УДК 004.89:004.852

Компромисс между точностью и справедливостью в алгоритмах искусственного интеллекта: стратегии оптимизации обучающих выборок

К.Д. Коновалов✉

Воронежский институт высоких технологий, Воронеж, Россия

Статья анализирует компромисс между точностью и справедливостью в алгоритмах искусственного интеллекта (ИИ) и машинного обучения, фокусируясь на оптимизации обучающих выборок. Рассматриваются причины предвзятости, включая дисбаланс классов и корреляции с чувствительными признаками. Описываются методы предобработки данных, балансировки классов и отбора признаков. Особое внимание уделяется метрикам справедливости (DPD, EOD) для минимизации дискриминации. Предложены рекомендации по сочетанию методов для достижения баланса между производительностью и объективностью. Статья ориентирована на специалистов, разрабатывающих этичные ИИ-системы.

Ключевые слова: искусственный интеллект, машинное обучение, предвзятость, справедливость, оптимизация данных, метрики справедливости.

Trade-Off Between Accuracy and Fairness in Artificial Intelligence Algorithms: Strategies for Optimizing Training Datasets

K.D. Konovalov✉

Voronezh Institute of High Technologies, Voronezh, Russia

The article explores the trade-off between accuracy and fairness in artificial intelligence (AI) and machine learning algorithms, emphasizing training dataset optimization. It examines sources of bias, such as class imbalance and correlations with sensitive attributes. Data preprocessing, class balancing, and feature selection methods are discussed. Special focus is given to fairness metrics (DPD, EOD) to reduce discrimination. Recommendations for combining methods to balance performance and objectivity are provided. The article targets specialists developing ethical AI systems.

Keywords: artificial intelligence, machine learning, bias, fairness, data optimization, fairness metrics.

Введение

Алгоритмы искусственного интеллекта (ИИ) и машинного обучения (МО) стали неотъемлемой частью современных технологий, находя применение в здравоохранении, финансах, государственном управлении, маркетинге, логистике и других областях. Эти технологии позволяют анализировать огромные массивы данных, выявлять сложные закономерности и принимать автоматизированные решения с высокой скоростью и точностью. С ростом влияния алгоритмов на общество всё более актуален вопрос их объективности [7]. Объективность определяется как способность алгоритмов выдавать результаты без систематических искажений, устойчивые к изменениям данных и воспроизводимые при повторном обучении. Предвзятость проявляется в снижении точности на новых данных, игнорировании редких сценариев и дискриминации по признакам, таким как пол, возраст, этнос или социальный статус [1].

Ключевым фактором, определяющим объективность моделей, является качество обучающих выборок [5]. Недостатки данных, такие как дисбаланс классов, недостаточное разнообразие, наличие шумов или аномалий, а также корреляции с чувствительными признаками, существенно усиливают предвзятость [6]. Попытки повысить метрики точности, такие как F1-score или ROC-AUC, могут привести к снижению справедливости, когда модель показывает высокую производительность на большем объёме данных, но ошибается в случае малочисленных групп или редких случаев [8]. Найти компромисс между точностью и справедливостью крайне важно, так как в социально значимых областях предвзятость может привести к несправедливым решениям и дискриминации, что влечёт за собой утрату доверия к технологиям.

Цель данной статьи – рассмотреть компромисс между точностью и справедливостью в алгоритмах ИИ и выполнить анализ стратегий оптимизации обучающих выборок, которые позволят достичь баланса между этими критериями. Особое внимание было уделено методам предобработки данных, балансировки классов, отбора признаков и выбору моделей машинного обучения.

Проблема компромисса между точностью и справедливостью

Точность моделей ИИ оценивается с помощью таких метрик качества, как Accuracy (доля правильных предсказаний), Precision (точность положительных предсказаний), Recall (полнота), F1-score (гармоническое среднее Precision и Recall) и ROC-AUC (площадь под кривой ошибок). Эти метрики позволяют эффективно измерить общую производительность алгоритма на обучающей и тестовой выборках, но не учитывают справедливость результатов обучения [7]. К примеру, модель с высокими показателями метрик точности может корректно классифицировать большинство данных, но иметь ошибки в работе с определёнными группами, имеющими чувствительные признаки такие, как пол, возраст или этнос, что делает её предвзятой [1].

