Вычислительные технологии, 2024, том 29, № 4, с. 4–23. © ФИЦ ИВТ, 2024
Computational Technologies, 2024, vol. 29, no. 4, pp. 4–23. © FRC ICT, 2024
ISSN 1560-7534
eISSN 2313-691X

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

DOI:10.25743/ICT.2024.29.4.002

Модели интерпретации данных полупроводниковых газовых сенсоров на основе методов машинного обучения

А. Д. Козьмин*, А. А. Редюк
Новосибирский государственный университет, 630090, Новосибирск, Россия
*Контактный автор: Козьмин Артем Дмитриевич, e-mail: a.kozmin@g.nsu.ru
Поступила 03 мая 2023 г., доработана 16 августа 2023 г., принята в печать 03 октября 2023 г.

Исследуется применение методов машинного обучения для восстановления концентрации угарного газа CO в воздухе по данным полупроводниковых датчиков газа (metal-oxide gas sensor — MOX). Концентрация CO критически важна при контроле качества воздуха, так как повышенные уровни могут нанести вред здоровью людей и животных. Проведен анализ выходных данных датчиков и созданы новые признаки, включая CO_h, учитывающий зависимость концентрации от времени суток. Построены модели множественной линейной и полиномиальной регрессии, а также нейронных сетей для восстановления концентраций CO. Исследовалось влияние регуляризации на точность интерпретации данных газовых сенсоров. Работа демонстрирует возможность использования методов машинного обучения для контроля качества воздуха.

Ключевые слова: полупроводниковый датчик газа, угарный газ, полносвязная нейронная сеть, регуляризация, линейная регрессия, полиномиальная регрессия.

Цитирование: Козьмин А.Д., Редюк А.А. Модели интерпретации данных полупроводниковых газовых сенсоров на основе методов машинного обучения. Вычислительные технологии. 2024; 29(4):4–23. DOI:10.25743/ICT.2024.29.4.002.

Введение

Основной механизм работы полупроводниковых газовых сенсоров основан на изменении электропроводности полупроводниковой пленки вследствие абсорбции и десорбции анализируемой газовой смеси на ее поверхности. Степень адсорбции зависит от концентрации газа, в результате этих эффектов изменяется электрическая проводимость сенсора. Таким образом, измерением сопротивления полупроводникового датчика возможно восстановить концентрацию газа. Однако точная интерпретация выходных данных газовых сенсоров затрудняется из-за неявной зависимости между показаниями датчика и концентрацией целевого газа. Кроме того, газовые сенсоры подвержены влиянию различных условий окружающей среды и проявляют высокую перекрестную чувствительность к другим газам, что дополнительно усложняет их использование.

Для разработки моделей интерпретации данных, получаемых от полупроводниковых газовых сенсоров, можно использовать эталонный газоанализатор на базе точного спектрометра и проводить полевые испытания. В исследовании, описанном в [1], оценена работа 24 идентичных коммерческих сенсорных платформ AQMesh при мониторинге таких газов, как: оксид азота (NO), диоксид азота (NO_2), оксид углерода (CO)

и озон (O_3). Полученные результаты показали, что отклик каждого датчика на одни и те же внешние условия уникален, а лабораторная калибровка не может корректировать работу датчиков в реальных условиях. В связи с этим необходимо проводить индивидуальную калибровку каждого датчика в полевых условиях.

Для обработки выходных данных газовых сенсоров применено большое количество различных алгоритмов. В одной из ранних работ [2] использовались нейронные сети с прямой связью (FFNN) для прогнозирования концентраций угарного газа (CO). В этой работе была достигнута относительная точность прогнозирования CO в 26 % при использовании гиперболического тангенса в качестве активационной функции нейрона и тренировочной выборки порядка 2000 ч. Кроме того, проведен анализ точности алгоритма в зависимости от размеров тренировочной выборки, который показал, что достаточный размер тренировочного набора составляет порядка двух недель. В более поздних работах [3, 4] доказано, что полевая калибровка с использованием методов обучения с учителем более эффективна, чем рассматриваемые методы линейной и полилинейной регрессии.

В статье [5] рассмотрены методы, позволяющие улучшить точность обработки данных газовых сенсоров при ограниченном количестве размеченных данных. Был использован метод обучения с частичным привлечением учителя, который существенно улучшил точность обработки данных при продолжительной работе газового сенсора. Метод позволяет использовать неразмеченные данные для обучения, что особенно полезно в случаях, когда доступ к размеченным данным ограничен. Результаты исследования показали, что применение обучения с частичным привлечением учителя может быть эффективным инструментом для повышения точности обработки данных газовых сенсоров при ограниченных ресурсах.

Другим перспективным подходом к интерпретации данных газовых сенсоров является использование рекуррентных нейронных сетей. В работе [6] рассмотрены два типа рекуррентых нейронных сетей — TDNN (time delay neural network) и NARX (nonlinear autoregressive exogenous model) и было проведено сравнение их результатов с нейронной сетью с прямой связью. Оказалось, что динамические нейронные сети демонстрируют более высокую точность, чем FFNN. В дальнейшем этот подход был развит в ансамблевые методы. В работе [7] исследованы ансамблевые модели рекуррентных нейронных сетей для мониторинга концентраций CO, O_3 и NO_2. Результаты исследования показали, что объединение четырех типов моделей (long short-term memory — LSTM, gated recurrent unit — GRU, bidirectional LSTM — Bi-LSTM, bidirectional GRU — Bi-GRU) дает лучший результат, чем каждая рекуррентная сеть по отдельности.

В последнее время кластерный подход становится все более популярным для решения проблемы дрейфа и воспроизводимости конкретного датчика. В работе [8] исследовался кластерный подход на примере анализа концентраций NO_2 и O_x с использованием медианного сигнала от шести аналогичных датчиков, что позволило отобрать четыре различных метода (multiple linear regression — MLR, boosted regression trees — BRT, Bayesian linear regression — BLR, Gaussian processes — GP). Кластеризация датчиков частично решила проблему дрейфа и невоспроизводимости показаний отдельных датчиков. В работе [9] исследовано четыре метода машинного обучения (MLR, regression based on k-nearest neighbors — KNN, random forest — RF, support vector regression — SVR) для калибровки датчиков O_3 на основе оксидов металлов. Векторная регрессия SVR оказалась наилучшим методом калибровки. Также проанализировано объединение данных нескольких датчиков для различных моделей. Это показало, что использование

от четырех до шести датчиков в методе SVR значительно улучшает среднеквадратичную ошибку.

Стоит отметить работу [10], где рассмотрены различные методы прогнозирования концентраций примесей внутри помещений. В рамках исследования авторы использовали четыре метода (rolling average, RF, gradient boosting, LSTM) и выявили, что концентрации сильно зависят от времени суток и дня недели. В ходе исследований оказалось, что использование времени суток в часах, а также дня недели значительно помогало в прогнозировании концентрации загрязняющих веществ внутри помещения для любого из рассмотренных алгоритмов.

В настоящей работе исследуются известные методы машинного обучения для восстановления концентрации угарного газа CO в окружающем воздухе по выходным данным полупроводникового газового сенсора. Измерение концентрации CO является критически важным для контроля качества воздуха внутри помещений и в окружающей среде, так как повышенная концентрация CO может вызвать серьезные проблемы со здоровьем людей и животных. Для достижения этой цели выполнены анализ структуры и корреляционный анализ выходных данных набора газовых датчиков CO. На основе результатов анализа созданы новые признаки, учитывающие зависимость концентрации газа CO от времени суток. Были построены различные модели множественной линейной и полиномиальной регрессии, а также несколько архитектур нейронных сетей с прямой связью для восстановления реальных значений концентраций CO. Также на этих моделях анализировалось влияние различных способов регуляризации на точность интерпретации данных газовых сенсоров.

1. Постановка задачи и анализ набора данных

В этом разделе представлена общая постановка задачи регрессии и обсуждаются основные меры качества (метрики) в задачах регрессионного анализа. Описан исходный набор данных и приводятся результаты первичного корреляционного анализа.

