Архитектура нейронной аддитивной модели для интерпретируемой оценки неопределенности предсказаний

Р. И. Думаев[1], С. А. Молодяков, Л. В. Уткин

Санкт-Петербургский политехнический университет Петра Великого

[1]dumaevrinat@gmail.com

***Аннотация*. Предложена модифицированная архитектура нейронной аддитивной модели для интерпретируемой оценки неопределенности предсказаний, основанная на параметрическом моделировании распределения целевой переменной. В отличие от классических нейронных аддитивных моделей, ориентированных на оценку среднего значения, рассматриваемый подход позволяет представлять параметры распределения как суммы нелинейных функций отдельных признаков. Для каждого параметра распределения обучаются независимые подсети, обеспечивающие декомпозицию вкладов признаков и их интерпретацию. Обучение осуществляется с использованием функции правдоподобия, что позволяет учитывать статистические свойства данных. Экспериментальная оценка показала сопоставимую точность с базовой архитектурой нейронной аддитивной модели и отсутствие статистически значимых различий, при этом предлагаемый подход расширяет возможности базовой модели за счет анализа неопределенности, обеспечивая объяснимость как предсказаний, так и связанных с ними характеристик распределения.**

Ключевые слова: нейронные аддитивные модели; объяснение на основе понятий; объяснимый искусственный интеллект

I. Введение

Применение методов глубокого обучения ограничено из-за отсутствия возможности интерпретации результатов их работы [1]. При решении различных прикладных задач, таких как построение систем поддержки принятия решений, помимо получения точных предсказаний, необходимо понимать взаимосвязь входных данных с результатом работы, то есть иметь объяснения полученных результатов, а также оценку связанной с ними неопределенности [2]. Таким образом, создание интерпретируемых моделей является ключевым направлением в развитии современных методов машинного обучения.

При этом такие подходы игнорируют статистические характеристики (например, форма распределения), которые существенны для задач оценки неопределенности предсказаний — в практических задачах информация о неопределенности может иметь сопоставимую значимость с точечным предсказанием, так как оценка неопределенности позволяет учитывать риск и степень надежности получаемых результатов. Существующие подходы к оценке неопределенности, к которым относятся ансамбли моделей или байесовские подходы, усложняют архитектуру моделей, снижают их интерпретируемость и приводят к дополнительным вычислительным затратам. При этом возникает необходимость в разработке методов, которые одновременно сохраняют прозрачность интерпретируемых моделей и позволяют учитывать неопределенность предсказаний в явном виде.

В данной работе предлагается модифицированная архитектура нейронной аддитивной модели, основанная на параметрическом моделировании распределения целевой переменной. Параметры распределения представляются как суммы нелинейных функций признаков с использованием независимых подсетей, что обеспечивает интерпретацию как предсказаний, так и характеристик распределения.

II. Обзор литературы

Существующие подходы к повышению объяснимости моделей глубокого обучения основываются на post-hoc методах объяснения, включая такие как LIME [3] или SHAP [4], которые аппроксимируют локальную значимость признаков. Данные методы обеспечивают лишь частичное понимание и лишены интерпретируемости. В связи с этим возрастает интерес к разработке изначально понятных моделей (интерпретируемых). Традиционные подходы, такие как обобщенные линейные модели и обобщенные аддитивные модели [5], обеспечивают прозрачность, однако их применение в контексте глубокого обучения связано с ограничениями. Развитием этих подходов к обеспечению интерпретируемости являются нейронные аддитивные модели (NAM), которые представляют предсказание в виде суммы нелинейных функций для отдельных признаков [6]. Архитектура NAM обеспечивает раздельное представление вкладов признаков и позволяет интерпретировать результат работы модели на глобальном уровне через анализ соответствующих функций формы, что дает возможность анализировать вклад каждого признака [7].

Работа выполнена при финансовой поддержке Российского Научного Фонда, проект № 25-11-00021