Для оценки справедливости применяются специализированные метрики, которые фокусируются на равенстве предсказаний для разных групп:

– Demographic Parity Difference (DPD): измеряет разницу в вероятности положительного предсказания между группами, выделенными по чувствительным признакам. Низкое значение DPD указывает на равномерное распределение предсказаний, минимизируя дискриминацию;

– Equalized Odds Difference (EOD): сравнивает показатели полноты (True Positive Rate, доля правильно классифицированных положительных примеров) и ложноположительной частоты (False Positive Rate, доля ошибочно классифицированных отрицательных примеров) между группами. Эта метрика оценивает, насколько модель равномерно ошибается для разных групп;

– Equal Opportunity Difference (EOpD): концентрируется на различиях в полноте для положительного класса, что особенно важно в задачах, где правильное распознавание положительных случаев имеет высокую значимость.

Эти метрики выявляют предвзятость, которая может оставаться скрытой при использовании только стандартных показателей качества [7]. Компромисс между точностью и справедливостью возникает, когда оптимизация для максимальной точности приводит к усилению предвзятости, отдавая предпочтение доминирующим классам или группам [1]. Например, модель, оптимизированная для высокой точности, может игнорировать редкие сценарии или меньшинства, увеличивая значения DPD, EOD или EOpD [8]. Этот компромисс особенно проблематичен в социально чувствительных приложениях, где несправедливые решения могут иметь серьёзные последствия [9].

Предвзятость в моделях обусловлена недостатками обучающих выборок:

– дисбаланс классов: неравное представительство категорий, из-за чего модель фокусируется на доминирующих классах, игнорируя редкие, но значимые сценарии;

– недостаток разнообразия: ограниченный набор примеров снижает способность модели к обобщению, особенно на данных, отличающихся от обучающей выборки;

– шумы и аномалии: ошибочные или некорректные данные искажают процесс обучения, усиливая предвзятость;

– корреляции с чувствительными признаками: переменные, напрямую или косвенно связанные с такими характеристиками, как пол или возраст, могут усиливать дискриминацию, даже если эти признаки явно исключены из модели.

Для минимизации предвзятости и достижения баланса между точностью и справедливостью необходимы системные подходы к оптимизации обучающих выборок, включающие предобработку данных, балансировку классов, отбор признаков и выбор подходящих моделей [5].

Стратегии оптимизации обучающих выборок

Оптимизация обучающих выборок направлена на устранение недостатков данных, повышение их репрезентативности и обеспечение устойчивости моделей. Рассмотрим основные стратегии и их влияние на компромисс между точностью и справедливостью [2].

1. Предобработка данных.

В предобработку данных включается преобразование поступающих в модель данных для повышения качества обучения и максимально возможное устранение причин предвзятости. Данный этап крайне важен, так как данные без обработки часто содержат пропуски, шумы или неравномерное распределение, которые негативно влияют на обучение [6]. Ключевые методы предобработки включают:

– базовая нормализация: заполнение пропущенных значений (например, медианой для числовых признаков или модой для категориальных) и кодирование категориальных переменных. One-hot-encoding создаёт бинарные переменные для каждой категории, что подходит для моделей, не работающих с категориальными данными напрямую, таких как логистическая регрессия. Label encoding присваивает числовые метки категориям, что полезно для моделей, работающих с порядковыми данными, например, деревьев решений. Базовая нормализация делает данные стабильнее и улучшает процесс обучения, но не устраняет корреляции с чувствительными признаками, что может сохранять предвзятость, особенно если исходные данные содержат скрытые зависимости [5];

– минимакс-нормализация: масштабирование числовых признаков в диапазоне [0, 1]. Позволяет использовать данные с алгоритмами, которые чувствительны к масштабу, как нейронные сети, метод опорных векторов (SVM) или логистическая регрессия. Этот метод повышает сходимость моделей, снижая влияние различий в масштабах признаков, и нередко повышает точность [2]. Однако минимакс-нормализация может повысить предвзятость в случае, если данные содержат признаки, связанные с чувствительными характеристиками (пол или возраст), поскольку масштабирование не устраняет эти зависимости [6];

– снижение размерности (PCA, t-SNE): Метод главных компонент (PCA) преобразует данные в новое пространство признаков, сохраняя максимальное распределение и сокращая размерность. Это позволяет исключить избыточные или предвзятые признаки, снижая вероятность переобучения и улучшает объективность. PCA особенно эффективен в задачах с большим числом признаков, где связь может быть

