На правах рукописи

МИЛОСЕРДОВ
Дмитрий Игоревич

МОДЕЛИ, МЕТОДЫ И АРХИТЕКТУРЫ ПРОГРАММНЫХ СИСТЕМ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ ТРУДНОФОРМАЛИЗУЕМЫХ СОБЫТИЙ С НЕПРЕРЫВНЫМ ОБУЧЕНИЕМ

Специальность 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2022

Работа выполнена в Федеральном государственном бюджетном учреждении науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН) в лаборатории технологий больших данных социокиберфизических систем

Научный руководитель:	**Осипов Василий Юрьевич**, доктор технических наук, профессор, директор СПИИРАН СПб ФИЦ РАН
Официальные оппоненты:	**Пророк Валерий Ярославович**, доктор технических наук, профессор, профессор кафедры программно-алгоритмического обеспечения автоматизированных систем управления ракетно-космической обороны ФГБВОУ ВО «Военно-космическая академия имени А.Ф. Можайского»
	Бахшиев Александр Валерьевич, кандидат технических наук, доцент Высшей школы автоматизации и робототехники Института машиностроения, материалов и транспорта ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого»
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)» (СПбГЭТУ «ЛЭТИ»)

Защита диссертации состоится 12 мая 2022г. в 16:00 часов на заседании диссертационного совета 24.1.206.01, созданном на базе Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН) по адресу: 199178, Санкт-Петербург, 14-я линия В.О., 39, каб. 401 e-mail: dc@spcras.ru, тел: (812)-328-33-11.

С диссертацией можно ознакомиться в отделе аспирантуры (каб. 402а) Федерального государственного бюджетного учреждения науки Санкт-Петербургского Федерального исследовательского центра Российской академии наук по адресу: 199178, Санкт-Петербург, 14-я линия В.О., 39 и на сайте http://www.spiiras.nw.ru/dissovet

Автореферат разослан «29» марта 2022 года

Ученый секретарь
диссертационного совета 24.1.206.01
кандидат технических наук  Абрамов Максим Викторович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. Прогнозирование событий – актуальная научно-практическая задача, представляющая интерес во многих областях. Повышение сложности анализируемых процессов и свойственным им событий требуют все более совершенных инструментов прогнозирования. Задачи по управлению транспортом, экономикой, финансами, социальной сферой, сложными техническими объектами невозможно эффективно решать без получения точной и своевременной информации о ближайшем будущем, что определяет **важность** и **значимость** решаемой научной задачи. Независимо от целей и задач прогнозирования, оно выполняется в условиях неопределенности ситуации, когда на интересующий параметр влияют прямые и косвенные факторы, изменяющиеся во времени. Информация об этих факторах зачастую не может быть представлена в виде набора простых аналитических моделей: она закодирована в большом массиве данных, а ее извлечение и обработка требуют новых, нестандартных подходов.

Степень разработанности темы. Прогнозирование событий и нейросетевую обработку информации исследовали А.Н. Аверкин, В.Ю. Осипов, Б.В. Соколов, Я.А. Холодов, Р.М. Юсупов, С.Ф. Яцун, P. Brockwell, E. Egrioglu, I. Goodfellow, L. Haitao, S. Haykin, H. Hu, S. Malik, F. Moretti, A. Sagheer, J. Schmidhuber, Z. Shen, K. Thurow, Y. Wu, B. Yang, T. Zhou и др. Разработано большое число моделей, методов и средств прогнозирования, однако они не всегда удовлетворяют требованиям по точности, оперативности, глубине прогнозирования. Главный недостаток существующих методов заключается в том, что они не обеспечивают должного уровня точности прогнозов событий в условиях высокой неопределенности связанных факторов. В настоящее время отсутствуют архитектуры, объединяющие в себе прогнозирование будущих событий и непрерывное обучение в целях формирования прогнозов в реальном времени с учетом вновь поступившей информации. Поэтому необходимо совершенствование научно-методического аппарата.

Целью диссертационной работы является повышение точности получаемых прогнозов трудноформализуемых событий.

Решаемая научная задача: разработка моделей, методов и архитектур программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением.

Цель работы достигается решением следующих **частных задач**:

1) анализ процесса прогнозирования трудноформализуемых событий;

2) разработка модели системы прогнозирования трудноформализуемых событий с непрерывным обучением;

3) разработка методов прогнозирования трудноформализуемых событий с непрерывным обучением и управлением направленностью вызова сигналов из ассоциативной памяти;

4) разработка архитектур программных систем, реализующих методы прогнозирования трудноформализуемых событий с непрерывным обучением;

5) оценивание полученных результатов, выработка рекомендаций по повышению точности и использованию разработанных моделей, методов и программных систем.

Объектом исследования является процесс прогнозирования трудноформализуемых событий рекуррентными нейронными сетями.

Предметом исследования выступает научно-методический аппарат нейросетевого прогнозирования трудноформализуемых событий.

Научную новизну диссертационной работы составляют:

1. Модель системы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением, отличающаяся своей структурой и правилами обработки сигналов, обеспечивающими оперативное прогнозирование с учетом изменений в законах проявления событий.

Модель содержит блок управления прогнозированием и две идентичные по своей структуре рекуррентные нейронные сети (РНС-1 и РНС-2), объединенные в систему. РНС-1 работает в режиме обучения, блок управления прогнозированием выполняет копирование обученной пространственно-временной модели событий из РНС-1 в РНС-2, а РНС-2 реализует

прогнозирование. Предложенная модель обеспечивает непрерывность процесса обучения при прогнозировании. Это позволяет обеспечить работу в реальном времени и возможность постоянного формирования прогнозов с учетом изменяющихся законов поведения временных рядов. Отсутствует необходимость переобучения сети при поступлении новых данных. Исключается искажение пространственно-временной модели РНС из-за смены режимов ее функционирования.

2. Методы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением: с временными сдвигами сигналов и без временных сдвигов сигналов, отличающиеся новыми правилами прогнозирования и управления ассоциативным вызовом информации из нейросетевой памяти и обеспечивающие высокую точность получаемых прогнозов трудноформализуемых событий.

Согласно методу с временными сдвигами, на вход РНС-1 подаются текущий и задержанный временные ряды. В процессе их прохождения вдоль слоев сети осуществляется ассоциативное пространственно-временное связывание прошлых и будущих событий. Блок управления прогнозированием копирует обученную модель из РНС-1 в РНС-2 и подает текущие входные данные в задержанный канал. В результате в текущем канале за счет вызова сигналов из ассоциативной памяти формируется прогноз будущих событий.

В методе прогнозирования без временных сдвигов на вход РНС-1 подается текущий временной ряд. При прохождении его по сети на ее элементах формируется модель событий, которая постоянно обновляется с учетом вновь поступающих данных. Блок управления прогнозированием копирует состояние РНС-1 в РНС-2 и запускает РНС-2 на формирование прогнозов по новым правилам, предусматривающим управление направленностью ассоциативного вызова сигналов из памяти нейронной сети. Согласно этим правилам, если обрабатываемая выборка признается короткой, то перед прогнозированием предлагается удлинять ее за счет ассоциативного вызова из памяти сети предшествующих значений.

3. Параллельная и буферная архитектуры программных систем, отличающиеся новой структурой и правилами функционирования программных систем прогнозирования с непрерывным обучением, обеспечивающие программную реализацию предложенных моделей и методов и расширение их функций.

В параллельной архитектуре эмулируется оба экземпляра нейронных сетей (РНС-1 и РНС-2). Буферная архитектура предполагает наличие только одного экземпляра (модуля) нейронной сети (РНС-1), называемого модулем эмуляции РНС-1 и РНС-2, а также входного буфера и модуля памяти для хранения состояний нейронов РНС-2. Новизна буферной архитектуры состоит в отказе от выделения памяти для хранения синапсов РНС-2, в выполнении квазипараллельного обучения и прогнозирования, а достигаемый эффект заключается в сокращении объемов требуемой памяти в общем случае в два раза.

4. Практические рекомендации по повышению точности и использованию программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением, обеспечивающие повышение точности прогнозов за счет определения наиболее эффективной конфигурации нейросетевых слоев применительно к задаче прогнозирования трудноформализуемых событий и разработки новых правил выбора метода и архитектуры в зависимости от условий, в которых функционирует система прогнозирования.

Теоретическая и практическая значимость работы. Теоретическая значимость полученных научных результатов состоит в развитии научно-методического аппарата прогнозирования трудноформализуемых событий рекуррентными нейронными сетями с непрерывным обучением. Практическая значимость этих результатов состоит в возможности повысить точность прогнозов возможных событий для различных приложений в условиях слабо формализуемых процессов с учетом большого числа неявно связанных факторов. Помимо повышения точности прогнозов, предложенные решения могут найти применение при проектировании перспективных интеллектуальных систем.

Методология и методы исследования. При выполнении диссертационного исследования использованы методы системного анализа и синтеза, интеллектуальной обработки данных, современная теория нейросетевого прогнозирования, а также методы оптимизации программных систем.

