УДК 65.011.56
doi:10.18720/SPBPU/2/id24-168

Костенко Дмитрий Андреевич [1],
аспирант;
Олейников Виталий Сергеевич [2],
ст. преподаватель;
Хохловский Владимир Николаевич [3],
доцент, канд. техн. наук, доцент

ПРЕДОБРАБОТКА ЗАШУМЛЕННЫХ ДАННЫХ ПЕРЕД НЕЙРОСЕТЕВЫМ АНАЛИЗОМ

[1, 2, 3] Россия, Санкт-Петербург, Санкт-Петербургский политехнический университет Петра Великого;
[1] kostenko_da@spbstu.ru, [2] oleinikov_vs@spbstu.ru, [3] hohlovskij_vn@spbstu.ru

Аннотация. В работе предложена методика уточнения нейросетевой модели производственного процесса, представленного зашумлёнными статистическими данными. Уточнение достигается путём разделения исходных данных и вычленения признаков, соответствующих разным состояниям оборудования, в отдельные наборы

данных. Отличие от известных подходов состоит в применении существующих алгоритмов обработки данных в комбинации с новым вариантом реализации нейросетевого анализа. Изложение иллюстрируется описанием методики подготовки данных на примере одного из производственных цехов.

Ключевые слова: нейросетевая модель, аппроксимация временных рядов, мультикритериальная аппроксимация, зашумленные данные, производительность, статистический анализ.

Dmitry A. Kostenko [1],
Postgraduate (PhD) Student;
Vitaly S. Oleinikov [2],
Senior Lecturer;
Vladimir N. Khokhlovskiy [3],
Candidate of Technical Sciences (PhD), Associate Professor

PREPOSSESSING OF NOISY DATA FOR NEURAL NETWORK ANALYSIS

[1, 2, 3] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia;
[1] kostenko_da@spbstu.ru, [2] oleinikov_vs@spbstu.ru, [3] hohlovskij_vn@spbstu.ru

Abstract. A methodology is proposed for refining neural network models of a production process available with noisy statistical data. The refinement is made by partitioning the source data and isolating parts corresponding to different states of equipment into separate datasets. The distinctive feature of the work lies in the use of known data science approaches in a combination with a new way of implementation of neural network analysis. The description of a production workshop illustrates the data preparation methodology.

Keywords: neural network model, time series approximation, multicriteria approximation, noisy data, performance, statistical analysis.

Введение

Актуальность вопроса экономии тепла и электроэнергии в промышленности не требует подтверждения [7]. В рамках этого направления ниже рассматривается задача прогнозирования потребления; чем точнее удастся решить эту задачу, тем больше потенциал экономии тепла и электроэнергии в конкретном производственном цикле «генерация — потребление — запрос на генерацию».

В качестве объекта управления рассмотрим цех предприятия, потребляющий пар. Требуется выполнить мультикритериальную аппроксимацию временного ряда [3, 4], чтобы спрогнозировать потребление пара на несколько часов вперёд. Имеется статистика влияющих параметров, получаемая с датчиков на оборудовании. Построение нейросетевой модели на основании необработанных данных дает результаты относительно невысокой точности. Для улучшения точности модели предлагается разделить входные данные на две категории. Первая, условно именуемая «стационарной», содержит статистику параметров за периоды, характерные для периодов устоявшегося режима работы системы. Вто-

рая, условно именуемая «переходной», содержит статистику параметров за периоды, характеризующие работу системы в переходных состояниях. Создаются две модели, каждая из которых будет обучена отдельно.

1. Анализ объекта прогнозирования

Для анализа с учетом конфигурации реального оборудования и отдельных особенностей физических процессов в нём был выбран один из цехов промышленного предприятия. Как видно из таблицы 1, набор данных по цеху содержит данные по четырём контурам и блок дополнительных параметров. Каждый контур описывается параметрами нескольких ёмкостей, обозначенных цифрами от 1 до 6. Параметры обозначены следующим образом: V — объём жидкости в ёмкости, pH — уровень кислотности в ёмкости, T — температура внутри ёмкости, m — масса осадка в ёмкости.

