УДК 004.89

РАЗРАБОТКА АЛГОРИТМОВ ВЫЯВЛЕНИЯ СКРЫТОЙ ПРЕДВЗЯТОСТИ В ОБУЧЕННЫХ МОДЕЛЯХ

Шмидт Г., студент направления Artificial Intelligence and Algorithms
MSc programme, II курс
Технический университет Дании, Копенгаген, Дания

Одной из самых актуальных задач в области машинного обучения и искусственного интеллекта является выявление и устранение скрытой предвзятости в обученных моделях [1].

Скрытая предвзятость может проявляться в разных формах: от искажений в обучающих данных до несоответствия выводов модели реальным условиям или этическим стандартам. Ее наличие не только снижает эффективность работы модели, но и может иметь серьезные последствия для пользователей, включая дискриминацию, несправедливость, нарушение приватности и доверия к системам ИИ. Поэтому разработка алгоритмов для обнаружения скрытой предвзятости в обученных моделях становится важным этапом в создании безопасных, этичных и справедливых искусственных интеллектов.

Процесс выявления предвзятости в моделях может быть многоэтапным и включать различные подходы, направленные на анализ как самих данных, так и поведения модели. Один из самых очевидных способов – это проверка исходных данных на наличие предвзятости. Это может быть сделано через статистические тесты, которые анализируют распределение целевых переменных по категориям, представляющим различные демографические или социальные группы. Например, в задачах классификации можно использовать методы анализа распределения классов по возрасту, полу, расовой принадлежности или другим социально значимым признакам, чтобы убедиться, что данные не содержат систематических искажений в сторону одних категорий и не ущемляют права других.

Однако сам факт, что данные могут быть сбалансированы или очищены от очевидных искажений, не гарантирует отсутствие предвзятости в модели. Даже в случае, когда данные оказываются статистически корректными, скрытая предвзятость может появляться в процессе обучения модели. Это связано с тем, что алгоритмы машинного обучения могут "захватывать" закономерности и зависимости в данных, которые являются результатом исторических и социальных факторов, в том числе тех, которые могут проявляться как несправедливые или неэтичные в контексте реальных приложений [2].

Например, модель, обученная на исторических данных о кредите, может научиться выявлять предвзятость в отношении определённых групп за счёт исторической дискриминации, даже если в самих данных это не будет явно выражено.

Алгоритмы выявления предвзятости в обученных моделях нацелены на обнаружение таких скрытых зависимостей. Для этого используется ряд подходов, таких как оценка параметров модели (например, коэффициентов в линейных моделях или важности признаков в деревьях решений) с точки зрения справедливости. Некоторые методы включают в себя тесты на справедливость, например, анализ разницы в точности предсказаний для различных групп, представленных в обучающих данных. Примеры таких тестов – это измерение демографической паритета, где проверяется, являются ли ошибки модели одинаковыми для разных демографических групп, или проверка на предвзятость в отношениях между переменными, где исследуется, не зависят ли важные выводы модели от признаков, которые могут быть связаны с предвзятостью.

Кроме того, одним из ключевых методов выявления предвзятости является использование методов из области интерпретируемого машинного обучения, таких как методы объяснения важности признаков (например, через значения Шепли или LIME), которые позволяют анализировать, как каждый признак влияет на принятие решения моделью, и выявлять признаки, которые могут быть связаны с предвзятостью. Визуализация важных признаков и их распределения по различным категориям данных помогает понять, какие аспекты данных или какие характеристики системы могут способствовать нечестным или неэтичным выводам [3].

Методы регуляризации и штрафов также играют важную роль в разработке алгоритмов для выявления и устранения предвзятости. Регуляризация может быть использована не только для уменьшения переобучения модели, но и для управления её поведением с точки зрения справедливости. Например, с помощью методов, таких как штрафы за использование предвзятых признаков, можно в процессе обучения модели минимизировать её склонность к решению, которое связано с несправедливыми зависимостями. Это позволяет строить модели, которые не только точны, но и обладают более этичными предсказаниями.

Другим подходом является использование контрфактических методов, которые помогают оценить поведение модели, если бы данные изменялись, чтобы устранить предвзятость. Например, метод контрфактических примеров позволяет моделям проверять, как бы они поступили в случае, если бы информация о группе, на которую могла быть направлена предвзятость, была бы другой. Это даёт возможность выявить участки данных, где модель может демонстрировать явные или скрытые искажения.

Для обеспечения максимальной точности и минимизации предвзятости алгоритм выявления предвзятости также должен включать процессы мониторинга в реальном времени, чтобы после внедрения модели в продакшн можно было постоянно отслеживать её поведение и выявлять случаи, когда модель начинает показывать нежелательные или неточные результаты, вызванные изменениями в данных или новых зависимостях, которые могли бы появиться после её развертывания.

Важным аспектом работы с предвзятостью является не только её выявление, но и удаление или минимизация этого эффекта. Для этого существует ряд методов, включая переподборку обучающих данных, создание справедливых взвешенных потерь (loss functions), а также улучшение качества данных, например, через алгоритмы балансировки классов или репрезентативности в выборках [4].

В заключение, разработка алгоритмов для выявления скрытой предвзятости в обученных моделях является необходимым элементом создания справедливых, этичных и доверенных искусственных интеллектов.

Это не только улучшает качество работы моделей, но и способствует созданию систем, которые могут быть использованы в реальных приложениях с высокой степенью уверенности в их безопасности и объективности.

Сложность задачи заключается в том, чтобы идентифицировать и устранить скрытые предвзятости, сохраняя при этом высокое качество предсказаний и доверие к системе, что требует комплексного подхода и постоянного совершенствования методов и инструментов [5].

Список литературы:

1. Mehrabi, N. A Survey on Bias and Fairness in Machine Learning / N. Mehrabi, F. Morstatter, N. Saxena [et al.] // ACM Computing Surveys. – 2021. – Vol. 54, № 6. – Article 115. – DOI 10.1145/3457607.

2. Suresh, H. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle / H. Suresh, J. V. Guttag // Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. – New York : ACM, 2021. – Article 17. – DOI 10.1145/3465416.3483305.

3. Buolamwini, J. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification / J. Buolamwini, T. Gebru // Proceedings of Machine Learning Research. – 2018. – Vol. 81. – P. 77–91.

4. Ntoutsi, E. Bias in Data-Driven Artificial Intelligence Systems - An Introductory Survey / E. Ntoutsi, P. Fafalios, U. Gadiraju [et al.] // WIREs Data Mining and Knowledge Discovery. – 2020. – Vol. 10, № 3. – e1356. – DOI 10.1002/widm.1356.

5. Olteanu, A. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries / A. Olteanu, C. Castillo, F. Diaz, E. Kıcıman // Frontiers in Big Data. – 2019. – Vol. 2. – Article 13. – DOI 10.3389/fdata.2019.00013.