1.1. Регрессионная модель

Регрессионный анализ — один из методов статистического анализа, который позволяет оценить связь между зависимой переменной и одной или несколькими независимыми переменными. Регрессионная модель задется функцией $f(X_i, \beta)$, где i — индекс строки данных, X_i — вектор независимых переменных, а β — неизвестные параметры модели. Предполагается, что зависимая переменная Y_i является суммой значений некоторой модели с добавленной случайной ошибкой ε_i:

$$Y_i = f(X_i, \beta) + \varepsilon_i. \tag{1}$$

Параметры модели настраиваются таким образом, чтобы модель наилучшим образом приближала зависимые переменные Y_i. Цель исследования — найти такую функцию f, которая будет наиболее точно соответствовать данным. Например, в случае многомерной линейной регрессии (MLR) функция (1) предполагается равной

$$f_1(X_i, \beta) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = \sum_{j=0}^{p} \beta_j X_{ji}, \tag{2}$$

где p — количество независимых переменных (наблюдаемых признаков), а $X_{0i} = 1$ для каждой i-й строки данных. Для нахождения оптимальных параметров модели β обычно используется метод наименьших квадратов, цель которого — минимизация суммы квадратов отклонений:

$$Q = \sum_i (Y_i - f(X_i, \beta))^2 = \sum_i \varepsilon_i^2 \to \min_\beta . \tag{3}$$

Для случая многомерной линейной регрессии решение имеет вид $\beta = (X^T X)^{-1} X^T \mathbf{Y}$, где X — матрица признаков, а $\mathbf{Y}$ — вектор зависимой величины. Для случая полиномиальной регрессии добавляются новые признаки в функцию f. Например, для двумерной полиномиальной регрессии второй степени функция имеет вид

$$f_2(X_i, \beta) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i} X_{1i} + \beta_5 X_{2i}^2. \tag{4}$$

1.2. Метрики

Выбор метрики является обязательным этапом при построении регрессионных моделей. Метрика необходима для оценки качества построенных моделей и позволяет сравнивать результат прогнозирования модели с истинными значениями. Одна из наиболее часто встречающихся в регрессионном анализе метрик — среднеквадратичная ошибка

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2, \tag{5}$$

где N — количество прогнозов, а Y_i, $\hat{Y}_i$ — наблюдаемое и предсказанное значения концентрации соответственно. Часто для анализа используют среднюю абсолютную процентную ошибку

$$\mathrm{MAPE} = \frac{100\ \%}{N} \sum_{i=1}^{N} \frac{|Y_i - \hat{Y}_i|}{|Y_i|}. \tag{6}$$

Также для нашей задачи мы ввели и использовали метрику GRE, которая определяется как процент некорректных прогнозов. Прогноз считается некорректным, если прогнозируемая концентрация лежит вне диапазона ± 25 % от истинного значения. В противном случае прогноз считается верным.

1.3. Используемый набор данных

Для построения и тестирования моделей интерпретации данных взяты сигналы полупроводниковых газовых сенсоров, собранные с помощью мультисенсорного устройства, разработанного компанией Pirelli Labs. Измерения проведены в центре города с интенсивным автомобильным движением в период с марта 2004 г. по апрель 2005 г. [2]. Набор данных содержит 9358 строк измерений с периодичностью одно измерение в час, усредненных по пяти различным датчикам, совместно с целевыми значениями газов. Кроме измерений датчика CO в наборе присутствуют данные датчика неметановых углеводородов NMHC, датчика общих оксидов азота NO_x, датчиков NO_2 и O_3. Проводились также измерения температуры воздуха T и его относительной влажности RH. Целевые

значения концентраций газов получены эталонным анализатором. После удаления пустых данных осталось 7344 строк, состоящих из измеряемых сопротивлений датчиков, температуры и влажности воздуха, а также целевой концентрации CO.

В ходе исследования моделировалась ситуация калибровки газового датчика. Он непрерывно калибруется в полевых условиях в течение некоторого периода времени, а далее используется по назначению без перекалибровки. С учетом этого при разделении данных в обучающую и тестовую выборки попадали данные с 10 марта 2004 г. по 24 июня 2004 г. и с 24 июня 2004 г. по 4 апреля 2005 г. соответственно (при размере тренировочной выборки в 2000 первых строк данных). Анализатор позволял измерять концентрацию целевого газа в пределах от 0.1 до 12 мг/м^3 с дискретностью в 0.1 мг/м^3.

1.4. Корреляционный анализ

Для выявления взаимного влияния различных газов на датчики использовалась корреляция Пирсона. Коэффициент корреляции r Пирсона характеризует наличие линейной зависимости между величинами X_1, X_2 и рассчитывается по формуле $r_{X_1X_2} = \mathrm{cov}_{X_1X_2}/(\sigma_{X_1}\sigma_{X_2})$, где cov — ковариация, а σ — среднеквадратичное отклонение. Значение коэффициента находится в диапазоне $[-1, 1]$ и интерпретируется следующим образом: сильная отрицательная зависимость при $r \approx -1$, отсутсвие линейной зависимости при $r \approx 0$ и сильная положительная зависимость при $r \approx 1$.

Корреляционная матрица для отфильтрованных данных представлена на рис. 1. Значение коэффициента корреляции между сопротивлением датчика угарного газа R_{CO} и сопротивлением датчика неметановых углеводородов R_{NMHC} составляет 0.89, что свидетельствует о наличии сильной линейной зависимости между ними. Это может быть вызвано как значительной зависимостью между примесями неметановых углеводородов

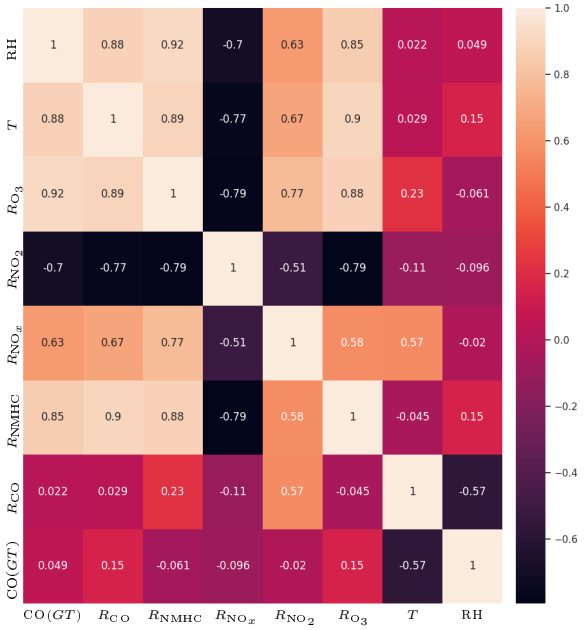

Рис. 1. Корреляционная матрица Пирсона: CO(GT) — целевая концентрация CO; R_{CO}, R_{NMHC}, R_{NO_x}, R_{NO_2}, R_{O_3} — сопротивления соответствующих датчиков; T, RH — температура и влажность воздуха

Fig. 1. Pearson correlation matrix: CO(GT) — target CO concentration; R_{CO}, R_{NMHC}, R_{NO_x}, R_{NO_2}, R_{O_3} — resistance of observed sensors; T, RH — air temperature and humidity

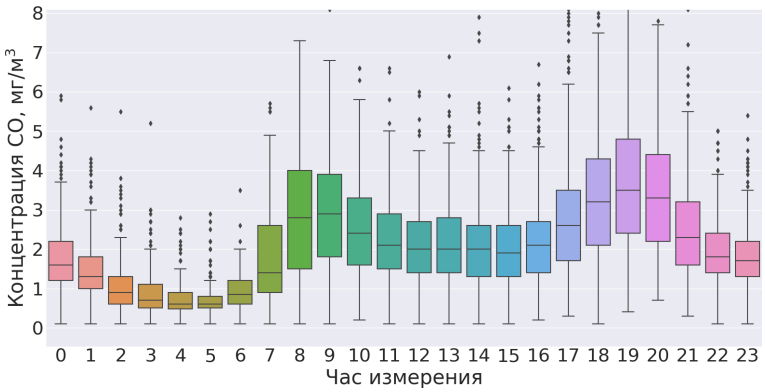

Рис. 2. Зависимость концентрации CO от часа наблюдения. Нижняя и верхняя границы “ящика” — первый и третий квартили, линия в середине “ящика” — медиана
Fig. 2. Dependence of CO concentration on the hour of observation. The lower and upper limits of the “box” are the first and third quartiles, the line in the middle of the “box”’ is the median

и угарного газа в самой атмосфере, так и плохой избирательностью (селективностью) самих датчиков. Следует отметить, что коэффициент корреляции между сопротивлением датчика R_{NMHC} и целевой концентрацией $\text{CO}(GT)$ равен 0.92, что выше, чем между сопротивлением датчика угарного газа R_{CO} и целевой концентрацией $\text{CO}(GT)$, равной 0.88. Как отмечено в работе [2], учет сопротивления датчика R_{NMHC} позволяет повысить точность прогнозирования концентрации угарного газа.

1.5. Создание новых признаков

Для улучшения точности предложенных моделей созданы новые признаки. Проанализирована зависимость концентрации CO от часа измерения, для чего было изучено распределение концентрации CO по эталонному газоанализатору в течение дня. На рис. 2 представлена ящичная диаграмма, отображающая зависимость целевой концентрации от часа наблюдения.

Исходя из диаграммы можно сделать вывод о сильной зависимости между временем измерения и уровнем концентрации угарного газа. Наибольшие медианные значения концентрации CO наблюдаются утром с 8 до 9 ч и в вечернее время с 18 до 20 ч, составляя около 3 мг/м^3. В дневное время медианные значения падают до уровня в 2 мг/м^3, а в ночное время наблюдается значительный спад примесей CO с медианными значениями около 0.5 мг/м^3 с 4 до 5 ч утра. Значительные изменения медианных значений концентраций угарного газа в течение суток могут быть вызваны временем наибольшей активности людей. Максимальные медианные значения концентрации наблюдались в часы пик движения транспорта.

Таким образом, было решено помимо исходных значений сопротивлений датчиков, температуры и влажности воздуха учитывать также час измерения концентрации. Для этого вычислены средние значения концентраций CO_h для каждого часа суток путем усреднения по всем данным эталонного газоанализатора. Полученный признак включен в рассматриваемые модели с целью повышения их точности.

2. Полиномиальная регрессия

Рассмотрим различные модели на основе линейной и полиномиальной регрессии, используя методы L_1- и L_2-регуляризации для преодоления проблем мультиколлинеар-

ности и переобучения. На основе L_1-регуляризации проводится отбор признаков для полиномиальной регрессии. Также исследуются особенности ложно предсказанных значений концентраций CO.

2.1. L_1-, L_2-регуляризация

Для моделей линейной и полиномиальной регрессии оптимальное решение для неизвестных параметров β может быть найдено путем минимизации суммы квадратов отклонений (3). Однако при наличии мультиколлинеарности возникают проблемы, когда матрица X^TX является вырожденной или близкой к этому состоянию. Это происходит, когда два или более признака сильно коррелируют между собой. В таких случаях обратная матрица $(X^TX)^{-1}$ может содержать экстремальные собственные значения. Для решения этих проблем необходимо использовать регуляризацию.

Регуляризация Тихонова, также известная как L_2-регуляризация, позволяет улучшить качество модели путем добавления нового члена в критерий качества:

$$Q_{L_2} = \sum_i \left(Y_i - f(X_i, \beta)\right)^2 + |\Gamma\beta|^2 \to \min_\beta,$$

где $\Gamma = \lambda E$, λ — неотрицательный гиперпараметр, а E — единичная матрица. После дифференцирования по β новое решение β^* имеет вид $\beta^* = (X^TX + \lambda^2 E)^{-1}X^T\mathbf{Y}$. Хотя это решение уменьшает дисперсию, оно становится смещенным. В линейных моделях регуляризация Тихонова позволяет избежать проблем мультиколлинеарности и переобучения.

L_1-регуляризация заключается в добавлении нового члена к критерию качества в виде

$$Q_{L_1} = \sum_i \left(Y_i - f(X_i, \beta)\right)^2 + \lambda|\beta| \to \min_\beta,$$

где λ — неотрицательный гиперпараметр. Эта регуляризация может зануллять значения некоторых параметров, что позволяет проводить отбор признаков.

2.2. Сравнение полиномиальных моделей

Базовой моделью линейной регрессии служит модель, состоящая из одного признака, а именно сопротивления датчика R_{CO}. Для обучения и тестирования моделей доступные данные были разделены на две выборки: тренировочную, состоящую из первых 2000 строк данных, и тестовую, содержащую оставшиеся 5344 строки данных. Кроме того, исследовались модели линейной регрессии с такими признаками, как сопротивление датчика неметановых углеводородов R_{NMHC}, температура воздуха T и созданный признак среднего значения концентрации CO_h. Сводные результаты для линейных моделей без регуляризации приведены в табл. 1. Как видно, при добавлении признаков T, CO_h и данных датчика неметановых углеводородов R_{NMHC} повышается точность линейных моделей без регуляризации. В таблице также представлены результаты моделей линейной регрессии с L_1-регуляризацией. Гиперпараметр регуляризации λ оптимизирован на логарифмической сетке в диапазоне от 10^{-4} до 10^2 с использованием перекрестной валидации для временных рядов TimeSeriesSplit с параметром валидации 10. Как видно из таблицы, регуляризация улучшает относительную точность MAPE и незначительно уменьшает процент выбросов при прогнозировании GRE. Однако при этом происходит незначительный рост ошибки MSE. При использовании L_2-регуляризации изменения менее значительные.

Т а б л и ц а 1. Результаты линейной регрессии
Table 1. Results of linear regression

Модель	Без регуляризации			L_1-регуляризация		
	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
R_{CO}	36.7	0.57	41.2	–	–	–
R_{CO}, T	40.4	0.66	50.2	–	–	–
R_{CO}, CO_h	36.3	0.52	39.9	–	–	–
R_{CO}, R_{NM}	32.9	0.37	33.6	32.2	0.38	33.0
R_{CO}, R_{NM}, T	**31.2**	**0.30**	**28.3**	**30.1**	**0.32**	**27.9**
R_{CO}, R_{NM}, CO_h	32.9	0.37	32.6	32.1	0.37	31.8
R_{CO}, R_{NM}, CO_h, T	**31.1**	**0.30**	**26.4**	**29.8**	**0.31**	**26.2**

Т а б л и ц а 2. Результаты полиномиальной регрессии
Table 2. Results of polynomial regression

Модель	Без регуляризации			L_1-регуляризация		
	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
$\text{Pol}_2(R_{\text{CO}}, R_{\text{NM}})$	26.8	0.30	28.1	26.8	0.30	28.0
$\text{Pol}_2(\dots) + \text{CO}_h$	26.0	0.29	24.9	26.0	0.29	24.8
$\text{Pol}_2(R_{\text{CO}}, R_{\text{NM}}, T)$	25.8	0.26	23.8	25.7	0.26	23.1
$\text{Pol}_2(\dots) + \text{CO}_h$	**25.5**	**0.24**	**20.7**	**25.5**	**0.24**	**20.5**

На следующем этапе строились различные полиномиальные модели второй и третьей степени. Наилучшие результаты показали многомерные полиномиальные модели второй степени, результаты которых приведены в табл. 2. Добавление новых признаков происходит согласно правилу (4), при этом среднее значение концентрации CO_h за час не использовалось в создании полиномиальных признаков. В таблице также приведены результаты нескольких полиномиальных моделей с L_1-регуляризацией. Настройка гиперпараметра λ проводилась на той же сетке, что и в случае линейной регрессии.

Согласно результатам, представленным в табл. 2, при учете часа измерения путем добавления нового признака CO_h в полиномиальной регрессии происходит улучшение точности рассматриваемых моделей. Использование L_1-регуляризации незначительно снижает ошибку GRE. Наименее существенными признаками в случае L_1-регуляризации являются квадрат температуры T^2 и квадрат сопротивления датчика неметановых углеводородов R^2_{NMHC}. Избавление от этих признаков в рассматриваемых моделях не улучшает их точности, как не улучшает точности и учет влажности воздуха RH. Рассмотрение полиномиальных моделей третьей степени и выше значительно увеличивало количество признаков, но не приводило к улучшению точности.

2.3. Анализ ошибочных предсказаний

Представим анализ некорректных прогнозов концентрации CO в модели полиномиальной регрессии второй степени $\text{Pol}_2(R_{\text{CO}}, R_{\text{NM}}, T) + \text{CO}_h$ при L_1-регуляризации (результаты выделены жирным шрифтом в табл. 2). На рис. 3, *а* показана точечная диаграмма, отображающая предсказанные значения концентрации в зависимости от истинных значений на тестовых данных. Синяя прямая на графике соответствует значениям, при которых предсказанные значения совпадают с истинными. Красные прямые выделяют внутреннюю область, где ошибка предсказания составляет менее 25 % относительно

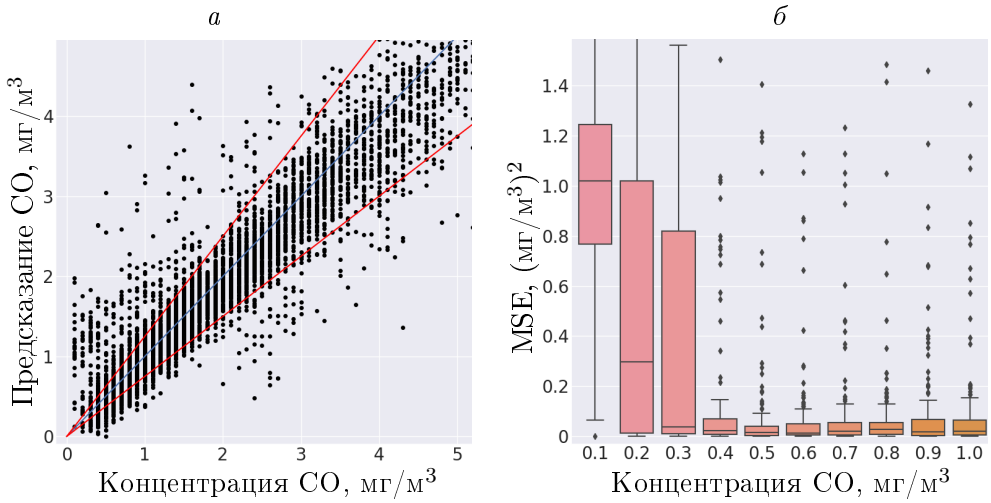

Рис. 3. Точечная диаграмма CO (*а*). По вертикали отложены результаты предсказаний концентрации, по горизонтали — истинные значения концентрации; ящичная диаграмма квадратичных отклонений MSE (*б*). По вертикальной оси — отклонение на тестовых данных, по горизонтальной — целевое значение концентрации

Fig. 3. The CO scatter plot (*а*). The results of concentration predictions are plotted vertically, the true values of concentration are plotted horizontally; the MSE squared deviation boxplot (*б*). On the vertical axis — the deviation on the test data, on the horizontal — the target value of the concentration

Т а б л и ц а 3. Результаты для модели $\mathrm{Pol}_2(R_{\mathrm{CO}}, R_{\mathrm{NM}}, T) + \mathrm{CO}_h$ с L_1-регуляризацией

Table 3. Results for the $\mathrm{Pol}_2(R_{\mathrm{CO}}, R_{\mathrm{NM}}, T) + \mathrm{CO}_h$ model with L_1 regularization

Условие на тест	Размер теста	MAPE, %	MSE, $(\text{мг/м}^3)^2$	GRE, %
Нет	5344	25.5	0.24	20.5
$\mathrm{CO} \leq 0.3$	118	361.5	0.58	88.9
$\mathrm{CO} > 0.3$	5226	17.9	0.23	18.9

истинного значения. Для обучения модели использовались первые 2000 строк данных, а оставшиеся 5344 строки — для тестирования.

Как видно из точечной диаграммы, значительная часть выбросов происходит при низких значениях концентрации CO. Для подтверждения этого факта построена ящичная диаграмма (рис. 3, *б*), которая показывает зависимость квадратичных отклонений MSE от истинных значений концентрации в диапазоне от 0.1 и до 2 мг/м^3. Как можно увидеть из диаграммы, при целевой концентрации от 0.1 и до 0.3 мг/м^3 наблюдаются значительные медианные значения квадратичных отклонений.

Таким образом, полиномиальная модель второй степени $\mathrm{Pol}_2(R_{\mathrm{CO}}, R_{\mathrm{NM}}, T) + \mathrm{CO}_h$ в случае L_1-регуляризации плохо описывает низкие значения целевых концентраций CO. При прогнозировании концентраций от 0.1 до 0.3 мг/м^3 модель склонна завышать значения в несколько раз. При этом на концентрациях газа выше 0.3 мг/м^3 наблюдается уменьшение как абсолютных, так и относительных отклонений. Сводные результаты приведены в табл. 3. Ошибки показаны как на всей тестовой выборке, так и на тестовом множестве с концентрацией CO ниже и выше 0.3 мг/м^3.

3. Нейронные сети прямого распространения

В этом разделе представлены модели прогнозирования концентрации CO на основе нейронных сетей прямого распространения. Были рассмотрены различные методы оптими-

зации, функции потерь и подходы к настройке скорости обучения. Также произведены сравнение различных функций активации нейронов и анализ количества необходимых скрытых слоев и нейронов в них. Рассмотрены различные методы регуляризации, которые могут быть использованы при обучении нейронных сетей.

3.1. Функция потерь

При обучении нейронной сети необходимо выбрать функцию потерь, которая будет являться мерой расхождения между истинными значениями концентрации CO и оценкой, полученной нейронной сетью. Задача оптимизации и обучения направлена на минимизацию этой функции. В качестве базовых функций потерь часто используются такие, как MSE (5) и MAPE (6), а также MAE:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |Y_i - \hat{Y}_i|,$$

где Y_i, $\hat{Y}_i$ — наблюдаемое и предсказанное значения.

Функция MSE хорошо работает в большинстве случаев, однако она может сильно реагировать на выбросы. При ее минимизации оптимизатор стремится как можно лучше описать значения больших выбросов, при этом жертвуя точностью в "хороших" точках. Ошибка MAPE менее требовательна к выбросам, но она не подходит для нашей задачи, так как относительные отклонения принимают экстремальные значения на низких концентрациях. Выбор данной функции приведет к чрезмерной точности для описания выбросов в области низких концентраций, при этом жертвуя точностью на больших значениях. Функция ошибок MAE менее чувствительна к большим абсолютным выбросам, чем MSE, и также не обладает выборочностью в предсказаниях низких концентраций, как MAPE.

Существуют и другие функции потерь для задач регрессии, которые стоит рассмотреть. Средняя квадратичная логарифмическая ошибка

$$\mathrm{MSLE} = \frac{1}{N}\sum_{i=1}^{N} \Big(\log(Y_i + 1) - \log(\hat{Y}_i + 1)\Big)^2$$

схожа с MAPE, однако имеет асимметричность для различных оценок: ее значение на недооценке выше, чем на переоценке. Если выбрать эту функцию потерь, наша модель будет склонна переоценивать значения концентрации.

Логарифм гиперболического косинуса

$$\mathrm{MLChE} = \frac{1}{N}\sum_{i=1}^{N} \log\Big(\cosh(Y_i - \hat{Y}_i)\Big).$$

При малых значениях x функция $\log(\cosh(x)) \approx x^2/2$, а при больших x ведет себя как $|x| - \log(2)$. Данная функция схожа с MSE, но менее чувствительна к существенно неправильным оценкам.

Линейно-экспоненциальная функция потерь

$$\mathrm{LINEX} = \frac{1}{N}\frac{2}{a^2}\sum_{i=1}^{N} \Big(e^{a(\hat{Y}_i - Y_i)} - a(\hat{Y}_i - Y_i) - 1\Big),$$

где a — параметр, отвечающий за асимметрию функции. Это асимметричная функция с гладкой производной. Если $a > 0$, то модель будет недооценивать концентрацию,

Т а б л и ц а 4. Результаты нейронной сети для различных функций потерь (среднее значение ± стандартное отклонение)
Table 4. Neural network results for various loss functions (average value ± standard deviation)

Функция потерь	a	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
MSE	Нет	24.7±0.2	0.27±0.003	20.6±1.1
MAE	Нет	**23.9±0.2**	**0.28±0.007**	**18.2±0.6**
MAPE	Нет	25.1±0.4	0.37±0.007	21.9±1.0
MSLE	Нет	24.6±0.2	0.31±0.008	19.9±0.9
MLChE	Нет	24.5±0.3	0.26±0.005	19.2±0.9
LINEX	0.1	24.6±0.2	0.27±0.004	19.7±1.2
MLEE	1.05	24.4±0.3	0.27±0.007	19.5±0.6

накладывая большую ошибку на превышающие значения, а если $a < 0$, то переоценивать. В случае, если $a > 0$, при малых $|\hat{Y}_i - Y_i|$ фунцкия ведет себя как MSE. При больших же значениях $|\hat{Y}_i - Y_i|$ поведение линейно в случае недооценки концентрации и экспоненциально для переоценки.

Новая гладкая асимметричная функция потерь

$$\text{MLEE} = \frac{1}{N}\sum_{i=1}^{N} \log\left(e^{a(\hat{Y}_i - Y_i)} + be^{-\frac{a}{b}(\hat{Y}_i - Y_i)} - b\right),$$

где $b = a^2/(2-a^2)$, $a \in (1, \sqrt{2})$ — настраиваемый параметр, отвечающий за асимметрию. При значениях a, близких к 1, функция практически не обладает асимметрией, с ростом же параметра a модель склонна к недооценке концентрации. При малых $x = \hat{Y}_i - Y_i$ функция ведет себя как x^2 при большой переоценке MLEE $\approx a|x|$, а при значительной недооценке MLEE $\approx \log(b) + a|x|/b$. Преимуществом данной функции потерь перед LINEX является отсутствие экспоненциального роста ошибки на недостаточных оценках концентрации.

Для сравнения различных функций потерь использована полносвязная нейронная сеть прямого распространения с одним скрытым слоем. В качестве входных значений нейронной сети использовались признаки R_{CO}, R_{NM}, T, CO_h. Количество нейронов в скрытом слое было выбрано 10, гиперболический тангенс использовался как функция активации. Тренировочная выборка состояла из первых 2000 строк данных, а для валидации из нее выделены последние 200 строк. Для тестового набора данных использовались последние 5344 строки. В качестве алгоритма оптимизации выбран ADAM, начальная скорость обучения равна $5{\cdot}10^{-3}$. Количество эпох обучения составляло 1000, а пакет данных (батч) включал 50 строк. Случайное задание начальных значений весов приводило к разбросам значений итоговых метрик. Для борьбы с этим каждая модель обучалась 10 раз, после чего вычислялись средние значения по каждой метрике, а также стандартные отклонения от них. Полученные результаты на тестовом наборе данных приведены в табл. 4. Из них можно сделать вывод, что наилучшей функцией потерь для обучения является MAE, худшие результаты соответствуют функции MAPE.

3.2. Учет дневных изменений концентрации

Для использования информации о суточных изменениях данных рассмотрены новые признаки, учитывающие средние значения сопротивления датчиков и температуры воз-

духа. Усреднение проводилось по предыдущим 24 значениям. Так, например, по значению сопротивления датчика R_{CO} создавались два новых признака — WR_{CO} и DR_{CO}:

$$WR_{\mathrm{CO}}[j] = \frac{1}{24}\sum_{i=0}^{23} R_{\mathrm{CO}}[j-i], \quad DR_{\mathrm{CO}}[j] = R_{\mathrm{CO}}[j] - WR_{\mathrm{CO}}[j].$$

Новые признаки создавались начиная с 24 строки. Первые 23 строки далее не использовались для построения моделей. Аналогичные признаки создавались для датчика неметановых углеводородов R_{NM}, а также температуры воздуха T.

Для исследования пригодности новых признаков использовалась полносвязная нейронная сеть прямого распространения с одним скрытым слоем и функцией потерь MAE. Количество нейронов в скрытом слое было выбрано 10, в качестве функции активации использовался гиперболический тангенс. Тренировочный набор состоял из 1977 строк, для тестового набора данных брались последние 5344 строки. В качестве алгоритма оптимизации выбран ADAM, начальная скорость обучения при этом равна $5{\cdot}10^{-3}$. Количество эпох обучения составляло 1000, размер одного пакета данных (батча) 50 строк. Усреднение проводилось по десяти реализациям обученных нейронных сетей.

Рассмотрены четыре различных варианта входных значений нейронной сети. В первом использовались признаки R_{CO}, R_{NM} и T, во втором добавлялся признак CO_h, в третьем входными признаками были WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT и WT. В четвертом варианте содержался также признак CO_h. Результаты представлены в табл. 5. Из таблицы можно сделать вывод, что добавление нового признака CO_h в нейронные сети приводит к улучшению точности моделей по всем метрикам. Разделение же сопротивлений и температуры на новые признаки приводит к существенному сокращению количества выбросов и уменьшению метрики GRE.

3.3. L_1- и L_2-регуляризация

Для каждой из четырех моделей применена L_1- и L_2-регуляризация на веса, соединяющие входной и скрытый слои. Основная цель использования регуляризации — уменьшение размера вектора весов, что может привести к уменьшению вероятности переобучения нейронной сети и повышению обобщающей способности модели.

Для этого из тренировочной выборки размером 1977 строк выделены последние 377 строк для валидации. Оптимальные гиперпараметры подобраны на валидационной выборке, а обучение проводилось на первых 1600 строках с функцией потерь MAE. Количество нейронов в скрытом слое выбрано равным 10, функция активации $\tanh(x)$. Оптимальное значение гиперпараметра регуляризации подобрано на логарифмической сетке в диапазоне от 10^{-4} до $10^{-0.5}$. После нахождения лучшего параметра обучение

Таблица 5. Результаты нейронных сетей для разных наборов входных признаков (среднее значение ± стандартное отклонение)
Table 5. Results of neural networks for different sets of input features (average value ± standard deviation)

Модель	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
1	25.8±0.3	0.30±0.009	24.6±1.0
2	24.0±0.2	0.28±0.006	18.3±0.4
3	24.7±0.7	0.28±0.011	19.8±0.8
4	**24.0±0.6**	**0.27±0.01**	**17.7±1.4**

проводилось на всей тренировочной выборке размером 1977 строк, а для тестирования использовались последние 5344 строки. Усреднение проводилось по десяти нейронным сетям. Средние значения метрик и стандартные отклонения от них на тестовом наборе данных представлены в табл. 6.

Анализ таблицы показывает, что добавление нового признака CO_h и разделение значений сопротивлений и температур на средние дневные значения и отклонения от них улучшают точность моделей. Кроме того, регуляризация позволяет бороться с переобучением и мультиколлинеарностью, что приводит к уменьшению ошибок при прогнозировании. Регуляризация L_1 показывает улучшение результата для метрик MAPE и GRE, регуляризация L_2 снижает ошибку MSE на тестовых данных.

Установлено, что в нейронных сетях с регуляризацией наблюдается улучшение стабильности обучения нейронных сетей. Выполнен анализ точности моделей в зависимости от значения гиперпараметра для L_1-регуляризации. На рис. 4 приведены результаты для двух моделей: с регуляризацией и без нее. В качестве входных признаков нейронной сети использовались WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. Для обучения использовались 1977 строк данных, для теста — последние 5344 строки. Результаты моделей приведены на тестовом наборе. В качестве алгоритма оптимизации

Т а б л и ц а 6. Результаты нейронных сетей с L_1- и L_2-регуляризацией (среднее значение ± стандартное отклонение)
Table 6. Results of neural networks with L_1- and L_2-regularization (average value ± standard deviation)

Модель	L_1-регуляризация			L_2-регуляризация		
	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
1	25.1±0.2	0.29±0.001	20.9±0.3	25.5±0.1	0.28±0.008	22.0±0.6
2	24.0±0.2	0.27±0.002	16.8±0.5	24.1±0.2	0.27±0.003	16.9±0.4
3	23.9±0.8	0.29±0.011	16.9±1.7	24.2±0.3	0.27±0.006	17.2±0.7
4	**23.4±0.1**	**0.28±0.002**	**15.5±0.2**	**23.7±0.1**	**0.26±0.005**	**15.8±0.4**

