УДК 004.89

ОБЪЯСНЕНИЕ ПОВЕДЕНИЯ АНСАМБЛЕВЫХ МОДЕЛЕЙ С ПОМОЩЬЮ АГРЕГИРОВАННЫХ МЕТРИК

Репьева Д.Н., студент гр. 24.М02-мкн, I курс
Санкт-Петербургский государственный университет, г. Санкт-Петербург

Ансамблевые методы машинного обучения – такие как случайный лес (Random Forest), градиентный бустинг, стохастические ансамбли нейронных сетей или байесовские комитеты – на протяжении последних десятилетий продемонстрировали выдающуюся эффективность во множестве задач, включая табличные данные, обработку изображений и текстов, а также анализ временных рядов. Их сила заключается в способности сочетать предсказания нескольких моделей, тем самым сглаживая индивидуальные ошибки и повышая устойчивость к переобучению [1].

Однако эта же структура, построенная на совокупности разнородных компонентов, делает интерпретацию ансамблевых моделей особенно сложной: поведение системы в целом может не совпадать с интуитивной логикой её отдельных элементов, а локальные решения – быть результатом комплексных взаимодействий между сотнями или тысячами базовых предсказателей [2].

В этом контексте объяснение поведения ансамблевых моделей требует использования агрегированных метрик, которые обеспечивают интерпретируемость без разрушения структурной сложности самой модели.

Основная задача в интерпретации ансамбля заключается в том, чтобы не просто определить важность отдельных признаков или выделить один из деревьев (или моделей), а в том, чтобы получить обобщённое, статистически устойчивое представление о работе всей совокупности, сохранив при этом информативность и когнитивную доступность. Агрегированные метрики позволяют решать эту задачу за счёт объединения локальных или компонентных объяснений в структурированные, обобщённые формы. Эти метрики могут быть как количественными (например, средняя важность признаков по всем моделям), так и качественными (например, частота появления признака в корневых узлах деревьев, или стабильность его вклада в разных регионах пространства признаков).

Наиболее широко используемой формой агрегированной интерпретации является оценка важности признаков на уровне ансамбля. Например, в случайном лесе или бустинге строится суммарная метрика, отражающая средний вклад каждого признака в уменьшение неопределённости (impurity) – будь то энтропия, джини или среднеквадратичная ошибка. Это даёт первичное

представление о структуре зависимости между входными переменными и результатом, и уже на этом уровне может быть построена простая модель объяснения – например, отранжированный список ключевых факторов, влияющих на итоговое решение. Однако такая оценка имеет глобальный характер и не отражает локальные вариации вклада признаков, а также не учитывает взаимодействия между ними.

Чтобы преодолеть эти ограничения, применяются агрегированные методы на основе SHAP (Shapley Additive Explanations). В частности, TreeSHAP – модификация метода Шепли, адаптированная под деревообразные ансамбли – позволяет вычислять вклад каждого признака для конкретного предсказания, с учётом всех возможных подмножеств других признаков. При этом значения SHAP могут быть агрегированы по множеству наблюдений, предоставляя как глобальную картину важности, так и контекстуализированную интерпретацию. Например, усреднение абсолютных значений SHAP по всей выборке позволяет выявить признаки с наибольшим стабильным вкладом в решение, в то время как анализ распределения значений SHAP по классам позволяет оценить селективность признака.

Агрегированные метрики могут быть полезны и для анализа взаимодействий признаков. Поскольку ансамбли автоматически выявляют нелинейные связи между признаками, то традиционные методы линейной интерпретации не улавливают этих эффектов. SHAP Interaction Values и производные от них метрики позволяют количественно оценить, как совместное присутствие двух признаков влияет на предсказание, в сравнении с их индивидуальными эффектами. Это позволяет выявить сложные паттерны, например: «признак A влияет только тогда, когда значение признака B выше определённого порога». Такие взаимодействия могут быть агрегированы в виде графов значимости, решающих таблиц или логических правил, обеспечивая интерпретируемость без упрощения модели до линейного приближения.

Кроме того, в практике анализа ансамблей используется агрегирование локальных объяснений, полученных на различных выборках или при разных конфигурациях модели. Это особенно полезно для оценки устойчивости интерпретации и выявления регионов пространства, где модель ведёт себя нестабильно [3].

Например, можно построить тепловую карту важности признаков по кластерам данных, тем самым выявив области, где объяснение радикально меняется. Это создаёт основу для более глубокой аналитики, например, при аудите предвзятости или проверке соответствия модели нормативным требованиям.

Для нейронных ансамблей или гибридных архитектур (например, когда объединяются сверточные сети с табличными моделями) используются агрегированные attention-метрики, визуализация согласованности градиентов или

плотности активаций, усреднённые карты значимости (saliency maps), а также ансамблевые оценки на основе контрфактического анализа. Такие методы позволяют формировать представление о доминирующих факторах в принятии решений ансамблем, при этом сохраняя нелинейную, высокоадаптивную структуру исходной модели.

Агрегированные интерпретации важны не только для понимания работы модели, но и для встроенного контроля качества, аудита и мониторинга. Например, если агрегированная метрика важности определённого признака со временем начинает резко меняться – это может указывать на дрейф данных или нарушение согласованности поведения модели. Сравнение агрегированных объяснений между тренировочной и продуктивной выборкой даёт механизм обнаружения критических отклонений. Более того, агрегированные интерпретации являются основой для формирования модел-карт (model cards) и документации для регуляторов, в которых необходимо кратко и формально описать, как и почему модель принимает решения [4].

Таким образом, объяснение поведения ансамблевых моделей с помощью агрегированных метрик представляет собой перспективное направление в области интерпретируемого машинного обучения. Оно позволяет сохранить сильные стороны ансамблей – точность, устойчивость, обобщающую способность – при этом предоставляя пользователям и разработчикам инструменты для системного и количественно обоснованного анализа внутренней логики модели [5].

Современные библиотеки, такие как SHAP, ELI5, InterpretML, LIT и Alibi, активно развивают соответствующую инфраструктуру, включая визуализацию, логгирование и автоматическую генерацию отчётов. Это делает агрегированную интерпретацию не просто удобным инструментом, но важным элементом инженерии доверенного ИИ.

Список литературы:

1. Chen, T. XGBoost : A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2016. – P. 785–794. – DOI 10.1145/2939672.2939785.

2. Prokhorenkova, L. CatBoost : Unbiased Boosting with Categorical Features / L. Prokhorenkova, G. Gusev, A. Vorobev [et al.] // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.

3. Ke, G. LightGBM : A Highly Efficient Gradient Boosting Decision Tree / G. Ke, Q. Meng, T. Finley [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

4. Dietterich, T. G. Ensemble Methods in Machine Learning / T. G. Dietterich // Multiple Classifier Systems. – Berlin : Springer, 2000. – P. 1–15. – DOI 10.1007/3-540-45014-9_1.

5. Caruana, R. Ensemble Selection from Libraries of Models / R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes // Proceedings of the 21st International Conference on Machine Learning. – New York : ACM, 2004. – P. 18.