

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Том 24
2018
№ 2

ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

Издается с ноября 1995 г.

УЧРЕДИТЕЛЬ
Издательство "Новые технологии"

СОДЕРЖАНИЕ

МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ

- Ковалев С. П., Шаймарданов М. Ю. Метод учета технологических ограничений при оптимизации топологии изделий машиностроения 75
- Чертовской В. Д. Методология математического описания и моделирования адаптивной автоматизированной системы управления производством ... 81

WEB-ТЕХНОЛОГИИ

- Лобов И. В., Готман В. Г. Использование контейнера Ogg для организации потоковой трансляции в реальном времени над протоколом HTTP методом опережающей загрузки 87

ПРОГРАММНАЯ ИНЖЕНЕРИЯ

- Мыльников Л. А., Селедкова А. Б. Способ выбора метода прогнозирования и горизонта планирования параметров с использованием оценки риска ... 97

БЕЗОПАСНОСТЬ ИНФОРМАЦИИ

- Осадчая Т. С., Щеглов А. Ю. Комплексное решение задачи защиты от атак с правами привилегированного пользователя 104

НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ

- Лютикова Л. А., Шматова Е. В. Логический подход к коррекции результатов работы СП-нейронных сетей 110
- Куликов С. В., Захаров О. С., Андреев Д. Ю. Исследование возможности совместного применения нейросетевого преобразователя биометрия—код и глубокой сверточной нейронной сети в распознавании лиц 116
- Горбатков С. А., Полупанов Д. В. Оптимальный отбор и агрегирование экзогенных переменных в нейросетевых моделях банкротств на основе функций Харрингтона 121

ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

- Авдошин С. М., Песоцкая Е. Ю. Интернет вещей: Транспорт 131
- Ле Ба Чунг, Холопов Ю. А. Асимметричный межмодульный интерфейс 138

Главный редактор:

СТЕМПКОВСКИЙ А. Л.,
акад. РАН, д. т. н., проф.

Зам. главного редактора:

ИВАННИКОВ А. Д., д. т. н., проф.
ФИЛИМОНОВ Н. Б., д. т. н., с.н.с.

Редакционный совет:

БЫЧКОВ И. В., акад. РАН, д. т. н.
ЖУРАВЛЕВ Ю. И.,
акад. РАН, д. ф.-м. н., проф.
КУЛЕШОВ А. П.,
акад. РАН, д. т. н., проф.
ПОПКОВ Ю. С.,
акад. РАН, д. т. н., проф.
РУСАКОВ С. Г.,
чл.-корр. РАН, д. т. н., проф.
РЯБОВ Г. Г.,
чл.-корр. РАН, д. т. н., проф.
СОЙФЕР В. А.,
акад. РАН, д. т. н., проф.
СОКОЛОВ И. А.,
акад. РАН, д. т. н., проф.
СУЕТИН Н. В., д. ф.-м. н., проф.
ЧАПЛЫГИН Ю. А.,
акад. РАН, д. т. н., проф.
ШАХНОВ В. А.,
чл.-корр. РАН, д. т. н., проф.
ШОКИН Ю. И.,
акад. РАН, д. т. н., проф.
ЮСУПОВ Р. М.,
чл.-корр. РАН, д. т. н., проф.

Редакционная коллегия:

АВДОШИН С. М., к. т. н., доц.
АНТОНОВ Б. И.
БАРСКИЙ А. Б., д. т. н., проф.
ВАСЕНИН В. А., д. ф.-м. н., проф.
ВИШНЕКОВ А. В., д. т. н., проф.
ДИМИТРИЕНКО Ю. И., д. ф.-м. н., проф.
ДОМРАЧЕВ В. Г., д. т. н., проф.
ЗАБОРОВСКИЙ В. С., д. т. н., проф.
ЗАГИДУЛЛИН Р. Ш., к. т. н., доц.
ЗАРУБИН В. С., д. т. н., проф.
КАРПЕНКО А. П., д. ф.-м. н., проф.
КОЛИН К. К., д. т. н., проф.
КУЛАГИН В. П., д. т. н., проф.
КУРЕЙЧИК В. В., д. т. н., проф.
ЛВОВИЧ Я. Е., д. т. н., проф.
МАРТЫНОВ В. В., д. т. н., проф.
МИХАЙЛОВ Б. М., д. т. н., проф.
НЕЧАЕВ В. В., к. т. н., проф.
ПОЛЕЩУК О. М., д. т. н., проф.
САКСОНОВ Е. А., д. т. н., проф.
СОКОЛОВ Б. В., д. т. н., проф.
ТИМОНИНА Е. Е., д. т. н., проф.
УСКОВ В. Л., к. т. н. (США)
ФОМИЧЕВ В. А., д. т. н., проф.
ШИЛОВ В. В., к. т. н., доц.

Редакция:

БЕЗМЕНОВА М. Ю.
ГРИГОРИН-РЯБОВА Е. В.
ЛЫСЕНКО А. В.
ЧУГУНОВА А. В.

Информация о журнале доступна по сети Internet по адресу <http://novtex.ru/IT>.

Журнал включен в систему Российского индекса научного цитирования и базу данных RSCI на платформе Web of Science.

Журнал входит в Перечень научных журналов, в которых по рекомендации ВАК РФ должны быть опубликованы научные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

INFORMATION TECHNOLOGIES

INFORMACIONNYYE TEHNOLOGII

Vol. 24
2018
No. 2

THEORETICAL AND APPLIED SCIENTIFIC AND TECHNICAL JOURNAL

Published since November 1995

ISSN 1684-6400

CONTENTS

MODELING AND OPTIMIZATION

- Kovalyov S. P., Shaymardanov M. Yu.** Methods of Applying Manufacturing Constraints in Topology Optimization 75
- Chertovskoy V. D.** Methodology of Mathematical Description and Modeling of Adaptive Automatized Manufacturing Control Systems 81

WEB TECHNOLOGIES

- Lobov I. V., Gotman V. G.** The Utilization of Ogg Multimedia Container Format for Live Streaming over HTTP Using Progressive Download Method 87

SOFTWARE ENGINEERING

- Mylnikov L. A., Seledkova A. B.** Selection of the Parameters Predicting Method in Production and Economic Systems Based on Risks Metrics 97

CRYPTOSAFETY INFORMATION

- Osadchaya T. S., Shcheglov A. Yu.** Comprehensive Solution of Protection against Attack with Rights of Privileged Users 104

NEUROTECHNOLOGIES

- Lyutikova L. A., Shmatova E. V.** A Logical Approach to the Correction of the Results of the Operation of $\Sigma\Pi$ -Neural Networks 110
- Kulikov S. V., Zakharov O. S., Andreev D. Yu.** Exploring the Possibility of Using Deep Convolutional Neural Network Paired with Neural "Biometric Image to Code" Converter in Face Recognition 116
- Gorbatkov S. A., Polupanov D. V.** Optimal Selection and Aggregation of Exogenous Variables in Neural Network Models of Bankruptcies based on Harrington Functions 121

APPLIED INFORMATION TECHNOLOGIES

- Avdoshin S. M., Pesotskaya E. Yu.** Internet of Things: Transportation 131
- Le Ba Chung, Holopov Ju. A.** Asymmetrical Inter-Module Interface 138

Editor-in-Chief:

Stempkovsky A. L., Member of RAS,
Dr. Sci. (Tech.), Prof.

Deputy Editor-in-Chief:

Ivannikov A. D., Dr. Sci. (Tech.), Prof.
Filimonov N. B., Dr. Sci. (Tech.), Prof.

Chairman:

Bychkov I. V., Member of RAS,
Dr. Sci. (Tech.), Prof.
Zhuravljov Yu. I., Member of RAS,
Dr. Sci. (Phys.-Math.), Prof.
Kuleshov A. P., Member of RAS,
Dr. Sci. (Tech.), Prof.
Popkov Yu. S., Member of RAS,
Dr. Sci. (Tech.), Prof.
Rusakov S. G., Corresp. Member of RAS,
Dr. Sci. (Tech.), Prof.
Ryabov G. G., Corresp. Member of RAS,
Dr. Sci. (Tech.), Prof.
Soifer V. A., Member of RAS,
Dr. Sci. (Tech.), Prof.
Sokolov I. A., Member of RAS,
Dr. Sci. (Phys.-Math.), Prof.
Suetin N. V.,
Dr. Sci. (Phys.-Math.), Prof.
Chaplygin Yu. A., Member of RAS,
Dr. Sci. (Tech.), Prof.
Shakhnov V. A., Corresp. Member of RAS,
Dr. Sci. (Tech.), Prof.
Shokin Yu. I., Member of RAS,
Dr. Sci. (Tech.), Prof.
Yusupov R. M., Corresp. Member of RAS,
Dr. Sci. (Tech.), Prof.

Editorial Board Members:

Avdoshin S. M., Cand. Sci. (Tech.), Ass. Prof.
Antonov B. I.
Barsky A. B., Dr. Sci. (Tech.), Prof.
Vasenin V. A., Dr. Sci. (Phys.-Math.), Prof.
Vishnekov A. V., Dr. Sci. (Tech.), Prof.
Dimitrienko Yu. I., Dr. Sci. (Phys.-Math.), Prof.
Domrachev V. G., Dr. Sci. (Tech.), Prof.
Zaborovsky V. S., Dr. Sci. (Tech.), Prof.
Zagidullin R. Sh., Cand. Sci. (Tech.), Ass. Prof.
Zarubin V. S., Dr. Sci. (Tech.), Prof.
Karpenko A. P., Dr. Sci. (Phys.-Math.), Prof.
Kolin K. K., Dr. Sci. (Tech.)
Kulagin V. P., Dr. Sci. (Tech.), Prof.
Kureichik V. V., Dr. Sci. (Tech.), Prof.
Ljvovich Ya. E., Dr. Sci. (Tech.), Prof.
Martynov V. V., Dr. Sci. (Tech.), Prof.
Mikhailov B. M., Dr. Sci. (Tech.), Prof.
Nechaev V. V., Cand. Sci. (Tech.), Ass. Prof.
Poleschuk O. M., Dr. Sci. (Tech.), Prof.
Saksonov E. A., Dr. Sci. (Tech.), Prof.
Sokolov B. V., Dr. Sci. (Tech.)
Timonina E. E., Dr. Sci. (Tech.), Prof.
Uskov V. L. (USA), Dr. Sci. (Tech.)
Fomichev V. A., Dr. Sci. (Tech.), Prof.
Shilov V. V., Cand. Sci. (Tech.), Ass. Prof.

Editors:

Bezmenova M. Yu.
Grigorin-Ryabova E. V.
Lysenko A. V.
Chugunova A. V.

Complete Internet version of the journal at site: <http://novtex.ru/IT>.

According to the decision of the Higher Certifying Commission of the Ministry of Education of Russian Federation, the journal is inscribed in "The List of the Leading Scientific Journals and Editions wherein Main Scientific Results of Theses for Doctor's or Candidate's Degrees Should Be Published"

МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ

MODELING AND OPTIMIZATION

УДК 004.9

С. П. Ковалев, вед. науч. сотр., e-mail: kovalyov@nm.ru,
ИПУ им. В. А. Трапезникова РАН,
М. Ю. Шаймарданов, студент, e-mail: mikhail.shaym@gmail.com,
МГТУ им. Н. Э. Баумана

Метод учета технологических ограничений при оптимизации топологии изделий машиностроения

Поставлена и решена задача разработки алгоритма топологической оптимизации с учетом технологических ограничений. Алгоритм позволяет автоматически получить оптимальные решения, учитывающие технологические направления роста, в цикле проектирования изделий машиностроения. Алгоритм получен путем дополнения классического SIMP-метода топологической оптимизации ограничениями типа направлений роста.

Ключевые слова: топологическая оптимизация, SIMP-метод, технологические ограничения, метод направлений роста, проектирование, машиностроение, алгоритм, вычислительный эксперимент

Введение

В последнее время активно исследуют возможность топологической оптимизации для построения оптимальной формы изделий. *Топологическая оптимизация* (англ. *topology optimization*) — метод оптимизации формы конструкции, с помощью которого получается рациональное распределение материала в деталях изделия в зависимости от заданных нагрузок, граничных условий и ограничений на массу и/или объем.

Топологическая оптимизация основывается на конечно-элементной модели проектной области. Каждый элемент модели характеризуется некоторой переменной — относительной плотностью элемента, которая определяет количество материала в этом элементе. Относительная плотность находится в диапазоне от 0 до 1.

В некоторых программных пакетах САПР возможность топологической оптимизации реализована как встроенная стандартная функция [1–3]. Однако применение такого рода оптимизации на практике — при разработке реальных деталей — все еще ограничено, поскольку существующие в настоящее время методы и алгоритмы топологической оптимизации зачастую приводят к формированию нетехнологичной формы модели — в детали появляются полости и ячеистые структуры. Детали с такими элементами достаточно сложны, и их невозможно либо дорого изготовить. Кроме того, сам процесс ин-

терпретации результатов топологической оптимизации оказывается сложным. Следовательно, существует потребность в разработке таких алгоритмов топологической оптимизации, которые бы учитывали технологические ограничения при синтезе оптимальной формы деталей.

В настоящей работе предложен модифицированный алгоритм топологической оптимизации, способный обеспечить учет технологических ограничений и при этом не приводящий к расходимости алгоритма или значительному росту требуемых вычислительных ресурсов. Для задания технологических ограничений используют понятие *направление роста* (англ. *growth direction*) — такое направление, вдоль которого плотность материала может убывать либо оставаться постоянной [4]. Предложенный алгоритм позволяет генерировать оптимальную форму изделия с учетом от одного до четырех и более направлений роста [5]. Изделия такой формы эффективны в изготовлении посредством широко распространенных технологий производства — литья, резания, штамповки.

В литературе выделяют три основных группы технологических ограничений:

- ограничения при производстве деталей литьем/резанием, штамповкой, ковкой, механической обработкой [4, 5];
- ограничения при производстве деталей с использованием аддитивных технологий (3D-печать) [6];
- ограничения сборочного производства [7].

В рамках работы рассматриваем только первую группу ограничений, поскольку технологии производства деталей типа литья/резания в настоящее время доминируют в машиностроении. Для простоты описания предложенного алгоритма рассматриваем топологическую оптимизацию плоских деталей постоянной фиксированной толщины. Таким образом, оптимизация проводится только в двух измерениях. Результаты работы непосредственно обобщаются на трехмерный случай [8].

Существуют следующие методы моделирования топологии детали с учетом ограничений производства при литье/резании:

- задание ограничений в виде минимального размера конечного элемента и применение геометрической симметрии проектной области с использованием метода движущихся асимптот и вейвлетов (при производстве деталей резанием) [9];
- задание ограничений в виде допустимых направлений вытягивания или угла отливки (при производстве деталей литьем) [10];
- использование ограничений по "направлениям роста" (при производстве деталей литьем/резанием, штамповкой, ковкой, механической обработкой) [4, 5].

Первые два метода [9, 10] предполагают ограничения только для указанных технологий производства. Ограничения по направлениям роста применимы для более широкого спектра технологий, поэтому именно они выбраны в качестве основных для представленной в работе модификации алгоритма топологической оптимизации.

Постановка задачи топологической оптимизации

В качестве целевой функции метода топологической оптимизации выступает *податливость* (англ. *compliance*) конструкции — способность конструкции сопротивляться деформации при внешнем воздействии. Основной характеристикой податливости является коэффициент жесткости, равный силе, которую необходимо приложить для единичного перемещения конструкции в точке приложения силы.

Проектной областью (англ. *design domain*) называем такую часть конструкции, топологию которой необходимо оптимизировать.

Классическим методом решения задачи топологической оптимизации является *SIMP-метод* (англ. *Solid Isotropic Material with Penalisation* — метод твердого изотропного материала со штрафованием) [11]. Этот метод предполагает, что модуль упругости материала нелинейно зависит от значения его относительной плотности.

Без штрафования относительная жесткость E_e/E_0 линейно зависит от плотности материала,

$E_e = \rho_e E_0$, где E_e — модуль Юнга для конечного элемента; E_0 — модуль Юнга материала изделия; $0 \leq \rho_e \leq 1$ — плотность. Для штрафования промежуточных плотностей представляем модуль Юнга материала элемента в виде

$$E_e(\rho_e) = E_{\min} + \rho_e^p (E_0 - E_{\min}),$$

где p — степень штрафования, как правило, равная 3; $E_{\min} = 10^{-9}$ — модуль Юнга для пустот, который вводится в целях предотвращения случаев, когда матрица жесткости становится вырожденной.

Использование степени $p > 1$ позволяет налагать штраф на элементы, в которых появляются промежуточные плотности материала в процессе оптимизации. Таким образом, в результате применения SIMP-метода отсекаются промежуточные плотности и получается бинарное решение — элементы конструкции принимают значения относительной плотности, равные либо единице, либо нулю.

На рис. 1 приведен пример решения задачи оптимизации топологии детали [5], проектная область которой имеет форму параллелепипеда. Налагаемое ограничение на перемещение — неподвижная заделка по всей левой границе. Сила, приложенная в центральной точке правой границы, направлена вниз. Необходимо сократить объем детали в 2 раза (относительное значение целевого объема равно 0,5).

Математическую задачу топологической оптимизации без учета технологических ограничений [1, 2] рассматриваем в следующей постановке:

$$\min_{\rho \in D_\rho} c(\rho) = \sum_{e=1}^E E_e(\rho_e) \mathbf{u}_e^T \mathbf{k}_e \mathbf{u}_e; \quad (1)$$

$$D_\rho = \begin{cases} V_T = V_0 f; \\ 0 \leq \rho_e \leq 1. \end{cases} \quad (2)$$

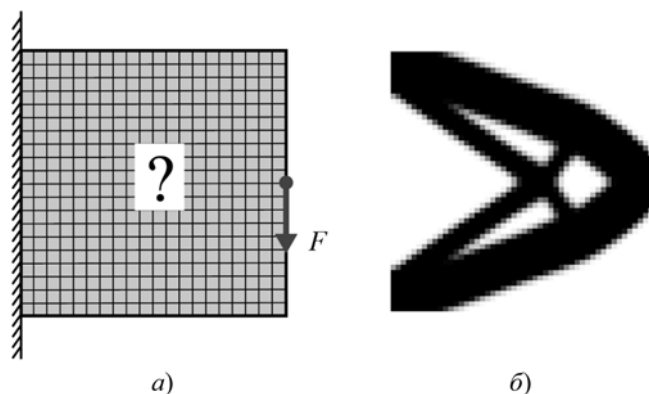


Рис. 1. Решение задачи и результат применения SIMP-метода топологической оптимизации [5]:

а — сечение проектной области с заданными нагрузкой и граничным условием; б — результат оптимизации после применения SIMP-метода

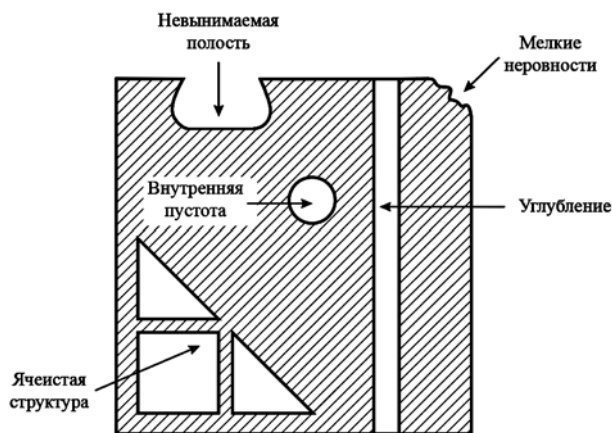


Рис. 2. Особенности топологической оптимизации, подлежащие устранению

Здесь $c(\rho)$ — целевая функция (податливость конструкции); ρ — вектор проектных переменных (т.е. относительных плотностей конечных элементов); E — общее число конечных элементов проектной области; ρ_e — относительная плотность e -го элемента; u_e — вектор перемещения e -го элемента; k_e — матрица жесткости этого элемента; V_0, V_T — исходный и целевой объемы проектной области соответственно; f — относительное значение целевого объема; D_p — ограничение задачи.

При заданных нагрузках, граничных условиях (рис. 1, а) и $f = 0,5$ конструкция балки в результате топологической оптимизации SIMP-методом принимает ячеистый вид (рис. 1, б). В более общих случаях появляются особенности конструкции, резко снижающие технологичность изготовления детали. Эти особенности относят [4] к следующим пяти типам (рис. 2).

- **Ячеистая структура конструкции.** Детали с ячеистой структурой очень сложно выполнить с приемлемыми затратами при использовании исключительно классических технологий производства.
- **Углубления в детали.** Если эти углубления являются необязательными, от них нужно отказываться.

- **Невынимаемые полости.** Для повышения технологичности деталь не должна иметь полости по типу показанной на рис. 2.
- **Внутренние пустоты.** Такие пустоты нужно ликвидировать, поскольку их реализация при производстве деталей механообработкой невозможна.
- **Мелкие неровности.** Точное воспроизведение мелких неровностей всегда сложно и дорого. Их необходимо сглаживать.

Метод учета ограничений производственных технологий

В предлагаемом нами методе ограничения по направлениям роста выступают в качестве дополнительных в SIMP-методе топологической оптимизации. Основная идея заключается в том, что чем ближе элемент геометрической модели располагается к границе раздела роста (англ. *growth surface*), тем выше должна быть плотность этого элемента. *Граница раздела роста* [4] — поверхность, которую выбирает проектировщик в соответствии с геометрическими характеристиками проектируемой детали. В качестве такой поверхности можно принять одну или несколько границ детали либо одну или несколько поверхностей, проходящих через деталь, например плоскость симметрии.

Пусть ρ_q — относительная плотность q -го элемента и этот элемент расположен ближе к границе раздела роста, чем r -й элемент. Ограничение по направлению роста в этом случае имеет вид неравенства $\rho_q \geq \rho_r$. На рис. 3 представлены рассмотренные ранее типы нетехнологичных форм деталей. В качестве границы раздела роста выступает плоскость yOz (след плоскости показан штриховой линией). Направления роста показаны жирными стрелками, начало которых находится на границе раздела роста.

Случай 1 (рис. 4, а). Ограничения по одному направлению роста

Если направление роста параллельно одной из сторон области проектирования (рис. 4, а), то

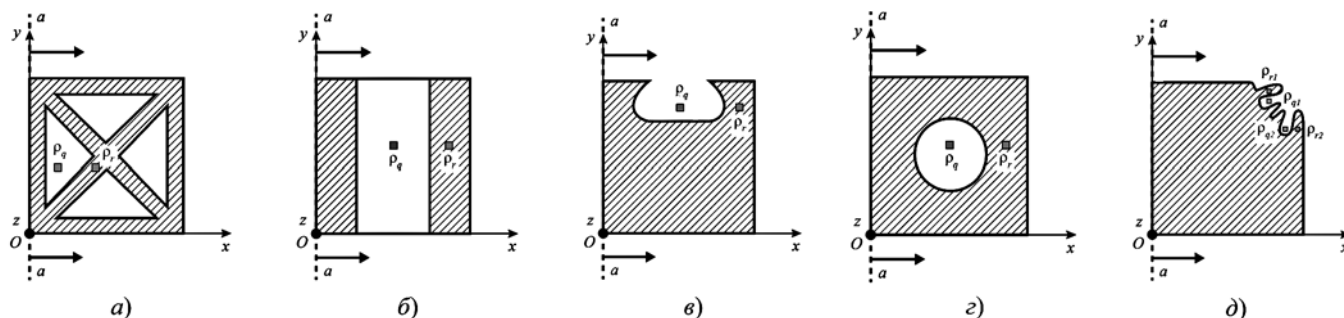


Рис. 3. Нарушение ограничений по направлениям роста $\rho_q \geq \rho_r$ для деталей с нетехнологичной формой:

а — ячеистая структура; б — углубление; в — внешняя полость; г — внутренняя пустота; д — мелкие неровности

ограничение по этому направлению представляем в виде

$$\rho_{1,m} \leq \rho_{2,m} \leq \dots \leq \rho_{N,m}, \quad m \in [1 : M], \quad (3)$$

где $\rho_{1,m}, \rho_{2,m}, \dots, \rho_{N,m}$ — относительные плотности элементов вдоль направления роста; N — число элементов вдоль этого направления; m, M — индексы и число элементов в направлении, перпендикулярном направлению роста, соответственно.

С учетом ограничения по одному направлению роста задачу топологической оптимизации (1), (2) формулируем в виде

$$\min_{\rho \in D_p} c(\rho) = \sum_{i=1}^N \sum_{j=1}^M E_{i,j}(\rho_{i,j}) \mathbf{u}_{i,j}^T \mathbf{k}_{i,j} \mathbf{u}_{i,j}, \quad (4)$$

$$D_p = \begin{cases} V_T = V_0 f, \\ 0 \leq \rho_{1,m} \leq \rho_{2,m} \leq \dots \leq \rho_{N,m} \leq 1, \\ m \in [1 : M]. \end{cases} \quad (5)$$

Предлагаем решить задачу (4), (5) с помощью итерационного метода *критерия оптимальности* [11], схема которого имеет вид

$$\begin{aligned} \rho_{i,j}^0 &= f; \\ \rho_{i,j}^{k+1} &= \begin{cases} \max(0, (1 - \mu)\rho_{i,j}^k), & \text{если } \rho_{i,j}^k B_{i,j}^\eta \leq \max(0, (1 - \mu)\rho_{i,j}^k); \\ \min(1, (1 + \mu)\rho_{i,j}^k), & \text{если } \rho_{i,j}^k B_{i,j}^\eta \leq \min(1, (1 + \mu)\rho_{i,j}^k); \\ \rho_{i,j}^k B_{i,j}^\eta & \text{в остальных случаях,} \end{cases} \\ i &\in [1 : N], j \in [1 : M]. \end{aligned}$$

Здесь k — номер итерации; переменная η — коэффициент затухания (принимаям $\eta = 0,5$); $\mu = 0,2$ — положительный предел перемещения конечного элемента, т.е. максимальное значение,

на которое может сдвинуться элемент при заданной нагрузке; $B_{i,j}$ определяем по формуле

$$B_{i,j} = \frac{-\frac{\partial c(\rho)}{\partial \rho_{i,j}}}{\lambda \frac{\partial V(\rho)}{\partial \rho_{i,j}}}, \quad (6)$$

где λ — множитель Лагранжа, который находим с помощью метода половинного деления [12]. Частные производные в выражении (6) определяем по формулам

$$\begin{aligned} \frac{\partial c(\rho)}{\partial \rho_{i,j}} &= -p \rho_{i,j}^{p-1} \mathbf{u}_{i,j}^T \mathbf{k}_{i,j} \mathbf{u}_{i,j}; \\ \frac{\partial V(\rho)}{\partial \rho_{i,j}} &= 1. \end{aligned}$$

В качестве условия окончания итераций используем условие

$$\max |c(\rho^{k+1}) - c(\rho^k)| \leq \varepsilon_c,$$

где ε_c — требуемая точность решения по c ; k — номер итерации.

Случай 2 (рис. 4, б). Ограничения по двум направлениям роста

Ограничения по двум противоположным направлениям имеют вид

$$\begin{aligned} \rho_{1,j} \leq \rho_{2,j} \leq \dots \leq \rho_{\frac{N}{2},j}; \\ \rho_{N,j} \leq \rho_{N-1,j} \leq \dots \leq \rho_{\frac{N}{2}+1,j}; \\ j \in [1 : M]. \end{aligned} \quad (7)$$

Здесь в качестве границы раздела роста принимаем плоскость $a-a$ (параллельна плоскости yOz). Конструкция "выращивается" от границы раздела вдоль двух противоположных направлений роста.

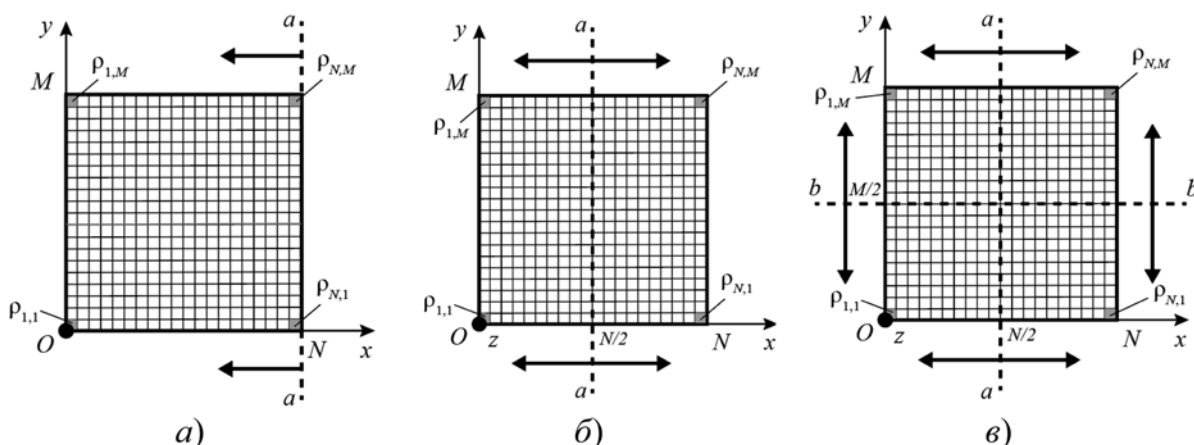


Рис. 4. Направления роста для прямоугольной области проектирования:

$a-a$, $b-b$ — следы границ раздела роста: a — одно направление роста; b — два направления роста; v — четыре направления роста

Ограничения по двум направлениям хорошо подходят для получения деталей, которые производят только с помощью литья без применения литейных стержней.

Случай 3 (рис. 4, в). Ограничения по четырем направлениям роста

В более общем случае требуются ограничения по четырем направлениям роста. Для этого добавляем к (7) аналогичные неравенства по вертикальным направлениям (вводим горизонтальную границу раздела роста $b-b$, параллельную плоскости xOz):

$$\begin{aligned} \rho_{i,1} &\leq \rho_{i,2} \leq \dots \leq \rho_{i,\frac{M}{2}}; \\ \rho_{i,M} &\leq \rho_{i,M-1} \leq \dots \leq \rho_{i,\frac{M}{2}+1}; \\ i &\in [1 : N]. \end{aligned} \quad (8)$$

Аналогичным путем (с соответствующей модификацией индексов элементов) могут быть сформулированы ограничения по произвольным направлениям роста [5].

Программная реализация и вычислительный эксперимент

В качестве программной среды для экспериментальной реализации разработанного алгоритма была выбрана система MATLAB. Нами предложена M-SIMP модификация программы, разработанной О. Sigmund [13, 14]. В классический алгоритм внедрены рассмотренные выше ограничения по четырем направлениям роста. Существует возможность изменять программу путем варьирования параметров SIMP-метода, ограничений конструкции и нагрузок. По ходу процесса топологической оптимизации в промежутках между итерациями классического SIMP-метода выполняем модификацию значений плотностей конечных элементов модели.

В вычислительном эксперименте мы решали задачу оптимизации топологии плоской детали. Проектная область детали имеет форму квадрата, налагаемое ограничение на перемещение — неподвижная заделка по левой границе, сила $F = 2000$ Н направлена вниз и приложена в центральной точке правой границы (см. рис. 1, а). Точность по податливости $\varepsilon_c = 0,01$.

На рис. 5 показаны промежуточные результаты работы программы M-SIMP-метода топологической оптимизации. Алгоритм сошелся на 58-й итерации (рис. 6). На рис. 6 представлен результат топологической оптимизации конструкции с учетом технологических ограничений.

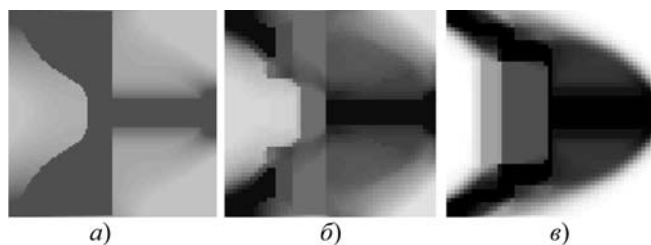


Рис. 5. Процесс оптимизации с учетом технологических ограничений:

а — итерация 1; б — итерация 2; в — итерация 5



Рис. 6. Результат топологической оптимизации, полученный с помощью программы M-SIMP

Заключение

Предложенный метод учета ограничения по нескольким направлениям роста позволяет в процессе топологической оптимизации получать оптимальную и в то же время технологичную форму изделия. В большинстве случаев достаточно применить ограничения по четырем направлениям роста (7), (8), чтобы сформированная деталь отличалась высокой технологичностью для традиционных процессов литья или резания.

С помощью разработанной программы M-SIMP можно достаточно эффективно проводить топологическую оптимизацию деталей проектирования, имеющих форму плоского параллелепипеда. При этом удалось сохранить сходимость процесса топологической оптимизации в целом и не допустить значительного роста потребности в ресурсах памяти и процессорного времени. Несложно распространить разработанный метод топологической оптимизации для учета рассмотренных ограничений при проектировании неплоских деталей. С этой целью можно использовать идею *областей видимости* [8].

В развитие данной тематики исследований авторы планируют разработать алгоритм топологической оптимизации с учетом технологических ограничений для сборных сварных конструкций.

Список литературы

1. Olason A., Tidman D. Methodology for Topology and Shape Optimization in the Design Process // Chalmers Publication Library. 2010. 66 p.
2. Blattman W. R. Generating CAD Parametric Features Based on Topology Optimization Results // All Theses and Dissertations. 2008. Paper 1371.
3. Johnsen S. Structural Topology Optimization. Basic Theory, Methods and Applications // Master Thesis. Norwegian University of Science and Technology. 2013.
4. Lu J. N., Chen Y. H. Manufacturable mechanical part design with constrained topology optimization // Journal of Engineering Manufacture. 2012. Vol. 226 (10). P. 1727–1735.
5. Ковалев С. П., Шаймарданов М. Ю. Учет технологических ограничений при топологической оптимизации в цикле проектирования изделий машиностроения // Системы проектирования, технологической подготовки производства и управления этапами жизненного цикла промышленного продукта (CAD/CAM/PDM-2016) // Труды 16-й международной конференции. М.: ООО "Аналитик", 2016. С. 399–402.
6. Brackett D., Ashcroft I., Harue R. Topology optimization for additive manufacturing. Loughborough University Institutional Repository. 2011.
7. Cui A., Zhang S., Xu L., Liu H. Topology Optimization and Robust Analysis of Welling Spot Layout for a Heavy-Duty Truck Cab Based on Element Strain Energy Density // Advanced Material Research. 2014. Vols. 887–888. P. 1284–1289.
8. Chen Y., Lu J., Wei Y. Topology optimization for manufacturability based on the visibility map // Computer-Aided Design and Applications. 2016. Vol. 13. P. 86–94.
9. Zuo K. T., Chen L. P., Zhang Y. Q., Yang J. Z. Manufacturing- and machining-based topology optimization // Int. J. Advd. Mfg. Technol. 2006. Vol. 27. P. 531–536.
10. Niclas S. Topology optimization of structures with manufacturing and unilateral contact constraints by minimizing an adjustable compliance—volume product // Struct Multidisc Optim. 2010. Vol. 42. P. 341–350.
11. Bendsoe M. P., Sigmund O. Topology optimization: theory, methods and applications. Springer-Verlag. 2003. 370 p.
12. Burden R. L., Faires J. D. The Bisection Method // Numerical Analysis. 2010. P. 48–56.
13. Sigmund O. A 99 line topology optimization code written in Matlab // Struct Multidisc Optim. 2001. Vol. 21. P. 120–127.
14. Andreassen E., Clausen A., Sigmund O., Lazarov B. S. Efficient topology optimization in Matlab using 88 lines of code // Struct Multidisc Optim. 2011. Vol. 43. P. 1–16.

S. P. Kovalyov, Leading Researcher, e-mail: kovalyov@nm.ru

V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

M. Yu. Shaymardanov, Student, e-mail: mikhail.shaym@gmail.com

Bauman Moscow State Technical University

Methods of Applying Manufacturing Constraints in Topology Optimization

In this paper, we formulated and solved the problem of development of topology optimization algorithm, which can be used to get optimization solutions automatically, taking into account the manufacturing constraints kind of growth directions in the cycle of engineering products design. The algorithm is obtained by the modification of the classic SIMP-method of topology optimization. We have modified the Matlab program that visualizes the optimization process in order to apply "four growth directions" constraints.

Keywords: topology optimization, SIMP-method, manufacturing constraints, growth direction method, design, engineering, algorithm, computational experiment

References

1. Olason A., Tidman D. Methodology for Topology and Shape Optimization in the Design Process, *Master Thesis. Chalmers Publication Library*, 2010. 66 p.
2. Blattman W. R. Generating CAD Parametric Features Based on Topology Optimization Results, *All Theses and Dissertations*, 2008. Paper 1371.
3. Johnsen S. Structural Topology Optimization. Basic Theory, Methods and Applications, *Master Thesis. Norwegian University of Science and Technology*, 2013.
4. Lu J. N., Chen Y. H. Manufacturable mechanical part design with constrained topology optimization, *Journal of Engineering Manufacture*, 2012, vol. 226 (10), pp. 1727–1735.
5. Kovalyov S. P., Shaymardanov M. Yu. Uchet tekhnologicheskikh ograniczeniy pri topologicheskoy optimizatsii v tsikle proektirovaniya izdeliy mashinostroeniya (Manufacturing constraints in topology optimization in the design cycle of engineering products), *Computer-Aided Design, Computer-Aided Manufacturing and Product Data Management (CAD/CAM/PDM-2016). Proceedings of the 16th international conference, Moscow, ООО "Analitic". 2016, pp. 399–402 (in Russian).*
6. Brackett D., Ashcroft I., Harue R. Topology optimization for additive manufacturing. *Loughborough University Institutional Repository*. 2011.
7. Cui A., Zhang S., Xu L., Liu H. Topology Optimization and Robust Analysis of Welling Spot Layout for a Heavy-Duty Truck Cab Based on Element Strain Energy Density, *Advanced Material Research*, 2014, vols. 887–888, p. 1284–1289.
8. Chen Y., Lu J., Wei Y. Topology optimization for manufacturability based on the visibility map, *Computer-Aided Design and Applications*, 2016, vol. 13, pp. 86–94.
9. Zuo K. T., Chen L. P., Zhang Y. Q., Yang J. Z. Manufacturing- and machining-based topology optimization, *Int. J. Advd Mfg Technol*, 2006, vol. 27, pp. 531–536.
10. Niclas S. Topology optimization of structures with manufacturing and unilateral contact constraints by minimizing an adjustable compliance—volume product, *Struct Multidisc Optim*, 2010, vol. 42, pp. 341–350.
11. Bendsoe M. P., Sigmund O. *Topology optimization: theory, methods and applications*, Springer-Verlag. 2003. 370 p.
12. Burden R. L., Faires J. D. 2.1 *The Bisection Method*, *Numerical Analysis* (9th edition). 2010, pp. 48–56.
13. Sigmund O. A 99 line topology optimization code written in Matlab, *Struct Multidisc Optim*, 2001, vol. 21, pp. 120–127.
14. Andreassen E., Clausen A., Sigmund O., Lazarov B. S. Efficient topology optimization in Matlab using 88 lines of code, *Struct Multidisc Optim*, 2011, vol. 43, pp. 1–16.

В. Д. Чертовской, д-р техн. наук, профессор, проф. каф., e-mail: vdchertows@mail.ru,
Санкт-Петербургский государственный университет морского и речного флота
имени адмирала С. О. Макарова

Методология математического описания и моделирования адаптивной автоматизированной системы управления производством

Исследованы иерархические адаптивные системы управления производством, в которых в оптимальном процессе функционирования меняется состав вектора цели. Проведен системный анализ процедур математического описания и компьютерной реализации. Предложены методы описания и получения числовой информации для моделей систем. Рассмотрены особенности и выбраны программная структура компьютерной реализации, методы согласования числовых данных.

Ключевые слова: системный анализ, адаптивная система управления, производство, планирование, описание, моделирование

Введение

Автоматизированные системы управления предприятием [1] с 60-х годов XX века прошли большой путь и получили широкое распространение. Сказанное в полной мере относится и к части таких систем — автоматизированным системам управления производством, охватывающим только бизнес-процесс "Производство" (рис. 1). В них уже не в полной мере удовлетворяются современные требования: в системах используется неоптимальный режим работы, а оперативной реакции на изменение состава вектора спроса не предусмотрено.

Указанные недостатки отсутствуют у адаптивных автоматизированных систем управления серийным производством (ААСУП), являющихся новым классом систем управления [2–7]. Новизна класса требует рассмотрения для него процедур математического описания и компьютерного моделирования.

Если вопросы описания нашли хорошее отражение в литературе [8, 9], то вопросы моделирования изучены явно недостаточно. Следует добавить, что между процедурами сильно взаимосвя-

зание, что делает необходимым их исследование с позиций системного анализа.

Постановка задачи. Целью настоящей работы является системное изучение ААСУП с позиций оптимизации режима работы и обеспечения оперативной реакции на быстрое изменение состава вектора рыночного спроса на продукцию. Использование адаптивных систем предполагает рассмотрение методологий математического описания и моделирования с последующим переходом к прикладному аспекту. Акцент в данной работе сделан не столько на формировании методологий, сколько на их применение в прикладном аспекте.

Необходимо, следовательно, системно рассмотреть процедуры описания и моделирования.

Математическое описание ААСУП

Цель функционирования системы определена. Она предполагает достаточно сложную структуру бизнес-процесса "Производство" серийного типа (рис. 2).

Такую структуру трудно описывать и тем более исследовать из-за большой разницы масштабов по времени на верхнем уровне (годы) и на нижнем уровне (часы и даже минуты). Ее целесообразно представить в виде трех блоков (рис. 2). Процессы в блоке 3 характерны для единичного типа производств, а процессы в блоках 1 и 2, представляющие наибольший интерес, — для серийного типа. Поскольку условия работы в двух последних блоках похожи, ограничимся рассмотрением блока 1.

