ScienceDirect
www.sciencedirect.com

Revista Iberoamericana de Automatica e Informática industrial 00 (2022) 1–12

R I A I
www.elsevier.es/RIAI

Применение методов машинного обучения с использованием суточной зависимости в калибровке газового датчика

Козьмин А.Д.[a,1,*], Редюк А.А.[b]

[a]Comita Espaaol de Automatica, Parc Tecnologic de Barcelona, Edifici U, C/ Llorens i Artigas, 4-6, 08028 Barcelona, Espaaa.
[b]Departamento de Automatica, Ingenieraa Electranica e Informatica, Universidad Politacnica de Madrid, C/ Josa Gutiarrez Abascal, na2, 28006, Madrid, Espaaa.

Resumen

Estas instrucciones constituyen una guaa para la preparacian de Copyright © XXXX CEA. Publicado por Elsevier Espaaa, S.L. Todos los derechos reservados.

Palabras Clave:
palabra 1, palabra 2, 5-10 palabras clave (tomadas de la lista del sitio web de IFAC).

1. Введение

Оксид углерода CO – бесцветный токсичный газ не имеющий запаха и вкуса. Он образуется в результате неполного сгорания углеродных соединений в условиях недостатка кислорода. CO может связываться с гемоглобином Hb в крови образовывая карбоксигемоглобин $HbCO$. Избыток $HbCO$ в крови может привести к кислородному голоданию и смерти. По имеющимся данным ежегодные отравления угарным газом в США получают 50 000 человек (Rose JJ et al., 2017). Всемирная организация здравоохранения считает, что уровни более 7 мг/м^3 потенциально токсичны в течение длительного периода времени. Известно, что при концентрациях свыше 14 мг/м^3 возрастает вероятность смерти от инфаркта миокарда. Согласно же санитарным правилам и нормам действующим на территории РФ (СанПиН, 2021), предельно допустимая среднесуточная концентрация CO в воздухе населённых мест ПДК$_{сс}$ = 3 мг/м^3, а максимальная разовая концетрация ПДК$_{мр}$ = 5 мг/м^3.

Точная оценка концентрации угарного газа крайне важна для контроля качества воздуха. Мониторинг загрязнения городов в конце 20-го века осуществлялся на базе промышленных спектрометров. Однако их размеры и стоимость делали неосуществимым плотное определение фактической концентрации в городе. В начале 21-го века стали популярными газовые сенсорные устройства. Это произошло из-за их более низкой стоимости, однако известные проблемы долговременной стабильности и дрейфа серьёзно ограничивали их надёжность и точность.

Основная функция газовых сенсоров это преобразование концентрации анализируемого вещества в электрический или другой сигнал. Калибровка газовых сенсоров затрудняется неявной зависимостью между показаниями датчика (ток или сопротивление) и целевой концентрацией. Кроме этого, газовые сенсоры подвержены влиянию условий окружающей среды. Также существует проблема неизбирательности датчиков, когда датчик регистрирует не только целевой газ, но и другие примеси. Например, на датчик угарного газа могут оказывать влияние неметановые углеводороды $NMHC$.

Как известно, для калибровки газового сенсора требуется использование устройства на базе точного спектрометра в полевых испытаниях. Так например в работе (Nuria Castell et al., 2017) провели оценку работы 24 идентичних комерческих сенсорных платформ AQMesh по мониторингу газов NO, NO_2, CO и O_3. Результаты показали, что отклик каждого датчика уникален. Кроме этого, было показано, что лабораторная калибровка не может скорректировать работу датчиков в реальных условиях. Это означает, что требуется проводить калибровку в полевых условиях для каждого датчика индивидуально.

*Autor en correspondencia.
Correos electrónicos: autor@cea-ifac.es (Козьмин А.Д.), autor2@cea-ifac.es (Редюк А.А.)
URL: a.kozmin@g.nsu.ru (Козьмин А.Д.), www2.cea-ifac.es (Редюк А.А.)
[1]Nota al pie para el autor 1

Для калибровки газовых сенсоров применялось большое количество различных алгоритмов. Ранее в работах были представлены результаты прогнозирования концентраций угарного газа с использованием нейронных сетей с прямой связью FFNN (S. De Vito et al., 2009). В данной работе при использовании гиперболического тангенса в качестве актцивационной функции нейрона удалось достичь относительной точности прогнозирования CO в 26% при тренировочной выборке порядка 2000 часов. Также был произведён анализ производительности в зависимости от размеров тренировочной выборки. Достаточный размер тренировочного набора при этом составил порядка 2 недель при относительной точности прогнозирования CO в 27%. Позже в работах (Spinelle L. et al., 2015, 2017) было показано, что полевая калибровка с использованием методов обучения с учителем более эффективна чем методы линейной LR и полилинейной MLR регрессии.

Исследовались также методы использующие неразмеченные данные при наличии небольшого количества размеченных данных. Так например применение методов обучения с частичным привлечением учителя SSL в работе (S. De Vito et al., 2012) привело к улучшению производительности при длительной непрерывной работе газового сенсора.

Следующим этапом развития калибровки газовых сенсоров было применение рекурентных нейронных сетей RNN. Так например, в работе (E. Esposito et al., 2016) рассматривалось два подхода динамических нейронных сетей TDNN и NARX, результаты сравнивались с нейронной сетью с прямой связью FFNN. Было показано, что динамические нейронные сети имеют лучшие результаты по сравнению FFNN. Развитием этого подхода послужили ансмблевымые методы. В работе (Wei-In Lai et al., 2022) исследуются ансамблевые модели рекуррентных нейронных сетей RNN для мониторинга концентраций CO, O_3, NO_2. Из результатов следует, что интеграция 4 типов моделей (LSTM, GRU, Bi-LSTM, Bi-GRU) показывает лучший результат, чем любая отдельная RNN.

Также в последнее время популярен класстерный подход в сенсорах для решения проблем дрейфа и воспроизводимости конкретного датчика. В работе (K.R. Smith et al, 2019) исследовался кластерный подход с использованием медианного сигнала от 6 аналогичных датчиков. Для анализа концентраций NO_2, O_x были выбраны 4 различных метода (MLR, BRT, BLR, GP). Кластеризация датчиков позволила бороться с проблемой дрейфа, а также решила вопрос невоспроизводимости показаний отдельных датчиков. В другой работе (Pau Ferrer Cid, 2019) исследовались 4 метода машинного обучения (MLR, KNN, RF, SVR) для калибровки датчиков O_3 на основе оксидов металлов. Наилучшим методом калибровки оказалась векторная регрессия SVR. Также анализировалось слияние датчиков для различных моделей. Было показано, что использование от 4 до 6 датчиков в методе SVR значительно улучшает среднеквадратичную ошибку (более 3 мг/м^3).

Также стоит отметить работу (Ahmad Mohammadshirazi et al., 2022), где рассматриваются различные методы прогнозирования концентраций примесей внутри помещений. Для этого авторы использовали четыре различных метода (Rolling Average, RF, Gradient Boosting, LSTM). Авторы подчеркивают сильную суточную зависимость концентраций от часа. В ходе иссследований оказалось, что использование времени суток в часах, а также дня недели значительно помогало в прогнозировании концентрации загрязняющих веществ внутри помещения для любого из рассмотреннных алгоритмов.

В соответсвии с этим, было решено на относительно простых алгоритмах MLR и FFNN исследовать создание новых признаков, учитывающих зависимость концентрации от часа. После чего были построены различные регрессионные модели для восстановления истинных значений концентраций угарного газа CO по данным, использующимся в статье (S. De Vito et al., 2009). Были построены модели линейной и полиномиальной регрессии, а также различные нейронные сети с прямой связью. Кроме этого анализировалось влияние различных способов регуляризаций на калибровку газовых сенсоров.

2. Теоретическая часть и анализ данных

В данном разделе рассматривается общая постановка задачи регрессии. Также обсуждаются основные метрики (меры качества) в задачах регрессионного анализа. Далее приводится описание исходного набора данных и проводится первичный корреляционный анализ.

2.1. Регрессионная модель

Регрессионные анализ представляет собой набор статистических процессов для оценки взаимосвязей между зависимой переменной (меткой) и одной или несколькими независимыми переменными. Регрессионная модель это функция $f(X_i, \beta)$ независимой переменной X_i и неизвестных параметров β, где i – строка данных. Предполагается, что зависимая переменная Y_i есть сумма 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 наиболее точно будет соответсвовать данным. Например для случая многомерной линейной регрессии MLR функция (1) предполагается равной:

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

где p – количество независимых переменных (наблюдаемых признаков), а $X_{0i} = 1$ для каждой i строки данных. Для нахождения оптимальных параметров модели $\hat{\beta}$ часто используется метод наименьших квадратов, целью которого является минимизация суммы квадратов отклонений(3):

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

Для случая многомерной линейной регрессии решение имеет вид:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad (4)$$

где X – матрица признаков, а Y – вектор зависимой величины.

Для случая полиномильной регрессии происходит добавление новых признаков в функцию f, например для случая двухмерной полиномиальной регрессии второй степени функция имеет вид:

$$f_2(X_i, \beta) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i} X_{1i} + \beta_5 X_{2i}^2. \quad (5)$$

2.2. Исходные данные

Для построения моделей были взяты данные собраные с помощью мультисенсорного устройства, разработанного Pirelli Labs. Измерения были проведены на одной из главных улиц в центре итальянского города с интенсивным автомобильным движением с марта 2004 года по апрель 2005 года.

Набор данных содержит 9358 строк усреднённых почасовых ответов 5 различных химических датчиков совместно с целевыми значениями газов. Датчик CO был изготовлен из оксида олова, датчик неметановых углеводородов $NMHC$ из оксида титана, общих оксидов азота NO_x из оксида вольфрама, датчик NO_2 из оксида вольфрама, O_3 из оксида индия. Кроме этого проводились измерения температуры воздуха T и его относительной влажности RH. Целевые значения концентраций газов были получены эталонным анализатором.

После отбраковки пустых данных осталось 7344 строк, состоящих из измеряемых сопротивлений датчиков, температуры и влажности воздуха, а также целевой концентрации CO. Анализатор позволял измерять концентрацию целевого газа в пределах от 0.1 и до 12.0 мг/м^3 с дискретностью в 0.1 мг/м^3. К важным особенностям данных можно отнести высокую корреляцию между сопротивлениями различных датчиков.

2.3. Корреляционный анализ

Для обнаружения взаимных влияний различных газов на датчики использовалась корреляция Пирсона, построенная с помощью библиотек pandas и seaborn. Коэффициент корреляции r – Пирсона характеризует существование линейной зависимости между двумя величинами X_1, X_2 и рассчитывается по формуле:

$$r_{X_1 X_2} = \frac{\mathbf{cov}_{X_1 X_2}}{\sigma_{X_1} \sigma_{X_2}}, \quad (6)$$

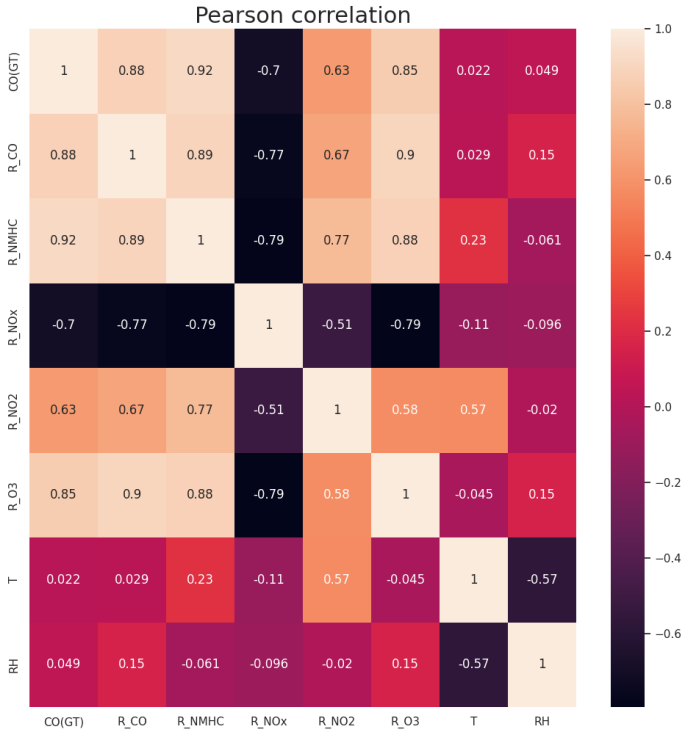

Рис. 1: Корреляционная матрица Пирсона. $CO(GT)$ – целевая концентрация; $R_{CO}, R_{NMHC}, R_{NOx}, R_{NO2}, R_{O3}$ – сопротивления соответствующих датчиков; T, RH– температура и влажность воздуха.

где **cov** – ковариация, а σ – среднеквадратичное отклонение. Значение коэффициентов лежит в диапазоне от $[-1, 1]$ 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 рисунке 1.

Значение коэффициента корреляции между сопротивлением датчика R_{CO} и сопротивлением R_{NMHC} составляет $r = 0.89$, что говорит о наличии сильной линейной зависимости. С одной стороны это может быть обусловлено значительной зависимостью между примесями неметановых углеводородов и угарного газа в самой атмосфере. С другой стороны это может быть вызвано неизбирательностью самих датчиков. Коэффициент кореляции между сопротивлением датчика R_{NMHC} и целевой концентрации $CO(GT)$ равен 0.92, что больше чем между сопротивлением датчика угарного газа R_{CO} и целевой концентрацией 0.88. Как отмечалось в работе (S. De Vito et al., 2009), дальнейший учёт сопротивления датчика R_{NMHC} позволит улучшить прогнозирование концентрации угарного газа.

2.4. Создание новых признаков

На следующем этапе была поставлена задача создания новых признаков для улучшения качества моделей. Была проанализирована зависимость концентрации CO

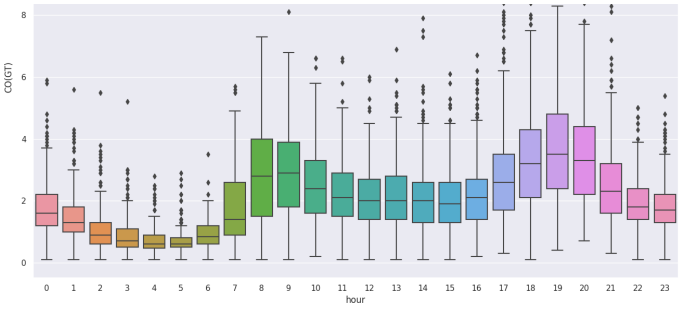

Рис. 2: Ящичные диаграммы. По вертикальной оси концентрация CO в мг/м^3, по горизонтальной оси час наблюдения. Нижняя и верхняя граница ящика - первый и третий квартили, линия в середине ящика - медиана. Концы усов соответсвуют краям статистически значимой выборки, данные выходящие за границы усов (выбросы) обозначены точками.

от часа измерения. Для этого были построены ящичные диаграммы для исходного набора данных. На рисунке 2 представлены ящичные диаграммы зависимости целевой концентрации от часа измерения.