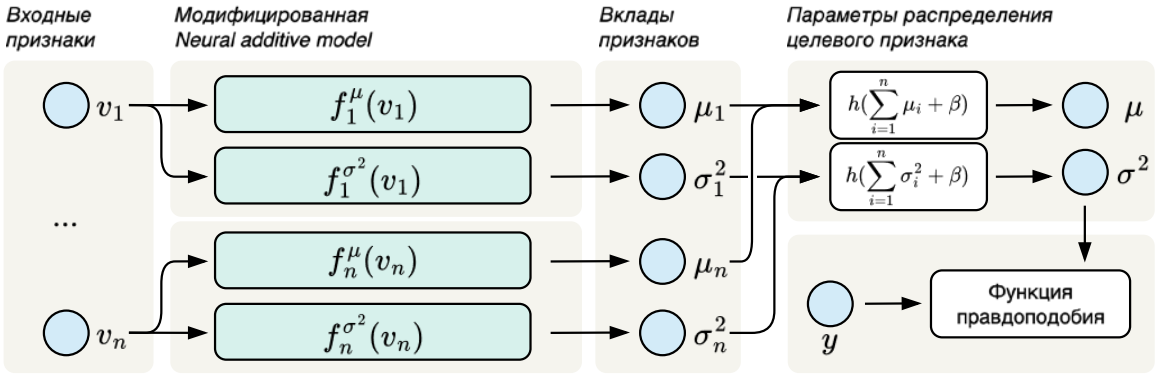

Рис. 1. Пример архитектуры модифицированной NAM для нормального распределения

Вместе с этим классические NAM ориентированы на оценку среднего значения целевой переменной и не учитывают ее распределение, тем самым такая архитектура не позволяет явно оценить неопределенность предсказаний. Существующие модификации архитектуры для учета неопределенности усложняют модель или снижают интерпретируемость: например, в ряде работ предпринимались попытки интегрировать в NAM учет неопределенности посредством байесовских методов [8]. Однако такие подходы ориентированы на эпистемическую неопределенность и не предоставляют механизма моделирования распределения целевой переменной через его параметры.

Другим же направлением является использование ансамблей NAM, а также методов на основе dropout для получения распределения предсказаний. Несмотря на простоту реализации, такие методы носят внешне-системный характер по отношению к архитектуре модели и не обеспечивают явной интерпретации влияния признаков на неопределенность. Таким образом, существующие подходы обеспечивают вероятностное моделирование без интерпретируемой структуры. Это указывает на недостаточную проработанность методов, объединяющих аддитивную интерпретируемую структуру с параметрическим описанием распределения целевого признака.

III. Предложенная архитектура

Предлагаемая архитектура направлена на объединение интерпретируемости нейронных аддитивных моделей с возможностью параметрического моделирования распределения целевой переменной. В отличие от классической постановки, в которой модель оценивает только условное математическое ожидание, рассматриваемый подход предполагает предсказание полного набора параметров выбранного распределения.

Пусть задана обучающая выборка:

$$S = \{(v_i, y_i)\}_{i=1}^{L}, \quad (1)$$

где L – число объектов, $v = \{v_1, v_2, ..., v_N\}$ – вектор N интерпретируемых признаков, а y_i соответствующее значение целевой переменной.

В отличие от классических нейронных аддитивных моделей, в которых предсказывается единственное значение (например, среднее), предлагаемый подход ориентирован на моделирование параметров распределения целевой переменной. Пусть θ_j, $j = 1, ..., J$ обозначают параметры выбранного распределения (например, μ и σ^2 в случае нормального распределения).

Для каждого параметра θ_j распределения строится независимая аддитивная подсеть, представленная в виде суммы нелинейных функций отдельных признаков. С этой целью для каждого признака v_n и каждого параметра θ_j обучается отдельная подсеть f_{nj}. Итоговое значение параметра определяется следующим образом:

$$\theta_j = h_j(\sum_{n=1}^{N} f_{nj}(v_n) + \beta_j), \quad (2)$$

где h_j – функция активации, обеспечивающая корректность области значений параметра (например, линейная функция для среднего и softplus для дисперсии), а β_j – смещение.

Такая структура модели обеспечивает аддитивную декомпозицию каждого параметра распределения по признакам. Значения функций f_{nj} интерпретируются как вклад признака v_n в формирование параметра θ_j, при этом вклад каждого признака не зависит от других признаков, что сохраняет интерпретируемость модели на глобальном уровне.

Переход к моделированию параметров распределения требует соответствующей постановки задачи обучения. В качестве функции потерь используется отрицательное логарифмическое правдоподобие (negative log-likelihood), которое позволяет учитывать вероятностную природу целевой переменной. Для нормального распределения функция потерь имеет вид:

$$L_{N(\mu,\sigma^2)} = \frac{1}{2}(log(\sigma^2) + \frac{(y-\mu)^2}{\sigma^2}), \quad (3)$$

где μ и σ^2 – предсказанные параметры распределения, а y – наблюдаемое значение. Минимизация данной функции осуществляется с использованием градиентных методов оптимизации. В общем случае выбор функции правдоподобия определяется предположениями о распределении целевой переменной и спецификой решаемой задачи. Так, для задачи бинарной классификации можно использовать логистическое распределение.