Построение прикладного описания проводится на основе методологии, поэтому приведем

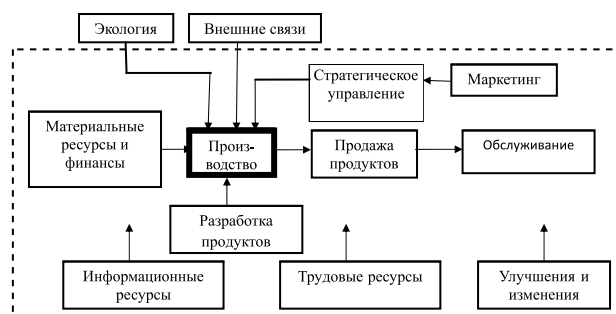


Рис. 1. Процедурное представление автоматизированной системы

ее краткое описание. Вопросам методологии посвящены работы [5–7]. В работе [5] рассматривается теоретико-множественное описание лишь двухуровневой системы. В работах [6, 7] четко не очерчен изучаемый тип производства, однако

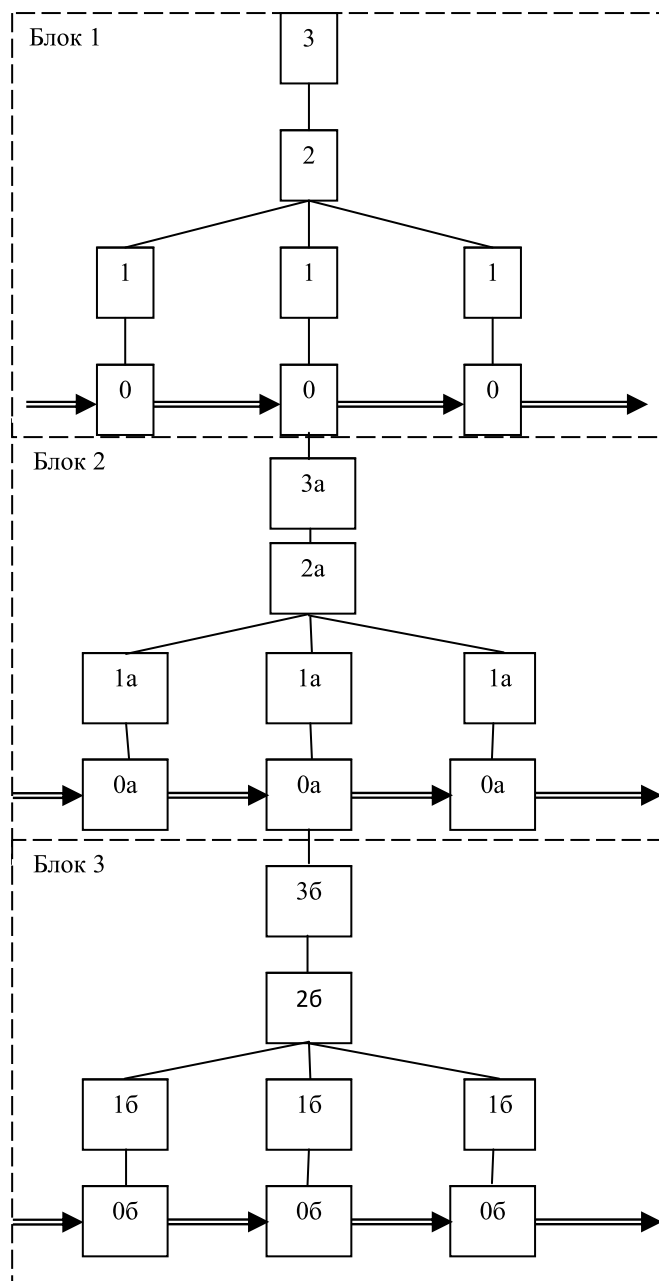


Рис. 2. Общая структура ААСУП:

0 — цехи; 1 — начальники цехов; 2 — диспетчер; 3 — руководство; 0а — участки; 1а — начальники участков; 2а — цепочки участков; 3а — начальник цеха; 0б — согласование работы элементов в пределах выпуска одного вида продукции или партии однотипной продукции (один цикл); 1б — согласование работы элементов в пределах выпуска партии однотипной продукции при наличии циклов для нескольких видов ресурсов; 2б — организация работы системы элементов в пределах выпуска одного вида продукции или партии разнотипной продукции с пересекающимися циклами для разных видов ресурсов; 3б — начальник участка

речь идет об управлении потоками (работами), т.е. о единичном типе производства. Описание процесса проводится фактически на теоретико-множественном уровне. Исследование взаимодействия элементов в двухуровневой системе с веерной структурой в одном масштабе времени на разных уровнях достаточно трудоемко. Управление осуществляется путем корректировки плана, при этом возмущения на рассчитываемом интервале времени определяются по результатам имитационного моделирования на основе данных за предыдущий период времени.

Методология описания представлена автором в работе [4], приведем ее основные положения. Концепция структурно-алгоритмического описания базируется на следующих основных принципах:

- первичность цели проектирования и структуры объекта управления по отношению к структуре управляющей части;
- рассмотрение цикла управления с выделением связанных процессов планирования и управления;
- описание процессов единым методом с использованием составного критерия, учитывающего экономические и динамические свойства;
- учет целенаправленности структурных элементов с согласованием их экономических интересов по вертикали и горизонтали;
- учет динамичности рынка.

Подходы к построению обобщенного описания определяются спецификой адаптивных систем. К специфическим особенностям относятся:

- трехуровневая структура системы;
- использование оптимального режима работы;
- применение адаптивного управления из-за динамичности рынка;
- интеграция описания проектирования и эксплуатации, функционирования и адаптации системы;
- специфика получения числовых данных.

Автором [4] сформировано трехуровневое обобщенное описание, в котором элементы представлены на теоретико-множественном уровне, а связи — с помощью графов. Там же приведена обобщенная технология работы с описанием, которую удобнее иллюстрировать в прикладной части.

При переходе к прикладному представлению и формированию математического описания была учтена наиболее важная особенность блока — изменение состава вектора цели (спроса), которое компенсируется изменением структурных связей.

С учетом особенностей были сформированы требования к методам описания иерархических систем [5]. Укажем основные требования:

- учет динамики процесса планирования;
- однотипное описание разных процессов планирования и управления;

- описание трехуровневой структуры;
- оптимизация процессов в системе.

Метод должен быть нагляден и обладать малым временем получения результата.

Анализ методов, существовавших на конец XX века, показал отсутствие таких методов для многоуровневых систем. Этот факт подтверждается и работами [10, 11]. Автору удалось построить два новых системных метода [4] (интегральный и однородный), удовлетворяющих поставленным требованиям. Однородный метод оказался более подходящим и получил большее распространение. Дополнительный анализ, проведенный в 2016 г. [12–14], подтвердил, что однородный метод жизнеспособен.

Однородный метод базируется на задаче динамического линейного программирования, состоящей из известной задачи статического линейного программирования (СЛП) и совокупности разностных уравнений.

Описание процесса управления этим методом для уровня цехов получает вид

$$\begin{aligned} \mathbf{z}_k(t_i) &= \mathbf{A}_k \mathbf{z}_k(t_{i-1}) + \mathbf{B}_k \mathbf{u}_k(t_{i-1}), \mathbf{z}_k(0) = \mathbf{z}_{k0}; \\ \mathbf{y}_k(t_i) &= \mathbf{C}_k \mathbf{z}_k(t_i); \\ \sum_{i=0}^{N-1} \mathbf{D}_1^m \mathbf{u}_{1k}(t_i) &\leq \mathbf{b}^m(0); \\ \mathbf{y}_k(T) &\leq \mathbf{Y}(T); \\ \mathbf{D}_k^m \mathbf{u}_k(t_i) &\leq \mathbf{b}_k^m(t_{i-1}); \\ \mathbf{D}_k^\Psi \mathbf{u}_k(t_i) &\leq \mathbf{b}_k^\Psi(t_{i-1}); \\ \mathbf{b}_k^\Psi(t_i) &\leq \mathbf{b}_k^\Psi(t_{i-1}) + \Delta \mathbf{b}_k^\Psi(t_{i-1}); \\ \boldsymbol{\varepsilon}_k(t_i) &= \mathbf{p}_k(t_i) - \mathbf{y}_k(t_i), \\ J_k &= \sum_{i=0}^N \{ \mathbf{C}_{1k} \boldsymbol{\varepsilon}_k(t_i) + \mathbf{C}_{2k} \mathbf{u}_k(t_i) \} \rightarrow \min; \\ i &= 0, N-1, t_i = iv, t_0 = 0, T = Nv, \end{aligned} \quad (1)$$

где $\mathbf{z}$, $\mathbf{y}$ — вектор-столбцы состояния и выхода; $\mathbf{u}$ — вектор-столбец управления; $\mathbf{p}$ — вектор-столбец плана; $\boldsymbol{\varepsilon}$ — вектор-столбец отклонений; $\mathbf{D}$ — матрица норм расходов, $\mathbf{b}$ — вектор-столбец наличного количества ресурсов; $\mathbf{b}_m(0)$ — вектор количества материальных ресурсов, которыми располагает уровень $h = 3$; $\Delta \mathbf{b}$ — поступление ресурсов; $\mathbf{Y}$ — вектор-столбец управления уровня $h = 3$; $\mathbf{C}_1$, $\mathbf{C}_2$ — вектор-строки потерь от отклонений и затрат на дополнительные ресурсы для управления; $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ — матрицы, отражающие динамику процесса управления v ; T — минимальный интервал времени и время моделирования; $m = 1, M$ — виды материальных ресурсов; $\psi = 1, \Psi$ — виды прочих ресурсов; $i = 1, I$ — моменты времени; $k = 1, K$ — номер подразделения;

N — целое число; J_k — критерий оптимизации, представляющий собой сумму из потерь за счет отклонения выпуска продукции от плана и из затрат на управление для ликвидации отклонения.

Процедура целевой адаптации начинается с изменения размерности вектора $\mathbf{p}_k(t_i)$ с последующим изменением структурных связей между ресурсами и продуктами.

На уровне диспетчера устанавливаются горизонтальные связи.

$$\mathbf{D}_k^m \mathbf{u}_k(t_{i+1}) \leq \mathbf{y}_{k-1}(t_i);$$

$$J_k = \sum_{k=0}^K J_k \rightarrow \min.$$

Уровень руководителя описывается подобно уровню $h = 1$ с заменой интервала $[t]$ на $[T] = m[t]$, где m — целое число.

Выражения для процесса управления можно использовать и для описания процесса планирования с заменой, показанной в табл. 1, в которой выражения

$$\begin{aligned} \mathbf{D}_k \mathbf{p}_k &\leq \mathbf{b}_k, \\ G_k &= \mathbf{F}_k \mathbf{P}_k \rightarrow \max \end{aligned}$$

системы (1) представляют собой прямую задачу СЛП, в которой $\mathbf{p}_k$, $\mathbf{P}_k$, G_k — ежедневный выпуск продукции, выпуск с накоплением, критерий оптимизации. Задача может быть записана иначе (табл. 2). Наряду с прямой задачей, потребуются обратная задача, запись которой представлена в табл. 2.

Таким образом, был построен достаточно удачный и относительно простой метод описания, который стал основой для изучения процедуры моделирования.

Таблица 1

Соотношение переменных

Управление	$\mathbf{y}(t_i)$	$\mathbf{u}(t_i)$	$\boldsymbol{\varepsilon}(t_i)$	J	C	$\boldsymbol{\varepsilon}_k(t_i) = \mathbf{p}_k(t_i) - \mathbf{y}_k(t_i)$
Планирование	$\mathbf{p}(t_i)$	$\mathbf{p}_i(t_i)$	0	G	F	$\mathbf{p}_k(t_i) = \mathbf{y}_k(t_i)$

Таблица 2

Виды задач СЛП

Задача	Даны	Найти
Прямая	Векторы $\mathbf{R}$, $\mathbf{b}$, $\mathbf{F}$, матрица $\mathbf{D}$	Вектор $\mathbf{P}$
Обратная	Векторы $\mathbf{R}$, $\mathbf{P}$, матрица $\mathbf{D}$	Вектор $\mathbf{F}$, $\mathbf{b}$

Моделирование ААСУП

Методология моделирования учитывает следующую систему концептуальных принципов.

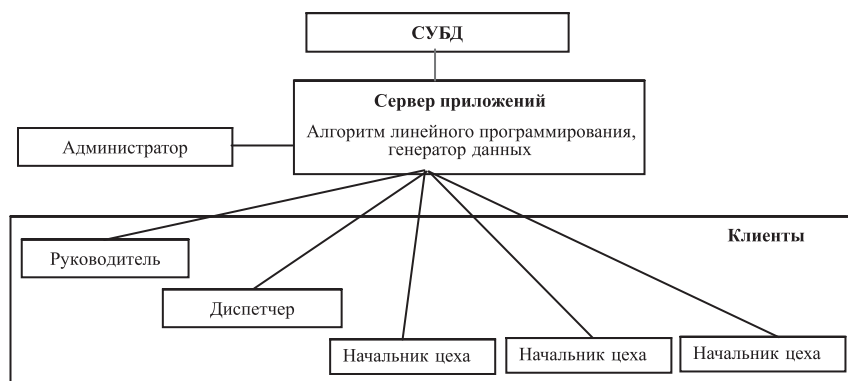


Рис. 3. Структура компьютерной реализации

1. Использование сетевой архитектуры "клиент — сервер" в силу иерархичности системы.
 2. Применение "тонкого" клиента с выделением составляющих: сервер базы данных, сервер приложений, программные средства клиента.
 3. Рассмотрение автоматизации управления, а не автоматизации документооборота.
- Подход к решению задачи определяется спецификой составляющих.

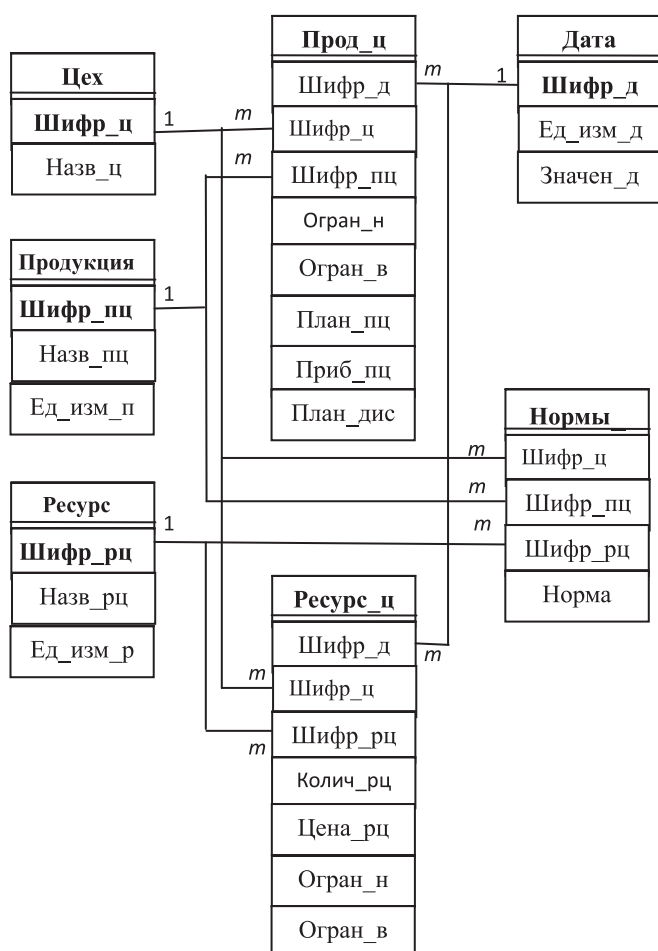


Рис. 4. Структура элементов базы данных уровней $h = 2$ и $h = 1$

Распределенная динамическая база данных должна обладать гибкостью доступа, простотой создания, широкой распространенностью, понятным интерфейсом.

Сервер приложений должен быть связан с СУБД, иметь открытый код, простую расширяемость, блочное построение, быть прост в настройке и широко распространен.

В программных средствах должны присутствовать алгоритмы и реализации разностных уравнений и оптимизации, модульность построения, оперативность апробации алгоритмов и способность стыковки с СУБД.

Обобщенная модель построена на основе обобщенного описания, а обобщенную технологию моделирования лучше иллюстрировать на прикладной модели.

В процедуре прикладного моделирования ААСУП [12] следует выделить построение и использование модели.

Построение модели. Здесь можно выделить построение структуры модели и программную работу с ней (табл. 3). Поскольку система является распределенной, целесообразно использовать архитектуру "клиент-сервер" с "тонким" клиентом (рис. 3).

Строится структура многоуровневой базы данных, которая для диспетчерского уровня может иметь вид, показанный на рис. 4.

Интерфейсы специфичны и требуют дополнительной проработки.

Локальные методы (табл. 3) включают в себя операции генерации/идентификации и обработку данных, связанные с их согласованием. Возможны различные сочетания алгоритмов операций, из которых представим один.

Обратимся к генерации данных. Первоначально генерируются согласованные данные (согласованный режим) для одного структурного

Таблица 3

Формирование структуры модели

Действие	Результат
Сетевое взаимодействие пользователей	Структура "клиент—сервер" и требования к программным средствам
Построение базы данных	Выбор СУБД
Формирование интерфейса	Интерфейсы администратора и клиентов
Локальные методы и коды используемых программ	Методы генерации/идентификации и согласования данных

элемента уровня $h = 1$ (цех), для чего используется алгоритм Р. Габасова для обратной задачи (см. табл. 2). Величины $\mathbf{P}$, $\mathbf{R}$, $\mathbf{A}$ задаются в диалоге; оптимальная оценка $\mathbf{DEL}$ и оптимальный потенциал $\mathbf{Y}$ генерируются функцией Rand:

$$\mathbf{b}_k = \mathbf{D}_k \mathbf{P}_k,$$

$$\mathbf{F}_{ko} = \mathbf{DEL}_k,$$

$$\mathbf{F}_k = \mathbf{F}_{ko} + \mathbf{D}_k \mathbf{Y}_k.$$

На уровне диспетчера ($h = 2$) вводятся горизонтальные связи

$$N_{k-1} = M_k, \quad (2)$$

$$\mathbf{D}_k^m \mathbf{p}_{1k} = \mathbf{p}_{k-1}, \quad k = \overline{2, K}, \quad (3)$$

где $N_{k-1} = M_k$ — размерность матрицы $\mathbf{D}_k$.

Далее вычисляются данные для уровня руководителя ($h = 3$).

$$\mathbf{D} = \prod_{k=1}^K \mathbf{D}_k;$$

$$\mathbf{G} = \sum_{k=1}^K \mathbf{F}_k \prod_{j=k+1}^K \mathbf{D}_j \mathbf{P}_K,$$

где

$$r = \begin{cases} k+1 & \text{при } k < K, \\ 0 & \text{при } k = K, \end{cases}$$

$\mathbf{D}_0 = \mathbf{I}$, $\mathbf{I}$ — единичная матрица.

Далее строится несогласованный режим. Для этого, в частности, достаточно нарушить выражения (2), (3).

Несогласованный режим получается, как правило, при специфической идентификации.

Согласование может осуществляться различными способами [13, 14], например, возвратом от несогласованных данных к согласованным с изменением коэффициентов целевых функций.

Следует особо отметить, что система методов генерации и согласования требует дополнительной детальной проработки.

Использование модели. Рассмотрим традиционный вариант работы. Первоначально администратором генерируются данные согласованного режима для уровня диспетчера. На их основе строятся данные для верхнего и нижнего уровней (см. рис. 2). Все данные, кроме оптимальных планов, передаются в базу данных. Клиенты запрашивают данные из базы данных (БД) для определения оптимальных планов, значения которых передаются в БД. Фактически осуществляется апробация модели.

Далее для тех же уровней формируются несогласованные данные и с ними проводятся те же операции.

Структурные элементы системы отличаются целенаправленностью в виде экономических интересов, которые проявляются в целевых функциях. В силу этого в процедуре функционирования требуется согласование интересов и данных. Согласование данных осуществляется сначала на уровне диспетчера, а потом — на других уровнях.

Данные, получаемые при идентификации, как правило, несогласованные, поэтому проводится и их согласование.

После завершения традиционного варианта те же операции осуществляют, переходя к адаптивному варианту. Однако следует предварительно удостовериться в целесообразности перехода на выпуск новой продукции.

Обсуждение результатов

Развитие традиционных автоматизированных систем управления производством привело к появлению и все большей потребности в новом классе — адаптивных систем, позволяющих оптимизировать процесс управления и автоматизировать процедуру оперативного перехода на выпуск новой продукции. Для прикладного применения таких систем необходимо системное изучение процедур их математического описания и моделирования.

Изучение процедуры описания позволило выявить на верхнем блоке трехуровневую структуру систем и (с учетом их особенностей) сформулировать требования к методам математического описания. Анализ, проведенный на основе требований, позволил построить новый метод описания — однородный метод, который лег в основу построения прикладной модели систем.

Для прикладной модели системы управления сформулированы требования к структуре программных средств (сервер базы данных, сервер приложений, программные файлы клиентов), позволившие выбрать архитектуру "клиент — сервер" с "тонким" клиентом. Специфичной оказалась процедура получения для модели числовых данных, которые, как показано, возможны в процедуре идентификации или генерации. При генерации числовых данных первоначально формируется согласованный режим, а затем — несогласованный. Идентификация числовых данных дает, как правило, несогласованные данные. Они согласуются с помощью предложенного локального метода.

Таким образом, автором проведено системное рассмотрение процедур от описания систем до их компьютерной реализации.

Полученный результат является фундаментом (каркасом) для развития системы локальных

методов функционирования адаптивных систем. Более того, он позволяет перейти к анализу процессов в нижнем блоке структуры системы.

Список литературы

1. Глушков В. М. Основы безбумажной информации. М.: Наука, 1982. 552 с.
2. Бобко И. М., Марчук Г. И., Аганбегян А. Г. Адаптивная АСУ производством // Под ред. Г. И. Марчука. М.: Статистика, 1981. 384 с.
3. Li Yihua. Анализ устойчивости АСУ // J. Chanhsha Univ. Elec. Powers, 1997, 12, N. 4. P. 408—411.
4. Советов Б. Я., Чертовской В. Д. Автоматизированное адаптивное управление производством. СПб.: Лань, 2002. 200 с.
5. Месарович М., Мако Д., Такахара Я. Теория иерархических систем / Пер. с англ. М.: Мир, 1973. 344 с.
6. Скурихин В. И., Забродский В. А., Копейченко Ю. В. Проектирование систем управления машиностроительным производством. Харьков: Вища школа, 1984. 206 с.
7. Скурихин В. И., Забродский В. А., Копейченко Ю. В. Адаптивные системы управления машиностроительным производством. М.: Машиностроение, 1989. 208 с.
8. Мизон В. А. Интеллектуальные методы управления предприятием. СПб.: Академия управления и экономики, 2008. 302 с.
9. Бурков В. Н. Основы математической теории активных систем. М.: Наука, 1977. 256 с.
10. Новиков Д. А. Теория управления организационными системами. М.: Физматгиз, 2012. 344 с.
11. Угольников Г. А., Усов А. Б. Трехуровневая система управления эколого-экономическими объектами // Проблемы управления. 2010. № 1. С. 26—32.
12. Чертовской В. Д. Моделирование процессов планирования в интеллектуальной системе управления производством // Материалы 9-й российской мультиконференции по проблемам управления "Информационные технологии в управлении (ИТУ-2016)" 4—6 октября. СПб.: Электроприбор, 2016. С. 536—544.
13. Чертовской В. Д. Анализ процесса согласованного автоматизированного планирования в иерархической системе управления производством // Информационные технологии. 2010. № 6 (166). С. 68—72.
14. Чертовской В. Д. Математическое описание и компьютерная реализация модели адаптивной автоматизированной системы управления производством // Информационно-управляющие системы. 2017. № 1 (86). С. 106—114.

V. D. Chertovskoy, Professor, e-mail: vdchertows@mail.ru,

Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg

Methology of Mathematical Description and Modeling of Adaptive Automatized Manufacturing Control Systems

Hierarchical adaptive manufacturing control systems was researched these work in optimal regime with change goal vector composition. System analyses of mathematical description and computer realization was made. Description method and method of obtain of numerical information for model of system were proposed. Singularities were considered and program structure of computer realization, methods of numeric data agreement were chosen.

Keywords: *system analyses adaptive manufacturing control system planning description modeling*

References

1. Glushkov V. M. *Osnovy bezbumazhnoy informacii* [Basis of Paperless Information]. Moscow: Nauka, 1982. 552 p. (in Russian).
2. Bobko I. M., Marchuk G. I., Aganbegjan A. G. *Adaptivnaya ASU proizvodstvom* [Adaptive Automatized Manufacturing Control System]. By edit G. I. Marchuk. Moscow: Statistika, 1981. 384 p. (in Russian).
3. Li Yihua. *Analiz ustoichivosti ASU* [Stability analysis of ACS] // J. Chanhsha Univ. Elec. Powers, 1997, vol. 12, no. 4, pp. 408—411.
4. Sovetov B. Y., Chertovskoy V. D. *Avtomatizirovannoe adaptivnoe upravlenie proizvodstvom* [Adaptive Automatized Manufacturing Control]. S.-Peterburg: Lan, 2002. 200 p. (in Russian).
5. Mesarovich M., Mako L., Takahara Y. *Teoriya ierarhicheskikh sistem*. [Hierarchical Systems Theory]. Per. s English. Moscow: Mir, 1973. 344 p. (in Russian).
6. Skurihin V. I., Zabrodskiy V. A., Koheychenko J. V. *Projektirovaniye sistemy adaptivnogo upravleniya mashinostroitelnykh proizvodstvom*. [Design of adaptive control systems of machine-building production]. Harkov: Visha shkola, 1984. 206 p. (in Russian).
7. Skurihin V. I., Zabrodskiy V. A., Koheychenko J. V. *Adaptivnye sistemy upravleniya mashinostroitelnykh proizvodstvom*. [Adaptive control systems of machine-building production]. Moscow, Mashinostroenie, 1989. 208 p. (in Russian).
8. Mison V. A. *Intellectualnye metody upravleniya predpriyatiem* [Intellectual Methods of Manufacturing Control]. S.-Peterburg: Academy of Control and Economic, 2008. 302 p. (in Russian).
9. Burkov V. N. *Osnovy teorii aktivnykh sistem*. [Basis of Active Systems Theory]. Moscow: Nauka, 1977. 256 p. (in Russian).
10. Novikov D. A. *Teoriya upravleniya organizatsionnymi sistemami* [Control Theory of Organization Systems]. Moscow: Fizmatgis, 2012. 344 p. (in Russian).
11. Ugolnitskiy G. A., Usov A. B. *Trehurovnevaya sistema ekologo-ekonomicheskimi objektami*. [Three-level Control System by Ecologically-economic Objects]. *Problemy upravleniya*, 2010, no. 1. pp. 26—32 (in Russian).
12. Chertovskoy V. D. *Modelirovaniye protsessov planirovaniya v intellektualnoy sisteme upravleniya proizvodstvom* [Modelling of Planning Processes in Intellectual Manufacturing Control System]. *Materialy 9 multikonferencii po problemam upravleniya "Informatsionnye tekhnologii v upravlenii (ITU-2016)" 4—6 oktyabrja*. Saint-Peterburg, Elektropribor, 2016. pp. 536—544 (in Russian).
13. Chertovskoy V. D. *Analiz processa soglasovannogo planirovaniya v ierarhicheskoy sisteme upravleniya proizvodstvom* [Automatized coordinated planning process analyses in hierarchical control system of manufacturing], *Informatsionnye tekhnologii*, 2010, no. 6 (166), pp. 68—72 (in Russian).
14. Chertovskoy V. D. *Matematicheskoe opisanie i komputernaya realizatsiya mjdeli adaptivnoy avtomatizirovannoy sistemy upravleniya proizvodstvom* [Mathematical description and computer realization of model for adaptive automatized manufacturing control system], *Informatsionno-upravljajucsie sistemy*, 2017, no. 1 (86), pp. 106—114 (in Russian).

УДК 004.773.5

И. В. Лобов, канд. физ.-мат. наук, ст. науч. сотр., e-mail: lobov@ihep.ru,
В. Г. Готман, мл. науч. сотр., e-mail: vladislav.gotman@ihep.ru,
НИИЦ "Курчатовский институт" ФГБУ ГНЦ РФ —
Институт физики высоких энергий, Московская обл., г. Протвино

Использование контейнера Ogg для организации потоковой трансляции в реальном времени над протоколом HTTP методом опережающей загрузки

Рассмотрены характеристики и общая схема организации потоковой трансляции в реальном времени над протоколом HTTP. Применительно к общей схеме дан обзор существующих лидирующих технологий в этой области. Рассмотрена технология использования контейнера Ogg для организации трансляции в реальном времени над протоколом HTTP методом опережающей загрузки. Проведено ее сравнение с существующими основными технологиями (Apple HLS, Adobe HDS, Microsoft Smooth Streaming, MPEG DASH).

Ключевые слова: трансляция, поток, опережающая загрузка, Ogg, Apple HLS, Adobe HDS, Microsoft Smooth Streaming, MPEG DASH

Введение

В последнее время в мире приобрели широкую популярность веб-трансляции презентаций в реальном времени, основой чему послужило бурное развитие технологий потоковых трансляций над протоколом HTTP. В настоящей работе делается сравнение предложенной в работе [1] идеи использования метода опережающей загрузки медиапотока формата Ogg [2] для организации живой трансляции с лидирующими технологиями живого потокового вещания. С этой целью выявлены общие свойства, присущие любой живой трансляции, а именно: разбиение трансляции на отдельные фрагменты, наличие у фрагмента метки времени и наличие метаданных для всей трансляции. Эти свойства определяют общую схему организации живой трансляции. Выполнен обзор существующих лидирующих технологий живого потокового вещания с точки зрения соответствия общей схеме. Показано, что технология трансляции медиапотока формата Ogg методом опережающей загрузки удовлетворяет общей схеме организации живой трансляции. Сделан вывод о логической эквивалентности всех рассмотренных технологий потоковых трансляций.

1. Общая схема организации трансляции в реальном времени

Под презентацией будем далее понимать действие, самостоятельно происходящее во времени, это может быть программа новостей, лекция, мастер-класс и т.п. Далее будет рассматриваться случай односторонних трансляций, когда клиент может только смотреть и слушать презентацию, но не участвовать в ней.

Задача трансляции презентации в реальном времени заключается в неискаженной передаче медиапотоков (синхронизованных между собой временных последовательностей видеок кадров и аудиосэмплов) от источника трансляции к конечному потребителю (клиенту) по каналу связи. Напрямую, в необработанном виде, эти потоки посылать нельзя ввиду большого объема информации в них и ограничений на пропускную способность канала связи. Поэтому используют различные алгоритмы сжатия медиапотоков. Соответственно, на стороне клиента после приема сжатых данных необходимо их декодировать и превратить в обычные видеок кадры и аудиосэмпы, которые уже можно воспроизводить.

На технологию трансляции презентации налагают различные требования, наиболее важные

из них: 1) возможность одновременного подключения к трансляции нескольких клиентов в произвольные моменты времени; 2) возможность продолжения воспроизведения трансляции после потери ее части при передаче по каналу связи; 3) синхронное воспроизведение аудио- и видеопотоков на стороне клиента. Из этих требований следует несколько важных выводов.

Во-первых, трансляция презентации не должна осуществляться непрерывным потоком октетов без дополнительной их организации. В противном случае клиент, подключающийся к середине трансляции, не сможет понять, где закончился один видеокادر (аудиосэмпл) и где начался другой. Это же относится и к случаю потери по линии связи части октетов. Если после потери части трансляции клиент может понять, где начинается первый кадр или сэмпл, то он сможет продолжить воспроизведение. Поэтому трансляцию удобно проводить порциями, имеющими четкие границы и снабженными (желательно) контрольной суммой. Далее будем называть эти порции *обобщенными фрагментами*.

Обобщенный фрагмент это передаваемая по каналу связи единица медиаконтейнера, содержащая часть закодированного медиапотока. Обобщенный фрагмент (далее просто *фрагмент*) может содержать как целое число медиаэлементов (видеокадров, аудиосэмплов), так и часть медиаэлемента. Фрагменты могут содержать один тип медиаэлементов (например, только видео) или несколько типов вместе (например, аудио и видео в одном фрагменте). Фрагменты являются независимыми друг от друга в том смысле, что фрагмент содержит всю необходимую информацию для извлечения из него закодированных медиаэлементов. Фрагмент можно передать клиенту за одну транзакцию (например, за один запрос GET [3, С. 35]), но это необязательно. Клиент

может получить фрагмент от сервера по частям, а затем "склеить" полученные части между собой.

Во-вторых, клиент должен знать, в какие моменты времени необходимо воспроизводить содержащиеся во фрагментах медиаэлементы. Для этого фрагмент снабжается *меткой времени*. При задержке или потере фрагментов метка времени предотвратит временной сдвиг и рассинхронизацию между различными медиапотоками (например, аудио и видео).

В-третьих, для возможности воспроизведения трансляции клиент должен получить *метаданные* медиапотока. Метаданные необходимы клиенту для первоначальной инициализации медиапотока. В метаданные входит информация о параметрах видео и аудио (размер кадра, частота кадров, частота дискретизации, применяемые кодеки и т.п.).

На рис. 1 изображена общая схема организации простейшей трансляции в реальном времени. От источника трансляции, например видеокамеры со встроенным микрофоном, на сервер поступают два потока медиаданных — видеокадры и аудиосэмплы. Каждый из этих потоков поступает со своей определенной частотой. Перед началом трансляции сервер создает метаданные трансляции и хранит их в отдельном буфере. По мере поступления медиаданных с источника трансляции, сервер их кодирует и помещает во фрагменты. Когда клиент подключается к трансляции, он вначале получает метаданные, затем начинает загружать фрагменты и воспроизводить их. На рис. 1 изображена гипотетическая ситуация, когда одному видеокадру всегда соответствуют два аудиосэмпла, а один фрагмент содержит три видеокадра и шесть аудиосэмплов.

2. Обзор важнейших технологий организации живого потокового вещания над протоколом HTTP

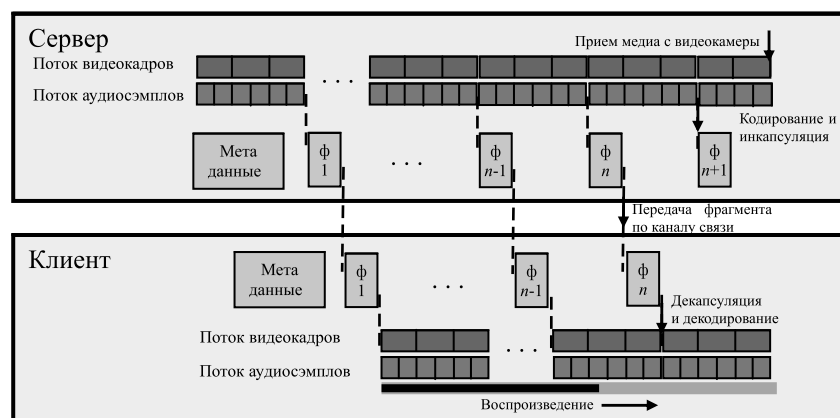


Рис. 1. Общая схема организации трансляции в реальном времени для случая, когда видео и аудио инкапсулируются в один фрагмент

На сегодняшний день лидирующими технологиями живого потокового вещания над протоколом HTTP являются *Apple HTTP Live Streaming* [4], *Adobe HTTP Dynamic Streaming* [5], *Microsoft Smooth Streaming* [6], *MPEG Dynamic Adaptive Streaming over HTTP* [7]. Дадим общую характеристику перечисленных технологий и проведем краткий обзор их существенных особенностей применительно к приведенной выше общей схеме. Адаптивное вещание в рамках настоящей статьи не обсуждается.

Общая характеристика обозреваемых технологий

Каждая из обсуждаемых далее технологий живого потокового вещания имеет свое собственное понятие минимальной единицы потока, передаваемой по каналу связи. Все эти определения соответствуют данному в разд. 1 понятию обобщенного фрагмента, дополненного требованием независимой загрузки. Приведем краткое описание процесса трансляции по этим технологиям. По мере прохождения трансляции сервер создает фрагменты. Клиент подключается к серверу и делает постоянные запросы для загрузки метаданных и фрагментов, используя HTTP-метод GET. С точки зрения клиента, каждый загружаемый фрагмент является файлом, хотя фрагменты необязательно должны физически находиться на сервере в виде отдельных файлов. Фрагмент содержит целое число видеок кадров и/или аудиосэмплов, закодированных видео- и аудиокодеками и инкапсулированных в контейнер.

Некоторые из этих технологий (*Apple HLS*, *Adobe HDS*) накладывают ограничение на обобщенный фрагмент — требование независимого декодирования фрагментов друг от друга.

В обсуждаемых технологиях роль метаданных играет файл, называемый манифестом. Сервер помещает в манифест общую информацию о свойствах воспроизводимого потока (применяемых кодеках, разрешении экрана и др.) и информацию о фрагментах (URL и метки времени). С каждым новым фрагментом сервер обновляет манифест.

Отличительной особенностью обозреваемых технологий является наличие у каждого фрагмента своего уникального URL, что делает фрагменты независимо загружаемыми. Манифест необходим клиенту не только для возможности декодирования принимаемого потока фрагментов, но также для получения URL отдельных фрагментов (посредством шаблонов или прямых ссылок). Поэтому для обновления списка доступных фрагментов клиенту необходимо периодически обновлять манифест. Отметим, что для технологии *Microsoft Smooth Streaming* манифест запрашивается клиентом только один раз — при подключении клиента к трансляции.

Схема взаимодействия клиента и сервера представлена на рис. 2.

Как видно из рис. 2, инициатором приема фрагментов у всех перечисленных технологий является клиент. Сервер при этом выполняет пассивную роль отсылки данных в ответ на запросы клиента.

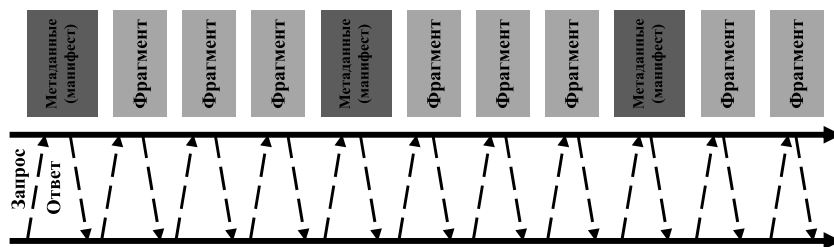


Рис. 2. Схема взаимодействия клиента и сервера, применяемая в существующих лидирующих технологиях организации живого потокового вещания над протоколом HTTP

Apple HTTP Live Streaming

Технология *HTTP Live Streaming (HLS)* разработана компанией *Apple* в 2009 г. и является одной из самых ранних и простых технологий ведения потоковых трансляций над *HTTP*. *HLS* входит составной частью в собственное программное обеспечение *Apple*, такое как *QuickTime*, *Safari*, *OS X* и *iOS*. *Apple HLS* может использовать в качестве клиента любые современные браузеры с поддержкой *HTML5* [8] и *Media Source Extensions*. Это дает возможность системам введения трансляций по технологии *HLS* поддерживать максимально широкое число клиентских устройств.

Во время ведения трансляции по технологии *HLS* презентация записывается в непрерывную цепочку файлов. В технологии *Apple HLS* данное в разд. 1 определение обобщенного фрагмента соответствует файлу формата *MPEG-2 TS* [9] с расширением *.ts*, называемым *медиафрагментом* [4, С. 5]. Видео и аудио содержатся в одном фрагменте. Для сжатия чаще всего используют кодеки — *H.264* [10] (видео) и *AAC* (аудио) [11].

Метаданные потока содержатся в манифесте — текстовом файле формата *m3u* или *m3u8*. Манифест содержит следующее:

- служебную информацию (EXT-X-STREAM-INF) о кодеках (CODECS), разрешении экрана (RESOLUTION), битрейте (BANDWIDTH) и т.д.;
- ссылки на фрагменты, доступные для скачивания клиентом с *web*-сервера, представляющие собой имена реальных файлов;
- метки времени (EXTINF), представляющие собой длительности фрагментов.

По мере прохождения презентации сервер кодирует новые видеок кадры и аудиосэмплы во фрагменты. С каждым новым готовым фрагментом сервер формирует новый манифест. Сервер может ограничивать число фрагментов в манифесте, удаляя из списка самые старые фрагменты, и поэтому общее число фрагментов будет постоянным. Наименьшее число фрагментов равно 3 [4, С. 32]. Пример файла-манифеста *Apple HLS* для живой трансляции:

```
#EXTM3U
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=
150000,RESOLUTION=416x234,
CODECS = "avc1.42e00a,mp4a.40.2"
#EXT-X-TARGETDURATION:10
#EXT-X-VERSION:3
#EXT-X-MEDIA-SEQUENCE:1
#EXTINF:10,
fileSequence1.ts
#EXTINF:10,
fileSequence2.ts
#EXTINF:10,
fileSequence3.ts
```

Здесь заданы три фрагмента, длительностью 10 с каждый.

После подключения к трансляции клиент вначале загружает манифест и тем самым получает список ссылок на доступные в данный момент фрагменты. Затем клиент начинает загружать сами фрагменты. После завершения загрузки всех фрагментов из полученного списка клиент запрашивает новый манифест, получает новый список фрагментов и опять начинает их загружать. Этот процесс продолжается до конца трансляции. Клиент воспроизводит загруженные фрагменты непрерывно, без каких-либо задержек между ними.