Исходя из диаграммы можно заключить, что существует зависимость между часом в которое происходит измерение и концентрацией угарного газа. Так например наибольшие медианные значения концентрации CO наблюдаются утром с 8 до 9 часов, а также в вечернее время с 18 до 20 часов. В эти времена суток медианные значения составляют около 3 мг/м^3. В дневное время происходит небольшой спад медианных значений концентраций угарного газа до уровеня в 2 мг/м^3. В ночное же время происходит значительный спад примесей CO. Минимальное количество токсичных газов наблюдается с 4 до 5 часов, медианная концентрация около 0.5 мг/м^3. Значительные изменение медианных значений концентраций угарного газа в течении суток может быть вызвано временем наибольшей активности людей. Максимальные медианные значения концентрации совпадают с час пиком движения транспорта.

Таким образом, было решено помимо исходных значений сопротивлений датчиков, температуры и влажности воздуха учитывать также час измерения концентрации. Для этого были вычеслены средние значения концентраций для каждого часа CO_h, усреднение проводилось по всем данным. В дальнейшем этот новый признак добавлялся в рассматриваемых моделях, что привело к повышению качества прогноза.

2.5. Метрики

Обязательным этапом в построении регрессионных моделей является выбор метрик. Метрика необходима для оценки качества построенных моделей. Одна из наиболее часто встречающихся в регрессионном анализе метрик это среднеквадратичная ошибка MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2, \tag{7}$$

где N – количество прогнозов, а $Y_i, \hat{Y}_i$ – наблюдаемое и предсказанное значение концентрации. Также для анализа используют среднюю абсолютную процентную ошибку MAPE:

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{|Y_i|}. \tag{8}$$

Для нашей задачи была введена метрика GRE, которая фактически указывала на процент неправильных прогнозов. Диапазон предсказанной концентрации должен был лежать в 25% ворота от целевого значения. В этом случае предсказание считалось верным.

3. Полиномиальная регрессия

В данном разделе рассматриваются различные модели линейной и полиномиальной регрессии. Применяются методы L_1, L_2 регуляризации для решения проблем мультиколлинеарности и переобучения. Кроме этого на основе L_1 регуляризации проводится отбор признаков для полиномиальной регрессии. После чего анализируются кривые обучения для различных линейных и полиномиальных моделей. Также исследуется особенности ложно предсказанных значений концентраций CO.

3.1. Регуляризация

Для линейных и полиномиальных моделей, исходя из критерия качества (3), минимизацией суммы квадратов отколнений находится решение для неизвестных параметров $\hat{\beta}$, которое имеет вид (4). Однако, если матрица X^TX - вырождена или регулярна, то возникают проблемы в обращении этой матрицы. В случае же мультиколлинеарности (когда два и более признака сильно коррелированы) появляются экстремальные собственные значения в обратной матрице $(X^TX)^{-1}$. Для решения этих проблем требуется регуляризация. Есть два основных типа регуляризации, это L_1 и L_2.

Регуляризация Тихонова или L_2, происходит добавлением нового члена к критерию качества, а именно:

$$Q_{L2} = \sum_i (Y_i - f(X_i,\beta))^2 + |\Gamma\beta|^2 \to \min_\beta. \tag{9}$$

где $\Gamma = \lambda E$, λ – неотрицательный гиперпараметр, а E – единичная матрица. Дифференциируя по β новое решение β^* имеет вид:

$$\beta^* = (X^TX + \lambda^2 E)^{-1}X^TY. \tag{10}$$

Данное решение уменьшает дисперсию, но становится смещённым. В линейных моделях регуляризация Тихонова позволяет избежать проблем мультиколлинеарности и переобучения.

Регуляризация через манхэттенское расстояние или L_1, происходит добавлением нового члена к критерию качества:

$$Q_{L1} = \sum_i (Y_i - f(X_i,\beta))^2 + \lambda|\beta| \to \min_\beta, \tag{11}$$

где λ – неотрицательный гиперпараметр. Данная регуляризация может занулять значения некоторых параметров, что позволяет проводить отбор признаков.

3.2. Сравнение полиномиальных моделей

Базовой моделью линейной регрессии служит модель состоящая из одного признака, а именно сопротивления датчика R_{CO}. Для обучения и тестирования данные были разделены на две последовательные выборки. Тренировачная выборка состояла из 2000 первых строк данных, тестовая выборка из оставшихся 5344 строк данных. Кроме этого исследовались модели линейной регрессии с такими признаками, как сопротивление датчика неметановых углеводородов R_{NMHC}, температура T воздуха и среднее значение концентрации CO_h. Сводные результаты для линейных моделей без регуляризации приведены в таблице (1).

Таблица 1: Результаты линейной регрессии без регуляризации. Ошибки показаны на всём тестовом множестве.

Модель	$MAPE, \%$	$MSE, (\text{мг}/\text{м}^3)^2$	$GRE, \%$
R_{CO}	36.7	0.57	41.2
R_{CO}, T	40.4	0.66	50.2
R_{NM}	35.0	0.4	38.5
R_{NM}, T	32.9	0.31	30.9
R_{CO}, R_{NM}	32.9	0.37	33.6
R_{CO}, R_{NM}, T	**31.2**	**0.30**	**28.3**
R_{CO}, CO_h	36.3	0.52	39.9
R_{CO}, CO_h, T	38.2	0.57	44.1
R_{NM}, CO_h	34.9	0.40	37.8
R_{NM}, CO_h, T	33.1	0.31	29.5
R_{CO}, R_{NM}, CO_h	32.9	0.37	32.6
R_{CO}, R_{NM}, CO_h, T	**31.1**	**0.30**	**26.4**

Как можно увидеть из таблицы при добавлении таких признаков как T и CO_h не происходит значительного улучшения качества линейных моделей без регуляризации.

В таблице (2) представлены результаты моделей линейной регрессии с L_1 регуляризацией. Гиперпараметр регуляризации искался на логарифмической сетке от 10^{-4} и до 10^{+2}. При этом использовалась перекрёстная валидация для временных рядов $TimeSeriesSplit$ с параметром валидации 10.

Таблица 2: Результаты линейной регрессии с L_1 регуляризацией. Ошибки показаны на всём тестовом множестве.

Модель	$MAPE, \%$	$MSE, (\text{мг}/\text{м}^3)^2$	$GRE, \%$
R_{CO}, R_{NM}	32.2	0.38	33.0
R_{CO}, R_{NM}, T	**30.1**	**0.32**	**27.9**
R_{CO}, R_{NM}, CO_h	32.1	0.37	31.8
R_{CO}, R_{NM}, CO_h, T	**29.8**	**0.31**	**26.2**

Как можно увидеть из таблицы (2) регуляризация улучшает относительную точность $MAPE$ и незначительно уменьшает процент выбросов при прогнозировании GRE. Однако, при этом происходит рост ошибки MSE. При использовании L_2 регуляризации изменения были менее значительными, поэтому данная таблица не приводится.

На следующем этапе строились различные полиномиальные модели, анализировались модели как 2, так и 3 степени. Однако, наилучшие показатели были у моделей 2 степени. В таблице (3) приведены результаты для различных многомерных полиномиальных моделей второй степени. Добавление новых признаков происходит согласно уравнению (5). При этом среднее значение концентрации за час CO_h не участвовало в создании полиномиальных признаков.

Таблица 3: Результаты полиномиальной регрессии без регуляризации. Ошибки показаны на всём тестовом множестве.