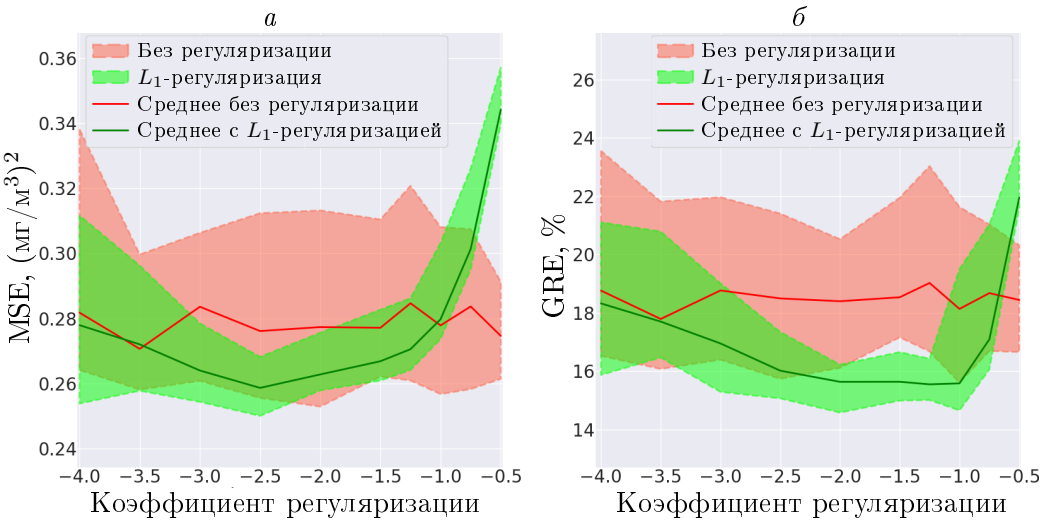

Рис. 4. Результаты нейронных сетей с регуляризацией и без нее; по горизонтали значение показателя степени с основанием 10 в регуляризации: *а* — ошибка MSE; *б* — ошибка GRE
Fig. 4. Results of neural networks with and without regularization; on the horizontal axis is the value of the exponent with base 10 in the regularization: *а* — MSE error; *б* — GRE error

выбран ADAM, начальная скорость обучения $5{\cdot}10^{-3}$, количество эпох 1000, размер батча 50. Усреднение проводилось по десяти нейронным сетям. Кроме средних значений указаны также границы, в которых изменялись данные метрики для каждого обучения. Поскольку модель без регуляризации не зависит от параметра, на графике видны незначительные изменения для средних значений на метриках GRE и MSE. Минимальные и максимальные же значения метрик для модели без регуляризации от точки к точке могут значительно изменяться, что обусловливается случайным заданием начальных весов. Границы закрашенных областей соответствуют максимальным и минимальным ошибкам на этих метриках. Видно, что при использовании регуляризации уменьшается разброс между минимальными и максимальными значениями метрик, что указывает на улучшение стабильности обучения.

3.4. Настройка гиперпараметров и архитектуры нейронной сети

Для алгоритма оптимизации ADAM скорость обучения α, экспоненциальные скорости затухания для первого и второго моментов β_1 и β_2, а также константа ε для численной стабильности являются гиперпараметрами, требующими подбора оптимальных значений [11]. Для настройки этих параметров использован метод поиска по сетке GridSearchCV на нейронной сети с входными признаками WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT и CO_h. Нейронная сеть содержала один скрытый слой с десятью нейронами. Были проанализированы функции активации $\tanh(x)$, $\mathrm{sigmoid}(x)$, $\mathrm{relu}(x)$ и $\mathrm{exponential}(x)$. Оптимальные значения гиперпараметров подбирались для первых 3500 строк данных, а также для всех данных. Количество эпох составляло 100, 500 и 1000, а параметр кроссвалидации KFold был равен 10. Для настройки использовалась метрика MAPE. В результате найдены оптимальные значения гиперпараметров и установлено, что наилучшей активационной функцией для данной задачи является $\tanh(x)$.

После настройки гиперпараметров и выбора функции активации построены кривые тренировки для нейронной сети из десяти нейронов с одним скрытым слоем. Тренировочная выборка состояла из первых 1200 строк данных, для валидационной выборки использовались следующие 777 строк. Кривые тренировки изображены на рис. 5. Как

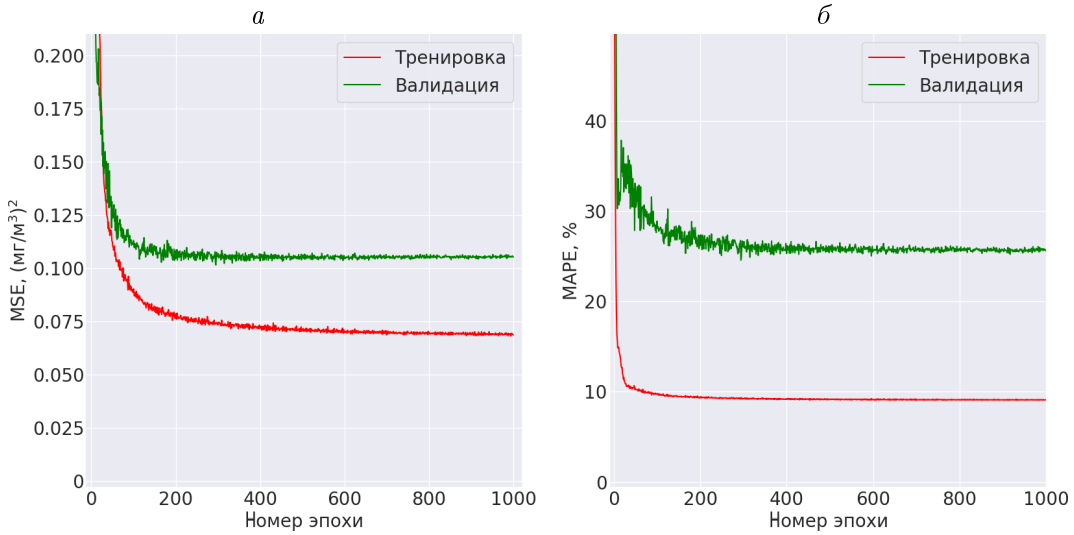

Рис. 5. Кривые тренировки для нейронной сети с L_1-регуляризацией: *а* — MSE; *б* — MAPE
Fig. 5. Training curves for a neural network with L_1 regularization: *а* — MSE; *б* — MAPE

Т а б л и ц а 7. Результаты нейронных сетей с L_1-регуляризацией при различных архитектурах (среднее значение ± стандартное отклонение)
Table 7. Results of neural networks with L_1-regularization for various architectures (average value ± standard deviation)

Число нейронов в слое	1 скрытый слой			2 скрытых слоя		
	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %	MAPE, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
3	22.95±0.02	0.246±0.001	14.6±0.1	22.87±0.04	0.264±0.002	14.3±0.1
7	23.0±0.1	0.247±0.002	14.7±0.2	22.89±0.06	0.265±0.003	14.3±0.1
10	23.0±0.2	0.240±0.002	14.6±0.3	22.91±0.03	0.265±0.002	14.3±0.1
13	23.0±0.2	0.247±0.003	14.6±0.2	22.89±0.06	0.266±0.003	14.3±0.1

видно из графика, достаточное количество эпох обучения для нейронной сети с одним скрытым слоем составляет порядка 600.

Для исследования различных архитектур нейронных сетей применен метод поиска по сетке GridSearchCV, с помощью которого найдено оптимальное количество слоев и нейронов в скрытых слоях. Нейронные сети обучались на тренировочной выборке из первых 1977 строк, а тестовый набор данных состоял из последних 5344 строк. Входными признаками нейронной сети служили WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. В качестве функции активации использовался гиперболический тангенс, а количество эпох было равно 1000. Усреднение результатов проводилось по десяти нейронным сетям. Анализ результатов проведен для нейронных сетей с количеством скрытых слоев от одного до пяти и количеством нейронов в каждом слое от трех до 15. Поиск наилучшей архитектуры выполнен для всех данных, в табл. 7 приведены результаты на тестовом наборе данных. В результате обнаружено, что оптимальное количество скрытых слоев составляет 2, а наилучшее количество нейронов в скрытом слое находится в диапазоне от трех до семи.

