Попова О. В., Григорьев Я. Ю., Жарикова Е. П., Григорьева А. Л.
O. V. Popova, Ya. Yu. Grigoriev, E. P. Zharikova, A. L. Grigorieva

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ ОЦЕНКИ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССОВ

APPLICATION OF MACHINE LEARNING METHODS IN PROCESS ASSESSMENT TASKS

Попова Ольга Владимировна – студент Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: olia.kms@mail.ru.
Olga V. Popova – Student, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: olia.kms@mail.ru.

Григорьев Ян Юрьевич – кандидат физико-математических наук, доцент, проректор по учебной работе Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: jan198282@mail.ru.
Yan Yu. Grigoryev – PhD in Physical and Mathematical Sciences, Associate Professor, Vice Rector for Academic Affairs, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: jan198282@mail.ru.

Жарикова Евгения Павловна – ведущий программист, ассистент кафедры прикладной математики Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: evgenia131094@mail.ru.
Evgeniya P. Zharikova – Assistant, Department of Applied Mathematics, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: evgenia131094@mail.ru.

Григорьева Анна Леонидовна – кандидат физико-математических наук, доцент, заведующий кафедрой прикладной математики Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: pm@knastu.ru.
Anna L. Grigorieva – PhD in Physical and Mathematical Sciences, Associate Professor, Head of the Department of Applied Mathematics, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: pm@knastu.ru.

Аннотация. В работе рассматриваются разные подходы к оценке технологических процессов на основе данных, считываемых в разные моменты времени, на разных этапах с учётом влияния входных параметров. Используются модели регрессионного моделирования с применением классических подходов и методов машинного обучения. Производится сравнение полученных решений, определяется необходимость применения методов интеллектуального анализа данных. Рассматриваются модель нейронной сети MLP и метод XGBoost, при обучении используются метрики MAE, MSE.

Summary. The paper discusses different approaches to assessing technological processes based on data read out at different points in time, at different stages, taking into account the influence of input parameters. Regression modeling models are used that implement classical approaches and machine learning methods. The comparison of the obtained solutions is made, the necessity of applying the methods of data mining is determined. The MLP neural network model and the XGBoost method are considered, the MAE, MSE metrics are used for training.

Ключевые слова: регрессия, нейронные сети, машинное обучение, производственный процесс.

Key words: regression, neural networks, machine learning, production process.

УДК 004.048

В настоящее время в производстве задействуется высокотехнологичное оборудование, оснащённое большим количеством различных датчиков, позволяющих получить большой объём данных о реализации технологического процесса. Для анализа и прогнозирования требуется обработка полученной информации. Существуют классические методы оценки статистических данных, к которым относятся методы регрессионного моделирования, имеющие автоматизацию в

различных средах MS Excel, Statistica и т.д. При построении регрессионных зависимостей возможны ситуации, когда вычислительный процесс становится достаточно трудоёмким. Это связано с видом выбранной зависимости и структурой набора данных. При этом полученный результат не всегда обеспечивает достаточную точность. Для решения таких проблем рассматриваются более сложные зависимости, что способствует ещё большему количеству вычислений. Эффективными методами решения такого класса задач являются методы машинного обучения, позволяющие достаточно эффективно обрабатывать большие объёмы информации и хорошо описывающие процессы в рамках диапазонов имеющихся данных.

При рассмотрении производственного процесса можно выделить две группы параметров. К первой группе относятся входные параметры, которые можно разделить на контролируемые и неконтролируемые. Контролируемые параметры могут изменяться пользователем с целью повышения эффективности процесса, неконтролируемые оказывают влияние на процесс, но не могут изменяться оператором. Вторая группа параметров определяет значения на выходе исследуемого процесса. Целями работы являются анализ технологического процесса на основе данных, установление зависимости и определение возможности влияния на результат с помощью изменения параметров.

