

Vitiugova I. Calibration of gas sensor data classification models
Калибровка моделей классификации данных газовых сенсоров

Vitiugova Iuliia

Lomonosov Moscow State University
Skobeltsyn Institute of Nuclear Physics
Laboratory of Adaptive Methods of Data Processing
Scientific supervisors: Sergey Dolenko, Alexander Efitorov
Витюгова Юлия

Московский государственный университет имени М.В.Ломоносова
Научно-исследовательского института ядерной физики имени Д.В. Скобельцына
Лаборатория адаптивных методов обработки данных
Научные руководители: С.А. Доленко, А.О. Ефиторов

***Abstract.** This scientific work focuses on the problem of storing and processing sensor data in modern microelectronics and their application in monitoring systems. An experiment was conducted using gas sensors as an illustration. The efficiency of calibration methods was examined using various machine learning algorithms to solve the problem of gas classification. In particular, the paper emphasizes the importance of model calibration in machine learning to enhance the accuracy of predictions and explain the linearity of the membership function.*

***Keywords:** sensor data, gas sensors, machine learning, big data analysis, model calibration.*

***Аннотация.** В данной научной работе рассматривается проблема хранения и обработки сенсорных данных в современной микроэлектронике и их применение в системах мониторинга. В качестве иллюстрации проведено исследование с использованием газовых датчиков. Рассматривалась эффективность методов калибровки при различных алгоритмах машинного обучения при решении задачи классификации газов. В особенности, статья выделяет важность калибровки моделей в машинном обучении для повышения эффективности предсказаний и прояснения линейности функции принадлежности.*

***Ключевые слова:** сенсорные данные, газовые датчики, машинное обучение, анализ больших данные, калибровка моделей.*

Введение

В настоящее время заметны весьма значительные успехи в развитии микроэлектроники и коммуникационных технологий, повсеместное проникновение цифровых технологий привели к появлению огромных объемов сенсорных данных. Сбор, хранение и анализ сенсорных данных, являются немаловажным направлением деятельности в современном мире (I. Shafer, 2013). Сенсорные данные от датчика представляют собой временные ряды. Примерами источников данных являются системы управления производством, системы эксплуатационного мониторинга телекоммуникационных систем, системы автоматизации научных исследований. Анализ сенсорных данных позволяет контролировать изменение состояния объекта мониторинга во времени, изучать закономерности поведения, выявлять тенденции, удаленно диагностировать неисправности, принимать обоснованные решения по управлению объектом. Задача хранения сенсорных данных, таким образом, является актуальной и уже имеет множество решений. Недавний систематический поиск программного обеспечения баз данных временных

рядов обнаружил 83 программных системы, и не похоже, чтобы на этом развитие темы закончилось (A. Bader, 2017). Разнообразие решений вызвано разнообразием условий задачи, в числе которых количество датчиков, темп поступления данных, длительность хранения, преобладание запросов на запись или на чтение, типы запросов на чтение, экономические ограничения. В исследуемой работе были рассмотрены данные с газовых сенсоров, представляющие собой простые временные ряды. Суть эксперимента заключалась в устранении утечек в трубопроводе с помощью создания газовых сенсоров. Проверка трубопровода основана на оценках: с помощью численных моделей, на основе параметров работы трубопровода, обнаружения акустической эмиссии и т.д. В нашем случае модель предсказывала значение одного из двух классов в ходе которых выступали газы: пропан и водород. Проводился анализ функции ответа выходов модели на валидационной выборке, приводя ее к линейному виду для того, чтобы оценить степень принадлежности примера к тому или иному классу. По итогу рассматривалась эффективность методов калибровки при различных алгоритмах машинного обучения при решении задачи бинарной классификации.

Модели машинного обучения помогают принимать решения в весьма сложных и ответственных задач, решение которых без участия человека казалось ранее недоступным. Наряду с этим одной из важнейших и до конца нерешённых теоретических и практических проблем в машинном обучении является калибровка моделей. На протяжении обучения в связи с разными причинами итоговая модель может демонстрировать разную эффективность для всевозможных примеров. Вытекающий фактор — функция числа, показывающая принадлежность примера к тому или иному классу не является линейной. Вместе с тем при выборе по порогу предполагается линейное и одинаковое поведение всех выходов сети. *Калибровкой модели* называют задачу приведения функции принадлежности к линейному виду.