Microsoft Smooth Streaming

Smooth Streaming (SS) является технологией ведения потоковых трансляций, разработанной компанией Microsoft. Впервые технология была анонсирована и применена для трансляции готового статического видео на Олимпийских играх в 2008 г. В 2010 г. была опубликована первая официальная версия протокола. Технология *Smooth Streaming* использует проигрыватель Microsoft Silverlight в качестве основного клиента.

Во время ведения трансляций по технологии Microsoft SS вся презентация записывается на сервере в один файл формата PIFF [12], который представляет собой расширение стандарта MPEG-4 Part 12 [13]. Формат файлов этого стандарта называют также фрагментированным MP4 (fMP4). Файлы формата fMP4 состоят из атомов — объектно-ориентированных строительных блоков, каждый из которых снабжен уникальным идентификатором типа этого блока [13, С. 2]. Файлы презентации содержат либо только видео .ismv, либо только аудио .isma, либо все вместе .ismv. Основными кодеками *Microsoft SS* являются H.264 и AAC.

Данному во введении определению обобщенного фрагмента в технологии *Microsoft SS* соответствует независимо загружаемая единица медиа, содержащая целое число видеок кадров или

аудиосэмплов, называемая в протоколе *Smooth Streaming* фрагментом [6]. Фрагмент состоит из двух атомов: moof с метаданными фрагмента и mdat с медиаданными. Физически фрагмент — это часть файла презентации. По мере прохождения презентации видеок кадры и аудиосэмпы сервер кодирует во фрагменты и добавляет к файлу презентации. Различные потоки, например аудио или видео, в *Microsoft SS* именуют треками. Фрагмент содержит только один трек.

Временная метка фрагмента в *Microsoft SS* есть абсолютное время первого видеок кадра или аудиосэмпа во фрагменте трека. Отличительной особенностью фрагмента *Microsoft SS* является то, что в нем содержатся две (или более) временные метки. Одна временная метка задает время самого фрагмента в треке. Она содержится в параметре "FragmentAbsoluteTime" атома Tfxd/moof [6]. Другая задает время начала следующего (или нескольких последующих) фрагмента трека. Она содержится в параметре "FragmentAbsoluteTime" атома Tfrf/moof фрагмента [6].

Метаданные потока содержатся в файле-манифесте клиента .ismc, имеющего формат XML [14]. Помимо метаданных, манифест *Microsoft SS* содержит шаблон построения ссылок для загрузки клиентом фрагментов с сервера. В общем виде шаблон имеет следующий вид:

```
http://{ service uri}/QualityLevels({bitrate})/Fragments(video=
{start time}), (1)
```

где service uri — базовый URL сервера; bitrate — битрейт потока; start time — временная метка фрагмента. Также в манифесте клиента содержатся временные метки и длительности нескольких самых свежих фрагментов.

При старте трансляции сервер создает несколько первоначальных фрагментов с медиаданными и манифест клиента, содержащий информацию об этих фрагментах. По мере поступления новых медиаданных сервер создает новые фрагменты и обновляет манифест клиента. Подключившемуся клиенту сервер передает сначала манифест, а потом запрашиваемые клиентом фрагменты. Для связи фрагментов с файлом презентации сервер использует специальный файл (.ism), называемый серверным манифестом.

При подключении к трансляции клиент скачивает манифест клиента, затем начинает загружать описанные в манифесте фрагменты, формируя ссылки на фрагменты по шаблону (1) и делая запросы серверу. После того, как клиент скачает все указанные в манифесте фрагменты, ему необходимо сформировать ссылку на следующий фрагмент, для этого он должен узнать метку времени следующего фрагмента. Как было сказано

выше, это есть параметр "FragmentAbsoluteTime" атома Tfrf последнего загруженного клиентом фрагмента. После получения каждого очередного фрагмента клиент определяет таким способом метки времени следующих фрагментов этого трека, формирует запросы по шаблону (1) и получает новые фрагменты. Таким образом, клиент загружает файл-манифест только один раз, все остальное время он поддерживает свой внутренний манифест.

Adobe HTTP Dynamic Streaming

HTTP Dynamic Streaming (HDS) является технологией ведения потоковых трансляций, разработанной компанией *Adobe*. Финальная версия 3.0 официального стандарта опубликована в 2013 г., хотя существуют официальные упоминания об HDS с 2010 года. Для просмотра презентации *Adobe HDS* использует собственный проигрыватель *Flash Player*.

Спецификация *Adobe HDS* [5, С. 4] определяет фрагмент, как наименьшую адресуемую единицу медиа, позволяющую автономное декодирование. Фрагмент является базовой нумерованной единицей медиа, передаваемой по каналу связи. Это определение соответствует данному в разд. 1 определению *обобщенного фрагмента*. Основными кодеками являются H.264 и AAC. В одном фрагменте могут содержаться аудио и/или видеоданные [5, С. 28]. Смежные фрагменты объединяются в сегменты, которые хранятся на диске в виде файлов с расширением .f4v. Сегменты используются для повышения эффективности хранения и загрузки фрагментов. На сервере вся презентация выглядит как набор одного или нескольких сегментов.

Фрагменты и сегменты в *Adobe HDS* имеют формат F4V, представляющий собой расширение стандарта MPEG-4 Part 12. Для каждого сегмента на HTTP-сервере присутствует индексный файл .f4x, в котором содержится байтовое смещение каждого фрагмента в сегменте. Индексный файл представляет собой атом произвольного доступа (*random access*) afra. Индексный файл необходим серверу для поиска требуемого фрагмента в сегменте при выполнении запросов клиента. По ходу ведения презентации индексная информация постоянно обновляется.

Метаданные потока содержатся в паре манифестов — set-level и stream-level — файлов формата XML с расширением .f4m.

Манифест set-level необходим клиенту для информирования о существующих видеопотоках, а также для адаптивного воспроизведения. Манифест set-level содержит ссылки на манифесты stream-level для каждого описанного в нем потока. Манифест stream-level содержит:

- метаданные потока — тип кодека, разрешение экрана и т.п.;
- информацию bootstrap, включающую в себя: номера сформированных фрагментов, их метки времени и номера сегментов, в которых инкапсулированы эти фрагменты. Информация bootstrap представляет собой три атома: abst, asrt и afrt, сформированных на сервере и закодированных в формат Base64.

Ниже приведены примеры файлов-манифестов.

1. Манифест set-level:

```
<?xml version = "1.0" encoding = "utf-8"?>
<manifest xmlns = "http://ns.adobe.com/f4m/1.0" version = "3.0">
<id>myVideo</id>
<baseURL>http://www.example.com/myvideo/</baseURL>
<streamType>live</streamType>
<media href = "stream250.f4m" bitrate = "250" />
<media href = "stream500.f4m" bitrate = "500" />
</manifest>
```

В нем указано:

- 1) базовый URL сервера <http://www.example.com/myvideo/>;
- 2) параметр трансляции live, трансляция происходит в реальном времени;
- 3) ссылки на два манифеста stream-level с битрейтами 250 и 500.

2. Манифест stream-level для потока с битрейтом 250:

```
<?xml version = "1.0" encoding = "utf-8"?>
<manifest xmlns="http://ns.adobe.com/f4m/1.0" version= "3.0">
<id>myStream1</id>
<baseURL>http://www.example.com/data/</baseURL>
<bootstrapInfo profile="named" id="boot1" fragmentDuration = "4">
(BASE64 encoding of bootstrap information)
</bootstrapInfo>
<media url = "stream250" bootstrapInfoId = "boot1"/>
</manifest>
```

В нем присутствуют:

- 1) базовый URL сервера <http://www.example.com/data/>;
- 2) идентификатор потока myStream1 (Movie-Identifier);
- 3) медиа URL потока stream250 (Quality-SegmentUrlModifier);
- 4) информация bootstrap.

По мере прохождения трансляции сервер создает фрагменты с медиаданными и добавляет их к сегменту. Число фрагментов в сегменте определяется параметрами длительностей фрагмента и сегмента, которые задаются в настройках сервера. Сервер создает информацию bootstrap, которая связывает между собой: а) время отображения видеок кадров/аудиосэмплов во фрагменте; б) номера

фрагментов; в) номера сегментов. Вместе с метаданными потоков информация bootstrap сервер помещает в манифест stream-level. При появлении новых фрагментов манифест stream-level и информация bootstrap в нем постоянно обновляются.

Для получения информации обо всех потоках клиент сначала единожды загружает с сервера set-level манифест, после чего запрашивает нужный stream-level манифест. Затем клиент расшифровывает информацию bootstrap и начинает запрашивать фрагменты выбранного потока в соответствии со следующим шаблоном:

```
http://<ServerBaseUrl>/<MovieIdentifier><QualitySegmentUrl  
Modifier>Seg<SegmentNumber>-Frag<FragmentNumber> (2)
```

Например, для того, чтобы получить фрагмент 210 сегмента 1, клиент должен сделать следующий запрос:

```
http://adobe.com/MyMovie/highSeg1-Frag210
```

HTTP-сервер, используя индексный файл .f4x, транслирует запрос в байтовую сдвигку запрашиваемого фрагмента (в примере это фрагмент 210) в сегменте (в примере это сегмент 1) и отправляет клиенту требуемый фрагмент. После загрузки всех доступных фрагментов, описанных в файле-манифесте, клиент загружает новый манифест stream-level, с bootstrap информацией о новых доступных фрагментах.

По умолчанию длительность фрагмента в технологии Adobe HDS составляет 4 с [15, С. 29], а рекомендованная длина буфера клиента — три фрагмента (утроенная длительность одного фрагмента) [5, С. 14]. С этими параметрами минимальная задержка живой трансляции будет отстоять от настоящего времени на 12 с.

Dynamic Adaptive Streaming over HTTP (MPEG-DASH)

MPEG-DASH является не готовой технологией, а стандартом [7], описывающим общие принципы организации ведения поточных трансляций над HTTP. Работа над созданием стандарта MPEG-DASH была начата в 2010 г., окончательный стандарт ISO/IEC 23009-1:2012 был опубликован в 2012 г. Существуют реализации стандарта MPEG-DASH в виде технологий, например *Azure Media Services* [16].

Данному в разд. 1 определению обобщенного фрагмента в терминологии MPEG-DASH соответствует сегмент — ассоциированная с URL единица медиаданных [7, С. 4].

Стандарт MPEG-DASH описывает два возможных формата фрагментов и вместе с тем по-

зволяет расширять набор форматов сторонними разработчиками. Форматы предложенных фрагментов MPEG-DASH:

1) MPEG-2 TS (базируется на стандарте ISO/IEC 13818-1 [17]).

2) fMP4 (базируется на стандарте MPEG-4 Part 12).

Фрагменты содержат закодированные видеокдры и/или аудиосэмпы. Фрагменты могут содержать как видео и аудио вместе, так и только аудио или только видео. Хотя стандарт MPEG-DASH не предписывает использование определенных кодеков, наиболее распространенными кодеками в существующих реализациях являются H.264 и AAC.

Метаданные потока содержатся в файле-манифесте формата XML с расширением .mpd. Манифест содержит как общую информацию о потоках — битрейты дорожек, типы дорожек и т.д., — так и ссылки на фрагменты, временные метки и длительности сегментов.

Ссылки бывают трех видов:

1. Прямая ссылка на одиночный медиафайл презентации в файле-манифесте, например:

```
<BaseURL>7657412348.mp4</BaseURL>
```

2. Список доступных фрагментов и их длительности, например:

```
<SegmentList duration = "10">  
<SegmentURL media = "seg-m1-C2view-1.mp4"/>  
<SegmentURL media = "seg-m1-C2view-2.mp4"/>  
<SegmentURL media = "seg-m1-C2view-3.mp4"/>  
</SegmentList>
```

3. План производства фрагментов, представляющий собой шаблон для построения ссылок, например:

```
http://cdn1.example.com/video/500000/$Time$.mp4v,  
t = "0" d = "96000" r = "432",
```

где \$Time\$ время начала будущего файла фрагмента; t — время начала; d — длительность одного фрагмента; r — число фрагментов. Пример использования шаблона (время начала первого фрагмента ноль):

```
http://cdn1.example.com/video/500000/0.mp4v  
http://cdn1.example.com/video/500000/180180.mp4v  
http://cdn1.example.com/video/500000/360360.mp4v
```

Временная метка фрагмента есть время начала его воспроизведения относительно начала презентации. Она содержится в файле-манифесте.

Для воспроизведения трансляции клиент подключается к серверу, делая запрос на манифест. Манифест дает возможность клиенту получить или сгенерировать список ссылок на фрагменты.

Манифест изменяется в случае появления новых фрагментов или изменения местоположения (ссылок) прежних фрагментов. В параметре файла-манифеста "minimumUpdatePeriod" указан минимальный период потенциальных изменений манифеста. С помощью этого параметра можно оптимизировать процесс периодического обновления манифеста клиентом.

В качестве примера реализации стандарта MPEG-DASH рассмотрим технологию *Azure Media Service*. Для просмотра презентации клиент подключается к серверу *Azure* и загружает манифест этой трансляции. *Azure* использует адресацию фрагментов в реальном времени и задает время фрагментов в манифесте. Поэтому клиент всегда знает время последнего доступного фрагмента и может подключиться к трансляции при минимальной задержке трансляции от реального времени. *Azure* минимизирует эту задержку путем уменьшения длительности одного фрагмента до 2 с. Типичная задержка трансляции клиентом составляет от нескольких до 30 с [18].

Особенностью *Azure* являются то, что параметр манифеста "minimumUpdatePeriod" равен нулю, поэтому клиент без дополнительной информации никогда сам не обновляет манифест. Фрагменты могут содержать атом *emsg* с параметром "publishTime" указывающим время, в которое манифест должен быть обновлен.

Таким образом, клиент сначала один раз запрашивает манифест, а потом запрашивает только фрагменты. Если на сервере добавились новые фрагменты или изменились ссылки на старые фрагменты, сервер заранее сообщает клиенту во фрагменте, когда ему необходимо обновить манифест. Этим достигается оптимизация Интернет-трафика. Клиент загружает манифест только по необходимости.

3. Технология организации живой потоковой трансляции над протоколом HTTP методом опережающей загрузки медиапотока формата Ogg

Суть изложенной в п. 1 общей схемы организации трансляции медиа в реальном времени (см. рис. 1) заключается в том, что медиапоток сжимается (кодируется), инкапсулируется в небольшие фрагменты и пересылается клиенту. Клиент извлекает данные из фрагмента, декодирует кадры или сэмплы, склеивает их в непрерывный медиапоток и воспроизводит его. Во всех приведенных в п. 2 технологиях передача

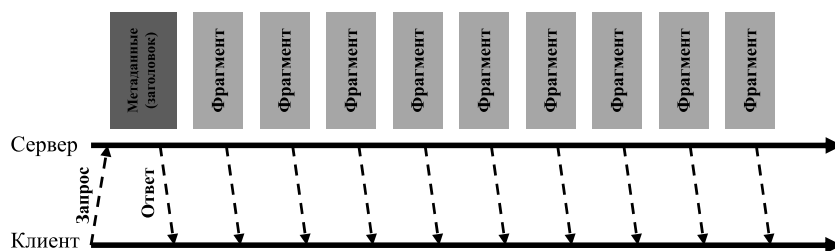


Рис. 3. Схема взаимодействия клиента и сервера, применяемая в технологии организации живой потоковой трансляции методом опережающей загрузки медиапотока поверх протокола HTTP

медиа от сервера к клиенту проводится только по инициативе последнего. Однако ничто не мешает использовать общую схему в случае, когда клиенту принадлежит только инициатива начала передачи медиа, после чего сервер начинает самостоятельную пересылку фрагментов клиенту. А это есть не что иное, как опережающая загрузка медиапотока с одновременным его воспроизведением (рис. 3).

Метод опережающей загрузки — это программная реализация возможности начала воспроизведения файла без необходимости ожидания завершения его полного скачивания. В рассматриваемом случае в роли загружаемого файла выступает презентация, т.е. медиапоток. Клиент начинает загрузку и воспроизведение презентации путем обращения к тегу `<video>`, в котором указывается URL презентации и атрибуты тега, которые передаются воспроизводящему презентацию проигрывателю браузера. При этом браузер клиента посылает однократно на сервер стандартный HTTP-запрос GET о подключении к трансляции. Получив запрос от клиента, сервер посылает метаданные, а затем начинает отсылать ему метаданные презентации. Браузер клиента, в свою очередь, начинает фоновую загрузку презентации. После накопления в буфере браузера некоторого достаточного объема метаданных презентации, начинается ее воспроизведение. Сравнивая рис. 2 и рис. 3, заметим, что метаданные отсылаются клиенту только однократно, а необходимости в периодической посылке манифестов нет, так как клиент в рассматриваемой схеме пассивен и ему не требуется знать URL загружаемых фрагментов.

Пересылаемые клиенту данные должны иметь фрагментированную структуру. После подключения клиента к серверу в середине презентации сервер попросту начнет транслировать медиа с очередного "свежего" фрагмента. Поскольку фрагмент снабжается меткой времени, воспроизведение клиентом презентации выполняется в "правильное" время, а нарушения синхронизации между видео и аудио при этом не будет. Для организации фрагментированного потока в настоя-

шей работе предлагается использование контейнера Ogg, который изначально разрабатывался как потоковый медиаконтейнер. Потоком в понимании Ogg является последовательность *пакетов*, которой присвоен уникальный идентификатор.

Пакет представляет собой данные, закодированные определенным кодеком. Например, для видеопотока пакетом может быть кадр, закодированный кодеком Theora [19]. Для аудиопотока пакетом может быть некоторое число аудиосэмплов, закодированных кодеком Vorbis [20]. Размер пакета Ogg может быть любым, в том числе и нулевым.

Медиафайл формата Ogg может содержать один или несколько потоков, например два потока — видео и аудио. Каждый поток в контейнере Ogg содержится в виде упорядоченной последовательности объектов одинаковой структуры — *страниц*. Страница Ogg имеет заголовок и тело, содержащее собственно данные. Заголовок страницы содержит:

- идентификатор медиапотока, которому принадлежит страница;
- порядковый номер страницы в потоке;
- тип страницы (страница является началом потока, страница является концом потока, страница содержит незавершенный пакет);
- маркер позиции в своем потоке последнего завершенного пакета в странице (*granule position*).

Естественно принять содержащую медиаданные страницу Ogg за *обобщенный фрагмент* потоковой трансляции. Каждый фрагмент содержит только один тип медиаэлементов (например только аудио или только видео). Отличительной особенностью страниц Ogg является то, что максимальный их размер ограничен значением 65,307 байт [2, С. 8], в то время как размер видеокадра (и соответственно пакет, в который он

закодирован) может быть больше этого значения. В этом случае один кадр будет содержаться в нескольких страницах Ogg, которые должны иметь один и тот же маркер позиции в своем потоке.

Маркер позиции сам по себе не является меткой времени. Но существует простой способ вычисления метки времени каждого кадра или сэмпла по известным маркеру позиции страницы и числу кадров (сэмплов) в секунду в потоке, схема вычисления зависит от кодека потока. Для кодека Theora маркер позиции является бинарной конструкцией из двух чисел-счетчиков ключевых (I) и разностных (P) кадров с начала трансляции медиапотока. Метка времени подсчитывается как число кадров от начала трансляции, деленное на число кадров в секунду. А в случае кодека Vorbis это просто число аудиосэмплов от начала трансляции, деленное на частоту дискретизации.

Метаданными потока являются пакеты, содержащие заголовки потока и инкапсулированные в несколько первых страниц Ogg. Они содержат общие данные о потоке, например разрешение экрана, число кадров (сэмплов) в секунду и т.д.

Для воспроизведения потоковой трансляции клиент подключается к серверу и делает однократный запрос, используя HTTP-метод GET, обращаясь ко всей презентации, как к файлу. В ответ клиент вначале получает страницы заголовка потока, а затем страницы с медиаданными, начиная с последней готовой страницы, сделанной сервером к моменту подключения клиента.

На рис. 4 изображена общая схема трансляции потока пакетов Ogg в реальном времени. В самом начале трансляции сервер создает первые три пакета (З), содержащие заголовки потока, и инкапсулирует их в страницы С1 и С2, которые хранит в памяти. При получении медиаданных от источника трансляции сервер кодирует их в пакеты (Д) и помещает в новые страницы, начиная с С3. В одной странице могут быть несколько пакетов, один пакет или часть пакета. На рис. 4 изображен случай, когда клиент подключается к серверу в момент, когда тот создал страницу $n - 1$. После подключения клиенту сначала передаются страницы С1 и С2 с заголовками потока, а затем страницы с медиаданными, начиная с номера $n - 1$.

Заключение

В настоящей работе предлагается использование контейнера Ogg для организации потоковой трансляции в реальном времени над протоколом

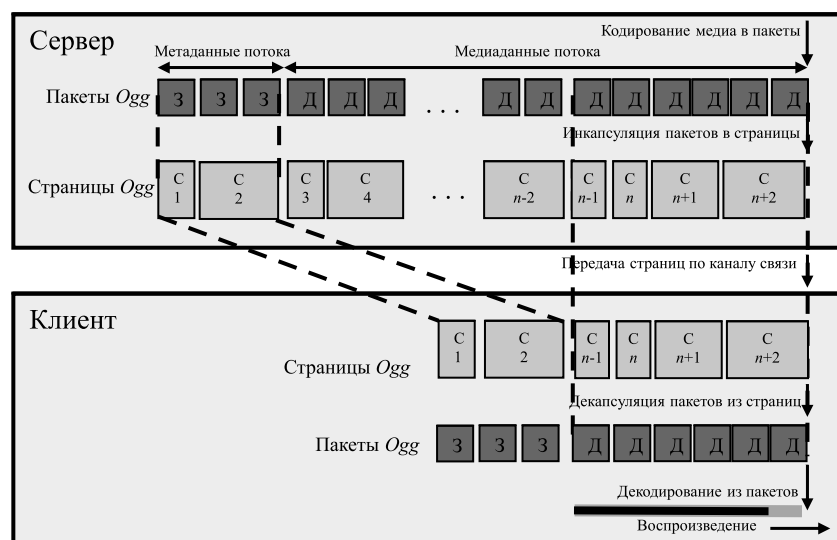


Рис. 4. Общая схема живой трансляции потока пакетов Ogg

HTTP методом опережающей загрузки. Несомненным плюсом использования формата медиаcontainers Ogg, кодеков Theora и Vorbis является то, что они обладают открытой лицензией и напрямую поддерживаются стандартом HTML5. Медиапоток, закодированный кодеками Theora и Vorbis и помещенный в медиаcontainer Ogg, доступен для воспроизведения во многих браузерах. Это позволяет клиенту отказаться от использования сторонних плагинов и расширений к браузерам.

В работе проведено сравнение предложенной в работе [1] технологии организации живой потоковой трансляции над протоколом HTTP методом опережающей загрузки медиапотока формата Ogg с лидирующими технологиями потоковых трансляций [4–7]. Для этой цели предложена общая схема, отражающая следующие важнейшие особенности потоковых трансляций:

- передача потока от сервера клиенту осуществляется фрагментами;
- трансляция имеет метаданные, предшествующие медиаданным потока;
- каждый фрагмент снабжен меткой времени.

Выявлено, что *Apple HLS*, *Adobe HDS*, *Microsoft Smooth Streaming*, *MPEG DASH* являются реализациями общей схемы и имеют следующие особенности:

- обобщенным фрагментом является имеющая URL единица медиа; инициатором начала трансляции и загрузки фрагментов является клиент, для чего ему необходимо предварительно получить от сервера манифест и затем периодически обновлять его;
- метаданные содержатся в файле-манифесте;
- метка времени фрагмента передается в манифесте.

Показано, что технология организации живой потоковой трансляции над протоколом HTTP методом опережающей загрузки медиапотока формата Ogg также удовлетворяет общей схеме и обладает следующими особенностями:

- фрагментом является страница Ogg с медиаданными; клиент пассивно принимает фрагменты от сервера;
- метаданные содержатся в заголовочных страницах Ogg и передаются клиенту только один раз после инициирования клиентом начала трансляции медиапотока;
- метка времени — это маркер позиции фрагмента (*granule position*).

Тот факт, что оба рассмотренных подхода ([4–7] и [1]) являются реализациями общей схемы, позволяет сделать вывод об их логической эквивалентности.

Заметим, что при использовании технологий потоковых трансляций клиент для плавного воспроизведения презентации должен иметь

в буфере несколько фрагментов. Поскольку для технологий потоковой трансляции [4–7] размер фрагмента, как правило, составляет несколько секунд, суммарная задержка отображения презентации от реального времени в таком подходе будет весьма значительной, например, для *Azure* она может быть до 30 с. В предложенной в работе [1] технологии потоковой трансляции размер видеофрагмента составляет один кадр. С учетом 25 кадров в секунду и 4–5 фрагментов в буфере ожидаемая задержка составляет несколько сотен миллисекунд.

На базе технологии, описанной в работе [1], был разработан медиасервер [21] для организации потоковых трансляций между технологическими подсистемами Ускорительного комплекса У-70. В качестве источников трансляции были использованы IP-камеры D-Link DCS-3430. Проведенные эксперименты по выявлению задержек воспроизведения презентаций от реального времени дали результат 250...500 мс. Это дает возможность использования предложенной в работе [1] технологии не только для односторонних трансляций, но и для организации общения клиентов в реальном времени между собой.

Список литературы

1. Лобов И. В., Готман В. Г. Трансляция мультимедиа в реальном времени над протоколом HTTP методом опережающей загрузки // Технологии и средства связи. 2016. № 5. С. 36–40.
2. Pfeiffer S. The Ogg Encapsulation Format Version 0. Request for Comments: 3533. 2003. 15 p.
3. Fielding R., Gettys J. Hypertext Transfer Protocol — HTTP/1.1. Request for Comments: 2616. 1999. 114 p.
4. Pantos R., May W. HTTP Live Streaming. draft-pantos-http-live-streaming-18. Apple Inc. 2015. 49 p.
5. HTTP Dynamic Streaming Specification Version 3.0 FINAL. Adobe Systems Incorporated. 2013. 31 p.
6. Smooth Streaming Protocol. [MS-SSTR] — v20160714. Microsoft Corporation. 2016. 64 p.
7. ISO/IEC 23009-1. Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats. Second edition. 2014. 144 p.
8. HTML Living Standard. The Web Hypertext Application Technology Working Group. 2017. 1177 p.
9. ITU-T H.222.0 (02/00). Information technology — Generic coding of moving pictures and associated audio information: systems. 2000. 153 p.
10. Wiegand T., Sullivan G., Luthra A. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC). 2002. 253 p.
11. ISO/IEC 13818-7:1997. Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC). 1997.
12. Bocharov J., Burns Q., Foltz F., Hughes K., Murching A., Olson L., Schnell P., Simmons J. Portable encoding of audio-video objects The Protected Interoperable File Format (PIFF). Microsoft Corporation. 2010. 28 p.

13. **ISO/IEC 14496-12:2015.** Information technology — Coding of audio-visual objects — Part 12: ISO base media file format. Edition: 5. 2015. 233 p.

14. **Extensible Markup Language (XML) 1.0.** W3C. Edition: 5. 2008. 32 p.

15. **Adobe HTTP Dynamic Streaming, User Guide.** Concentric Cloud Solutions An Xo Company. 2012. 37 p.

16. **Azure Media Services overview and common scenarios.** Microsoft Corporation 2017. 986 p.

17. **ISO/IEC 13818-1.** Information technology — Generic coding of moving pictures and associated audio information: Systems. Edition: 2. 2000. 154 p.

18. **Hughes K.** DASH Live Streaming with Azure Media Service. Azure Media Services. 2014.

19. **Theora Specification.** Xiph.Org Foundation. 2011. 196 p.

20. **Vorbis I specification.** Xiph.Org Foundation. 2015. 74 p.

21. **Lobov I., Gotman V.** Media server for video and audio exchange between the U-70 accelerator complex control rooms // Proceedings of RuPAC2014. RUPAC. Obninsk. 2014. P. 368—388.

I. V. Lobov, Senior Researcher, e-mail: lobov@ihep.ru

V. G. Gotman, Junior Researcher, e-mail: vladislav.gotman@ihep.ru

Institute for High Energy Physics, National Research Center "Kurchatov Institute", MO, Protvino

The Utilization of Ogg Multimedia Container Format for Live Streaming over HTTP Using Progressive Download Method

The general scheme of the media live streaming over HTTP has been proposed. It was shown that the general scheme has three common basic features: generalized fragments, metadata and fragment timestamp. The generalized fragment is the media container unit being transmitted over network. The fragment should contain encoded part of the media stream. The metadata stores common information (codec type, frame rate, image dimensions, bit rate, sample rate) which necessary for the client to decode and play the media stream back. The timestamp is the fragment property necessary to correct playback and synchronize for several media tracks — particularly for video and audio streams. The analysis of existing basic technologies (Apple HLS, Adobe HDS, Microsoft Smooth Streaming, MPEG DASH) of live streaming over HTTP was made. It was revealed that all these technologies meet this scheme. All of them have their own concept for the small unit of the coded media stream which matches to the proposed generalized fragment with one significant addition — the fragment has independent download possibility. The metadata and timestamps are stored into the manifest file. The technology of the media container Ogg utilization for live streaming over HTTP by progressive download method was proposed. This technology is quite different from the existing basics technologies. Yet it meets the general scheme. The comparison of the proposed technology with existing basic technologies has been made.

Keywords: live streaming, progressive download, Ogg format, Apple HLS, Adobe HDS, Microsoft Smooth Streaming, MPEG DASH

References

1. **Lobov I. V., Gotman V. G.** Translyatsiya multimedia v realnom vremeni nad protokolom HTTP metodom operezhayushey zagruzki (Real time media streaming over HTTP using progressive download method). *Tehnologii i sredstva svyazi*. 2016. no 5. pp. 36—40.

2. **Pfeiffer S.** The Ogg Encapsulation Format Version 0. Request for Comments: 3533. 2003. 15 p.

3. **Fielding R., Gettys J.** Hypertext Transfer Protocol — HTTP/1.1. Request for Comments: 2616. 1999. 114 p.

4. **Pantos R., May W.** HTTP Live Streaming. draft-pantos-http-live-streaming-18. Apple Inc. 2015. 49 p.

5. **HTTP Dynamic Streaming Specification Version 3.0 FINAL.** Adobe Systems Incorporated. 2013. 31 p.

6. **Smooth Streaming Protocol.** [MS-SSTR] — v20160714. Microsoft Corporation. 2016. 64 p.

7. **ISO/IEC 23009-1.** Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats. Second edition. 2014. 144 p.

8. **HTML Living Standard.** The Web Hypertext Application Technology Working Group. 2017. 1177 p.

9. **ITU-T H.222.0 (02/00).** Information technology — Generic coding of moving pictures and associated audio information: systems. 2000. 153 p.

10. **Wiegand T., Sullivan G., Luthra A.** Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC). 2002. 253 p.

11. **ISO/IEC 13818-7: 1997.** Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC). 1997.

12. **Bocharov J., Burns Q., Folta F., Hughes K., Murching A., Olson L., Schnell P., Simmons J.** Portable encoding of audio-video objects The Protected Interoperable File Format (PIFF). Microsoft Corporation. 2010. 28 p.

13. **ISO/IEC 14496-12: 2015.** Information technology — Coding of audio-visual objects — Part 12: ISO base media file format. Edition: 5. 2015. 233 p.

14. **Extensible Markup Language (XML) 1.0.** W3C. Edition: 5. 2008. 32 p.

15. **Adobe HTTP Dynamic Streaming, User Guide.** Concentric Cloud Solutions An Xo Company. 2012. 37 p.

16. **Azure Media Services overview and common scenarios.** Microsoft Corporation 2017. 986 p.

17. **ISO/IEC 13818-1.** Information technology — Generic coding of moving pictures and associated audio information: Systems. Edition: 2. 2000. 154 p.

18. **Hughes K.** DASH Live Streaming with Azure Media Service. Azure Media Services. 2014.

19. **Theora Specification.** Xiph.Org Foundation. 2011. 196 p.

20. **Vorbis I specification.** Xiph.Org Foundation. 2015. 74 p.

21. **Lobov I., Gotman V.** Media server for video and audio exchange between the U-70 accelerator complex control rooms. *Proceedings of RuPAC2014. RUPAC. Obninsk*, 2014, pp. 368—388.

УДК 33.27, 65.054

Л. А. Мыльников, канд. техн. наук, доц., e-mail: leonid.mylnikov@pstu.ru,
А. Б. Селедкова, магистрант, e-mail: aleksandrasedkova@yandex.ru,
Пермский национальный исследовательский политехнический университет

Способ выбора метода прогнозирования и горизонта планирования параметров с использованием оценки риска

Рассмотрены особенности прогнозирования параметров производственных систем, в том числе некоторая специфика, связанная с подготовкой данных, и разница в использовании методов, базирующихся на дополнительной информации о характере данных, и методов, не использующих такие сведения. Предложен способ выбора метода прогнозирования значений параметров с использованием интегральной оценки накопления величины оценки риска и выбора горизонта планирования на примере использования таких методов прогнозирования, как регрессионный анализ, авторегрессия, метод опорных векторов, вейвлет-анализ, прогнозирование с использованием фракталов.

Ключевые слова: геометрическое покрытие, раскрой, многосвязный ортогональный полигон, метод матричной декомпозиции, метод ограниченной декомпозиции

Введение

Согласно международному рейтингу производственной конкурентоспособности за 2016 г., использование предиктивных моделей в управлении оказывает наибольшее влияние на конкурентоспособность промышленного производства, что позволило США и КНР занять лидирующие позиции в течение последних пяти лет. Существенно возрастает роль использования прогнозов в задачах управления в связи с замещением методов статистической обработки данных когнитивными методами.

Использование прогнозов дает новый стимул к развитию методов математической формализации для решения задач управления материалами, частями (детальями), операциями, задач выбора поставщиков [1] и учета в них стохастических факторов, а также вероятностных подходов к оценке рисков с учетом разных характеров рассматриваемых событий (совместных, взаимозависимых, несовместных и взаимонезависимых) для решения задач планирования с учетом динамики рассматриваемых процессов.

При этом использование прогнозов приводит к появлению вероятностных моделей, основанных на использовании оценок рисков [2], теоремы Байеса [3] или метода Монте-Карло [4].

Такие подходы позволяют перейти от оценки рисков для отдельных случаев и отдельных задач

к рассмотрению проектов, процессов и производственных систем в целом.

В связи с этим выбор метода прогнозирования становится важной задачей, непосредственно влияющей на достоверность получаемых результатов.

Первоначально анализ временных рядов осуществлялся с использованием моделей, в которых влияние временного параметра проявлялось только в систематической составляющей [5] (в таких моделях предполагается, что течение времени никак не отражается на случайной составляющей).

Особенностью большинства параметров, характеризующих экономическую составляющую производственных систем, является возможность описания изменения значений многих параметров с использованием инновационных и S-образных кривых или периодическими кривыми. Однако в последнее время мы наблюдаем увеличение подвижности значений параметров, характеризующих производственные системы. Для прогнозирования в таких условиях необходимо использование методов, позволяющих быстро реагировать на изменение значений и самокорректироваться с учетом тенденций, которым подвержены изучаемые параметры. Таким образом, при рассмотрении отдельных параметров (не используя методы комплексного прогнозирования ряда параметров, таких как PLS-регрессия [6] и когнитивные карты [7]) задача прогнозиро-

вания будет сводиться к выбору и применению следующих методов:

- методы, используемые для прогнозирования данных, описываемых специальными кривыми (как правило, инновационными и S -образными ([6], [9]);
- методы, использующие обучение без учителя в широком классе задач обработки данных, в которых известны только описания множества объектов (обучающей выборки) и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами; основываются на применении метода кластеризации.
- методы, применяемые для анализа: метод k -средних, метод иерархической кластеризации, нейронные сети, метод авторегрессии [10];
- методы, использующие обучение с учителем. Данные методы основываются на использовании статистических данных для определения параметров методов, таких как метод опорных векторов [11], алгоритм k -ближайших соседей, метод смеси нормальных распределений, наивный байесовский классификатор, деревья решений, классификатор на радиальных базисных функциях и т.п.;
- полуконтролируемое обучение (группа методов машинного обучения, которые одновременно используют структурированные и дополнительные неструктурированные данные на основе таких подходов, как самообучение, совместное обучение, алгоритмы на основе теории графов, генеративные модели, методы кластеризации и разметки неструктурированных данных [12]).

При анализе данных о проектах, реализуемых в производственных системах, эти данные представляют собой, как правило, либо статистические выборки за небольшой период времени, либо обобщенные выборки данных, соответствующие группе товаров, модификаций изделия и т.п. (т.е. неразделенным данным). Данные могут характеризоваться разной частотой (разными интервалами времени) для разных параметров, могут содержаться в разных источниках и быть неполными или ошибочными.

Такая ситуация со статистическими данными может быть разрешена использованием либо дополнительной информации об анализируемых данных (например, закономерности, которой они описываются), либо методов увеличения размера статистических выборок.

Дополнительная информация выражается в сведениях о типе кривой, которой описываются данные или участки данных. Знания закономерностей изменения значений позволяют определить вид функции и свести решение задачи к по-

иску коэффициентов функций для наилучшего описания данных.

Для использования методов прогнозирования (без учета данных о виде функции) часто необходимо решать задачи, связанные с увеличением объема статистической выборки в связи с нерегулярностью и разными периодами собираемых данных.

В случае если данные имеют разную частоту выборки, то для увеличения объема данных можно использовать минимальный временной шаг собираемых данных. Для этого данные, которые имеют меньшую частоту выборки (число значений), можно развернуть на основании сезонных коэффициентов для каждой рассматриваемой модели. Сезонные колебания характеризуются специальными показателями, которые называются индексами сезонности (I_s).

Численное прогнозирование временных рядов

При использовании различных методов прогнозирования можно получить разные результаты. Методы прогнозирования используют ретроспективные данные. Для повышения точности получаемых результатов эти данные разбивают на две выборки: тренировочную (для построения модели прогнозирования) и тестовую (которая является последним во времени участком ретроспективных данных и служит для проверки модели). После принятия решения о применении модели тестовые данные добавляются к тренировочным и параметры модели уточняются для использования в реальных условиях.

Для выбора метода и его параметров часто используют экспертные оценки, методы проверки адекватности получаемых результатов и т.п. Однако такие подходы могут носить субъективный характер или не учитывать взаимовлияния нескольких параметров, значения которых получаются на основе прогнозирования в прикладных задачах, или оценивать горизонт прогнозирования, который можно считать достоверным. В связи с этим практическое использование получаемых результатов может быть затруднено и будет требовать последующего привлечения экспертов в рассматриваемой предметной области или иной постобработки результатов.

Рассмотрим несколько наиболее распространенных методов прогнозирования.

Одним из самых первых и простых методов, который используют для построения прогнозов, является метод регрессионного анализа. Его точность зависит от вида функции, и если данные имеют сложные закономерности, функциональное описание которых не поддается описанием

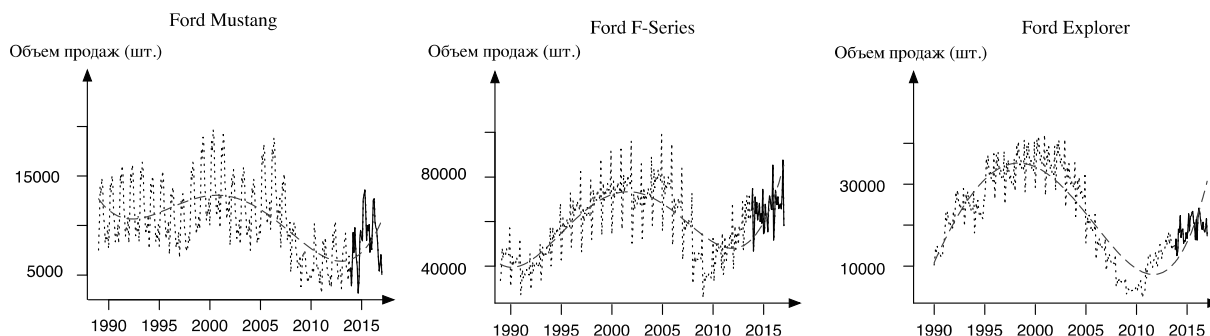


Рис. 1. Прогнозирование объема продаж методом регрессии: исходные данные (пунктир), данные для проверки (сплошная линия), прогноз (штриховая линия)

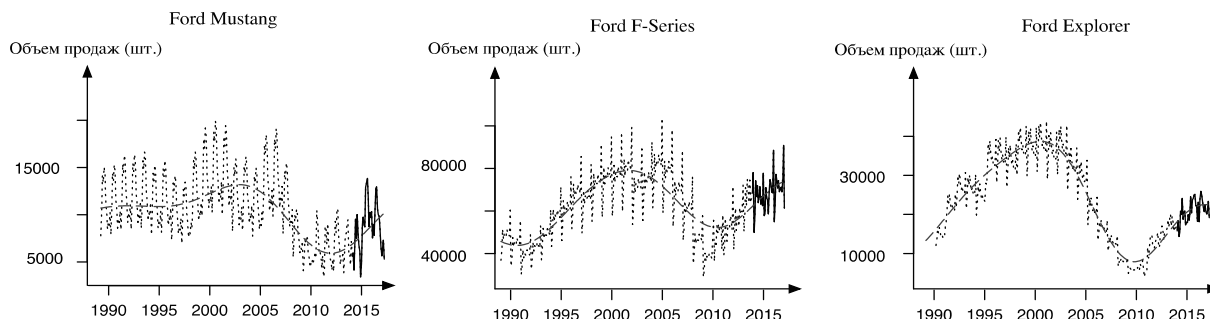


Рис. 2. Прогнозирование объема продаж методом опорных векторов: исходные данные (пунктир), данные для проверки (сплошная линия), прогноз (штриховая линия)

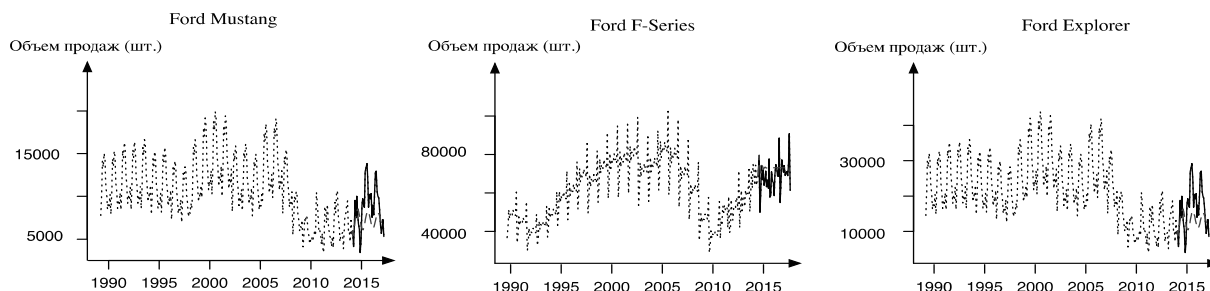


Рис. 3. Прогнозирование объема продаж методом авторегрессии: исходные данные (пунктир), данные для проверки (сплошная линия), прогноз (штриховая линия)

простых функций, используют степенные ряды, которые дают низкую точность за пределами рассматриваемого интервала с данными (рис. 1) [13].

