СПИСОК ЛИТЕРАТУРЫ

1. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский. – М.: Горячая линия-Телеком, 2008. – 452 с.

2. Саймон, Д. Алгоритмы эволюционной оптимизации / Д. Саймон. – М.: ДМК Пресс, 2020. – 940 с.

3. Ротач, В.Я. Теория автоматического управления / В.Я. Ротач. – М.: Изд-во МЭИ, 2004. – 399 с.

4. https://python-control.readthedocs.io/en/latest/generated/control.step_response.html (дата обращения: 10.03.2025).

5. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.differential_evolution.html (дата обращения: 10.03.2025).

КРИТИКА ДАТАСЕТА – НЕОБХОДИМАЯ ЧАСТЬ РАБОТЫ С ПРОМЫШЛЕННЫМИ БОЛЬШИМИ ДАННЫМИ

Елисейкин М.М. – аспирант
Очков В.Ф. – д-р техн. наук, профессор
ФГБОУ ВО «НИУ «МЭИ»

АННОТАЦИЯ. Рост использования машинного обучения и ИИ в промышленности, а также увеличения числа людей, занятых в анализе промышленных больших данных, требуют адаптации методов работы с данными. Междисциплинарный подход к проблеме корректности и безопасности работы с промышленными большими данными позволяет добавить в методы работы аналитиков принцип «нулевое доверие» и этап «критика датасета».

КЛЮЧЕВЫЕ СЛОВА: анализ данных, большие данные, критика датасета, метрологический дефицит.

ABSTRACT. The growing use of machine learning and AI in industry, along with the rising number of professionals engaged in the analysis of industrial big data, requires the adaptation of data-handling methods. An interdisciplinary approach to ensuring correctness and security in working with industrial big data allows to introduce the "zero trust" principle and a dedicated "dataset critique" stage into analysts' workflows.

KEYWORDS: data analysis, big data, dataset critique, metrological deficit.

Расширение промышленного использования технологий машинного обучения и искусственного интеллекта увеличивает остроту проблемы качества исходных данных, которые используются при обучении системы. Зачастую, увлекшись самой возможностью анализировать датасеты, исследователи не учитывают вопрос о том, какой физический смысл имеют собранные данные и насколько на результаты их работы влияют физические явления, связанные с процессом сбора этих данных. На одной из недавних конференций, где авторы озвучивали проблему метрологического дефицита в промышленных данных [1, 2], из зала был задан вопрос: «Что делать?». На этот простой и очевидный

вопрос был дан такой же простой и очевидный ответ: дорабатывать методы работы с большими данными в соответствии с изменяющимися условиями.

При рассмотрении данной проблемы необходимо учитывать, что она является универсальной и не ограничивается лишь использованием технологий, связанных с анализом данных в промышленности. Например, подобная проблема настолько остро стоит при анализе данных, собранных в результате геофизического мониторинга, что Институт физики Земли РАН посвятил ей специальный выпуск в 2024 году [3, 4].

Вот как описывают там свое видение проблемы: «И хотя многие приемы такой предварительной обработки являются общепринятыми и широко используются на практике, вопрос о принципах построения таких процедур и “границе допустимого” при их применении до сих пор остается открытым. При этом активность его обсуждения в публикациях далеко не соответствует тому влиянию, которое он потенциально оказывает на результаты».

Таким образом, если проблема физического смысла анализируемых данных является не узкотематической, а междисциплинарной, то при ее решении можно использовать подходы из других научных областей.

Кибербезопасность – «нулевое доверие» к датасету

В настоящий момент широко распространены методы предварительной обработки данных, включающие в себя проверку корректности формата записи и очистку от единичных выбросов. Однако такой подход основан исключительно на математическом смысле данных и не учитывает их физический смысл.

Проблема в том, что в промышленном применении датасеты, использованные при обучении, выступают, по сути, в роли управляющих сигналов для АСУТП. Если по каким-то причинам при обучении будут использованы некорректные данные, то в процессе эксплуатации промышленной установки может произойти авария. А значит, рассматривая датасеты как часть процесса функционирования киберфизической системы, мы вправе применить к ним принципы, используемые в области кибербезопасности.