Данные

Используемые в данной работе экспериментальные данные были получены на химическом факультете МГУ кафедры неорганической химии в лаборатории химии и физики полупроводниковых и сенсорных материалов, руководитель группы ст.н.с., к.х.н. В.В.Кривецкий. В ходе эксперимента проводились измерения и запись данных с газовых датчиков для мониторинга атмосферного воздуха с целью обнаружения утечек. Использовался синтез газочувствительных материалов с помощью одностадийного метода пиролиза с распылением пламени для изготовления датчиков. В настоящем исследовании были использованы три материала: чистый SnO₂, модифицированный золотом SnO₂ и биметаллические Au и Pd,

модифицированные SnO₂. Датчики помещали в изолированную проточную камеру из политетрафторэтилена. Поток воздуха отбирался снаружи здания химического факультета МГУ через трубку из ПТЭФ, снабженную керамическим пылевым фильтром с помощью мембранного вакуумного насоса. Наружный поток был зафиксирован с помощью дроссельной заслонки. Газовые баллоны использовались для добавления газов с помощью регуляторов расхода. Генератор чистого воздуха использовался для подачи воздуха, разбавления потока углеводородов, а также для формирования потока воздуха через камеру датчика без примеси метана или пропана. Смешивание газов продолжалось 12 ч и состояло из трех этапов по 1 ч каждый с поступлением газа в различных концентрациях как для метана, так и для пропана. Каждая стадия поступления газа отделялась от следующей потоком окружающего воздуха в течение 1 ч. Приготовление газовой смеси было полностью автоматизировано с использованием контроллеров массового расхода и заслонок, управляемых специальным программным обеспечением для ПК. Датчики работали в режиме модуляции рабочей температуры, с линейными ступенями нагрева и охлаждения и стабильными температурными интервалами 5 секунд при низкой и высокой температуре. Высокая температура была установлена равной 500°C для всех 3 датчиков, низкая температура была установлена равной 100°C. Стадия нагрева длилась 35 секунд, охлаждения – 15 секунд. Таким образом, один цикл модуляции температуры длился 60 секунд. Примененный протокол смешивания газа позволил получить 60 циклов измерений концентрации газа. Соппротивление чувствительного слоя измеряли в режиме постоянного тока смещения. Для реализации программы режима работы датчика и сбора данных использовался лабораторный электронный блок управления.

Оценка прогнозирования

При калибровке модели мы пытаемся масштабировать предсказания модели так, чтобы распределение и поведение оценки вероятности для каждого класса было аналогично распределению и поведению вероятности для каждого класса, наблюдаемым в обучающих данных.

Рассмотрим задачу классификации с k – количеством классов, входными данными $x \in X$ и меткой $y \in Y = \{1, \dots, k\}$. Пусть h – нейронная сеть и при ее действии на X , получаем: $h(X) = (\hat{Y}, \hat{P})$, где \hat{Y} – предсказание класса, а \hat{P} – вероятность достоверности предсказания, подлежащая калибровке. Введем понятие *идеальной калибровки*, которая определяется как:

$$\mathbb{P}(\hat{Y} = Y | \hat{P} = P) = p, \quad \forall p \in [0,1]$$

При идеальной калибровке значения оценки принадлежности должны

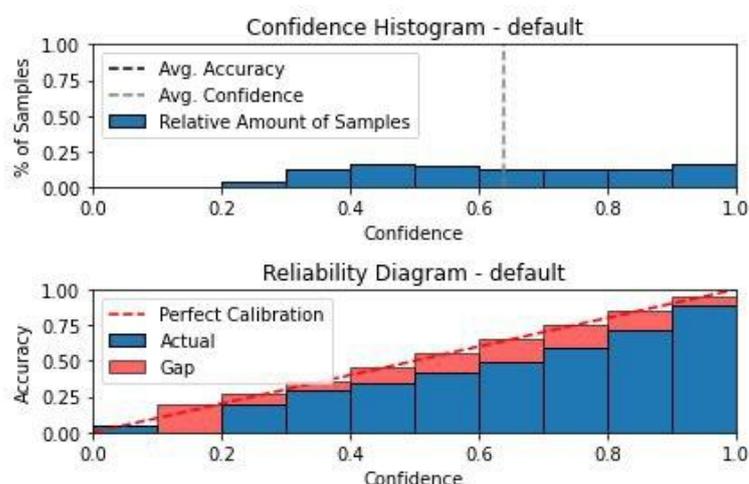
соответствовать доле правильно классифицированных значений. Например, при рассмотрении 100 примеров, каждый с вероятностью правильного ответа 0.8, ожидая, что 80 из них должны быть правильно классифицированы. В реальных условиях достичь идеальной калибровки практически невозможно, но качество калибровки возможно оценить с помощью некоторых метрик:

Диаграмма калибровки

Диаграмма калибровки является распространенным средством для иллюстрации свойств вероятностных систем прогнозирования. Для построения диаграммы калибровки требуется оценить *точность* и *уверенность* модели по конечным выборкам, затем оба значения визуализируются в виде гистограммы с идентификационной линией (Рис.5). Если модель идеально откалибрована, то диаграмма должна построить функцию идентичную идеальной калибровки. В случае, если идентификационная линия ниже диагонали, тогда предсказанные оценки модели слишком велики, если ниже — слишком малы. Для удобства дальнейшего изложения введем понятие *биннинга*. Значения принадлежности объекта к классу модели группируются по оценки вероятности в интервал ячеек M и рассчитывается точность каждой ячейки. Данный алгоритм называется *биннингом*. Далее создается множество индексов выборок B_m , значения предсказаний, которые попадают в интервал $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$. Рассчитать уверенность и точность модели можно по следующим формулам:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) ,$$

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$



Методы калибровки

Калибровка функции принадлежности — это операция масштабирования, которая применяется после того, как предсказания были сделаны моделью. Рассмотрим выходы нейронной сети: модель, получая входные данные x_i и метки y_i , получает прогнозируемые метки \hat{y}_i для класса k и выводит оценки принадлежности \hat{p}_i объектов к классам, а также логит-функции $z_i \in \mathbb{R}$.

Биннинг гистограмм (Zadrozny, 2001) — это непараметрический метод калибровки. Вслед за биннингом некалиброванных предсказаний \hat{p}_i , следует задача оптимизации по параметрам θ . Каждой ячейке B_1, \dots, B_M присваивается калиброванный параметр θ , таким образом, чтобы минимизировать функцию потерь:

$$\min_{\theta_1, \dots, \theta_M} \sum_{m=1}^M \sum_{i=1}^n \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1}) (\theta_i - y_i)^2$$

$$B_m \in (a_m, a_{m+1}]$$

где $\mathbf{1}$ - функция индикатора. Учитывая фиксированные границы ячеек, решение приводит к θ_m , которые соответствуют среднему числу образцов положительного класса в ячейке B_m .

Изотонная регрессия (Zadrozny, 2001) — это наиболее распространенный непараметрический метод калибровки, изучающий кусочно-постоянную функцию f для преобразования некалиброванных выходов: $\hat{q} = f(\hat{p})$. Решение задачи оптимизации идентично биннингу гистограмм, но существует дополнительное условие:

$$\min_{\substack{\theta_1, \dots, \theta_M \\ \alpha_1, \dots, \alpha_{M+1}}} \sum_{m=1}^M \sum_{i=1}^n \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1}) (\theta_i - y_i)^2$$

$$0 = \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_{M+1} = 1,$$

$$\theta_1 \leq \theta_2 \leq \theta_M$$

Где M — число интервалов ячеек, $\alpha_1, \dots, \alpha_{M+1}$ — границы интервалов. При такой параметризации изотонная регрессия является строгим обобщением биннинга гистограмм, в котором границы ячеек и предсказания ячеек совместно оптимизируются.

Байесовский Биннинг (Naeini, 2015) является ансамблевым методом. Он осуществляет комбинацию нескольких биннингов с разным числом ячеек, создавая схему биннинга s , которая представляет собой пару (M, \mathcal{L}) из числа

ячеек M и соответствующего разбиения \mathcal{L} . Если биннинг гистограмм и изотоническая регрессия дают единую схему биннинга, то Байесовский биннинг рассматривает пространство S всех возможных схем биннинга для набора данных валидации D . Оценка принадлежности выводится с помощью линейной комбинации оценок биннингов с весами, применяя байесовское усреднение вероятностей:

$$\begin{aligned}\mathbb{P}(\hat{q}_{te}|\hat{p}_{te}, D) &= \sum_{s \in S} \mathbb{P}(\hat{q}_{te}, S = s | \hat{p}_{te}, D) \\ &= \sum_{s \in S} \mathbb{P}(\hat{q}_{te} | \hat{p}_{te}, S = s, D) \mathbb{P}(S = s | D)\end{aligned}$$

где $\mathbb{P}(\hat{q}_{te} | \hat{p}_{te}, S = s, D)$ — калиброванная вероятность с использованием схемы биннинга S .