Модель	$MAPE, \%$	$MSE, (\text{мг}/\text{м}^3)^2$	$GRE, \%$
$Pol_2(R_{CO}, T)$	40.7	0.65	51.3
$Pol_2(...) + CO_h$	35.7	0.5	40.6
$Pol_2(R_{NM}, T)$	27.3	0.27	23.6
$Pol_2(...) + CO_h$	28.3	0.27	22.9
$Pol_2(R_{CO}, R_{NM})$	26.8	0.30	28.1
$Pol_2(...) + CO_h$	26.0	0.29	24.9
$Pol_2(R_{CO}, R_{NM}, T)$	**25.8**	**0.26**	**23.8**
$Pol_2(...) + CO_h$	**25.5**	**0.24**	**20.7**

В таблице (4) приведены результаты полиномиальных моделей с L_1 регуляризацией. Настройка гиперпараметра проводилась на той же сетке, что и в случае линейной регрессии. Параметр перекрёстной валидации $TimeSeriesSplit$ 10.

Таблица 4: Результаты полиномиальной регрессии с L_1 регуляризацией. Ошибки показаны на всём тестовом множестве.

Модель	$MAPE, \%$	$MSE, (\text{мг}/\text{м}^3)^2$	$GRE, \%$
$Pol_2(R_{CO}, R_{NM})$	26.8	0.30	28.0
$Pol_2(...) + CO_h$	26.0	0.29	24.8
$Pol_2(R_{CO}, R_{NM}, T)$	25.7	0.26	23.1
$Pol_2(...) + CO_h$	**25.5**	**0.24**	**20.5**

Согласно результатам таблиц (3),(4) при учёте часа измерения путём добавления нового признака CO_h в полиномиальной регресси происходит улучшение качества большинства моделей. Использование L_1 регуляризации незначительно снижает ошибку GRE. Наименее существенными признаками в случае L_1 регуляризации были квадрат температуры T^2 и квадрат сопротивления датчика неметановых углеводородов R^2_{NMHC}. Однако, избавление от этих признаков в рассматриваемых моделях не позволило улучшить их качества. Не позволило улучшить качество и учёт влажности воздуха RH. Рассмотрение полиномиальных моделей 3 степени и выше, значительно увеличивало количество признаков, но не приводило к улучшению качества.

3.3. Кривые обучения

Кривые обучения показывают величину ошибок на тренировочных и тестовых данных, в зависимости от

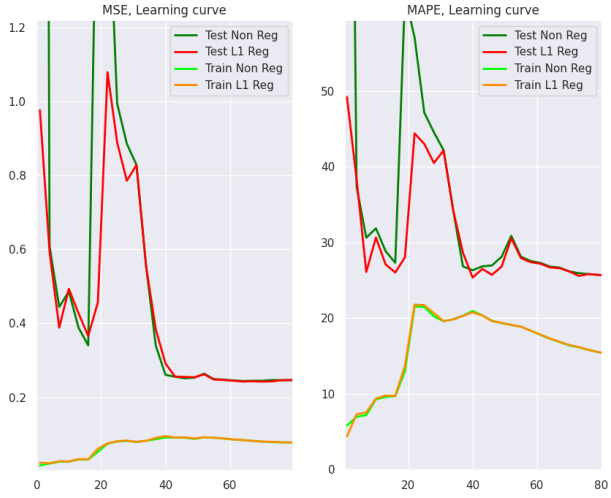

Рис. 3: Кривые обучения для полиномиальной регрессии с признаками $Pol_2(R_{CO}, R_{NM}, T) + CO_h$ для случаев с L_1 регуляризацией и без. По вертикальной оси слева ошибка MSE в $(\text{мг/м}^3)^2$, по вертикальной оси справа ошибка $MAPE$ в процентах. По горизонтальной оси размер тренировочной выборки в днях.

размера тренировочной выборки. Из этого графика можно понять насколько рассматриваемая модель выиграет от увеличения количества данных, а также выяснить какого размера тренировочной выборки будет достаточно для требуемого качества модели на тестовых данных. Кроме этого, из сравнения ошибок на тестовых и тренировочных данных можно выяснить в какую сторону следует изменять сложность модели (количество признаков).

Для построения кривых обучения проводилось разбиение данных на несколько выборок. Наша сутиация моделировала анализ необходимой длительности калибровки сенсора для высокой точности непрерывного прогнозирования значений концентрации CO в течении последних 9 месяцев. В качестве тестовых данных использовались измерения с конца июня 2004 года по апрель 2005 года, а именно последние 5344 последних строк. Тренировочная выборка изменялась в диапазоне от 1 дня и до 80 дней. Она состояла из данных с середины марта по конец июня 2004 года. При этом данные выбирались в хронологическом порядке и без пропусков, сразу после окончания тренировочных данных следовали тестовые. Для анализа была выбрана модель полиномиальной регрессии $Pol_2(R_{CO}, R_{NM}, T) + CO_h$ с L_1 регуляризацией и без. Гиперпараметр регуляризации в каждом случае находился отдельно на сетке от 10^{-5} и до 10^{+2}. Параметр перекрёстной валидации $k-Fold$ был равен 10. На рисунке 3 представлены кривые обучения для метрик MSE слева и $MAPE$ справа.

Как видно из рисунка 3 необходимое время калибровки состовляет порядка 40 дней, при этом существенно в какое время года проводится калибровка. Например, при добавлении новых тренировочных майских данных к данным полученным в июне происходило значительное ухудшение качества модели. Увеличение ошибки при более длительной калибровке совпадало с попаданием в набор тренировочных данных низких значений концентраций газа CO. Стоит отметить, что модель с L_1 регуляризацией лучше справлялась с калибровкой, чем та же модель, но без регуляризации.

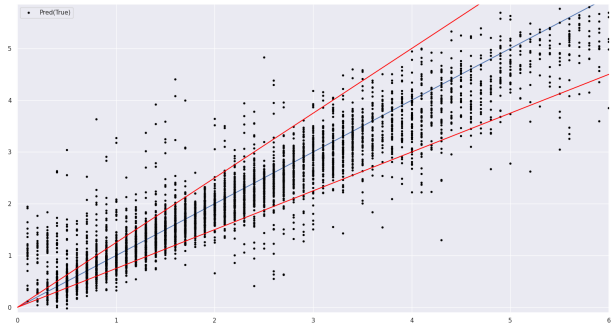

Рис. 4: Точечная диаграмма CO. По вертикали результаты предсказаний концентрации в мг/м^3, по горизонтали истинные значения концентрации в мг/м^3.

3.4. Анализ ошибок

В данном разделе приводится анализ некорректных прогнозов CO в модели полиномиальной регрессии с признаками $Pol_2(R_{CO}, R_{NM}, T) + CO_h$ в случае L_1 регуляризации. На графике 4 представлена точечная диаграмма предсказанных значений от истинных значений концентрации CO на тестовых данных. Синяя прямая на графике показывает где прогноз совпадет с истинным значением. Красные прямые выделяют внутреннюю область, где ошибка предсказания меньше 25% относительно истинного значения. Первые 2000 строк данных использовались для тренировки, а оставшиеся 5344 строк для теста.

Как можно увидеть из диаграммы 4 значительная часть выбросов происходит при низких значениях концентрации CO. Для подтверждения этого факта были построены две ящичные диаграммы 5 и 6. Первая диаграмма (5) показывала зависимость квадратичных отклонений от истинных значений концентрации в диапазоне от 0.1 и до 2.0 мг/м^3. Вторая диаграмма (6) показывала зависимость относительных отклонений от истинных значений концентрации в том же диапазоне.