В работе рассматривается технологический процесс преобразования некоторого вида сырья в конечный полезный продукт с примесями. Данные о процессе формируются на предприятии за заданный период времени с помощью датчиков и замеров, проводимых операторами процесса. Целью исследования является отбор параметров, оказывающих влияние на процесс, изменение которых может способствовать повышению эффективности последнего.

На первом этапе проводится анализ данных классическими методами. Анализ большого количества различных функциональных зависимостей показывает низкое качество разрабатываемых моделей. Ниже приводятся варианты зависимостей с лучшими полученными показателями.

Функции:

$$y = a + b * \ln(x) + \frac{c}{\ln(x)} + \frac{d}{\ln(x)^2} + \frac{e}{\ln(x)^2} + \frac{f}{\ln(x)^3} + \frac{g}{\ln(x)^3} + h * \ln(x)^4$$

(a = -6 3100 614, b = -1 8748 367, c = -1.2026, d = -3 031 927, e = -1.22, f = -221 662.8, g = -51 962 215, h = -3028.43) с коэффициентом детерминации, равным 0,28;

$$y = a + bx^{0.5} + cx + dx^{1.5} + ex^2 + fx^{2.5} + gx^3 + hx^{3.5} + ix^4$$

(a = 67 715 929, b = -2.03, c = 2.67, d = -1.99, e = 9.27, f = -2.7, g = 5.06, h = -5.3, i = 2.42) с коэффициентом детерминации, равным 0,38.

Подобрать вид функциональной связи, обеспечивающей надёжную зависимость, не представляется возможным. Даже достаточное усложнение вида регрессионной зависимости не позволяет обеспечить высокие значения коэффициента детерминации и допустимые показатели при оценке надёжности с помощью критерия Фишера.

На следующем этапе исследования рассматриваются методы машинного обучения, которые достаточно требовательны к набору исходных данных и позволяют находить хорошие приближения в указанном диапазоне. К этим методам относятся: искусственные нейронные сети (распознавание, кластеризация, прогноз); эволюционное программирование; генетические алгоритмы (оптимизация); ассоциативная память (поиск аналогов, прототипов); нечёткая логика; деревья решений; системы обработки экспертных знаний [9]. Эти методы позволяют лучше описывать соответствующие процессы и имеют возможность динамически адаптироваться к изменению потоков обрабатываемой информации за счёт пополнения обучающей выборки и т.д. Помимо имеющихся преимуществ, они обладают недостатками, связанными с узкой направленностью применения полученных моделей в производственных задачах.

Для построения модели производится подготовка данных. Первоначальный набор данных разделяется случайным образом на обучающую часть (70 %), валидационную (10 %) и тестовую (20 %). Обучающая и валидационная части используются непосредственно в процессе обучения, а

тестовая часть необходима для того, чтобы проверить качество обучения посредством нижеописанных метрик.

Данные для обучения хранятся в csv файле, имеющем разметку, отражённую в табл. 1.

Таблица 1

Разметка csv файла

Время работы автоклава, ч	Железо, %	Мышьяк, %	...	Хлориды, г/кг
24	17.3	4.01	...	3.70
24	16.8	4.5	...	4.90
23.9	20.1	7.3	...	8.05

Для решения поставленной задачи методами машинного обучения проводится сравнение качества решения при помощи XGBoost и нейронных сетей.

Рассмотрим метод XGBoost. Он заключается в реализации деревьев решений с градиентным усилением, разработанных для увеличения скорости и производительности. Результаты метрики обученной модели представлены в табл. 2.

Таблица 2

Результаты метрик обученной модели XGBoost

Метрика	Значение
MSE	575.65
MAE	15.57
r2	0.97

Приводится вариант решения с применением нейронной сети MLP (multi layer perceptron). Архитектура MLP может содержать один или несколько скрытых слоёв, обеспечивающих уровни абстракции. Прогнозы делаются на выходном слое, называемом видимым (см. рис. 1).

