics" 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: ssi@frccsc.ru (to Svetlana Strigina)
http://www.ipiran.ru/english/journal_systems.asp

SYSTEMS AND MEANS OF INFORMATICS (SISTEMY I SREDSTVA INFORMATIKI)

SCIENTIFIC JOURNAL

Volume 31 No.1 Year 2021

Editor-in-Chief and Chair of Editorial Council
Academician I. A. Sokolov

I N T H I S I S S U E: