МЕТОДИЧЕСКОЕ И ПРОГРАММНО-ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ФУНКЦИОНИРОВАНИЯ АВТОМАТИЗИРОВАННЫХ СИСТЕМ

METHODOLOGICAL AND SOFTWARE-INFORMATION SUPPORT FOR THE FUNCTIONING OF AUTOMATED SYSTEMS

УДК 004.896
DOI: 10.17586/0021-3454-2024-67-11-951-957

АВТОМАТИЗАЦИЯ СОЗДАНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

В. А. Соболевский

Санкт-Петербургский федеральный исследовательский центр Российской академии наук,
Санкт-Петербург, Россия
arguzd@yandex.ru

Аннотация. Для автоматизации создания моделей машинного обучения, прогнозирующих временные ряды, предложено использовать AutoML-подход. Рассматриваются алгоритмы и технологии, позволяющие реализовать автоматизацию создания моделей. Выбрана гибридная архитектура машинного обучения, которая использовалась при решении задач автоматизации структурно-параметрического синтеза моделей и оптимизации гиперпараметров, а также при автоматическом выборе показателей оценки качества синтезированных моделей. Пользовательский интерфейс для данной системы реализован на основе платформы AutoGenNet, воплощающей концепцию No-Code разработки, которая позволяет скрыть от пользователей сложность процессов создания и обучения моделей, что обеспечивает снижение порога вхождения для работы с программой. Использование платформы AutoGenNet позволяет реализовать механизм генерации программных оболочек для эксплуатации обученных моделей, а также обеспечить автоматизацию процессов генерации и обучения гибридных моделей, что упрощает и ускоряет процесс решения задач прогнозирования временных рядов с помощью моделей машинного обучения. Полученные результаты могут быть масштабированы и использованы для создания моделей прогнозирования временных рядов в различных прикладных задачах.

Ключевые слова: *машинное обучение, прогнозирование временных рядов, AutoML, линейная регрессия, XGBoost*

Благодарности: исследования, выполненные по данной тематике, проводились при частичной финансовой поддержке бюджетной темы № FFZF-2022-0004.

Ссылка для цитирования: *Соболевский В. А.* Автоматизация создания моделей машинного обучения для решения задач прогнозирования временных рядов // Изв. вузов. Приборостроение. 2024. Т. 67, № 11. С. 951–957. DOI: 10.17586/0021-3454-2024-67-11-951-957.

AUTOMATING THE CREATION OF MACHINE LEARNING MODELS FOR SOLVING TIME SERIES FORECASTING PROBLEMS

V. A. Sobolevsky

St. Petersburg Federal Research Center of the RAS,
St. Petersburg, Russia
arguzd@yandex.ru

Abstract. It is proposed to use the AutoML approach to automate the creation of machine learning models predicting time series. Algorithms and technologies allowing to implement the automation of model creation are considered. A hybrid architecture of machine learning used in solving problems of automating the structural-parametric synthesis of models and optimizing hyperparameters, as well as in the automatic selection of indicators for assessing the quality of synthesized models, is chosen. The user interface for this system is implemented on the basis of the AutoGenNet platform, which embodies the No-Code development concept, which allows hiding the complexity of the processes of creating and training models from users, which reduces the entry threshold for working with the program. Using the

AutoGenNet platform allows implementing a mechanism for generating software shells for operating trained models, as well as automating the processes of generating and training hybrid models, which simplifies and speeds up the process of solving time series forecasting problems using machine learning models. The results obtained can be scaled and used to create time series forecasting models in various applied problems.

Key words: machine learning, time series forecasting, AutoML, linear regression, XGBoost

Acknowledgments: the research conducted on this topic was carried out with partial financial support from budget topic No. FFZF-2022-0004.

For citation: Sobolevsky V. A. Automating the creation of machine learning models for solving time series forecasting problems. *Journal of Instrument Engineering*. 2024. Vol. 67, N 11. P. 951–957 (in Russian). DOI: 10.17586/0021-3454-2024-67-11-951-957.

Введение. В современном мире объем данных, генерируемых и накапливаемых различными системами, стремительно растет. Важным аспектом анализа данных является способность предсказывать будущие значения параметров исследуемых сложных объектов (СлО) на основе имеющихся данных. Очень часто в прикладных задачах требуется прогнозировать значения, имеющие привязку ко времени и/или к очередности, в рамках которых указанные данные были получены.

Прогнозирование временных рядов играет ключевую роль в принятии решений на основе данных прошлого и настоящего. Эффективная система прогнозирования временных рядов позволяет предсказывать будущие тенденции, выявлять циклы и сезонность, а также проводить анализ аномалий. С другой стороны, для решения задач прогнозирования временных рядов все более активно применяются модели машинного обучения [1–3]. При этом широкое распространение получают гибридные модели машинного обучения [4–6], которые представляют собой комбинацию нескольких алгоритмов, что позволяет эффективно объединить их сильные стороны и компенсировать ограничения отдельных методов. Достоинства гибридных моделей проявляются в различных аспектах задач машинного обучения и аналитики данных, в том числе в задачах анализа временных рядов. Гибридные модели обеспечивают повышение значений показателей стабильности и надежности СлО по сравнению с значениями данных показателей при использовании отдельных моделей. Когда моделирование осуществляется с применением нескольких алгоритмов, результаты могут быть сверены между собой, что уменьшает вероятность ошибок при решении конкретных прикладных задач. Примером таких алгоритмов может быть гибридное обучение с использованием случайного леса и нейронных сетей. Комбинирование этих алгоритмов позволяет получить к более точные результаты по сравнению с использованием только одного из них. Нейронные сети могут уловить сложные паттерны, тогда как случайный лес помогает справляться с переобучением и недообучением.

Также важным преимуществом гибридных моделей является способность адаптироваться к различным видам шумов и аномалий в наборе данных, что делает их применимыми для решения различных классов прикладных задач, где рассматриваемые данные часто содержат аномальные значения и/или неполноту. Такая адаптивность приводит к уменьшению риска переобучения и повышает способность модели к обобщению на новых данных, что особенно актуально при анализе временных рядов.

Однако применение на практике алгоритмов глубокого обучения, и в особенности гибридных моделей, имеет существенные ограничения. Одна из основных проблем — потребность в обширных наборах обучающих данных, которые необходимы для поиска моделью закономерностей в данных и последующего точного прогнозирования на этой основе. Качество обучающих данных может существенно влиять на точность алгоритмов машинного обучения, поэтому для успешного их использования критически важны правильный сбор и подготовка данных.

Еще одной значительной проблемой машинного обучения являются высокие требования к специалистам, осуществляющим создание и интеграцию соответствующих моделей, — специалист должен обладать знаниями в областях программирования, математического анализа и анализа больших данных.

AutoML-подход к автоматизации машинного обучения. Для преодоления описанных ограничений сформировалась научно-техническая область — автоматизированное машинное обучение (AutoML — Automated Machine Learning). Основные задачи AutoML включают разработку и применение методов автоматизированного создания и обучения моделей машинного обучения. Эта научно-техническая область сформировалась только в последнее десятилетие, но уже существует множество научных и проектных групп, работающих в этом направлении.

На данный момент разработаны разные варианты программного обеспечения, которое позволяет автоматизированно создавать модели машинного обучения для различных прикладных задач. Примерами такого программного обеспечения могут служить auto-sklearn [7], Auto-Keras [8] и другие похожие решения. Однако эти программные системы часто не являются кроссплатформенными, что ограничивает спектр предназначенных для их решения задач, или представлены в виде программных библиотек, что не позволяет использовать их как автономные программные продукты. Например, существует расширение для MatLab [9], которое автоматизирует процессы обучения моделей и компилирует их в исполняемые файлы на C++, но для работы с таким программным решением необходимо наличие специалиста по машинному обучению с опытом работы в среде MatLab.

В настоящее время наиболее масштабные и универсальные системы AutoML создаются крупными компаниями. Примером такой системы может служить Google Cloud AutoML [10], разработанная компанией Alphabet Inc. Основой этой системы является облачная программная платформа, предназначенная для автоматизированного создания моделей машинного обучения, решающих разные прикладные задачи. В программной системе также предусмотрен графический пользовательский интерфейс, облегчающий ее использование. Платформа основана на принципах масштабируемости, что позволяет интегрировать созданные модели в сторонние программные комплексы. Также следует отметить, что облачный сервис, являющийся результатом работы этой системы, требует доработки для интеграции с другим программным обеспечением. Google Cloud AutoML имеет все недостатки облачной платформы. Она критически зависит от подключения к Интернету, так что без надежного соединения использование платформы может быть затруднено или невозможно. Поскольку данные хранятся на серверах, принадлежащих сторонней компании, существует риск несанкционированного доступа или утечки информации. Пользователи имеют ограниченный контроль над инфраструктурой и ресурсами, используемыми для хранения и обработки данных в облаке. В дальнейшем может оказаться сложным или даже невозможным перенос модели на другую платформу или перемещение обученной модели на локальные серверы. Есть также риск потери доступа к обученной модели из-за стихийных бедствий или иных непредвиденных обстоятельств.

В общем и целом на сегодняшний день уже существуют разработки и исследования в области AutoML, которые могут быть использованы для автоматизации создания моделей машинного обучения, решающих задачу прогнозирования временных рядов. Однако большинство этих систем представляют собой инструменты для профессионалов в области машинного обучения и не могут быть использованы неспециалистами. В представленной статье предлагается программный комплекс, решающий эту проблему.

Применение AutoML-подхода к решению задачи прогнозирования временных рядов. В качестве основной модели, используемой для решения задач прогнозирования временных рядов, была выбрана гибридная модель, комбинирующая линейную регрессию и экстремальный градиентный бустинг (eXtreme Gradient Boosting — XGBoost) [11].

Линейная регрессия является методом статистического моделирования, который используется для оценки отношения между зависимой переменной (целевой переменной) и одной или несколькими независимыми переменными (признаками), представленными линейной зависимостью. В гибридной модели линейная регрессия используется для прогнозирования направления изменений (трендов) рассматриваемых процессов. За счет простоты, интерпретируемости и низких требований к вычислительным ресурсам данная модель позволяет эффективно находить тренды в рамках работы гибридной модели.

Вторая модель — XGBoost — это модифицированная версия алгоритма градиентного бустинга, созданная в целях повышения эффективности и скорости обучения модели. Данный алгоритм реализован в виде программной библиотеки с открытым кодом, которая позволяет на практике добиться повышения точности моделей при решении прикладных задач. XGBoost также справляется с задачами регрессии, где необходимо прогнозировать непрерывные числовые значения. В гибридной модели XGBoost используется для прогнозирования разницы между трендом линейного процесса и точным значением параметра в конкретный момент времени.

Конфигурация гибридной модели обусловлена тем, что линейная регрессия хорошо справляется с задачей экстраполяции трендов, но не способна отслеживать нелинейные зависимости между параметрами. С другой стороны, XGBoost хорошо справляется с задачей регрессионного анализа нелинейных зависимостей в данных, но не может экстраполировать тренды и поэтому плохо подходит для анализа и прогнозирования временных рядов. Применение комбинации этих алгоритмов позволяет нивелировать их недостатки и использовать достоинства для повышения точности прогнозирования временных рядов.

В качестве базовых задач AutoML, которые будут решаться при автоматизации обучения гибридной модели, были выбраны следующие: структурно-параметрический синтез модели и оптимизация гиперпараметров; создание пользовательских интерфейсов для AutoML; автоматический выбор показателей оценки; создание программных оболочек для эксплуатации моделей.

Полный перечень задач AutoML гораздо больше, но в представленной работе было принято решение сконцентрироваться только на перечисленных выше. Это обусловлено тем, что AutoML является относительно новым направлением, в котором на данный момент работает лишь небольшое число научных и проектных групп. В связи с этим отсутствуют устоявшиеся подходы к описанию и решению задач в области AutoML. Каждая из отдельных задач AutoML сама по себе сложно формализована и требует индивидуального подхода к поиску решения. В настоящее время появляются алгоритмы, способные решать эти задачи по отдельности, но еще не существует решения, которое обеспечивало бы комплексную автоматизацию всех этапов создания моделей машинного обучения. Поэтому в представленном исследовании был сделан акцент на тех этапах разработки моделей, которые присутствуют в большинстве типовых задач прогнозирования временных рядов. Автоматизация именно этих этапов позволит снизить порог подготовки, требующейся от специалистов при разработке программ на базе моделей машинного обучения, что, в свою очередь, приведет к экономии времени и ресурсов.

Данное исследование ориентировано на унифицированную работу с разными задачами прогнозирования временных рядов, поэтому для комплексной автоматизации процессов их генерации и обучения должен быть использован унифицированный алгоритм. В качестве такого алгоритма выбрана модификация генетического алгоритма [12], зарекомендовавшего себя при решении ряда практических задач [13, 14]. В представленной статье этот алгоритм был адаптирован для работы с выбранной гибридной моделью машинного обучения.

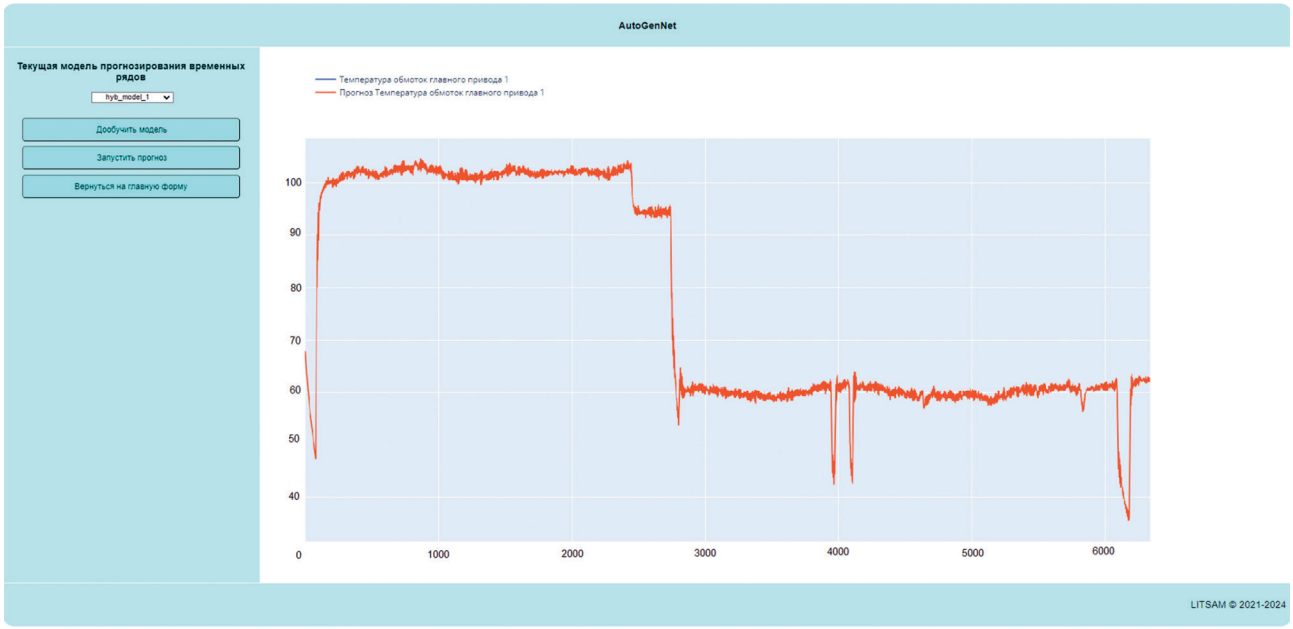
Введенные модификации, отличающие данный алгоритм от классического генетического алгоритма, позволяют использовать его для автоматизированного обучения принципиально разных архитектур с различными наборами гиперпараметров. Использование данного алгоритма позволяет решить задачи AutoML, связанные со структурно-параметрическим синтезом моделей и оптимизацией гиперпараметров, а также с автоматическим выбором показателей оценки (поскольку данный алгоритм позволяет инкапсулировать и заменять методы оценивания точности моделей).

Программный комплекс автоматизации создания моделей машинного обучения. Для того чтобы упростить процесс создания гибридных моделей, данная модель и алгоритм автоматизации процессов генерации и обучения гибридных моделей были запрограммированы и интегрированы в существующую программную систему комплексной автоматизации синтеза искусственных нейронных сетей AutoGenNet [15, 16].

Отличительная особенность данной программной системы заключается в автоматизации следующих этапов проектирования и создания моделей машинного обучения: подготовка

данных, выбор гиперпараметров модели, настройка процесса обучения модели и генерация программных оболочек для созданных моделей.

Данное расширяемое и масштабируемое программное обеспечение позволяет интегрировать различные архитектуры моделей машинного обучения и автоматизировать процесс их обучения. Также AutoGenNet предоставляет графический пользовательский интерфейс, который содержит необходимый минимум инструментов для автоматизации процесса создания различных видов моделей. Таким образом, следуя концепции No-Code разработки, сложность процессов создания и обучения моделей машинного обучения скрывается от пользователей. Это дает возможность использовать программное обеспечение не только специалистам в данной области, но и специалистам из сторонних областей, не имеющим достаточного количества знаний в области машинного обучения для создания моделей вручную. По вышеописанным причинам гибридная модель и алгоритм автоматизации процессов ее генерации и обучения были интегрированы в программный комплекс AutoGenNet. Это позволило решить задачи AutoML, связанные с созданием пользовательских интерфейсов, упрощением работы с моделями машинного обучения, а также генерацией программных оболочек для эксплуатации обученных моделей. Пример интерфейса взаимодействия с гибридными моделями прогнозирования временных рядов представлен на рисунке.