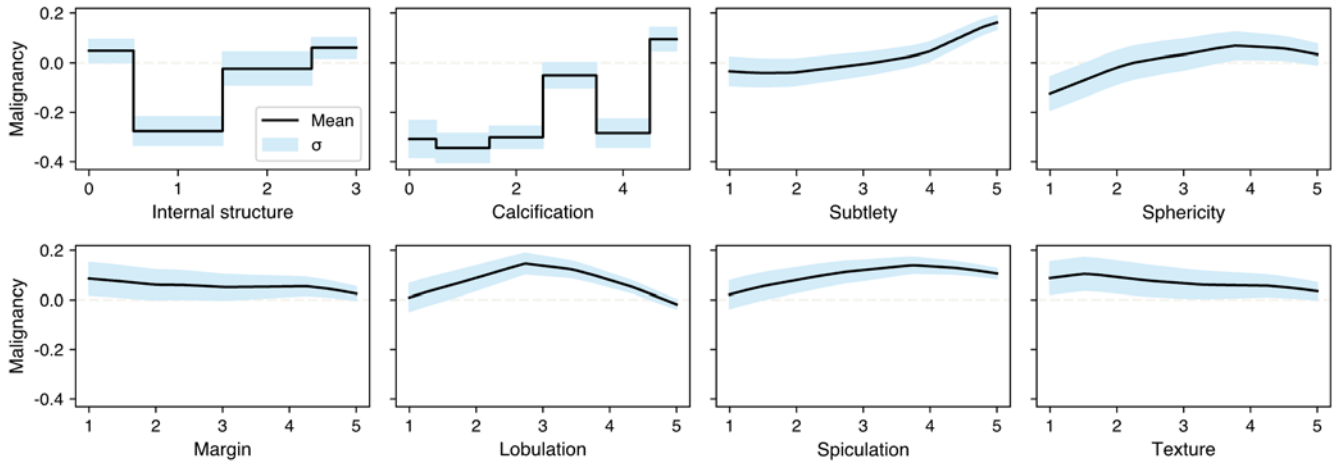

Рис. 2. Полученные функции влияния признаков для набора данных LIDC-IDRI

Предложенная архитектура позволяет получать не только точечные оценки, но и характеристики неопределенности предсказаний, заданные через параметры распределения. При этом интерпретируемость сохраняется для каждого параметра отдельно, что дает возможность анализировать влияние признаков как на ожидаемое значение, так и на меру разброса.

Дополнительно, для количественной оценки влияния признаков может использоваться мера важности, основанная на вариации значений функций формы на рассматриваемом интервале значений признака. Такая мера позволяет сравнивать относительный вклад различных признаков в формирование параметров распределения.

Таким образом, предложенная модель объединяет интерпретируемую аддитивную структуру с вероятностным описанием целевой переменной, обеспечивая возможность анализа как предсказаний, так и связанной с ними неопределенности на уровне отдельных признаков.

IV. Эксперименты

A. *Наборы данных*

Мы оцениваем предложенную архитектуру на нескольких общедоступных наборах данных из разных областей с различным размером пространства признаков:

- Набор данных осложнений инфаркта миокарда (MICD), который содержит подробную информацию о пациентах с инфарктом миокарда и предназначен для тестирования методов прогнозирования внутрибольничных осложнений [9].
- Набор данных MIMIC, который содержит комплексные клинические записи пациентов отделений интенсивной терапии [10].
- Набор данных диагностики рака молочной железы (WDBC), включающий признаки, полученные методом тонкоигольной аспирации опухоли молочной железы [11].
- Набор данных LIDC, содержащий размеченные аннотации узлов в легких, выполненные несколькими радиологами, и используется для задач анализа и диагностики легочных новообразований [12].

Мы оцениваем точность обученных моделей с использованием пятикратной перекрестной проверки. Набор данных случайным образом разделялся на обучающую выборку (80%) и тестовую выборку (20%), образуя пять разбиений. Пересечение выборок (одни и те же примеры) между обучающими и тестовыми данными в пределах одного разбиения не допускалось.

Для оценки качества прогнозирования предлагаемой модели и архитектуры NAM, а также их сравнения использовались следующие метрики: площадь под кривой «точность–полнота» (AUC) для задач бинарной классификации и среднеквадратичная ошибка (RMSE) для задач регрессии.

B. *Детали реализации*

Все эксперименты и сама архитектура модифицированной NAM были реализованы с использованием библиотеки PyTorch. Каждая подсеть признака реализована в виде многослойного перцептрона (MLP). Оптимизация выполняется с использованием оптимизатора AdamW. Все модели обучаются в течение фиксированного числа эпох. Для сети признаков были заданы следующие параметры: скрытые слои размерностью (256, 128, 64, 32), функция активации leaky ReLU для скрытых слоев, линейная активация для среднего значения и softplus активация для дисперсии с целью предсказывания только положительных значений.

C. *Результаты*

В табл. I представлены результаты работы каждого подхода вместе со значением p-критерия Вилкоксона.

Выбор непараметрического критерия Вилкоксона для связанных выборок обусловлен особенностями сравнения и свойствами полученных данных: значения метрик для обеих моделей вычислялись на одних и тех же разбиениях выборки, наблюдения являются связными. Также применение данного критерия позволяет оценить наличие различий между моделями без дополнительных предположений о виде распределения ошибок.

