УДК 004.89

МЕТОДЫ ДИАГНОСТИКИ И ВИЗУАЛИЗАЦИИ ИСТОЧНИКОВ ПРЕДВЗЯТОСТИ В СЛОЖНЫХ МОДЕЛЯХ

Лобода Л.Д., студент гр. ИТб-212, IV курс
Научный руководитель: Кулак И. В., старший преподаватель
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

Методы диагностики и визуализации источников предвзятости в сложных моделях машинного обучения представляют собой одну из наиболее актуальных и значимых задач в области искусственного интеллекта, особенно в контексте обеспечения справедливости и прозрачности алгоритмических решений [1].

Сложные модели, такие как глубокие нейронные сети и ансамбли деревьев решений, становятся всё более распространёнными в практических приложениях, от медицины до финансов и правосудия. Однако одна из ключевых проблем заключается в их недостаточной интерпретируемости и трудности в выявлении факторов, которые могут привести к несправедливым или предвзятым результатам.

В этом контексте диагностика и визуализация источников предвзятости становятся не только необходимыми, но и фундаментальными для развития этичных систем машинного обучения.

Одной из первых задач является определение, что именно является источником предвзятости в модели. Предвзятость может возникать на разных уровнях: начиная с самой модели, через алгоритмы обучения, до исходных данных, на которых она обучается. Основными источниками предвзятости могут быть, например, непропорциональное представление определённых групп в обучающих данных, использование чувствительных признаков (пол, возраст, этническая принадлежность) или даже алгоритмические особенности, такие как неправильная настройка гиперпараметров модели. Методы диагностики должны быть направлены на то, чтобы эффективно выявлять и изолировать эти источники.

Одним из самых распространённых подходов для диагностики предвзятости является анализ важности признаков. В сложных моделях машинного обучения определение, какие признаки играют наибольшую роль в принятии решения, может помочь понять, откуда именно исходит предвзятость. Например, если модель принимает решения, основываясь на определённых

признаках, которые могут быть косвенно связаны с чувствительными атрибутами, такими как социальный статус или этническая группа, это может привести к предвзятости. Для этого можно использовать такие методы, как SHAP (Shapley Additive Explanations) и LIME(Local Interpretable Model-agnostic Explanations), которые помогают выявить влияние каждого признака на результат модели, а также дают возможность понять, какие именно признаки влияют на предсказания для различных групп [2].

Другим важным инструментом диагностики является контрфактический анализ, который позволяет определить, как изменение одного или нескольких признаков влияет на предсказания модели.

Например, если предсказания модели изменяются при изменении полового признака, это может свидетельствовать о наличии предвзятости, и контрфактический анализ поможет это выявить.

В случае с чувствительными признаками, такими как раса или возраст, важно, чтобы изменения этих признаков не приводили к значительным изменениям в предсказаниях модели, что является показателем её справедливости.

Одним из методов визуализации предвзятости является построение тепловых карт важности признаков, где можно наглядно показать, какие области данных или какие признаки наиболее сильно влияли на принятие решения моделью. Визуализация таких карт может помочь обнаружить «горячие точки», где модель начинает делать ошибки или показывать признаки предвзятости. Тепловые карты могут быть полезными для визуализации зависимостей между входными признаками и выходами модели, что особенно важно при работе с высокоразмерными данными. Такие визуализации могут помочь исследователям и разработчикам быстро и наглядно оценить, какие именно признаки приводят к проблемам с точностью или справедливостью [3].

Также важно использовать методы снижения размерности, такие как PCA (Principal Component Analysis) или t-SNE (t-Distributed Stochastic Neighbor Embedding), для того чтобы выявить скрытые паттерны и структуры в данных, которые могут быть источником предвзятости. Эти методы позволяют преобразовать многомерные данные в двумерные или трёхмерные пространства, что даёт возможность визуализировать данные и легко заметить аномалии или сбои в поведении модели. При этом визуализация в низкоразмерном пространстве может помочь в выявлении скрытых зависимостей между признаками, которые способствуют несправедливости.

Другим важным аспектом является мониторинг модели после её внедрения в реальное использование. Даже если на стадии разработки модели не были выявлены признаки предвзятости, важно продолжать её тестирование и анализ в процессе эксплуатации, поскольку изменения в данных или окружении могут привести к появлению новых источников предвзятости. Для этого могут быть

использованы методы онлайн-мониторинга и обратной связи, которые позволяют отслеживать, как модель ведет себя в реальных условиях и выявлять новые проблемы, связанные с несправедливостью.

Для более глубокого анализа предвзятости важно применять методы контрфактической справедливости, которые исследуют, как модель реагирует на изменения в чувствительных признаках и как это влияет на её предсказания [4].

Например, при анализе данных о здоровье модели, которые используют такие признаки, как пол или возраст, можно изучить, как предсказания изменяются при контрфактическом изменении этих признаков. Это позволяет не только выявить скрытые формы предвзятости, но и понять, насколько модель зависит от этих признаков для принятия решений.

Кроме того, для улучшения диагностики и визуализации предвзятости важным шагом является использование методов "объясняемого ИИ" (XAI), которые помогают разработать более интерпретируемые модели и дают возможность пользователям понять, как именно модель принимает свои решения. Применяя такие методы, можно не только выявить предвзятость, но и минимизировать её, делая модели более прозрачными и доступными для анализа.

Подводя итог, методы диагностики и визуализации источников предвзятости в сложных моделях машинного обучения являются необходимым инструментом для создания этичных и справедливых систем. Эти методы помогают выявить скрытые формы предвзятости, обусловленные как данными, так и алгоритмами, и предложить подходы для их устранения.

Использование контрфактического анализа, интерпретируемых моделей и методов визуализации важности признаков позволяет эффективно диагностировать и устранить предвзятость, что в свою очередь способствует повышению доверия к моделям и их результатам, особенно в таких ответственных сферах, как здравоохранение, финансы, правосудие и другие [5].

Список литературы:

1. Goldstein, A. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation / A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin // Journal of Computational and Graphical Statistics. – 2015. – Vol. 24, № 1. – P. 44–65. – DOI 10.1080/10618600.2014.907095.

2. Apley, D. W. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models / D. W. Apley, J. Zhu // Journal of the Royal Statistical Society: Series B. – 2020. – Vol. 82, № 4. – P. 1059–1086. – DOI 10.1111/rssb.12377.

3. Greenwell, B. M. pdp: An R Package for Constructing Partial Dependence Plots / B. M. Greenwell // The R Journal. – 2017. – Vol. 9, № 1. – P. 421–436. – DOI 10.32614/RJ-2017-016.

4. Cabrera, A. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning / A. Cabrera, W. Epperson, F. Hohman [et al.] // IEEE Conference on Visual Analytics Science and Technology. – 2019. – P. 46–56. – DOI 10.1109/VAST47406.2019.8986948.

5. Cai, C. J. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making / C. J. Cai, S. Winter, D. Steiner [et al.] // Proceedings of the CHI Conference on Human Factors in Computing Systems. – 2019. – P. 1–14. – DOI 10.1145/3290605.3300234.