Заключение. Рассмотрены пути решения задачи автоматизации процесса генерации и обучения моделей машинного обучения для задачи прогнозирования временных рядов. Использованный в работе AutoML-подход позволяет существенно упростить и ускорить внедрение этих моделей в различные области человеческой деятельности. Применение представленной программной системы приводит к снижению затрат и ускорению разработки программных комплексов на основе моделей машинного обучения, предназначенных для решения различных прикладных задач прогнозирования временных рядов. Результаты тестирования программного обеспечения демонстрируют повышение уровня автоматизации процесса создания данных моделей за счет существенного снижения требований к специалистам, занимающимся их разработкой.

Планы дальнейшего развития включают промышленное тестирование программного обеспечения для оценки требуемых вычислительных ресурсов и расширения набора архитектур машинного обучения, которые можно использовать для прогнозирования временных рядов. Благодаря универсальности используемого алгоритма автоматизации процессов генерации и обучения моделей, а также гибкости и масштабируемости программной платформы

AutoGenNet результаты данного исследования можно использовать в качестве основы для создания многомодельного программного комплекса, решающего широкий спектр задач анализа временных рядов.

СПИСОК ЛИТЕРАТУРЫ

1. *Houndekindo F., Ouarda T. B. M. J.* Prediction of hourly wind speed time series at unsampled locations using machine learning // Energy. 2024. Vol. 299, N 131518.
2. *Moreno F. P., Rodriguez F. I., Comendador V. F. G., Jurado R. D.-A., Suarez M. Z., Valdes R. M. A.* Prediction of air traffic complexity through a dynamic complexity indicator and machine learning models // Journal of Air Transport Management. 2024. Vol. 119, N 102632.
3. *Cooper C., Zhang J., Ragai I., Gao R. X.* Multi-sensor fusion and machine learning-driven sequence-to-sequence translation for interpretable process signature prediction in machining // Journal of Manufacturing Systems. 2024. May. DOI:10.1016/j.jmsy.2024.04.010.
4. *Kebede Y.B., Yang M.-D., Huang C.-W.* Real-time pavement temperature prediction through ensemble machine learning // Engineering Applications of Artificial Intelligence. 2024. Vol. 135, N 08870.
5. *Naeini S. S., Snaiki R.* A physics-informed machine learning model for time-dependent wave runup prediction // Ocean Engineering. 2024. Vol. 295, N 116986.
6. *Castillo A. F., Garibay M. V., Diaz-Vazquez D., Yebra-Montes C., Brown L. E., Johnson A., Garcia-Gonzalez A., Gradilla-Hernandez M. S.* Improving river water quality prediction with hybrid machine learning and temporal analysis // Ecological Informatics. 2024. Vol. 82, N 102655.
7. *Feurer M., Klein A., Eggensperger K., Springenberg T. J., Blum M., Hutter F.* Auto-sklearn: Efficient and Robust Automated Machine Learning // Automated Machine Learning. 2019. P. 113–134.
8. *Jin H., Song Q., Hu X.* Auto-Keras: An Efficient Neural Architecture Search System // arXiv:1806.10282 [cs]. 2019.
9. *Shure L.* Building Optimized Models in a few steps with AutoML. 2020 [Электронный ресурс]: https://blogs.mathworks.com/loren/2020/06/13/building-optimized-models-in-a-few-steps-with-automl/, 26.06.2024.
10. *Zeng Y., Zhang J.* A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision // Computers in Biology and Medicine. 2020. Vol. 122, N 103861.
11. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // arXiv:1603.02754 [cs]. 2016.
12. *McCall J.* Genetic algorithms for modelling and optimization // Journal of Computational and Applied Mathematics. 2005. Vol. 184, iss. 1. P. 205–222.
13. *Sobolevskii V. A.* The system of convolution neural networks automated training // CEUR Workshop Proceedings. 2020. P. 100–106.
14. *Михайлов В. В., Пономаренко М. Р., Соболевский В. А.* Моделирование влияния климатических факторов на динамику надземной фитомассы растительных сообществ тундры // Глобальные климатические изменения: региональные эффекты, модели, прогнозы: Материалы Междунар. науч.-практ. конф. Воронеж: Изд-во „Цифровая полиграфия“, 2019. Т. 2. С. 106–109.
15. Свид. о рег. программ для ЭВМ № 2021668925. Программа автоматизированной генерации и обучения искусственных нейронных сетей / *Б. В. Соколов, В. А. Соболевский.* 22.10.2021.
16. *Соболевский В. А.* Использование технологий AutoML для решения задач мониторинга // Информатизация и связь. 2024. № 1. С. 90–97.