Анализ проводится с целью выяснения степени зависимости основного параметра, «Расхода пара», от вышеуказанных влияющих параметров и построения модели для аппроксимации временных рядов.

Исходный набор данных охватывает один месяц, замеры выполняются с периодом один замер в минуту. Таким образом, набор включает 44 640 векторов по 45 параметров, в виде двухмерного массива. Для проверки нейросетевой модели зависимости основного параметра от влияющих набор данных разбит на последовательные блоки по 1 000 значений, для каждого из которых выполнен анализ на глубину в 360 значений. Из исходных 44 640 замеров получилось 41 эксперимент, для каждого из которых подсчитан параметр MAPE — Mean Absolute Percentage Error. Эта метрика позволяет рассчитать отклонение в относительных единицах, что упрощает оценку полученной модели [5].

Таблица 1

Параметры технологического процесса

Цех				
Контур 1	Контур 2	Контур 3	Контур 4	Дополнительно
*V*1	*T*1	*V*1	*V*1	*T*1
*T*1	*m*1	pH1	pH1	*T*2
pH2	pH2	*T*1	*T*1	*T*3
*T*2	*T*2	*V*2	*V*2	*T*4
*V*2	*V*2	pH2	pH2	*T*5
*V*3	pH 3	*T*2	—	—
*V*4	*T*3	*V*3	—	—
pH5	*V*3	pH3	—	—
*T*5	*V*4	*T*3	—	—
*V*5	*V*5	*V*4	—	—
*V*6	*m*2	pH4	—	—
—	—	*T*4	—	—

Использовались два набора данных, первый (см. рис. 1), составленный по данным из таблицы 1, и второй (см. рис. 2), для которого данные из колонки «Дополнительно» в таблице 1 была убрана.

Рис. 1. График MAPE для полного набора данных 1

Рис. 2. График MAPE для уменьшенного набора данных 2

На основании полученных графиков было получено общее представление о точности работы модели с заданным набором данных, и принято решение провести дополнительный анализ данных в наборе, чтобы найти и удалить из набора данных значения, приводящие к зашумлению результата и снижению итогового значения MAPE.

Анализ графиков, полученных при помощи оригинального набора данных (см. рис. 3), при помощи набора данных, усеченного до 6 замеров в минуту (см. рис. 4), и при помощи усечённого набора данных с приме-

нением фильтра Савицкого-Голея (см. рис. 5) не позволил выявить характерных зависимостей, позволяющих улучшить набор данных.