На защиту выносятся следующие положения:

1. Модель системы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением.
2. Методы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением и управлением направленностью вызова сигналов из ассоциативной памяти.
3. Параллельная и буферная архитектуры программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением.
4. Практические рекомендации по повышению точности и использованию программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением.

Соответствие диссертации паспорту научной специальности. Представленные результаты соответствуют паспорту специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Высокая степень достоверности научных положений обеспечена анализом текущего уровня исследований в данной области, корректным использованием апробированного математического аппарата, согласованностью теоретических выводов с результатами вычислительных экспериментов, сравнением предложенных решений с известными аналогами и одобрением основных положений диссертационной работы на международных и всероссийских научных конференциях.

Апробация и реализация результатов. Основные положения диссертационной работы представлялись на 6 международных конференциях «Digital Transformation And Global Society (DTGS-2019)» (Санкт-Петербург, 19-21 июня 2019 г.), «Experimental Economics and Machine Learning (EEML-2019)» (Пермь, 25-26 сентября 2019 г.), 5-я Международная научно-практическая конференция «Технологическая перспектива-2019» (Санкт-Петербург, 7-8 ноября 2019 г.), «Digital Transformation And Global Society-2020» (Санкт-Петербург, 17-19 июня 2020 г.), 6-я Международная научно-практическая конференция «Технологическая перспектива-2020» (Санкт-Петербург, 12-13 ноября 2020 г.), 7-я Международная научно-практическая конференция «Технологическая перспектива-2021» (Санкт-Петербург, 11-12 ноября 2021 г.) и всероссийской конференции «Информационные технологии в управлении» (ИТУ-2020) (Санкт-Петербург, 7-8 октября 2020 г.).

Результаты диссертационной работы использованы в НИР СПб ФИЦ РАН №0073-2019-0001 «Теоретические основы и алгоритмические модели когнитивного управления, взаимодействия и анализа состояния групп гетерогенных робототехнических комплексов», а также в ЦСАМ АО «НТЦ РЭБ» при проведении научных исследований по обнаружению и траекторному сопровождению малоразмерных беспилотных летательных аппаратов для прогнозирования радиолокационной обстановки и выявления аномальных радиосигналов.

Публикации. По научным результатам диссертационного исследования опубликовано 12 работ, в том числе 2 публикации в журналах из «Перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук», одна из статей в указанном перечне опубликована без соавторов, 5 публикаций в зарубежных изданиях, индексируемых в Scopus/WoS (в том числе 2 публикации в журналах Q1), два свидетельства о государственной регистрации программы для ЭВМ, одно из которых зарегистрировано без соавторов.

Личный вклад соискателя. Автором лично разработаны архитектуры и правила функционирования программных систем нейросетевого прогнозирования с непрерывным

обучением. Лично автором разработаны практические рекомендации по повышению точности прогнозов и использованию программных систем нейросетевого прогнозирования с непрерывным обучением. Модель и методы нейросетевого прогнозирования с непрерывным обучением разработаны в соавторстве с научным руководителем, причем вклад соискателя в совместных публикациях был значительным.

Структура и объем работы. Текст работы состоит из следующих структурных элементов: введение; основная часть, включающая четыре главы; заключение; список сокращений и условных обозначений; список литературы, содержащий 157 наименований; список иллюстративного материала; три приложения, содержащие список публикаций соискателя по теме диссертации, копии полученных свидетельств об интеллектуальной собственности, а также копии актов внедрения результатов диссертационной работы. Общий объем диссертационной работы – 145 страниц. Работа включает в себя 38 рисунков, 14 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цели исследования, решаемая научная задача и положения, выносимые на защиту, отражена суть и новизна основных научных результатов.

В первой главе диссертации проведен анализ целей и условий прогнозирования событий. Отмечена необходимость учета изменений законов проявления событий, взаимосвязей различных факторов, прямо или косвенно влияющих на эти законы, а также наличия в анализируемых временных рядах дефектов.

Проведен анализ известных методов прогнозирования временных рядов (рис. 1). Показано, что известные методы не в полной мере удовлетворяют имеющимся требованиям. Поставлена научная задача разработки новых моделей, методов и программных архитектур нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением.

Во второй главе в качестве основы разрабатываемой системы прогнозирования предлагается использовать известные РНС с управляемыми элементами. Такое прогнозирование предполагает обучение РНС и последующее предсказание с применением сформированной пространственно-временной нейросетевой модели.

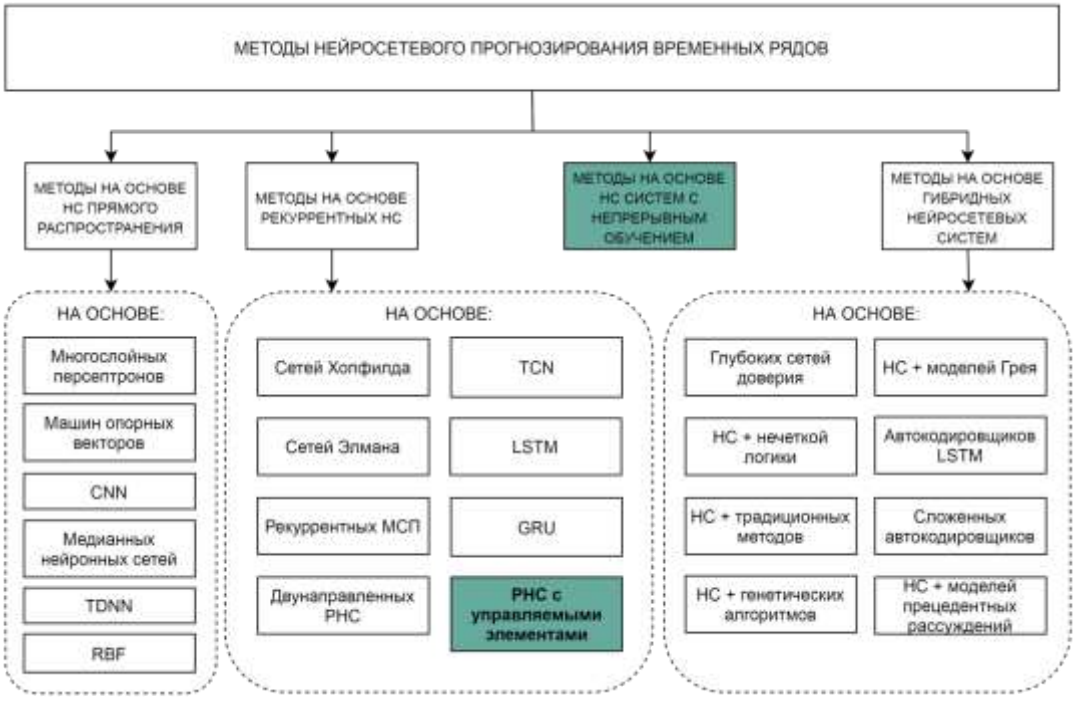


Рисунок 1 — Классификация методов нейросетевого прогнозирования временных рядов

Для решения задач прогнозирования трудноформализуемых событий с учетом возможных условий предлагается обобщенная модель системы нейросетевого прогнозирования событий с непрерывным обучением (рис. 2).

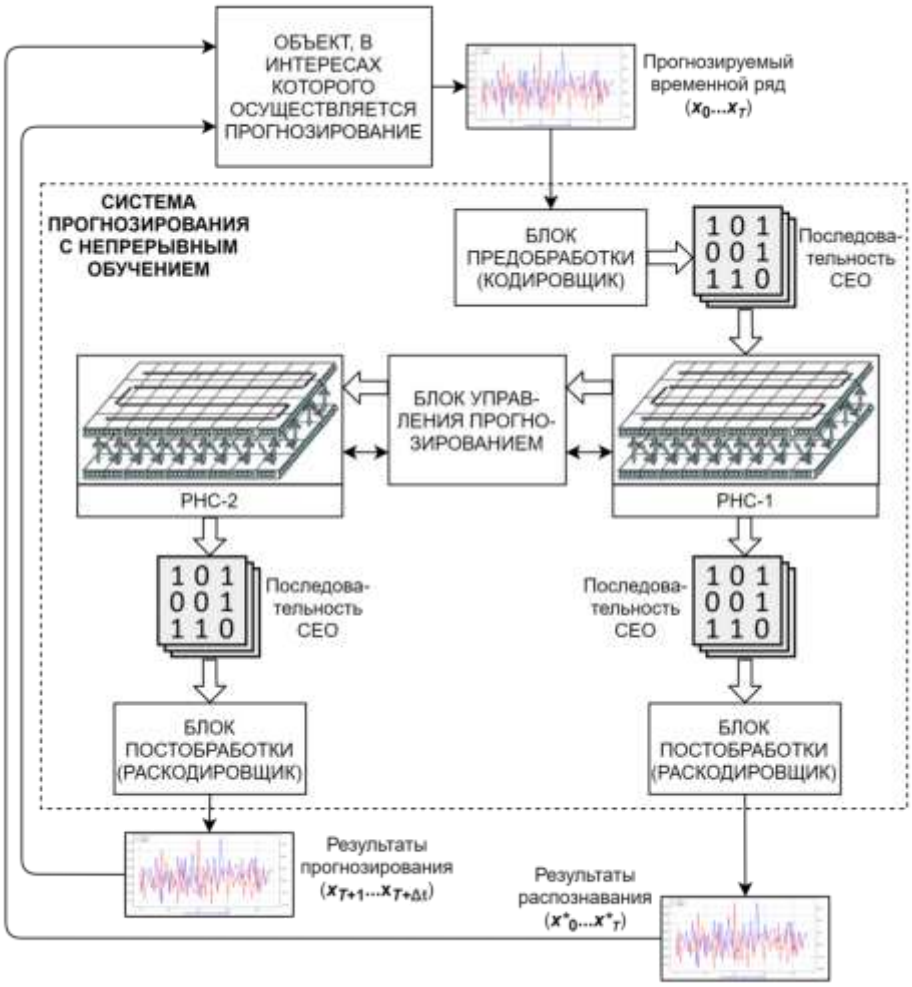


Рисунок 2 — Обобщенная модель системы нейросетевого прогнозирования событий с непрерывным обучением

Согласно рис. 2, на вход системы подается временной ряд (обозначен $x_0 \ldots x_T$), который в блоке предобработки преобразуются в последовательность совокупностей единичных образов (СЕО). Последовательность СЕО обрабатывается в РНС-1. В процессе этой обработки на синапсах РНС-1 формируется пространственно-временная модель наблюдаемых событий. При этом за счет обученной модели возможно распознавание и частичное восстановление обрабатываемой в РНС-1 выборки (фильтрация, очистка от шумов и т.п.). Данные, снимаемые с РНС-1, раскодируются в блоке постобработки, в результате чего на выходе системы прогнозирования имеется распознанный временной ряд ($x^*_0 \ldots x^*_T$). Когда необходимо выполнить прогноз, блок управления прогнозированием копирует состояние РНС-1 в РНС-2 и запускает прогнозирование на РНС-2, а РНС-1 продолжает обучаться. В процессе обработки в РНС-2 формируются спрогнозированные СЕО, которые после раскодирования во втором блоке постобработки передаются на второй выход системы и представляют собой результат прогнозирования ($x_{T+1} \ldots x_{T+\Delta t}$).

Предложенная обобщенная модель конкретизируется в зависимости от задач и условий прогнозирования. Приводятся 4 типовых оптимизационных задачи поиска методов нейросетевого прогнозирования трудноформализуемых событий.

1. Когда требуется найти целесообразный способ S_0 прогнозирования трудноформализуемых событий на области поиска допустимых способов Q, позволяющий минимизировать его ошибку E_0 на заданном горизонте $T_{\text{упр. зад.}}$ при ограничениях на время формирования прогноза $T_{\text{рсч. зад.}}$, необходимо решать задачу:

$$E_0(S_0) = \min_{i \in Q} E_i(S_i),\ T_{\text{упр. }i}(S_i) = T_{\text{упр. зад.}},\ T_{\text{рсч. }i}(S_i) \le T_{\text{рсч. зад.}}$$

2. В случае, когда нужно найти целесообразный способ S_0 прогнозирования трудноформализуемых событий при минимуме ошибки прогнозирования на заданный горизонт и ограничениях на объем используемой памяти $H_{\text{рсч. зад.}}$, решаемая задача имеет вид:

$$E_0(S_0) = \min_{i \in Q} E_i(S_i),\ T_{\text{упр. }i}(S_i) = T_{\text{упр. зад.}},\ H_{\text{рсч. }i}(S_i) \le H_{\text{рсч. зад.}}$$

3. Если необходимо найти целесообразный способ S_0 прогнозирования трудноформализуемых событий для заданного горизонта за минимальное время $T_{\text{рсч. }0}$ при ограничении на величину ошибки $E_{\text{зад.}}$, то условия задачи представляются в виде:

$$T_{\text{рсч. }0}(S_0) = \min_{i \in Q} T_{\text{рсч. }i}(S_i),\ T_{\text{упр. }i}(S_i) = T_{\text{упр. зад.}},\ E_i(S_i) \le E_{\text{зад.}}$$

4. Когда требуется найти целесообразный способ S_0 прогнозирования трудноформализуемых событий при минимизации объема требуемой памяти $H_{\text{рсч.0}}$ системы и ограничениях на ошибку прогнозирования с заданным горизонтом, требуется решать задачу:

$$H_{\text{рсч.0}}(S_0) = \min_{i \in Q} H_{\text{рсч.}i}(S_i),\ T_{\text{упр. }i}(S_i) = T_{\text{упр. зад.}},\ E_i(S_i) \le E_{\text{зад.}}$$

Для решения сформулированных задач предложено два метода на основе модели прогнозирования с непрерывным обучением: прогнозирование с временными сдвигами сигналов (рис. 3) и без временных сдвигов с управлением направленностью вызова сигналов из ассоциативной памяти (рис. 4).

Согласно методу прогнозирования с временными сдвигами, временной ряд $\boldsymbol{X}_t$ сдвигается на интервал τ и формируется ряд $\boldsymbol{X}_{t-\tau}$. В РНС на вход подается пара $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-\tau}$. В процессе прохождения пары временных рядов РНС формирует пространственно-временную ассоциативную модель путем изменения весов синапсов.

Для прогнозирования информация о весах синапсов в первой РНС копируются во вторую РНС ($\boldsymbol{W}_t \to \boldsymbol{W^*}_t$). Затем на второй вход этой РНС вместо задержанных сигналов подаются текущие сигналы ($\boldsymbol{X^*}_t$). В результате на слоях нейросети происходит ассоциативный вызов информации о будущих событиях ($\boldsymbol{X}_{t+\tau}$).

С формальной точки зрения процесс обучения и прогнозирования по методу с временными сдвигами представляет собой выражения:

$$\boldsymbol{W}_{t+1} = \varphi(\boldsymbol{X}_t, \boldsymbol{X}_{t-\tau}, \boldsymbol{W}_t),\ \boldsymbol{X}_t = \psi(\boldsymbol{X}_{t-\tau}, \boldsymbol{W}_t),$$

$$\Downarrow$$

Копирование весов синапсов РНС-1 в РНС-2

$$\Downarrow$$

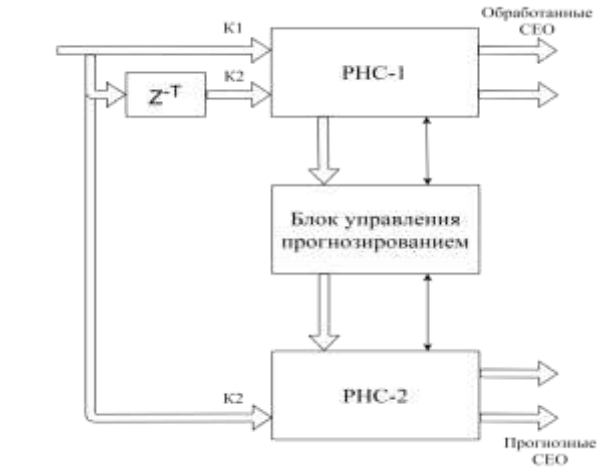
$$\boldsymbol{X}_{t+\tau} = \psi^*(\boldsymbol{X^*}_t, \boldsymbol{W^*}_t),$$

где φ – функция изменения весов синапсов; ψ – функция переходов процесса из одних состояний в другие, $\boldsymbol{W}_t$ – пространственно-временная модель событий.

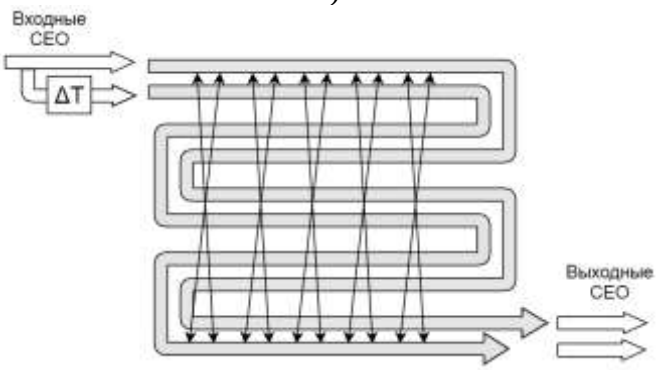
Структурная схема системы прогнозирования, реализующей указанный метод, представлена на рис. 3а, демонстрация метода на уровне нейросетевых каналов – на рис.3б. В данном случае демонстрируется РНС с конфигурацией слоев в виде спирали постоянного радиуса. Нейросетевые каналы на рис. 3б обозначены широкими стрелками, а узкими стрелками показаны ассоциативные связи, формируемые в процессе обучения и используемые при прогнозировании. На рис. 3в представлена блок-схема, описывающая выполнение метода прогнозирования с непрерывным обучением системы с временными сдвигами. Жирным контуром выделены блоки, содержащие научную новизну.