Одним из подобных принципов является «нулевое доверие» (Zero Trust) – рассмотрение в качестве потенциально опасных любых управляющих сигналов, даже если они происходят изнутри периметра безопасности [5]. Применительно к рассматриваемой проблеме это означает, что возможность использования конкретного датасета в конкретной задаче нужно подвергать сомнению и явным образом подтверждать.

У нас уже есть этапы разработки, связанные с безопасностью функционирования АСУТП, – валидация модели и проверка модели на адекватность. Однако эти этапы затрагивают лишь конечный этап разработки и не включают в себя анализ исходных данных. Учитывая особенность технологий машинного обучения и ИИ, на конечном этапе разработки мы можем не иметь возможности обнаружить, что проблема кроется в физическом смысле исходных данных.

История – критика датасета

Также полезный подход можно взять из дисциплины, которая достаточно далека от функционирования АСУТП, – из истории.

В процессе своего развития современная история как научная дисциплина прошла путь от простого «механического» объединения разных исторических источников в сводный текст до критики исторического источника [6], включающую в себя как внешнюю (анализ происхождения), так и внутреннюю (анализ смысла) критику.

На первый взгляд, может показаться, что нет ничего общего между функционированием современных киберфизических систем и историческими источниками возрастом в несколько веков или даже тысячелетий. Однако даже сейчас, находясь лишь в начале пути применения больших данных в промышленности, мы уже сталкиваемся с тем, что датасеты могут содержать данные возрастом в несколько десятилетий. А если рассматривать ситуацию в науке в целом, то исследователям уже приходится иметь дело с датасетами, составленными из данных собранных за несколько веков [7].

Таким образом, критика датасета как целенаправленный поиск подтверждения его целостности, физической осмысленности и возможности использования в конкретном случае должна стать обязательной частью работы с датасетом.

Заключение

С каждым днем в профессии, связанные с анализом датасетов, приходит все больше людей. Эти люди могут не иметь глубоких знаний в предметной области или времени и желания на глубокий анализ используемых ими данных. А значит, они будут делать свою работу в соответствии с установившимся практиками и популярными рекомендациями.

И для повышения качества анализа промышленных данных и безопасности функционирования киберфизических систем, настроенных в соответствии с результатами этого анализа, было бы полезно ввести в практику обязательный этап «Критика датасета», в процессе которого в обязательном порядке будут даваться ответы на следующие вопросы.

1. Применимы ли собранные данные к конкретной установке?

2. Собирались ли эти данные без смены измерительных приборов, методик измерения и других событий, влияющих на метрологические характеристики собранных данных?

3. Собирались ли эти данные без перезапуска установки, ее ремонта, модернизации и других событий, влияющих на режим работы установки?

4. Являются ли анализируемые данные исходными или они были подвергнуты нормализации, очистке от выбросов, заполнению пробелов, обогащению дополнительными данными и другим действиям, влияющим на полноту и достоверность анализируемых данных?

5. Если датасет содержит данные из разных источников, то не могли ли эти данные быть собраны в настолько разных местах, разных условиях и разными

измерительными приборами, что объединение их в общий датасет не имело бы смысла?

Можно возразить, что хороший аналитик будет задаваться этими вопросами без напоминаний. Однако, кроме хороших аналитиков, анализом промышленных больших данных будут заниматься и обычные аналитики, которые просто будут делать свою работу так, как их этому научили. И если посмотреть на описания учебных курсов, в большом количестве представленных в Интернете, то можно увидеть, что начальный процесс работы с датасетом ограничивается лишь предобработкой, состоящей из нормализации, очистки от выбросов, заполнении пробелов и прочих действий, не относящихся к физическому смыслу данных. Критика датасета, как постановка под сомнение его физического смысла и применимости в конкретном случае, в рекомендациях отсутствует.

