ОТЗЫВ

научного руководителя на диссертацию
Цзоу Цзиньин «Алгоритмы интерпретируемого искусственного интеллекта на основе значений Шепли в задачах высокой размерности», представленную на соискание ученой степени кандидата технических наук по специальности
1.2.2. Математическое моделирование, численные методы и комплексы программ.

Диссертация посвящена изучению и развитию методов объяснимого искусственного интеллекта применительно к высокоразмерным системам ИИ. В основе исследования лежит классический алгоритм теории кооперативных игр - значение Шэпли. В данной диссертации предложены четыре алгоритма на основе значения Шэпли для решения задач в высокоразмерных интерпретируемых моделях искусственного интеллекта. Первый метод, Bi-Level based Shapley Calculation, вводит двухуровневую структуру в сочетании с вычислением значения Шэпли для снижения вычислительных затрат, связанных с большим количеством игроков в игре. Этот подход определяет два типа игроков: первый тип рассматривает отдельные признаки как игроков, а второй тип рассматривает группы признаков как игроков. Все признаки первоначально группируются, а значение Шэпли вычисляется в два этапа. На первом этапе вычисляются только значения Шэпли групповых игроков, а на втором этапе вычисляются значения Шэпли отдельных игроков признаков внутри каждой группы. Такая иерархическая структура значительно снижает общую вычислительную сложность. Второй метод, двухуровневая аппроксимация Шэпли на основе ограниченных K-средних, совершенствует двухуровневый подход, вводя ограниченные k-средние для более эффективной группировки игроков при ограничении размеров групп. В оригинальном двухуровневом методе неэффективность может возникнуть, если некоторые группы содержат слишком много игроков, что приводит к увеличению вычислительных затрат. Данный метод решает эту проблему, обеспечивая одинаковое количество игроков во всех группах, тем самым минимизируя общие вычислительные затраты и повышая эффективность.Третий метод, метод Шэпли на основе графов, совершенствует исходный подход к выборке Шэпли, при котором формирование выборки происходит совершенно случайно. Этот метод использует коэффициент корреляции Пирсона для создания матрицы отношений и построения графа отношений между игроками. Для уточнения графа ребра с меньшими весами удаляются на основе заранее заданных пороговых значений. Затем используется алгоритм поиска для случайной генерации путей в графе, руководствуясь весами ребер. Набор узлов вдоль каждого пути используется в качестве наблюдений для вычисления значения Шэпли. Генерируя таким образом более квалифицированные выборки, метод оптимизирует качество выборки и значительно повышает скорость сходимости алгоритма. Четвертый метод, метод последовательного обновления Шэпли, вводит механизм байесовского обновления для уточнения стратегии выборки. В отличие от исходных подходов, в которых все выборки генерируются заранее, этот метод динамически генерирует наблюдения на основе предварительных распределений вероятностей. После каждой итерации вычисления значений Шэпли вероятность участия каждого игрока обновляется, определяя, будет ли он включен в следующее наблюдение. Эта

итеративная корректировка позволяет более эффективно осуществлять выборку и улучшает общий вычислительный процесс.

Общая структура данной диссертации состоит из введения, четырех основных глав, заключения и ссылок на литературу.

В первой главе представлен интерпретируемый метод обнаружения аномальных логов, основанный на вычислении значений Шэпли в сочетании с двухуровневым подходом. Ключевым новшеством этого метода является разработка двухуровневого фреймворка для решения проблемы высокой вычислительной сложности, вызванной большим количеством коалиций в моделях, основанных на последовательности, что позволяет эффективно вычислять значение Шэпли. Алгоритм применен в DeepLog, системе обнаружения аномалий на основе нейронных сетей. Он эффективно устраняет ограничения подходов, основанных на признаках, при анализе вкладов и предназначен для обработки аномалий, вызванных последовательностями признаков. Этот подход повышает эффективность вычислений и дает такие преимущества, как поддержка объяснений последовательностей.

Во второй главе предлагается двухуровневый метод (Bi-level) в сочетании с ограниченным подходом k-means для дальнейшего повышения скорости вычислений. Основное новшество заключается в разумной группировке признаков и ограничении размера группы для решения проблем эффективности, вызванных непоследовательной длиной групп в оригинальном двухуровневом методе Шэпли, что открывает новые перспективы для решения высокоразмерных задач обнаружения аномалий. Предложенный метод применяется в системе обнаружения раковых аномалий на основе Isolation Forest и сравнивается с подходом, основанным на семплировании. Результаты показывают, что метод сопоставим по производительности и предлагает альтернативное решение для интерпретации высокоразмерных задач в сложных моделях ИИ.