Как можно увидеть из диаграммы 5 при целевой концентрации от 0.1 и до 0.3 мг/м^3 наблюдаются значительные медианные значения квадратичных отклонений. На диаграмме 6 медианные значения относительной ошибки при прогнозировании целевого газа с концентрацией в 0.1 мг/м^3 около 1000%, а на концентрации в 0.2 мг/м^3 около 250%.

Таким образом, полиномиальная модель с признаками $Pol_2(R_{CO}, R_{NM}, T) + CO_h$ в случае L_1 регуляризации плохо описывает низкие значения целевых концентраций CO. При прогнозировании целевых концентрациях от 0.1 и до 0.3 мг/м^3 модель склонна завышать значения концентрации в несколько раз. При этом на кон-

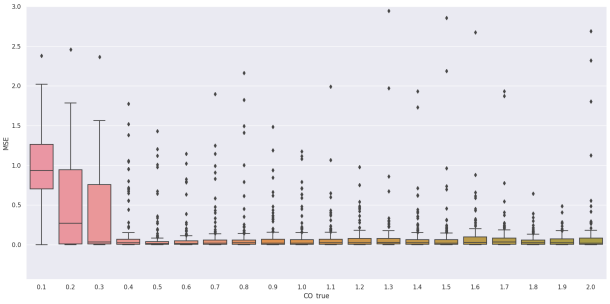

Рис. 5: Ящичная диаграмма квадратичных отклонений. По вертикальной оси квадратичные отклонения на тестовых данных в $(мг/м^3)^2$, по горизонтальной оси целевое значение концентрации угарного газа в $мг/м^3$.

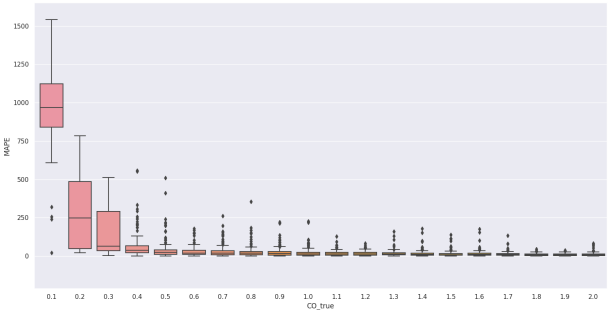

Рис. 6: Ящичная диаграмма относительных отклонений. По вертикальной оси относительные отклонения на тестовых данных в процентах, по горизонтальной оси целевое значение концентрации угарного газа в $мг/м^3$.

центрациях газа выше 0.3 $мг/м^3$ наблюдается уменьшение как абсолютных, так и относительных отклонений. Сводные результаты приведены в таблице (5). Ошибки показаны на всём тестовом множестве, а также на тестовом множестве с концентрацией CO ниже и выше чем 0.3 $мг/м^3$.

Таблица 5: Результаты для модели полиномиальной регрессии с L_1 регуляризацией.

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

4. Нейронные сети прямого распространения

В данном разделе рассматриваются нейронные сети прямого распространения FFNN. В начале рассматриваются различные функции потерь. После чего обсуждается выбор оптимизатора, а также настройка скорости обучения. Далее приводится сравнение различных активационнных функций и производится анализ количества скрытых слоёв и нейронов в скрытом слое. Кроме этого обсуждается создание новых признаков. В заключении обсуждаются различные базовые способы регуляризации нейронных сетей, а также рассматриваются преимущества нейронных сетей с L_1 регуляризацией.

4.1. Функции потерь

Для обучения нейронной сети обязательным является выбор функции потерь. Функция потерь в нашем случае является мерой расхождения между иситнными значениями концентрации CO и оценкой, полученной нейронной сетью. Дальнейшная задача оптимизации и обучения направлена на минимазцию этой функции. В качестве базовых функций потерь часто используются такие, как MSE (7) и $MAPE$ (8), а также MAE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|, \tag{12}$$

где $Y_i, \hat{Y}_i$ – наблюдаемое и предсказанное значение.

К особенностям функции MSE можно отнести требовательное отношение к большим выбросам. При минимизации данной функции оптимизатор стремится как можно лучше описать значения больших выбросов, при этом жертвуя точностью в хороших точках. Ошибка $MAPE$ менее требовательна к выбросам, однако она так же не подходит для нашей задачи. Это связано с тем, что относительные отклонения принимают экстремальные значения на низких концентрациях. Выбор данной функции приведёт к чрезмерной подгонке выбросов в области низких концентраций, при этом жертвуя точностью на больших значениях. Функция ошибок MAE менее чувствительна к большим абсолютным выбросам чем MSE, а также не обладает выборочностью в предсказаниях низких концентраций как $MAPE$.

Существуют также и другие функции потерь для задач регрессии. Одной из них является средняя квадратичная логарифмическая ошибка $MSLE$:

$$MSLE = \frac{1}{N}\sum_{i=1}^{N}\Big(log(Y_i + 1) - log(\hat{Y}_i + 1)\Big)^2. \tag{13}$$

Данная функция схожа с $MAPE$, однако $MSLE$ имеет асимметрию для различных оценок. Она наказывает за недостаточную оценку больше, чем за превышающую истинные значения. При выборе данной функции наша модель будет склонна преувеличивать значения концентрации. Ещё один пример функции потерь это логарифм гиперболического косинуса $MLChE$:

$$MLChE = \frac{1}{N}\sum_{i=1}^{N} log\Big(cosh(Y_i - \hat{Y}_i)\Big). \tag{14}$$

При малых значениях x функция $log(cosh(x)) \approx x^2/2$, а при больших x ведёт себя как $|x| - log(2)$. Исходя из этого, данная функция схожа с MSE, однако она не будет так сильно зависить от существенно неправильных оценок.

Следующая рассматриваемая функция это линейно-эскпоненциальная функция потерь *LINEX*. Это ассиметричная функция с гладкой прозводной:

$$LINEX = \frac{1}{N}\frac{2}{a^2}\sum_{i=1}^{N}[e^{a(\hat{Y}_i - Y_i)} - a(\hat{Y}_i - Y_i) - 1], \quad (15)$$

где $a-$ параметр отвечающий за ассиметрию функции. Если $a > 0$, то модель будет недооценивать концентрацию, накладывая большую ошибку на превышающие значения. Если $a < 0$, то переоценивать, сильнее наказывая недостаточную оценку. В случае если $a > 0$ при малых $|\hat{Y}_i - Y_i|$ фунцкия ведёт себя как MSE. При больших же значениях $|\hat{Y}_i - Y_i|$ поведение линейно в случае недооценки концентрации и экспоненциально для переоценки.

Также была применена новая гладкая ассиметричная функция потерь *MLEE*:

$$MLEE = \frac{1}{N}\sum_{i=1}^{N} log\Big(e^{a(\hat{Y}_i - Y_i)} + be^{-\frac{a}{b}(\hat{Y}_i - Y_i)} - b\Big), \quad (16)$$

где $b = \frac{a^2}{2-a^2}$, настраиваемый параметр $a \in (1, \sqrt{2})$ отвечает за ассиметрию. При значениях a близких к 1 функция практически не обладает ассиметрией, с ростом же параметра a модель склонна к недооценке концентрации. При малых $x = \hat{Y}_i - Y_i$ фунцкия квадратична x^2. При большой переоценке поведение линейно $a|x|$, при значительной недооценке $log(b)+a|x|/b$. Преимуществом данной функции потерь перед *LINEX* является отсутсвие экспоненциального роста ошибки на недостаточных оценках концентрации.

