УДК 004.89

МЕТОДЫ ОБНАРУЖЕНИЯ ПЕРЕКОСОВ В РАСПРЕДЕЛЕНИИ ОШИБОК МЕЖДУ ГРУППАМИ ПОЛЬЗОВАТЕЛЕЙ

Угву Ч.К., студент гр. BIT15, II курс
Университет Кади Айяд, Марокко, г. Марракеш

Методы обнаружения перекосов в распределении ошибок между группами пользователей играют ключевую роль в обеспечении справедливости и прозрачности алгоритмических систем, особенно в тех областях, где решения, принимаемые на основе машинного обучения, могут существенно повлиять на жизни людей. Перекосы в распределении ошибок могут возникать, когда модель систематически ошибается чаще для одной группы пользователей, чем для другой, что может привести к дискриминации или несправедливому обращению. Такие ситуации могут проявляться, например, в задачах кредитования, здравоохранения, правосудия или трудоустройства, где отклонения в предсказаниях могут не только ухудшить результаты для отдельных индивидов, но и усиливать социальное неравенство [1].

Основной задачей в данном контексте является выявление и анализ таких перекосов в ошибках модели, чтобы понять, как различные группы пользователей могут быть неравномерно затронуты ошибками, которые модель делает. В первую очередь, важно различать два типа ошибок: ложноположительные и ложные отрицания. Ложноположительная ошибка возникает, когда модель ошибочно классифицирует отрицательный исход как положительный, а ложная отрицательная ошибка – наоборот, когда модель классифицирует положительный исход как отрицательный. Различие в распределении этих ошибок между группами пользователей может свидетельствовать о наличии перекосов и предвзятости в модели.

Для анализа перекосов в ошибках между группами пользователей часто применяются метрики справедливости, такие как демографический паритет и равенство ошибок. Демографический паритет оценивает, насколько одинаково модель распределяет положительные исходы между разными группами. Если, например, одна группа пользователей (по признаку пола, расы или социального статуса) получает положительный результат (например, одобрение кредита) реже, чем другая группа, это может указывать на предвзятость модели. Однако на практике этой метрики бывает недостаточно, так как она не учитывает различия в ошибках между группами, что делает её менее чувствительной к более тонким формам предвзятости [2].

Равенство ошибок (Equalized Odds) является более точной метрикой для выявления перекосов, поскольку она оценивает, насколько равномерно распределяются ошибки между группами. Если для одной группы модель имеет более высокую вероятность как ложных положительных, так и ложных отрицательных результатов, это является важным индикатором предвзятости. Для задач с высокими рисками, таких как медицинская диагностика, такие различия могут быть особенно опасными, так как они могут привести к нежелательным последствиям для одной из групп. Например, если модель чаще ошибается в диагнозах для женщин или представителей этнических меньшинств, это может привести к неверным медицинским решениям, что является серьёзным нарушением принципов справедливости и этики.

Методы вычисления и сравнения ошибок на разных подгруппах данных играют важную роль в выявлении таких перекосов. Для этого данные могут быть разделены на группы, исходя из демографических характеристик, таких как возраст, пол, этническая принадлежность, социальный статус и другие чувствительные признаки. Каждая из этих групп анализируется на предмет различий в распределении ошибок, и если для одной из групп наблюдается значительное увеличение ошибок (например, ложных положительных или ложных отрицательных), это является сигналом для дальнейшего анализа и коррекции модели. Таким образом, важно не только вычислять общую точность модели, но и проводить погрупповой анализ ошибок, чтобы точно понять, как модель работает для разных категорий пользователей.

Для более детального анализа перекосов в распределении ошибок используются графические методы. Например, можно построить ROC-кривую (Receiver Operating Characteristic curve) для каждой группы пользователей и сравнить её с другими. Если кривые для разных групп сильно различаются, это может свидетельствовать о наличии перекоса в ошибках. Также применяются гистограммы ошибок или тепловые карты, которые наглядно показывают, как модель ошибается для различных подгрупп данных. Визуализация помогает не только выявить перекосы, но и лучше понять, какие признаки оказывают влияние на решение модели [3].

Одним из методов, который активно используется для борьбы с перекосами, является коррекция данных на этапе обучения модели. Когда выявляются значительные различия в ошибках между группами, можно применить подходы, такие как переподборка выборки или перераспределение весов, чтобы сбалансировать данные и минимизировать влияние чувствительных признаков, таких как пол, возраст или раса. Этот метод предполагает, что в обучающем наборе данных будут учтены все группы в равной степени, что позволяет модели быть более справедливой при принятии решений для всех категорий пользователей.

Кроме того, важным инструментом является регуляризация модели, которая направлена на уменьшение её склонности к переподгонке (overfitting) под одну группу данных, в ущерб другим. Регуляризация помогает модели стать менее чувствительной к небольшим изменениям в данных, что снижает вероятность возникновения перекосов в ошибках.

Кроме работы с данными, для устранения перекосов в распределении ошибок можно использовать методы постобработки. После того как модель обучена, можно провести её пересмотр и провести калибровку вероятностей для разных групп, чтобы сделать ошибки более равномерными. Это позволяет привести к более сбалансированным предсказаниям, уменьшая различия между группами в распределении ошибок [4].

Не менее важным шагом является мониторинг модели на этапе её эксплуатации. Модели, которые успешно работают в условиях обучающих данных, могут демонстрировать различные результаты в реальных условиях, когда сталкиваются с новыми, ранее не встречавшимися данными. Важно постоянно отслеживать поведение модели, чтобы выявить, не происходит ли новых перекосов в распределении ошибок с течением времени, и оперативно исправлять возможные сбои.

В заключение, методы обнаружения перекосов в распределении ошибок между группами пользователей являются важным шагом в обеспечении справедливости и этичности моделей машинного обучения. Анализ ошибок, использование метрик справедливости, коррекция данных и регуляризация – все эти подходы помогают создавать модели, которые работают справедливо для всех групп пользователей, минимизируя риск дискриминации и укрепляя доверие к алгоритмическим системам. В конечном итоге, такой подход способствует созданию более этичных и справедливых технологий, которые могут быть использованы для улучшения качества жизни людей, без создания дополнительных барьеров или неравенства [5].

Список литературы:

1. Woodworth, B. Learning Non-Discriminatory Predictors / B. Woodworth, S. Gunasekar, M. I. Ohannessian, N. Srebro // Proceedings of the 2017 Conference on Learning Theory. – 2017. – Vol. 65. – P. 1920–1953.

2. Hashimoto, T. Fairness without Demographics in Repeated Loss Minimization / T. Hashimoto, M. Srivastava, H. Namkoong, P. Liang // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 1929–1938.

3. Dutta, S. An Information-Theoretic Quantification of Discrimination with Exempt Features / S. Dutta, P. Venkatesh, P. Mardziel [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Vol. 34, № 3. – P. 3825–3833. – DOI 10.1609/aaai.v34i03.5811.

4. Kim, M. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification / M. Kim, O. Reingold, G. Rothblum // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. – 2019. – P. 247–254. – DOI 10.1145/3306618.3314287.

5. Kallus, N. Residual Unfairness in Fair Machine Learning from Prejudiced Data / N. Kallus, A. Zhou // Proceedings of the 37th International Conference on Machine Learning. – 2020. – Vol. 119. – P. 2439–2448.