Другим методом, получившим широкое распространение, является метод опорных векторов (рис. 2). Данный метод хорошо использовать для получения средней оценки. Когда наблюдается большой разброс значений, его результаты могут сильно отличаться от реальных. Кроме этого, его точность напрямую зависит от объема рассматриваемой выборки [11].

Для прогнозирования периодически изменяемых данных хорошо подходит метод авторегрессии, однако он плохо описывает данные, имеющие большие флуктуации. Результаты, полученные с его использованием, приведены на рис. 3 [14].

При работе с данными, имеющими большое число периодических составляющих, в конце 90-х годов получил распространение метод вейв-

лет анализа. Точность метода зависит от того, какие периодические составляющие будут учитываться. Для их выбора используют спектрограмму (рис. 4, см. вторую сторону обложки), по которой экспертно или с использованием дополнительных методов выбирают наиболее значимые частоты. При этом получаемые вейвлет-спектрограммы отличаются от обычных спектров Фурье тем, что дают четкую привязку спектра различных особенностей сигналов ко времени. После выбора частот для каждой из реконструированных кривых (кривых, описывающих одну из частотных составляющих, получаемых вейвлет-методом) проводится прогнозирование одним из методов, хорошо зарекомендовавшим себя при описании периодических данных (метод авторегрессии) [15].

Получаемые для различных частот данные для учета всех составляющих можно усреднить [16]. Для рассматриваемого примера получим данные,

Результаты оценки точности получаемых прогнозов для автомобилей Ford Mustang, Ford F-Series и Ford Explorer

Параметр	Метод регрессии	Метод опорных векторов	Метод авторегрессии	Вейвлет-анализ	Фрактальный анализ
Ford Mustang					
Среднеквадратическая ошибка	2792,051	2586,582	2780,096	4444,728	2902,137
Средняя ошибка аппроксимации, %	30,05322	28,54587	26,8457	59,44638	28,5975
Ford F-Series					
Среднеквадратическая ошибка	10543,86	7769,33	8167,399	9522,45	12945,55
Средняя ошибка аппроксимации, %	13,04173	9,30563	9,527343	12,13911	15,67242
Ford Explorer					
Среднеквадратическая ошибка	5355,316	2371,569	4883,227	17161,69	3131,141
Средняя ошибка аппроксимации, %	22,35523	10,77976	20,56545	90,50585	13,18923

приведенные на рис. 5 (см. вторую сторону обложки).

Для описания флуктуаций и кривых со сложными периодическими сигналами в последнее время широкое распространение получают фрактальные методы (рис. 6, см. вторую сторону обложки).

При этом так же, как и в других методах, большое влияние на точность прогнозирования оказывают его параметры. Основным параметром фрактальных методов является размерность используемого фрактала.

Метод прогнозирования можно выбирать исходя как из опыта, так и оценки точности (использование критериев согласия (например, χ^2 , Колмогорова—Смирнова и др.), погрешность (относительная, абсолютная, среднеквадратичная)).

Сравним получаемые результаты со среднеквадратичной ошибкой

$$\delta_{\text{отн}} = \sqrt{\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n}},$$

где n — длина временного ряда; y_t и $\hat{y}_t$ — прогнозное и точное (ретроспективное статистическое) значения, и средней ошибкой аппроксимации

$$\bar{\Delta F} = \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{n}.$$

Тогда для рассматриваемого примера видно, что одни методы показывают лучшие результаты, чем другие (табл. 1). При этом не стоит забывать о том, что разные результаты могут быть получены при использовании одного и того же метода с разными параметрами.

Однако такой подход к выбору метода прогнозирования не дает ответа на вопрос о лучшем с точки зрения точности методе для использования в какой-либо задаче, имеющей математическую формализацию и горизонт планирования.

2. Выбор метода прогнозирования на основе оценки горизонта планирования

В производственных и экономических системах большую роль играет использование прогнозов в моделях поддержки принятия решений. Широкое распространение получили методы с критериальными оценками.

Рассмотрим одну из возможных формализаций задачи объемно-календарного планирования [17]:

$$\sum_i C_i x_i \rightarrow \max, \quad (1)$$

где $C_i, i = \overline{1, N}$, — чистая прибыль от производства i -го товара (если предположить, что предприятие не производит продукцию себе в убыток и норма прибыли всей продукции примерно одинакова, то в качестве C_i можно использовать стоимость единицы продукции); x_i — объем производства i -го товара.

На задачу могут накладываться ограничения на производственные мощности, на ключевые материалы, ограничения по рынку продаж:

$$\sum_i x \leq G^q, q = \overline{1, Q},$$

где G^q — ограничение по рынку сбыта, определяемое на основании прогноза. При рассмотрении задачи объемно-календарного планирования C_i и G^q становятся зависимыми от времени и определяются на основании данных прогнозов. В некоторых случаях, если нет информации о планируемом изменении цен, то в качестве C_i следует использовать последнее известное значение.

В зависимых от времени параметрах в качестве значений используют результаты прогнозирования, что поднимает вопрос, связанный с достоверностью получаемых результатов в моделях управления и поддержки принятия решений в производственных и экономических системах. На

Число итераций до резкого роста значения оценки риска

Тип автомобиля	Метод регрессии	Метод опорных векторов	Метод авторегрессии	Вейвлет анализ	Фрактальный анализ
Ford Mustang	5	5	10	8	9
Ford F-Series	1	10	28	1	2
Ford Explorer	7	25	18	1	19

$$r = \left| 1 - \frac{a}{a^*} \right|,$$

где a — прогнозное значение величины оцениваемого фактора; a^* — точное значение оцениваемого фактора.

Для определения горизонта планирования будем использовать тестовую выборку данных. Когда мы рассматриваем задачу управления, в которой используются данные прогнозирования для нескольких переменных, то так как в качестве переменных, как правило, выбирают независимые переменные, а значения они принимают в каждый момент времени, то эти события стоит рассматривать как независимые.

Для независимых параметров c_1 и c_2 справедлива зависимость $P(c_1c_2) = P(c_1)P(c_2)$, где $P(c)$ — вероятность наступления события c . Основываясь на формуле $P = 1 - r$, где P — вероятность, получаем выражение для определения оценки риска

$$\begin{aligned} r(c_1c_2) &= 1 - P(c_1c_2) = 1 - P(c_1)P(c_2) = \\ &= 1 - (1 - r(c_1))(1 - r(c_2)) = \\ &= 1 - 1 + r(c_1) + r(c_2) - r(c_1)r(c_2) = \\ &= r(c_1) + r(c_2) - r(c_1)r(c_2). \end{aligned}$$

Значения прогнозов каждого параметра представляют собой временной ряд. Учитывая то, что эти значения рассчитываются одновременно, их стоит рассматривать как совместные события.

Тогда оценку риска можно вычислить нарастающим итогом, используя выражение для вероятности наступления события, при котором значения параметра c_1 примут значения c_{1_1} и c_{1_2} — $P(c_{1_1} + c_{1_2}) = P(c_{1_1}) + P(c_{1_2}) - P(c_{1_1})P(c_{1_2})$. Отсюда получаем $r(c_{1_1} + c_{1_2}) = 1 - P(c_{1_1} + c_{1_2}) = 1 - 1 + r(c_{1_1}) - 1 + r(c_{1_2}) + 1 - r(c_{1_1}) - r(c_{1_2}) + r(c_{1_1})r(c_{1_2}) = r(c_{1_1})r(c_{1_2})$.

Для вычисления последующих значений оценки риска будем использовать формулу

$$r(c_{1_{i-2}} + c_{1_{i-1}} + c_{1_i}) = r(c_{1_{i-2}} + c_{1_{i-1}}) + r(c_{1_{i-1}})r(c_{1_i}).$$

Применив приведенные выкладки к данным рассмотренного выше примера, получим, что с некоторого момента оценка риска начинает резко

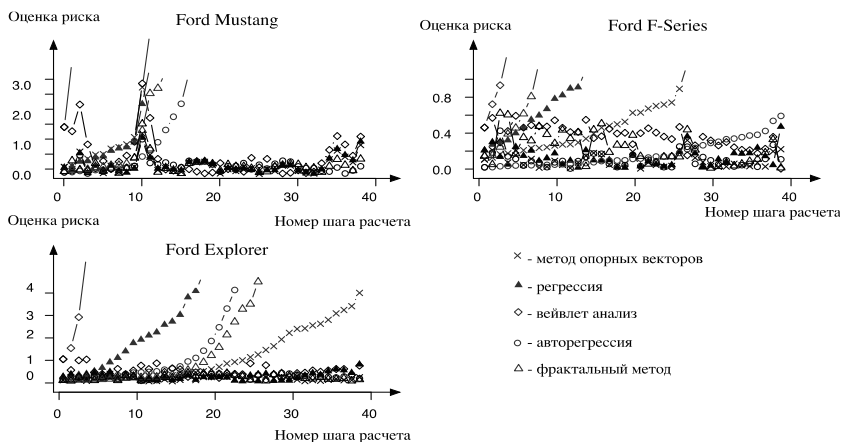


Рис. 7. Моментальные величины оценки риска при использовании прогнозов и значения нарастающим итогом

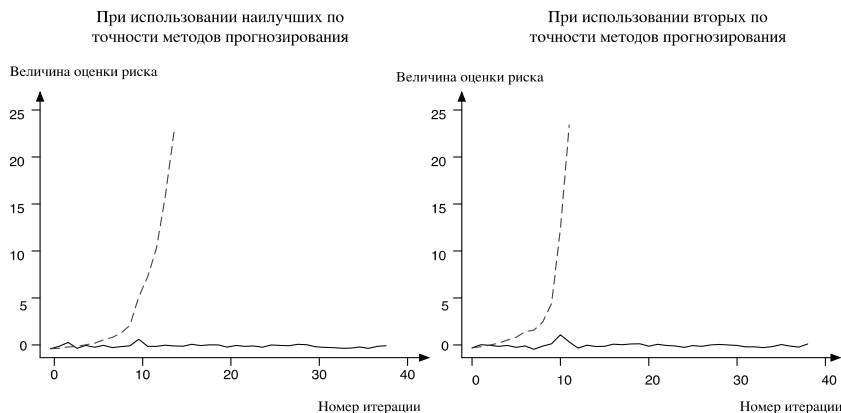


Рис. 8. Значения оценок риска (сплошная линия) и оценок риска нарастающим итогом (штриховая линия) для решений задачи объемно-календарного планирования на основе значений прогнозов с тремя товарами

расти, что, в свою очередь, позволяет нам определить горизонт прогнозирования для каждого метода и выбирать метод для прогнозирования исходя из максимально удаленного во времени горизонта планирования (рис. 7).

Как видно из табл. 1 и 2, выбираемые с использованием предложенного подхода методы прогнозирования совпадают с лучшими методами прогнозирования по показателю точности.

При использовании прогнозов, получаемых методами, выбранными с помощью табл. 2, в задаче (1) [16] оценки риска будут принимать значения, приведенные на рис. 8.

В результате приведенных рассуждений можно сформулировать следующую последовательность действий для решения задачи определения горизонта планирования в прогностических моделях для решения задач управления производственными системами:

- 1) формирование набора ретроспективных данных для решения задачи прогнозирования;
- 2) разделение данных на две последовательности: обучающую и тестовую выборки;
- 3) экспертный выбор нескольких методов прогнозирования для независимых параметров рассматриваемой задачи управления и получения предварительных прогнозов;
- 4) выбор методов прогнозирования на основании экспертной оценки горизонта прогнозирования и допустимой для задачи оценки риска, связанного с использованием данных прогнозирования;
- 5) определение общей оценки риска для рассматриваемой задачи управления; если она оказывается допустима для использования получаемых результатов, то определение горизонта планирования.

Заключение

Использование данных прогнозов в задачах оптимального управления открывает новые возможности исследования процессов, протекающих в производственных системах и обусловленных внедрением товарных проектов, а также экономико-математических моделей и методов управления этими процессами. Управление рисками в получаемых моделях позволяет учитывать вероятностную природу процессов, протекающих вне рассматриваемой производственной системы, но непосредственно влияющих на ее деятельность.

Предложенный в статье подход связанный с использованием прогнозов в задачах планирования и управления и учетом оценок рисков в этих задачах является развитием некоторых идей, заложенных такими учеными как Ж. Тироль, а также Г. Марковиц, Ж. Линтер, Ж. Моссинь, но в отличие от них позволяет учитывать фактор времени, что проявляется в возможности выбора горизонта планирования на основе оценок риска и определения связанных с выбираемыми решениями оценок риска.

Поскольку рассмотрены только единичные примеры и расчеты, то при реализации проектов с учетом получаемых на модели значений может возникнуть ситуация, когда реальные значения будут отличаться от полученных в результате прогнозирования. Решение этой проблемы связано с необходимостью дополнительного исследова-

ования получаемых решений и оценок риска в зависимости от точности значений прогнозов. Для исследования такой зависимости необходимо моделирование отклонений значений прогнозов в доверительных интервалах с учетом их плотности вероятности. Такие оценки станут предметом дальнейших исследований и будут помогать принимать решения в условиях ограниченного времени. Их применение позволит не только оценить риски, связанные с использованием прогнозов, но и позволит оценить точность и чувствительность используемых для принятия решений моделей. Кроме этого, предлагаемый подход может быть распространен на оценку производственных рисков при проведении многократных расчетов и учете рисков, связанных с износом оборудования [19].

Авторы благодарят Правительство Пермского края за поддержку исследований в рамках научно-го проекта № С-26/058.

Список литературы

1. Aissaoui N., Haouari M., Hassini E. Supplier selection and order lot sizing modeling: a review // Computers & Operations Research — 2007. Т. 34. № 12. С. 3516—3540.
2. Абдуллаев А. Р., Мыльников Л. А., Васильева Е. Е. Риски в инновационных проектах: причины появления, интегральные риски, экспертиза проектов с учетом рисков // Экономический анализ: теория и практика. 2012. № 40 (295). С. 41—49.
3. Tajbakhsh S. D., Castillo E. del, Rosenberger J. L. A Bayesian Approach to Sequential Optimization based on Computer Experiments // Quality and Reliability Engineering International. 2015. Т. 31. № 6. С. 1001—1012.
4. Moghaddam K. S. Supplier selection and order allocation in closed-loop supply chain systems using hybrid Monte Carlo simulation and goal programming // International Journal of Production Research. 2015. Т. 53. № 20. С. 6320—6338.
5. Андерсон Т. Статистический анализ временных рядов. М.: Мир, 1976. 755 с.
6. A primer on partial least squares structural equation modeling (PLS-SEM) / ред. J. F. Hair. Los Angeles: Sage, 2017.
7. Кулинич А. А. Компьютерные системы моделирования когнитивных карт: подходы и методы // Проблемы управления. 2010. № 3. С. 2—16.
8. Мыльников Л. А. Поддержка принятия решений при управлении инновационными проектами. Пермь: Перм. гос. техн. ун-т, 2011. 145 с.
9. Файзрахманов Р. А., Мыльников Л. А., Алкдиру Р. Х., Суслова А. А. Методика определения и уточнения значений коэффициентов связей когнитивных карт на примере анализа взаимосвязи объемов финансирования научных исследований и патентной активности // Экономический анализ: теория и практика. 2013. № 30 (333). С. 43—51.
10. Афанасьев В. Н., Юзбашев М. М. Анализ временных рядов и прогнозирование. М.: ИНФРА-М, 2010. 320 с.
11. Jose C., Goyal P., Aggrwal P. Local deep kernel learning for efficient non-linear SVM prediction // In Proceedings of the International Conference on Machine Learning, Atlanta, Georgia, June 2013. P. 486—494.
12. Turan J., Ratnov L., Bengio Y. Word representations: a simple and general method for semi-supervised learning Association for Computational Linguistics, 2010. pp. 384—394.
13. Садовникова Н. А., Шмойлова Р. А. Анализ временных рядов и прогнозирование, 2016. 151 с.
14. Третьуб А. В., Третьуб И. В. Методика построения модели ARIMA для прогнозирования динамики временных

рядов // Вестник Московского государственного университета леса — Лесной вестник. 2011. № 5. С. 179—184.

15. **Мифтахова Н. Э., Панасюк М. В.** Вейвлет-анализ динамики региональной системы // Ученые записки Казанского университета. Серия "Естественные науки". 2009. № 1 (151). С. 247—262.

16. **Алексеев В. И.** Анализ и прогнозирование циклических временных рядов с использованием вейвлетов и нейросетевых нечетких правил вывода // Вестник Югорского государственного университета. 2013. № 3 (30). С. 3—10.

17. **Мыльников Л. А., Файзрахманов Р. А.** Роль имитационного моделирования и вычислительных программных комплексов для принятия управленческих решений при

реализации товарных инновационных проектов в производственно-экономических системах // Программные системы и вычислительные методы. 2015. Т. 4. № 4. С. 390—406.

18. **Mylnikov L.** Risk Evaluation in Manufacturing Organization Tasks for Product Technological Projects and Establishment of Project Portfolio for Production Systems // Proceedings of the 2016 International Conference on Applied mathematics, simulation and modelling Beijing, China, 2016. P. 399—402.

19. **Pan E., Liao W., Xi L.** A joint model of production scheduling and predictive maintenance for minimizing job tardiness // The International Journal of Advanced Manufacturing Technology. 2012. Vol. 60. N 9—12. P. 1049—1061.

L. A. Mylnikov, Associate Professor, e-mail: leonid.mylnikov@pstu.ru,

A. B. Seledkova, Master Student, e-mail: aleksandrasedkova@yandex.ru,

Perm National Research Polytechnic University

Selection of the Parameters Predicting Method in Production and Economic Systems Based on Risks Metrics

The article describes the choice features of methods for predicting the parameter values of production and economic systems. For this, the data features collected and used are considered, the main problem in the projects' data analysis implemented in production and economic systems will be either a small amount of statistical data or data describing a number of products, product modifications, etc. (I.e., undivided data). The main approaches used for predicting parameters are considered. These are approaches described by special curves, supervised learning, unsupervised learning and semi-supervised learning. Features of the forecasted data and also forecasts based on regression analysis, the method of support vector machine, auto regression and wavelet analysis are constructed. For the predicted values, the method of choosing the forecasting method based on the risk assessment of the introduced forecasts is given. The choice is proposed to be made on the basis of an estimate of the exact prediction for a maximum period of time. Since, starting with a certain forecast value, risk value assessment begins to increase sharply, which allow choosing the method of forecasting. Also, the application of forecasts in decision-making tasks is considered.

Keywords: Forecast, risk metrics, management parameters, regression analysis, reference vector method, wavelet analysis

References

1. **Aissaoui N., Haouari M., Hassini E.** Supplier selection and order lot sizing modeling: A review, *Computers & Operations Research*, 2007, no. 12 (34), pp. 3516—3540.

2. **Abdullaev A. R., Mylnikov L. A., Vasileva E. E.** Riski v innovacionnyh proektah: prichiny pojavleniya, integral'nye riski, ehkspertiza proektov s uchetom riskov, *Jekonomicheskij analiz: teorija i praktika*, 2012, no. 40 (295), pp. 41—49 (in Russian).

3. **Tajbakhsh S. D., Castillo E. del, Rosenberger J. L.** A Bayesian Approach to Sequential Optimization based on Computer Experiments, *Quality and Reliability Engineering International*, 2015, no. 6 (31), pp. 1001—1012.

4. **Moghaddam K. S.** Supplier selection and order allocation in closed-loop supply chain systems using hybrid Monte Carlo simulation and goal programming, *International Journal of Production Research*, 2015, no. 20 (53), pp. 6320—6338.

5. **Anderson T.** Statisticheskij analiz vremennyh rjadov. Moscow: Mir, 1976. 755 p. (in Russian).

6. **A primer** on partial least squares structural equation modeling (PLS-SEM). Los Angeles: Sage, 2017.

7. **Kulinich A. A.** Komp'yuternye sistemy modelirovaniya kognitivnyh kart: podhody i metody, *Problemy upravleniya*. 2010. no. 3, pp. 2—16 (in Russian).

8. **Mylnikov L. A.** Podderzhka prinjatija reshenij pri upravlenii innovacionnymi proektami. Perm: PSTU, 2011, 146 p. (in Russian).

9. **Fayzrahmanov R. A., Mylnikov L. A., Suslova A. A.** Metodika opredelenija i utocnenija znachenij koeficientov svyazej kognitivnyh kart na primere analiza vzaimosvyazi ob'emov finansirovaniya nauchnyh issledovanij i patentnoj aktivnosti, *Jekonomicheskij analiz: teorija i praktika*, 2013, no. 30 (333), pp. 43—51 (in Russian).

10. **Afanas'ev V. N., Juzbashev M. M.** Analiz vremennyh rjadov i prognozirovanie. Moscow: INFRA-M, 2010. 320 p. (in Russian).

11. **Jose C., Goyal P., Aggrwal P.** Local deep kernel learning for efficient non-linear SVM prediction. In Proceedings of the

International Conference on Machine Learning, Atlanta, Georgia, June 2013, pp. 486—494.

12. **Turan J., Ratnoff L., Bengio Y.** Word representations: a simple and general method for semi-supervised learning Association for Computational Linguistics, 2010, pp. 384—394.

13. **Sadovnikova N. A., Shmojlova R. A.** Analiz vremennyh rjadov i prognozirovanie, 2016, 151 p. (in Russian).

14. **Tregub A. V., Tregub I. V.** Metodika postroenija modeli ARIMA dlja prognozirovaniya dinamiki vremennyh rjadov, *Vestnik Moskovskogo gosudarstvennogo universiteta lesa—Lesnoj vestnik*. 2011. no. 5. pp. 179—184 (in Russian).

15. **Miftahova M. Je., Panasjuk M. V.** Veyvlet-analiz dinamiki regional'noj sistemy, *Uchenye zapiski Kazanskogo universiteta. Serija Estestvennye nauki*. 2009. no. 1 (151). pp. 247—262 (in Russian).

16. **Alekseev V. I.** Analiz i prognozirovanie ciklicheskih vremennyh rjadov s ispol'zovaniem veyvletov i nejrosetevykh nechetkiy pravil vyvoda, *Vestnik Jugorskogo gosudarstvennogo universiteta*, 2013, no. 3 (30), pp. 3—10 (in Russian).

17. **Mylnikov L. A., Fayzrahmanov R. A.** Rol' imitacionnogo modelirovaniya i vychislitel'nyh programmnyh kompleksov dlja prinjatija upravlencheskih reshenij pri realizacii tovarnyh innovacionnyh proektov v proizvodstvenno-jekonomicheskijh sistemah, *Programmnye sistemy i vychislitel'nye metody*, 2015, no. 4 (4), pp. 390—406 (in Russian).

18. **Mylnikov L.** Risk Evaluation in Manufacturing Organization Tasks for Product Technological Projects and Establishment of Project Portfolio for Production Systems, *Proceedings of the 2016 International Conference on Applied mathematics, simulation and modelling* Beijing, China, 2016, pp. 399—402.

19. **Pan E., Liao W., Xi L.** A joint model of production scheduling and predictive maintenance for minimizing job tardiness, *The International Journal of Advanced Manufacturing Technology*, 2012, no. 9—12 (60), pp. 1049—1061.

БЕЗОПАСНОСТЬ ИНФОРМАЦИИ

CRYPTOSAFETY INFORMATION

УДК 004.056.53

Т. С. Осадчая, инженер, e-mail: taniaosadchaya6@gmail.com, А. Ю. Щеглов, д-р техн. наук, проф.,
Университет ИТМО, Санкт-Петербург, Россия

Комплексное решение задачи защиты от атак с правами привилегированного пользователя

Работа посвящена решению задачи комплексной защиты от атак с правами привилегированного пользователя. Рассматриваются следующие источники данной угрозы безопасности: легальные пользователи, имеющие привилегированные права, и вредоносные программы, запущенные с привилегированными правами. Рассмотрен подход, обеспечивающий контроль и усечение действий привилегированных пользователей, в том числе усечение их возможностей по администрированию, а также исключение возможности влияния вредоносных программ, запущенных с привилегированными правами, на систему.

Ключевые слова: привилегированные пользователи, инсайдеры, контроль доступа, права доступа, усечение возможностей пользователей по администрированию, механизм самозащиты, вредоносная программа

Введение

В настоящее время одной из актуальных угроз безопасности является наличие привилегированных прав, поскольку привилегированные пользователи имеют доступ к важным, имеющим стратегическое значение, данным, а также к конфиденциальной информации о компании или ее сотрудниках [1].

Существуют два источника этой угрозы безопасности:

- действия пользователя, имеющего привилегированные права (например, системного администратора);
- запуск тем или иным образом вредоносной программы, предполагающей некое воздействие на систему, с привилегированными правами.

Необходимо отметить, что привилегированный пользователь является одним из самых опасных инсайдеров — внутренних пользователей, которые преднамеренно или случайно могут допустить утечку конфиденциальной информации [2, 3].

В случае неправомерных действий пользователей, обладающих привилегированными правами, может быть нанесен серьезный ущерб, так как такие пользователи имеют доступ к системным настройкам, управлению пользователями и управлению разрешениями, т. е. имеют широкий доступ к критически важным приложениям, ключевым системам и информации [4].

Сегодня средством любой атаки, в частности целевой, направленной против конкретной компании или государственного органа и предполагающей использование уязвимостей в различных программных средствах, является запуск какой-либо сторонней вредоносной программы с правами привилегированного пользователя. Исследования показывают, что 80 % целевых атак предполагают взлом привилегированной учетной записи [5].

Следовательно, **задачами средств защиты информации** являются:

- контроль привилегированных пользователей и возможность воздействия на их права в целях снижения вероятности осуществления данными пользователями инсайдерской атаки;
- предотвращение влияния программ, запущенных с привилегированными правами, на систему.

Поскольку в основе архитектуры защиты современных универсальных операционных систем (в частности, систем семейств Windows и Unix) лежит принцип "полного доверия к администратору", их встроенные механизмы защиты не могут обеспечить эффективного противодействия внутренним угрозам: ими не предоставляются возможности отслеживания и защиты от действий привилегированных пользователей [6]. Следовательно, контроль действий привилегированных пользователей возможен только по-

средством применения специализированных решений — средств контроля привилегированных пользователей.

Указанные выше задачи необходимо решать в комплексе. Необходимо как контролировать и разграничивать права доступа привилегированного пользователя (администратора), так и предотвращать возможность влияния на систему программ, которые могут быть наделены вредоносными свойствами и запущены под привилегированной учетной записью.

Для этого субъект доступа должен содержать два компонента: "учетная запись" и "процесс". Кроме того, защита, связанная с усечением и контролем прав доступа, должна решаться на системном уровне для противодействия любому воздействию привилегированных пользователей на систему защиты (удаление, останов служб и т. д.).

В настоящее время развито направление, связанное с контролем действий привилегированных пользователей. Англоязычными аналогами наименования решений по управлению привилегированными пользователями являются Privileged User Management (PUM), Privileged Identity Management (PIM) и др. [7].

Принцип работы средств контроля привилегированных пользователей представляет собой проверку и подтверждение полномочий привилегированных пользователей, выявление их подозрительной активности с помощью регистрации всех действий, уведомление ответственных лиц о такой активности и при необходимости принудительный разрыв сессии.

Основными задачами, решаемыми средствами контроля привилегированных пользователей, являются:

- централизованное управление привилегированными учетными записями, автоматическое обнаружение привилегированных учетных записей;
- управление аутентификацией и авторизацией (включая управление парольной защитой) для привилегированных учетных записей для точной идентификации пользователя, работающего под учетной записью администратора в конкретный момент времени;
- аудит действий, выполняемых привилегированными пользователями (определение времени обращения к конфиденциальным данным и имени пользователя, ведение журнала выполненных действий, их отслеживание в режиме реального времени), в том числе видеофиксация всех подключений и сессий;
- создание политик доступа для привилегированных пользователей к корпоративным ресурсам (типы сеансов, протоколы доступа, время доступа и т.п.), так данные пользователи

ли получают только необходимый и достаточный доступ к обслуживаемым информационным системам.

Таким образом, существующие в настоящее время решения, посвященные контролю действий привилегированных пользователей, реализуются внешними средствами и в первую очередь ориентированы на задачи контроля привилегированных пользователей [8, 9].

В составе подавляющей части существующих PUM-решений отсутствуют клиентская часть и собственные драйверы. Как следствие, данные решения могут реализовывать только настройку встроенных возможностей операционной системы, при этом средство защиты становится "шлюзом" между машиной администратора и целевой машиной для контроля действий привилегированных пользователей.

PUM-решения ставят своей основной целью регистрацию всех действий привилегированных пользователей для их дальнейшего анализа. Однако данными средствами не решается в полной мере задача по усечению возможностей пользователей, имеющих повышенные привилегии, по администрированию. Также в них отсутствует защита от целевых атак, более того, вносится дополнительная угроза уязвимости самого PUM-решения.

Комплексная система защиты информации "Панцирь+" для ОС Microsoft Windows (КСЗИ "Панцирь+") на сегодняшний день является единственной системой, в которой субъект доступа идентифицируется парой "пользователь — процесс". Данная возможность является запатентованной [10]. При этом КСЗИ работает на системном уровне и запускается как системная служба, права которой выше, чем права администратора. КСЗИ "Панцирь+" решает задачи, отличные от задач PUM-решения, что позволяет апробировать предлагаемый в данной статье подход к защите.

Целью исследования является повышение уровня безопасности и сохранение конфиденциальности и целостности системных объектов, настроек операционной системы и данных пользователей.

Основными задачами, решаемыми в комплексе в рамках реализации этой цели, являются:

контроль и усечение действий легальных привилегированных пользователей;

предотвращение влияния на систему вредоносных программ, запущенных с привилегированными правами.

Идея подхода к защите заключается в необходимости:

- создания разграничительной политики доступа для пользователей с привилегированными правами;

- исключения возможности воздействия любыми сторонними программами на системные объекты и объекты, хранящие данные.

Необходимо отметить, что данный подход базируется на реализации механизма самозащиты, обеспечивающего невозможность влияния на функционирование средства защиты информации привилегированными пользователями любым способом. Данный механизм не позволит привилегированному пользователю, не являющемуся администратором безопасности, каким-либо образом повлиять на запущенную службу и драйверы средства защиты, модифицировать его настройки.

Кроме того, средством защиты должны реализовываться два уровня иерархии администраторов: администратор безопасности и остальные привилегированные пользователи. Администратор безопасности определяет права других привилегированных пользователей, разрешая им требуемый набор функций администрирования из предоставленных операционной системой. Средством защиты при этом может управлять только администратор безопасности. Без реализации подобной иерархии осуществить усечение прав пользователей, имеющих повышенные привилегии, невозможно.

Контроль и усечение действий легальных привилегированных пользователей

В общем случае задача защиты информации от привилегированных пользователей заключается в том, чтобы данные пользователи (например, системные администраторы) не имели доступа к файлам, создаваемым другими пользователями [11]. Данная задача решается благодаря созданию разграничительной политики доступа с использованием механизма контроля доступа к создаваемым

файлам. Основу контроля доступа к создаваемым файлам составляет их автоматическая разметка. При создании субъектом нового файла он автоматически размечается: им наследуется учетная информация субъекта доступа, создавшего этот файл [12]. Используя данный механизм, можно задать такие правила, что привилегированный пользователь будет иметь доступ только к тем файлам, которые были созданы им.

Для контроля действий привилегированных пользователей необходимо настроить параметры аудита, задав объекты, доступ к которым будет регистрироваться в журналах аудита.

Существуют следующие виды контроля действий:

- контроль отказов — в аудите отображаются выполненные пользователем действия, которые запрещены ему разграничительной политикой (рис. 1);
- контроль действий — в аудите отображаются выполненные пользователем действия (рис. 2).

Кроме того, возможна локализация аудита: благодаря настройке параметров аудита можно контролировать не все действия привилегированных пользователей, а, например, только работу с различными оснастками панели управления, редактирование автозагрузки, повышение привилегий.

Для усечения возможностей привилегированных пользователей по администрированию следует использовать подход, основанный на разрешении доступа к защищаемому объекту файловой системы или реестра только соответствующей библиотеке, которая включает в себя интерфейс настройки. Таким образом, создается разграничительная политика для процессов, после чего привилегированному пользователю разрешается либо запрещается запуск данной библиотеки.

В рамках данного подхода сначала выбирается требуемая функция администрирования, после

Номер	Время	Процесс	Пользователь	Режим дс	Имя объекта	Им	Разное
1	Пн 20/03/2017 13...	D:\Windows\System32\mmc.exe	TEST-PC\Администратор	Ч	D:\Windows\System32\dmocx.dll		ДОСТУП ЗАПРЕЩЕН!
2	Пн 20/03/2017 13...	D:\Windows\System32\svchost.exe	TEST-PC\Администратор ...	Ч	D:\Windows\System32\vdslldr.exe		ДОСТУП ЗАПРЕЩЕН!
3	Пн 20/03/2017 13...	D:\Windows\System32\svchost.exe	TEST-PC\Администратор ...	Ч	D:\Windows\System32\vdslldr.exe		ДОСТУП ЗАПРЕЩЕН!
4	Пн 20/03/2017 13...	D:\Windows\explorer.exe	TEST-PC\Администратор	ЧЗ	D:\Windows\System32\Tasks		ДОСТУП ЗАПРЕЩЕН!
5	Пн 20/03/2017 13...	D:\Windows\explorer.exe	TEST-PC\Администратор	ЧЗ	D:\Windows\System32\Tasks\GoogleUpdateTaskMachineUA		ДОСТУП ЗАПРЕЩЕН!

Рис. 1. Пример аудита (контроль отказов)

Номер	Время	Процесс	Пользователь	Режим доступа	Имя объекта
1	Чт 09/03/2017 ...	D:\Windows\System32\mmc.exe	TEST-PC\Администратор	Ч И	D:\Windows\System32\dmdlgls.dll
2	Чт 09/03/2017 ...	D:\Windows\System32\svchost.exe	TEST-PC\Администратор	Ч И	D:\Windows\System32\vdslldr.exe
3	Чт 09/03/2017 ...	D:\Windows\System32\mmc.exe	TEST-PC\Администратор	Ч И	D:\Windows\System32\dmview.ocx
4	Чт 09/03/2017 ...	D:\Windows\System32\vdslldr.exe	TEST-PC\Администратор	Ч И	D:\Windows\System32\vdutil.dll
5	Чт 09/03/2017 ...	D:\Windows\System32\vdslldr.exe	TEST-PC\Администратор	Ч И	D:\Windows\System32\vds_ps.dll

Рис. 2. Пример аудита (контроль действий)

чего определяется библиотека или исполняемый файл, с помощью которого возможна работа с определенным интерфейсом, связанным с выполнением рассматриваемой функции.

Затем обеспечивается возможность осуществления записи только из выбранного интерфейса: осуществлять запись в защищаемый объект файловой системы или реестра должен только один определенный процесс, всем остальным процессам запись в данный объект запрещается.

После этого при создании разграничительной политики для привилегированного пользователя ему разрешается или запрещается запуск этой библиотеки или исполняемого файла.

Примеры разграничительных политик, используемые далее в данной работе, были созданы с использованием КСЗИ "Панцирь +" для ОС Microsoft Windows.

Пример разграничительной политики для защиты от действий легальных привилегированных пользователей

Были заданы правила доступа привилегированному пользователю к создаваемым файлам, которые запрещают ему доступ к файлам, созданным и обрабатываемым интерактивными пользователями (рис. 3).

Субъект осуществляющий доступ	Субъект-создатель файла	Режим доступа	Режим аудита
 Администратор	любой	-Ч-З-И-У-П	ЧСИП:ЧСИП
 Администратор	Администратор	+Ч+З-И+У+П	--И--:--И--

Рис. 3. Пример разграничительной политики доступа к создаваемым файлам

Примеры разграничительных политик

Возможность по администрированию	Объект	Права доступа
Открытие элемента панели управления "Учетные записи пользователей"	C:\Windows\System32\usercpl.dll	-И
Отображение вкладки "Безопасность" в свойствах папок и файлов	C:\Windows\System32\authz.dll	-И
Открытие элемента панели управления "Дата и время"	C:\Windows\System32\timedate.cpl; HKLM\System\CurrentControlSet\Control\Time-ZoneInformation; HKLM\SYSTEM\ControlSet001\Control\Time-ZoneInformation	-И; +Ч-З-У-П; +Ч-З-У-П

В результате реализации данной разграничительной политики привилегированный пользователь сможет иметь доступ только к файлам, созданным им. При этом ему запрещено запускать созданные им исполняемые файлы.

Благодаря запрету на запуск соответствующих динамических библиотек были усечены возможности привилегированного пользователя по администрированию.

Рассмотрим примеры политик безопасности, соответствующих усечению различных возможностей привилегированных пользователей по администрированию. Ниже представлены правила, относящиеся к привилегированным пользователям. До их создания необходимо сначала запретить любым процессам чтение/исполнение требуемых динамических библиотек или исполняемых файлов, а также запретить запись в соответствующие объекты файловой системы и реестра. Примеры разграничительных политик представлены в таблице.

Здесь "-И" — запрет на исполнение, "+Ч" — разрешение на чтение, "-З" — запрет на запись, "-У" — запрет на удаление, "-П" — запрет на переименование.

В результате реализации данных разграничительных политик (рис. 4—6) Администратор не сможет открыть элемент панели управления

Профиль:  Администратор			
Правила доступа для выбранного профиля			
Тип	Объект файловой системы	Режим доступа	Режим аудита
	C:\Windows\System32\usercpl.dll	-И	--И--:--И--

Рис. 4. Пример разграничительной политики (открытие элемента панели управления)

Профиль:  Администратор			
Правила доступа для выбранного профиля			
Тип	Объект файловой системы	Режим доступа	Режим аудита
	C:\Windows\System32\authz.dll	-И	--И--:--И--

Рис. 5. Пример разграничительной политики (отображение вкладки "Безопасность")

Профиль:  Администратор			
Правила доступа для выбранного профиля			
Тип	Объект файловой системы	Режим доступа	Режим аудита
	C:\Windows\System32\timedate.cpl	-И	--И--:--И--
Тип	Объект реестра	Режим доступа	Режим аудита
	HKLM\System\CurrentControlSet\Control\Time-ZoneInformation	+Ч-З-У-П	ЧСИП:ЧСИП
	HKLM\SYSTEM\ControlSet001\Control\Time-ZoneInformation	+Ч-З-У-П	ЧСИП:ЧСИП

Рис. 6. Пример разграничительной политики (открытие элемента панели управления)

"Учетные записи пользователей", увидеть вкладку "Безопасность", открыв свойства какой-либо папки или файла, открыть элемент панели управления "Дата и время".

Исключение возможности влияния на систему вредоносной программы, запущенной с привилегированными правами

Реализация рассматриваемого подхода к защите состоит в выборе объектов, запись в которые позволена только определенным процессам. При этом должна предотвращаться возможность их модификации любыми другими процессами.

Рассматриваемые подходы реализованы на практике и апробированы при построении комплексной системы защиты информации "Панцирь+" для ОС Microsoft Windows.

Пример разграничительной политики для защиты от влияния на систему вредоносных программ, запущенных с привилегированными правами

Был выбран объект "%SystemRoot%\system32*"; всем процессам, находящимся в директории "%SystemRoot%\system32*", был предоставлен полный доступ к данному объекту; всем остальным процессам (используется маска "*") модификация выбранного объекта была запрещена (т.е. им разрешается только чтение и исполнение объектов, находящихся в выбранной директории).

Таким образом, в разграничительной политике были созданы следующие правила (рис. 7).

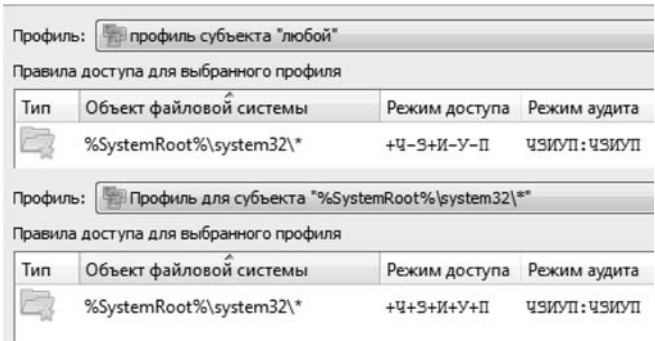


Рис. 7. Пример разграничительной политики для защиты от влияния вредоносных программ: "Ч" — чтение, "З" — запись, "И" — исполнение, "У" — удаление, "П" — переименование

В результате реализации данной разграничительной политики любая сторонняя программа не будет иметь возможности провести запись или модификацию выбранных объектов.