Для сравнения различных фунцкий потерь была выбрана 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. В качестве тренировки использовались первые 2000 строк, для валидации из них выделялось 200 последних строк. Валидационная выборка использовалась для нахожденния оптимального значения гиперпараметра a для *LINEX* на сетке от 0.01 и до 2, а для функции потерь *MLEE* на сетке от 1 и до 1.41. После нахождения оптимального гиперпараметра весь тренировочный набор использовался для обучения. Для тестового набора данных использовались последние 5344 строк. В качестве алгоритма оптимизации был выбран *ADAM*, начальная скорость при этом $5 \cdot 10^{-3}$, коэффициент распада $5 \cdot 10^{-3}$. Количество эпох 1000, размер батча составлял 50 строк. Случайное задание начальных 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 на них. Получившиеся результаты на тестовом наборе данных приведены в таблице (6).

Таблица 6: Результаты FFNN для различных функций потерь. Первое значение в столбце соответсвует среднему значению на метрике, а второе значение стандартному отклонению от него.

Функция потерь	a	$MAPE$, %	MSE, $(мг/м^3)^2$	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

Исходя из результатов таблицы (6) можно сделать вывод, что лучшей функцией потерь является *MAE*, худшие результаты получились на функции *MAPE*.

4.2. Учёт дневных изменений и регуляризация

В данном разделе рассматривается создание новых признаков, учитывающих средние значения сопротивления датчиков и температуры воздуха. Усреднение проводилось по предыдущим 24 строкам. Так например значение сопротивления датчика R_{CO} заменялось на два новых признака WR_{CO} и DR_{CO}:

$$WR_{CO}[j] = \frac{1}{24}\sum_{i=0}^{23} R_{CO}[j-i], \quad (17)$$

$$DR_{CO}[j] = R_{CO}[j] - WR_{CO}[j]. \quad (18)$$

При этом 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.

В качестве анализа использовалась полносвязная нейронная сеть прямого распространения с одним скрытым слоем. Функция потерь была выбрана *MAE*. Количество нейронов в скрытом слое было выбрано 10, гиперболический тангенс использовался как функция активации. В качестве тренировки использовались 1977 строк. Для тестового набора данных использовались последние 5344 строк. В качестве алгоритма оптимизации был выбран *ADAM*, начальная скорость при этом $5 \cdot 10^{-3}$, коэффициент распада $5 \cdot 10^{-3}$. Количество эпох 1000, размер батча составлял 50 строк. Усреднение проводилось по 10 нейронным сетям.

Рассматривались 4 различных модели. Первая использовала признаки R_{CO}, R_{NM}, T в качестве входных значений. Вторая отличалась от первой добавлением признака CO_h. В третьей входные призаки были WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM} и DT, WT. Последняя содержала также CO_h. Результаты представлены в таблице (7).

Как можно увидеть из таблицы (7) добавление нового признака CO_h в нейронные сети приводит к улучшении качества моделей. Разделение же сопротивлений и

Таблица 7: Результаты нейронных сетей без регуляризации. Первое значение в столбце соответсвует среднему значению на метрике, а второе значение стандартному отклонению.

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 ± 0.010**	**17.7 ± 1.4**

тепературр на новые признаки приводит к уменьшению количества выбросов, метрика GRE.

Для тех же 4 моделей была применена L_1, и L_2 регуляризация для наложения штрафа на ядро скрытого слоя. Для этого из тренировачной выборки размером 1977 строк было выделено последних 377 строк для валидации. Нахождение оптимальных гиперпараметров проводилось на валидационной выборке. При этом обучение проводилось на первых 1600 строках с функцией потерь MAE. Количество нейронов в скрытом слое было выбрано равным 10, функция активации $tanh$. Оптимальный гиперпараметр регуляризации искался на логарифмической сетке от 10^{-4} и до $10^{-0.5}$. После нахождения лучшего параметра весь тренировочный набор данных размером 1977 строк использовался для обучения. Для теста использовались последние 5344 строк. Средние значения метрик и стандартные отклонения от них на тестовом наборе данных представлены в таблице (8). Усреднение проводилось по 10 нейронным сетям.

Таблица 8: Результаты нейронных сетей с L_1, L_2 регуляризацией. Первое значение в столбце соответсвует среднему значению на метрике, а второе стандартному отклонению.

Модель, Тип	Параметр	$MAPE$, %	MSE, $(\text{мг}/\text{м}^3)^2$	GRE, %
1 L_1	10^{-1}	25.1±0.2	0.29±0.001	20.9±0.3
2 L_1	10^{-1}	24.0±0.2	0.27+0.002	16.8±0.5
3 L_1	10^{-1}	23.9±0.8	0.29±0.011	16.9±1.7
4 L_1	10^{-1}	**23.4 ± 0.1**	**0.28 ± 0.002**	**15.5 ± 0.2**
1 L_2	10^{-2}	25.5±0.1	0.28±0.008	22.0±0.6
2 L_2	10^{-2}	24.1±0.2	0.27±0.003	16.9±0.4
3 L_2	10^{-2}	24.2±0.3	0.27±0.006	17.2±0.7
4 L_2	10^{-2}	**23.7 ± 0.1**	**0.26 ± 0.005**	**15.8 ± 0.4**

Из результатов таблицы можно понять, что добавление нового признака CO_h и разделение сопротивлений и температур на средние дневные значения и отклонения от них улучшает точность моделей. Кроме этого, регуляризация с наложением штрафа на ядро скрытого слоя позволяет бороться с переобучением и мультиколлинеарностью, что приводит к уменьшению ошибок при прогнозировании. Регуляризация L_1 показывает лучший результат для метрик $MAPE$ и GRE, регуляризация L_2 снижает ошибку MSE на тестовых данных.

Также было замечено, что в нейронных сетях с регуляризацией наблюдается улучшение в стабильности

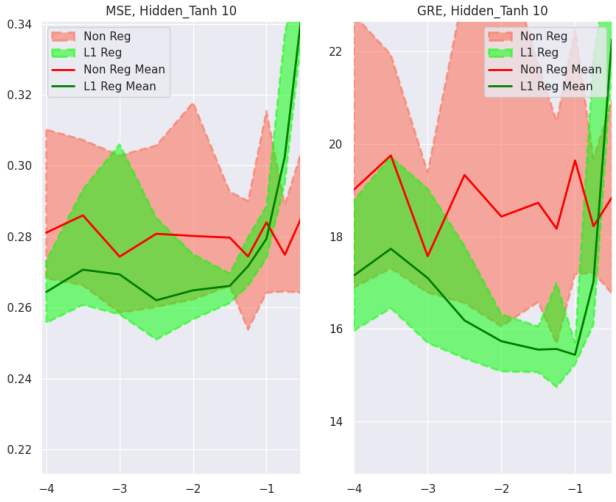

Рис. 7: Результаты нейронных сетей без регуляризации и с регуляризацией в зависимости от параметра регуляризации. По горизонтальной оси значение показателя степени с основанием 10 в регуляризации. По вертикальной оси на первом графике ошибка MSE в $(\text{мг}/\text{м}^3)^2$, на втором ошибка GRE в процентах.

обучения нейронных сетей. Был проведён анализ качества моделей для L_1 регуляризации в зависимости от значения гиперпараметра регуляризации. На рисунке 7 приведены результаты для двух моделей с регуляризацией и без неё. В качестве входных признаков нейронной сети служили WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM} и DT, WT, а также CO_h. Для обучения использовались 1977 строк данных, для теста последние 5344 строк данных. Результаты моделей приведены на тестовом наборе. В качестве алгоритма оптимизации был выбран $ADAM$, начальная скорость при этом $5 \cdot 10^{-3}$, коэффициент распада $5 \cdot 10^{-3}$. Количество эпох 1000, размер батча 50. Усреднение проводилось по 10 нейронным сетям. Кроме средних значений указаны также границы в которых изменялись данные метрики. 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 обуславливается случайным заданием начальных весов. Границы закрашенных областей соотвествуют максимальным и минимальным ошибкам на этих метриках.