3.5. Кривые обучения

Построены кривые обучения для архитектуры нейронной сети с двумя скрытыми слоями по 5 нейронов в каждом и активационной функцией гиперболический тангенс. Входными признаками нейронной сети служили WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. Для построения кривых обучения использовались данные с конца июня 2004 г. по апрель 2005 г. — последние 5344 строки данных. Тренировочная выборка изменялась в диапазоне от одного до 80 дней и состояла из данных с середины марта по конец июня 2004 г. Для того чтобы наилучшим образом моделировать калибровку газовых сенсоров, данные выбирались в хронологическом порядке, т. е. сразу после окончания тренировочных данных следовали тестовые. На рис. 6 представлены кривые обучения для ошибок MSE и MAPE.

Как показано на рис. 6, при малом объеме обучающей выборки (от одного до нескольких дней) результаты нейронных сетей на тестовом наборе демонстрируют большие значения MSE и MAPE. При увеличении размера обучающей выборки до 35 суток значения почти не изменяются, тогда как при дальнейшем ее увеличении наблюдается резкое уменьшение их значений с выходом на постоянное значение в области 45 дней. Из этого можно сделать вывод, что для эффективного обучения модели и стабильного

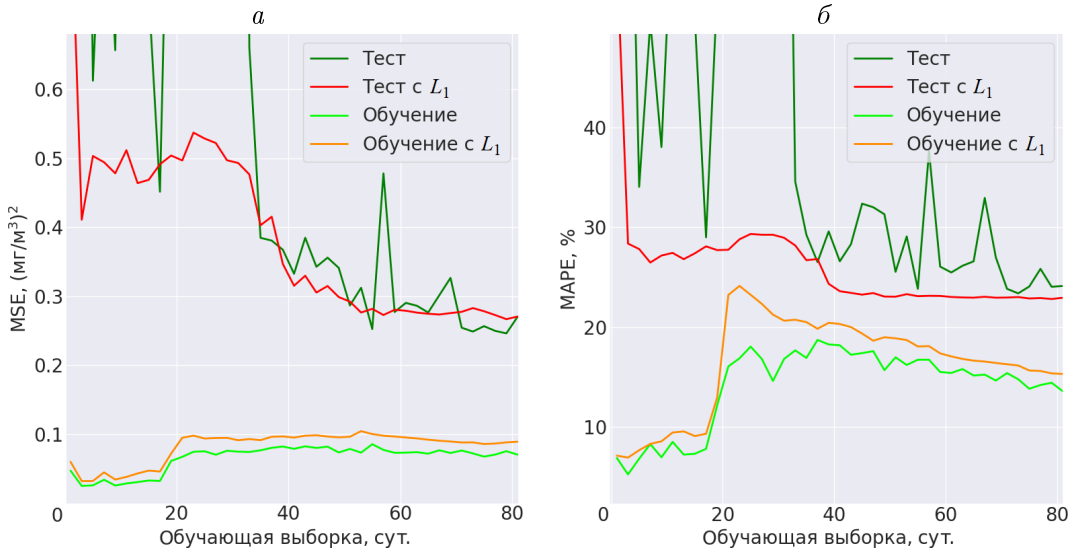

Рис. 6. Кривые обучения для нейронной сети с двумя скрытыми слоями с L_1-регуляризацией и без нее: *а* — MSE; *б* — MAPE. По горизонтальной оси отложен размер тренировочной выборки в сутках

Fig. 6. Learning curves for a neural network with two hidden layers with and witnout L_1-regularization: *а* — MSE; *б* — MAPE. The size of the training sample in days is plotted along the horizontal axis

прогнозирования концентраций угарного газа на использованном наборе данных достаточно 40 дней непрерывных измерений. Отметим, что использование L_1-регуляризации в нейронной сети демонстрирует более стабильное обучение, чем без нее.

Заключение

Исследована возможность применения методов машинного обучения для восстановления концентрации угарного газа по выходным данным полупроводниковых газовых сенсоров. Выполнены анализ структуры и корреляционный анализ выходных данных набора датчиков, на основе результатов которых созданы новые признаки, в частности CO_h, учитывающий зависимость концентрации газа CO от времени суток. С использованием имеющихся и созданных признаков построены различные модели множественной линейной и полиномиальной регрессии, а также несколько простых архитектур нейронных сетей с прямой связью для восстановления реальных значений концентраций угарного газа CO по данным датчиков. На этих моделях анализировалось влияние различных способов регуляризации на точность их обучения. Наилучший результат среди моделей на основе полиномиальной регрессии показала модель $\mathrm{Pol}_2(R_{\mathrm{CO}}, R_{\mathrm{NM}}, T)$ с дополнительным признаком CO_h и L_1-регуляризацией. К недостаткам использования нового признака CO_h можно отнести снижение точности прогноза концентрации при нестандартных внешних условиях, таких, например, как перекрытие улицы для проезда транспорта и снижение уровня CO.

Из анализа графика на рис. 3 следует, что основной вклад в погрешность восстановления концентрации целевого газа вносят данные, соответствующие низким значениям реальной концентрации CO. Так, разброс квадратичных отклонений на концентрациях меньше 0.4 мг/м^3 значительно больше, чем на более высоких концентрациях целевого газа. Это можно объяснить низкой дискретностью референсного анализатора, которая

в анализируемых данных составляла 0.1 мг/м^3. Кроме того, значительные расхождения в предложенных моделях в области пониженных концентраций могут быть связаны с большой погрешностью измерений эталонного анализатора. Погрешность измерения в указанном диапазоне концентраций составляет порядка 0.3 мг/м^3.

Выполнен анализ применения нейронных сетей прямого распространения для прогнозирования концентрации CO. Установлено, что наилучшей функцией потерь для обучения является абсолютная ошибка MAE, а лучшей активационной функцией нейрона — гиперболический тангенс $\tanh(x)$. Создание новых признаков из исходных с помощью дневного усреднения значений сопротивлений датчиков и температуры, а также отклонений от средних значений за день позволило уменьшить значения метрик MAPE и GRE и сократить набор тренировочных данных, необходимых для достижения минимальной погрешности восстановления концентрации газа. Наилучшая архитектура нейронной сети состояла из двух скрытых слоев по пять нейронов в каждом слое с L_1-регуляризацией ядра скрытого слоя. Подбор оптимального гиперпараметра регуляризации, а также настройка оптимизатора ADAM с помощью метода GridSearchCV позволили улучшить точность метрики MAPE на тестовых данных до значений 22.9±0.1 %, а значение GRE — до 14.3±0.1 %.

Полученные результаты возможно сравнить с результатами, представленными в статье [2], в которой обучение производилось на том же наборе даных, что и в данной работе. В качестве модели для восстановления концентрации газа CO использовалась нейронная сеть с активационной функцией $\tanh(x)$. При обучении на выборке из 2000 строк данных ошибка MAPE составила 26 %.

Благодарности. Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (проект № FSUS-2021-0015).

Список литературы

[1] **Castell N., Dauge F.R., Schneider P., Vogt M., Lerner U., Fishbain B., Broday D., Bartonova A.** Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International. 2017; (99):293–302. DOI:10.1016/j.envint.2016.12.007. Available at: `https://www.sciencedirect.com/science/article/pii/S0160412016309989`.

[2] **De Vito S., Piga M., Martinotto L., Di Francia G.** CO, NO_2 and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. Sensors and Actuators B: Chemical. 2009; (143):182–191. DOI:10.1016/j.snb.2009.08.041. Available at: `https://www.sciencedirect.com/science/article/abs/pii/S092540050900673X`.

[3] **Spinelle L., Gerboles M., Villani G., Aleixandre M., Bonavitacola F.** Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensors and Actuators B: Chemical. 2015; (215):249–257. DOI:10.1016/j.snb.2015.03.031. Available at: `https://www.sciencedirect.com/science/article/pii/S092540051500355X`.

[4] **Spinelle L., Gerboles M., Villani G., Aleixandre M., Bonavitacola F.** Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2. Sensors and Actuators B: Chemical. 2017; (238):706–715. DOI:10.1016/j.snb.2016.07.036. Available at: `https://www.sciencedirect.com/science/article/pii/S092540051631070X`.