В качестве альтернативы предлагается метод прогнозирования без временных сдвигов последовательностей СЕО (рис. 4). Как и в предыдущем случае, используется пара РНС с управляемыми элементами. Однако в данном случае используется только один нейросетевой

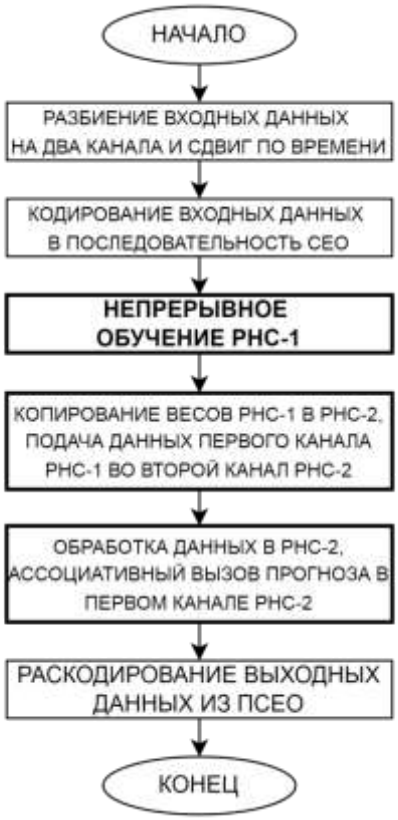
канал. Пространственно-временное связывание осуществляется внутри этого канала, а прогнозирование основывается на ассоциативном вызове будущих значений за счет предыдущих СЕО.



а)



б)



в)

Рисунок 3 — Схемы, поясняющие метод нейросетевого прогнозирования с временными сдвигами: а – структурная схема системы прогнозирования; б – канальная структура РНС во время обучения, в – блок-схема, описывающая выполнение метода

С формальной точки зрения описание процесса обучения и прогнозирования нейросетевой системой выглядит следующим образом:

$$\boldsymbol{W}_{t+1} = \varphi(\boldsymbol{X}_t, \boldsymbol{W}_t), \boldsymbol{X}_{t+\tau} = \psi(\boldsymbol{X}_t, \boldsymbol{W}_t),$$

$$\Downarrow$$

Копирование весов синапсов РНС-1 в РНС-2

$$\Downarrow$$

$$\boldsymbol{X}_{t+\tau} = \psi^*(\boldsymbol{X^*}_t, \boldsymbol{W^*}_t).$$

При обучении на вход подается ряд $\boldsymbol{X}_t$. В процессе прохождения его по слоям РНС-1 формируется пространственно-временная модель событий $\boldsymbol{W}_t$ в виде матрицы весов синапсов и текущих состояний нейронов. Когда необходимо произвести прогноз на горизонт $\tau = \overline{1, T_{\text{упр.}}}$, выполняется копирование пространственно-временной модели $\boldsymbol{W}_t \to \boldsymbol{W^*}_t$ и копирование состояний слоев $\boldsymbol{X}_t \to \boldsymbol{X^*}_t$ из РНС-1 в РНС-2, после чего на РНС-2 запускается ускоренная обработка данных с заданным коэффициентом ускорения. В результате на выходе РНС-2 получаем спрогнозированный временной ряд $\boldsymbol{X}_{t+\tau}$.

Суть метода прогнозирования в условиях малых выборок, наличия шума и пропусков, являющегося составной частью метода прогнозирования без временных сдвигов и определяющего правила управления направленностью вызова сигналов из ассоциативной памяти нейронной сети, продемонстрирована на рисунке 5. Она заключается в оценке параметров обрабатываемой выборки $T_1 \ldots T_N$ и, если та признается короткой, осуществляется удлинение ее за счет усиления ассоциативного вызова сигналов в направлении выхода РНС и получение таким

путем предыдущих значений $T_{-M} \dots T_0$. Получение же прогноза $T_{N+1} \dots T_{N+K}$ предполагается за счет использования уже дополненного ряда $T_{-M} \dots T_N$ при усилении ассоциативного вызова сигналов в направлении входа РНС.

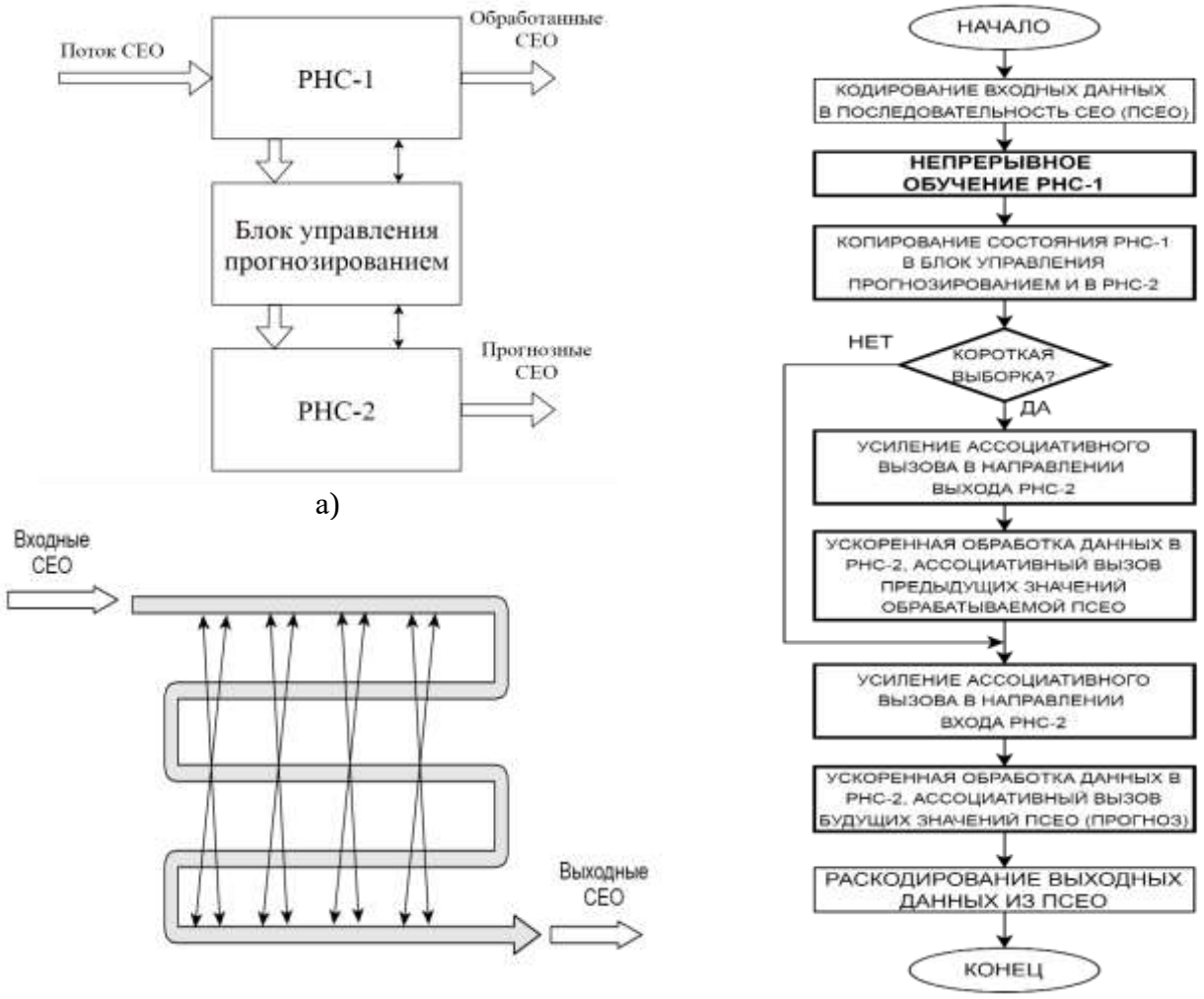


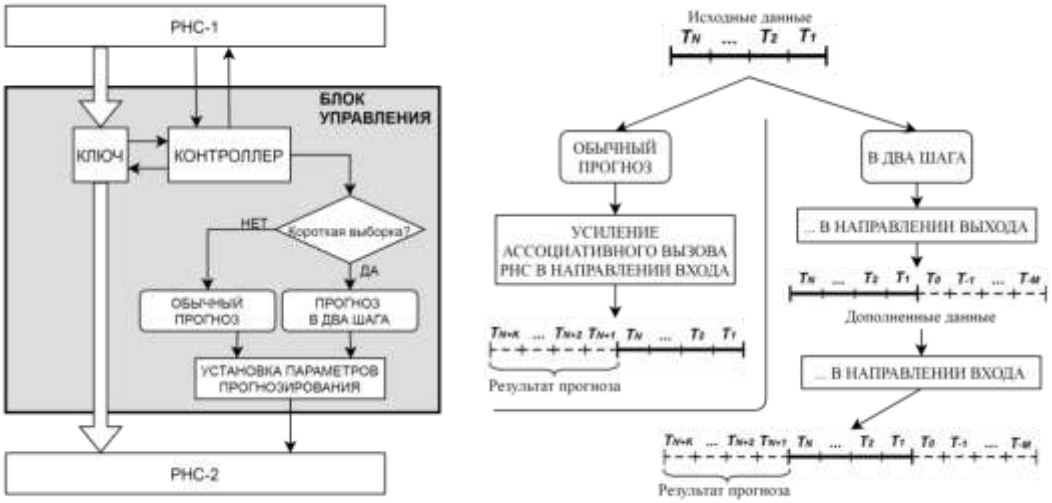
Рисунок 4 — Схемы, поясняющие метод нейросетевого прогнозирования без временных сдвигов: а – структурная схема системы прогнозирования; б – канальная структура РНС во время обучения, в – блок-схема, описывающая выполнение метода

