УДК 004.8

Гурбанова Лейла Гадировна, студентка кафедры прикладной математики,

Российский технологический университет МИРЭА,

г. Москва

Gurbanova Leila Gadirovna, Student, Department of Applied Mathematics,

RTU MIREA,

Moscow

ИНТЕРПРЕТИРУЕМОСТЬ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ: МЕТОДЫ ОБЪЯСНЕНИЯ ОТ КЛАССИЧЕСКИХ ПОДХОДОВ ДО SHAP И LIME

Аннотация. *В статье рассматривается проблема интерпретируемости моделей машинного обучения в контексте развития современных методов искусственного интеллекта. Увеличение сложности алгоритмов, включая ансамблевые методы и глубокие нейронные сети, приводит к снижению прозрачности принимаемых решений и формированию эффекта «черного ящика». В работе проводится систематизация подходов к интерпретации моделей, включающая классические методы анализа (коэффициенты линейной регрессии, Feature Importance, деревья решений) и современные пост-hoc методы объяснения (LIME, SHAP, Partial Dependence Plots, ICE, Anchors). Рассматриваются их теоретические основы, математические принципы и сравнительные характеристики по уровню интерпретируемости, вычислительной сложности и устойчивости результатов. Особое внимание уделяется практическому применению методов интерпретации в задачах медицинской диагностики, кредитного скоринга и обнаружения мошеннических операций. Показано, что интерпретируемость моделей является ключевым фактором повышения доверия к системам искусственного интеллекта, обеспечения их*

прозрачности и соответствия нормативным требованиям. Сделан вывод о необходимости развития гибридных подходов, объединяющих высокую точность моделей и возможность их объяснения.

Ключевые слова: интерпретируемость моделей, машинное обучение, объяснимый искусственный интеллект, SHAP, LIME, Feature Importance, post-hoc методы, «черный ящик», анализ данных, классификация, кредитный скоринг, медицинская диагностика

INTERPRETABILITY OF MACHINE LEARNING MODELS: EXPLANATION METHODS FROM CLASSICAL APPROACHES TO SHAP AND LIME

Abstract. *The article addresses the problem of interpretability of machine learning models in the context of the development of modern artificial intelligence methods. The increasing complexity of algorithms, including ensemble methods and deep neural networks, leads to reduced transparency of decision-making processes and the emergence of the "black box" effect. The study provides a systematization of model interpretation approaches, including classical analytical methods (linear regression coefficients, Feature Importance, decision trees) and modern post-hoc explanation techniques (LIME, SHAP, Partial Dependence Plots, ICE, Anchors). Their theoretical foundations, mathematical principles, and comparative characteristics in terms of interpretability level, computational complexity, and stability of results are considered. Special attention is given to practical applications of interpretability methods in medical diagnosis, credit scoring, and fraud detection tasks. It is shown that model interpretability is a key factor in increasing trust in artificial intelligence systems, ensuring their transparency, and meeting regulatory requirements. The study concludes that there is a need to develop hybrid approaches that combine high predictive accuracy with explainability.*

Keywords: *model interpretability, machine learning, explainable artificial intelligence, SHAP, LIME, Feature Importance, post-hoc methods, black box, data analysis, classification, credit scoring, medical diagnosis*

Введение

Современные методы машинного обучения характеризуются высокой вычислительной сложностью и необходимостью обработки больших объёмов данных. Рост размерности задач искусственного интеллекта требует поиска новых вычислительных подходов, способных обеспечить повышение производительности и ускорение анализа информации. Одним из наиболее перспективных направлений является использование квантовых вычислений в задачах интеллектуальной обработки данных. Квантовые вычислительные системы основаны на принципах квантовой механики, включая суперпозицию и запутанность состояний. В отличие от классических битов, кубиты способны одновременно находиться в нескольких состояниях, что обеспечивает возможность параллельной обработки информации. Благодаря этому квантовые алгоритмы потенциально способны значительно ускорять решение отдельных вычислительных задач. На пересечении искусственного интеллекта и квантовых вычислений сформировалось направление квантового машинного обучения, в рамках которого разрабатываются квантовые нейронные сети. Такие модели рассматриваются как перспективная альтернатива классическим нейросетевым архитектурам при решении задач классификации, прогнозирования и оптимизации.

