УДК 004.89

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОДХОДОВ К ИНТЕРПРЕТАЦИИ РЕШЕНИЙ СЛОЖНЫХ МОДЕЛЕЙ

Некрасова В.С., студент гр. ИУ12-21М, III курс
Московский государственный технический университет имени Н. Э. Баумана,
г. Москва

В последние годы модели машинного обучения достигли беспрецедентных успехов в решении сложных задач, охватывающих широкий спектр областей – от компьютерного зрения и обработки естественного языка до медицины и финансов. При этом растёт и важность не просто высокой точности, но и возможности интерпретировать и понимать решения, которые эти модели принимают. В особенности это актуально для сложных моделей, таких как глубокие нейронные сети, ансамбли, трансформеры и другие гибридные архитектуры, чья внутренняя логика зачастую остаётся «чёрным ящиком». Для обеспечения доверия пользователей, выполнения нормативных требований и поддержки принятия обоснованных решений в прикладных системах возникает необходимость сравнения и оценки различных подходов к интерпретации таких моделей [1].

Интерпретация решений сложных моделей – это совокупность методов, направленных на объяснение, каким образом входные данные преобразуются в конкретный выход. При этом подходы к интерпретации существенно различаются по нескольким ключевым аспектам: уровню локализации объяснений (глобальные vs локальные), модели взаимодействия (модель-зависимые vs модель-агностичные), способу представления результата (числовые метрики, визуализации, текстовые отчёты) и по степени формализации. Важным критерием является и степень сохранения оригинальной функциональности модели при интерпретации, а также возможность масштабирования на большие объёмы данных и сложные архитектуры.

Одним из наиболее распространённых и понятных классов являются методы локальной интерпретации, которые концентрируются на объяснении конкретного предсказания для отдельного объекта. Такие методы, например LIME и SHAP, позволяют понять, какие признаки или входные элементы внесли наибольший вклад в итоговое решение. Они обладают преимуществом интуитивной понятности, возможности универсального применения к любым моделям и наглядной визуализации. Однако локальные методы страдают от недостатков – их вычислительная стоимость часто высока, объяснения могут быть

нестабильны при малейших изменениях данных, а также они не всегда отражают глобальную логику модели.

В отличие от них, глобальные методы интерпретации стремятся охарактеризовать поведение модели в целом. К таким методам относятся анализ важности признаков на уровне всей выборки, построение упрощённых моделей-заместителей (surrogate models), например решающих деревьев или линейных моделей, а также визуализация внутренних структур модели – например, карты активации в сверточных сетях или визуализация attention в трансформерах.

Глобальные методы позволяют выявить общие закономерности, понять ключевые признаки и оценить степень обобщения модели, но, как правило, уступают локальным методам в детализации и специфичности объяснений для отдельных случаев.

Отдельное направление занимают модель-зависимые методы интерпретации, которые используют специфику внутренней структуры модели для извлечения объяснений. Это, например, методы градиентного анализа (Integrated Gradients, DeepLIFT), визуализация attention-механизмов, обратное распространение активаций, а также специализированные техники для ансамблей и графовых нейросетей. Такие методы зачастую обеспечивают более точные и «глубокие» объяснения, но требуют доступа к внутренностям модели и могут быть неприменимы к закрытым или проприетарным системам.

В противоположность им, модель-агностичные методы (model-agnostic) работают поверх модели, рассматривая её как «чёрный ящик». Они включают в себя, помимо LIME и SHAP, методы пермутационной оценки важности признаков, локальное приближение функций, генерацию контрфактических объяснений и другие. Главным преимуществом является универсальность и независимость от архитектуры, но они могут страдать от сниженной точности, высокой вычислительной нагрузки и ограниченной интерпретируемости в сложных сценариях [2].

Кроме классических подходов, всё более актуальными становятся гибридные методы, которые комбинируют преимущества различных техник для достижения лучшего баланса между точностью, стабильностью и понятностью. Например, объединение локальных объяснений с глобальной визуализацией, внедрение доменной семантики и онтологий для обогащения интерпретаций, использование интерактивных инструментов, позволяющих пользователю исследовать и модифицировать объяснения в режиме реального времени [3].

Также стоит отметить важность метрик качества интерпретации, которые позволяют объективно сравнивать разные подходы. К ним относятся устойчивость и стабильность объяснений при изменении входных данных, согласованность с другими методами, когнитивная доступность, полнота объяснения и

вычислительная эффективность. Наличие таких метрик помогает не только выбрать оптимальный инструмент для конкретной задачи, но и развивать стандарты и практики построения доверенных систем ИИ [4].

Важным элементом сравнения выступает и масштабируемость методов, особенно в условиях растущих объёмов данных и сложности моделей. Например, некоторые методы, обладающие высокой точностью, оказываются непригодными для работы в режиме реального времени или для больших производственных систем. В таких случаях приходится искать компромиссы между скоростью и качеством интерпретации.

Практические кейсы демонстрируют, что выбор подхода к интерпретации зависит от множества факторов: области применения, типа модели, целей пользователя и контекста принятия решений. В медицине, например, ценится высокая точность и возможность локального объяснения с понятным языком, в финансовой сфере – соответствие нормативным требованиям и формальная проверяемость, в промышленности – оперативность и интеграция с мониторингом.

Таким образом, сравнительный анализ подходов к интерпретации решений сложных моделей не сводится к простому ранжированию «лучше-хуже», а представляет собой многомерную оценку, учитывающую разнообразие методологических, когнитивных и прикладных аспектов [5].

Продвижение в этой области требует комплексного подхода – развития новых методов, создания универсальных фреймворков, стандартизации оценок и глубокого взаимодействия между исследователями, разработчиками и конечными пользователями [6].

Только в таком контексте возможно построение действительно доверенных, прозрачных и эффективных систем машинного обучения, способных обеспечить не только качество, но и объяснимость своих решений [7].

Список литературы:

1. Guidotti, R. A Survey of Methods for Explaining Black Box Models / R. Guidotti, A. Monreale, S. Ruggieri [et al.] // ACM Computing Surveys. – 2018. – Vol. 51, № 5. – Article 93. – DOI 10.1145/3236009.

2. Doshi-Velez, F. Towards a Rigorous Science of Interpretable Machine Learning / F. Doshi-Velez, B. Kim // arXiv. – 2017. – arXiv:1702.08608.

3. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead / C. Rudin // Nature Machine Intelligence. – 2019. – Vol. 1. – P. 206–215. – DOI 10.1038/s42256-019-0048-x.

4. Lipton, Z. C. The Mythos of Model Interpretability / Z. C. Lipton // Queue. – 2018. – Vol. 16, № 3. – P. 31–57. – DOI 10.1145/3236386.3241340.

5. Barredo Arrieta, A. Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI / A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser [et al.] // Information Fusion. – 2020. – Vol. 58. – P. 82–115. – DOI 10.1016/j.inffus.2019.12.012.

6. Adadi, A. Peeking Inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI) / A. Adadi, M. Berrada // IEEE Access. – 2018. – Vol. 6. – P. 52138–52160. – DOI 10.1109/ACCESS.2018.2870052.

7. Samek, W. Explainable Artificial Intelligence : Understanding, Visualizing and Interpreting Deep Learning Models / W. Samek, G. Montavon, A. Vedaldi [et al.]. – Cham : Springer, 2019. – 435 p. – DOI 10.1007/978-3-030-28954-6.