В третьей главе предложены программные архитектуры, реализующие рассмотренные во второй главе модели и методы нейросетевого прогнозирования с непрерывным обучением. Разработана архитектура программной системы, называемая далее «параллельной» (рис. 6).

Предлагаемая архитектура является развитием методов в части адаптации их к программной реализации, а также расширения их функций. Так, в ней определено трехпоточное окружение, обеспечивающее параллельную работу блока управления прогнозированием и экземплярами РНС-1 и РНС-2. Конкретизирована структура блока управления прогнозированием. Определены модули чтения-записи параметров РНС, элементы визуализации процесса прогнозирования и управления программной системой, включающие кнопочные панели и графические модули для каждого из экземпляров РНС. Уточнены программные структуры данных, передаваемые между элементами архитектуры. Для нее разработана и приведена в тексте диссертации UML диаграмма классов.

Для рационального использования памяти предложена «буферная» архитектура (рис. 7). Ее UML диаграмма классов представлена в тексте диссертации. В отличие от параллельной архитектуры, в которой используется пара РНС, в буферной архитектуре вводится модуль эмуляции РНС-1 и РНС-2. Структурно этот модуль представляет собой экземпляр РНС,

наделенный возможностями по смене режимов функционирования (обучение/работа). Дополнительно вводятся модуль памяти для хранения состояний нейронов РНС, а также входной буфер и новые связи между модулями.



а) б)

Рисунок 5 — Пояснения к методу нейросетевого прогнозирования в условиях коротких выборок, шумов и пропусков: а – выбор режима прогнозирования блоком управления прогнозированием; б – особенности обычного прогнозирования и прогнозирования в два шага.

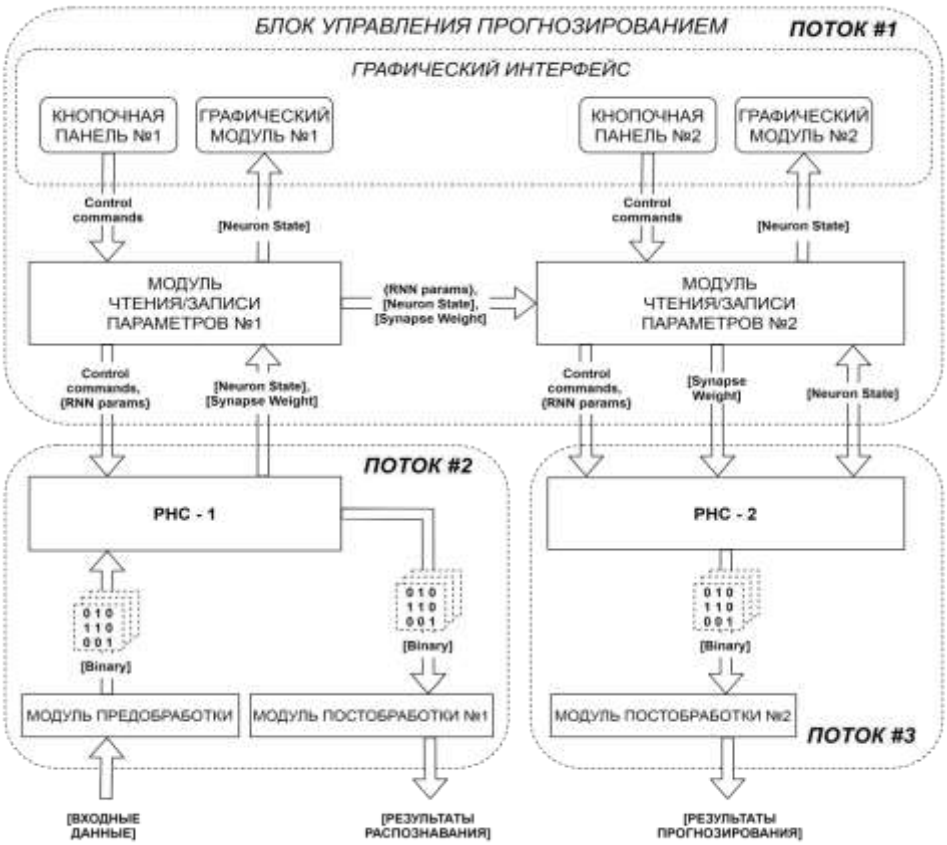


Рисунок 6 — Архитектура программной системы нейросетевого прогнозирования с непрерывным обучением. Здесь Control commands – управляющие команды, {RnnParams} – программная структура с информацией о параметрах функционирования РНС, [Binary], [Neuron

State], [Synapse Weight] – массивы с бинарными данными СЕО, состояниями нейронов и весами синапсов соответственно.

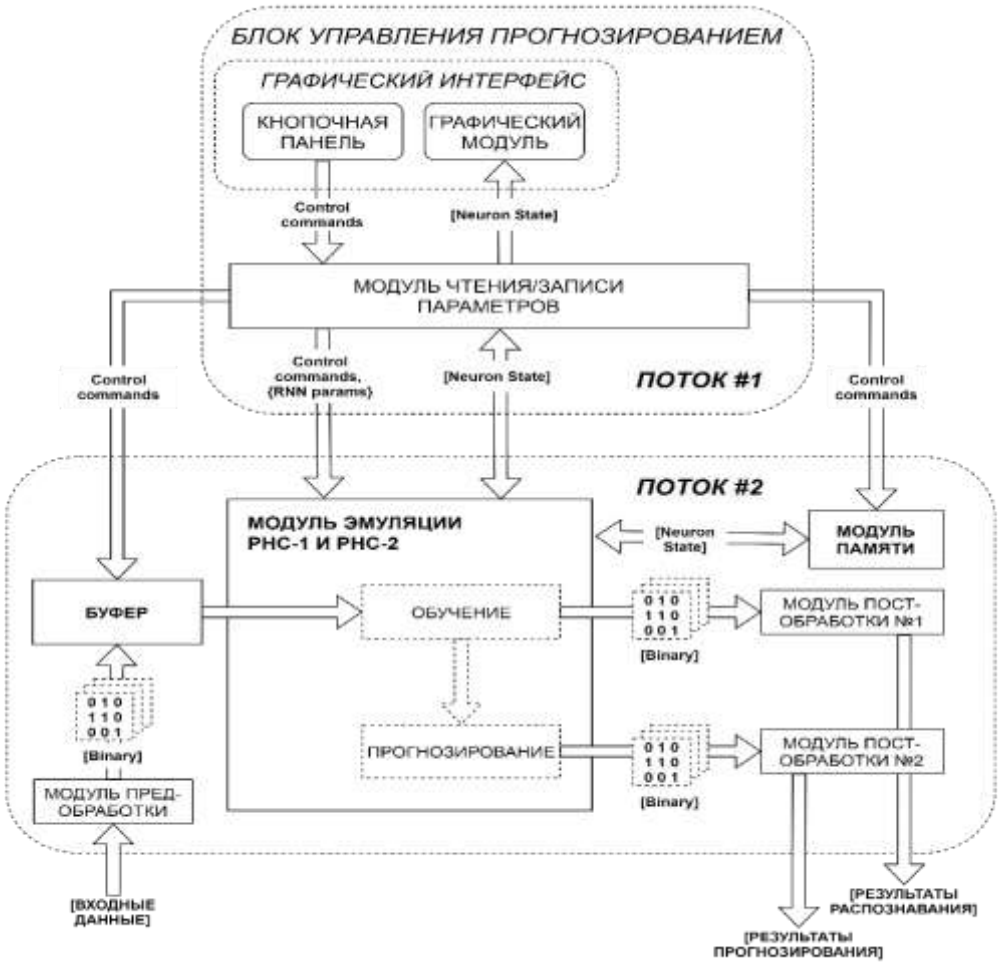


Рисунок 7 — Буферная архитектура программной системы нейросетевого прогнозирования. Обозначения структур данных идентичны рис. 6.