Рассмотрим следующий пример: пусть некая программа Procmon64.exe запущена с правами привилегированного пользователя в целях изменения выбранных файлов.

Поскольку данная программа не расположена в директории "%SystemRoot%\system32", при обращении ее к объектам, находящимся в указанной выше директории, возможность влияния данной программой на систему будет исключена. В журнале контроля действий появятся сообщения, представленные на рис. 8.

Заключение

Подводя итог, следует отметить, что в результате осуществления рассмотренного подхода к реализации разграничительной политики при определенных условиях поставленные задачи могут быть решены эффективно. Удастся не только защитить информацию, обрабатываемую на компьютере, от привилегированных пользователей и усечь возможности данных пользователей по администрированию, но и исключить возможность влияния на систему вредоносных программ, запущенных с привилегированными правами. При реализации рассматриваемого подхода возможности привилегированного пользователя будут определяться системой защиты, а любая сторонняя программа не будет иметь возможности провести запись или модификацию выбранных системных объектов.

Рассматриваемое решение было апробировано при разработке комплексной системы защиты информации "Панцирь+" для ОС Microsoft Windows.

Такое решение позволяет серьезно разграничить права привилегированных пользователей. Поэтому при использовании подобной системы актуальной становится задача защиты от повышения прав администратора до прав системы. Данный вопрос будет рассмотрен в следующей статье.

Список литературы

1. Жан-Ноэль де Гальзан. CEO WALLIX — Комплексная политика безопасности должна включать эф-

Журнал управления доступом к файловой системе						
Номер	Время	Процесс	Пользователь	Режим доступа	Имя объекта	Им. Разное
1	Пн 17/04/2017...	C:\Users\Admin_8\AppData\Local\Temp\Procmon64.exe	TATbЯHA-ПK\Admin_8	ЧЗ	C:\Windows\System32\drivers\PROCMON23.SYS	ДОСТУП ЗАПРЕЩЕН
2	Пн 17/04/2017...	C:\Users\Admin_8\AppData\Local\Temp\Procmon64.exe	TATbЯHA-ПK\Admin_8	ЧЗ	C:\Windows\System32\drivers\PROCMON23.SYS	ДОСТУП ЗАПРЕЩЕН
3	Пн 17/04/2017...	C:\Users\Admin_8\AppData\Local\Temp\Procmon64.exe	TATbЯHA-ПK\Admin_8	З	C:\Windows\System32\drivers\PROCMON23.SYS	ДОСТУП ЗАПРЕЩЕН

Рис. 8. Отказы в доступе

фективное управление привилегированными пользователями. URL: http://it-bastion.com/wp-content/uploads/2016/10/JN_interview_2013RU.pdf (дата обращения: 20 ноября 2017 г.).

2. **Беляева М., Синельников А.** Защита от инсайдеров — миф или реальность? // *Information Security / Информационная безопасность*. 2008. № 3.

3. **Угроза** привилегированных пользователей и методы по ее устранению. URL: <http://ppt.ru/guide/news/136623> (дата обращения: 20 ноября 2017 г.).

4. **Сердюк В., Романов М.** Обратная сторона привилегий // *BIS JOURNAL — Информационная безопасность банков*. 2017. № 1.

5. **Крис Брук.** 88 % сетей уязвимы к взлому привилегированных аккаунтов. URL: <https://threatpost.ru/88-percent-of-networks-susceptible-to-privileged-account-hacks/13219/> (дата обращения: 20 ноября 2017 г.).

6. **Щеглов А. Ю.** Компьютерная безопасность. Противодействие внутренним ИТ-угрозам. Часть 1. Угроза хищения данных с использованием мобильных накопителей. URL: <http://www.npp-itb.spb.ru/publications/16.html> (дата обращения: 20 ноября 2017 г.).

7. **Гридасов В.** Контроль привилегированных пользователей (PUM) — обзор мирового и российского рынка. URL: https://www.anti-malware.ru/reviews/privileged_user_management_market_russia_2016# (дата обращения: 20 ноября 2017 г.).

8. **Романов М.** Системы контроля привилегированных пользователей // *Information Security / Информационная безопасность*. 2015. № 4.

9. **Шабанов И.** Проблемы контроля привилегированных пользователей и их решение на примере Wallix AdminBastion. URL: <http://www.anti-malware.ru/node/12023#> (дата обращения: 20 ноября 2017 г.).

10. **Патент** на изобретение № 2534599 / Щеглов А. Ю., Щеглов К. А. Система контроля доступа к ресурсам компьютерной системы с субъектом доступа "пользователь, процесс".

11. **Щеглов А. Ю., Щеглов К. А.** Анализ и проектирование защиты информационных систем. Контроль доступа к компьютерным ресурсам. Методы, модели, технические решения. СПб.: Профессиональная литература, 2017.

12. **Патент** на изобретение № 2524566 / Щеглов А. Ю., Щеглов К. А. Система контроля доступа к файлам на основе их автоматической разметки.

T. S. Osadchaya, Engineer, e-mail: taniaosadchaya6@gmail.com, **A. Yu. Shcheglov**, Ph. D., Professor, ITMO University, St. Petersburg, Russia

Comprehensive Solution of Protection against Attack with Rights of Privileged Users

The work is dedicated to solvation of the problem of complex protection against attacks with privileged user rights. The following sources of this security threat are considered: legal users with privileged rights, and malicious programs running with privileged rights. The new approach of protection is described. It provides control and delineation of the actions of privileged users, including limitation of their administration capabilities, as well as excluding the possibility of the influence of malicious programs running with privileged rights on the system.

Keywords: privileged users, insiders, access control, access rights, limitation of users' administration capabilities, self-defense mechanism, malicious software

References

1. **Zhan-Nojel' de Gal'zan**, CEO WALLIX. Kompleksnaja politika bezopasnosti dolzhna vkljuchat' jeffektivnoe upravlenie privilegirovannymi pol'zovateljami. Available at: http://it-bastion.com/wp-content/uploads/2016/10/JN_interview_2013RU.pdf (accessed: 20 November 2017) (in Russian).

2. **Beljaeva M., Sinel'nikov A.** Zashhita ot insajderov — mif ili real'nost'? *Information Security*, 2008, no. 3, pp. 34—35 (in Russian).

3. **Ugroza** privilegirovannyh pol'zovatelej i metody po ejo ustraneniu. Available at: <http://ppt.ru/guide/news/136623> (accessed: 20 November 2017) (in Russian).

4. **Serdjuk V., Romanov M.** Obratnaja storona privilegij. *BIS JOURNAL — Informacionnaja bezopasnost' bankov*, 2017, no. 1, pp. 32—33 (in Russian).

5. **Kris Bruk.** 88 % setej ujazvimy k vzlomu privilegirovannyh akkauntov. Available at: <https://threatpost.ru/88-percent-of-networks-susceptible-to-privileged-account-hacks/13219/> (accessed: 20 November 2017) (in Russian).

6. **Shcheglov A. Yu.** Komp'juternaja bezopasnost'. Protivodejstvie vnutrennim IT-ugrozam. Chast' 1. Ugroza hishhenija dannyh s ispol'zovaniem mobil'nyh nakopitelej. Available at: <http://www.npp-itb.spb.ru/publications/16.html> (accessed: 20 November 2017) (in Russian).

[npp-itb.spb.ru/publications/16.html](http://www.npp-itb.spb.ru/publications/16.html) (accessed: 20 November 2017) (in Russian).

7. **Gridasov V.** Kontrol' privilegirovannyh pol'zovatelej (PUM) — obzor mirovogo i rossijskogo rynka. Available at: https://www.anti-malware.ru/reviews/privileged_user_management_market_russia_2016# (accessed: 20 November 2017) (in Russian).

8. **Romanov M.** Sistemy kontrolja privilegirovannyh pol'zovatelej. *Information Security*, 2015, no.4, pp. 18—19 (in Russian).

9. **Shabanov I.** Problemy kontrolja privilegirovannyh pol'zovatelej i ih reshenie na primere Wallix AdminBastion. Available at: <http://www.anti-malware.ru/node/12023#> (accessed: 20 November 2017) (in Russian).

10. **Patent** na izobretenie № 2534599. Shcheglov A. Yu., Shcheglov K. A. Sistema kontrolja dostupa k resursam komp'juternoj sistemy s sub#ektom dostupa "pol'zovatel", process". (in Russian).

11. **Shcheglov A. Yu., Shcheglov K. A.** Analiz i proektirovanie zashhity informacionnyh sistem. Kontrol' dostupa k komp'juternym resursam. *Metody, modeli, tehnicheckie reshenija*. 2017. St.-Petersburg: Professional'naja literatura (in Russian).

12. **Patent** na izobretenie № 2524566. Shcheglov A. Yu., Shcheglov K. A. Sistema kontrolja dostupa k fajlam na osnove ih avtomaticheskoy razmetki (in Russian).

УДК 519.7

Л. А. Лютикова, канд. физ.-мат. наук, зав. отделом, e-mail: lylarisa@yandex.ru,

Е. В. Шматова, мл. науч. сотр., e-mail: lenavsh@yandex.ru,

Институт прикладной математики и автоматизации — филиала Федерального государственного научного учреждения "Федеральный научный центр "Кабардино-Балкарский научный центр Российской академии наук" (ИПМА КБНЦ РАН), г. Нальчик

Логический подход к коррекции результатов работы ΣΠ-нейронных сетей¹

Рассматривается метод построения логического корректора для работы ΣΠ-нейронных сетей при решении задач распознавания. Предлагается метод обнаружения неявных закономерностей по структуре ΣΠ-нейрона при котором повышаются адаптивные свойства распознающей системы. Утверждается, что комбинированный подход к организации работы распознающей системы повышает ее эффективность и позволяет в случае возникновения некорректного ответа ΣΠ-нейрона в качестве решения указать наиболее близкие по запрашиваемым признакам объекты.

Ключевые слова: логический анализ, анализ данных, алгоритм, ΣΠ-нейрон, обучающая выборка, деревья решений, корректирующие операции

Введение

Распознавание образов — это целый раздел теоретической информатики, разрабатывающий принципы и методы классификации, а также идентификации предметов, явлений, процессов, сигналов, ситуаций — всех тех объектов, которые могут быть описаны конечным набором некоторых признаков или свойств, характеризующих объект [1].

Существует ряд методов для решения данных задач, каждый из которых обладает как своими преимуществами, так и своими недостатками [2].

Комбинированный подход применяется, когда требуется скорректировать работу нескольких различных алгоритмов, каждый из которых безошибочно классифицирует лишь часть обучаемых объектов. Цель коррекции — сделать так, чтобы ошибки одних алгоритмов были скомпенсированы другими и качество результирующего алгоритма оказалось лучше, чем каждого из базовых алгоритмов в отдельности [1, 3, 4].

На практике существуют различные подходы к построению комбинированных корректирующих алгоритмов, сочетание которых дает практически значимые результаты [4—9].

В данной работе рассматривается подход, основанный на логическом анализе данных, используемый для коррекции работы ΣΠ-нейронной сети. ΣΠ-нейрон (сигма-пи-нейрон) является обобщением классической модели формального нейрона с линейной функцией суммирования $sp(x_1, \dots, x_n)$ входных сигналов.

Можно предположить, что правила опознавания объекта ΣΠ-нейронная сеть содержит в весовых коэффициентах, но правила эти не явные, и определить причину ошибки бывает затруднительно. Построение корректора для осуществления контрольных и корректирующих функций работы ΣΠ-нейрона предлагается с использованием методов логического анализа, позволяющего по весовым коэффициентам восстановить обучающую выборку, провести ее анализ, построить базу знаний, минимизировать ее и в случае ошибочной работы ΣΠ-нейрона скорректировать ответ относительно построенных правил. В статье рассматривается метод обнаружения в структуре ΣΠ-нейрона ранее не известных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений.

¹ Работа выполнена при поддержке гранта РФФИ № 15-01-03381 и фундаментального научного проекта ОНИТ РАН.

Постановка задачи

Пусть $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in \{0, 1, \dots, k_i - 1\}$, где $k_i \in [2, \dots, N]$, $N \in \mathbb{N}$, — множество признаков. $Y = \{y_1, y_2, \dots, y_m\}$ — множество объектов, каждый объект y_i характеризуется соответствующим набором признаков $x_1(y_i), \dots, x_n(y_i) : y_i = f(x_1(y_i), \dots, x_n(y_i))$.

Или пусть $X = \{x_1, x_2, \dots, x_n\}$, где $x_i \in \{0, 1, \dots, k_r - 1\}$, $k_r \in [2, \dots, N]$, $N \in \mathbb{Z}$ — обрабатываемые входные данные $X_i = \{x_1(y_i), x_2(y_i), \dots, x_n(y_i)\}$, $i = 1, \dots, n$, $y_i \in Y$, $Y = \{y_1, y_2, \dots, y_m\}$ — выходные данные:

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & \dots & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & \dots & x_n(y_2) \\ \dots & \dots & \dots & \dots \\ x_1(y_m) & x_2(y_m) & \dots & x_n(y_m) \end{pmatrix} \rightarrow \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}.$$

Вид функции $Y = f(x)$ не задан.

Пусть $\Sigma\Pi$ -нейрон представлен следующей структурой:

$$\text{sp}(x_1, \dots, x_n) = \sum w_i \prod x_i,$$

где $\{w_1, w_2, \dots, w_k\}$ — множество весов данного $\Sigma\Pi$ -нейрона, распознающего k элементов заданной предметной области $Y = \{y_1, y_2, \dots, y_k\}$, сформированной соответствующим набором признаков $\{X_1, \dots, X_k\}$. И пусть существует набор признаков $X_l \notin \{X_1, \dots, X_k\}$ такой, что результат работы $\Sigma\Pi$ -нейрона на данном запросе некорректен.

Будем говорить, что результат работы $\Sigma\Pi$ -нейрона некорректен, если:

1) он не опознал соответствующий запросу элемент, т.е.

$$\text{sp}(X_i) \neq y_i;$$

2) он опознал объект, не принадлежащий данной предметной области, т.е.

$$\text{sp}(X_i) = y_l, y_l \notin \{y_1, \dots, y_n\};$$

3) при запросе $X_j \notin \{X_1, X_2, \dots, X_k\}$ найден элемент, принадлежащий предметной области, но в обучающей выборке существует элемент с большим числом совпадающих признаков, т.е. $\text{sp}(X_i) = y_l, y_l \in \{y_1, \dots, y_n\}$, и существует y_j с соответствующим вектором признаков x_j такой, что

$$\dim |X_i - X_j| < \dim |X_i - X_l|,$$

где $\dim |X_l - X_i|$ — число несовпадающих признаков; X_l — вектор признаков y_l .

Необходимо построить функцию, которая при ошибке работы $\Sigma\Pi$ -нейрона сможет подобрать

наиболее близкий элемент или класс элементов из обучающей выборки по заданным признакам:

$$F(X_l) : F(X_l) = y_i, y_i \in Y, \dim |X_l - X_i| \rightarrow \min.$$

Основные сведения о $\Sigma\Pi$ -нейронных сетях

$\Sigma\Pi$ -нейрон представляет собой *алгебраическую модель* нейрона, отражающую процессы обработки информации в аксо-дендритной системе нейрона [10–12], которая обладает лучшими способностями по аппроксимации зависимостей.

Один $\Sigma\Pi$ -нейрон в своем классическом варианте имеет полилинейную функцию суммарного сигнала

$$\text{sp}(X) = \text{out} \left(\theta + \sum w_k \prod_{i=1}^n x_i \right),$$

где θ — константа, out — функция вывода.

Обучить логико-арифметический $\Sigma\Pi$ -нейрон можно за один проход обучающей выборки.

В процессе обучения строится последовательность логико-арифметических $\Sigma\Pi$ -нейронов $\{\text{spn}_k(X)\}$, $\text{spn}(X) = \text{sngsp}(X)$. Процесс построения рекуррентный:

1) вначале $\text{sp}_0(X) = 0$,

2) на k -м шаге $\text{sp}_k(X) = \text{sp}_{k-1}(X) + w_k \prod_{i \in \mathbf{i}_k^0} x_i$, где $\mathbf{i}_k^0 = \{i : x_{ki} = 1\}$. Вес w_k находится следующим образом:

$$w_k = \begin{cases} s_k - \text{sp}_{k-1}(X_k), & \text{если } \text{spn}_{k-1}(X_k) \neq y_k; \\ 0, & \text{если } \text{spn}_{k-1}(X_k) = y_k, \end{cases}$$

где s_k — произвольное значение, такое что $\text{sp}(s_k) = y_k$.

После N шагов искомый логико-арифметический $\Sigma\Pi$ -нейрон построен. Этот очень простой алгоритм обучения в случае, когда out — пороговая функция, был первоначально предложен Тимофеевым А. В. [13].

ПРИМЕР 1. Пусть в виде таблицы задана следующая предметная область:

x_1	x_2	x_n	y
0	0	1	2
0	1	0	4
0	1	1	8
1	1	1	128

Множество признаков X представлено следующими значениями:

$$X = \{X_1 = (0, 0, 1), X_2 = (0, 1, 0), \\ X_3 = (0, 1, 1), X_4 = (1, 1, 1)\},$$

а множество объектов $Y = \{2, 4, 8, 128\}$.

В результате обучения по данным таблицы ΣП-нейрон будет иметь вид

$$\text{sp}(x_1, x_2, x_3) = 4x_2 + 2x_3 + 2x_2x_3 + 120x_1x_2x_3.$$

Введем запрос (1, 1, 0), который в точности не совпадает ни с одним из заданных наборов. Результат обученного ΣП-нейрона: $\text{sp}(1, 1, 0) = 4$. Действительно, объект "4" из множества Y совпадает с запросом в двух точках, но в двух точках с запросом совпадает и объект "128", причем этот объект выделяется во всей выборки значением переменной x_1 . Корректировку подобных результатов работы ΣП-нейрона можно провести с помощью логического анализа данных.

Логический подход к исследованию обучающей выборки построения решающей функции

В данном разделе рассматриваются логические методы анализа данных. Предлагается метод построения логической функции классификатора, рассматриваются ее свойства. Анализируются получаемые результаты [14, 15].

Данные, с которыми приходится иметь дело при решении задач распознавания, как известно, являются неполными, неточными, неоднозначными. Однако получаемые решения должны соответствовать закономерностям, явно и неявно присутствующим в рассматриваемой предметной области. Использование логических методов позволяет достаточно хорошо проанализировать данные, выделить существенные и несущественные признаки, выявить минимальный набор правил, необходимый для того, чтобы полностью восстановить исходные закономерности. В результате можно получить более компактное и надежное представление исходной информации, которая обрабатывается надежнее и быстрее.

Будем говорить, что построенная система правил является полной, если она обеспечивает вывод всех решений в рассматриваемой области.

Группу объектов, выделенных по определенному признаку (группе признаков), будем называть классом.

Каждый объект может быть представителем одного или нескольких классов, каждый класс определяется набором однотипных признаков.

Для объяснения логической связи между понятиями предметной области будем использовать правило продукции, предложенное Э. Постом. В нашем случае правило продукции представляет собой подстановку следующего вида:

$$x_1(y_i) \& x_2(y_i) \& \dots \& x_n(y_i) \rightarrow P(y_i),$$

где $P(y) = \begin{cases} 1, & y = y_i; \\ 0, & \text{иначе,} \end{cases}$
 $x_1(y_i), x_2(y_i), \dots, x_n(y_i)$ — конечное число признаков, характеризующих элемент y_i .

Это позволяет выразительно представить зависимости между объектом и его признаками:

$$\bigwedge_{j=1}^m x_j(y_i) \rightarrow P(y_i), i = 1, \dots, l; x_j(y_i) \in \{0, 1, \dots, k-1\},$$

где предикат $P(y_i)$ принимает значение "истина", т.е. $P(y_i) = 1$ в случае, если $y = y_i$, и $P(y_i) = 0$, если $y \neq y_i$. Или

$$\bigvee_{j=1}^n \bar{x}(y_j) \vee P(y_j), j \in [1, \dots, m].$$

Решающей функций назовем конъюнкцию всех решающих правил:

$$\bigwedge_{j=1}^m x_j(y_i) \rightarrow P(y_i), i = 1, \dots, l; x_j(y_i) \in \{0, 1, \dots, k-1\},$$

или $f(X) = \bigwedge_{j=1}^m \left(\bigvee_{i=1}^n \bar{x}_i(y_j) \vee P(y_j) \right).$ (1)

Функцию (1) можно интерпретировать следующим образом.

Если обучающую выборку, состоящую из k элементов, описать булевой функцией

$$F(x_1(y_i), \dots, x_n(y_i), P^\sigma(y_1), \dots, P^\sigma(y_n)),$$

где

$$P^\sigma(y_i) = \begin{cases} \overline{P(y_i)} & \text{при } \sigma = 0; \\ P(y_i) & \text{при } \sigma = 1, \end{cases}$$

то данная функция принимает значения "0" на наборах $(x_1(y_i), \dots, x_n(y_i), P^\sigma(y_1), \dots, \overline{P(y_i)}, \dots, P^\sigma(y_n))$ и "1" на всех остальных наборах, т.е. она допускает любые отношения между признаками и объектами, кроме отрицания объекта на множестве характеризующих этот объект признаков в обучающей выборке.

Функция (1) выражает зависимость между характеристиками объекта и самим объектом, находит все возможные классы в заданной области, вплоть до классов объектов, объединенных по единственному признаку, допускает включение новых правил продукции (модифицируема), при вводе значения любого объекта из исследуемой области определяет этот объект. Если вводимые данные точно не определены в области, по которой была построена функция, то функция определяет наиболее подходящий объект или классы объектов, идентифицируемые по части входных данных.

Все свойства функции (1) подробно рассмотрены в работе [16].

Поскольку функция — это дизъюнкция конъюнкций разной длины переменных, она может быть подвергнута сокращению.

Логическое описание класса K_j — это дизъюнкт решающей функции (1) состоящий из набора предикатов, характеризующих наличие или отсутствие какого-либо элемента, и переменных, характеризующих общие признаки для этих элементов.

В результате функция (1) будет состоять из переменных, сочетание которых не является характеристикой каких-либо классов или отдельных объектов, и объектной части (дизъюнктов), которые содержат предикаты объектов. Алгоритм построения объектной части решающей функции описан в работе [16].

Решающая функция для примера 1 будет иметь следующий вид:

$$\begin{aligned} f(X) &= (\overline{x_1} \vee \overline{x_2} \vee \overline{x_3} \vee P(128)) \& \\ &\& (x_1 \vee \overline{x_2} \vee \overline{x_3} \vee P(8)) \& \\ &\& (x_1 \vee x_2 \vee \overline{x_3} \vee P(2)) \& (x_1 \vee \overline{x_2} \vee x_3 \vee P(4)) = \\ &= \overline{x_1} \overline{x_2} \vee \overline{x_1} \overline{x_2} \vee \overline{x_1} \overline{x_3} \vee \overline{x_3} P(4) \vee \overline{x_2} P(2) \vee x_1 P(128) \vee \\ &\vee P(128) P(8) P(2) x_3 \vee P(128) P(8) P(4) x_2 \vee \\ &\vee P(4) P(8) P(2) \overline{x_1}. \end{aligned}$$

Объектная часть функции представима следующими дизъюнктами:

$$\begin{aligned} f(X) &= \overline{x_3} P(4) \vee \overline{x_2} P(2) \vee x_1 P(128) \vee \\ &\vee P(128) P(8) P(2) x_3 \vee \\ &\vee P(128) P(8) P(4) x_2 \vee P(4) P(8) P(2) \overline{x_1}. \end{aligned}$$

При запросе, где вектор признаков представлен значениями (1, 1, 0), логическая функция выдаст следующий результат: $f(1, 1, 0) = P(4) \vee P(128)$.

В отличие от нейронной сети она в качестве ответа предложит объект "4" и объект "128".

В результате применения данного алгоритма, как и функции (1), мы получаем не единственный ответ, а возможные варианты, подходящие для данного запроса по соответствующим признакам. Анализ полученных результатов может быть подвергнут частотному анализу, процедуре взвешенного голосования или оценке эксперта. Эксперт может установить несколько порогов и соответственно ранжировать полученные закономерности по степени достоверности.

Выделение правил из обученной нейронной сети

В данной работе рассматривается $\Sigma\Pi$ -нейронная сеть, результатом обучения которой будет являться полином вида

$$\text{sp}(x_1, \dots, x_n) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_p x_1 x_2 \dots x_n,$$

где $\{w_1, w_2, \dots, w_k\}$ — множество весов данного $\Sigma\Pi$ -нейрона.

Обучающая выборка, по которой проводилась настройка весов, вообще говоря, может быть неизвестна. Требуется восстановить обучающую выборку, обнаружить логические закономерности и использовать их для корректировки результата работы исходного $\Sigma\Pi$ -нейрона.

Восстановление обучающей выборки по полиному — это, по сути, процедура, обратная обучению.

На нижнем уровне находятся переменные $\{x_1, x_2, \dots, x_n\}$. Все коэффициенты полинома (веса $\Sigma\Pi$ -нейрона) разбиваются на уровни в зависимости от числа переменных, входящих вместе с ним в слагаемое. То есть на первом слое веса с наименьшее число переменных, на втором соответственно на одну переменную больше и т.д., в верхнем слое находятся все коэффициенты с наибольшим числом переменных. Элемент каждого верхнего уровня связан с соответствующими элементами нижнего слоя по общим переменным.

Веса первого слоя $\{w_1, w_2, \dots, w_r\}$ соответственно и есть объекты $\{y_1, y_2, \dots, y_r\}$ на каждом следующем слое $y_{k+1} = w_{k+1} + \sum y_i$, где i — индексы соответствующих объектов, переменные которых входят в качестве сомножителей в элемент на y_{k+1} .

ПРИМЕР 2. Пусть обученный $\Sigma\Pi$ -нейрон имеет следующий вид

$$\text{sp}(x_1, x_2, x_3) = 4x_2 + 2x_3 + 2x_2x_3 + 120x_1x_2x_3.$$

Восстановим объекты обучающей выборки и найдем обобщающие логические правила (рис. 1).

Построение неявных правил. Для построения по дереву неявных правил рассмотрим путь из каждой вершины y_k к каждой переменной x_i . Если такой путь существует, то множество вершин принадлежащих соответствующему пути, т.е. множество объектов, определяемых переменной, образуют класс по этой переменной x_i . Та-

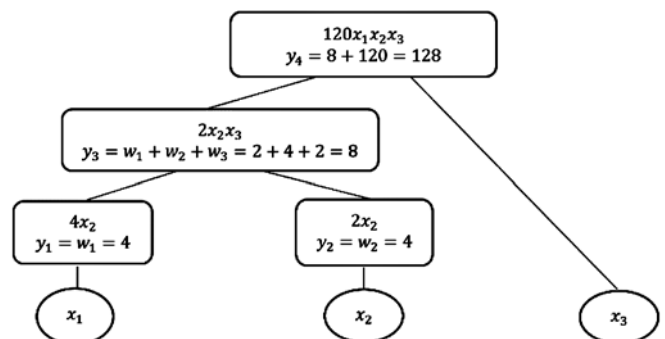


Рис. 1

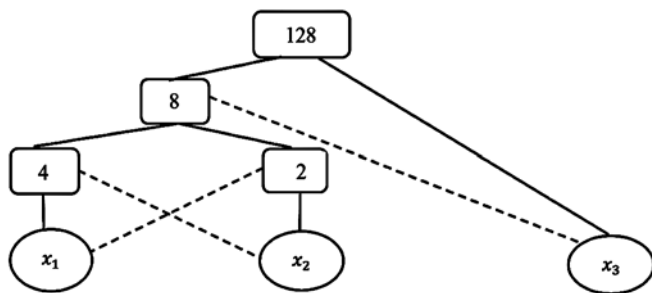


Рис. 2

кой путь можно представить в виде логической формулы с помощью операции конъюнкции, т.е.

$$P(y_k) \& P(y_{k-1}) \& \dots \& P(y_i) \& x_i.$$

Если пути от заданной вершины y_k к вершине x_i не существует, то данный путь (класс объектов) определяется отрицанием соответствующей переменной. Совокупность всевозможных путей, как существующих, так и несуществующих, представима в виде дизъюнкции (рис. 2). Таким образом, можно полностью восстановить объектную часть логической функции (1).

Штриховой линией обозначены связи с отрицанием переменной.

$$\begin{aligned} f(X) = & \overline{x_3}P(4) \vee \overline{x_2}P(2) \vee x_1P(128) \vee \\ & \vee P(128)P(8)P(2)x_3 \vee P(128)P(8)P(4)x_2 \vee \\ & \vee P(4)P(8)P(2)\overline{x_1}, \end{aligned}$$

что полностью совпадает с объектной частью функции (1).

ТЕОРЕМА. Пусть $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in \{0, 1, \dots, k_i - 1\}$, где $k_i \in [2, \dots, N]$, $N \in \mathbb{N}$ — множество признаков. $Y = \{y_1, y_2, \dots, y_m\}$ — множество объектов, каждый объект y_i характеризуется соответствующим набором признаков $x_1(y_i), \dots, x_n(y_i)$, и пусть $\text{sp}(X) = Y$, тогда по соответствующей структуре $\text{sp}(X)$ можно получить все возможные решения на заданном пространстве данных.

Доказательство. Как было показано выше, по структуре нейрона можно восстановить обучающую выборку, для нее возможно построить логическую решающую функцию (1), которая строит полную систему правил для заданной предметной области.

Предложенная процедура построения неявных правил упрощает задачу.

Поскольку с помощью логической функции (1) можно провести полный анализ обучающей выборки и найти все возможные связи, то объектная часть логической функции может выступить в качестве корректора для $\Sigma\P$ -нейрона. В соответствии с предложенным выше алгорит-

мом корректор можно построить, опираясь на структуру самого $\Sigma\P$ -нейрона, даже не имея обучающей выборки.

В окончательном виде модель $\Sigma\P$ -нейрона вместе с его корректором имеет следующий вид:

$$Y = P(\text{sp}(X)) \vee f(X),$$

где $f(X)$ является корректором работы нейрона. Тогда наилучшее решение можно выбрать между результатом работы нейрона и корректором. Если ответ нейрона был правильным и объект, ранее не принадлежал обучающей выборке, то в силу простой модифицируемости функции $f(X)$, новое правило будет принято корректором и расширит базу знаний.

Заключение

В работе рассмотрены вопросы поиска логических закономерностей по структуре обученного $\Sigma\P$ -нейрона. Предложено совместное использование нейросетевых технологий и методов логического вывода как средства выявления логических закономерностей и представления более точного результата в задачах распознавания. Также рассмотрены преимущества логического анализа данных и построена логическая функция для выявления всех закономерностей исследуемой предметной области. Предложена процедура построения деревьев решений на основе обученного $\Sigma\P$ -нейрона, которая не предъявляет никаких требований к архитектуре, алгоритму обучения, входным и выходным значениям и другим параметрам сети. Построение дерева осуществляется по структуре нейрона, результат выявляет ряд скрытых логических закономерностей данных, строит логическую функцию, которая корректирует работу нейрона в случае неточных зашумленных данных, позволяя указать наиболее правдоподобные (близкие к эталонным) ответы в рамках сделанного запроса, а также легко модифицировать в случае правильного ответа нейрона, не предусмотренного обучающей выборкой. В результате существенно повышается качество автоматизированного решения интеллектуальных задач, надежность их функционирования, точность достижения верного решения за счет использования наиболее эффективных систем анализа исходных данных и разработки более точных методов их обработки.

Список литературы

1. Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5—68.

2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: МДК Пресс, 2015. 400 с.
3. Воронцов К. В. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // Журнал вычислительной математики и математической физики. 2000. Т. 40, № 1. С. 166—176.
4. Абламейко С. В., Бирюков А. С., Докукин А. А., Дьяконов А. Г., Журавлев Ю. И., Краснопрошин В. В., Образцов В. А., Романов М. Ю., Рязанов В. В. Практические алгоритмы алгебраической и логической коррекции в задачах распознавания по прецедентам // Журнал вычислительной математики и математической физики. 2014. Т. 54. № 12. С. 1979.
5. Тимофеев А. В., Косовская Т. М. Нейросетевые методы логического описания и распознавания сложных образов // Труды СПИИРАН, 27 (2013). С. 144—155.
6. Дюкова Е. В., Журавлев Ю. И., Прокофьев П. А. Методы повышения эффективности логических корректоров // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1555—1583.
7. Гридин В. Н., Солодовников В. И., Евдокимов И. А., Филиппов С. В. Построение деревьев решений и извлечение правил из обученных нейронных сетей // Искусственный интеллект и принятие решений. 2013. № 4. С. 26—33.
8. Zhiting Hu, Xueze Ma, Zhengzhong Liu, Eduard Hovy, Eric Xing. Harnessing Deep Neural Networks with Logic Rules // Computer Science Learning. 2016. P. 2410—2420. arXiv: 1603.06318.

9. Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka GrabskaBarwinska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 2016. № 538 (7626). P. 471—476.
10. Шибзухов З. М. О поточно корректных операциях над алгоритмами распознавания и прогнозирования // Доклады РАН. 2013, Т. 450, № 1. С. 24—27.
11. Shibzhukhov Z. M. Correct Aggregation Operations with Algorithms // Pattern Recognition and Image Analysis. 2014, Vol. 24, N. 3. P. 377—382.
12. Шибзухов З. М. О некоторых конструктивных и корректных классах алгебраических СП-алгоритмов // Доклады РАН, 2010. Т. 432, № 4. С. 465—468.
13. Тимофеев А. В., Пшибихов В. Х. Алгоритмы обучения и минимизации сложности полиномиальных распознающих систем // Известия АН СССР. Техническая кибернетика. 1974. № 7. С. 214—217.
14. Тимофеев А. В., Лютикова Л. А. Развитие и применение многозначных логик и сетевых потоков в интеллектуальных системах // Труды СПИИ РАН. Вып. 2, 2005. С. 114—126.
15. Лютикова Л. А., Шматова Е. В. Анализ и синтез алгоритмов распознавания образов с использованием переменнo-значной логики // Информационные технологии. 2016. Т. 22, № 4. С. 292—297.
16. Лютикова Л. А. Использование математической логики с переменной значностью при моделировании систем знаний // Вестник Самарского государственного университета. Естественная серия. 2008. № 6 (65). С. 20—27.

L. A. Lyutikova, Head of the Department, e-mail: lylarisa@yandex.ru,

E. V. Shmatova, Junior Researcher, e-mail: lenavsh@yandex.ru,

Institute of Applied Mathematics and Automation of Kabardin-Balkar Scientific Centre of RAS
(IAMA KBSC RAS), Nalchik

A Logical Approach to the Correction of the Results of the Operation of $\Sigma\Pi$ -Neural Networks

The paper considers the method of constructing a logical corrector for the operation of $\Sigma\Pi$ -neural networks when solving recognition problems. A method is proposed for detecting implicit regularities, according to the structure of the $\Sigma\Pi$ -neuron, which can enhance the adaptive properties of the recognition system. It is argued that the combined approach to the organization of the recognition system increases its effectiveness and allows in cases of an incorrect response of the $\Sigma\Pi$ -neuron as a solution to indicate the objects closest to the requested attributes.

Keywords: logical analysis, data analysis, algorithm, $\Sigma\Pi$ -neuron, training sample, decision trees, corrective operations

References

1. Zhuravljov Ju. I. Ob algebraicheskom podhode k resheniju zadach raspoznavaniya ili klassifikacii, *Problemy kibernetiki*, 1978, vol. 33, pp. 5—68 (in Russian).
2. Flah P. *Mashinnoe obuchenie. Nauka i isskustvo postroeniya algoritmov, kotorye izvlekajut znaniya iz dannyh*, Moscow, MDK Press, 2015. 400 p. (in Russian).
3. Voroncov K. V. Optimizacionnye metody linejnoy i monotonnoj korrekcii v algebraicheskom podhode k probleme raspoznavaniya, *Zhurnal vychislitel'noj matematiki i matematicheskoy fiziki*, 2000, vol. 40, no. 1, pp. 166—176 (in Russian).
4. Ablamejko S. V., Birjukov A. S., Dokukin A. A., D'jakonov A. G., Zhuravlev Ju. I., Krasnoproschin V. V.,

- Obrazcov V. A., Romanov M. Ju., Rjazanov V. V. Prakticheskie algoritmy algebraicheskoy i logicheskoy korrekcii v zadachah raspoznavaniya po precedentam, *Zhurnal vychislitel'noj matematiki i matematicheskoy fiziki*, 2014, vol. 54, no. 12, pp. 1979. (in Russian).
5. Timofeev A. V., Kosovskaja T. M. Nejrosetevye metody logicheskogo opisaniya i raspoznavaniya slozhnyh obrazov, *Trudy SPIIRAN*, 2013, 27, pp. 144—155 (in Russian).
6. Djukova E. V., Zhuravlev Ju. I., Prokof'ev P. A. Metody povysheniya jeffektivnosti logicheskikh korrektorov, *Mashinnoe obuchenie i analiz dannyh*, 2015, vol. 1, no. 11, pp. 1555—1583 (in Russian).
7. Gridin V. N., Solodovnikov V. I., Evdokimov I. A., Filipkov S. V. Postroenie derev'ev reshenij i izvlechenie pravil iz obuchennyh nejronnyh setej, *Iskusstvennyj intellekt i prinjatje reshenij*, 2013, no. 4, pp. 26—33.

8. Zhiting Hu, Xueze Ma, Zhengzhong Liu, Eduard Hovy, Eric Xing. Harnessing Deep Neural Networks with Logic Rules, *Computer Science Learning*, 2016. P. 2410—2420. arXiv: 1603.06318.
9. Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka GrabskaBarwinska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou et al. Hybrid computing using a neural network with dynamic external memory, *Nature*, 2016, no. 538 (7626), pp. 471—476.
10. Shibzuho Z. M. O potochetchno korrektnyh operacijah nad algoritmami raspoznavanija i prognozirovanija, *Doklady RAN*, 2013, vol. 450, no. 1, pp. 24—27 (in Russian).
11. Shibzuchov Z. M. Correct Aggregation Operations with Algorithms, *Pattern Recognition and Image Analysis*, 2014, vol. 24, no. 3, pp. 377—382.
12. Shibzuho Z. M. O nekotoryh konstruktivnyh i korrektnyh klassah algebraicheskikh $\Sigma\Pi$ -algoritmov, *Doklady RAN*, 2010, vol. 432, no. 4, pp. 465—468 (in Russian).
13. Timofeev A. V., Pshibihov V. H. Algoritmy obuchenija i minimizacii slozhnosti polinomial'nyh raspoznajushhih sistem, *Izvestija AN SSSR. Tehniceskaja kibernetika*, 1974, no. 7, pp. 214—217 (in Russian).
14. Timofeev A. V., Ljutikova L. A. Razvitie i primenenie mnogoznachnyh logik i setevykh potokov v intellektual'nyh sistemah, *Trudy SPII RAN*, 2005, vyp. 2, pp. 114—126 (in Russian).
15. Ljutikova L. A., Shmatova E. V. Analiz i sintez algoritmov raspoznavanija obrazov s ispol'zovaniem peremennno-znachnoj logiki, *Informacionnye tehnologii*, 2016, vol. 22, no. 4, pp. 292—297 (in Russian).
16. Ljutikova L. A. Ispol'zovanie matematicheskoj logiki s peremennoj znachnost' pri modelirovanii sistem znaniy, *Vestnik Samarskogo gosudarstvennogo universiteta. Estestvennonauchnaja serija*, 2008, no. 6 (65), pp. 20—27 (in Russian).

УДК 004.93

С. В. Куликов, науч. сотр., e-mail: kulikov@deepmark.ru,
О. С. Захаров, науч. сотр., e-mail: zakharov@deepmark.ru,
Д. Ю. Андреев, ген. директор, e-mail: andreev@deepmark.ru,
ООО "Лаборатория умных технологий", г. Пенза

Исследование возможности совместного применения нейросетевого преобразователя биометрия—код и глубокой сверточной нейронной сети в распознавании лиц¹

Проводится анализ возможности использования нейросетевых преобразователей биометрия—код (НПБК), отвечающих требованиям серии стандартов ГОСТ Р 52633, в задаче извлечения стабильного ключа из изображения лица. НПБК используется в качестве последнего слоя заранее обученной глубокой сверточной нейросетевой модели. По методике, соответствующей ГОСТ Р 52633.1—2009, проводится оценка показателей стабильности, уникальности и качества параметров, получаемых на выходе глубокой нейросетевой модели, для анализа возможности обучения НПБК на выходных параметрах глубокой нейросетевой модели. Проведено тестирование ряда конфигураций НПБК (соотношение числа нейронов и числа входов каждого нейрона), выбранных согласно ГОСТ Р 52633.5—2011, и сравнение полученных ROC-кривых с аналогичными кривыми, полученными для шаблонов на базе Евклидова расстояния и машин опорных векторов.

Ключевые слова: распознавание лиц, криптографический ключ, глубокая сверточная нейронная сеть, нейросетевой преобразователь биометрия—код, НПБК, машина опорных векторов, Евклидово расстояние, показатель уникальности, показатель стабильности, показатель качества, ROC-кривая

Введение

Глубокие сверточные нейронные сети в настоящий момент не имеют конкурентов в области распознавания лиц. Тем не менее, проблема извлечения стабильного ключа с высокой энтропией из изображения лица остается открытой, хотя необходимость использования глубокой нейронной сети в этой задаче и не вызывает сомнений.