скрытой. t-SNE (t-distributed Stochastic Neighbor Embedding) используется для выявления кластеров в данных, что помогает обнаружить аномалии или несбалансированные группы. Однако снижение размерности приводит к потере информации, что снижает точность, особенно в случае с высокоразмерными данными;

– кодирование с учётом целевой переменной (Target Encoding): заменяет категориальные признаки средними значениями целевой переменной для каждой категории. Target Encoding повышает точность за счёт учёта зависимостей между признаками и целевой переменной, но может усиливать корреляции с чувствительными признаками, если они связаны с целевой переменной, что снижает справедливость;

– аугментация данных: генерация дополнительного объёма данных, который повышает разнообразие и улучшает обобщающую способность моделей. При обработке текста аугментация может включать замену слов синонимами [3], а в случае с обработкой сигналов – добавление шума. Аугментация помогает моделям лучше обрабатывать редкие классы, но некорректная реализация может, наоборот, усилить предвзятость, если синтетические данные наследуют корреляции или искажения.

Предобработка подразумевает тщательный выбор методов, опираясь на текущую задачу. PCA и t-SNE наиболее эффективны для повышения справедливости за счёт исключения избыточных признаков, тогда как минимакс-нормализация и Target Encoding лучше подходят для задач, где важна максимальная точность. Комбинирование методов, таких как нормализация и снижение размерности, часто даёт оптимальный баланс [2].

2. Балансировка классов.

Дисбаланс классов является одной из основных причин предвзятости, особенно в случаях, где редкие классы имеют высокую значимость. Модели часто отдают предпочтение преобладающим классам, игнорируя редкие, в случаях, когда баланс между категориями сильно нарушен. Это снижает объективность и способность к обобщению. Ключевые методы балансировки включают:

– SMOTE (Synthetic Minority Oversampling Technique): генерация синтетических примеров для малочисленного класса и внесение данных между существующими точками в пространстве признаков. SMOTE увеличивает представленность редких классов, улучшая способность модели распознавать их, что повышает полноту и общую точность. Однако синтетические данные могут усиливать корреляции с чувствительными признаками, если они присутствуют в исходной выборке, что снижает справедливость. Кроме того, SMOTE может создавать нереалистичные примеры в областях с низкой плотностью данных [6];

– RandomUnderSampler: случайно сокращает количество примеров преобладающего класса до уровня малочисленного, выравнивая распределение классов. RandomUnderSampler прост в реализации и эффективно снижает предвзятость, так как модель перестаёт отдавать предпочтение определенным классам [2]. Однако удаление данных может привести к потере большого объёма информации, особенно в малых выборках, что снижает точность;

– ADASYN (Adaptive Synthetic Sampling): улучшенная версия SMOTE, которая фокусируется на генерации синтетических примеров вблизи сложных для классификации областей, где малочисленный класс плохо отделяется от главенствующего. ADASYN адаптивно определяет количество сгенерированных примеров для каждой точки, что повышает точность для редких классов. Однако, как и SMOTE, ADASYN может усиливать предвзятость, если данные содержат зависимости с чувствительными признаками, и требует осторожного применения, чтобы избежать переобучения [6];

– Tomek Links: этот метод удаляет пары примеров (один из преобладающего класса, другой из малого), которые ближе всего друг к другу в пространстве признаков. Удаление таких «пограничных» примеров повышает чёткость разделения классов, улучшая объективность модели. Tomek Links эффективен при умеренном дисбалансе, но может быть недостаточным при сильном, так как удаляет лишь ограниченное число примеров [5];

– отсутствие балансировки: сохранение исходного распределения классов оправдано, если данные естественно сбалансированы или редкие классы не имеют критической значимости. Однако в большинстве случаев это приводит к игнорированию миноритарных классов, усиливая предвзятость [1].

Выбор метода балансировки зависит от объёма данных, степени дисбаланса и приоритетов задачи. SMOTE и ADASYN подходят для повышения точности в задачах с редкими классами, тогда как RandomUnderSampler и Tomek Links эффективны для минимизации предвзятости. Комбинирование методов, например, SMOTE с Tomek Links, может обеспечить баланс между точностью и объективностью.

3. Отбор признаков.