Целью данной работы является анализ теоретических основ квантовых нейронных сетей, исследование архитектур квантового машинного обучения и оценка перспектив их применения в анализе данных.

Теоретические основы квантовых нейронных сетей

Квантовые нейронные сети представляют собой вычислительные модели, объединяющие методы искусственного интеллекта и принципы квантовой механики. Основой таких систем являются кубиты — квантовые элементы информации, способные находиться в состоянии суперпозиции. В отличие от классического бита, принимающего значения (0) или (1), состояние кубита описывается линейной комбинацией:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

где (α) и (β) — коэффициенты вероятности квантового состояния.

Суперпозиция позволяет квантовой системе одновременно представлять множество состояний, что формирует основу квантового параллелизма. Дополнительным преимуществом является квантовая запутанность, обеспечивающая сложные корреляции между кубитами и повышающая эффективность вычислительных процессов. Большинство современных квантовых нейронных сетей реализуется в виде параметризованных квантовых схем. Такие схемы состоят из квантовых вентилей, параметры которых изменяются в процессе обучения аналогично весовым коэффициентам классических нейронных сетей. Для оптимизации параметров применяются гибридные алгоритмы, объединяющие квантовые вычисления и классические методы градиентного спуска. В задачах классификации качество работы модели оценивается с помощью функции потерь. Одной из наиболее распространённых метрик является accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

где (TP) — истинно положительные результаты, (TN) — истинно отрицательные, (FP) — ложноположительные, (FN) — ложноотрицательные классификации.

Архитектуры и применение квантового машинного обучения

Современные архитектуры квантового машинного обучения в большинстве случаев строятся как гибридные квантово-классические системы. В таких моделях квантовый процессор выполняет вычислительно сложные операции, а классические алгоритмы используются для оптимизации параметров и управления процессом обучения.

Одной из наиболее распространённых архитектур являются вариационные квантовые схемы (Variational Quantum Circuits, VQC). В этих моделях входные данные кодируются в квантовые состояния, после чего над системой выполняются последовательности квантовых преобразований. Параметры квантовых вентилей оптимизируются в процессе обучения, что позволяет модели адаптироваться к структуре данных. Для оценки качества работы модели используется функция потерь, например среднеквадратичная ошибка:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

где y_i — истинное значение, а $\widehat{y_i}$ — предсказание модели.

Квантовые нейронные сети находят применение в различных задачах анализа данных. В области классификации они используются для распознавания изображений, обработки медицинских данных и анализа финансовых показателей. Благодаря квантовому параллелизму такие модели способны эффективно работать с высокоразмерными пространствами признаков.

Дополнительным направлением является использование квантовых алгоритмов в задачах оптимизации и кластеризации. Исследуются квантовые версии алгоритмов k-means и методов снижения размерности данных. Предполагается, что в перспективе квантовые вычисления смогут существенно ускорить обработку больших массивов информации. Однако

практическое применение квантового машинного обучения пока ограничивается характеристиками современных квантовых устройств. Большинство существующих квантовых процессоров обладают ограниченным числом кубитов и высоким уровнем шумов, что снижает стабильность вычислений и усложняет реализацию глубоких квантовых моделей. Тем не менее развитие квантовых архитектур и гибридных алгоритмов позволяет рассматривать квантовые нейронные сети как перспективное направление развития систем искусственного интеллекта.