При функционировании на этапе обучения входные данные кодируются в формат РНС и проходят без задержки через буфер и через модуль эмуляции, который в это время обучается. В результате на синапсах РНС модуля эмуляции выстраивается пространственно-временная модель наблюдаемых событий. В отличие от параллельной архитектуры, в предлагаемой системе при получении команды на прогноз происходит заморозка синаптических весов модуля эмуляции, сохранение состояний его нейронов в модуле памяти, после чего модуль эмуляции переходит в режим прогнозирования и выполняет его. Если во время прогнозирования на вход системе поступают данные, то они задерживаются в буфере. Когда прогнозирование окончено, веса РНС модуля эмуляции размораживаются, а состояния нейронов восстанавливаются из модуля памяти. Таким образом, модуль эмуляции приводит свое состояние к такому, в каком он находился на момент получения команды на прогнозирование. После этого продолжается обучение. Если в буфере накопились данные, то внутреннее время модуля эмуляции ускоряется, и он и обрабатывает эти данные в ускоренном режиме, чтобы синхронизировать свое состояние с внешней средой.

Главным достоинством буферной архитектуры является сокращение объемов памяти, требуемых для функционирования системы. Эффект достигается тем, что в предлагаемом подходе не требуется хранить синапсы второго экземпляра РНС. Поскольку количество синапсов растет в квадратичной зависимости от количества нейронов РНС, то предлагаемая архитектура позволяет снизить потребление памяти в общем случае в два раза.

Блок-схемы, раскрывающие правила функционирования параллельной и буферной архитектур, приведены на рисунках 8 и 9 соответственно.

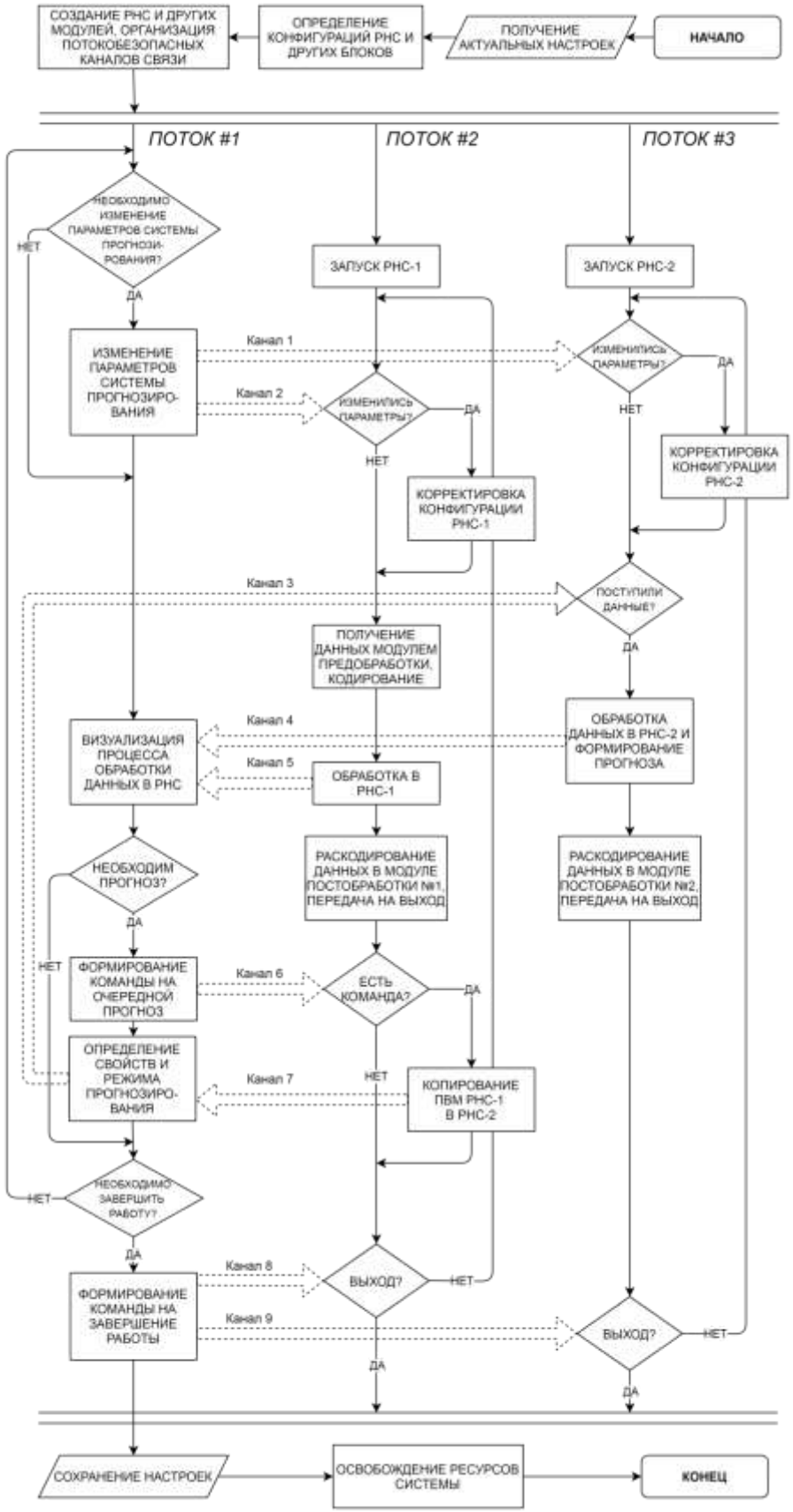


Рисунок 8 — Блок-схема, поясняющая правила функционирования параллельной архитектуры программной системы нейросетевого прогнозирования с непрерывным обучением. Каналы 1-9 отвечают за передачу данных между параллельными потоками.

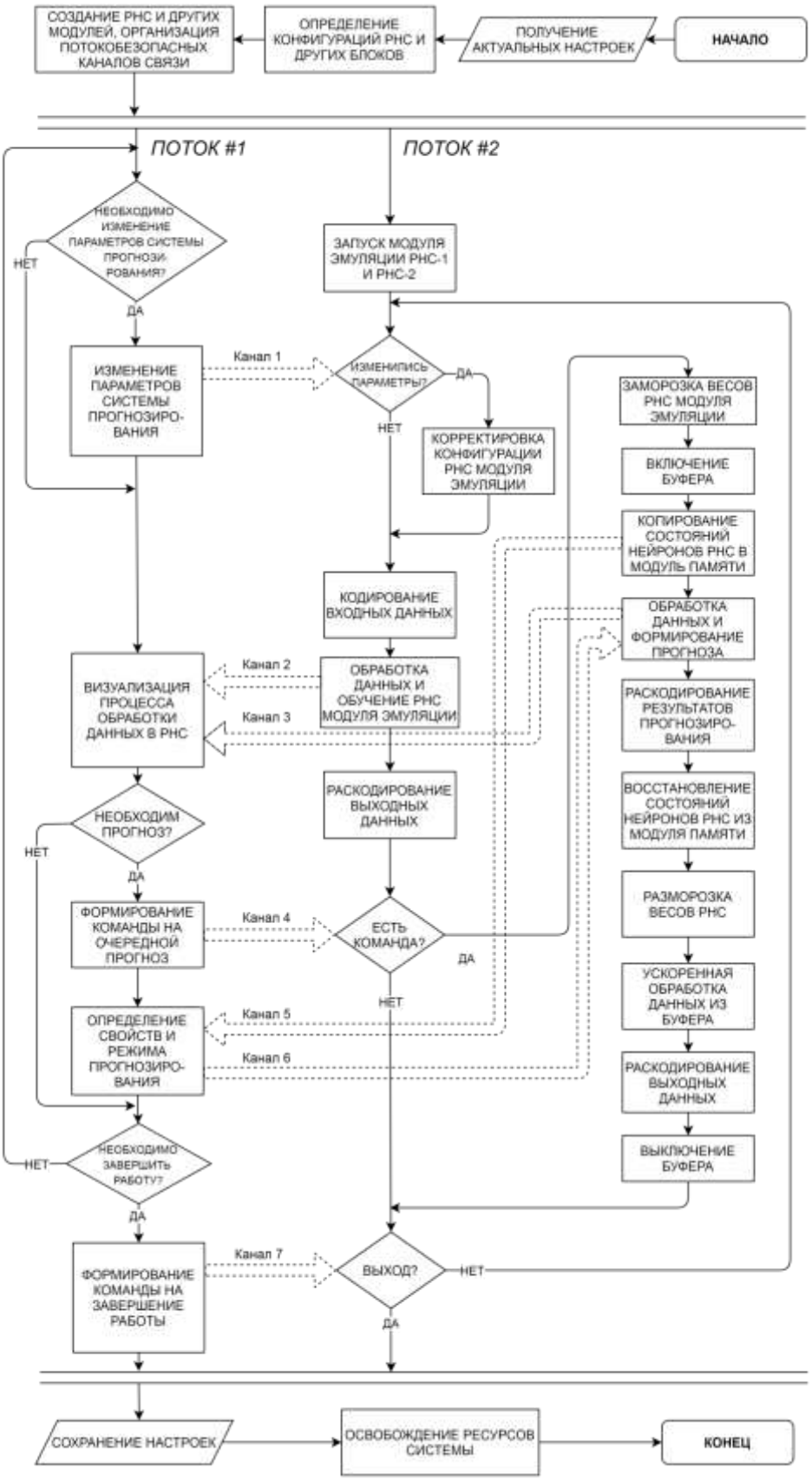


Рисунок 9 — Блок-схема, поясняющая правила функционирования буферной архитектуры программной системы нейросетевого прогнозирования с непрерывным обучением. Каналы 1-7 отвечают за передачу данных между параллельными потоками.

