УДК 004.89

МЕТОДЫ ОБНАРУЖЕНИЯ ДИСКРИМИНАЦИИ ПО ЧУВСТВИТЕЛЬНЫМ ПРИЗНАКАМ В МАШИННОМ ОБУЧЕНИИ

Парк Э., студент направления Industrial Production Engineering
Bachelor’s degree program, II курс
Миланский технический университет, Милан, Италия

Одной из наиболее серьёзных проблем в области машинного обучения и искусственного интеллекта является дискриминация, связанная с чувствительными признаками, такими как пол, возраст, раса, этническая принадлежность, религия и другие характеристики, которые могут не только быть юридически и этически неприемлемыми, но и существенно снижать доверие к системам ИИ. В процессе обучения моделей, особенно когда речь идёт о больших данных, важно контролировать, чтобы алгоритмы не наследовали предвзятости, которые могут быть скрытыми или незаметными на этапе подготовки данных, но могут оказывать серьёзное влияние на конечные решения модели. Поэтому методы обнаружения дискриминации по чувствительным признакам становятся ключевым элементом в обеспечении справедливости, прозрачности и этичности в искусственном интеллекте [1].

Дискриминация по чувствительным признакам происходит тогда, когда модель или система принимает решения, основанные на признаках, которые представляют собой или могут служить индикаторами дискриминации. Такие признаки могут включать явные (например, пол или расовую принадлежность) и неявные (например, почтовый индекс, который может быть связан с социальным статусом или этнической группой) характеристики. В некоторых случаях эти признаки могут не напрямую влиять на предсказания, но их наличие в данных может привести к созданию моделей, которые воспроизводят исторические предвзятости или дискриминационные практики.

Одним из первых подходов к решению этой проблемы является мониторинг на уровне данных. При этом анализируется, насколько данные, используемые для обучения модели, сбалансированы и корректны с точки зрения присутствия чувствительных признаков. Это включает в себя проверку распределений категориальных переменных и количественных признаков по различным социальным группам, а также анализ их возможных взаимосвязей с целевой переменной. К примеру, в задаче кредитного скоринга важно убедиться, что переменные, такие как уровень дохода или образование, не оказываются косвенно

связанными с признаками, такими как расовая или этническая принадлежность, что может привести к несправедливым решениям.

Для более глубокого анализа дискриминации используются методы статистического тестирования и измерения справедливости. Одним из таких подходов является оценка демографической паритета, когда проверяется, что ошибки модели (например, ложные отрицания или ложные положительные результаты) одинаково распределены по различным группам, не приводя к систематическим ошибкам для какой-либо из категорий. Такой тест позволяет выявить скрытые предвзятости, даже если признаки, по которым происходит дискриминация, не очевидны [2].

Другим важным методом является проверка на равенство шансов (equal opportunity), которая анализирует, обеспечивают ли модели одинаковые шансы для различных групп на достижение правильных предсказаний. Это особенно важно в контексте задач, связанных с принятиями решений, влияющих на жизни людей, таких как медицина, правосудие и трудоустройство. Если модель слишком часто делает ошибки для определённой группы (например, для женщин или людей определённой расы), это может свидетельствовать о наличии скрытой дискриминации.

Для обнаружения дискриминации и предвзятости активно используются методы из области интерпретируемого машинного обучения. Например, с помощью алгоритмов интерпретации можно анализировать важность признаков, чтобы выяснить, оказывает ли какой-либо из чувствительных признаков значительное влияние на принятие решения модели. Это может быть выполнено с использованием различных техник, таких как анализ важности признаков через значения Шепли, LIME или SHAP, которые позволяют точно локализовать вклад каждого признака в предсказания модели и выявить те, которые могут быть связаны с дискриминационными тенденциями.

Также широко применяются методы изменения модели и данных, направленные на устранение предвзятости. Одним из таких методов является обучение модели с учётом справедливости, при котором в процессе оптимизации целевой функции модели добавляются штрафы за дискриминацию. Например, можно использовать регуляризацию, которая поощряет модели за равенство ошибок между различными группами. Это позволяет одновременно минимизировать предвзятость и сохранить высокое качество предсказаний. В некоторых случаях можно использовать методы, такие как генерация контрфактических данных, когда с помощью специально созданных примеров, в которых чувствительные признаки искусственно изменяются, исследуется, как это влияет на результаты модели [3].

Одним из наиболее инновационных подходов к устранению дискриминации является трансформация данных перед обучением. Методы, такие как

репаративная генерация данных, позволяют изменить данные так, чтобы чувствительные признаки не влияли на решение модели, или чтобы влияние этих признаков было сведено к минимуму. Например, может быть проведена обработка признаков таким образом, чтобы сохранить всю важную информацию для предсказания, но исключить информацию, которая может быть использована для дискриминации.

Ещё одним инструментом является переподборка данных – когда в процессе подготовки данных используется балансировка классов, что позволяет избежать ситуаций, когда одни группы преобладают в данных, а другие оказываются недопредставленными, что может привести к искажению решений модели. Это особенно важно в случаях, когда определённые демографические группы представлены в данных менее чем другие, что может привести к снижению точности предсказаний для этих групп [4].

Важно отметить, что все эти методы и подходы не являются универсальными и требуют индивидуальной настройки в зависимости от конкретной задачи и области применения. Обнаружение и устранение дискриминации по чувствительным признакам является не только технической, но и этической задачей, и требует комплексного подхода, включая взаимодействие специалистов по данным, юристов, этиков и других экспертов.

Таким образом, методы обнаружения дискриминации по чувствительным признакам в машинном обучении играют ключевую роль в обеспечении справедливости и прозрачности ИИ-систем.

Разработка и внедрение таких методов помогают не только повысить точность и надёжность моделей, но и обеспечить их соответствие высоким этическим стандартам, что особенно важно в сферах, затрагивающих права и интересы людей [5].

Список литературы:

1. Kusner, M. J. Counterfactual Fairness / M. J. Kusner, J. Loftus, C. Russell, R. Silva // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4066–4076.

2. Zemel, R. Learning Fair Representations / R. Zemel, Y. Wu, K. Swersky [et al.] // Proceedings of the 30th International Conference on Machine Learning. – 2013. – Vol. 28. – P. 325–333.

3. Pedreschi, D. Discrimination-Aware Data Mining / D. Pedreschi, S. Ruggieri, F. Turini // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2008. – P. 560–568. – DOI 10.1145/1401890.1401959.

4. Kamishima, T. Fairness-Aware Classifier with Prejudice Remover Regularizer / T. Kamishima, S. Akaho, H. Asoh, J. Sakuma // Machine Learning and

Knowledge Discovery in Databases. – Berlin : Springer, 2012. – P. 35–50. – DOI 10.1007/978-3-642-33486-3_3.

5. Luong, B. Discovering Discrimination in Ranking Systems / B. Luong, S. Ruggieri, F. Turini // Proceedings of the VLDB Endowment. – 2011. – Vol. 5, № 2. – P. 109–120. – DOI 10.14778/2078331.2078336.