Глубокие сверточные нейронные сети — специальный класс многослойных нейронных се-

тей, который наиболее широко применяется для распознавания изображений. Такие сети состоят из большого числа слоев, обычно — несколько десятков, и содержат среди прочих слоев, применяющие к входным данным операцию свертки.

Существует класс задач, в которых требуется преобразовывать биометрический образ в код, обладающий свойствами криптографического ключа (*biometric encryption*). Возможность получения криптографического ключа из биометрических данных позволяет обеспечивать дополнительную безопасность и обезличенность при организации биометрических систем. В частности, в такой системе база персональных данных, в том числе биометрических, может быть орга-

¹ Работа выполнена при финансовой поддержке Фонда содействия инновациям (договор № 1554ГСИ/24419).

низована таким образом, чтобы обеспечить обезличенность путем шифрования всех данных на извлекаемом из биометрических данных ключе. Утечки из такой базы менее опасны, чем из базы с открытым хранением персональных данных (или с шифрованием на мастер-ключе).

Биометрические данные лица человека не исключение, и их тоже используют для получения подобных ключей [1]. Само собой, технология преобразования изображения лица в код более других подвержена атакам с использованием муляжей (фотографий и видеозаписей) [2]. Имея фотографию атакуемого пользователя и зашифрованные персональные данные, злоумышленник имеет возможность расшифровать их, если технология не предусматривает защиту от фотографии на уровне преобразования в код. Теоретическая возможность такой защиты существует, но выходит за рамки этой статьи. В любом случае технология преобразования биометрических данных лиц в код имеет право на существование как минимум в виде составной части мультибиометрических систем.

В России технология преобразования биометрических данных в код попадает под определение "средства высоконадежной биометрической аутентификации" и регламентируется серией стандартов ГОСТ Р 52633. Задачей данной работы является анализ возможности совместного применения глубокой сверточной нейронной сети — де-факто стандарта для распознавания лица, и нейросетевого преобразователя биометрия—код по ГОСТ Р 52633.0—2006 [3]. Критерием возможности совместного применения является снижение/увеличение дифференцирующей способности модели при использовании НПБК, определяемое по ROC-кривой (*receiver operating characteristic*), в сравнении с машиной опорных векторов и шаблонами на базе Евклидова расстояния. Для построения ROC-кривых проводится численное моделирование работы всех трех вышеупомянутых технологий на тестовой базе лиц объемом 454736 примеров для 10575 субъектов.

1. Архитектура сетей

В качестве глубокой сверточной нейронной сети была выбрана относительно компактная модель со слоями "Max Feature Map" [4]. Архитектура сети была выбрана с тем расчетом, чтобы производительность сети позволяла использовать ее для работы на встраиваемых контроллерах низкой производительности (что является одним из основных применений схем *biometric encryption*) и чтобы получить наиболее практически значимые результаты.

Сеть имеет пять групп слоев, по пять слоев в каждой группе, слой регуляризации *dropout* и слой классификации *softmax* на выходе. Сеть

принимает на вход черно-белое изображение лица размером 128×128 пикселей. Перед подачей на глубокую сверточную нейронную сеть лицо проходит процедуру выравнивания по особым точкам (два угла рта, кончик носа, два центра глаз). Для обнаружения лиц на кадрах/фотографиях как в рамках обучения глубокой модели, так и в рамках исследований, описываемых в этой работе, использовали каскадный нейросетевой детектор, аналогичный MtCNN [5]. Выравнивание лица осуществляли по горизонтальной линии глаз, масштаб выбирали пропорционально расстоянию между горизонтальной линией глаз и центром рта (рис. 1, см. третью сторону обложки), положение точки, соответствующей центру рта, не было ограничено вертикальной линией центра изображения.

Для обучения модели была использована собственная база лиц объемом 1,5 млн изображений для 30 тысяч субъектов. Для регуляризации модели при обучении выровненное изображение лица произвольно смещали на ± 8 пикселей по обеим осям и была установлена вероятность выключения нейрона на слое *dropout*, равная 70 %. Полученная за ~ 2 млн итераций сеть имеет показатель EER (вероятность верной классификации при одинаковой вероятности ошибок), равный 98,8 % на базе LFW [6].

Нейросетевые преобразователи биометрия—код в соответствии с ГОСТ Р 52633.0—2006 в качестве входных данных используют "частично случайный вектор входных биометрических параметров", в концепцию которого отлично вписывается дескриптор, получаемый на выходе (точнее, на слое, предыдущем по отношению к слою *softmax*) глубокой нейронной сети. Таким образом, НПБК представляет собой некий аналог шаблона машины опорных векторов, применяемой к выходному дескриптору глубокой нейросетевой модели. НПБК обучается на нескольких примерах образа "Свой" и заданном коде, после чего теоретически должен выдавать максимально случайные выходные коды для образов "Чужой" и заданный код для образов "Свой". Обучение НПБК, а именно — соотношение числа нейронов и числа входов каждого нейрона, процедуру вычисления весов нейронной сети регламентирует ГОСТ Р 52633.5—2011.

В приложении А ГОСТ Р 52633.0—2006 рекомендуется для технологии распознавания лица выбирать длину выходного кода НПБК, равной 7...14 бит. То есть распознавание лица считается слабой технологией, не способной обеспечить достаточно низкую вероятность ошибок второго рода (FPR, *false positive rate*) при приемлемом уровне ошибок первого рода (FNR, *false negative rate*). Это объясняется тем, что ГОСТ Р 52633.0—2006 был написан в 2006 г., когда технологии распознавания лиц были на несколько порядков сла-

бее современных, и с тех пор это приложение стандарта не актуализировалось. Для проведения исследований число нейронов и соответственно длина выходного кода НПБК была выбрана равной 32 и 128 бит. По мнению авторов статьи, современные технологии распознавания лица по вероятностям ошибок находятся на уровне технологий распознавания отпечатка пальца, поэтому целесообразно отойти от устаревших требований стандарта и использовать длину ключа в 32 бита, которая согласно приложению А ГОСТ Р 52633.0—2006 может быть использована для технологии распознавания отпечатка пальца. Длина кода в 128 бит согласно стандарту может применяться только для технологий, содержащих тайный образ (секретную фразу или подпись), и была выбрана для анализа поведения НПБК в условиях, не соответствующих требованиям стандарта, но имеющих определенное практическое применение (криптографические ключи 128 бит могут считаться надежными в определенных условиях, а ключи 7, 14 или 32 бит — определенно нет).

2. Исследование НПБК

Тестирование НПБК проводили на дескрипторах, полученных из выровненных лиц базы CASIA-WebFace. Использовалась очищенная база, содержащая 454736 лиц 10575 субъектов. Зависимость типа *ROC* вероятности действительного положительного решения (TPR, *True Positive Rate*) от ложно-положительного решения (FPR, *False Positive Rate*) для тестовой базы CASIA-WebFace показана на рис. 2.

Согласно приведенной в ГОСТ Р 52633.1—2009 [8] методике расчета для используемой базы был проведен анализ стабильности, уникальности и качества биометрических параметров, используемых в качестве входных параметров НПБК. Гистограммы распределения образов по классам средней стабильности, уникальности и качеству параметров представлены на рис. 3.

Согласно проведенному анализу входных биометрических параметров, показатель среднего качества является низким, однако согласно ГОСТ Р 52633.5—2011 [2], НПБК способен обучаться на низкокачественных входных параметрах при использовании достаточного числа входов у каждого нейрона. Изменяя число входов

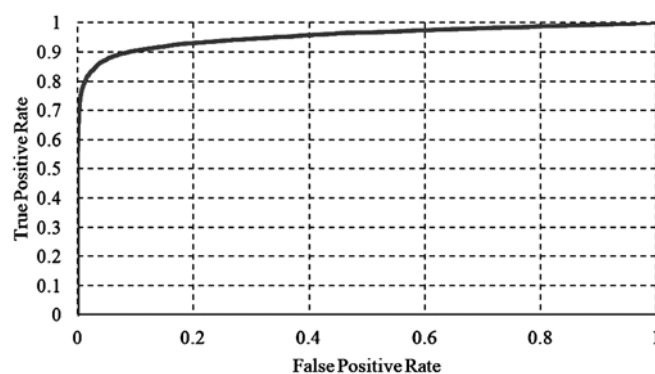


Рис. 2. ROC-кривая для используемой глубокой нейронной сети на базе LFW

у нейрона, можно регулировать значение выходного качества данного нейрона и, следовательно, регулировать вероятность появления ошибок первого (P_1) и второго рода (P_2). Чем выше выходное качество нейрона, тем ниже вероятность появления ошибок P_1 и P_2 .

Согласно ГОСТ Р 52633.5—2011 НПБК может использовать однослойную, либо двухслойную нейронную сеть. Для проведения исследования была выбрана однослойная нейронная сеть с 32 и 128 нейронами. Число входов у нейронов изменялось от 6 до 48. Изменением числа входов у нейронов осуществляется выбор рабочей точки на ROC-кривой для НПБК в целом.

Для выбора оптимальной конфигурации сети и исследования зависимости появления ошибок первого и второго рода была проведена серия экспериментов по обучению и тестированию обученной нейронной сети. На каждом шаге задавали число входов у нейрона, генерировалась таблица связей и выполнялось обучение однослойной нейронной сети из 32 или 128 нейронов. Обучение выполнялось для 100 первых субъектов в базе CASIA-WebFace; таким образом, на каждом шаге выполнялось обучение 100 однослойных нейронных сетей. Во время тестирования вычисляли чувствительность алгоритма (TPR) и специфичность алгоритма (FPR). В терминах ГОСТ Р 52633.5—2011 TPR соответствует $(1-P_1)$, FPR соответствует P_2 . Для вычисления TPR использовали примеры образа "Свой", не участвующие в обучении. Вычисление FPR выполняли на всех образах базы за исключением обучающе-

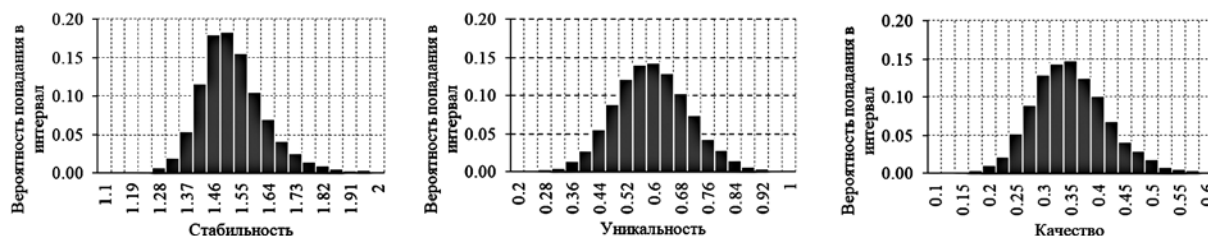


Рис. 3. Распределение образов "Свой" по классам средней стабильности, уникальности, качества параметров

го образа. Зависимость TPR от FPR для разных конфигураций сети (числа входов нейронов) показана на рис. 4 (см. третью сторону обложки).

На рис. 4 показаны четыре группы точек, каждая группа показывает значения TPR и FPR для 100 обучающих образов. Из рис. 4 видно, что при увеличении числа входов нейронов от 12 до 48 происходит увеличение значений TPR и FPR.

На рис. 5 приведена кривая изменения средних значений TPR и FPR для различных конфигураций сети. Точками показаны значения TPR и FPR для различных однослойных нейронных сетей при увеличении числа входов нейронов сети от 12 до 48.

Проведенные эксперименты показали, что для распознавания лиц предпочтительнее использовать сеть, содержащую не 32, а 128 нейронов: сеть с такой конфигурацией показывает лучшую дифференцирующую способность.

Для оценки точности работы НПБК в качестве шаблона аналогичная кривая была построена для машин опорных векторов (SVM) и шаблона, основанного на использовании Евклидова расстояния.

НПБК принимает решение о принадлежности примера образу "Свой", если выходной код совпадает с исходным кодом. Для обучения SVM в качестве негативов (термин "Все Чужие" по ГОСТ Р 52633.0—2006) использовалась база LFW [6]. Использовался не бинарный выход SVM,

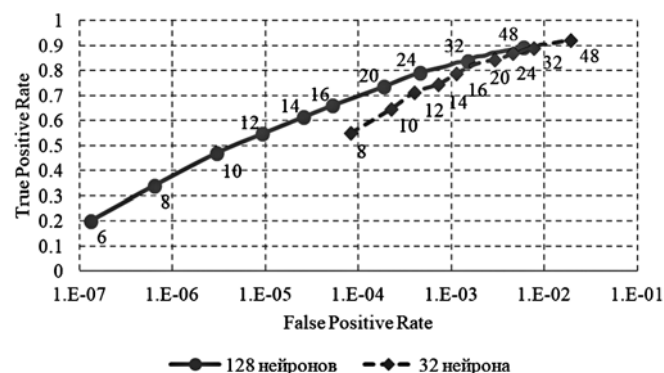


Рис. 5. ROC-кривая изменения средних значений TPR и FPR для различных конфигураций сети

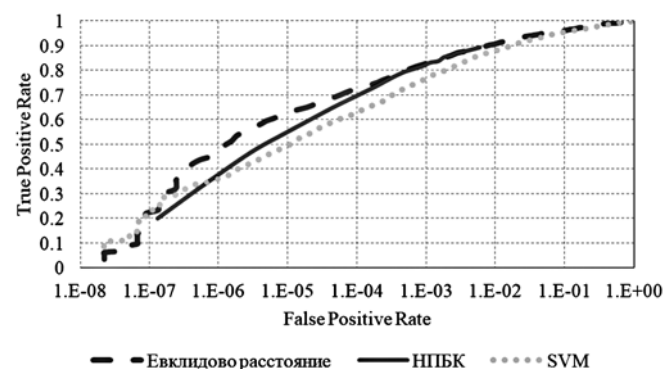


Рис. 6. ROC-кривые для трех технологий распознавания лиц

а вероятность принадлежности классу "Свой", решение о принадлежности принималось сравнением вероятности с порогом. Шаблон на базе Евклидова расстояния использовал последнее для определения расстояния между дескриптором рассматриваемого тестового примера и всеми обучающими примерами образа "Свой". Решение о принадлежности к образу "Свой" принималось в зависимости от числа обучающих примеров, расстояние до которых было меньше заданного. Для получения ROC-кривой для SVM и шаблонов на базе Евклидова расстояния изменяли порог принятия решения, при этом использовались те же обучающие примеры и те же образы (субъекты базы), что и для тестирования различных конфигураций НПБК. Сравнение ROC-кривых НПБК со 128 нейронами с аналогичными кривыми для шаблонов на базе Евклидова расстояния и SVM выполнено на рис. 6.

Заключение

Разрешающая способность НПБК применительно к задаче распознавания лиц в целом оказалась на уровне шаблонов и SVM. Если быть точнее, НПБК показывает несколько большее качество распознавания, чем SVM, при обучении на той же базе негативов. SVM специфична к обучающей базе негативов, при наличии смещения в этой базе относительно условий тестирования разрешающая способность снижается. Этим объясняется лучшее качество распознавания для простых шаблонов на базе Евклидова расстояния относительно и SVM, и НПБК. В то же время НПБК по сути является модифицированной SVM, отличающейся дополнительной регуляризацией — правилом обучения, которое гарантирует равновероятность бинарных значений на выходах нейронов. Дополнительная регуляризация снижает эффект переобучения на базе негативов, что и дает некоторое улучшение качества распознавания по сравнению с обычной SVM.

Что касается длины выходного кода, было установлено, что НПБК с кодом длиной 128 бит имеет лучшее качество распознавания, чем с кодом 32 бита. При этом выходной код длиной 128 бит соответствует требованиям ГОСТ Р 52633.0—2006 к качеству выходного "белого шума" НПБК: соблюдается требование к равновероятности состояний и корреляции разрядов выходного кода, это легко объясняется тем, что средний модуль корреляции 256 входных биометрических параметров НПБК, полученных на выходе глубокой сверточной сети, находится на уровне 0,1. Следовательно, положение ГОСТ Р 52633.0—2006, регламентирующее длину ключа для технологии распознавания лица на уровне 7—14 бит как минимум нуждается в пояснении, а более вероятно — в уточнении.

Список литературы

1. Lu Haiping, Martin Karl, Francis Bui, Plataniotis K. N., Hatzinakos D. Face recognition with biometric encryption for privacy-enhancing self-exclusion // Digital Signal Processing, 2009 16th International Conference (5–7 July 2009), Santorini-Hellas, Greece. P. 440–450.
2. Toli C.-A. Provoking Security: Spoofing Attacks against Crypto-Biometric Systems // World Congress on Internet Security (October 2015), Dublin, Ireland. URL: [https://lirias.kuleuven.be/bitstream/123456789/506050/1/Provoking + Security + Spoofing + Attacks + against + Crypto-Biometric + Systems.pdf](https://lirias.kuleuven.be/bitstream/123456789/506050/1/Provoking+Security+Spoofing+Attacks+against+Crypto-Biometric+Systems.pdf).
3. ГОСТ Р 52633.0—2006. Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации. М.: Гос. стандарт, 2007. 26 с.
4. Wu Xiang, He Ran, Sun Zhenan, Tan Tieniu. A Light CNN for Deep Face Representation with Noisy Labels // Chinese Academy of Sciences, Beijing, P. R. China. URL: <https://arxiv.org/abs/1511.02683> (дата обращения: 12.08.2017).
5. Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, Qiao Yu. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. URL: <https://arxiv.org/abs/1604.02878> (дата обращения: 12.08.2017).
6. Learned-Miller E., Huang B. G., RoyChowdhury A., Li Haoxiang, Hua Gang. Labeled Faces in the Wild: A Survey. // Advances in Face Detection and Facial Image Analysis, Springer Publishing Company, Incorporated, 2016. P. 189–248.
7. ГОСТ Р 52633.5—2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия—код доступа. М.: Гос. стандарт, 2012. 20 с.
8. ГОСТ Р 52633.1—2009. Защита информации. Техника защиты информации. Требования к формированию баз естественных биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации. М.: Гос. стандарт, 2010. 24 с.

S. V. Kulikov, Researcher, e-mail: kulikov@deepmark.ru,
O. S. Zakharov, Researcher, e-mail: zakharov@deepmark.ru,
D. Yu. Andreev, Director, e-mail: andreev@deepmark.ru,
LLC "Laboratoriya umnykh tekhnologiy", Penza, Russia

Exploring the Possibility of Using Deep Convolutional Neural Network Paired with Neural "Biometric Image to Code" Converter in Face Recognition

The article focuses on the possibility of using neural "biometric image to code" converter (NBCC) according to standards GOST R 52633.X in face-based biometric encryption. NBCC is used as last layer in pretrained deep convolutional neural network. The article describes the evaluation of indexes defined in GOST R 52633.1—2009 such as stability, uniqueness and quality of features extracted with deep convolutional neural network to analyze the possibility of training NBCC on features mentioned above. A number of NBCC configurations selected in accordance to GOST R 52633.5—2011 are evaluated and compared by ROC-curves to Euclidian distance based templates and SVMs.

Keywords: face recognition, biometric encryption, deep convolutional neural network, features, stability index, uniqueness index, quality index, support vector machine, Euclidian distance, ROC-curve

References

1. Lu Haiping, Martin Karl, Francis Bui, Plataniotis K. N., Hatzinakos Dimitris. Face recognition with biometric encryption for privacy-enhancing self-exclusion, *Digital Signal Processing, 2009 16th International Conference*, 5–7 July 2009, Santorini-Hellas, Greece, 2009, pp. 440–450.
2. Toli C.-A. Provoking Security: Spoofing Attacks against Crypto-Biometric Systems, *World Congress on Internet Security, October 2015, Dublin, Ireland*, available at: [https://lirias.kuleuven.be/bitstream/123456789/506050/1/Provoking + Security + Spoofing + Attacks + against + Crypto-Biometric + Systems.pdf](https://lirias.kuleuven.be/bitstream/123456789/506050/1/Provoking+Security+Spoofing+Attacks+against+Crypto-Biometric+Systems.pdf).
3. ГОСТ Р 52633.0—2006. *Zashhita informacii. Tehnika zashhity informacii. Trebovaniya k sredstvam vysokonadezhnoj biometricheskoj autentifikacii* (Information security. Information security technologies. Requirements for highly reliable biometric authentication instruments), Moscow: State standart, 2007, 26 p. (in Russian).
4. Wu Xiang, He Ran, Sun Zhenan, Tan Tieniu. *A Light CNN for Deep Face Representation with Noisy Labels*, Chinese Academy of Sciences, Beijing, P. R. China, available at: <https://arxiv.org/abs/1511.02683> (date of access: 12.08.2017).
5. Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, Qiao Yu. *Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks*, available at: <https://arxiv.org/abs/1604.02878> (date of access: 12.08.2017).
6. Learned-Miller E., Huang B. G., RoyChowdhury A., Li Haoxiang, Hua Gang. *Labeled Faces in the Wild: A Survey*, *Advances in Face Detection and Facial Image Analysis*, Springer Publishing Company, Incorporated, 2016, pp. 189–248.
7. ГОСТ Р 52633.5—2011. *Zashhita informacii. Tehnika zashhity informacii. Trebovaniya k sredstvam vysokonadezhnoj biometricheskoj autentifikacii. Avtomaticheskoe obuchenie nejrosetevykh preobrazovatelej biometrija-kod dostupa* (Information security. Information security technologies. Automating training of neural "biometric image to code" converter), Moscow: State standart, 2012, 20 p. (in Russian).
8. ГОСТ Р 52633.1—2009. *Zashhita informacii. Tehnika zashhity informacii. Trebovaniya k formirovaniyu baz estestvennykh biometricheskikh obrazov, prednaznachennykh dlja testirovaniya sredstv vysokonadezhnoj biometricheskoj autentifikacii*. (Information security. Information security technologies. Requirements for gathering of inartificial biometric datasets, designed for highly reliable biometric authentication instruments evaluation), Moscow: State standart, 2010, 24 p. (in Russian).

С. А. Горбатков, д-р техн. наук, проф.,
проф. каф. математики и информатики, e-mail: sgorbatkov@mail.ru,
Финансовый университет при Правительстве Российской Федерации, Уфимский филиал, г. Уфа,
Д. В. Полупанов, канд. техн. наук, доц.,
доц. каф. информационных технологий и компьютерной математики, e-mail: demetrious@mail.ru,
Башкирский государственный университет, г. Уфа

Оптимальный отбор и агрегирование экзогенных переменных в нейросетевых моделях банкротств на основе функций Харрингтона

Предложены два метода сокращения размерности факторного пространства при построении нейросетевых моделей банкротств: оптимального отбора факторов и агрегирования с использованием обобщенной функции желательности Харрингтона. Общей оригинальной концепцией для обоих методов, отличающейся от известных методов компрессии переменных, является взаимосвязь с операцией регуляризации обратной задачи обучения нейросетей на байесовском ансамбле.

Ключевые слова: оптимальный отбор, агрегирование факторов, нейросетевые модели банкротств, системный подход к моделированию

Введение.

Актуальность исследования

Работа посвящена одному из важных вопросов совершенствования нейросетевых технологий в сложных условиях моделирования (неточности, неполноты, неопределенности данных), характерному для задач банкротств. Широко известен факт, что предобработка данных и, в частности компрессия факторного пространства, предопределяет качество получаемой нейросетевой модели (НСМ), особенно в условиях сильной зашумленности данных и малого объема наблюдений [1]. Экономические приложения и более узко — модели банкротств не являются исключением [2—5]. Однако алгоритмы оптимального отбора факторов, а также компрессии факторного пространства в классе обратных задач для НСМ банкротств до сих пор не исследованы, точнее отсутствует общепринятый теоретический подход в этом вопросе. Исключение составляет работа Риссанена [6], где получен теоретически обоснованный обобщенный результат, базирующийся на принципе минимальной длины описания в выбранной модели-гипотезе h из ансамбля моделей, который выражается через информационную числовую меру Кульбака—Лейблера.

Подход Риссанена имеет существенное ограничение: он базируется на допущении об априорном знании закона распределения шумов в данных, который в экономических приложениях зачастую неизвестен. В настоящей работе мы отказались от этого допущения в целях приближения НСМ к практике.

Подчеркнем также важность системного подхода к разработке алгоритма предобработки данных. Если рассматриваемые НСМ восстановления зависимостей являются классом моделей, управляемых данными [7], то следует ожидать, что учет взаимосвязи алгоритмов предобработки данных (включая оптимальный отбор, компрессию факторного пространства) с алгоритмами обучения модели и ее регуляризации породит положительный эмерджентный эффект. Этот вопрос в литературе не рассматривался.

В литературе также не рассматривался вопрос об агрегировании переменных в НСМ на основе обобщенной функции желательности Харрингтона, что позволяет уже на стадии предобработки данных эффективно учитывать взаимодействие факторов, т. е. их влияние друг на друга.

Проведение исследований по указанным выше малоизученным актуальным вопросам применения нейросетей (НС) в моделях банкротств и послужило посылком к написанию данной статьи.

1. Постановка задачи исследования

Вначале оговорим экономическую постановку задачи, которая состоит в обслуживании банком своего кредитного портфеля в рамках выделенного примерно однородного кластера корпораций-заемщиков, сходных по отрасли экономики, условиям деятельности на рынке, экономико-политической обстановке в регионе и др. Фактор масштаба корпораций учитывается приведением экономических показателей стандартной бухгалтерской отчетности к безразмерному виду. Примером служат так называемые "финансовые коэффициенты" [3, 4], отражающие ликвидность корпораций, платежеспособность, рентабельность бизнеса, деловую активность и другие показатели финансово-экономического состояния. Число этих коэффициентов может достигать до нескольких десятков. Требуется построить модель для указанного кластера корпораций-заемщиков, которая позволяла бы оценивать вероятность банкротства конкретного заемщика и, главное, определять способы реструктуризации задолженности в зависимости от достигнутой стадии банкротства. При этом для построения модели предполагается использование ретроспективных данных аналогичного кластера заемщиков, содержащих для каждой корпорации информацию типа "банкрот — не банкрот". Для этих целей могут быть использованы базы данных самой кредитной организации, базы "Бюро кредитных историй Российской Федерации", налоговых органов.

Теперь изложим информационно-математическую постановку задачи исследования. Будем рассматривать обратную задачу (ОЗ) восстановления зависимости вероятности банкротства P от вектора экзогенных переменных $\mathbf{x} = (x_1, x_2, \dots, x_j, \dots, x_n)$, скрытой в данных. Эту зависимость будем определять в виде логистической функции, предложенной Ольсоном [8]:

$$P(t) = 1/[1 + \exp(-\hat{y}(\mathbf{x}(t), t))], P \in [0, 1], \quad (1)$$

где t — время, а показатель экспоненты $\hat{y}(\mathbf{x}(t), t)$, играющий роль аргумента в (1), восстанавливается с помощью НС-отображения:

$$\hat{y}(\mathbf{x}, t) = F(\mathbf{x}, \mathbf{W}, t), F: \mathbf{x} \in R^{(n)} \rightarrow \hat{y} \in R^{(1)}, \quad (2)$$

где $\mathbf{W}$ — множество синаптических весов НС; $F(\cdot)$ — оператор НС-отображения.

Отметим сразу особенности логистической НСМ (1)–(2).

1. Модель способна восстанавливать любые (даже очень сильные) нелинейные зависимости $y(\mathbf{x}, t)$, что выгодно отличает ее от регрессионных моделей Альтмана [3] и его последователей. В частности, в модели (1)–(2) исчезает проблема мультиколлинеарности факторов.

2. Логистическое отображение (1) является сжимающим в том смысле, что интервал для аргумента $\hat{y}(\mathbf{x}, t) \in [-7, 7]$ отображается в соответствующий интервал значения функции $P \in [0, 1]$, т. е. коэффициент сжатия составляет примерно 14. Следовательно, если НСМ (1)–(2) уже получена, протестирована и проэкзаменована, то ошибки в задании вектора факторов $\mathbf{x}$ будут "сжиматься" оператором логистического отображения (1) при вычислении вероятности P . Однако задача обучения НС, т. е. нахождения синаптических весов $\mathbf{W}$, является ОЗ, некорректно поставленной по Адамару [9], что требует специальных мер по регуляризации ее решения. Одним из путей является подход введения некорректной ОЗ в класс корректно поставленных по Тихонову [9]. Согласно теории решения некорректных задач, разработанной А. Н. Тихоновым и представителями его научной школы, показано: если в пространстве искомых решений Z выделить более узкое множество (компакт) $Z' \subset Z$, такое, что выполняются условия:

- *a priori* известно, что на компакте $\exists \mathbf{z} \in Z'$ (решение ОЗ существует);
- $\forall \mathbf{z} \in Z' : \mathbf{z} = A^{-1}(\mathbf{u})$ (решение единственно);
- $\forall \delta \exists \xi : (\Delta \mathbf{u} < \delta) \Leftrightarrow (\Delta \mathbf{z} < \xi)$, если $(\mathbf{z} + \Delta \mathbf{z}) \in Z'$ (решение устойчиво),

то можно регуляризовать решение.

Подчеркнем основную идею теории А. Н. Тихонова: для регуляризации ОЗ следует сужать пространство Z искомых решений. В работе [9] разработаны такие конструктивные методы, в частности стабилизирующие функционалы $\Omega(\mathbf{z})$.

Применительно к рассматриваемой ОЗ восстановления параметров НСМ банкротств (1)–(2) роль искомой функции $\mathbf{z}$ играет множество параметров НС, т. е. синаптических весов $\mathbf{W}$, аналогом обратного оператора A^{-1} является обратный оператор НС-отображения F^{-1} из выражения (2). В качестве наблюдаемых экспериментальных характеристик объекта $\mathbf{u}$ выступают кортежи (вектор-строки) таблицы выборки данных $\langle \mathbf{x}, y \rangle_{ig}$, $i = \overline{1, N}$, $g = \overline{1, G}$, где y_g — метки "банкрот — не банкрот" для корпорации-заемщика; g — номер

объекта; i — номер наблюдения. В качестве стабилизирующего функционала Тихонова в нашей задаче может быть выбран $\Omega(\mathbf{W}) = \|\mathbf{W}\|_{E_n}^2$, где $\|\cdot\|_{E_n}$ — евклидова норма матрицы $\mathbf{W}$. Такой пример рассмотрен авторами работы [4].

Однако даже при теоретической корректности ОЗ возможно появление практической некорректности. Причинами могут быть наличие шумов в экспериментальных данных (потеря устойчивости решения) и дискретность набора экспериментальных точек (неединственность решения) — наблюдаемые характеристики объекта могут быть описаны несколькими различными функциями $\mathbf{z}_1 = F_1^{-1}(\mathbf{x}, y)$, $\mathbf{z}_2 = F_2^{-1}(\mathbf{x}, y)$ и др. В случае НС $F_1(\cdot)$, $F_2(\cdot)$ — это различные операторы НС отображения (2).

Так, в работе [3] авторами на модельном примере НС-аппроксимации искусственно зашумленной детерминированной зависимости было показано, что существует критический уровень зашумленности и ее объема в данных по доле зашумленных вектор-столбцов $\{x_{ij}\}$ и y_i , когда НС теряет устойчивость.

Следовательно, требуется специальный алгоритм регуляризации ОЗ обучения НС. Этот вопрос применительно к НСМ банкротств практически не изучен. Авторы статьи разработали на базе подхода С. А. Шумского [10] алгоритм регуляризации обучения НС, названный "квазибайесовским" (см. ниже в разд. 2 настоящей работы). Этот термин означает упомянутый выше во введении отказ от допущения об априорном задании вида закона распределения шумов в данных и использует парадигму А. Н. Тихонова сужения пространства искомых решений.

3. При формировании обучающего множества НС возникает проблема, которую мы назвали "динамической неполнотой данных". До исследования авторов [3] решение этой проблемы не рассматривали. Подчеркнем, что "динамическая неполнота данных" при обучении НС — это одно из наглядных свойств НСМ как модели, управляемой данными. Суть этой проблемы состоит в следующем.

Пусть имеются ретроспективные данные наблюдения входного вектора $\mathbf{x} = (x_1, \dots, x_j, \dots, x_n)$ в нескольких временных срезах $\{t_k\}$, $k = 1, N$. Обучающая выборка $\langle \mathbf{x}_{gk}, y_{gk} \rangle$, $g = 1, G$ содержит наблюдения для G предприятий. При этом в последнем временном срезе ($k = N$) известны как векторы значения факторов $\{\mathbf{x}_{gN}\}$, так и значения выходной переменной $y = \arg(P(\mathbf{x}_{gN}, t_N))$. Это

позволяет обучить НС и оценить вероятность риска банкротства P в последнем временном срезе ($t = t_N$). Однако для всех предшествующих временных срезов ($t_1, t_2, \dots, t_{N-1}$) значения выходной (эндогенной) переменной $y_{gk} = \arg(P(\mathbf{x}_{gk}, t_k))$, $k = 1, N-1$ могут быть неизвестны, поскольку к моменту $t \leq t_N$ для части корпораций-заемщиков процесс кризиса развивается и они еще не признаны банкротами. В этом и проявляется глубокая специфика моделируемых объектов. Нужно построить специальный итерационный процесс восстановления эндогенных переменных $\{y_{gk}\}$, $k = 1, N-1$, $g = 1, G$ во всех временных срезах и для всех объектов. Соответствующий нейросетевой логистический динамический метод (НЛДМ), базирующийся на системном законе инерционности экономических процессов, разработан авторами в работе [3]. Описание этого метода выходит за рамки статьи. Поэтому далее будем считать, что данные во всех временных срезах, т. е. кортежах $\langle \mathbf{x}_{gk}, y_{gk}, t_k \rangle$, восстановлены и являются полными и НСМ рассматривается как статическая задача для каждого фиксированного среза ($y = t_k = \text{const}$).

4. Целью исследований в работе является разработка оригинального нейросетевого статического логистического метода (НСЛМ) моделирования банкротств в сложных условиях зашумленности и неполноты данных с оптимальным отбором факторов для формирования исходных данных, а также агрегированием факторов в виде обобщенных функций желательности Харрингтона [11]. Цель направлена на повышение качества НСЛМ-моделей.

5. Для достижения цели было необходимо:

А. Решить задачу формализации и оптимизации отбора факторов в исходных данных.

Б. Изучить возможность эффективного использования в построении НСМ банкротств агрегирующих функций Харрингтона.

В. Разработать алгоритм регуляризации решения обратной задачи обучения НС, оценить в этом алгоритме влияние агрегирования факторов в виде обобщенных функций Харрингтона.

2. Результаты исследований. Количественные оценки

Сформулируем идеи, связанные с решением задач А, Б, В по разработке НСЛМ с оптимальным отбором факторов, а также с агрегированием факторов с указанием конкретной научной

и прикладной новизны этих идей, а затем уже детализируем соответствующие операции алгоритмов указанных выше задач и приведем количественные оценки.

Задача А. Здесь осуществляется оптимальный отбор факторов. Формируется экспертным способом стартовый набор факторов, т. е. "сырые данные" D и по ним строится байесовский ансамбль вспомогательных нейросетевых субмоделей (НСМ), где НС-гипотезы $\{h_q(\mathbf{x}, \mathbf{W}, \mathbf{s})\}$ отличаются друг от друга видом активационных функций и параметрами структуры $\mathbf{s}$ (числом скрытых слоев нейронов и оптимальным числом нейронов в этих слоях). Согласно байесовской концепции регуляризации обучения [4, 10] все НС-гипотезы $\{h_q\}$, $q = \overline{1, Q}$ о происхождении данных D должны относиться к одному и тому же классу Ω , т. е. иметь одну и ту же "метагипотезу" $\forall h_q \in \Omega, q = \overline{1, Q}$. В данной работе для выбора Ω мы использовали парадигму многослойных персептронов с обратным распространением ошибки при обучении (Multilayer Perseptron — Back Propagation (MLP-BP)). В алгоритме оптимального отбора факторов используется критерий $\bar{\Theta}$, усредненный на апостериорном, т. е. отфильтрованном байесовском ансамбле НС $\{h_q|D, \mathbf{W}, H\}$, $q = \overline{1, Q^*}$, $Q^* \leq Q$. При этом критерий отбора $\bar{\Theta}$ выражается прямым способом через число ошибок первого и второго рода идентификации сетью объектов типа "банкрот — не банкрот". Организуется итерационный процесс оптимального отбора факторов, т. е. двойной фильтрации факторов исходных данных D : "внешних" итераций по индексу j (номеру фактора), $j = \overline{1, n}$ и "внутренних" итераций по индексу q (номеру сети) в байесовском ансамбле НС-гипотез $\{h_q\}$. Внешние итерации осуществляются как случайный выбор фактора из выборки D . Если фактор x_j оказывается достаточно информативным по критерию $\bar{\Theta}$ распознавания меток "банкрот — не банкрот" на выборке D , то он возвращается в выборку. В противном случае фактор исключается из выборки, поскольку он не несет полезной информации и ведет себя как шум. Такой случайный выбор охватывает перебор всех факторов. Критерий $\bar{\Theta}$ выражается следующими формулами:

$$\begin{aligned} \Theta_{qj} &= (N^{(1)} / N) + (N^{(11)} / N); \\ \bar{\Theta}_j &= \left[\sum_{q=1}^{Q^*} \Theta_{qj} \right] / N; \quad j = \overline{1, n}, \end{aligned} \quad (3)$$

где N — общее число корпораций-заемщиков в исходной выборке $\langle \mathbf{x}_j, y \rangle_{ig}$, $i = \overline{1, N}$, $g = \overline{1, G}$; $N^{(1)}$, $N^{(11)}$ — число ошибок первого и второго рода при идентификации обученной и протестированной НС предъявленных примеров по всей выборке; Q^* — число НСМ на байесовском ансамбле, отфильтрованном во внутреннем итерационном цикле по индексу $q = \overline{1, Q}$.

Правило отбора очередного j -го фактора в выборке:

$$x_j^* : \bar{\Theta}_j \leq \eta_1 | \Theta_{jq} \leq \eta_2. \quad (4)$$

Здесь звездочка "*" в факторе x_j^* означает, что он возвращается в выборку D при отборе; η_1 и η_2 — экспертно задаваемые числа, характеризующие качество ансамблевой идентификации для НСМ. Если для банка-кредитора важно учесть в оценке (3)—(4) различие ошибок первого и второго рода с точки зрения кредитной политики, то можно это сделать введением коэффициентов предпочтения [3] Фишберна в формуле (3). Отметим, что критерий $\bar{\Theta}$, в отличие от критерия, принятого в [10], не требует априорного задания аналитического вида функции правдоподобия, что упрощает расчеты и приближает их к практике.

Во внутреннем итерационном процессе фильтрации НС-гипотез $\{h_q\}$, $q = \overline{1, Q}$ при каждом фиксированном факторе x_j , ($j = \text{const}$) используется тот же критерий отбора $\bar{\Theta}$ по (3)—(4). НС-гипотеза h_q остается в байесовском ансамбле, если для нее выполнено условие $\Theta_{jq} \leq \eta_2$. В противном случае h_q исключается для данного j -го шага внешних итераций из ансамбля.

Таким образом, оптимальный отбор факторов в итоге двойной фильтрации во внешнем (по индексу j) и внутреннем (по индексу q) цикле можно представить формулой

$$\mathbf{x}_{v.\text{опт}} = \arg \left[\min_{v=\overline{1, n}} \{ \bar{\Theta}_v \} \middle| h_q(\mathbf{x}, y, \mathbf{W}) \in H \right], \quad (5)$$

где $\mathbf{x}_{v.\text{опт}}$ — оптимальный вектор (набор факторов) с номером v при случайном переборе всех факторов ($j = \overline{1, n}$); на этом наборе критерий $\bar{\Theta}_{\min}$ по (3)—(4) имеет минимальное значение.

Если множество $\mathbf{x}_{v.\text{опт}}$ не пусто, то (5) и есть мера эффекта предварительной компрессии факторов за счет их оптимального отбора. Поскольку алгоритм (5) включает в себя за счет внутренних итераций на байесовском ансамбле реализацию НСМ, то эту компрессию можно рассматривать

как "предрегуляризацию" основной НСМ в итоге решения последующих задач Б и В.

Теперь мы можем на концептуальном уровне сформулировать научную новизну основной идеи, заложенной в алгоритм (3)–(5) решения задачи А, алгоритм отличается от известных алгоритмов следующими признаками.

1. Оптимальный отбор факторов осуществляется не изолированно от построения основной (рабочей) НСМ, а в тесной взаимосвязи с нею с использованием ансамбля ВНСМ из того же класса Ω и того же критерия качества (3), (5), что и в основной НСМ.

2. Одновременно с оптимальным отбором факторов осуществляется регуляризация ВНСМ на байесовском ансамбле НС, которую в составе НСЛМ можно трактовать как "предрегуляризацию" всей модели.

Порождаемый этими новыми признаками эмерджентный эффект — предварительная оптимальная компрессия факторного пространства и, соответственно, облегчение построения основной НСМ в аспекте обеспечения ее требуемых прогностических свойств.

