УДК 004.89

ПОСТРОЕНИЕ ТЕСТОВ НА ПРЕДВЗЯТОСТЬ В УСЛОВИЯХ НЕПОЛНЫХ ИЛИ ЗАШУМЛЁННЫХ ДАННЫХ

Мохамед А.А., студент гр. 03-220, II курс
Арабская нефтегазовая академия, г. Абу-Даби

Построение тестов на предвзятость в условиях неполных или зашумленных данных является одной из самых сложных и важных задач в области машинного обучения, особенно когда речь идет о применении моделей в реальных условиях. В реальной жизни данные часто оказываются неполными, зашумленными или неструктурированными, что создает дополнительные трудности при оценке их качества, а также при разработке эффективных и справедливых алгоритмов [1].

В то же время, эти проблемы могут быть значительно усугублены существующей предвзятостью в данных, что приводит к несправедливым или неточным предсказаниям модели. В таких условиях необходимость создания тестов на предвзятость становится критической для обеспечения того, чтобы модели работали справедливо и эффективно для всех групп пользователей, несмотря на неполноту или шум в данных.

Проблема неполных и зашумленных данных заключается в том, что модель может начинать извлекать скрытые зависимости, которые на самом деле являются артефактами этих данных, а не реальными закономерностями. Например, если определённые признаки недоступны для части данных (например, информация о доходах в некоторых записях отсутствует), модель может начать полагаться на другие признаки, которые могут не быть связанными с реальным процессом.

В таких ситуациях важно точно понимать, насколько сильно модель зависит от чувствительных признаков, таких как возраст, пол или этническая принадлежность, и как эти признаки могут быть использованы для предвзятых решений [2].

Создание тестов на предвзятость в условиях неполных или зашумленных данных требует применения нескольких подходов и методов. Один из таких методов – это использование методов устойчивости модели. Устойчивость модели к неполным данным и зашумлённым признакам можно проверить, применяя техники, такие как кросс-валидация с пропусками (например, k-fold cross-validation), при которых для различных подмножеств данных случайным образом пропускаются части признаков или изменяются их значения, что позволяет

проверить, как сильно предвзятость или шум в данных влияют на итоговые результаты модели. Если модель продолжает демонстрировать признаки предвзятости при отсутствии некоторых признаков, это может свидетельствовать о том, что она использует ненадёжные или неэтичные зависимости для принятия решений.

Для построения тестов на предвзятость также используется анализ на основе контрфактических сценариев. В условиях неполных данных это позволяет оценить, как изменение одного или нескольких признаков (например, пола или расы) в недостающих частях данных влияет на предсказания модели. Контрфактические тесты могут выявить, зависит ли результат от отсутствующих признаков, и если да, то в какой степени это зависит от данных, которые, возможно, не были представлены в исходных обучающих наборах. Это позволяет выявить скрытые формы предвзятости, которые могут быть следствием неполных данных, но, тем не менее, оказывают влияние на конечный результат.

Кроме того, создание тестов на предвзятость требует применения методов регуляризации, которые помогают минимизировать зависимость модели от чувствительных признаков, даже если данные неполны или зашумлены. Регуляризация может включать в себя различные подходы, такие как L1/L2 регуляризация, которая способствует уменьшению весов признаков, не имеющих реального значения, или обучение с добавлением шума, что позволяет модели стать более устойчивой к неполноте или зашумлению в данных. Применение таких техник позволяет улучшить качество модели, минимизируя влияние признаков, которые могут быть искажены из-за неполноты данных [3].

Одним из основных вызовов при построении тестов на предвзятость в условиях зашумленных данных является способность модели выявлять важные зависимости среди множества ненадёжных и искажённых признаков. Для решения этой задачи могут быть использованы алгоритмы обработки зашумленных данных, такие как методы устойчивого обучения или методы фильтрации шума. Эти алгоритмы позволяют очистить данные от лишнего шума, при этом сохраняя их ценную информацию, что улучшает способность модели выявлять реальные закономерности и сводит к минимуму влияние шумных данных на предсказания.

Важным аспектом в построении таких тестов является оценка стабильности предсказаний при различных изменениях в обучающих данных.

Если модель, обученная на неполных или зашумленных данных, демонстрирует устойчивость к незначительным изменениям в данных, это может свидетельствовать о её более высокой справедливости и адекватности в реальных условиях. Однако если модель начинает делать значительные отклонения в предсказаниях при небольших изменениях в данных, это может быть

индикатором того, что она чрезмерно чувствительна к шуму в данных, и её предсказания могут быть ненадёжными или несправедливыми.

Для построения эффективных тестов на предвзятость также важно использовать методы анализа важности признаков. В условиях зашумленных данных это помогает понять, какие признаки модель считает наиболее важными для принятия решения, и как это соотносится с этическими принципами. Если модель чрезмерно полагается на чувствительные или потенциально дискриминирующие признаки, такие как пол, возраст или этническая принадлежность, это может указывать на наличие предвзятости, которую необходимо устранить. В таких случаях использование методов интерпретируемого ИИ, таких как SHAP или LIME, помогает выявить, какие именно признаки вносят наибольший вклад в предсказания модели и имеют ли они этическое обоснование.

Кроме того, важно учитывать, что периодический аудит модели и её поведения на новых, зашумлённых данных может помочь обеспечить долгосрочную корректность и этичность её работы. В реальных условиях данные постоянно обновляются, и изменения в входных данных могут повлиять на результаты предсказаний модели. Регулярные проверки и тестирование модели на зашумленных и неполных данных обеспечат её актуальность и корректность на протяжении всего жизненного цикла.

Таким образом, построение тестов на предвзятость в условиях неполных или зашумленных данных является важной частью разработки этичных и справедливых систем машинного обучения. Это требует применения различных методов, таких как регуляризация, контрфактический анализ, устойчивое обучение и аудиты, которые помогают выявить скрытые зависимости и минимизировать влияние нежелательных факторов на результаты предсказаний.

Важно помнить, что такие подходы обеспечивают не только повышение точности и стабильности моделей, но и гарантируют, что алгоритмы будут принимать справедливые решения, не зависимые от искажённых или неполных данных, что в свою очередь способствует улучшению доверия к системам машинного обучения в реальных приложениях.

Список литературы:

1. Quadrianto, N. Recycling Privileged Learning and Distribution Matching for Fairness / N. Quadrianto, V. Sharmanska // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 677–688.

2. Blum, A. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? / A. Blum, K. Stangl // Proceedings of the 1st Symposium on Foundations of Responsible Computing. – 2020. – P. 1–22.

3. Coston, A. Fair Transfer Learning with Missing Protected Attributes / A. Coston, A. Ramamurthy, D. Wei [et al.] // Proceedings of the AAAI/ACM

Conference on AI, Ethics, and Society. – 2019. – P. 91–98. – DOI 10.1145/3306618.3314236.