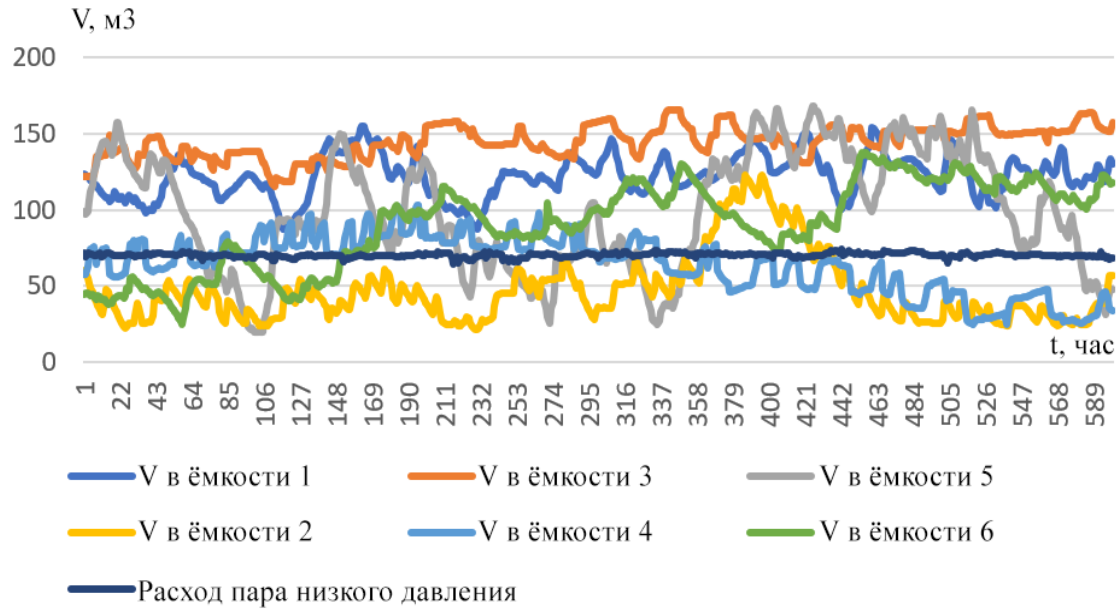

Рис. 3. График потребления пара и уровня в ёмкостях, 6 замеров в минуту

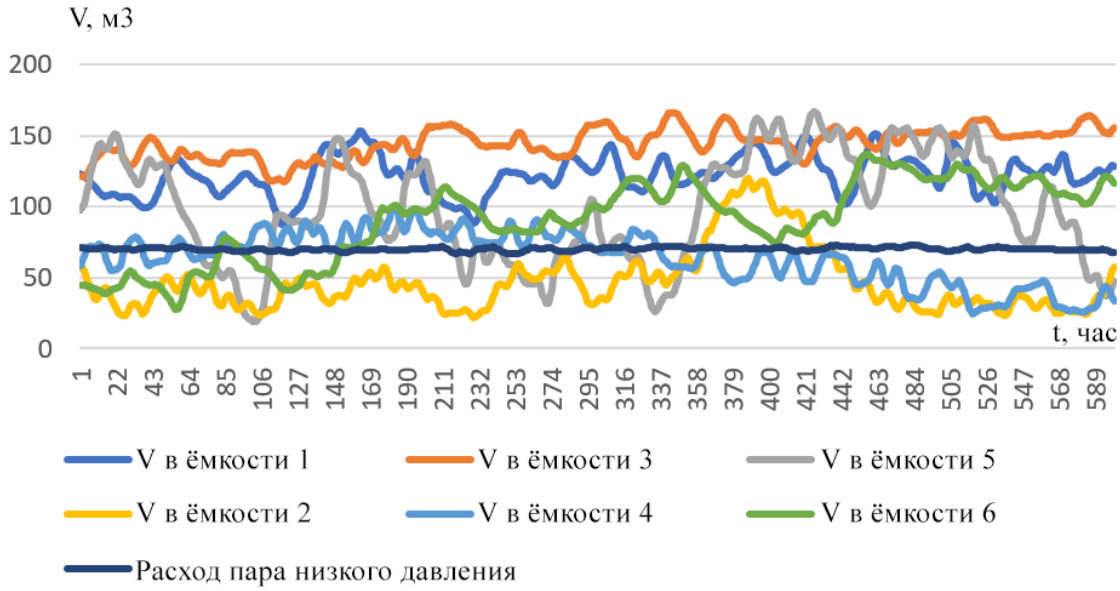

Рис. 4. График потребления пара и уровня в ёмкостях, 6 замеров в минуту и фильтр Савицкого-Голея

2. Предлагаемое решение

Чтобы улучшить результат аппроксимации, было принято решение разделить набор данных на две части: одна должна включать данные, характеризующие работу системы в устоявшемся режиме, а другая —

в режиме переходном. Для этого на основании исходного набора данных была построена матрица значений [1], показывающая изменение значений параметра «Расход пара» (см. табл. 2). Значения округлены до целого.

Таблица 2

Матрица значений

	10:00	11:00	12:00	13:00	14:00	15:00	16:00
1 день	16	16	14	15	15	15	15
2 день	14	14	14	13	14	14	15
3 день	16	16	16	16	16	16	14
4 день	16	16	14	15	15	15	14

С её помощью была составлена матрица изменения значений (см. табл. 3). Она показывает разницу в значении Расхода пара между соседними периодами изменений.