Задача Б. Здесь осуществляется основной процесс комплексного (ступенчатого) агрегирования факторов, отличающийся от известных алгоритмов компрессии факторов тем, что вначале на эвристическом уровне по функциональному экономическому признаку формируются группы (кластеры) факторов, а затем внутри каждого u -го кластера факторы агрегируются в виде обобщенной функции желательности Харрингтона (ОФХ) [11] $\{H_u\}$; $u = \overline{1, M}$, и НС строится по агрегированным переменным в (2):

$$\hat{y}(H_u, t_k) = F(H_u, \mathbf{W}, t_k), k = \overline{1, N}, u = \overline{1, M}. \quad (6)$$

Достигаемые за счет этого признака НСЛМ эмерджентные эффекты:

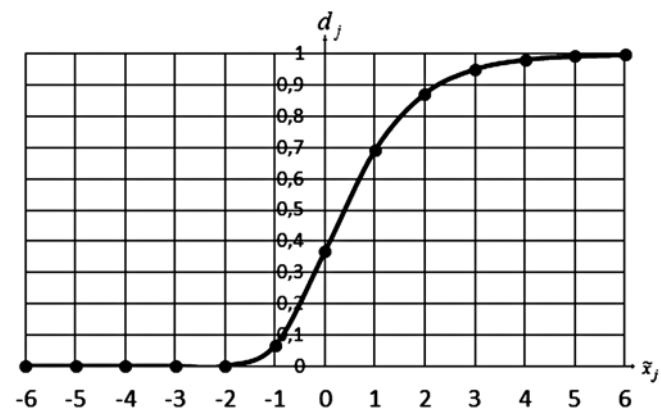
- возможность внесения в алгоритм предобработки данных априорной эвристической информации, аккумулирующей профессиональный опыт аналитика, разрабатывающего НС при образовании кластеров факторов;
- сильная компрессия данных и облегчение обучения НС в аспекте повышения ее прогностической силы согласно рекомендациям Рисанена [6];
- возможность учета в явной форме нелинейного взаимного влияния факторов $x_j, x_k, k \neq j$ друг на друга внутри структур функции $\{H_u\}$;

- возможность наглядной экономической интерпретации зависимостей вероятности банкротства от агрегатов $\{H_u\}$, получаемых с помощью (1) и НСМ. Например, $\{H_u\}$ могут оценивать рентабельность бизнеса, ликвидность корпораций-заемщиков, платежеспособность и другие свойства.

Сделаем краткие комментарии по ОФХ. Более подробное описание содержится в работе [11]. Одним из наиболее удобных способов агрегирования как экзогенных x_j , так и эндогенных y_m , по мнению авторов статьи, является построение обобщенной функции желательности Харрингтона H . В основе построения этой функции лежит преобразование натуральных значений переменных в безразмерную шкалу желательности или предпочтительности: $x_j, y_m \rightarrow d_j$, где d_j — частная функция желательности.

Шкала желательности (см. рисунок) относится к психофизическим шкалам. Ее назначение — установление соответствия между физическими и психологическими (лингвистическими) величинами. Под психологическими величинами понимаются чисто субъективные (экспертные) оценки желательности того или иного значения переменных x_j, y_m . В данной работе все количественные оценки относятся к факторам $\{x_j\}$, $j = \overline{1, n}$, хотя полученные результаты в принципе справедливы и для агрегированных экзогенных переменных $\{y_m\}$.

Чтобы получить шкалу желательности, удобно пользоваться готовыми разработанными таблицами между отношениями предпочтения в числовой системе (отметка d_j на шкале частных функций желательности) и лингвистической шкале (табл. 1) [11].



Частная функция желательности $d_j(x_j)$

Таблица 1

Стандартные отметки на шкале желательности

Желательность (лингвистическая оценка переменной x_j)	Отметка на шкале желательности d_j
Очень хорошо	0,8...1,0
Хорошо	0,63...0,8
Удовлетворительно	0,37...0,63
Плохо	0,2...0,37
Очень плохо	0...0,2

Более точные результаты, чем табл. 1, дает аналитическое представление функции желательности:

$$d_j = \exp(-\exp(-\tilde{x}_j)); d_j \in [0, 1]; \tilde{x}_j \in [-6, 6]. \quad (7)$$

Заметим, что ширина интервала нормированной переменной $\tilde{x}_j$ должна выбираться так, чтобы значение $\tilde{x}_j = -6$ соответствовало в табл. 1 лингвистической оценке "Очень плохо" ($d_j = 0$), а $\tilde{x}_j = 6$ должно соответствовать оценке "Очень хорошо" ($d_j = 1,0$). Ось абсцисс на рисунке можно сузить, например $\tilde{x}_j \in [-3, 3]$. Тогда функция желательности $d_j(\tilde{x}_j)$ будет иметь более крутой вид и точность графических оценок несколько снизится.

Харрингтоном был предложен единый комплексный мультипликативный показатель, а именно обобщенная функция желательности (среднее геометрическое в каждом u -м кластере):

$$H_u = \sqrt[n]{\prod_{j=1}^n d_j}, \quad u = \overline{1, M}, \quad (8)$$

значения которой также могут быть интерпретированы по шкале, представленной в табл. 1. График ОФХ (8) будет таким же, как и для частных функций желательности d_j (7) (см. рисунок). В исследуемой нами задаче число агрегируемых факторов в каждом кластере будет различным, т. е. в (8) вместо n следует подставлять $\{n_u\}$.

Использование функции Харрингтона по (7), (8) требует выполнения следующих правил.

- Все факторы в фиксированном кластере факторов ($u = \text{const}$) должны быть однонаправленными, т. е. уменьшению вероятности банкротства P в логистической функции (1) должно соответствовать увеличение фактора x_j и,

соответственно, рост показателя экспоненты $\hat{y}(\mathbf{x})$. Если некоторые из факторов $\{x_k\}$ в кластере имеют инверсный характер, т. е. рост этих факторов увеличивает вероятность банкротства P в формуле (1), то их следует преобразовать, например: $x_k \rightarrow x'_k = 1/x_k$.

- ОФХ H по (8) устроена так, что, если одна из частных функций желательности $d_j(\tilde{x}_j)$ внутри кластера ($x_j \in H_u, u = \text{const}$) равна нулю, то обнуляется и весь агрегат H_u с потерей его информативности. Поэтому авторы статьи, учитывая малую чувствительность функции частных желательностей $d_j(\tilde{x}_j)$ по (7) вблизи нуля, предлагают несколько загрузить оценку, принимая

$$d'_j = \begin{cases} \varepsilon, & \text{если } d_j \leq 0,1, \\ d_j & \text{в противном случае.} \end{cases} \quad (9)$$

Здесь ε — задаваемое аналитиком малое число, например $\varepsilon = 0,05$.

Задача В. По сути квазибайесовский алгоритм регуляризации (КБАР) решения ОЗ нахождения параметров НСМ, т. е. множества синаптических весов $\mathbf{W}$, был описан выше в алгоритме задачи А. Более подробно этот вопрос изложен в работе [4]. Остается сформулировать основную идею КБАР и конкретизировать его новизну.

КБАР использует ту же парадигму регуляризации решения ОЗ, что и в теории А. Н. Тихонова — это сужение пространства искомых решений $Z' \subset Z$, где Z' — некоторый "компакт". Однако способ сужения Z до компакта Z' в КБАР другой, отличный от построения стабилизаторов Тихонова $\Omega(\mathbf{z})$.

В КБАР сужение пространства Z осуществляется последовательно в три этапа:

Этап 1. Выбор класса (метегипотезы Ω ; $h_q(\mathbf{x}, y, \mathbf{W}) \in \Omega$) и вида НС-гипотез $\{h_q\}$ внутри Ω . Как было отмечено выше, в качестве Ω был выбран класс MLP-ВР, а в качестве варьируемых внутри Ω активационных функций для $\{h_q\}$ — гиперболический тангенс или логистическая функция; в качестве параметров структуры $\mathbf{s}$ сети выбрано число скрытых слоев и оптимальное число нейронов в этих слоях.

Этап 2. Апостериорная фильтрация НС-гипотез байесовского ансамбля $\{h_q\}$, $q = \overline{1, Q}$ по прямому критерию ошибок первого и второго рода при идентификации объектов выборки в (4), (5).

Этап 3. Усреднение критерия Θ на отфильтрованном байесовском ансамбле Q^* по (3).

Байесовский ансамбль НСМ $\{h_q\}$

№ НСМ	Число скрытых слоев	Вид активационной функции в скрытых слоях
1	1	(10)
2	2	(10) в обоих слоях
3	2	(10) в первом слое, (11) — во втором слое
4	1	(11)
5	2	(11) в обоих слоях
6	2	(11) в первом слое, (10) — во втором слое

табл. 2. В скрытых слоях НС использованы активационные функции логистической сигмоиды и гиперболического тангенса:

$$f(s) = \frac{1}{1 + e^{-as}}, \quad a > 0; \quad (10)$$

$$f(s) = \text{th}(bs), \quad b > 0. \quad (11)$$

Активационная функция выходного слоя во всех НСМ — линейная.

Указанные выше характеристики ансамбля были единичными в количественных оценках для всех трех задач А, Б, В (см. ниже).

Экспертно задаваемое число η_1 в правиле отбора факторов выбирали равным 0,1 (или 90 % правильно идентифицируемых объектов). Результаты оптимального отбора факторов показаны в табл. 3. Видно, что все шесть НС-гипотез $\{h_q\}$ в байесовском ансамбле успешно прошли фильтрацию и усредняемый критерий качества НСМ $\bar{\Theta}$ по (3) вычисляется по всем шести НС, т. е. $Q^* = 6$. Оптимальный вариант отбора соответствует пятой итерации и соответствует набору из 11 факторов: $L_1, L_2, F_1, F_2, R_1, R_2, R_3, A_2, A_4, A_6$ при минимальном значении критерия качества по (3) $\bar{\Theta}_{\min} = 0,0444$ (или 95,56 % верно идентифицированных объектов).

Таким образом, применение байесовского ансамбля в сочетании с отбором признаков приводит к улучшению модели: критерий $\bar{\Theta}$ уменьшился по сравнению с исходным значением ($v = 0$), равным 0,0778, до значения $\bar{\Theta} = 0,0444$, т. е. в 1,75 раз, при сокращении числа факторов с 15 до 11 (в 1,36 раз). Это весьма ощутимый выигрыш за счет оптимального отбора факторов, причем одновременно реализуется процесс регуляризации основной НСМ.

КБАР отличается от известного алгоритма С. А. Шумского [10] двумя новыми признаками:

1) отказом от допущения об известном априори виде закона распределения шумов (в [10] используются гауссовы и лапласовы распределения);

2) регуляризация решения ОЗ для основной НСМ выполняется взаимосвязано с алгоритмом оптимального отбора факторов: для одинакового класса $h_q \in \Omega$ вспомогательных субмоделей и основной модели, а также одинакового критерия качества (3), (4).

Получаемый эмерджентный эффект от этих новых признаков — упрощение расчетов (не требуется выполнять расчет функции правдоподобия распределения шумов, которая обычно неизвестна), обеспечение работоспособности (устойчивости) НСМ в сложных условиях моделирования (за счет процедур предрегуляризации — регуляризации ОЗ), приближение НСМ к практике.

Количественные оценки. В качестве исходных данных D использовали ретроспективные данные корпораций-заемщиков, одной из наиболее распространенных отраслей экономики — строительной отрасли, полученные фирмой "Bereua Van Dijk" [12]. База данных содержала 90 наблюдений. Использовалась система из 15 удельных показателей, широко применяемых в задачах оценки банкротств [13]: L_1 — быстрый коэффициент ликвидности; L_3 — коэффициент покрытия запасов; P_1 — текущий коэффициент ликвидности; F_1 — коэффициент финансовой зависимости; F_2 — коэффициент автономии собственных средств; F_3 — обеспеченность запасов собственными оборотными средствами; F_4 — индекс постоянного актива; R_1 — общая рентабельность; R_2 — рентабельность активов; R_3 — рентабельность собственного капитала; R_5 — рентабельность оборотных активов; A_2 — оборачиваемость активов; A_4 — оборачиваемость кредиторской задолженности; A_5 — оборачиваемость дебиторской задолженности; A_6 — оборачиваемость запасов.

Формулы для расчета этих удельных показателей содержатся в работе [3] и оперируют с данными стандартной бухгалтерской отчетности.

Был реализован алгоритм (5), описанный выше для задачи А оптимального отбора факторов с применением регуляризации на байесовском ансамбле ВНСМ.

Характеристики ансамбля: метаягипотеза Ω — MLP-BP, реализованный на программном продукте NeuroSolutions 5.0 (демоверсия). Характеристики НС-гипотез $\{h_q\}$, $q = 1, Q$ приведены в

Итерационный процесс оптимального отбора факторов

№ шага итерации v	Набор факторов $\{x_{vj}\}$	Исключение (–) или включение (+) факторов	$N^{(I)}$	$N^{(II)}$	$\bar{\Theta}$ по (3)
0	$L_1, L_2, P_1, F_1, F_2, F_3, F_4, R_1, R_2, R_3, R_5, A_2, A_4, A_5, A_6$	—	4	3	0,0778
1	$L_1, L_2, P_1, F_1, F_2, F_3, F_4, R_1, R_2, R_3, A_2, A_4, A_5, A_6$	$-R_5$	2	3	0,0556
2	$L_1, L_2, P_1, F_1, F_2, F_3, R_1, R_2, R_3, A_2, A_4, A_5, A_6$	$-F_4$	3	1	0,0444
3	$L_1, L_2, P_1, F_1, F_2, F_3, R_1, R_2, R_3, A_4, A_5, A_6$	$-A_2$	4	1	0,0556
4	$L_1, L_2, P_1, F_1, F_2, F_3, R_1, R_2, R_3, A_2, A_4, A_6$	$+A_2, -A_5$	3	1	0,0444
5	$L_1, L_2, F_1, F_2, F_3, R_1, R_2, R_3, A_2, A_4, A_6$	$-P_1$	3	1	0,0444
6	$L_1, L_2, F_1, F_2, R_1, R_2, R_3, A_2, A_4, A_6$	$-F_3$	3	3	0,0667

Таблица 4

Результаты нейросетевого моделирования при агрегировании факторов на основе функций Харрингтона и сравнения с другими моделями

Номер модели	Показатель	Номер НС-гипотез в ансамбле						Среднее на ансамбле
		1	2	3	4	5	6	
1	$N^{(I)}$	2	7	5	0	2	2	1
	$N^{(II)}$	4	8	4	1	3	3	2
	Θ	0,0667	0,1667	0,1	0,0111	0,0556	0,0556	0,0333
2	$N^{(I)}$	4	4	3	3	2	3	4
	$N^{(II)}$	3	4	2	2	4	2	3
	Θ	0,0778	0,0889	0,0556	0,0556	0,0667	0,0556	0,0778
3	$N^{(I)}$	3	4	4	2	1	3	3
	$N^{(II)}$	2	3	1	1	2	1	1
	Θ	0,0556	0,0778	0,0556	0,0333	0,0333	0,0444	0,0444
4	$N^{(I)}$	2	3	2	2	1	1	2
	$N^{(II)}$	12	11	10	7	5	9	8
	Θ	0,1556	0,1556	0,1333	0,1	0,0667	0,1111	0,1111

Количественные оценки для задач Б и В сведены в итоговую табл. 4. Предлагаемый метод агрегирования на базе ОФХ (модель 1), который был подробно описан выше в задачах Б и В, сравнивался с широко применяемыми другими моделями: вариантами расчета на 15 факторах (без агрегирования, модель 2), с рассмотренным выше оптимальным отбором факторов (модель 3) и также с агрегированием факторов по методу нечеткой матричной свертки [13] (модель 4).

Заметим, что эффект фильтрации на апостериорном байесовском ансамбле (обученных и протестированных НС) проявился четко для модели 1 (отсеяна НС с номером 2 при $\Theta > \eta = 0,1$) и в модели 4 (НС с номерами 1, 2, 3 и 6).

Наилучший результат ($\bar{\Theta} = 0,0333$) соответствует предлагаемой модели 1 с агрегированием факторов по четырем кластерам по ОФХ $\{H_u\}$,

$u = 1, 2, 3, 4$. В кластере H_1 агрегируются факторы ликвидности; в кластере H_2 — факторы финансовой устойчивости; в кластере H_3 — факторы рентабельности и в кластере H_4 — факторы деловой активности. Таким образом, позиционированный эмерджентный эффект предлагаемого метода в обеих задачах Б и В подтвержден на реальных зашумленных данных строительной отрасли.

Заключение

В предложенном методе оптимального отбора факторов при формировании базы исходных данных D (см. табл. 3) достигается эмерджентный эффект — улучшение качества НСМ $\bar{\Theta}$ по (5) в 1,75 раза при компрессии факторов в 1,35 раз, что достигается за счет взаимосвязи алгоритма отбора факторов с регуляризацией ВНСМ

на байесовском ансамбле. Механизм указанной "взаимосвязи", по мнению авторов статьи, состоит в учете нелинейного влияния факторов друг на друга в процессе обучения ВНСМ в методе отбора факторов, что повышает информативность данных D и качество всей рабочей НСМ по прогностическому критерию Θ .

В предложенном методе комплексного (ступенчатого) агрегирования факторов на основе функций Харрингтона достаточно высокое прогностическое качество НСМ в сложных условиях моделирования банкротств (см. табл. 4) обеспечивается основным новым признаком метода: операция агрегирования проводится взаимосвязанно с операцией регуляризации решения ОЗ на байесовском ансамбле путем поэтапного сужения пространства искомых решений ($Z \rightarrow Z'$, где Z' — компакт).

Предложенные теоретические идеи двух оригинальных методов компрессии переменных подтверждены результатами на реальных данных строительной отрасли.

В аспекте дальнейшего развития предложенных двух методов представляет интерес их последовательное сочетание при большом числе факторов (сотни) — вначале оптимального отбора факторов, затем агрегирования факторов на основе функций Харрингтона.

Список литературы

1. Галушкин А. И. Нейронные сети: основы теории. М.: Горячая линия — Телеком, 2012. 469 с.

2. Галушкин А. И. Применение нейрокомпьютеров в финансовой деятельности. URL: <http://www.hardline.ru/3/37/1484/> (дата обращения: 28.08.2017).

3. Белолипцев И. И., Горбатков С. А., Романов А. Н., Фархива С. А. Моделирование управленческих решений в сфере экономики в условиях неопределенности. М.: ИНФРА-М, 2015. 299 с.

4. Горбатков С. А., Полупанов Д. В., Макеева Е. Ю., Бирюков А. Н. Методологические основы разработки нейросетевых моделей экономических объектов в условиях неопределенности. М.: Издательский дом "Экономическая газета", 2012. 494 с.

5. Матвеев М. Г., Свиридов А. С., Алейникова Н. А. Модели и методы искусственного интеллекта. Применение в экономике: учеб. пособие. М.: Финансы и статистика; ИНФРА-М, 2008. 448 с.

6. Rissanen J. Modeling by shortest data description // *Automatica*. 1978. Vol. 14. P. 465—471.

7. Доленко С. А. Нейросетевые методы решения обратных задач // XV Всероссийская научно-техническая конференция "Нейроинформатика — 2013": Лекции по нейроинформатике. М.: НИЯУ МИФИ, 2013. С. 214—269.

8. Ohlson J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy // *Journal of Accounting Research*, 1980, № 18 (1), p. 109—113.

9. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. М.: Наука. Физматлит, 1986. 288 с.

10. Шумский С. А. Байесова регуляризация обучения // Научная сессия МИФИ 2002. IV Научно-техническая конференция "Нейроинформатика — 2002": Лекции по нейроинформатике. Часть 2. М.: МИФИ, 2002. С. 30—93.

11. Адлер Ю. П., Маркова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 1976. 279 с.

12. Makeeva E. U., Neretina E. A. Binary model versus discriminant analysis relating to corporate bankruptcies: The Case of Russian Construction Industry // *Journal of Accounting, Finance and Economics*. 2013. Vol. 3. № 1. P. 65—76.

13. Шевченко И. В., Халафян А. А., Васильева Е. Ю. Создание виртуальной клиентской базы для анализа кредитоспособности российских предприятий // *Финансы и кредит*. 2010. № 1 (385). С. 13—18.

S. A. Gorbatkov, D. Sc. Professor, sgorbatkov@mail.ru,
Financial university under the government of the Russian Federation, Ufa (branch), Ufa 450015, Russia,

D. V. Polupanov, Ph. D., Associate Professor, demetrius@mail.ru,
Bashkir State University, Ufa, 450076, Russia

Optimal Selection and Aggregation of Exogenous Variables in Neural Network Models of Bankruptcies Based on Harrington Functions

The article is devoted to the issue of constructing and further improving neural network models in conditions of noise, inaccuracy, incompleteness and uncertainty of the data. A practical application is the development of an original neural network logistics method for assessing the probability of enterprise bankruptcies. In order to improve the quality, adequacy, accuracy and predictive properties of neural network models, two methods of reducing factor space are proposed: 1) an iterative method of optimal factor selection; 2) aggregation of factors using the generalized Harrington desirability function. A common original concept of the proposed methods, which differs from the known methods of exogenous variables compression, is a complex (systemic) study of the interrelation between the operations of optimal factor selection and their aggregation with

the operation of regularizing the inverse problem of training neural networks in a Bayesian ensemble. The generated emergent effect is the compression of the factor space and, accordingly, facilitating the construction of a neural network model in terms of providing its required prognostic properties. The proposed theoretical ideas of the two original methods of variable compression are confirmed by the results on real data of construction industry enterprises. Quantitative estimates are presented comparing the proposed methods with the construction of neural network models on the initial set of factors, as well as with the aggregation of factors by the fuzzy matrix convolution method.

Keywords: optimal selection, aggregation of factors, neural network models of bankruptcies, system approach to modeling

References

1. Galushkin A. I. *Nejronnye seti: osnovy teorii* [Neural networks: the fundamentals of the theory], Moscow, Gorjachaja linija — Telekom, 2012. 469 p. (in Russian).
2. Galushkin A. I. *Primenenie nejrokomput'juterov v finansovoj dejatel'nosti* [The use of neurocomputers in financial activities]. URL: <http://www.hardline.ru/3/37/1484/> (data of access: 28.08.2017). (in Russian).
3. Belolipcev I. I., Gorbatkov S. A., Romanov A. N., Farhieva S. A. *Modelirovanie upravlencheskih reshenij v sfere jekonomiki v uslovijah neopredelennosti*, Moscow, INFRA-M, 2015, 299 p. (in Russian).
4. Gorbatkov S. A., Polupanov D. V., Makeeva E. U., Birjukov A. N. *Metodologicheskie osnovy razrabotki nejrosetevyh modelej jekonomicheskikh objektov v uslovijah neopredelennosti* [The methodological basis for the development of neural network models of economic entities in conditions of uncertainty], Moscow, Izdatel'skij dom "Ekonomicheskaja gazeta", 2012, 494 p. (in Russian).
5. Matveev M. G., Sviridov A. S., Alejnikova N. A. *Modeli i metody iskusstvennogo intellekta. Primenenie v jekonomike: uchebnoe posobie*. [Models and methods of artificial intelligence. Application in Economics: Textbook], Moscow, Finansy i statistika; INFRA-M, 2008, 448 p. (in Russian).
6. Rissanen J. Modeling by shortest data description, *Automatica*, 1978, vol. 14, pp. 465—471.
7. Dolenko S. A. *Nejrosetevye metody reshenija obratnykh zadach* [Neural network methods for solving inverse problems], XV Vserossijskaja nauchno-tehnicheskaja konferencija "Nejroinformatika — 2013": Lekcii po nejroinformatike, Moscow, NIJaU MIFI, 2013, pp. 214—269 (in Russian).
8. Ohlson J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 1980, no. 18 (1), pp. 109—113.
9. Tihonov A. N., Arsenin V. Ja. *Metody reshenija nekorrektnykh zadach* [Methods for solving ill-posed problems]. Moscow, Nauka. Fizmatlit, 1986, 288 p. (in Russian).
10. Shumskij S. A. Bajesova reguljarizacija obuchenija [Bayesian regularization of learning], *Nauchnaja sessija MIFI 2002. IV Nauchno-tehnicheskaja konferencija "Nejroinformatika — 2002": Lekcii po nejroinformatike. Chast' 2*, Moscow, MIFI, 2002, pp. 30—93 (in Russian).
11. Adler Ju. P., Markova E. V., Granovskij Ju. V. *Planirovanie jeksperimenta pri poiske optimal'nykh uslovij*. [Planning an experiment in searching for optimal conditions], Moscow, Nauka, 1976, 279 p. (in Russian).
12. Makeeva E. U., Neretina E. A. Binary model versus discriminant analysis relating to corporate bankruptcies: The Case of Russian Construction Industry, *Journal of Accounting, Finance and Economics*, 2013, vol. 3, no. 1, pp. 65—76.
13. Shevchenko I. V., Halafjan A. A., Vasil'eva E. Ju. *Sozdanie virtual'noj klientskoj bazy dlja analiza kreditosposobnosti rossijskikh predpriyatij* [Creating a virtual customer base for the analysis of the creditworthiness of Russian companies], *Finansy i kredit*, 2010, no. 1 (385), pp. 13—18 (in Russian).

УДК 004.05

С. М. Авдошин, канд. техн. наук, проф., руководитель департамента
программной инженерии факультета компьютерных наук НИУ ВШЭ, savdoshin@hse.ru,
Е. Ю. Песоцкая, канд. экон. наук, доц.
факультета компьютерных наук НИУ ВШЭ, epesotskaya@hse.ru,
Национальный исследовательский университет "Высшая школа экономики" (НИУ ВШЭ)

Интернет вещей: Транспорт

Рассматривается возможность использования Интернета вещей в транспортной отрасли, перспективы развития Интернета вещей при перевозке грузов и пассажиров в России. Детально рассмотрены эффекты, относящиеся к транспортным перевозкам, "умному транспорту" и потребительским приложениям, которые можно достичь в транспортной отрасли за счет Интернета вещей. Авторы исследуют перспективные направления развития Интернета вещей в транспортной отрасли, связанные с развитием цифровых и мобильных технологий. Выполнен анализ преимуществ и получения дополнительной выгоды в процессах перевозки пассажиров и грузов, обслуживания транспорта, управления маршрутами, логистики. При формировании рекомендаций были учтены основные мировые тренды, связанные с развитием Интернета вещей в транспорте.

Ключевые слова: Интернет вещей, мобильные технологии, мобильные приложения, перевозки, транспорт, "умный транспорт", интеллектуальные транспортные системы

Введение

С ростом числа используемых устройств, подключенных к сети Интернет, которые оснащены датчиками и могут выполнять команды пользователя, понятие "Интернет вещей" становится все более популярным и широко используемым. Происходит переход от "Интернета людей" к "Интернету вещей" (Internet of Things, IoT), который представляет собой широкую сеть подключенных к сети устройств, обеспечивающих постоянную связь и обмен информацией в режиме реального времени.

Аналитики Gartner подсчитали, что к концу 2017 г. по всему миру будет насчитываться 8,4 млрд подключенных к Сети устройств. По сравнению с прошлым годом их число вырастет на 31 %. Предполагается, что к 2020 г. число подобных устройств достигнет 20,4 млрд штук. [1]. По итогам 2017 г. объем рынка Интернета вещей в денежном выражении составит 1,7 трлн долл. против 1,4 трлн долл. в 2016 г. К 2020 г. по всему миру будет насчитываться от 40 до 50 млрд устройств, подключенных к Интернету в соответствии с исследованием AIG. [2].

Интернет вещей — ключевой элемент цифровой трансформации и нового 4-го этапа цифро-

вой революции, на пороге которого стоит сегодняшнее общество. В отличие от предыдущего этапа, который отличался быстрым проникновением Интернета в жизнь потребителей, новый этап готов обеспечивать проникновение гораздо более широкого спектра цифровых сервисов, продуктов и систем. Фактически речь идет об "Интернете всего" и предполагает подключенность не только людей, но и механизмов к глобальной сети за счет всеобщей подключенности и динамичного роста сенсорных устройств и данных.

Помимо людей, Интернетом сегодня "пользуются" около 10 млрд машин и механизмов — устройств, датчиков и приборов, а к 2020 г. прогнозируется двукратное увеличение этого числа. 99 % мировых данных уже оцифровано, и более 50 % имеет IP-адрес. По прогнозам в дальнейшем объем данных будет удваиваться каждые два года [3].

Все больший оборот набирает промышленный Интернет вещей который учитывает отраслевую и производственную специфику, объединяет все промышленные объекты в единую сеть с помощью M2M (*machine-to-machine*), Big Data ("большие" данные), облачных технологий. Основной задачей, как правило, является рост производительности за счет развития новых источников дохода и сокращения простоев, дефектов,

затрат на ремонт. Что же касается транспортной отрасли — это может быть оптимизация логистических затрат и поиск наилучших условий доставки, снижение расходов топлива, сокращение простоев транспорта и времени технического обслуживания, автоматизация диспетчерских служб. Так, например, американская компания *GE Aviation* [4] производит авиадвигатели, на которых установлены сенсоры, позволяющие удаленно получать данные об эксплуатации и на их основе выявлять оптимальные алгоритмы обслуживания самолетов, что позволило в 7 раз сократить затраты на обслуживание. А горнодобывающая компания *Rio Tinto* в Австралии [5] использует для перевозки породы беспилотные карьерные самосвалы, работающие непрерывно и управляемые из оперативного центра на расстоянии 1 200 км.

Массовое внедрение интеллектуальных датчиков в оборудование и производственные линии глубоко проникает во многие индустрии: промышленность, металлургия, энергетика, сельское хозяйство, ритейл, ЖКХ, транспорт и другие. Сегодня компании, которые решили использовать Интернет вещей и анализировать данные, которые для них собирают сервисы и устройства, умеют строить предиктивные модели и принимать решения, основанные на обработанных данных. То есть все собранные данные и отчеты превращаются в идеи и рекомендации, используемые компаниями для принятия эффективных и взвешенных решений.

До недавнего времени многие компании работали по принципу "делаем, что мы считаем нужным" или "делаем, как умеем". Решения принимались на основании собственного опыта и интуиции. Такой подход не всегда себя оправдывал, поскольку для принятия решений зачастую не хватало объективных и актуальных данных. Сегодня во многих отраслях преобладает автоматизированный сбор и обработка данных. С каждым днем в мире появляется все больше различных данных, следовательно, больше возможностей для получения полезных инсайтов и ускорения роста. Большое количество релевантных данных позволяет компании смотреть как на ситуацию в целом, так и на мельчайшие детали, что является существенным преимуществом. Поэтому своевременный сбор и обработка нужной информации — чрезвычайно важная задача. Согласно прогнозам компании IDC, мировой парк установленных точек Интернета вещей, который в конце 2016 г. составлял 14,9 млрд, в 2025 г. превысит 82 млрд. При таких темпах Интернет вещей вскоре может стать столь же незаменимым, как и сама сеть Интернет. По оценке аналитического агентства Gartner, количество

"вещей", подключенных к глобальной сети, составит 50 млрд к 2025 г. По данным McKinsey, экономический эффект от развития индустрии к 2025 г. составит 6,2 триллионов долл. США [6].

Однако успешность Интернета вещей отрасли будет, скорее, зависеть от человеческого фактора, нежели от специальных датчиков, сенсоров, способов анализа данных, алгоритмов или инструментов. По данным исследования Digital IQ [7], промышленным компаниям необходимо создавать условия для распространения цифровой культуры и обеспечивать прямую заинтересованность в успешной трансформации со стороны руководства, привлекать и обучать специалистов из числа представителей "цифрового поколения".

Интернет вещей в транспортной отрасли

Существенное удешевление технических средств и повсеместная распространенность сетевых подключений позволили технологии Интернета вещей войти практически в каждую отрасль экономики, и транспорт не стал исключением. В отрасли, где протяженность различных видов путей превышает 1,6 млн км, а число грузового транспорта (автомобильного, железнодорожного и прочих) — 7 млн единиц [8], крайне важно использовать системы удаленного мониторинга, анализа информации в режиме реального времени, контроля состояния транспорта и путей.

В отчете IDC дается прогноз, что расходы на Интернет вещей в 2017 г. вырастут на 17 %, а IoT-расходы заказчиков в 2017 г. составят чуть более 800 млрд долл., что говорит о том, что заказчики во всем мире готовы платить за внедрение технологий Интернета вещей. Ожидается, что объем инвестиций в транспортные перевозки составит 85 млрд долл. Наибольшие инвестиции запланированы на контроль за транспортировкой грузов [9].

Интернет вещей в транспортной отрасли позволяет обеспечивать эффективность использования транспортных средств, отслеживая потребности в техническом обслуживании, поддерживать постоянную связь и подключение к сети для оптимизации маршрутов доставки и контроля работы транспорта, повышать качество и эффективность коммуникаций для получения своевременной информации о состоянии на дорогах, в том числе влияющей на безопасность в пути, а также преобразовать транспортные узлы в общественные центры. Любое транспортное средство, подключенное к сети Интернет, автоматически становится сегодня частью Интернета вещей и может участвовать в динамичной цифровой экосистеме.

Начальным этапом развития рынка Интернета вещей для транспорта можно считать по-

явление первых подключенных устройств для навигации и оптимизации маршрута. К таким устройствам можно отнести автомобильные навигаторы с возможностью подключения по каналам сотовой связи, а также спутниковые трекеры для мониторинга и управления автопарком Fleet Management [10].

С ростом проникновения Интернета и появлением подключенных ноутбуков, планшетов, смартфонов спрос на персональные навигаторы существенно снизился. На смену им пришли GSM-автосигнализации, позволяющие удаленно наблюдать за автомобилем и управлять им через мобильное приложение на смартфоне/планшете.

Самое очевидное продолжение развития Интернета вещей в транспорте, за которое уже взялись Tesla и BMW — внедрение автопилота, который, в отличие от человека, не ошибается, не устает и не засыпает, и в перспективе поможет снизить число аварий на дорогах по вине водителя. Уже к 2020 г. Япония планирует вывести на улицы Токио множество самоуправляемых автомобилей, в том числе электромобилей. В России Камаз и Россельмаш также работают над тем, чтобы научить свои машины двигаться без участия водителя [11].

Сегодня больше половины активностей в транспортной отрасли, связанных с Интернетом вещей, можно отнести к "умному транспорту", потребительским приложениям, транспортным перевозкам. Толчком к развитию послужило распространение смартфонов, доля которых в 2017 г. приблизилась к 50 % сотовых устройств в России [8]. Именно за счет смартфонов, которые берут водители с собой в дорогу, стало возможно использовать системы навигации и мониторинг загруженности дорог. Ряд эффектов можно достичь и при производстве транспорта, в этом случае эффекты будут сравнимы с аналогичными в других отраслях: снижение издержек за счет своевременной идентификации брака, прогнозов отклонений, корректировки настроек производственного цикла в режиме реального времени, а также благодаря своевременным изменениям и контролю качества. Рассмотрим каждую из областей применения Интернета вещей, характерных для транспортной отрасли, более подробно.

Транспортные перевозки. Говоря о транспортных перевозках, стоит рассматривать категории воздушного, автомобильного, железнодорожного и морского транспорта. В целом во всех этих категориях Интернет вещей направлен на то, чтобы предоставлять пользователям/пассажирам актуальную информацию, улучшать реагирование в чрезвычайных ситуациях и осуществлять более точное планирование загрузки транспортного средства и его маршрута.

Существенным технологическим прорывом для мониторинга транспортных средств послужило появление и развитие облачных сервисов SaaS (Software as a Service — программное обеспечение как услуга). Благодаря этим сервисам пользователям стали доступны сервисы мониторинга своих транспортных средств из любой точки мира и через любое устройство, подключенное к сети Интернет.

Следующий шаг для интеллектуального мониторинга в транспортной отрасли стал возможен благодаря установке систем контроля за расходом топлива и удаленного мониторинга передвижения транспорта на базе датчиков GPS и системы навигации ГЛОНАСС. С помощью датчиков мониторинга стало возможно снижение затрат за счет контроля простоев, оптимизации маршрута и топлива, анализа данных маршрута и его более тщательного планирования. В России, в целях безопасности, нормативно закреплены обязанности устанавливать системы дистанционного мониторинга и контроля движения транспорта для коммерческих перевозок пассажиров и перевозки опасных грузов, а также оснащать все автомобили системой экстренного оповещения "ЭРА-ГЛОНАСС" [12].

Для транспортной отрасли характерны понятия пассажирского пути, пути грузоперевозок, ЖД, авиа и морских перевозок. Этот путь описывает все процессы, начиная с момента возникновения потребности и оценки вариантов перевозки, заканчивая состоянием при перевозке, прибытием, контролем в точке прибытия, доставкой в конечную точку. Имея информацию о таком пути, можно продумать улучшения на основе анализа мониторинга и анализа данных на каждом этапе движения пассажира или груза. Для этого в транспорте широко используют системы отслеживания маршрута транспорта, мониторинг грузоперевозок, контроль отгрузки и складирования.

Множество решений разрабатывается для более безопасной эксплуатации транспортных средств. Так, технология Connected Cars ("подключенные автомобили") позволяет использовать системы экстренного вызова скорой помощи со встроенной SIM-карты, что значительно увеличивает скорость оказания первой помощи при ДТП. Датчики эксплуатации и контроля состояния могут быть интегрированы в системы поездов и общественного транспорта. Получение актуальной информации с датчиков движения транспорта позволит сократить расход топлива за счет мониторинга скорости движения в зависимости от маршрута, обеспечить контроль движения, мониторинг нештатных ситуаций.

Установка датчиков загрузки является актуальной задачей для всех видов транспорта. Тщатель-

ный мониторинг показателей датчиков позволяет максимально правильно загрузить транспорт, что снижает расходы на топливо и положительно влияет на общие затраты, оптимизацию логистики. Датчики исправности, отвечающие за оповещение об ошибках, также повышают безопасность, снижают затраты на ремонт и диагностику транспортного средства.

Недавно компания *Orange* совместно с *Peugeot Citroën* [13] разработала решение для управления автопарком, которое поможет анализировать стиль вождения, отслеживать потребление топлива и сокращать расходы. Система будет уведомлять о возникновении неполадок и необходимости технического обслуживания, а также отслеживать местоположение водителей, активов и товаров с помощью GPS.

Удаленный мониторинг вождения пользователей также может быть полезен компаниям автострахования для расчета оптимальной стоимости полиса в зависимости от стиля и манеры вождения, аккуратности на дорогах, соблюдения скоростного режима и правил дорожного движения.

Железнодорожные перевозки также стали объектом для применения Интернета вещей. Так, крупнейший железнодорожный оператор *Trenitalia* [14], который ежедневно перевозит 2 млн пассажиров в Италии, использует Интернет вещей для анализа данных от тысяч встроенных датчиков в реальном времени, получая возможность радикально повысить качество процессов технического обслуживания. В Нидерландах на железных дорогах используется передовое программное обеспечение, которое анализирует 56 тыс. переменных, включая состояние железных дорог и уровень пассажирского спроса. Благодаря анализу этих данных перевозчик отправляет более 5 тыс. поездов в день, повышает операционную эффективность на 6 % и экономит около 20 млн евро в год [15]. Такой эффект достигается за счет анализа расхода топлива в целях определения оптимального режима скорости по маршруту при разном тоннаже и числе остановок.

В авиации Интернет вещей также используется преимущественно для отслеживания перевозок в режиме реального времени, контроля состояния износа оборудования, что положительно влияет на снижение затрат на техническое обслуживание и управление оборудованием воздушного флота.

"Умный" транспорт ("умный" город). Интернет вещей в части умного транспорта можно условно разделить на "умные" транспортные средства и городские магистрали, беспилотный транспорт и электромобили.

Говоря об "умных" городе и транспорте, отдельно стоит упомянуть и "умную парковку". В современных больших городах парковка для

многих водителей является большой проблемой. Водители вынуждены кружить в поисках свободного парковочного места, что приводит к потере топлива, выбросам углекислого газа и нервным срывам водителей. "Умные парковки", оснащенные специальными датчиками, осуществляют поиск свободных мест, собранные данные отправляют на сервер, который в реальном времени через мобильное приложение информирует пользователей о свободных парковочных местах. Такая информация может быть связана с автомобильными навигационными системами и позволит оптимизировать трафик. Жители крупных городов уже знают, что оплачивать парковку, указывать продолжительность стоянки, отслеживать время можно через специальное приложение на смартфоне.

Кроме того, Интернет вещей активно применяют для оптимизации потоков транспорта в городах. Например, в Москве в 2007 г. появились первые интеллектуальные светофоры на опытном участке протяженностью 7,5 км [16]. Каждый интеллектуальный светофор является частью большой сети, принимая информацию с видеокамер о плотности потока и получая возможность пропускать машины с минимальной задержкой в зависимости от загрузки магистрали. Информация передается в единый центр управления системой по беспроводной связи и используется для оптимального регулирования транспортного потока. Такая система позволяет двигаться потоку машин более эффективно, существенно экономить время и деньги автомобилистов. Так, в США анализ работы системы на девяти пилотных перекрестках показал уменьшение времени ожидания сигнала светофора на 40 %, уменьшение времени поездки на 26 %, и сокращение выбросов на 21 % [17].

Все активнее в интеллектуальных транспортных системах используют детекторы транспортного потока, которые позволяют наблюдать за движением транспорта, оценивать его скорость и состояние дорог. Такие системы могут быть оснащены микроволновым радаром для измерения скорости, ультразвуковым детектором для оценки габаритных размеров и классификации транспортных средств по классам, многоканальным инфракрасным детектором для обеспечения подсчета автомобилей и определения интенсивности движения, измерения загруженности транспортных каналов и их оптимизации. Скоро с помощью подключенных датчиков станет возможным создавать наиболее эффективную транспортную систему, ответить на вопросы, где строить новую развязку, в каком направлении запускать новый маршрут.

Что касается беспилотного транспорта, стоит отметить значительный прогресс многих компаний. Так, на улицах Вашингтона с 2016 г. испы-

тывают самоуправляемый автобус Olli, который разработан компанией *Local Motors* при сотрудничестве с IBM [18]. Автобус умеет выполнять своевременное распознавание дорожных знаков с помощью высокотехнологичной стереокамеры, избегать столкновений ему помогают современные радары.