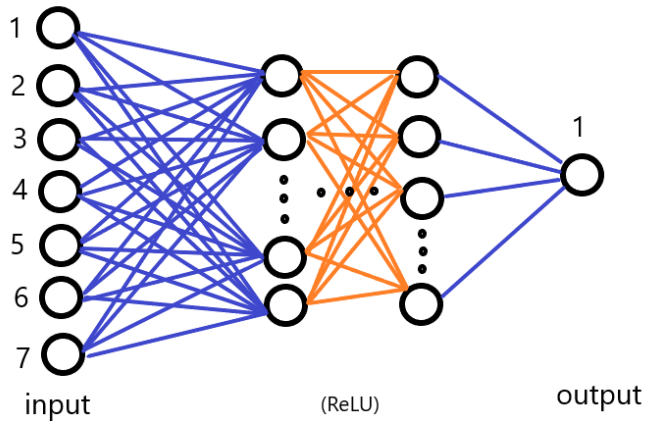

Рис. 1. Архитектура MLP

На входе сети 7 параметров, соответствующих значениям «Fe», «As», «Time» и т.д. Экспериментальным путём определяется функция активации. Лучшие результаты обеспечивает функция ReLU. Определяется функция потерь MSE. Реализуется алгоритм оптимизации Adam (adaptive moment estimation). Обучение производится с размером пакета 512, скоростью обучения 0.0001, со случайной заморозкой 50 % связей на каждой итерации и нормализацией для каждого пакета.

График изменения функции потерь на обучающей и валидационной частях набора данных представлен на рис. 2.

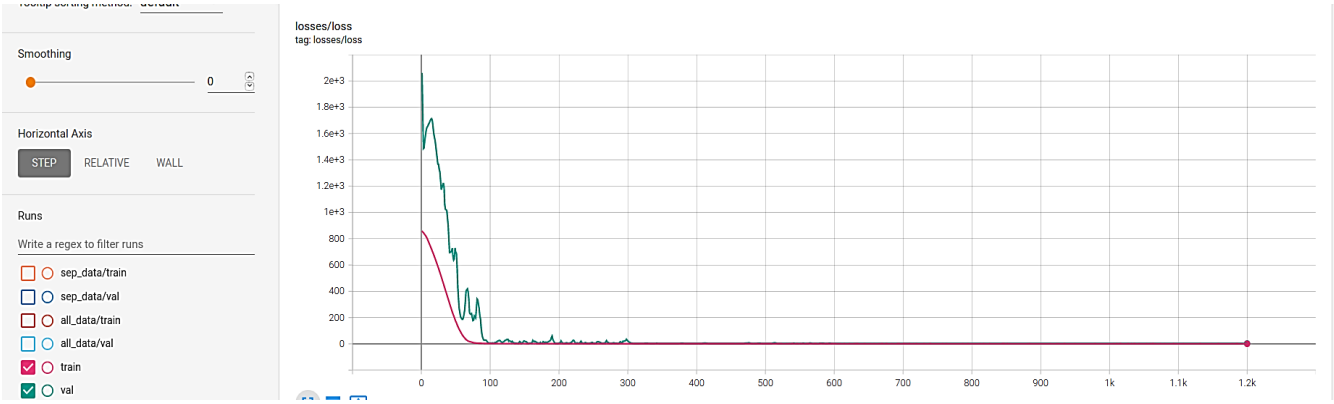

Рис. 2. График изменения функции потерь во время обучения нейронной сети на фактических данных

Результаты метрики обученной модели представлены в табл. 3.

Таблица 3

Результаты метрики обученной модели MLP

Метрика	Значение
MSE	1.41
MAE	0.04
r2	0.003

В качестве метрик для анализа качества методов используются MAE и MSE. MAE (Mean Absolute Error) измеряет среднюю сумму абсолютной разницы между фактическим значением и прогнозируемым значением:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|, \quad \text{где } e_t = original_t - predict_t.$$

MSE (Mean Squared Error) измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}e_t^2, \qquad \text{где } e_t = original_t - predict_t.$$

Исследование показывает, что использование нейронной сети даёт более качественный результат на заданном диапазоне значений и обеспечивает возможность решения поставленной задачи. Сравнительные данные представлены в табл. 4.

