УДК 004.852

Д.А. Волуйский

студент, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

З.М. Альбекова

канд. пед. наук, доцент Департамента цифровых, робототехнических систем и электроники Института перспективной инженерии, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

Ф.В. Самойлов

канд. техн. наук, доцент Департамента цифровых, робототехнических систем и электроники Института перспективной инженерии, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

СРАВНИТЕЛЬНЫЙ АНАЛИЗ КАЛИБРОВКИ УВЕРЕННОСТИ КЛАССИФИКАЦИОННЫХ МОДЕЛЕЙ И ЭФФЕКТИВНОСТИ ПОСТ-ХОК МЕТОДОВ КАЛИБРОВКИ

Аннотация. Рассматривается проблема смещённости вероятностных оценок классификационных моделей и оценивается результативность трёх пост-хок методов калибровки: температурного масштабирования, сигмоидной калибровки Платта и изотонической регрессии. Эксперименты выполнены на пяти моделях машинного обучения и четырёх открытых наборах данных при тридцати случайных разбиениях выборки на обучающую, калибровочную и тестовую части. Качество оценивалось по метрикам ECE, MCE, Brier Score и NLL, значимость различий проверялась парным критерием Уилкоксона с поправкой Бонферрони–Холма. Установлена сильная положительная связь между исходным уровнем калибровки модели и величиной её улучшения после калибровки. Показано, что температурное масштабирование является наиболее безопасным универсальным инструментом, тогда как калибровка Платта при малых калибровочных выборках способна существенно ухудшать качество

вероятностных прогнозов. Сформулированы практические рекомендации по выбору метода калибровки в зависимости от типа модели и объёма данных.

Annotation. The paper addresses the bias of probabilistic estimates produced by classification models and evaluates the effectiveness of three post-hoc calibration techniques: temperature scaling, Platt scaling, and isotonic regression. Experiments cover five machine learning models and four public datasets with thirty random train/calibration/test splits. Quality is assessed by ECE, MCE, Brier score and NLL; significance is tested with the paired Wilcoxon test and the Bonferroni–Holm correction. A strong positive relationship is revealed between the initial calibration level of a model and the magnitude of improvement after calibration. Temperature scaling turns out to be the safest universal approach, while Platt scaling may considerably degrade probabilistic forecasts when calibration sets are small. Practical recommendations on method choice depending on model type and data volume are provided.

Ключевые слова: калибровка вероятностей; Expected Calibration Error; Temperature Scaling; Platt Scaling; изотоническая регрессия; критерий Уилкоксона

Keywords: probability calibration; Expected Calibration Error; temperature scaling; Platt scaling; isotonic regression; Wilcoxon test

Введение

Современные классификаторы возвращают не только метку класса, но и вероятностную оценку уверенности, на которую опираются системы медицинской диагностики, автономного управления и другие ответственные приложения. Однако такие оценки часто смещены: модель может сообщать уверенность 0,9, тогда как фактическая доля верных ответов составляет лишь около 0,7. Это расхождение называют мискалибровкой, и оно опасно в системах с высокой ценой ошибки.

Для исправления мискалибровки предложено множество пост-хок методов, однако большинство исследований сосредоточено на глубоких нейронных сетях,

оставляя в стороне классические алгоритмы: случайный лес, градиентный бустинг, SVM и логистическую регрессию. Кроме того, эффективность методов редко проверяется строгим статистическим инструментарием с учётом множественных сравнений. Цель работы — оценить, при каких условиях пост-хок калибровка действительно улучшает качество вероятностных прогнозов, а когда она бесполезна или даже вредна. Для этого проверялись три гипотезы: о связи исходной калибровки модели с величиной эффекта калибровки; о значимых различиях между методами; о зависимости эффективности методов от размера калибровочной выборки.

Материалы и методы

Экспериментальная база включала четыре открытых набора данных библиотеки scikit-learn, охватывающих бинарную и многоклассовую классификацию (таблица 1). В качестве моделей использованы логистическая регрессия с L2-регуляризацией, SVM с вероятностным выходом, случайный лес из 200 деревьев, градиентный бустинг на 100 итераций и многослойный перцептрон с одним скрытым слоем из 64 нейронов; гиперпараметры зафиксированы для воспроизводимости.

Таблица 1 — Характеристики наборов данных

Датасет	Образцов	Признаков	Классов	Тип задачи
Digits	1797	64	10	Многоклассовая
Breast Cancer	569	30	2	Бинарная
Wine	178	13	3	Многоклассовая
Iris	150	4	3	Многоклассовая

Сравнивались три метода. Температурное масштабирование делит логиты на единственный параметр — температуру, влияя лишь на распределение уверенности и не меняя метки классов. Сигмоидная калибровка Платта обучает логистическую функцию на калибровочной выборке; её недостаток — высокая дисперсия оценки параметров на малых данных. Изотоническая регрессия строит

кусочно-постоянную монотонную функцию, минимизирующую квадратичную ошибку; она гибка, но склонна к переобучению.