В четвертой главе приводятся результаты моделирования, демонстрирующие работоспособность предложенных методов и программных архитектур, на примере задачи прогнозирования транспортных потоков и задачи прогнозирования лексического содержания новостных лент.

Для проведения экспериментов использовался компьютер под управлением ОС Windows 7 x64, с процессором Intel Core i7-4790 CPU 3.6GHz, 32 ГБ ОЗУ. ПО разработано на языке программирования C++ (Qt 5.14.1).

В качестве исходных данных для прогнозирования транспортных потоков использовалось три набора данных. Первый набор содержит данные сервиса «Яндекс.Пробки» о скоростях движения по дорогам в г. Санкт-Петербург в период 14-27 февраля 2019 года. Второй и третий наборы – данные, собираемые Лабораторией исследования транспортных данных (TDRL), содержит информацию о занятости дорог и объеме трафика соответственно, в период 12-24 января 2018 года.

Для оценки точности использовались метрики средней абсолютной ошибки (MAE), средней абсолютной процентной ошибки (MAPE) и среднеквадратической ошибки (RMSE):

$$MAE = \frac{1}{N}\sum_{p=1}^{N}\left|y_p - y_p^*\right|,\ MAPE = \frac{1}{N}\sum_{p=1}^{N}\frac{\left|y_p - y_p^*\right|}{y_p}\cdot 100\%,\ RMSE = \sqrt{\frac{1}{N}\sum_{p=1}^{N}(y_p - y_p^*)^2},$$

где y_p – реальные значения временных рядов; y_p^* – спрогнозированные значения.

Полученные показатели точности прогнозирования транспортных потоков для предлагаемых методов и известных аналогов представлены в таблице 1.

Таблица 1 — Сравнение показателей эффективности прогнозирования транспортных потоков

Метод	**MAE (км/ч)**	**MAPE (%)**	**RMSE (км/ч)**
Интернет-сервис «Яндекс.Пробки»	3.64	28.60	4.96
Модель ARIMA	4.18	30.00	4.90
Нейронная сеть LSTM	4.22	26.10	5.15
Предлагаемый метод с временными сдвигами	2.89	23.20	4.35
Предлагаемый метод без временных сдвигов	3.30	23.60	4.36

Анализ таблицы 1 показывает, что предложенные методы в метриках MAE, MAPE, RMSE превосходят известные решения. Несколько точнее показал себя метод прогнозирования с временными сдвигами, однако метод без временных сдвигов позволяет получать прогнозы более оперативно (31 мс против 74 мс при использовании метода с временными сдвигами).

Отдельная серия экспериментов была проведена с целью определения оптимальной конфигурации нейросетевых слоев для выработки рекомендаций по повышению точности прогнозирования событий. Используемые РНС могут иметь слои различной конфигурации: линейные, петлевые, спиральные. При прочих равных параметрах РНС была оценена возможность прогнозирования с различными конфигурациями слоев (табл. 2).

Таблица 2 — Результаты оценки точности прогнозов, полученных с применением РНС с различными конфигурациями слоев

Вид структуры РНС	**MAE**	**MAPE (%)**	**RMSE**
Линейная	0,79	20,30	1,18
Спираль с постоянным диаметром	0,67	17,00	1,06
Петлевая	0,74	18,90	1,15
Спираль сходящаяся	0,71	17,10	1,08
Спираль расходящаяся	0,66	17,00	1,06

ARIMA	0,91	24,40	1,35

Результаты показывают, что все пять конфигураций продемонстрировали достаточно высокую точность прогнозирования объема дорожного трафика. Она выше точности ARIMA. Наилучшие результаты характерны спиральным структурам, которые оказались предпочтительнее, чем структуры линейные.

В целях прогнозирования лексического содержания новостных информационных лент в период 16-22 марта 2019 г. через каждые 15 мин. собиралась информация с сайта Lenta.ru. Результаты моделирования и оценки точности прогнозирования лексического содержания новостных лент приведены в табл. 3, где показатели MAE, MAPE и RMSE определяются по формулам:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\sum_{p=1}^{M}\left|(\delta_{tp} - \delta_{tp}^{*})\cdot TF\_IDF_{tp}\right|,\ MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{\sum_{p=1}^{M}\left|(\delta_{tp}-\delta_{tp}^{*})\cdot TF\_IDF_{tp}\right|}{\sum_{p=1}^{N}\delta_{tp}\cdot TF\_IDF_{tp}}\cdot 100\%,$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\sum_{p=1}^{M}\left|(\delta_{tp} - \delta_{tp}^{*})\cdot TF\_IDF_{tp}\right|\right)^{2}},$$

где δ_{tp} – булевы значения присутствия ($\delta_{tp} = 1$) либо отсутствия ($\delta_{tp} = 0$) p-го слова из словаря размером M на момент времени t; δ_{tp}^{*} - прогноз присутствия этого слова; $TF\_IDF_{tp} = f_{tp}\cdot\log\left(\frac{W}{w_p}\right)$ – коэффициент важности p-го слова из словаря, вычисляемое как частота встречаемости f_{tp} этого слова в тексте на момент времени t, умноженное на логарифм отношения количества исследуемых наборов слов W к количеству наборов w_p, в которых есть p-е слово.

Анализ табл. 3 показывает, что как на 1-часовом, так и на 4-часовом горизонтах предлагаемые подходы превосходят известную архитектуру LSTM.

Таблица 3 — Результаты прогнозирования лексического содержания новостных лент.

Размер словаря	306 слов		1000 слов		1000 слов	
Период прогнозирования	1 час		1 час		4 часа	
Интервал прогнозирования	15 мин		15 мин		60 мин	
Метод прогнозирования	**Предл. м-д**	**LSTM**	**Предл. м-д**	**LSTM**	**Предл. м-д**	**LSTM**
MAPE	0.25	0.39	0.23	0.36	0.46	0.57
MAE	21.5	29.1	35.6	39.6	41.6	44.5
RMSE	0.26	0.40	0.25	0.38	0.48	0.59

На основе результатов моделирования выработаны рекомендации по применению систем нейросетевого прогнозирования с непрерывным обучением:

1. Рекомендация по выбору метода прогнозирования: когда точность прогнозирования имеет решающую роль, необходимо использовать метод прогнозирования с временными сдвигами. Если важнее экономия ресурсов, то предпочтительно использовать метод без временных сдвигов.

2. Рекомендация по выбору архитектуры: при существенных ограничениях на объем памяти, выделяемый системе прогнозирования, необходимо использовать буферную архитектуру, в противных случаях используется архитектура параллельная.

3. Рекомендация по выбору конфигураций слоев РНС: при прогнозировании трудноформализуемых событий предпочтительно использовать конфигурацию слоев РНС в виде спирали.

В заключении приведены выводы и результаты, полученные в ходе выполнения работы, даны рекомендации по применению разработанных моделей, методов и архитектур, определены перспективы дальнейшей разработки темы.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена задача разработки моделей, методов и архитектур программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением, поставленная цель повышения точности получаемых прогнозов трудноформализуемых событий достигнута. Решенная задача имеет важное значение для совершенствования моделей, методов и средств прогнозирования событий в условиях слабо формализуемых процессов с учетом большого числа неявно связанных факторов.

Основные научные результаты, составляющие **итоги** исследования:

1. Предложена модель системы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением, отличающаяся своей структурой и правилами обработки сигналов, обеспечивающими оперативное прогнозирование с учетом изменений в законах проявления событий.

Модель содержит блок управления прогнозированием и две идентичные по своей структуре рекуррентные нейронные сети (РНС-1 и РНС-2), объединенные в систему. РНС-1 работает в режиме обучения, блок управления прогнозированием выполняет копирование обученной пространственно-временной модели событий из РНС-1 в РНС-2, а РНС-2 реализует прогнозирование. Предложенная модель обеспечивает непрерывность процесса обучения при прогнозировании. Это позволяет обеспечить работу в реальном времени и возможность постоянного формирования прогнозов с учетом изменяющихся законов поведения временных рядов. Отсутствует необходимость переобучения сети при поступлении новых данных. Исключается искажение пространственно-временной модели РНС из-за смены режимов ее функционирования.