ТАБЛИЦА I. СРАВНЕНИЕ ПРЕДЛОЖЕННОЙ АРХИТЕКТУРЫ С NAM

Набор данных	**Метрика**	**Предложенная архитектура**	**NAM**	**p-value**
MICD	AUC	0.697 ± 0.028	0.672 ± 0.034	0.125
MIMIC	AUC	0.832 ± 0.005	0.831 ± 0.005	0.437
WDBC	AUC	0.985 ± 0.009	0.984 ± 0.009	0.375
LIDC	RMSE	0.142 ± 0.007	0.142 ± 0.008	0.812

Анализ представленных в табл. I результатов демонстрирует, что предложенная модифицированная архитектура обеспечивает сопоставимое качество прогнозирования по сравнению с базовой архитектурой NAM на всех рассматриваемых наборах данных. Для задач бинарной классификации значения метрики AUC находятся на близком уровне, при этом в ряде случаев наблюдается незначительное преимущество предложенного подхода.

Статистический анализ с использованием критерия Вилкоксона показал отсутствие статистически значимых различий между результатами моделей на всех наборах данных. Это позволяет сделать вывод о том, что предложенная архитектура не уступает базовой NAM по качеству прогнозирования. Важным результатом является то, что при этом предложенный подход расширяет функциональные возможности модели за счет явного моделирования распределения целевой переменной, что позволяет получать не только точечные оценки, но и интерпретируемые характеристики неопределенности. Пример таких полученных функций влияния признаков показан на рис. 1. Благодаря аддитивной декомпозиции параметров распределения достигается возможность анализа вклада отдельных признаков как в предсказание среднего значения, так и в оценку неопределенности.

V. ЗАКЛЮЧЕНИЕ

В работе предложена модифицированная архитектура нейронной аддитивной модели, обеспечивающая параметрическое моделирование распределения целевой переменной при сохранении интерпретируемой аддитивной структуры. Подход позволяет получать как точечные предсказания, так и интерпретируемые оценки неопределенности за счет декомпозиции параметров распределения по признакам.

Экспериментальные результаты показали сопоставимое качество прогнозирования с базовой NAM при отсутствии статистически значимых различий, что подтверждает эффективность предложенной архитектуры. При этом модель расширяет функциональные возможности за счет явного учета неопределенности без потери интерпретируемости.

Предложенный подход может быть использован в задачах, требующих точности и интерпретируемой оценки надежности предсказаний, и представляет интерес для дальнейшего развития в направлении более сложных вероятностных моделей.

СПИСОК ЛИТЕРАТУРЫ

[1] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. 2020.

[2] Yu J., Wang Z., Vasudevan V., Yeung L., Seyedhosseini M., Wu Y. Coca: Contrastive captioners are image-text foundation models //arXiv preprint arXiv:2205.01917. 2022.

[3] Ribeiro M. T., Singh S., Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. С. 1135-1144.

[4] Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions //Advances in neural information processing systems. 2017. Т. 30.

[5] Hastie T.J. Generalized additive models //Statistical models in S. 2017. С. 249-307.

[6] Agarwal R., Melnick L., Frosst N., Zhang X., Lengerich B., Caruana R., Hinton G.E. Neural additive models: Interpretable machine learning with neural nets //Advances in neural information processing systems. 2021. Т. 34. С. 4699-4711.

[7] Dumaev R.I., Molodyakov S.A., Utkin L.V. A model for explainable malignancy assessment of pulmonary nodules on CT images //Artificial Intelligence and Decision Making. 2024. №. 4. С. 123-134.

[8] Bouchiat K., Immer A., Yèche H., Rätsch G., Fortuin V. Improving neural additive models with bayesian principles //arXiv preprint arXiv:2305.16905. 2023.

[9] Golovenkin S.E., Shulman V.A., Rossiev D.A., Shesternya P.A., Nikulina S.Y., Orlova Y.V., Voino-Yasenetsky V. F. Myocardial infarction complications. UCI Machine Learning Repository [Электронный ресурс].

[10] Saeed M., Villarroel M., Reisner A.T., Clifford G., Lehman L.W., Moody G., Mark R.G. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database //Critical care medicine. 2011. Т. 39. №. 5. С. 952-960.

[11] Street W.N., Wolberg W.H., Mangasarian O.L. Nuclear feature extraction for breast tumor diagnosis //Biomedical image processing and biomedical visualization. SPIE, 1993. Т. 1905. С. 861-870.

[12] Samuel G. The Lung Image Database Consortium (LIDC) and Image Database resource initiative (IDRI): A completed reference database of lung nodules on CT scans //Medical physics. 2011. Т. 38. С. 2.