Температурное шкалирование придумано специально для калибровки моделей нейронных сетей. Метод температурного шкалирования считается самым простым в реализации, при этом дает весьма хорошие результаты. Идея метода заключается в том, что он работает непосредственно на логит-функциях z_i , а не на оценках принадлежности \hat{p}_i . Логит-функции масштабируются с помощью скалярного параметра $T > 0$ для всех классов. Оптимальный параметр T получается из минимизации отрицательного логарифмического правдоподобия для отложенного набора данных. T — называют

«температурой», и она как бы «смягчает» функцию softmax (т. е. повышая выходную энтропию), T не зависит от результата классификации, поэтому точность калибруемого алгоритма не меняется. Калиброванное значение принадлежности может быть получено путем:

$$\hat{q}'_i = \max_k \sigma_{SM} \left(\frac{z_i}{T} \right)^{(k)}$$

Результаты

В данном вычислительном эксперименте были рассмотрены выходы четырех алгоритмов машинного обучения: Многослойного персептрона (МП), Метода опорных векторов (МОВ), Стохастического градиентного спуска (СГС) и Случайного леса (СЛ). Используемые данные были разделены на обучающий и тестовый набор, которые содержали 5292, 2808 примеров. В качестве базовых параметров обучения модели нейронной сети были взяты следующие: персептрон с логистическая сигмоидальной передаточной функцией для скрытых слоев, для выходного слоя использовалась линейная. Далее к моделям были применены методы калибровки: Биннинг гистограмм, Изотонная регрессия, Байесовский биннинг и Температурное шкалирование. Рассчитаны оценочные значения эффективности (площадь под графиком ROC — AUC) моделей до и после применения методов калибровки.

	<i>Original model without calibration</i>	<i>Histogram Binning calibration method</i>	<i>Isotonic Regression calibration method</i>	<i>Bayesian Binning calibration method</i>	<i>Temperature Scaling calibration method</i>
<i>MP</i>	0.910	0.884	0.892	0.899	0.910
<i>MOV</i>	0.917	0.911	0.911	0.912	0.917
<i>SGD</i>	0.928	0.922	0.930	0.928	0.928
<i>RF</i>	0.925	0.925	0.927	0.926	0.925

Заключение

При нейросетевом решении задачи бинарной классификации, полученные метки и оценки принадлежности примеров к классам прошли постобработку с помощью методов калибровки: Гистограммного биннинга, Изотонной регрессии, Байесовского биннинга и Температурного шкалирования. По результатам выполнения данной работы можно сделать следующие выводы:

1. Для модели МП классификатора лучшие оценки эффективности показали методы Температурного шкалирования и Байесовского биннинга, наблюдалось ухудшение качества при методах Изотонной регрессии и Гистограммного биннинга.

2. Для MOV модели высокие значения получены при применении метода Температурного шкалирования, другие методы показали менее высокие результаты.

3. Для классификатора СГС самая высокая оценка эффективности была при применении метода Изотонной регрессии.

4. При применении метода Изотонной регрессии, получены наилучшие результаты для модели СЛ.

References

1. I. Shafer, R.R. Sambasivan, A. Rowe, G.R. Ganger, Specialized storage for big numeric time series, 2013
2. A. Bader, O. Kopp, M. Falkenthal. "Survey and comparison of open-source time series databases, 2017.
3. Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger. On Calibration of Modern Neural Networks, 2017.
4. Max Kuhn. Applied Predictive Modeling, 2017.
5. Valeriy V. Krivetskiy, Matvei D. Andreev, Aleksandr O. Efitov, Alexander M. Gaskov. Statistical shape analysis of temperature modulated metal oxide gas sensor response for improved selectivity of hydrocarbons detection in real atmospheric conditions, 2020.

6. Jochen Bröcker and Leonard A. Smith. Increasing the Reliability of Reliability Diagrams, 2007.
7. Alexandru Niculescu-Mizil, Rich Caruana. Predicting Good Probabilities with Supervised Learning, 2005.
8. Bianca Zadrozny Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, 2001.
9. Tim Leathart, Maksymilian Polaczuk, Temporal Probability Calibration, 2020.
10. Taejong Joo, Uijung Chung, Min-Gwan Seo, Being Bayesian about Categorical Probability, 2020.
11. Tim Leathart Eibe Frank Bernhard Pfahringer Geoffrey Holmes, On Calibration of Nested Dichotomies, 2018.

Приложение