2. Предложены методы нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением: с временными сдвигами сигналов и без временных сдвигов сигналов, отличающиеся новыми правилами прогнозирования и управления ассоциативным вызовом информации из нейросетевой памяти и обеспечивающие высокую точность получаемых прогнозов трудноформализуемых событий.

Согласно методу с временными сдвигами, на вход РНС-1 подаются текущий и задержанный временные ряды. В процессе их прохождения вдоль слоев сети осуществляется ассоциативное пространственно-временное связывание прошлых и будущих событий. Блок управления прогнозированием копирует обученную модель из РНС-1 в РНС-2 и подает текущие входные данные в задержанный канал. В результате в текущем канале за счет вызова сигналов из ассоциативной памяти формируется прогноз будущих событий.

В методе прогнозирования без временных сдвигов на вход РНС-1 подается текущий временной ряд. При прохождении его по сети на ее элементах формируется модель событий, которая постоянно обновляется с учетом вновь поступающих данных. Блок управления прогнозированием копирует состояние РНС-1 в РНС-2 и запускает РНС-2 на формирование прогнозов по новым правилам, предусматривающим управление направленностью ассоциативного вызова сигналов из памяти нейронной сети. Согласно этим правилам, если обрабатываемая выборка признается короткой, то перед прогнозированием предлагается удлинять ее за счет ассоциативного вызова из памяти сети предшествующих значений.

3. Разработаны параллельная и буферная архитектуры программных систем, отличающиеся новой структурой и правилами функционирования программных систем прогнозирования с непрерывным обучением, обеспечивающие программную реализацию предложенных моделей и методов и расширение их функций.

В параллельной архитектуре эмулируется оба экземпляра нейронных сетей (РНС-1 и РНС-2). Буферная архитектура предполагает наличие только одного экземпляра (модуля) нейронной

сети (РНС-1), называемого модулем эмуляции РНС-1 и РНС-2, а также входного буфера и модуля памяти для хранения состояний нейронов РНС-2. Новизна буферной архитектуры состоит в отказе от выделения памяти для хранения синапсов РНС-2, в выполнении квазипараллельного обучения и прогнозирования, а достигаемый эффект заключается в сокращении объемов требуемой памяти в общем случае в два раза.

4. Разработаны практические рекомендации по повышению точности и использованию программных систем нейросетевого прогнозирования трудноформализуемых событий с непрерывным обучением, обеспечивающие повышение точности прогнозов за счет определения наиболее эффективной конфигурации нейросетевых слоев применительно к задаче прогнозирования трудноформализуемых событий и разработки новых правил выбора метода и архитектуры в зависимости от условий, в которых функционирует система прогнозирования.Разработанные методы и модели позволяют выполнять оперативные и точные прогнозы, устойчивые к зашумленности входных данных и/или их недостатку, учитывающие специфику и условия прогнозирования.

Результаты моделирования показывают, что предлагаемые методы превосходят по точности прогнозирования известные аналоги. Так, применительно к задаче прогнозирования средних скоростей движения по городским дорогам в сравнении с моделями интернет-сервиса «Яндекс.Пробки», ARIMA и LSTM получено преимущество более чем на 10% по показателю MAPE. При прогнозировании лексического содержания новостных лент как на 1-часовом, так и на 4-часовом горизонтах предлагаемый метод превосходит архитектуру LSTM. Так, для словаря из 306 слов общая процентная ошибка типа «0» и типа «1» для LSTM выше на 12.2%, а MAPE выше на 7.6%. Для словаря из 1000 слов и 1-часового горизонта прогноза эти показатели составляют 20.0% и 4.0% соответственно, а для 4-часового горизонта они составляют 10.4% и 2.9%, соответственно.

Разработанные архитектуры могут немного уступать известным подходам по времени, затрачиваемому на прогнозирование, но выигрывают за счет отсутствия необходимости в переобучении системы. В известных методах полный цикл переобучения занимал от 19229 до 28674 мс. Предлагаемый подход отличается гибкостью и может быть настроен для работы с различными задачами и возможностями.

Даны **рекомендации** по использованию результатов исследования для прогнозирования событий в приложениях, требующих прогнозирования временных рядов со сложной динамикой и влиянием большого числа неявно связанных факторов. Они могут быть использованы в перспективных НИР и ОКР, а также в учебном процессе.

В качестве **перспектив дальнейшей разработки темы** можно указать дальнейшее повышение точности прогнозов за счет совершенствования правил управления ассоциативным вызовом из памяти нейронных сетей, внедрение подходов активного обучения, увеличение скорости нейросетевой обработки информации за счет реализации алгоритмов на графических процессорах.

Полученные результаты соответствуют паспорту специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в журналах из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёной степени кандидата наук, на соискание учёной степени доктора наук:

1. Осипов В. Ю., Милосердов Д. И. Нейросетевое прогнозирование событий для роботов с непрерывным обучением // Информационно-управляющие системы. – 2020. – №5(108). – С. 33-42. DOI: 10.31799/1684-8853-2020-5-33-42

2. Милосердов Д. И. Архитектурные особенности программных систем нейросетевого прогнозирования с непрерывным обучением // Информационные технологии. – 2020. – Т. 26, № 11. – С. 641-647. DOI: 10.17587/it.26.641-647

В зарубежных изданиях, индексируемых в WoS/Scopus:

3. Osipov V., Nikiforov V., Zhukova N., Miloserdov D. Urban traffic flows forecasting by recurrent neural networks with spiral structures of layers // Neural Computing and Applications. – 2020. – Vol. 32(209). DOI: 10.1007/s00521-020-04843-5

4. Osipov V., Kuleshov S., Zaytseva A., Levonevskiy D., Miloserdov D. Neural network forecasting of news feeds // Expert Systems with Applications – 2020. – Vol. 169. DOI: 10.1016/j.eswa.2020.114521

5. Osipov V., Miloserdov D. Neural Network Forecasting of Traffic Congestion // Digital Transformation and Global Society, DTGS 2019. – In Communications in Computer and Information Science. – 2019. – Vol. 1038. DOI: 10.1007/978-3-030-37858-5_20

6. Osipov V., Zhukova N., Miloserdov D. Neural Network Associative Forecasting of Demand for Goods // Experimental Economics and Machine Learning, EEML 2019. – 2019. – Vol. 2479.

7. Miloserdov I., Miloserdov D. Development of Stability Control Mechanisms in Neural Network Forecasting Systems // Journal of Physics: Conference Series, 2021. DOI: 10.1088/1742-6596/1864/1/012105

В других изданиях:

8. Милосердов Д. И. Программный комплекс нейросетевого прогнозирования временных рядов // 5-я Международная научная конференция «Технологическая перспектива в рамках евразийского пространства: новые рынки и точки экономического роста». – 2019. – С. 166-169.

9. Милосердов И.В., Милосердов Д.И. Разработка механизмов обеспечения устойчивости в нейросетевых системах прогнозирования (Материалы конференции «Информационные технологии в управлении», 2020 г.) URL: -https://itc.etu.ru/assets/files/itc-2020/papers/198.pdf

10. Милосердов Д.И. Нейросетевое прогнозирование событий для интеллектуальных роботов с непрерывным обучением // Технологические тренды и наукоемкая экономика: бизнес, отрасли, регионы. Коллективная монография. Под редакцией О.Н. Кораблевой [и др.]. – 2020. – С. 27-37. DOI: 10.53115/ 9785001880134

Интеллектуальная собственность:

11. Свидетельство о государственной регистрации программы для ЭВМ №2019662053. Осипов В. Ю., Милосердов Д. И. Программа прогнозирования событий на основе рекуррентных нейронных сетей с управляемыми элементами. 2019.

12. Свидетельство о государственной регистрации программы для ЭВМ №2020616182. Милосердов Д. И. Программа прогнозирования событий с непрерывным обучением на основе рекуррентной нейронной сети с управляемыми элементами. 2020.

Автореферат диссертации

МИЛОСЕРДОВ
Дмитрий Игоревич

МОДЕЛИ, МЕТОДЫ И АРХИТЕКТУРЫ ПРОГРАММНЫХ СИСТЕМ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ ТРУДНОФОРМАЛИЗУЕМЫХ СОБЫТИЙ С НЕПРЕРЫВНЫМ ОБУЧЕНИЕМ

Текст автореферата размещен на сайтах:
Высшей аттестационной комиссии Министерства образования
и науки Российской Федерации
https://vak.minobrnauki.gov.ru/
Федерального государственного бюджетного учреждения науки Санкт-Петербургского Федерального исследовательского центра Российской академии наук (СПб ФИЦ РАН)
http://www.spiiras.nw.ru/dissovet/

Подписано в печать «11» марта 2022 г.
Формат 60x84 1/16. Бумага офсетная. Печать офсетная.
Усл.печ.л. 1,0. Тираж 100 экз.
Заказ №