В третьей главе предлагается подход к семплированию по Шэпли на основе графа. Оригинальный метод семплирования заимствован из открытого фреймворка SHAP, где стратегия семплирования является случайной. Инновационный аспект этой главы заключается в построении графа отношений между игроками на основе характеристик данных и определении весов ребер для представления корреляции между игроками. Заменив оригинальную стратегию случайного семплирования алгоритмом поиска для генерации высококачественных образцов, метод достигает более быстрой сходимости. Этот подход был применен к данных обнаружения рака и продемонстрировал значительное снижение вычислительных затрат по сравнению с оригинальным алгоритмом.

В четвертой главе предлагается метод последовательного обновления Шэпли, основанный на байесовской модели. Основным новшеством этого метода является введение концепции последовательных байесовских обновлений в алгоритм сэмплирования значений Шэпли. По сравнению с методом сэмплирования по Шэпли на основе графа, предложенным в главе 3, и оригинальным алгоритмом случайного сэмплирования в рамках SHAP, этот метод динамически генерирует сэмплы, используя исторические вклады игроков, и корректирует сэмплы на основе результатов каждой итерации. Он обладает большей гибкостью и способностью к обобщению, обеспечивая более адаптивное решение для применения значения Шэпли в объясняемом ИИ.

Во время учебы в аспирантуре Цзоу Цзиньин опубликовал 7 научных работ, в том числе:

- Explainable AI: Using Shapley Value to Explain the Anomaly Detection System Based on Machine Learning Approaches.ПРОЦЕССЫ УПРАВЛЕНИЯ И УСТОЙЧИВОСТЬ, 2020, 355-360.

- Explainable AI: Using Shapley Value to Explain Complex Anomaly Detection ML-Based Systems//Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020. – 2020. – Т. 332. – С. 152. (cited 40)

- High-Dimensional Explainable AI for Cancer Detection. 1 Sep 2021, In: International Journal of Artificial Intelligence. 19, 2, p. 195-217 23 p. (cited 41)

- "Explainable AI: Graph Based Sampling Approach for High Dimensional AI System." *International Conference on Intelligent Information Technologies for Industry*. Cham: Springer Nature Switzerland, 2023.

- Explainable AI: Efficiency Sequential Shapley Updating Approach," in *IEEE Access*, vol. 12, pp. 166414-166423, 2024, doi: 10.1109/ACCESS.2024.3495543.

- XAI evaluation: evaluating black-box model explanations for prediction." 2021 II International conference on neural networks and neurotechnologies (NeuroNT). IEEE, 2021.

- Dynamic shapley value in the game with perishable goods." Contributions to Game Theory and Management 14 (2021): 273-289.

Из всех опубликованных работ 2 были опубликованы в высокорейтинговых журналах (Scopus Q1), в том числе:

- Zou, J., Xu, F., Zhang, Y., Petrosian, O. & Krinkin, K. High-Dimensional Explainable AI for Cancer Detection. 1 Sep 2021, In: International Journal of Artificial Intelligence. 19, 2, p. 195-217 23 p.
- Explainable AI: O. Petrosian and J. Zou, "Explainable AI: Efficiency Sequential Shapley Updating Approach," in *IEEE Access*, vol. 12, pp. 166414-166423, 2024, doi: 10.1109/ACCESS.2024.3495543.

Результаты исследований были представлены на следующих международных конференциях и семинарах:

- Control Processes and Stability (CPS'20)
- The International Conference on Machine Learning and Intelligent Systems (MLIS2020)
- Intelligent Information Technologies for Industry(IITI23)

В 2019 году Цзоу Цзиньин завершил магистерскую программу «Game Theory and Operations Research» Санкт-Петербургского государственного университета и в том же году поступил в аспирантуру по программе «Системный анализ, информатика и управление». С момента получения степени магистра и по настоящее время Цзоу Цзиньин занимается академическими исследованиями и практической работой в этой области как в университетах, так и в коммерческих компаниях.

Считаю, что диссертационная работа «Алгоритмы интерпретируемого искусственного интеллекта на основе значений Шепли в задачах высокой размерности» Цзоу Цзиньин соответствует специальности 1.2.2. Математическое моделирование, численные методы и комплексы программ и удовлетворяет требованиям, предъявляемым Санкт-Петербургским государственным университетом к работам на соискание учёной степени кандидата технических наук. Автор заслуживает присуждения учёной степени кандидата технических наук по специальности 1.2.2. Математическое моделирование, численные методы и комплексы программ.

Петросян Ованес Леонович
Научный руководитель,
Доктор физико-математических наук,
Профессор кафедры математического моделирования энергетических систем
Санкт-Петербургского государственного университета