4.3. Настройка гиперпараметров и архитектуры нейронной сети

Для нейронной сети с входными признаками WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM} и DT, WT, а также CO_h был выбран алгоритм оптимиазции $ADAM$. Обозначим за θ параметры сети, а функцию потерь за $J(\theta)$. Гиперпараметрами для данного оптимизатора являются скорость обучения α, экспоненциальные скорости затухания для первого и второго момента $\beta_1, \beta_2 \in [0, 1)$, а также константа ε для численной стабильности. Основные вычис-

ления алгоритма можно представить в следующем порядке (Diederik P. Kingma et al., 2015), в начале вычисляется градиент $g_t = \nabla_\theta J_t(\theta_{t-1})$. После чего обновляются оценки первого и второго момента:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1 g_t), \tag{19}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2. \tag{20}$$

Далее выполняется поправка на смещение для моментов $\hat{m}_t = m_t/(1-\beta_1^t)$ и $\hat{v}_t = v_t/(1-\beta_2^t)$. И наконец обновление параметров сети:

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}. \tag{21}$$

Кроме этого существует также возможность изменять скорость обучения. В данной работе уменьшение происходило с помощью коэффициента распада *decay* по следующей формуле:

$$\alpha_t = \alpha_0 \frac{1}{1 + t * decay}, \tag{22}$$

где α_0 начальная скорость обучения. Обновление скорости происходило на каждой итерации t.

Для настройки гиперпараметров в алгоритме *ADAM* был применён метод *GridSearchCV* поиска по сетке. Поиск проводился для нейронной сети с входными признаками WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. Нейронная сеть состояла из одного скрытого слоя, количество нейронов 10. В качестве функции активации анализировались *tanh*, *sigmoid*, *relu* и *exponential*. Поиск оптимальных гиперпараметров проводился для первых 3500 строк данных, а также всех данных. Количество эпох составляло 100, 500 и 1000. Параметр кроссвалидации *KFold* был равен 10. Анализ производился с помощью метрики *MAPE*. В ходе исследований было обнаружено, что оптимальные значения первого момента $\beta_1 = 0.9$, а второго $\beta_2 = 0.999$. Данные значения совпадают со значениями по умолчанию в алгоритме *ADAM*.

Важными параметрами для настройки являлись α_0, *decay*, ε и функция активации. В случае нейронных сетей с регуляризацией настраивался также параметр регуляризации l_1. Оптимальная скорость обучения и коэффициент распада получились $\alpha_0 = 0.01, decay = 0.001$. Оптимальный параметр регуляризации $l_1 = 0.1$, а константа $\varepsilon = 0.1$. Наилучшей активационной функцией для данной задачи оказалась *tanh*.

После настройки гиперпараметров и выбора функции активации были построены кривые тренировки для нейронной сети с одним скрытым слоем в 10 нейронов. Активационная функция *tanh*. Тренировочная выборка состояла из первых 1200 строк данных, для валидационной выборки использовались следующие 777 строк данных. Кривые тренировки изображены на рисунке 8.

Из графика 8 видно, что достаточное количество эпох для нейронной сети с одним скрытым слоем составляет порядка 600.

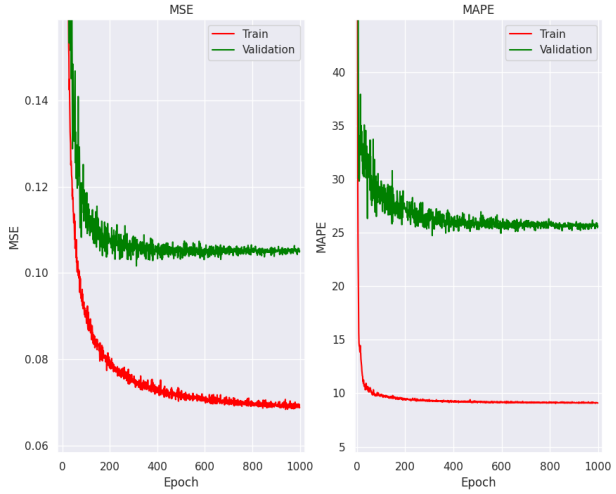

Рис. 8: Кривые тренировки для нейронной сети с L_1 регуляризацией. По горизонтальной оси количество эпох. По вертикальной оси на первом графике ошибка MSE в $(мг/м^3)^2$, на втором ошибка *MAPE* в процентах.

На следующем этапе исследовались различные архитектуры нейронной сети. Для этого с помощью метода поиска по сетке методом *GridSearchCV* находилось оптимальное количество слоёв и нейронов в скрытом слое. Для этого анализировались результаты нейронных сетей с количеством скрытых слоёв от 1 до 5 и количеством нейронов в каждом скрытом слое от 3 до 15. Поиск наилучшей архитектуры проводился для всех данных. Количество эпох составляло 1000 и 2000. Параметр кроссвалидации *KFold* был равен 10. В ходе поиска было обнаружено что оптимальное количество скрытых слоёв нейронной сети было равно 2, а наилучшее количество нейронов в скрытом слое состовляло от 3 до 7. Однако результаты данных моделей слабо отличались от результатов нейронной сети с одним скрытым слоем и 10 нейронами. В таблице 9 приведены результаты на тестовом наборе данных в зависимости от количества скрытых слоёв и нейронов в одном скрытом слое. Обучение проводилось на тренировочной выборке из 1977 первых строк. Для тестового набора использовались последние 5344 строк. Входными признаками нейронной сети служили WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. Гиперболический тангенс использовался в качестве активационной функции. Количество эпох было равно 1000. Усреднение результатов проводилось по 10 нейронным сетям.

4.4. Кривые обучения

В данном разделе приводятся кривые обучения для архитектуры нейронной сети с двумя скрытыми слоями по 5 нейронов в каждом. Активационная функция гиперболический тангенс. Входными признаками нейронной сети служили WR_{CO}, DR_{CO}, WR_{NM}, DR_{NM}, DT, WT, а также CO_h. В качестве тестовых данных использовались измерения с конца июня 2004 года по апрель

Таблица 9: Результаты нейронных сетей с L_1 регуляризацией при различных архитектурах. Первое значение в столбце соответсвует среднему значению, а второе стандартному отклонению.

скрытых слоёв	нейронов	$MAPE$, %	MSE, $(\text{мг/м}^3)^2$	GRE, %
1	3	22.95±0.02	0.246±0.001	14.6±0.1
1	5	22.93±0.02	0.246±0.001	14.5±0.1
1	7	23.0±0.1	0.247±0.002	14.7±0.2
1	10	23.0±0.2	0.246±0.002	14.6±0.3
1	13	23.0±0.2	0.247±0.003	14.6±0.2
2	3	22.87±0.04	0.264±0.002	14.3±0.1
2	5	22.89±0.06	0.264±0.002	14.3±0.1
2	7	22.89±0.06	0.265±0.003	14.3±0.1
2	10	22.91±0.03	0.265±0.002	14.3±0.1
2	13	22.89±0.06	0.266±0.003	14.3±0.1

Рис. 9: Кривые обучения для нейронной сети с двумя скрытыми слоями с L_1 регуляризацией и без. По вертикальной оси слева ошибка MSE в $(\text{мг/м}^3)^2$, по вертикальной оси справа ошибка $MAPE$ в процентах. По горизонтальной оси размер тренировочной выборки в днях.