СВЕДЕНИЯ ОБ АВТОРЕ

Владислав Алексеевич Соболевский — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория информационных технологий в системном анализе и моделировании; мл. научный сотрудник; E-mail: arguzd@yandex.ru

Поступила в редакцию 23.07.24; одобрена после рецензирования 01.08.24; принята к публикации 23.09.24.

REFERENCES

1. Houndekindo F., Ouarda T.B.M.J. *Energy*, 2024, vol. 299, art. no. 131518.
2. Moreno F.P., Rodriguez F.I., Comendador V.F.G., Jurado R. D.-A., Suarez M.Z., Valdes R.M.A. *Journal of Air Transport Management*, 2024, vol. 119, art. no. 102632.
3. Cooper C., Zhang J., Ragai I., Gao R.X. *Journal of Manufacturing Systems*, 2024, no. 1–4(75), https://doi.org/10.1016/j.jmsy.2024.04.010.
4. Kebede Y.B., Yang M.D., Huang C.-W. *Engineering Applications of Artificial Intelligence*, 2024, vol. 135, art. no. 08870.
5. Naeini S.S., Snaiki R. *Ocean Engineering*, 2024, vol. 295, art. no. 116986.
6. Castillo A.F., Garibay M.V., Diaz-Vazquez D., Yebra-Montes C., Brown L.E., Johnson A., Garcia-Gonzalez A., Gradilla-Hernandez M.S. *Ecological Informatics*, 2024, vol. 82, art. no. 102655.
7. Feurer M., Klein A., Eggensperger K., Springenberg T.J., Blum M., Hutter F. *Automated Machine Learning*, 2019, pp. 113–134, DOI:10.1007/978-3-030-05318-5_6.
8. Jin H., Song Q., Hu X. *Proc. of the 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining — KDD'19*, 2019, DOI:10.1145/3292500.3330648.
9. Shure L. *Building Optimized Models in a few steps with AutoML*, 2020, https://blogs.mathworks.com/loren/2020/06/13/building-optimized-models-in-a-few-steps-with-automl/.
10. Zeng Y., Zhang J. *Computers in Biology and Medicine*, 2020, vol. 122, art. no. 103861.
11. Chen T., Guestrin C. *22nd ACM SIGKDD International Conference*, August 2016, DOI:10.1145/2939672.2939785.
12. McCall J. *Journal of Computational and Applied Mathematics*, 2005, no. 1(184), pp. 205–222.
13. Sobolevskii V.A. *CEUR Workshop Proceedings*, 2020, vol. 2803, pp. 100–106.
14. Mikhailov V.V., Ponomarenko M.R., Sobolevsky V.A. *Global'nyye klimaticheskiye izmeneniya: regional'nyye effekty, modeli, prognozy* (Global Climate Change: Regional Effects, Models, Forecasts), Proceedings of the Intern. Sci. and Pract. Conf., Voronezh, 2019, vol. 2, pp. 106–109. (in Russ.)
15. Certificate on the state registration of the computer programs 2021668925, *Programma avtomatizirovannoy generatsii i obucheniya iskusstvennykh neyronnykh setey* (Program for Automated Generation and Training of Artificial Neural Networks), B.V. Sokolov, V.A. Sobolevsky, Priority 22.10.2021. (in Russ.)
16. Sobolevsky V.A. *Informatization and communication*, 2024, no. 1, pp. 90–97. (in Russ.)

DATA ON AUTHOR

Vladislav A. Sobolevsky — PhD; St Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Information Technologies in System Analysis and Modeling; Junior Researcher; E-mail: arguzd@yandex.ru

Received 23.07.24; approved after reviewing 01.08.24; accepted for publication 23.09.24.