Каждый датасет делился на обучающую (60 %), калибровочную (20 %) и тестовую (20 %) части при 30 случайных разбиениях. Качество оценивалось по четырём показателям: ожидаемой ошибке калибровки ECE, максимальной ошибке MCE, показателю Брайера и отрицательному логарифмическому правдоподобию NLL. Для каждой из 60 комбинаций «модель × датасет × метод» парный критерий Уилкоксона сопоставлял ECE до и после калибровки; 60 p-значений корректировались поправкой Бонферрони–Холма ($\alpha = 0{,}05$). Дополнительно на всех 20 парах «модель × датасет» варьировалась доля калибровочной выборки (5–100 %) с десятью повторами на конфигурацию.

Результаты

Базовые значения ECE варьировались от 0,014 до 0,222 при среднем 0,069. Наиболее сильные эффекты получены для моделей с плохой исходной калибровкой: случайный лес на Digits снизил ECE с 0,222 до 0,016 — более чем в десять раз. Калибровочные кривые до и после обработки иллюстрирует рисунок 1.

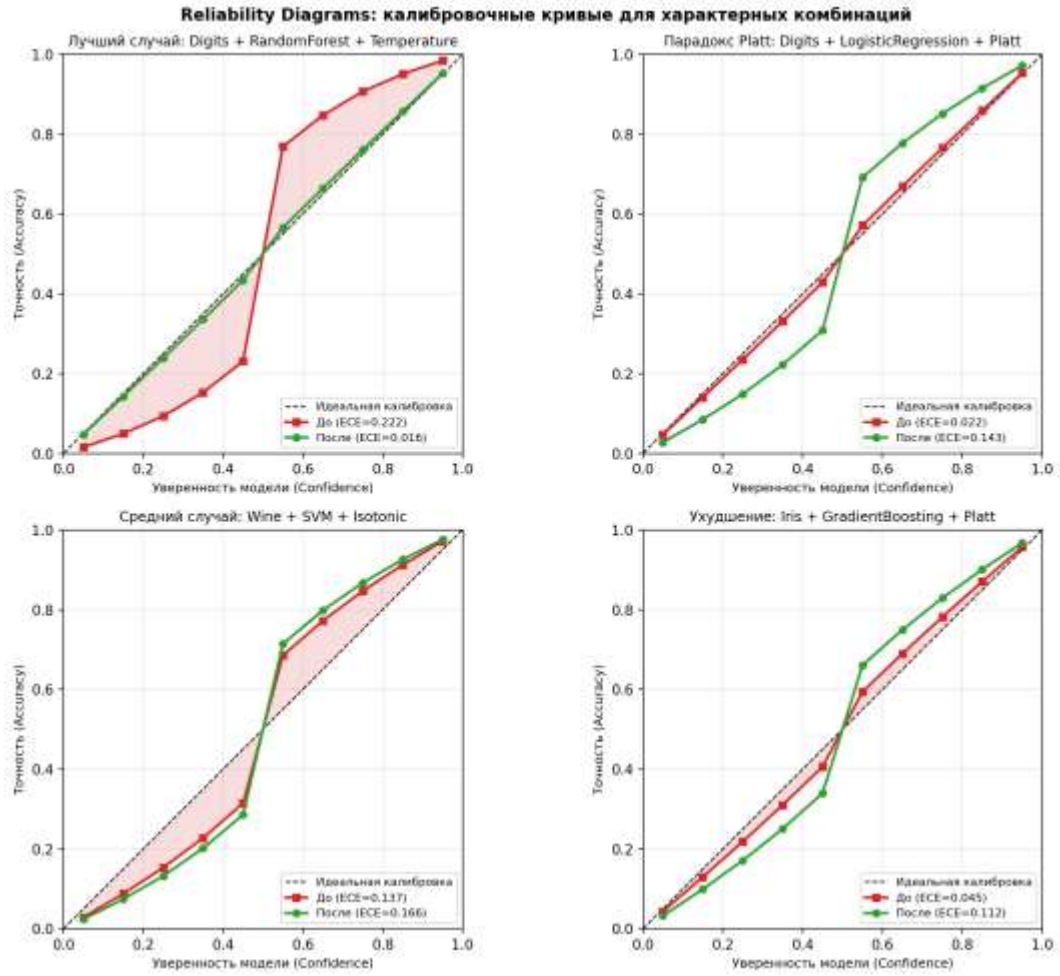

Рисунок 1 — Калибровочные кривые (reliability diagrams) для характерных комбинаций

Между исходным уровнем ECE модели и средним улучшением после калибровки обнаружена сильная положительная корреляция Пирсона ($r = 0{,}767$, $p < 0{,}001$). Случайный лес со средней исходной ECE 0,108 получает наибольший выигрыш (+0,057), тогда как хорошо откалиброванные логистическая регрессия и градиентный бустинг в среднем не улучшаются и даже деградируют (таблица 2).