Таблица 3

Матрица изменения значений

	10:00	11:00	12:00	13:00	14:00	15:00	16:00
1 день	0	0	-2	-1	0	0	0
2 день	-1	0	0	-1	+1	0	+1
3 день	+1	0	0	0	0	0	-2
4 день	+2	0	-2	+1	0	0	-1

Построение матриц и анализ их содержимого позволил определить требования к методу очистки нового набора данных. В частности, установлено, что наибольшее количество времени система проводит в состоянии, где значение Расхода находится в пределах от 13 до 16 или с округлением — от 14 до 18. Если система работает в таком режиме более 60 минут, то режим работы можно считать устоявшимся. После составления визуальных таблиц обработка данных была автоматизирована при помощи пользовательских подпрограмм на языке Python [2].

На основании вышесказанного было принято решение разделить набор данных на два: первый набор «стационарный», куда попадают значения из устоявшегося режима работы. Второй — «переходный», куда попадают остальные значения.

Обучая одну нейронную сеть на устоявшемся режиме, можно уменьшить значения MAPE; вторая сеть, обученная на выбросах, должна быть эффективнее в работе с переходным режимом [6].

Нейронные сети были обучены и получившиеся наборы данных проанализированы (см. рис. 5 и 6). Для целей данного исследования показатель MAPE менее 15 считается удовлетворительным. Сеть обучена на 14.000 последних значений.

Рис. 5. График MAPE для набора данных устоявшегося режима работы

Полученные графики не могут прямо накладываться на исходные с рисунков 1 и 2, так как размеры наборов данных не совпадают — разделение наборов по режимам работы системы выявило соотношение 58 на 42 (устоявшийся режим / переходный режим). Однако среднее значение MAPE для устоявшегося режима опустилось до 6.99 с учётом пика, обусловленного внешним фактором (точка 23 на рис. 6).

Рис. 6. График MAPE для набора данных переходного режима работы

Заключение

Как показали проведенные эксперименты, предложенная методика позволяет эффективно увеличить точность работы нейронной сети с заданным набором данных в т. ч. в случае, когда корреляция между влияющим и зависимыми параметрами задана неявно. Удалось добиться устойчивого результата при обучении на 14 000 значений. Планируется

получение прогнозов по двум моделям параллельно и выборе стратегии исходя из оценок снизу и сверху. Автоматизированная интерпретация результата на основе определения текущего состояния системы является естественным продолжением данной работы.

Отдельные параметры процесса не могут быть корректно зафиксированы и представляют собой нуль-значения. Дополнительная кластеризация по единицам оборудования позволяет улучшить результат, но выходит за рамки данной публикации.

Благодарности

Работа выполнена в рамках гранта РНФ № 23-29-00551 от 13.01.2023 г.

Список литературы

1. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science. 50 важнейших понятий. – СПб.: БХВ, 2018. – 416 с.

2. Грас Д. Data Science. Наука о данных с нуля. – СПб.: БХВ, 2020. – 400 с.

3. Нильсен Э. Практический анализ временных рядов. Прогнозирование со статистикой и машинное обучение. – М.: Диалектика-Вильямс, 2021. – 544 с.

4. Сидоров С.Г, Никологорская А.В. Анализ временных рядов как метод построения прогноза потребления электроэнергии // Вестник ИГЭУ. – 2010. – №3.

5. Шишков Е.М., Проничев А.В., Савельев А.А. Прогнозирование временных рядов с применением методов машинного обучения на примере графика выдачи мощности электрической станции // МНИЖ. – 2022. – № 2-1 (116).

6. Шмойлова Р.А., Садовникова Н.И. Анализ временных рядов и прогнозирование. – М.: Синергия, 2016. – 152 с.

7. Юсупов О.Я. Актуальность проблемы экономии электроэнергии в современных условиях развития экономики промышленности // Экономика и социум. – 2021. – № 2-2 (81).