Компания *Cognitive Technologies* совместно с Камаз до конца 2017 г. обещает выпустить промышленную версию грузовика с элементами искусственного интеллекта, который сможет самостоятельно передвигаться по скоростным магистралям с помощью подключенных датчиков для повышения своей энергетической эффективности и безопасности [19]. Все грузовики автопарка будут иметь встроенный GPS, накапливать данные друг от друга и окружающей среды вокруг них и таким образом смогут реагировать на внешние события быстрее и безопаснее, чем люди. По прогнозам экспертов McKinsey [20], беспилотные грузовые и легковые автомобили, которые будут передвигаться более медленно и соблюдая правила, могут привести к снижению аварий на 90 %, если получат широкое применение. Использование мониторинга с помощью датчиков Интернета вещей позволит более точно контролировать и выбирать наиболее оптимальную скорость, сократить расходы на топливо и выбрать лучшие маршруты на основании данных о загруженности дорог.

По прогнозам компании A. T. Kearney, к 2030 г. объем рынка беспилотных автомобилей и связанных с ними сервисов должен составить 282 млрд долл. США. По оценкам *Nissan*, развитие беспилотных автомобилей будет ежегодно добавлять 0,15 % к годовым темпам роста ВВП в Европе и к 2050 году принесет экономике стран Евросоюза 17 трлн евро [21].

Приложения для Интернета вещей. Потребительские приложения для Интернета вещей, как правило, позволяют пользователям, в том числе промышленным компаниям, получить доступ к мониторингу своих "вещей" в любое время и в любом месте через устройство, подключенное к Интернет. "Вещами" могут быть любые устройства, транспортные средства, дороги, склады и помещения, светофоры, детали, элементы производственного цикла (машины, агрегаты), даже люди. Такие "вещи" передают информацию датчиков, сообщают о своем состоянии, позволяя пользователю анализировать полученную информацию и принимать решения с помощью удобных интерфейсов с подключенными устройствами.

Существует множество примеров пользовательских приложений в транспорте, начиная от навигационных приложений и средств мониторинга и видеонаблюдения за транспортным средством, заканчивая приложениями, позво-

ляющими осуществить заказ блюд на борту самолета или предсказать расход топлива, предложить оптимальную скорость для его экономии. Однако стоит отметить, что наибольшую пользу представляют так называемые промышленные приложения — приложения, относящиеся к промышленному Интернету вещей.

Известный пример, который наглядно демонстрирует преимущества промышленного приложения для Интернета вещей, — это крупнейший порт Германии в Гамбурге, где уже много лет существуют транспортная и логистическая проблемы. Экспедиционные компании теряют много времени в пробках, в поисках парковочных мест, пропускной способности дорог в порту недостаточно для огромного объема перевозок, а возможности для строительства новых дорог сильно ограничены. В результате около 70 % грузовиков приезжали раньше времени, создавались пробки, что являлось главным ограничением роста грузоперевозок. Оценивая процесс с точки зрения клиента порта, рабочая группа SAP [22] предположила, что если они смогут отслеживать движение транспорта задолго до прибытия в порт, то это существенно упростит логистику и расширит возможности диспетчеров. Было исследовано распределение рейсов транспортных компаний и выявлено, что примерно треть фур едет с территории Германии. Благодаря тому, что отслеживание местоположения фур делается в момент начала движения, удалось рассчитать оптимальное расписание служб порта: управление въездами, выездами, парковками грузовиков, контейнерными погрузчиками.

Теперь управляющие структуры получили возможность с помощью датчиков, установленных в мобильных устройствах, следить за каждым грузовиком, контейнером или судном и в режиме реального времени полностью контролировать движение в порту через мобильное приложение. Информация о местоположении фур передается с мобильных устройств в единый центр обработки, который развернут на платформе SAP Cloud Platform. Разные пользователи системы: дорожная служба порта, распределитель контейнеров, служба парковки получают информацию через веб-интерфейс на свои устройства. На основе информации о заказах на обслуживание и о местоположении фур службы порта могут в режиме реального времени оптимизировать распределение мест погрузки и разгрузки грузовиков. За счет внедрения системы логистические компании экономят топливо и уменьшают время доставки грузов.

В целом и пользовательские, и промышленные приложения для Интернета вещей позволяют пользователю дистанционно наблюдать, получать и анализировать точные данные, управ-

лять транспортными объектами через мобильное приложение на смартфоне/планшете, что в итоге позитивно сказывается на экономии затрат всех транспортных и логистических процессов.

Заключение

В России достаточно много игроков, участвующих как в развитии Интернета вещей, так и в транспортной отрасли. Вместе с уникальными новыми возможностями приходят и новые проблемы — это безопасность, бесперебойное функционирование в режиме реального времени, необходимость высококвалифицированных специалистов для поддержки новых цифровых технологий. Чтобы компании смогли получить максимум выгоды от Интернета вещей, они должны быть готовы оценить возможные риски — от кибератак и рисков информационной безопасности, потери персональных данных, несанкционированного доступа к подсоединенным машинам и устройствам до риска неграмотного управления удаленными устройствами, что может привести к негативным последствиям.

Возможности использования новых технологий поднимают множество вопросов в отношении ответственности, например, кто или какое устройство несет ответственность в случае аварии беспилотного транспортного средства? Как должен быть запрограммирован беспилотный автомобиль, который, чтобы избежать возможного столкновения, с высокой вероятностью провоцирует другое ДТП? Этот вопрос еще до конца не изучен, существует множество споров по поводу ответственности устройств, управляемых дистанционно.

Конечно, Интернет вещей дает массу преимуществ. Так, например, "умный" автомобиль самостоятельно отслеживает износ и интервалы замены деталей, сообщает о неисправностях, но при этом не решает проблему. Конечная ответственность все равно лежит на владельце процесса и продукта, технологии лишь предоставляют больше информации для принятия наиболее верного решения.

Стоит отметить, что технология Интернета вещей становится все более доступна обычному пользователю. Как утверждают аналитики компании Goldman Sachs, стоимость датчиков — важная составляющая Интернета вещей — резко упала, более чем на 50 % за последнее десятилетие, при том, что они стали более эффективными [23]. Судя по всему, такая тенденция будет развиваться, и вскоре нас будут окружать умные дороги, машины и приложения, которые будут подсказывать следующий шаг. И тогда перед пользователем встанет другой вопрос — сможет

ли он обработать такое большое количество информации, которое предоставляют ему различные датчики и устройства, насколько он сможет делегировать машинам принятие решений и что он оставит себе.

Список литературы

1. **Meulen R.** Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016 // Gartner, 2017 [Официальный сайт]. URL: <http://www.gartner.com/newsroom/id/3598917> (дата обращения: 15.07.2017).
2. **DuBravac S., Ratti C.** The Internet of Things: Evolution or Revolution? // American International Group (AIG), 2016 [Официальный сайт]. URL: <https://www.aig.com/content/dam/aig/america-canada/us/documents/insights/aig-white-paper-iot-english-digital-brochure.pdf> (дата обращения: 15.07.2017).
3. **Банке Б., Бутенко В., Котов И., Рубин Г., Тушен Ш., Сычева Е.** Россия онлайн? Догнать нельзя отстать // The Boston consulting Group (BCG), 2016 [Официальный сайт]. URL: http://image-src.bcg.com/Images/BCG-Russia-Online_tcm27-152058.pdf (дата обращения: 15.07.2017).
4. **Jenner G.** Analysis: How airlines are tapping into the Internet of Things // GE [Официальный сайт]. URL: <https://www.ge.com/digital/press-releases/how-airlines-are-tapping-internet-things> (дата обращения: 15.07.2017).
5. **Driving** productivity in the Pilbara // RioTinto, 2016 [Официальный сайт]. URL: http://www.riotinto.com/ourcommitment/spotlight-18130_18328.aspx (дата обращения: 15.07.2017).
6. **Тренина А.** Вещь в сети. Промышленный интернет // Коммерсант, 2016 [Официальный сайт]. URL: <http://www.kommersant.ru/doc/3036782> (дата обращения: 15.07.2017).
7. **Puthiyamadam T., Garrett D., Curran C.** 2017 Digital IQ How organizations can maximize and profit from technology investments // PwC, 2017 [Официальный сайт]. URL: <https://www.pwc.com/us/en/advisory-services/digital-iq.html> (дата обращения: 15.07.2017).
8. **Пуха Ю., Сидоров Г., Учуваткин М., Стапран Д.** Перспективы развития "Интернета вещей" в России // PwC, 2017 [Официальный сайт]. URL: http://www.pwc.ru/ru/communications/assets/the-internet-of-things/PwC_Internet-of-Things_Rus.pdf (дата обращения: 15.07.2017).
9. **О'Доннелл Л.** IDC: IoT-рынок — кладезь возможностей для VAR'ов // CRN, 2017 [Официальный сайт]. URL: <https://www.crn.ru/news/detail.php?ID=120078> (дата обращения: 15.07.2017).
10. **Ермолич П.** Подключенные Автомобили (Connected Cars). Перспективы российского рынка M2M/IoT в транспортной отрасли до 2020 года // JSON, 2017 [Официальный сайт]. URL: http://json.tv/ict_telecom_analytics_view/perspektivy-rossiyskogo-rynka-m2miot-v-transportnoy-otrasli-do-2020-goda-20151211043329 (дата обращения: 15.07.2017).
11. **Рагимова С.** Точка зрения машины. Интернет вещей, *Коммерсант. Business Guide*. 12/2016 [Официальный сайт]. URL: <https://www.kommersant.ru/doc/3168705> (дата обращения: 15.07.2017).
12. **Вопросы и ответы, Эра-глонасс**, 2017 [Официальный сайт]. URL: <http://glonassunion.ru/era-glonass/info> (дата обращения: 15.07.2017).
13. **Kitney R.** 5 областей, которые изменит "Интернет вещей" // Orange Business Services, 2017 [Официальный сайт]. URL: <http://www.orange-business.com/ru/blogs/get-ready/mobilnost/5-oblastey-kotorye-izmenit-internet-veshchey> (дата обращения: 15.07.2017).
14. **Magyar J.** Trenitalia's IoT Strategy Makes the Trains Run on Time // SAP [Официальный сайт]. URL: <http://news.sap>

com/trenitalias-iot-strategy-makes-the-trains-run-on-time/ (дата обращения: 15.07.2017).

15. **Большие** данные в разных отраслях: сценарии применения // CNews, 2014 [Официальный сайт]. URL: http://www.cnews.ru/articles/bolshie_dannye_v_raznyh_otraslyah_stsenarii (дата обращения: 15.07.2017).

16. **Ермолич П.** "Умная дорога" — российский рынок M2M/IoT в области дорожной инфраструктуры [Электронный ресурс] // JSON, 2017 [Официальный сайт]. URL: http://json.tv/ict_telecom_analytics_view/umnaya-doroga-rossiyskiy-rynok-m2miot-v-oblasti-dorojnoj-infrastruktury-20160314040956 (дата обращения: 15.07.2017).

17. **Heineman K.** Intelligent Traffic Signals Reduce Intersection Wait 40 Percent // Tchnologist, GE, 2017 [Официальный сайт]. <http://txchnologist.com/post/63090236446/intelligent-traffic-signals-reduce-intersection-online-magazine> (дата обращения: 15.07.2017).

18. **Smigala H., Kress A.** Local Motors Debuts "Olli", the First Self-driving Vehicle to Tap the Power of IBM Watson // IBM, 2016 [Официальный сайт]. URL: <https://www-03.ibm.com/press/us/en/pressrelease/49957.wss> (дата обращения: 15.07.2017).

19. **Первый камаз** с искусственным интеллектом выпустят уже в 2017 году // Камаз, 2016 [Официальный сайт]. <http://www.kamaz-volga.ru/about/smi/39350/> (дата обращения: 15.07.2017).

20. **Интернет вещей** // American International Group (AIG), 2016 [Официальный сайт]. <https://www.aig.ru/why-aig/internet-of-things> (дата обращения: 15.07.2017).

21. **Дембинская Н.** Как беспилотные автомобили изменят экономику // РИА Новости, 2016 [Официальный сайт]. URL: <https://ria.ru/economy/20161205/1482890294.html> (дата обращения: 15.07.2017).

22. **Simpler and Smarter Connections at Germany's Largest Seaport** // SAP, 2017 [Официальный сайт]. URL: <https://www.sap.com/about/customer-testimonials/public-sector/hamburg-port-authority.html> (дата обращения: 15.07.2017).

23. **Jankowski S., Covello J., Bellini H., Ritchie J., Costa D.** The Internet of Things: Making sense of the next mega-trend // Goldman Sachs, 2014 [Официальный сайт]. URL: <http://www.goldmansachs.com/our-thinking/outlook/internet-of-things/iot-report.pdf> Goldman Sachs Global Investment Research 2014 (дата обращения: 15.07.2017).

S. M. Avdoshin, Ph. D., Professor, Head of Software Engineering School,
Faculty of Computer Science, savdoshin@hse.ru,

E. Yu. Pesotskaya, Associate Professor, Faculty of Computer Science, epesotskaya@hse.ru
National Research University Higher School of Economics (HSE), Moscow, 105187, Russia

Internet of Things: Transportation

The paper describes the possibility of using the Internet of Things in transport industry and processes while all the life cycle of transporting passengers and goods. Authors provide detailed research of effects and benefits by using Internet of Things while transportation, "smart cars and vehicles", customer applications that drives cost optimization and economic effect. The paper provides the analysis of potential trends in development of Internet of Things in transportation sector forced by information and digital technologies development recent years. Also the productivity and cost savings aspects are carefully examined as well as their impact to the company's transportation and logistics business processes. While providing recommendations on business processes improvement and costs and risks reducing initiatives, the global trends are considered.

Keywords: Internet of Things, mobile technologies, mobile applications, transportation, "smart cars", intellectual transport systems

References

1. **Meulen R.** Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016, Gartner, 2017 [Official website]. URL: <http://www.gartner.com/newsroom/id/3598917> (accessed 15.07.2017).

2. **DuBravac S., Ratti C.** The Internet of Things: Evolution or Revolution? American International Group (AIG), 2016 [Official website]. URL: <https://www.aig.com/content/dam/aig/america-canada/us/documents/insights/aig-white-paper-iot-english-digital-brochure.pdf> (accessed 15.07.2017).

3. **Banke B., Butenko V., Kootov I., Rubin G., Toushen S., Sycheva E.** Russia online? Catch up impossible to fall behind, *The Boston Consulting Group (BCG)*, 2016 [Official website]. URL: http://image-src.bcg.com/Images/BCG-Russia-Online_tcm27-152058.pdf (accessed 15.07.2017).

4. **Jenner G.** Analysis: How airlines are tapping into the Internet of Things, *GE* [Official website]. URL: <https://www.ge.com/digital/press-releases/how-airlines-are-tapping-internet-things> (accessed 15.07.2017).

5. **Driving** productivity in the Pilbara, *RioTinto*, 2016 [Official website]. URL: http://www.riotinto.com/ourcommitment/spotlight-18130_18328.aspx (accessed 15.07.2017).

6. **Tronina A.** Thinks in Net. Industrial Internet, *Kommersant*, 2016 [Official website]. <http://www.kommersant.ru/doc/3036782> (accessed 15.07.2017) (in Russian).

7. **Puthiyamadam T., Garrett D., Curran C.** 2017 Digital IQ How organizations can maximize and profit from technology investments, *PwC*, 2017 [Official website]. URL: <https://www.pwc.com/us/en/advisory-services/digital-iq.html> (accessed 15.07.2017).

8. **Puha Yu., Sidorov G., Uchvatkin M., Stapran D.** Perspectives of "Internet of Things" in Russia, *PwC*, 2017. URL: http://www.pwc.ru/ru/communications/assets/the-internet-of-things/PwC_Internet-of-Things_Rus.pdf (accessed 15.07.2017) (in Russian).

9. **O'Donnell L.** IDC: IoT-market — great possibilities for VAR, *CRN*, 2017 [Official website]. URL: <https://www.crn.ru/news/detail.php?ID=120078> (accessed 15.07.2017) (in Russian).

10. **Ermolich P.** Connected Cars. Perspectives of Russian market M2M/IoT in transportation till 2020, *JSON*, 2017 [Official website]. URL: http://json.tv/ict_telecom_analytics_view/perspektivy-rossiyskogo-rynka-m2miot-v-transportnoy-otrasli-do-2020-goda-20151211043329 (accessed 15.07.2017) (in Russian).

11. **Ragimova S.** Machine Point of View. Internet of Things, *Kommersant Business Guide*, 12/2016 [Official website]. URL:

<https://www.kommersant.ru/doc/3168705> (accessed 15.07.2017) (in Russian).

12. **Questions** and answers, *Era-Glonass*, 2017 [Official website]. URL: <http://glonassunion.ru/era-glonass/info>, Era-Glonass (accessed 15.07.2017).

13. **Kitney R.** 5 oblastey, kotorye ismenyat "internet veshey", Orange Business Services, 2017 [Official website]. URL: <http://www.orange-business.com/ru/blogs/get-ready/mobilnost/5-oblastey-kotorye-izmenit-internet-veshchey> (accessed 15.07.2017) (in Russian).

14. **Magyar J.** Trenitalia's IoT Strategy Makes the Trains Run on Time, *SAP* [Official website]. URL: <http://news.sap.com/trenitalias-iot-strategy-makes-the-trains-run-on-time/> (accessed 15.07.2017).

15. **Big Data** in different industries: application scenarios [Electronic resource], *CNews*, 2014, [Official website]. URL: http://www.cnews.ru/articles/bolshie_dannye_v_raznyh_otraslyah_stsenarii (accessed 15.07.2017) (in Russian).

16. **Ermoloh P.** "Smart Road" — Russian market of M2M/IoT in transport infrastructure, *JSON*, 2017 [Official website]. URL: http://json.tv/ict_telecom_analytics_view/umnaya-doroga-rossiyskiy-rynok-m2miot-v-oblasti-dorojnoj-infrastruktury-20160314040956 (accessed 15.07.2017) (in Russian).

17. **Heineman K.** Intelligent Traffic Signals Reduce Intersection Wait 40 Percent, *Txchnologist*, GE, 2017 [Official website]. <http://txchnologist.com/post/63090236446/intelligent-traffic-signals-reduce-intersection-online-magazine> (accessed 15.07.2017).

18. **Smigala H., Kress A.** Local Motors Debuts "Olli", the First Self-driving Vehicle to Tap the Power of IBM Watson [Electronic resource], *IBM*, 2016 [Official website]. URL: <https://www-03.ibm.com/press/us/en/pressrelease/49957.wss> (accessed 15.07.2017).

19. **Perviy** Kamaz s iskustvennym intellektom vipystyat yje v 2017 godu [Electronic resource], *KAMAZ*, 2016 [Official website] <http://www.kamaz-volga.ru/about/smi/39350/> (accessed 15.07.2017) (in Russian).

20. **Internet** of Things [Electronic resource], *American International Group (AIG)*, 2016 [Official website]. <https://www.aig.ru/why-aig/internet-of-things> (accessed 15.07.2017).

21. **Dembinskaya N.** Kak bespilotnie avtomobili izmenyat economy, *RIA Novosti*, 2016 [Official website]. URL: <https://ria.ru/economy/20161205/1482890294.html> (accessed 15.07.2017) (in Russian).

22. **Simpler** and Smarter connections at Germany's Largest Seaport [Electronic resource], *SAP*, 2017 [Official website]. URL: <https://www.sap.com/about/customer-testimonials/public-sector/hamburg-port-authority.html> (accessed 15.07.2017).

23. **Jankowski S., Covello J., Bellini H., Ritchie J., Costa D.** The Internet of Things: Making sense of the next mega-trend, *Goldman Sachs*, 2014 [Official website]. URL: <http://www.goldmansachs.com/our-thinking/outlook/internet-of-things/iot-report.pdf> Goldman Sachs Global Investment Research 2014 (accessed 15.07.2017).

УДК 004.057.4

Ле Ба Чунг, аспирант, e-mail: chungbaumanvietnam@gmail.com,
Московский физико-технический институт (ГУ),

Ю. А. Холопов, вед. инженер,
Институт точной механики и вычислительной техники им. С. А. Лебедева
Российской академии наук, Москва

Асимметричный межмодульный интерфейс

Рассмотрены межмодульные связи в цифровой системе управления, построенной по принципу "вынесенной руки". Предложен простой асимметричный межмодульный интерфейс для передачи информации между периферийными устройствами и вычислителем. Интерфейс отличается несложным механизмом управления обменом, простой структурой пакетов и высокой плотностью данных в пакете, что позволяет организовать быстрые "прозрачные" связи между удаленными устройствами системы управления и центральным вычислителем.

Ключевые слова: цифровая система управления, принцип "вынесенной руки", межмодульные связи, центральный вычислитель, периферийные устройства

Введение

Аппаратные средства цифровых систем управления (ЦСУ), как правило, конструктивно распределены в нескольких модулях. Модуль реализуется как узел с четко выделенной функцией и минималистичностью интерфейсов. В любой ЦСУ присутствуют модули исполнительных устройств и датчиков, размещенные непосредственно на объекте управления. В точках их размещения воз-

можны неблагоприятные условия эксплуатации [1], такие как высокая температура, вибрация, агрессивная среда и т.п., а сами точки размещения после окончательной сборки объекта управления становятся физически недоступными, что исключает возможность изменения алгоритмов управления модулями через локальные технологические интерфейсы. Тем не менее возможность коррекции управляющих программ должна быть предусмотрена, и функция коррекции должна

быть доступной в течение всего жизненного цикла объекта управления.

Информационная среда ЦСУ с нужными свойствами может быть реализована широко распространенным способом — совмещением функций технологических и рабочих интерфейсов в едином — комплексном — интерфейсе. Но, по мнению авторов, правильным подходом к построению информационной среды с корректируемыми управляющими программами является выбор структуры информационной среды, принципиально не требующей изменения функций удаленных узлов.

Нужным свойством обладает информационная среда ЦСУ, построенная по принципу "вынесенной руки" (далее — ЦСУВР), с единым централизованным вычислителем. В такой системе управления датчики и исполнительные устройства должны быть функционально независимыми от алгоритмов управления, а алгоритмы управления всех уровней реализуются в центральном вычислителе. Периферийные устройства ЦСУВР (датчики и актуаторы) имеют сходный функциональный состав и поэтому могут взаимодействовать с центральным вычислителем через однотипные унифицированные интерфейсы.

Целью работы является создание простого, типового интерфейса для передачи информации между периферийными устройствами и центральным вычислителем в цифровых централизованных системах управления. Свойства разрабатываемого интерфейса, как мы покажем ниже, существенно отличаются от свойств существующих стандартных интерфейсов связи с периферийными устройствами.

Система управления, построенная по принципу "вынесенной руки"

Идеологически концепция "вынесенной руки" означает возможность дистанционного управления объектом или его разнесенными составными частями в том же масштабе реального времени, что и при непосредственном управлении. То есть время передачи сигналов управления объектом из центрального вычислителя намного меньше времени реакции объекта на их изменение. Аналогичное требование предъявляется для передачи центральному вычислителю параметров состояния объекта управления.

Цифровая система управления с центральным вычислителем может быть реализована несколькими

способами. Уменьшение времени обменов может быть достигнуто за счет использования для управления объектом высокоуровневых команд управления (ВКУ). В такой иерархической системе управления непосредственно с исполнительными устройствами, объектом управления и окружающей средой взаимодействуют контуры регулирования нижнего уровня, реализованные в виде автономных модулей. В наборе параметров состояния, передаваемых в центральный вычислитель (далее — "центр") могут практически отсутствовать данные с датчиков, а сам контур регулирования будет иметь свойства высокоавтономного, почти независимого исполнителя. Алгоритмы функционирования ЦСУ с ВКУ отлаживаются поэтапно: сначала — функции узлов нижнего уровня до их установки на объекте управления и с использованием для перепрограммирования технологических интерфейсов. Затем общие алгоритмы управления объектом (в терминах ВКУ) отлаживаются после монтажа вынесенных узлов на их штатные места. Изменение алгоритмов работы узлов на этом этапе становится невозможным.

Более гибкая реализация ЦСУ — система управления, построенная по принципу "вынесенной руки", — предполагает использование центрального вычислителя как единственного исполнителя алгоритмов управления всех уровней. Необходимость распараллеливания вычислений по нескольким процессорам при использовании современных программно-управляемых вычислителей отсутствует, так как скорость обработки единственным процессором практически не ограничена.

В такой системе управления функциональный состав оконечных устройств минимален (рис. 1) и включает в себя трансиверы межмодульного интерфейса и локальные подсистемы: измерительную и исполнительную, а также не изображенный на рисунке блок быстродействующей автономной защиты от критичных режимов. Реализация защиты такого типа через "центр" может оказаться невозможной из-за критичности

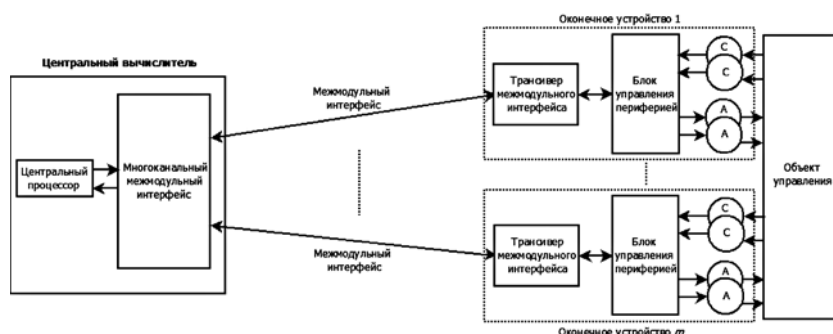


Рис. 1. Структура цифровой системы управления, построенной по концепции "вынесенной руки"

дополнительных задержек в межуровневых связях. Для центрального вычислителя вынесенные периферийные устройства выполняют функции ввода-вывода, а вся переменная алгоритмическая часть, связывающая параметры состояния и управления, оказывается сконцентрированной только в нем. Теперь изменение любого алгоритма в любой подсистеме управления ЦСУВР осуществляется только в центральном вычислителе.

Средства ввода-вывода ЦСУВР размещаются в оконечных устройствах, непосредственно взаимодействующих с объектом управления. Они расположены удаленно относительно центрального вычислителя, но скорость взаимодействия "центра" с периферийными устройствами должна быть такой, чтобы в системе не была заметна эта удаленность.

Традиционно для удаленного подключения используются сетевые или подобные им интерфейсы. Рассмотрим подробнее свойства стандартных интерфейсов сетевого типа и покажем их недостатки при использовании в роли периферийных, межмодульных интерфейсов в системе управления, построенной по принципу "вынесенной руки".

Анализ существующих стандартных интерфейсов

Разрабатываемый межмодульный интерфейс должен обладать функциональным свойством "прозрачного", практически асинхронного средства пакетного обмена потоками данных между узлами (рис. 2). Объясним это утверждение подробнее. Межмодульный интерфейс осуществляет обмен параметрами состояния и управления между узлами: центральным вычислителем и периферийными устройствами распределенной системы управления. Из-за отсутствия прямой связи между источниками и приемниками параметров обновление переменной в приемнике происходит с задержкой относительно момента ее изменения источником. Кроме того, в при-

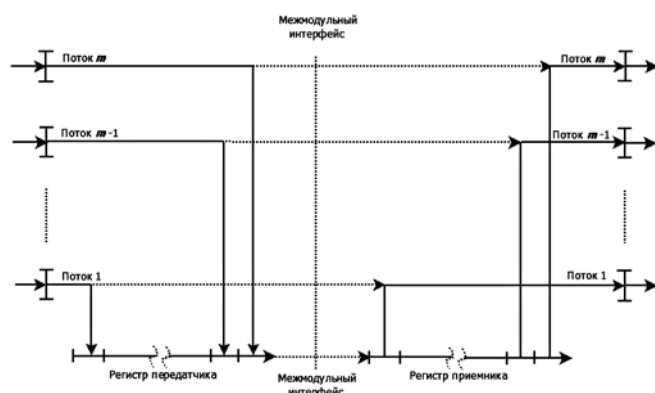


Рис. 2. Пакетный обмен потоками данных

нятой концепции построения связей между узлами ЦСУ ведущий узел не имеет возможности определять момент использования переменных, принятых узлом ведомым. Поэтому стратегия обменов информацией сводится к организации как можно более частого "логического замыкания" связей между всеми источниками параметров и их приемниками. Скоростные характеристики современных приемопередатчиков позволяют организовать обновление информации между узлами с частотой в несколько килогерц, что вполне достаточно, например, для прямого управления ШИМ-генераторами исполнительного устройства из центрального вычислителя.

Прозрачность интерфейса реализуется через цикличность обменов, происходящих с максимально возможной частотой. Другие свойства интерфейса — следствия, вытекающие из этого основного свойства. Поэтому дальнейшее сравнение интерфейсов будет проводиться в терминах оценки их пригодности для реализации описанного выше свойства.

Большая часть стандартных сетевых интерфейсов построена в соответствии с идеологией шинной архитектуры [2], т. е. использует общую среду передачи данных в режиме разделения времени (рис. 3).

Шинная топология связей не позволяет получить максимальную скорость обменов из-за сложностей согласования параметров длинных линий. Поэтому в современных сетевых интерфейсах связи между узлами реализуются в топологии "точка—точка", но на логическом уровне присутствуют следующие структурные свойства, снижающие скорость обменов потоками данных:

- ✓ механизм вхождения в связь;
- ✓ функция восстановления связи в случае возникновения сбоя;
- ✓ механизм разделения пакетов с помощью временных интервалов;
- ✓ наличие лишних полей (адресации, длины данных и т.д.) в коммуникационных пакетах [3];
- ✓ отсутствие цикличности обменов.

Цикличность в таких интерфейсах может быть реализована с использованием внешнего управления: аппаратными средствами, что усложняет интерфейс, или программными, что снижает частоту обменов.

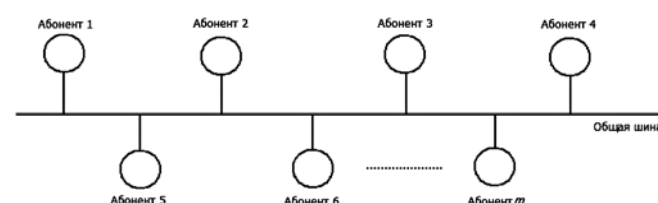


Рис. 3. Шинная топология

Наиболее близким по свойствам к междомодульному интерфейсу является широко распространенный интерфейс SPI [4]. В этом интерфейсе легко реализуются цикличность и переменная длина поля данных. Как недостаток интерфейса SPI можно отметить избыточное число сигнальных связей.

Сформулируем требования к междомодульному интерфейсу, учитывающие его функциональное назначение и свободные от описанных выше недостатков. Разрабатываемый междомодульный интерфейс должен иметь следующие технические характеристики:

- аппаратно реализованную цикличность обменов;
- последовательную пакетную передачу данных с минимальным числом сигнальных связей между абонентами;
- максимальную пропускную способность обменов потоками данных в полудуплексном режиме за счет отсутствия пауз и упрощенного формата пакетов;
- стабильность периода разовой циклической транзакции;
- стабильность фазы транзакций.

Некоторые из перечисленных характеристик не являются очевидными. Их актуальность будет пояснена ниже.

Асимметричный междомодульный интерфейс

На основе названных требований был разработан "асимметричный" междомодульный интерфейс [5], предназначенный для обменов информацией между периферийными устройствами и центральным вычислителем.

На рис. 4 показана событийно-временная диаграмма циклического взаимодействия ведущего (центрального вычислителя) и ведомого (оконечного устройства). Каждая транзакция обмена информацией между ведущим и ведомым распадается на восемь фаз.

На рис. 4 приняты следующие обозначения:

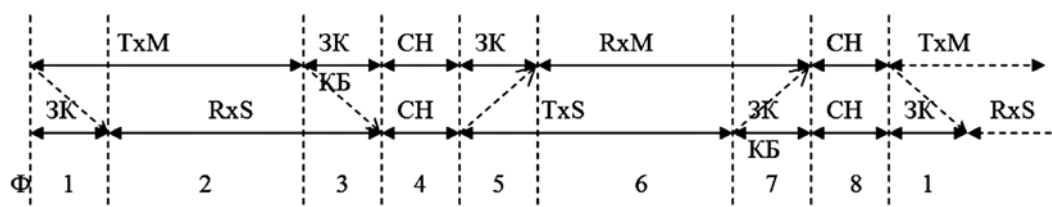
- ✓ Ф — фазы;

- ✓ ЗК — задержка канала, выбрана равной длительности одного битового интервала;
- ✓ TxM — данные, передаваемые ведущим;
- ✓ RxS — данные, принимаемые ведомым;
- ✓ КБ — стоповый бит, равный 1, для заряда линии связи в состояние "пауза" перед переключением направления;
- ✓ СН — битовый интервал смены направления передачи информации;
- ✓ RxM — данные, принимаемые ведущим;
- ✓ TxS — данные, передаваемые ведомым.

Принцип работы междомодульного интерфейса заключается в следующем: ведущий передает ведомому пакеты вместе с контрольным кодом. Далее наступает межпакетная пауза длительностью в один битовый интервал, которая нужна для смены направления передачи информации. В это время линия находится в состоянии единицы, а ведомый анализирует контрольный код. Если контрольный код правильный, то он выдает в линию нулевой уровень — стартовый бит ответного пакета. Обнаружение ведущим стартового бита подтверждает наличие синхронизации машин состояний абонентов и возможность приема ведущим входящего пакета от ведомого. После приема ведущим ответного пакета сразу начинается очередная транзакция.

Свойство асимметричности междомодульного интерфейса объясняется централизованной структурой системы управления, построенной по принципу "вынесенной руки". В "центре" системы — центральном вычислителе — реализованы не только функция расчетной подсистемы ЦСУ, но и функция управления связью с периферией. В случае сбоя междомодульных интерфейсов реакция на сбой реализуется в центральном вычислителе, следовательно, расположенный в нем трансивер междомодульного интерфейса должен обладать свойством ведущего устройства, а трансивер междомодульного интерфейса в периферийном устройстве — роль ведомого. Поэтому интерфейс имеет в названии определение "асимметричный".

Машина состояния ведущего:



Машина состояния ведомого:

Рис. 4. Событийно-временная диаграмма взаимодействия ведущего и ведомого



Рис. 5. Структура межмодульного пакета

Межмодульные пакеты управления и состояния имеют одинаковую структуру, показанную на рис. 5.

Каждый межмодульный пакет состоит из четырех полей: стартового бита, поля данных, поля контрольного кода (CRC) и стопового бита (паузы).

Стартовый бит от ведущего является признаком начала очередной циклической транзакции. Активный уровень стартового бита от ведомого сигнализирует ведущему, что он корректно принял пакет и машины состояний узлов работают согласованно. В случае сбоя в битовом интервале стартового бита пакета от ведомого выставляется пассивный уровень, и ведущий понимает, что необходимо запустить процедуру восстановления синхронизации машин состояний.

Длина *поля данных* определяется числом потоков передаваемых данных и их разрядностью и выбирается фиксированной в ходе проектирования системы.

Интерфейс контролирует *контрольным кодом*. Набор полей, которые должны циркулировать между двумя модулями, фиксирован для конкретного устройства. Поэтому требование о стабильном периоде транзакции не является для нас ограничивающим.

Функцией *стопового бита* является заряд линии данных до единичного уровня. После окончания передачи единичный уровень сохраняется в линии в течение всего времени смены направления обменов, и на его фоне можно четко распознать нулевой уровень стартового бита ответного пакета.

По сравнению с интерфейсами других типов отдельное поле адреса в пакете межмодульного интерфейса отсутствует. Поскольку выбор абонента выполняет "центр" и обмен всегда осуществляется с конкретным абонентом, адресация проводится по "подключению". Процедура доступа к среде передачи данных отсутствует по той же причине.

Рассмотрим подробнее процедуру восстановления синхронизации машин состояний абонентов. Из-за наличия в интерфейсе одной сигнальной линии для передачи информации о нормализации можно использовать только временные

параметры. При восстановлении синхронизации последовательность фаз переключения машины состояний ведущего сохраняется. Ведомый синхронизируется с ведущим в рамках пакетной транзакции такой же длительности, как рабочая. Для передачи поля синхронизации используется интервал поля данных рабочего пакета.

Заключение

Асимметричный межмодульный интерфейс предоставляет возможность реализации концепции "вынесенной руки" с небольшими затратами аппаратных средств. Возможность организации быстрых связей с удаленными периферийными устройствами позволяет упростить их до уровня, не требующего модернизации в жизненном цикле системы управления, и исключает тем самым необходимость изменения алгоритмов управления на уровне периферии.

Предлагаемый интерфейс работает циклично и непрерывно, имеет простую структуру пакетов, несложный механизм управления обменами, максимальную пропускную способность в полудуплексном режиме и возможность согласованной работы с информационной средой цифровых систем управления.

Список литературы

1. Клячко Л. М., Острецов Г. Э. Проектирование высоконадежных систем автоматического управления движением корабля. М.: Физмалит, 2010. 136 с.
2. НОУ ИНТУИТ. Топология шина. Основы локальных сетей. М.: НИЯУ "МИФИ", 2005. 4 с.
3. Холопов Ю. А., Ле Ба Чунг, Нгуен Тхань Чунг, Чан Ван Хань. Особенности реализации распределенных систем управления на основе аппаратных сетевых решений // Информационные технологии. 2016. Т. 22, № 7 (22). С. 494—498.
4. Новицкий А. Синхронный последовательный интерфейс SPI в микроконтроллерах "от А до Я" и его реализация на примере ADuC70xx фирмы Analog Devices. Часть 1 // Компоненты и технологии. 2009. № 3. 7 с.
5. Ле Ба Чунг, Холопов Ю. А. Межмодульный интерфейс // Тезисы докладов 43-й Международной молодежной научной конф. "Гагаринские чтения". Москва, 05—20 апреля 2017 г." М.: МАИ, 2017. 758 с.

Le Ba Chung, Graduate Student, e-mail: chungbaumanvietnam@gmail.com,
Moscow Institute of Physics and Technology (State University),
Ju. A. Holopov, Leader Engineer,
Lebedev Institute of Precision Mechanics and Computer Engineering

Asymmetrical Inter-Module Interface

Inter-module communications in a digital control system based on the principle of "outstretched hand" are considered in this article. A simple asymmetric inter-module interface for transferring information between peripheral devices and a computer is proposed. The interface is characterized by non-complex exchange control mechanism, simple packet structure and high data density in the packet, which allow quick "transparent" communications between remote control system devices and a central computer.

Keywords: digital control system, "outstretched hand" principle, inter-module communication, central computer, peripherals

References

1. **Kljachko L., Ostrecov G.** Proektirovanie vysokonadezhnyh sistem avtomaticheskogo upravlenija dvizheniem korablja, Moscow, Fizmalit, 2010. 136 p. (in Russian).
2. **NOU INTUIT.** Topologija shina. Osnovy lokal'nyh setej, Moscow, "MIFI". 2005, 4 p. (in Russian).
3. **Holopov Ju. A., Le Ba Chung, Nguyen Thanh Trung, Tran Van Khanh.** Osobennosti realizacii raspredelennyh sistem

upravljenija na osnove apparatnyh setevykh reshenij, *Informacionnye tehnologii*, vol. 22, no. 7, pp. 494—498 (in Russian).

4. **Novickij A.** Sinhronnyj posledovatel'nyj interfejs SPI v mikrokontrollerah "ot A do Ja" i ego realizacija na primere ADuC70xx firmy Analog Devices. Chast' 1, *Komponenty i tehnologii*, 2009, no. 3 (in Russian).

5. **Le Ba Chung, Holopov Ju. A.** Mezhdovol'nyj interfejs, *Tezisy dokladov 43-j Mezhdunarodnoj molodezhnoj nauchnoj konferencii "Gagarinskie chtenija"*. Moskva, 05—20 aprelja 2017 g., Moscos, MAI, 2017, 758 p. (in Russian).



2—6 апреля 2018 г. в Донском государственном техническом университете (г. Ростов-на-Дону) состоится международная научная конференция

"Параллельные вычислительные технологии (ПаВТ) 2018",

двенадцатая в серии ежегодных конференций,
посвященных развитию и применению параллельных
вычислительных технологий в различных областях науки и техники.

Главная цель конференции — предоставить возможность для обсуждения перспектив развития параллельных вычислительных технологий и представления результатов, полученных ведущими научными группами в использовании суперкомпьютерных технологий для решения задач науки и техники.

Организаторы конференции:

- Федеральное агентство научных организаций России;
- Суперкомпьютерный консорциум университетов России.

Тематика конференции покрывает все аспекты применения высокопроизводительных вычислений в науке и технике, включая приложения, аппаратное и программное обеспечение, специализированные языки и пакеты.

В первый день работы конференции будет объявлена **28-я редакция списка Top50** самых мощных компьютеров СНГ.

Во все дни работы конференции будет действовать **суперкомпьютерная выставка**, на которой ведущие производители аппаратного и программного обеспечения представят свои новейшие разработки в области высокопроизводительных вычислений.

Языки конференции: русский, английский.

Конференция проводится при поддержке **Российского фонда фундаментальных исследований**.

Официальный сайт конференции: <http://agora.guru.ru/pavt2018/>

Адрес редакции:

107076, Москва, Стромынский пер., 4

Телефон редакции журнала **(499) 269-5510**

E-mail: it@novtex.ru

Технический редактор *Е. В. Конова*.

Корректор *Е. В. Комиссарова*.

Сдано в набор 08.12.2017. Подписано в печать 26.01.2018. Формат 60×88 1/8. Бумага офсетная.

Усл. печ. л. 8,86. Заказ ИТ218. Цена договорная.

Журнал зарегистрирован в Министерстве Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.

Свидетельство о регистрации ПИ № 77-15565 от 02 июня 2003 г.

Оригинал-макет ООО "Авансед солюшнз". Отпечатано в ООО "Авансед солюшнз".
119071, г. Москва, Ленинский пр-т, д. 19, стр. 1.