УДК 004.89

ИСПОЛЬЗОВАНИЕ ТЕСТОВ СПРАВЕДЛИВОСТИ ДЛЯ АНАЛИЗА ПОВЕДЕНИЯ КЛАССИФИКАТОРОВ

Задбоев В.А., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза
С. М. Буденного, г. Санкт-Петербург

Использование тестов справедливости для анализа поведения классификаторов является важным направлением в области машинного обучения и искусственного интеллекта, особенно в тех сферах, где решения, принимаемые моделями, могут иметь значительные последствия для пользователей. Применение классификаторов в таких областях, как медицина, финансовые услуги, правосудие или социальное обеспечение, требует не только высокой точности, но и справедливости в принятии решений. Тесты справедливости предоставляют важные инструменты для оценки, насколько объективны и беспристрастны модели, и помогают выявить скрытые предвзятости, которые могут привести к дискриминации или неравенству между различными социальными группами [1].

Прежде всего, следует отметить, что справедливость в контексте машинного обучения означает отсутствие предвзятости по отношению к определённым группам, основанным на чувствительных признаках, таких как пол, раса, возраст, этническая принадлежность, религия или другие социально значимые характеристики. Даже если модель демонстрирует высокую точность и минимальные ошибки в своих предсказаниях, она может иметь скрытые формы предвзятости, если её решения систематически отличаются для разных групп. Это может происходить даже в тех случаях, когда явная дискриминация не наблюдается в процессе обучения модели.

Для анализа справедливости классификаторов разрабатываются различные метрики справедливости, которые позволяют оценить, насколько равномерно и честно модель принимает решения относительно различных групп. Оценка справедливости может быть проведена на нескольких уровнях: на уровне предсказаний модели, на уровне её ошибок (ложных положительных и ложных отрицательных результатов) и на уровне параметров самой модели.

В этой связи тесты справедливости становятся мощным инструментом для проверки не только качества модели, но и её социальной приемлемости [2].

Одним из самых популярных тестов справедливости является демографический паритет (demographic parity), который предполагает, что

вероятность положительного исхода (например, одобрения кредита или принятия на работу) должна быть одинаковой для разных групп. Если для одной группы вероятность положительного исхода значительно выше, чем для другой, это может свидетельствовать о наличии предвзятости. Например, если кредитная модель систематически одобряет кредиты мужчинами, а женщинам – чаще отказывает, это будет являться нарушением принципа демографического паритета. При этом стоит отметить, что этот критерий подходит не ко всем типам задач и может быть не всегда применим в контекстах, где важно учитывать дополнительные параметры.

Другой важной метрикой является равенство возможностей (equal opportunity), которая оценивает, равны ли шансы разных групп на получение положительных исходов в задачах классификации. Этот тест особенно актуален в тех случаях, когда одна из групп может иметь более высокий риск ошибки при ложноположительных или ложноотрицательных предсказаниях. Например, если медицинская модель неправильно диагностирует заболевание у одной группы пациентов, тогда шансы на получение корректной диагностики для этой группы будут ниже, что также будет свидетельствовать о предвзятости модели.

Метрика равенства шансов (equalized odds) используется для анализа и устранения предвзятости, связанной с как ложноположительными, так и ложноотрицательными ошибками. В отличие от равенства возможностей, равенство шансов требует, чтобы ошибки модели были одинаково распределены для всех групп, независимо от того, какую ошибку она сделает. Этот подход помогает гарантировать, что модель не будет чрезмерно ошибаться для какой-либо из групп, что особенно важно в ситуациях, когда ошибки могут иметь серьёзные последствия [3].

Помимо этих традиционных метрик, существует и ряд более сложных методов справедливости, которые учитывают более тонкие аспекты работы классификаторов. Например, метрики, учитывающие перекрытие и коррелированность признаков, исследуют, как различные факторы могут взаимодействовать и приводить к искаженному результату для определённых групп. Такие метрики позволяют более глубоко понять, как различные параметры данных могут влиять на решение модели и выявить случаи, когда предвзятость проявляется через сложные взаимосвязи между признаками.

Для проведения анализа справедливости важно также учитывать интерпретируемость моделей, которая позволяет понять, какие именно признаки или данные влияют на предсказания модели. Методы интерпретации, такие как анализ важности признаков, LIME (Local Interpretable Model-agnostic Explanations) или SHAP (Shapley Additive Explanations), помогают понять, как каждый признак влияет на итоговое решение модели. Это может быть полезно

не только для проверки справедливости, но и для обнаружения скрытых форм предвзятости. Например, если модель, несмотря на высокие показатели точности, делает предсказания, сильно зависимые от определённых признаков, таких как расовая принадлежность или пол, это будет являться индикатором наличия предвзятости.

Кроме того, следует отметить важность контрфактического тестирования моделей, когда мы анализируем поведение модели на изменённых данных. Например, изменив только один или несколько признаков в данных, можно увидеть, как это повлияет на решения модели. Такой подход помогает выявить скрытые зависимости и предвзятости, которые могут не быть видны на уровне обычных тестов справедливости, но которые могут существенно искажать результаты модели в реальных приложениях [4].

Важно, что тесты справедливости в контексте классификаторов не всегда предполагают нахождение простых решений. Иногда справедливость может быть противопоставлена точности модели. Например, чтобы достичь демографического паритета, модель может потребовать значительных изменений в данных или алгоритмах, что может повлиять на её общую точность. В таких случаях важно находить баланс между достижением справедливости и сохранением высокого уровня точности и надёжности модели.

В заключение, использование тестов справедливости для анализа поведения классификаторов является ключевым шагом в обеспечении этичного и справедливого применения моделей машинного обучения. Эти тесты позволяют выявлять скрытые предвзятости, анализировать, как различные группы влияют на результаты модели, и обеспечивать равные шансы для всех категорий.

Современные метрики справедливости и методы интерпретируемости становятся важными инструментами для создания моделей, которые не только точны, но и соответствуют социальным, этическим и юридическим стандартам [5].

Список литературы:

1. Friedler, S. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning / S. A. Friedler, C. Scheidegger, S. Venkatasubramanian [et al.] // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 329–338. – DOI 10.1145/3287560.3287589.

2. Agarwal, A. A Reductions Approach to Fair Classification / A. Agarwal, A. Beygelzimer, M. Dudik [et al.] // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 60–69.

3. Zafar, M. B. Fairness Constraints: Mechanisms for Fair Classification / M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. – 2017. – Vol. 54. – P. 962–970.

4. Donini, M. Empirical Risk Minimization under Fairness Constraints / M. Donini, L. Oneto, S. Ben-David [et al.] // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 2791–2801.

5. Corbett-Davies, S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning / S. Corbett-Davies, S. Goel // arXiv preprint. – 2018. – arXiv:1808.00023.