УДК 004.89

ИНСТРУМЕНТЫ ИНТЕРПРЕТАЦИИ В ЗАДАЧАХ С ПЕРЕКРЫВАЮЩИМИСЯ И КОРРЕЛИРОВАННЫМИ ПРИЗНАКАМИ

Поехалов К.В., студент гр. 602, II курс
Университет ИТМО, г. Санкт-Петербург

В задачах машинного обучения и анализа данных часто приходится сталкиваться с особенностями исходных признаков, которые могут значительно усложнять процесс интерпретации моделей и их предсказаний. Одной из таких особенностей является наличие перекрывающихся и сильно коррелированных признаков, характерных для многих прикладных областей – от биомедицины и финансов до промышленного анализа и обработки текстов.

Эти корреляции и взаимозависимости создают дополнительные сложности при попытках выделить индивидуальный вклад каждого признака в итоговое решение модели, что негативно сказывается на прозрачности, объяснимости и доверии к системе.

В связи с этим разработка специализированных инструментов интерпретации, способных адекватно работать в условиях мультиколлинеарности и высокой взаимосвязанности признаков, является одной из актуальных и важных задач в области объяснимого искусственного интеллекта.

Традиционные методы интерпретации, такие как оценка важности признаков с помощью градиентов, пермутационной значимости или значений Шепли, часто основываются на предположении о независимости признаков. В реальности же перекрывающиеся и коррелированные признаки приводят к тому, что один и тот же эффект может быть «разделён» между несколькими признаками, что усложняет адекватное определение их индивидуальной роли [1].

Это затрудняет не только локальную интерпретацию конкретных предсказаний, но и глобальный анализ модели в целом. Без учёта корреляций можно получить вводящие в заблуждение или неполные объяснения, что особенно критично в сферах с высокими требованиями к прозрачности и безопасности [2].

Одним из подходов к решению этой проблемы является разработка и применение методов, учитывающих структуру взаимосвязей между признаками.

Например, расширенные методы на основе значений Шепли, адаптированные для коррелированных данных, используют специальные алгоритмы для распределения вклада признаков с учётом их совместного влияния и корреляций. Такие методы позволяют более точно оценить, какой вклад вносит

конкретный признак, не игнорируя при этом сложные взаимодействия. Кроме того, для снижения размерности и устранения избыточности часто применяются техники предварительной обработки – факторный анализ, методы главных компонент (PCA), кластеризация признаков, которые помогают выявить группы взаимосвязанных переменных и работать с ними как с обобщёнными признаками [3].

Еще одним перспективным направлением являются инструменты, комбинирующие интерпретацию с визуальным анализом корреляций и кластеров признаков.

Интерактивные визуализации, такие как матрицы корреляций, дендрограммы кластеризации, тепловые карты и параллельные координаты, позволяют аналитикам и экспертам по предметной области понять структуру данных, выделить группы схожих или взаимозависимых признаков и сопоставить их с результатами модели. В таких системах можно не только увидеть, какие признаки наиболее влияют на предсказания, но и как они взаимодействуют между собой, что существенно повышает качество и глубину интерпретации.

В задачах с высокой размерностью и сложными корреляциями эффективными оказываются методы локальной интерпретации, основанные на обучении упрощённых моделей в локальных окрестностях входных данных. Такие локальные модели, например, локальные линейные аппроксимации или деревья решений, позволяют «распутывать» сложные взаимосвязи и выявлять релевантные признаки в конкретном контексте, где влияние коррелированных переменных становится более однозначным. Это даёт возможность получать более понятные и контекстуализированные объяснения для отдельных объектов и ситуаций.

Особое внимание уделяется разработке алгоритмов, способных выявлять и устранять эффект мультиколлинеарности при формировании интерпретаций. Среди таких методов – регуляризация с учётом корреляций, инкапсуляция взаимозависимых признаков в новые комплексные переменные, а также обучение специальных объяснительных моделей, чувствительных к структуре данных. Это позволяет не только повысить качество интерпретаций, но и сделать их более стабильными и воспроизводимыми при изменениях данных и переобучении моделей.

Кроме того, современные исследования направлены на интеграцию инструментов интерпретации с методами аудита и тестирования моделей на предвзятость, устойчивость и справедливость, которые особенно актуальны в условиях коррелированных признаков. Корреляции могут скрывать системные ошибки или неявные дискриминационные эффекты, которые традиционные методы интерпретации не выявляют. Использование комплексных

аналитических платформ с поддержкой корреляционных структур данных помогает обнаруживать такие проблемы и своевременно их устранять.

С технической стороны разработка инструментов интерпретации в условиях перекрывающихся и коррелированных признаков реализуется с применением продвинутых библиотек и фреймворков, таких как SHAP, LIME, Alibi, которые расширяются и модифицируются для учёта мультиколлинеарности. Важную роль играют также визуализационные библиотеки и интерактивные платформы, позволяющие объединять численные оценки важности с наглядными представлениями данных и взаимосвязей.

Таким образом, инструменты интерпретации в задачах с перекрывающимися и коррелированными признаками представляют собой сложный, но крайне важный класс методов и средств, обеспечивающих глубокое и достоверное понимание поведения моделей машинного обучения [4].

Их развитие способствует повышению прозрачности, устойчивости и доверия к ИИ-системам, особенно в тех областях, где точность и ответственность принятия решений имеют критическое значение [5].

Список литературы:

1. Strobl, C. Bias in Random Forest Variable Importance Measures : Illustrations, Sources and a Solution / C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn // BMC Bioinformatics. – 2007. – Vol. 8. – Article 25. – DOI 10.1186/1471-2105-8-25.

2. Strobl, C. Conditional Variable Importance for Random Forests / C. Strobl, A.-L. Boulesteix, T. Kneib [et al.] // BMC Bioinformatics. – 2008. – Vol. 9. – Article 307. – DOI 10.1186/1471-2105-9-307.

3. Gregorutti, B. Correlation and Variable Importance in Random Forests / B. Gregorutti, B. Michel, P. Saint-Pierre // Statistics and Computing. – 2017. – Vol. 27. – P. 659–678. – DOI 10.1007/s11222-016-9646-1.

4. Nicodemus, K. K. The Behaviour of Random Forest Permutation-Based Variable Importance Measures under Predictor Correlation / K. K. Nicodemus, J. D. Malley, C. Strobl, A. Ziegler // BMC Bioinformatics. – 2010. – Vol. 11. – Article 110. – DOI 10.1186/1471-2105-11-110.

5. Song, E. Shapley Effects for Global Sensitivity Analysis : Theory and Computation / E. Song, B. L. Nelson, J. Staum // SIAM/ASA Journal on Uncertainty Quantification. – 2016. – Vol. 4, № 1. – P. 1060–1083. – DOI 10.1137/15M1048070.