Таблица 2 — Сравнительная эффективность методов калибровки

Метод	Ср. ΔECE	Значимых улучшений	Значимых ухудшений	Без эффекта
Temperature Scaling	0,0248	6	0	14

Isotonic Regression	0,0124	5	5	10
Platt Scaling	−0,0318	2	12	6

Температурное масштабирование оказалось наиболее безопасным методом: ни одного значимого ухудшения из двадцати комбинаций и шесть значимых улучшений. Изотоническая регрессия занимает промежуточное положение. Сигмоидная калибровка Платта значимо ухудшила ECE в двенадцати случаях из двадцати — прежде всего для исходно хорошо откалиброванных моделей; этот феномен условно назван «парадоксом Платта». Ранжирование методов по показателю Брайера и NLL согласуется с ранжированием по ECE.

Эксперимент с долей калибровочной выборки выявил резкое расхождение методов. Температурное масштабирование достигает почти полной эффективности уже при 5 % пула (ECE ≈ 0,028 против базовых 0,076), изотоническая регрессия стабильна примерно с 10 %. Сигмоидная калибровка при 5 % даёт катастрофический рост ECE до 0,54 и выходит на приемлемый уровень лишь с 50 % выборки. Причиной служит соотношение смещения и дисперсии: стандартная ошибка коэффициентов логистической регрессии растёт с уменьшением выборки, и на нескольких десятках образцов параметрическая оценка перестаёт быть состоятельной.

Обсуждение

Результаты согласуются с литературой: превосходство температурного масштабирования для нейронных сетей показано в [1], а избыточность калибровки для современных хорошо откалиброванных моделей — в [2]. Настоящее исследование переносит эти выводы на классические модели и количественно характеризует «парадокс Платта» средним ухудшением −0,032 по ECE.

Из 60 комбинаций значимое улучшение после поправки зафиксировано в 13 случаях, ухудшение — в 17, отсутствие эффекта — в 30; без поправки числа составляли 14, 20 и 26. Четыре комбинации перешли в разряд незначимых, что

подтверждает необходимость коррекции при массовом тестировании: калибровка не является универсально полезной операцией. Практические рекомендации: случайный лес калибровать всегда; при исходной ECE ниже 0,05 от калибровки воздержаться; температурное масштабирование выбирать по умолчанию; изотоническую регрессию — при не менее 200 калибровочных образцов; сигмоидную калибровку — только при выборке не менее 500 образцов и ECE выше 0,05.

Заключение

Выполнен систематический сравнительный анализ трёх пост-хок методов калибровки на пяти моделях и четырёх наборах данных со строгим статистическим контролем. Установлено, что величина эффекта определяется исходной калибровкой модели, температурное масштабирование является самым безопасным методом, а сигмоидная калибровка опасна при малых выборках. Перспективы: расширение базы данных, включение глубоких архитектур и разработка адаптивной стратегии выбора калибровки.

Список литературы

1. Guo C., Pleiss G., Sun Y., Weinberger K.Q. On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 1321–1330.

2. Minderer M., Djolonga J., Villegas R. et al. Revisiting the Calibration of Modern Neural Networks // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 15682–15694.

3. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods // Advances in Large Margin Classifiers. – Cambridge : MIT Press, 1999. – P. 61–74.

4. Zadrozny B., Elkan C. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers // Proceedings of the 18th International Conference on Machine Learning. – 2001. – P. 609–616.

5. Naeini M.P., Cooper G., Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning // Proceedings of the 29th AAAI Conference on Artificial Intelligence. – 2015. – P. 2901–2907.

6. Nixon J., Dusenberry M.W., Zhang L., Jerfel G., Tran D. Measuring Calibration in Deep Learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2019. – P. 38–41.

References

1. Guo C., Pleiss G., Sun Y., Weinberger K.Q. On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 1321–1330.

2. Minderer M., Djolonga J., Villegas R. et al. Revisiting the Calibration of Modern Neural Networks // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 15682–15694.

3. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods // Advances in Large Margin Classifiers. – Cambridge : MIT Press, 1999. – P. 61–74.

4. Zadrozny B., Elkan C. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers // Proceedings of the 18th International Conference on Machine Learning. – 2001. – P. 609–616.

5. Naeini M.P., Cooper G., Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning // Proceedings of the 29th AAAI Conference on Artificial Intelligence. – 2015. – P. 2901–2907.

6. Nixon J., Dusenberry M.W., Zhang L., Jerfel G., Tran D. Measuring Calibration in Deep Learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2019. – P. 38–41.