Таблица 4

Сравнительная таблица значений метрик методов MLP и XGBoost

Метрика	Метод	
	MLP	XGBoost
MSE	1.41	575.65
MAE	0.04	15.57
r2	0.003	0.97

Вывод. В результате исследования, проводимого на наборе данных, представленных предприятием, реализующим конкретный технологический процесс, была показана эффективность применения нейронных сетей в сравнении с классическими методами и рядом других методов машинного обучения (XGBoost при оценках эффективности методов, базирующихся на метриках MSE и MAE) при поиске решений высокой точности на заданном рабочем диапазоне. Увеличение выборки и добавление данных может способствовать выявлению других альтернатив.

ЛИТЕРАТУРА

1. Обеспечение информационной безопасности при решении задач мониторинга подстилающей поверхности с помощью беспилотных летательных аппаратов / Е. П. Жарикова, И. А. Трещев, Я. Ю. Григорьев, А. Л. Григорьева // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2020. – № I-1 (41). – С. 4-8.
2. Челухин, В. А. Применение математических моделей в информационных технологиях / В. А. Челухин, В. А. Абрамсон // Материалы I Международной научно-практической конференции молодых учёных. – Комсомольск-на-Амуре: ФГБОУ ВО «КнАГУ», 2021. – С. 117-120.
3. Амосов, О. С. Использование генетического алгоритма для настройки глубоких нейронных сетей в задачах классификации компьютерных атак / О. С. Амосов, С. Г. Амосова, Д. С. Магола // Информатика и системы управления. – 2020. – № 4 (66). – С. 104-117.
4. Исследование архитектур глубоких нейронных сетей со сверточными и рекуррентными слоями для задач распознавания аномалий сетевого трафика в компьютерных системах / О. С. Амосов, С. Г. Амосова, Ю. С. Иванов, С. В. Жиганов // Управление развитием крупномасштабных систем (MLSD'2019): материалы двенадцатой международной конференции / под общ. ред. С. Н. Васильева, А. Д. Цвиркуна. – М.: Международный научно-исследовательский институт проблем управления РАН, 2019. – С. 995-1005.
5. Амосов, О. С. Оценивание случайных последовательностей с использованием нечётких систем и кластеризации / О. С. Амосов, Д. С. Магола, Е. А. Малашевская // Информатика и системы управления. – 2012. – № 1 (31). – С. 146-155.
6. Khanghah, E. R. Study of the relationship between overinvestment, leverage, and liquidity in companies listed in tehran stock exchange / E. R. Khanghah, B. Partovi // Revista Genero & Direito. – 2020. – № 9 (2). – P. 558-572.
7. Pivnicka, K. Fish species-diversity in small Czech and Moravian streams / K. Pivnicka, J. Poupe; M. Svatora // Zivocisna Vyroba. – 1995. – № 40 (4). – P. 177-180.
8. Андрейчиков, А. З. Интеллектуальные информационные системы / А. З. Андрейчиков, О. Н. Андрейчикова. – М.: Финансы и статистика, 2004. – 424 с.
9. Бардин, А. К. Особенности оценки эффективности производственных процессов при их дифференциации / А. К. Бардин, В. Е. Сайкинов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. – 2016. – № 124. – С. 1493-1504.
10. Zharikova, E. P. Applications of computer vision in cross-sectoral tasks / E.P. Zharikova, Y.Y. Grigoriev, A. L. Grigorieva // Lecture notes in networks and system. – 2021. – № 200. – P. 415-426.
11. Математическое моделирование процесса получения энергии при использовании перепада температур день-ночь на основе нейронных сетей и искусственного интеллекта / В. А. Челухин, М. А. Кружаев, Пьей Зон Аунг, Е. В. Абрамсон // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2020. – № I-1 (41). – С. 18-24.