Ограничения и перспективы развития квантовых моделей

Несмотря на значительный теоретический потенциал квантовых нейронных сетей, их практическое применение в анализе данных ограничено рядом фундаментальных и технологических факторов. Эти ограничения связаны как с физической природой квантовых систем, так и с текущим уровнем развития квантового аппаратного обеспечения. Одной из ключевых проблем является декогеренция — процесс потери квантовой когерентности вследствие взаимодействия системы с внешней средой. Декогеренция приводит к разрушению квантовых состояний и снижению точности вычислений, что особенно критично при выполнении многокубитных операций в нейросетевых моделях. Дополнительным ограничением выступает шумность современных квантовых устройств. Большинство существующих квантовых процессоров относится к классу NISQ (Noisy Intermediate-Scale Quantum), что означает ограниченное число кубитов и высокую чувствительность к ошибкам. В таких условиях реализация глубоких квантовых нейронных сетей становится затруднительной. Существенной проблемой также является эффект «плато градиентов» (barren plateau), при котором значения градиентов функции потерь стремятся к нулю при увеличении размерности квантовой схемы. Это значительно усложняет процесс обучения и снижает эффективность оптимизационных алгоритмов.

Несмотря на указанные ограничения, квантовые нейронные сети обладают значительным потенциалом развития. Одним из перспективных направлений является создание гибридных квантово-классических систем, в которых квантовые вычисления используются для обработки наиболее сложных частей модели, а классические алгоритмы обеспечивают устойчивость и интерпретируемость. Также активно развиваются методы квантовой коррекции ошибок, направленные на повышение стабильности вычислений и снижение влияния шумов. В долгосрочной перспективе ожидается увеличение числа доступных кубитов и улучшение качества квантовых операций, что позволит реализовать более сложные архитектуры квантового машинного обучения.

Заключение

В ходе работы рассмотрены основные принципы построения квантовых нейронных сетей, их архитектурные особенности, а также возможные направления применения в задачах анализа данных. Показано, что квантовые вычислительные модели основаны на использовании фундаментальных свойств квантовой механики — суперпозиции и запутанности, что обеспечивает принципиально новые вычислительные возможности по сравнению с классическими подходами. Анализ существующих квантовых алгоритмов демонстрирует, что квантовые нейронные сети могут быть эффективно использованы в задачах классификации, оптимизации и обработки многомерных данных. Особый интерес представляют гибридные квантово-классические модели, которые позволяют компенсировать ограничения современных квантовых устройств и обеспечивают более устойчивое обучение.

В то же время выявлены ключевые ограничения практического применения квантовых нейросетевых моделей, включая декогеренцию, шумность вычислений и проблему масштабируемости. Эти факторы

существенно замедляют переход от теоретических моделей к промышленным квантовым системам анализа данных. Тем не менее развитие технологий квантовых вычислений, а также совершенствование методов квантового машинного обучения позволяют рассматривать квантовые нейронные сети как перспективное направление исследований. В будущем ожидается их более глубокая интеграция в системы искусственного интеллекта и обработку больших данных, что может привести к появлению новых вычислительных парадигм.

Литература

1. Aeppli G., Rosenbaum T. Quantum Annealing and Related Optimization Methods / Ed. by A. Das, K. Chakrabarti. Lecture Notes in Physics. Heidelberg: SpringerVerlag, 2007.
2. Altaisky M. Quantum Neural Network: Tech. Report. arxiv.org:quant-ph/0107012. 2001.
3. Altaisky M., Rao V. Inverted Mexican Hat Potential in Activation of Receptor Cells // Nonlin. Analysis B. 2009.
4. Beck F. Synaptic Quantum Tunnelling in Brain Activity // Neuroquantology. 2008.
5. Beck F., Eccles J. Quantum Aspects of Brain Activity and the Role of Consciousness // PNAS. 1992.
6. Behera L., Kar I., Elitzur A. A Recurent Quantum Neural Network Model to Describe Eye Tracking of Moving Targets // Found. Phys. Lett. 2005.
7. Chrisley R. Learning in Non-Superpositional Quantum Neurocomputers // Brain, Mind and Physics / Ed. by Pylkkéanen, P. Pylkkéo. IOS Press, 1997.