Удаление избыточных или смещенных переменных является основным направлением отбора признаков, что повышает точность и справедливость моделей. Этот шаг особенно важен в задачах с чувствительными признаками, где корреляции могут усиливать дискриминацию [1]. Далее приведены основные методы:

– SelectKBest: метод выбирает указанное количество наиболее значимых признаков на основе статистических показателей, таких как взаимная информация или хи-квадрат. Например, выбор 10 лучших признаков позволяет сосредоточиться на переменных, наиболее тесно связанных с целевой переменной, исключая те, которые косвенно коррелируют с чувствительными признаками, такими как пол или возраст. SelectKBest прост в реализации, эффективен для повышения точности и справедливости, но требует правильного выбора метрики, чтобы избежать потери важной информации;

– VarianceThreshold: удаление признаков с низкой дисперсией (почти постоянной) снижает избыточность данных, упрощая модель и снижая риск переобучения. Этот метод эффективен для больших выборок со множеством признаков, но менее полезен для устранения предвзятости, поскольку не учитывает корреляции с целевой переменной или чувствительными признаками;

– Recursive Feature Elimination (RFE): метод последовательно удаляет наименее значимые признаки, используя базовую модель (например, логистическую регрессию или SVM) для оценки их важности. RFE улучшает интерпретируемость модели, сохраняя только наиболее релевантные переменные, и может уменьшить предвзятость, устраняя коррелированные признаки. Однако RFE является вычислительно затратным, особенно для больших выборок, и его эффективность зависит от качества базовой модели;

– L1-регуляризация (Lasso): применение L1-регуляризации к линейным моделям, таким как логистическая регрессия, устанавливает веса неинформативных признаков на ноль, выполняя выбор признаков во время обучения. Lasso эффективен для задач с большим числом признаков, поскольку одновременно снижает сложность модели и убирает предвзятые переменные [1]. Но он может быть менее точным, если данные содержат сложные или нелинейные связи;

– SHAP-значения (SHapley Additive exPlanations): метод оценивает вклад каждого признака в прогнозы модели, позволяя идентифицировать переменные, вызывающие предвзятость. Например, признаки, сильно влияющие на предсказания для определённых групп, могут быть исключены для повышения справедливости [8]. SHAP

обеспечивает глубокое понимание влияния признаков, но требует значительных вычислительных ресурсов, особенно для больших моделей.

Отбор признаков имеет решающее значение для задач, где справедливость является приоритетом [7]. SelectKBest и SHAP эффективно исключают предвзятые переменные, тогда как Lasso и RFE подходят для задач с большим числом признаков. Комбинирование методов, таких как SelectKBest с SHAP, позволяет достичь оптимального баланса между точностью и объективностью [2].

Выбор моделей и их влияние

Выбор модели МО существенно влияет на компромисс между точностью и справедливостью. Разные алгоритмы имеют уникальные характеристики, которые определяют их устойчивость к предвзятости и способность к обобщению [6]. Рассмотрим основные модели:

– логистическая регрессия: эта линейная модель предполагает линейную зависимость между признаками и целевой переменной, что делает её простой и интерпретируемой. Логистическая регрессия устойчива к шуму и хорошо подходит для задач, где важна справедливость, так как её предсказания легко анализировать [1]. Однако её способность фиксировать сложные нелинейные связи ограничена, что может снизить точность при решении задач большой размерности. Предварительная обработка, такая как нормализация и отбор признаков, значительно повышает её производительность;

– дерево решений: этот нелинейный алгоритм создает иерархические вычислительные алгоритмы путем разделения данных на основе результатов. Деревья решений устойчивы к выбросам и категориальным данным, что делает их пригодными для неоднородных выборок. Однако они подвержены переобучению, особенно при глубокой структуре. Балансировка классов и ограничение глубины дерева помогают повысить справедливость;

– случайный лес: ансамблевый метод, который объединяет несколько деревьев решений, построенных на основе случайной выборки данных и признаков. Случайный лес обеспечивает высокую точность и надежность прогнозов из-за усреднения, что делает его подходящим для решения сложных задач. Однако он может привести к предвзятости, если данные содержат коррелированные признаки или классовые несоответствия. Выбор и балансировка элементов улучшают их точность;

– градиентный бустинг: ансамблевый метод, который постоянно тренирует деревья, исправляя предыдущие ошибки. Такие алгоритмы, как XGBoost или LightGBM, обеспечивают высокую точность и гибкость, но чувствительны к дисбалансам и шуму в данных. Градиентный бустинг требует тщательной настройки гиперпараметров и оптимизации выбора для минимизации искажений. Использование SHAP-значений помогает интерпретировать его прогнозы и устранять предвзятые признаки [8];