[5] **De Vito S., Fattoruso G., Pardo M., Tortorella F., Di Francia G.** Semisupervised learning techniques in artificial olfaction: a novel approach to classification problems and drift counteraction. IEEE Sensors Journal. 2012; 12(11):3215–3224. DOI:10.1109/JSEN.2012.2192425. Available at: https://ieeexplore.ieee.org/document/6176193.

[6] **Esposito E., De Vito S., Salvato M., Bright V., Jones R.L., Popoola O.** Dynamic neural network architectures for on field stochastic calibration of indicative lowcost air quality sensing systems. Sensors and Actuators B: Chemical. 2016; (231):701–713. DOI:10.1016/j.snb.2016.03.038. Available at: https://www.sciencedirect.com/science/article/pii/S092540051630332X.

[7] **Lai W.I., Chen Y.Y., Sun J.H.** Ensemble machine learning model for accurate air pollution detection using commercial gas sensors. Sensors. 2022; (22):4393. DOI:10.3390/s22124393. Available at: https://www.researchgate.net/publication/361260797_Ensemble_Machine_Learning_Model_for_Accurate_Air_Pollution_Detection_Using_Commercial_Gas_Sensors.

[8] **Smith K.R., Edwards P.M., Ivatt P.D., Lee J.D., Squires F., Dai C., Peltier R.E., Evans M.J., Sun J., Lewis A.C.** An improved low-power measurement of ambient NO_2 and O_3 combining electrochemical sensor clusters and machine learning. Atmospheric Measurement Techniques. 2019; (12):1325–1336. DOI:10.5194/amt-12-1325-2019. Available at: https://amt.copernicus.org/articles/12/1325/2019.

[9] **Cid P.F.** Calibration of low-cost air pollutant sensors using machine learning techniques. Universitat Politecnica de Catalunya. 2019; Available at: https://upcommons.upc.edu/bitstream/handle/2117/168918/143249.pdf?sequence=1.

[10] **Mohammadshirazi A., Kalkhorani V.A., Humes J., Speno B., Rike J., Ramnath R., Clark J.D.** Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: a case study. Building and Environment. 2022; (213). DOI:10.1016/j.buildenv.2022.108833. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0360132322000816.

[11] **Kingma D.P., Ba J.L.** Adam: a method for stochastic optimization. Conference Paper at the 3rd International Conference for Learning Representations, San Diego. 2015; DOI:10.48550/arXiv.1412.6980. Available at: https://arxiv.org/abs/1412.6980.

Вычислительные технологии, 2024, том 29, № 4, с. 4–23. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 4, pp. 4–23. © FRC ICT, 2024 eISSN 2313-691X

MATHEMATICAL MODELLING

DOI:10.25743/ICT.2024.29.4.002

Interpretation models for data of metal-oxide gas sensors based on machine learning methods

A. D. Kozmin*, A. A. Redyuk

Novosibirsk State University, 630090, Novosibirsk, Russia
*Corresponding author: Artem D. Kozmin, e-mail: a.kozmin@g.nsu.ru

Received May 03, 2023, revised August 16, 2023, accepted October 03, 2023.

Abstract

The study examines the application of machine learning methods for determining the concentration of carbon monoxide (CO) in the air based on data from metal-oxide (MOX) gas sensors. High

levels of concentration are hazardous for human and animal health, making air quality control critically important. The output data from the sensors were investigated, and new features were created to account for the daily temporal variation of gas concentration's. Multiple linear and polynomial regression models, as well as neural networks, were developed to predict CO concentration. The impact of regularization on the accuracy of gas sensor data interpretation was also explored. The analysis revealed that the primary source of error in CO concentration recovery was the data with low concentration values. Creating new features through daily averaging of resistance sensor values and temperature, as well as deviations from the mean values for the day, improved the results of the MAPE and GRE metrics. It was found that the best loss function for training neural networks is the absolute error (MAE), and the best activation function for a neuron is the hyperbolic tangent function ($\tanh(x)$). The study demonstrates the potential use of machine learning methods for air quality control.

Keywords: MOX gas sensor, carbon monoxide, fully connected neural network, regularization, linear regression, polynomial regression.

Citation: Kozmin A.D., Redyuk A.A. Interpretation models for data of metal-oxide gas sensors based on machine learning methods. Computational Technologies. 2024; 29(4):4–23. DOI:10.25743/ICT.2024.29.4.002. (In Russ.)

Acknowledgements. This research was funded by the Ministry of Science and Higher Education of the Russian Federation (Project No. FSUS-2021-0015).

References

1. **Castell N., Dauge F.R., Schneider P., Vogt M., Lerner U., Fishbain B., Broday D., Bartonova A.** Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International. 2017; (99):293–302. DOI:10.1016/j.envint.2016.12.007. Available at: https://www.sciencedirect.com/science/article/pii/S0160412016309989.
2. **De Vito S., Piga M., Martinotto L., Di Francia G.** CO, NO_2 and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. Sensors and Actuators B: Chemical. 2009; (143):182–191. DOI:10.1016/j.snb.2009.08.041. Available at: https://www.sciencedirect.com/science/article/abs/pii/S092540050900673X.
3. **Spinelle L., Gerboles M., Villani G., Aleixandre M., Bonavitacola F.** Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensors and Actuators B: Chemical. 2015; (215):249–257. DOI:10.1016/j.snb.2015.03.031. Available at: https://www.sciencedirect.com/science/article/pii/S092540051500355X.
4. **Spinelle L., Gerboles M., Villani G., Aleixandre M., Bonavitacola F.** Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2. Sensors and Actuators B: Chemical. 2017; (238):706–715. DOI:10.1016/j.snb.2016.07.036. Available at: https://www.sciencedirect.com/science/article/pii/S092540051631070X.
5. **De Vito S., Fattoruso G., Pardo M., Tortorella F., Di Francia G.** Semisupervised learning techniques in artificial olfaction: a novel approach to classification problems and drift counteraction. IEEE Sensors Journal. 2012; 12(11):3215–3224. DOI:10.1109/JSEN.2012.2192425. Available at: https://ieeexplore.ieee.org/document/6176193.
6. **Esposito E., De Vito S., Salvato M., Bright V., Jones R.L., Popoola O.** Dynamic neural network architectures for on field stochastic calibration of indicative lowcost air quality sensing systems. Sensors and Actuators B: Chemical. 2016; (231):701–713. DOI:10.1016/j.snb.2016.03.038. Available at: https://www.sciencedirect.com/science/article/pii/S092540051630332X.
7. **Lai W.I., Chen Y.Y., Sun J.H.** Ensemble machine learning model for accurate air pollution detection using commercial gas sensors. Sensors. 2022; (22):4393. DOI:10.3390/s22124393. Available at: https://www.researchgate.net/publication/361260797_Ensemble_Machine_Learning_Model_for_Accurate_Air_Pollution_Detection_Using_Commercial_Gas_Sensors.
8. **Smith K.R., Edwards P.M., Ivatt P.D., Lee J.D., Squires F., Dai C., Peltier R.E., Evans M.J., Sun J., Lewis A.C.** An improved low-power measurement of ambient NO_2 and O_3 combining electrochemical sensor clusters and machine learning. Atmospheric Measurement Techniqu-

es. 2019; (12):1325–1336. DOI:10.5194/amt-12-1325-2019. Available at: https://amt.copernicus.org/articles/12/1325/2019.

9. **Cid P.F.** Calibration of low-cost air pollutant sensors using machine learning techniques. Universitat Politecnica de Catalunya. 2019; Available at: https://upcommons.upc.edu/bitstream/handle/2117/168918/143249.pdf?sequence=1.
10. **Mohammadshirazi A., Kalkhorani V.A., Humes J., Speno B., Rike J., Ramnath R., Clark J.D.** Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: a case study. Building and Environment. 2022; (213). DOI:10.1016/j.buildenv.2022.108833. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0360132322000816.
11. **Kingma D.P., Ba J.L.** Adam: a method for stochastic optimization. Conference Paper at the 3rd International Conference for Learning Representations, San Diego. 2015; DOI:10.48550/arXiv.1412.6980. Available at: https://arxiv.org/abs/1412.6980.