Целью данных тезисов является приглашение к обсуждению и совместной выработке решения озвученной проблемы.

СПИСОК ЛИТЕРАТУРЫ

1. Елисейкин, М.М. Метрологический дефицит в больших данных / М.М. Елисейкин, В.Ф. Очков // 32-я Международная конференция «Математика. Компьютер. Образование», Пущино, 27–31 января 2025 г. – Ижевск: АНО «Ижевский институт компьютерных исследований». – С. 9.

2. Елисейкин, М.М. Метрологический дефицит в промышленных «больших данных» / М.М. Елисейкин, В.Ф. Очков // Законодательная и прикладная метрология. – 2024. – № 4. – С. 19–24. DOI: 10.32446/2782-5418.2024-4-19-24.

3. Дещеревский, А.В. Проблема качества данных при режимном геофизическом мониторинге: кто виноват и что делать? / А.В. Дещеревский // Наука и технологические разработки. – 2024. – Т. 103. – № 3. – С. 3–26. DOI: 10.21455/std2024.3-1.

4. Елисейкин, М.М. Фрагментация временных рядов – не аномалия, а норма / М.М. Елисейкин, В.Ф. Очков // Наука и технологические разработки. – 2024. – Т. 103. – № 4. – С.39–46. DOI: 10.21455/std2024.4-3.

5. Нулевое доверие (Zero Trust) [Электронный ресурс] / Энциклопедия «Касперского». Глоссарий. – URL: https://encyclopedia.kaspersky.ru/glossary/zero-trust/ (дата обращения: 09.03.2025).

6. Источниковедение: Теория. История. Метод. Источники рос. истории: учеб. пособие для гуманит. спец. / И.Н. Данилевский, В.В. Кабанов, О.М. Медушевская, М.Ф. Румянцева. – М.: РГГУ, 1998. – 701 с.

7. Елисейкин, М.М. О метрологических характеристиках исторических данных / М.М. Елисейкин, В.Ф. Очков // Законодательная и прикладная метрология. – 2024. – № 5. – С. 47–51. DOI: 10.32446/2782-5418.2024-5-47-51.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГБОУ ВО «НИУ «МЭИ»
Филиал ФГБОУ ВО «НИУ «МЭИ» в г. Волжском
ПАО «ЛУКОЙЛ»
ПАО «РусГидро»
ПАО «Россети Юг»
АО «Системный оператор Единой энергетической системы»
ПАО «Мосэнерго»

МЕЖДУНАРОДНАЯ НАУЧНО-ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ «ЭНЕРГЕТИКА БУДУЩЕГО: ИНЖИНИРИНГ И ЦИФРОВИЗАЦИЯ», ПОСВЯЩЕННАЯ 30-ЛЕТИЮ ФИЛИАЛА ФГБОУ ВО «НИУ «МЭИ» В Г. ВОЛЖСКОМ

г. Волжский
16 мая 2025 года

ТЕЗИСЫ ДОКЛАДОВ

ВОЛЖСКИЙ МЭИ 2025

УДК 620.9+621.3+681.5
ББК 31

Международная научно-практическая конференция «Энергетика будущего: инжиниринг и цифровизация», посвященная 30-летию филиала ФГБОУ ВО «НИУ «МЭИ» в г. Волжском, 16 мая 2025 г.: тезисы докладов. – Волжский: Филиал ФГБОУ ВО «НИУ «МЭИ» в г. Волжском, 2025. – 98 с.

ISBN 978-5-94721-179-5

Тезисы докладов Международной научно-практической конференции «Энергетика будущего: инжиниринг и цифровизация» освещают актуальные проблемы в области энергетики. Тексты тезисов, представленные авторами, сверстаны и при необходимости сокращены. Как правило, сохранена авторская редакция.

Печатается по решению Учебно-методического совета филиала ФГБОУ ВО «НИУ «МЭИ» в г. Волжском.

УДК 620.9+621.3+681.5
ББК 31

ISBN 978-5-94721-179-5