

УДК 004.89

АЛГОРИТМЫ ПОИСКА ЛОКАЛЬНЫХ ОБЛАСТЕЙ НЕСПРАВЕДЛИВОСТИ В ВЫСОКОРАЗМЕРНЫХ ПРОСТРАНСТВАХ

Герман К.О., студент гр. ИТб-212, IV курс
Научный руководитель: Асанов С. А., старший преподаватель
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

Алгоритмы поиска локальных областей несправедливости в высокоразмерных пространствах являются ключевым инструментом для обнаружения скрытых и часто незаметных предвзятостей в сложных моделях машинного обучения, работающих с многомерными данными. Сложность подобных моделей возрастает с увеличением числа признаков, что делает их трудными для анализа и интерпретации, особенно когда необходимо выявить несправедливость или предвзятость [1].

В высокоразмерных пространствах возможна ситуация, когда модель работает хорошо в большинстве случаев, но в локальных областях данных проявляется несправедливость, которая может оказывать пагубное влияние на принятие решений. Алгоритмы поиска таких областей играют важную роль в устранении этих предвзятостей и обеспечении справедливости в процессе принятия решений.

Одним из основных вызовов работы с высокоразмерными данными является то, что с увеличением числа признаков пространство решений становится экспоненциально более сложным. Это приводит к так называемому проклятию размерности, когда данные теряют свою интуитивную интерпретируемость и становятся трудными для анализа. В таких условиях традиционные методы поиска несправедливости или предвзятости, которые эффективны в более низкоразмерных пространствах, оказываются менее применимыми.

Высокоразмерные пространства требуют новых подходов, которые могут эффективно обнаруживать локальные аномалии, даже если в глобальном контексте модель демонстрирует хорошую производительность [2].

Алгоритмы поиска локальных областей несправедливости могут использовать несколько подходов для решения этой проблемы. Во-первых, можно использовать методы локальной интерпретируемости, такие как LIME (Local Interpretable Model-agnostic Explanations) или SHAP (Shapley Additive Explanations), которые позволяют интерпретировать поведение модели в локальных областях данных. Эти методы помогают понять, какие признаки

вносят наибольший вклад в предсказания модели, и, если один или несколько чувствительных признаков (например, пол, раса или возраст) оказывают значительное влияние, это может быть индикатором предвзятости в локальных областях. Эти методы дают возможность исследовать поведение модели в малых подмножествах данных, что помогает находить несправедливости, которые не видны в общей картине.

Однако для эффективного поиска локальных областей несправедливости в высокоразмерных пространствах требуется использование более специализированных методов, таких как кластеризация и снижение размерности. В высокоразмерных данных многие важные связи и закономерности могут быть скрыты из-за сложности взаимодействий между признаками. Методы снижения размерности, такие как PCA (Principal Component Analysis) или t-SNE (t-Distributed Stochastic Neighbor Embedding), могут помочь уменьшить количество признаков, выделяя наиболее важные компоненты, что позволяет лучше визуализировать данные и выявить области, где модель может демонстрировать несправедливое поведение. Эти методы особенно полезны для поиска таких областей, где классическое моделирование или анализ может не обнаружить явных аномалий.

Ключевым моментом в поиске локальных областей несправедливости является метод локального поиска аномалий, который позволяет выделить подмножества данных, для которых модель показывает нестабильные или неоправданно предвзятые результаты. Для этого могут быть использованы такие алгоритмы, как DBSCAN (Density-Based Spatial Clustering of Applications with Noise), который позволяет найти плотные области данных и аномалии в этих областях, или методы, основанные на определении расстояний, такие как k-nearest neighbors (k-NN), которые помогают найти схожие объекты, но с различными результатами предсказания [3].

Например, если два похожих пациента с идентичными медицинскими данными получают разные результаты в зависимости от их этнической принадлежности, это может указывать на предвзятость модели в локальной области.

Ещё одним подходом является использование многокритериальных методов для анализа данных, которые позволяют учитывать не только саму ошибку модели, но и её последствия для различных групп. В таких методах оценка справедливости производится с учётом множества факторов, таких как последствия ошибок для разных этнических, социальных или возрастных групп. Это помогает лучше понять, какие локальные области данных оказываются несправедливыми и в какой степени модель может быть предвзята для разных категорий пользователей.

Для дальнейшего улучшения поиска локальных областей несправедливости может быть использовано усиленное обучение с дополнительной

регуляцией. Например, можно обучить модель, которая будет стараться минимизировать не только стандартные потери, но и показатели несправедливости на локальных подмножествах данных. Это даст возможность модели не только предсказывать правильные результаты, но и повышать свою справедливость, особенно в тех областях, где она может демонстрировать недостаточную производительность или предвзятость. Такой подход будет полезен для обучения с учётом справедливости, при котором модель получает обратную связь не только о качестве своих предсказаний, но и о том, как её предсказания влияют на разные группы пользователей.

Методы поиска локальных областей несправедливости также требуют использования гибридных подходов, которые совмещают несколько моделей для анализа. Например, в некоторых случаях можно использовать комбинацию линейных моделей для оценки влияния отдельных признаков и более сложных нелинейных моделей, таких как деревья решений или нейросети, для анализа взаимосвязей между признаками. Это даёт возможность повысить точность и эффективность поиска локальных аномалий и несправедливости, поскольку разные типы моделей могут быть лучше настроены на разные аспекты данных и их характеристики.

В завершение, поиск локальных областей несправедливости в высокоразмерных пространствах является важной задачей для обеспечения справедливости и этичности в системах машинного обучения.

Эффективное использование методов кластеризации, снижения размерности, локального анализа аномалий, а также многокритериальных подходов позволяет выявить скрытые предвзятости и уменьшить их влияние на итоговые решения. Это не только повышает качество предсказаний, но и способствует созданию более справедливых и ответственных алгоритмических систем, что особенно важно в таких областях, как здравоохранение, финансы и правосудие, где ошибки могут иметь значительные последствия для людей.

Список литературы:

1. Jabbari, S. Fairness in Reinforcement Learning / S. Jabbari, M. Joseph, M. Kearns [et al.] // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 1617–1626.
2. Kearns, M. An Empirical Study of Rich Subgroup Fairness for Machine Learning / M. Kearns, S. Neel, A. Roth, Z. S. Wu // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 100–109. – DOI 10.1145/3287560.3287592.
3. Zhang, M. Fairness-Aware Subgroup Discovery / M. Zhang, S. Neill, D. B. Neill // Proceedings of the 2021 SIAM International Conference on Data Mining. – Philadelphia : SIAM, 2021. – P. 271–279. – DOI 10.1137/1.9781611976700.31.