2005 года, а именно последние 5344 последних строк. Тренировочная выборка изменялась в диапазоне от 1 дня и до 80 дней. Она состояла из данных с середины марта по конец июня 2004 года. Данные выбирались в хронологическом порядке, то есть сразу после окончания тренировочных данных следовали тестовые. Такой способ разделения данных наилучшим образом симулирует калибровку газовых сенсоров. На рисунке 9 представлены кривые обучения для ошибки MSE слева и $MAPE$ справа.

Как можно заключить из графика 9 для стабильного прогнозирования концентраций угарного газа с конца июня по апрель достаточно 40 дней калибровки с середины мая по июнь. Однако, исходя из метрики $MAPE$ при калибровке сенсора от 3 до 30 дней в модели нейронной сети с L_1 регуляризацией не происходит значительного изменения в качестве прогнозов тестовых данных. При этом ошибка $MAPE$ на этом интервале изменяется незначительно и лежит в диапазоне от 26% и до 29%. При калибровке от 50 дней ошибка $MAPE$ выходит на значение в 23%.

5. Заключение

В ходе проделанной работы были построены различные модели для прогнозов значений концентрации угарного газа. В разделе 3 были построены модели линейной и полиномиальной регресии второго порядка. Из анализа было выяснено, что добавление нового признака CO_h улучшает точность моделей. При этом наилучшая MLR модель содержала признаки $Pol_2(R_{CO}, R_{NM}, T)$ а также CO_h и имела L_1 регуляризацию. При этом средняя относительная ошибка $MAPE$ на тестовом наборе из 5344 последних строк данных оказалась равной 25.5%, а ошибка GRE была равна 20.7%. Для данной модели требовалась калибровка газового сенсора в течении порядка 40 дней (рис. 3). При этом, добавление новых тренировочных данных в межсезонье приводило к значительному росту ошибок на тестовом наборе данных. Так например, добавление данных в майский период 2004 года к июньским данным ухудшало $MAPE$ на тестовом наборе более чем на 10%.

Из анализа графика 5 становится очевидным, что проблемным местом в прогнозировании концентрации являются низкие целевые значения концентрации CO. Так например, разброс квадратичных отклонений на концентрациях меньших 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.

В разделе 4 анализировались нейронные сети прямого распространения. Наилучшей функцией потерь оказалась абсолютная ошибка MAE, а лучшей активационной функцией нейрона гиперболический тангенс $tanh$. Создание новых признаков из исходных с помощью дневного усреднения значений сопротивлений датчиков, а также температуры и отклонений от средних значений за день позволило улучшить точность $MAPE$ и GRE. Добавление нового признака CO_h также привело к улучшению качества прогнозов и уменьшению времени калибровки газового сенсора. Наилучшая архитектура нейронной сети состояла из двух скрытых слоёв по 5 нейронов в каждом, с активационной функцией нейрона $tanh$ и L_1 регуляризацией ядра скрытого слоя. Подбор оптимального гиперпараметра регуляризации, а также настройка оптимизатора $ADAM$ с помощью метода $GridSearchCV$, позволило улучшить точность $MAPE$ на тестовых данных из последних 5344 строк до значений $22.9 \pm 0.1\%$, а значение GRE до $14.3 \pm 0.1\%$.

Сравнивая результаты кривых обучения для нейронной сети (рис. 9) и для модели полиномиальной регресии (рис. 3) можно сделать вывод о превосходстве нейронных сетей прямого распространения для калибровки газовых сенсоров. Так например, при добавлении

в тренировочный набор к июньским данным майских у нейронных сетей не происходит значительного увеличения относительной ошибки *MAPE* на всём тестовом наборе, в отличие от модели полиномиальной регресии. Ошибка для нейронной сети *MAPE* изменялась в границах от 26% до 29%. Достаточным временем калибровки при этом можно считать измерения концентрации в течении 3 дней для выхода на среднюю относительную ошибку *MAPE* в 27.5 ± 1.5%. При калибровке газового сенсора от 40 дней средняя ошибка *MAPE* опускалась до стабильных 23.5 ± 0.5%.

Список литературы

De Vito, S., Piga, M., Martinotto, L., and Di Francia, G.: CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, ensor. Actuat. B-Chem., 143, 182–191, https://doi.org/10.1016/j.snb.2009.08.041, 2009. Sensors and Actuators B: Chemical.
DOI: 10.1016/j.snb.2009.08.041

S. De Vito, G. Fattoruso, M. Pardo, F. Tortorella and G. Di Francia, "Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction,"in IEEE Sensors Journal, vol. 12, no. 11, pp. 3215-3224, Nov. 2012, doi: 10.1109/JSEN.2012.2192425.
DOI: 10.1109/JSEN.2012.2192425

Diederik P. Kingma, Jimmy Ba., Adam: A Method for Stochastic Optimization. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015 https://doi.org/10.48550/arXiv.1412.6980
DOI: 10.48550/arXiv.1412.6980

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L., and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low-cost air quality sensing systems, Sensor. Actuat. B-Chem., 231, 701–713, https://doi.org/10.1016/j.snb.2016.03.038, 2016.
DOI: 10.1016/j.snb.2016.03.038

Nuria Castell, Franck R Dauge, Philipp Schneider, Matthias Vogt, Uri Lerner, Barak Fishbain, David Broday, and Alena Bartonova. Can com- mercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment international, 99:293–302, 2017. https://doi.org/10.1016/j.envint.2016.12.007
DOI: 10.1016/j.envint.2016.12.007

Spinelle, L., Gerboles, M., Villani, M., Aleixandre, M., Bonavitacola, F., 2017. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. Sensors Actuators B Chem. 238, 706–715. https://doi.org/10.1016/j.snb.2016.07.036
DOI: 10.1016/j.snb.2016.07.036

Spinelle, L., Gerboles, M., Gabriella Villani, M., Aleixandre, M., Bonavitacola, F., 2015. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensors Actuators B Chem. 215, 249–257. https://doi.org/10.1016/j.snb.2015.03.031
DOI: 10.1016/j.snb.2015.03.031

Rose JJ, Wang L, Xu Q, et al. Carbon Monoxide Poisoning: Pathogenesis, Management, and Future Directions of Therapy [published correction appears in Am J Respir Crit Care Med. 2017 Aug 1;196 (3):398-399]. Am J Respir Crit Care Med. 2017;195(5):596-606. doi:10.1164/rccm.201606-1275CI
DOI: 10.1164/rccm.201606-1275CI

Smith, K. R., Edwards, P. M., Ivatt, P. D., Lee, J. D., Squires, F., Dai, C., Peltier, R. E., Evans, M. J., Sun, Y., and Lewis, A. C.: An improved low-power measurement of ambient NO2 and O3 combining electrochemical sensor clusters and machine learning, Atmos. Meas. Tech., 12, 1325–1336, https://doi.org/10.5194/amt-12-1325-2019, 2019.
DOI: 10.5194/amt-12-1325-2019

Pau Ferrer Cid, Calibration of Low-cost Air pollutant sensors using Machine Learning techniques, Universitat Politècnica de Catalunya, 2019

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. https://doi.org/10.3390/s22124393
DOI: 10.3390/s22124393

Ahmad Mohammadshirazi, Vahid Ahmadi Kalkhorani, Joseph Humes, Benjamin Speno, Juliette Rike, Rajiv Ramnath, Jordan D. Clark, Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: A case study,Building and Environment,Volume 213,2022, https://doi.org/10.1016/j.buildenv.2022.108833.
DOI: 10.1016/j.buildenv.2022.108833

СанПиН "Гигиенические нормативы содержания загрязняющих веществ в атмосферном воздухе городских и сельских поселений"от 28 декабря 2021 года, 1.2.3685-21