– нейронные сети: глубокие модели, состоящие из множества слоёв нейронов, способны улавливать сложные нелинейные зависимости, что делает их мощным инструментом для задач с высокоразмерными данными [4]. Однако сложность делает их склонными к переобучению и усилению предвзятости, особенно если данные содержат искажения. Нейронные сети требуют большого объёма данных, тщательной предобработки и регуляризации (например, dropout) для достижения объективности;

– SVM (Support Vector Machines): метод опорных векторов строит гиперплоскость, максимально разделяющую классы в пространстве признаков. SVM эффективен для задач с чёткой разделимостью классов и устойчив к шуму, но чувствителен к масштабу данных и дисбалансу. Использование ядерных функций

(например, RBF-ядра) позволяет улавливать нелинейные зависимости, но увеличивает вычислительную сложность. Нормализация и балансировка данных критически важны для SVM.

Каждая модель требует адаптации обучающей выборки. Логистическая регрессия и SVM выигрывают от нормализации и отбора признаков, случайный лес и градиентный бустинг устойчивы к необработанным данным, но чувствительны к дисбалансу, а нейронные сети требуют аугментации и больших объёмов данных.

Заключение

Оптимизация обучающих выборок является ключевым инструментом для достижения баланса между точностью и справедливостью в алгоритмах ИИ. Методы предобработки (PCA, минимакс-нормализация, аугментация), балансировки классов (SMOTE, RandomUnderSampler, ADASYN), отбора признаков (SelectKBest, SHAP) и выбор моделей (логистическая регрессия, случайный лес, нейронные сети) позволяют минимизировать предвзятость, сохраняя высокую производительность. Интеграция метрик справедливости, таких как DPD и EOD, и визуализация компромиссов помогают разработчикам создавать устойчивые и этичные системы. Дальнейшие исследования в области адаптивных методов, новых метрик и глубокого обучения расширят возможности справедливого МО, делая ИИ более ответственным и соответствующим этическим стандартам.

СПИСОК ИСТОЧНИКОВ

1. Харитонова Ю.С. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права / Ю.С. Харитонова, В.С. Савина, Ф. Паньини // Вестник Пермского университета. Юридические науки. – 2021. – № 53. – С. 488–515.

2. Рюмина Е.В. Сравнительный анализ методов устранения дисбаланса классов эмоций в видеоданных выражений лиц / Е.В. Рюмина, А.А. Карпов // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20, № 5. – С. 683–691.

3. Преображенский А.П. О проблемах нейросетевого анализа текстовой информации / А.П. Преображенский, Д.В. Меняйлов // Радиоэлектронные устройства и системы для инфокоммуникационных технологий («РЭУС-2022»): доклады Всероссийской конференции (с международным участием): Выпуск: LXXVII. – Москва: Российское научно-техническое общество радиотехники, электроники и связи им. А.С. Попова, 2022. – С. 155–159.

4. Питолин А.В. Прогнозирование характеристик рассеяния объектов сложной формы на основе нейросетевых технологий / А.В. Питолин, Р.П. Юров, А.П. Преображенский // Антенны. – 2007. – № 8 (123). – С. 59–61.

5. Балалаева Ю.С. «AI bias»: «предвзятость искусственного интеллекта» или результат деятельности разработчиков? / Ю.С. Балалаева // Юристъ-Правоведъ. – 2023. – № 3 (106). – С. 7–14.

6. Справедливость и предвзятость в машинном обучении // Политическое образование [Электронный ресурс]. – URL: https://lawinrussia.ru/spravedlivost-i-predvzyatost-v-mashinnom-obuchenii/ (дата обращения: 03.06.2025).

7. Caton S., Haas Ch. Fairness in Machine Learning: A Survey // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2010.04053 (дата обращения: 03.06.2025).

8. Hort M., Chen Zh., Zhang J.M., Harman M., Sarro F. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2207.07068 (дата обращения: 03.06.2025).

9. Bias in Machine Learning: Identifying, Mitigating, and Preventing Discrimination // GeeksforGeeks [Электронный ресурс]. – URL: https://www.geeksforgeeks.org/bias-in-machine-learning-identifying-mitigating-and-preventing-discrimination/ (дата обращения: 03.06.2025).

ИНФОРМАЦИЯ ОБ АВТОРЕ

Коновалов Кирилл Дмитриевич, студент, Воронежский институт высоких технологий, Воронеж, Россия.
e-mail: llyriq@yandex.ru