

УДК 004.89

МЕТОДЫ СРАВНЕНИЯ ПРЕДСКАЗАНИЙ МОДЕЛЕЙ ДЛЯ ГРУПП С РАЗЛИЧНЫМИ ХАРАКТЕРИСТИКАМИ

Сычужников В.Б., старший научный сотрудник
Военная академия связи имени Маршала Советского Союза
С. М. Буденного, г. Санкт-Петербург

Методы сравнения предсказаний моделей для групп с различными характеристиками играют ключевую роль в обеспечении справедливости и прозрачности в применении машинного обучения, особенно когда речь идет о задачах, где предсказания могут существенно повлиять на жизнь людей. Такие задачи могут включать кредитный скоринг, принятие решений о трудоустройстве, судебные разбирательства, медицинскую диагностику и другие области, где ошибки могут привести к серьезным последствиям для отдельных людей или групп [1].

Основная цель методов сравнения предсказаний заключается в том, чтобы выявить, насколько модель принимает одинаковые решения для групп, отличающихся по определенным характеристикам, таким как пол, возраст, раса, этническая принадлежность, социальный статус и другие чувствительные признаки.

Одним из первых и наиболее распространенных методов в этом контексте является анализ ошибок модели. Ошибки модели, такие как ложноположительные и ложные отрицательные результаты, могут различаться для разных групп, что является важным индикатором предвзятости. Например, в задаче диагностики заболевания важно убедиться, что модель не ошибается чаще для определенных групп пациентов (например, для женщин или людей определенной расы).

Для этого проводится сравнение частоты ошибок модели между различными группами и анализируются возможные различия в точности предсказаний. Если ошибки модели распределяются неравномерно и приводят к систематическим сбоям для одной из групп, это может свидетельствовать о скрытой предвзятости или структурной дискриминации, которую необходимо устранить [2].

Другим важным методом является использование метрик справедливости для сравнения предсказаний для разных групп. Среди таких метрик можно выделить демографический паритет, который проверяет, одинаково ли распределяются положительные исходы для разных групп. Например, в задаче

кредитного скоринга важно, чтобы вероятность одобрения кредита была одинаковой для разных этнических групп или полов, если эти признаки не являются релевантными для принятия решения. Если эта метрика показывает значительные различия между группами, то это указывает на наличие предвзятости в модели, что требует её пересмотра и корректировки.

Равенство шансов (Equal Opportunity) и равенство ошибок (Equalized Odds) также являются важными метриками для анализа и сравнения предсказаний. Эти метрики исследуют, насколько одинаковы ошибки модели для разных групп. Например, в задачах правосудия или здравоохранения важно, чтобы ошибка ложного отрицания (когда модель неправильно классифицирует позитивный случай как негативный) была одинаково низкой для всех групп. Если одна из групп имеет более высокую вероятность ложных отрицаний, это может свидетельствовать о том, что модель несправедливо обрабатывает эту группу, что требует дополнительных усилий по её корректировке.

Контрфактические тесты также играют важную роль в анализе различий в предсказаниях для групп с различными характеристиками. Эти тесты позволяют проверить, как изменение одного или нескольких признаков влияет на результаты модели. Например, если у модели наблюдается заметное различие в предсказаниях для мужчин и женщин при условии, что все остальные признаки остаются неизменными, это может быть индикатором того, что модель использует чувствительный признак (например, пол) для принятия решения. В таком случае можно применить контрфактическое изменение (например, заменить пол мужчины на женский в наборе данных), чтобы проверить, изменится ли результат модели. Если предсказания изменяются, это может свидетельствовать о предвзятости модели, и необходимо пересмотреть её обучение и архитектуру [3].

Для более глубокого анализа различий в предсказаниях моделей используются методы визуализации и анализ важности признаков. С помощью этих инструментов можно наглядно представить, какие признаки или комбинации признаков имеют наибольшее влияние на принятие решений модели. Например, с помощью методов визуализации внимания (для моделей на основе трансформеров) можно увидеть, как модель фокусируется на различных частях входных данных, и проверить, использует ли она чувствительные признаки (например, расу или возраст) для принятия решений. Это может помочь выявить скрытые формы предвзятости, которые не всегда очевидны при стандартных тестах.

Также стоит отметить важность адаптивного тестирования моделей в реальных условиях. Для этого применяются методы, которые могут адаптировать модель к изменениям в данных или условиях эксплуатации, обеспечивая справедливость предсказаний для различных групп. Например, в задачах прогнозирования, где данные постоянно обновляются, важно использовать методы,

которые позволяют модели автоматически учитывать изменения в демографической структуре или других важных факторах. Это требует создания системы, которая будет регулярно анализировать производительность модели для различных групп и в случае выявления предвзятости – корректировать алгоритмы или данные.

Для оценки устойчивости модели к предвзятости также применяются статистические методы, такие как проверка гипотез и анализ дисперсии. Эти методы позволяют определить, существует ли статистически значимая разница в предсказаниях для разных групп, и, если такая разница существует, на каком уровне её можно минимизировать [4].

Например, в задаче, связанной с трудоустройством, важно провести тестирование модели для различных групп кандидатов и выявить, существуют ли значимые различия в вероятности получения работы для разных этнических групп, мужчин и женщин.

Кроме того, для улучшения сравнения предсказаний для разных групп, активно исследуются методы балансировки данных, которые обеспечивают более равномерное представление различных категорий в обучающем наборе данных. Это особенно важно, когда одна из групп в данных недопредставлена, что может привести к перекосу в результатах модели. В таких случаях применяются техники, такие как переподборка выборки, синтетическое добавление данных или перераспределение весов, которые помогают обеспечить более справедливое обучение модели и снизить её предвзятость.

В результате, использование методов сравнения предсказаний моделей для групп с различными характеристиками является важным шагом на пути к созданию справедливых, этичных и прозрачных систем машинного обучения. Эти методы не только помогают улучшить точность моделей, но и обеспечивают их соответствие стандартам справедливости, что особенно важно в тех областях, где решения, принятые ИИ-системами, могут напрямую повлиять на судьбы людей [5].

Список литературы:

1. Feldman, M. Certifying and Removing Disparate Impact / M. Feldman, S. A. Friedler, J. Moeller [et al.] // Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2015. – P. 259–268. – DOI 10.1145/2783258.2783311.
2. Kearns, M. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness / M. Kearns, S. Neel, A. Roth, Z. S. Wu // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 2564–2572.
3. Hebert-Johnson, U. Multicalibration: Calibration for the (Computationally-Identifiable) Masses / U. Hebert-Johnson, M. Kim, O. Reingold, G. Rothblum //

Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 1939–1948.

4. Williamson, R. C. Fairness Risk Measures / R. C. Williamson, A. K. Menon // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 6786–6797.

5. Diana, E. Minimax Group Fairness: Algorithms and Experiments / E. Diana, W. Gill, M. Kearns [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Vol. 35, № 8. – P. 6993–7001.