Причинно-следственные связи в объяснимом искусственном интеллекте

Н. В. Шевская[1], Е. С. Охримук[2], Н. В. Попов[3]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]nvrazmochaeva@etu.ru, [2]cat.oxrimuck@yandex.ru, [3]nvpopov@stud.eltech.ru

***Аннотация*. Вопросы объяснимости искусственного интеллекта стали все чаще возникать в современной науке. Системы искусственного интеллекта достигли такого уровня сложности, что объяснение причин принятия определенных решений приняло нечеловекоразмерный масштаб. Большая часть попыток объяснить, как работают модели искусственного интеллекта, сводится к анализу пространства признаков и оценки важности (значимости) признаков в конкретной предметной области. Важную роль играет понимание, есть ли связь между существующими методами объяснимого искусственного интеллекта и методами построения истинных причинно-следственных связей модели (при условии их наличия). В настоящей работе демонстрируется классический подход: подготовка данных, обучение моделей и объяснение результатов, а также осуществляется постановка задачи поиска причинно-следственных связей.**

Ключевые слова: искусственный интеллект; объяснимый искусственный интеллект; XAI; причинно-следственные связи; причинность

I. Введение

Системы искусственного интеллекта (например, глубокие нейросетевые модели) научились обрабатывать огромные массивы данных, выявлять скрытые закономерности, существенно сокращая время работы и время принятия важных решений (например, в медицине и в биологии). Вместе с этим, системы искусственного интеллекта достигли такой сложности, что человеку становится невозможно понять, как они работают и почему принимают то или иное решение. Факт непонимания порождает недоверие, а это, в свою очередь, создает сопротивление внедрению таких систем как на юридическом (кто понесет ответственность в случае ошибки), так и на этическом (как последствия скажутся на человеке) [1].

Наблюдаемая тенденция наращивания мощностей в системах искусственного интеллекта рано или поздно приведет к отторжению таких систем обществом, что, вероятно, может затормозить развитие технологий и социально значимых структур. Непонимание, как работают мощные системы искусственного интеллекта, является краеугольным камнем развития самих систем и поэтому заслуживает рассмотрения [2].

Проблема объяснимости моделей искусственного интеллекта уже долгое время решается классическими методами объяснения, порожденными еще более классическими методами из области анализа пространства признаков. Такой подход показывает, какие из параметров наблюдаемых объектов в исходном наборе данных оказывают наибольшее влияние на принимаемое решение (например, в задачах классификации снимков МРТ головного мозга по наличию заболевания) [3]. Однако в ответе на вопрос о параметрах, оказывающий наибольшее влияние на принимаемое решение, нет ответа на вопрос о причинах принимаемого решения (часто врачу необходимо много времени, чтобы объяснить пациенту необходимость того или иного действия, например, операционного вмешательства). Задача определения значимости параметров известна благодаря богатой истории решений в области анализа пространства признаков, и не является по существу новой.

Один и тот же принцип, лежащий в основе методов объяснимого ИИ и методах анализа пространства признаков поднимает новую проблему – проблему определения причинно-следственных связей [4]. Определение веса или значимости признаков, конечно же, не дает представления о наличии или отсутствии причинно-следственных связей в самой модели.

II. Прогностические модели и их обучение

Содержательная постановка задачи: необходимо построить для дальнейшего объяснения прогностическую модель на медицинских данных. Формально, данные будут о сердечно-сосудистых заболеваниях.

A. Описание набора данных

В качестве тестируемых данных был выбран датасет из открытого источника [5].

При выборе датасета основными критериями были:

- размер датасета – для обучения прогностических моделей необходимо большое количество объектов, чтобы не допустить недообучения.
- тип задачи – методы интерпретации, которые рассмотрены в данной работе, применены для задач классификации, то есть таких задач, где целевая переменная принадлежит конечному множеству.

Выбранный датасет представляет собой набор медицинских исследований реальных пациентов на предмет сердечно-сосудистых заболеваний. Набор данных содержит 70000 объектов, 11 признаков и 1 целевую переменную, отвечающую болеет человек или нет (0- человек здоров, 1 – человек болен).

Информация о признаках, описывающих данные о пациенте и его принадлежность к одному из классов сердечно-сосудистых заболеваний: Age (возраст), Gende (пол), Height (рост), Weight (вес), Ap_h (верхняя граница

давления), Ap_lo (нижняя граница давления), Cholesterol (содержание холестерина: 1 – нормально, 2 – выше нормы, 3 – значительно выше нормы), Gluc (глюкоза: 1 – нормально, 2 – выше нормы, 3 – значительно выше нормы), Smoke (курит человек или нет), Alco (употребление алкоголя, субъективная характеристика), Active (физическая активность, субъективная характеристика), Cardio (наличие или отсутствие сердечно-сосудистых заболеваний, целевая переменная).

Перед обучением необходимо проверить данные на наличие пропущенных значений, выбросов и аномалий.

B. Обработка набора данных

Перед обучением прогностических моделей выбранный датасет был исследован и преобразован для улучшения качества работы моделей. В датасет был проведен поиск пропущенных значений и категориальных переменных. Таких значений не было обнаружено. Признак “age” представлен в днях. Для удобства понимания и оценивая данный признак был преобразован в года. Был создан новый признак “BMI” – индекс массы тела, который рассчитывается как вес, деленный на рост в квадрате. Этот признак будет хранить в себе информацию об обоих признаках и может заменить их, тем самым может ускорить обучение прогностических моделей и интерпретацию их результатов. Далее был проведен поиск аномальных значений в выбранном наборе данных. Аномальными считались те значений, которые являются либо выбросами, либо некорректно заполненными (определяли по правилу трех сигм). Были удалены все отрицательные значения показателей верхней и нижней границ артериального давления (“ap_hi”, “ ap_lo”) и значений, выходящие за общепризнанную норму (нижняя от 15 до 220, верхняя от 50 до 310).Был добавлен дополнительный признак “len”, который рассчитывается как разница верхней и нижней границ артериального давления. Это было сделано для поиска таких аномальных значений, где значение нижней границы больше значений верхней границы, что невозможно для реальных пациентов. Далее для признаков “height”, “weight”, “len” была произведена очистка от аномальных значений, которые не попадают в промежуток трех сигм. Этот способ разработан на основе статистик [6].

В итоге получили новый преобразованных датасет, который составляет теперь 67905 объектов.

C. Выбор прогностических моделей

Было принято решение выбирать самые популярные прогностические модели для обучения классификации, находящиеся в откроой библиотеке scikit-learn: (1) ExtraTreesClassifier, (2) Explainable Boosting Machine, (3) Linear SVC, (4) RidgeClassifier, (5) Naive Bayes, (6) Logistic Regression, (7) Stochastic Gradient Decent, (8) Support Vector Machines, (9) AdaBoostClassifier, (10) XGBClassifier, (11) k-Nearest Neighbors, (12) Random Forest, (13) BaggingClassifier, (14) Decision Tree Classifier.

Данные модели машинного обучения являются проверенными и эффективными, с их подробным описанием можно ознакомиться в документации [7].

Для обучения и тестирования данные разбились на тестовую и обучающую выборки в соотношении 30 к 70.

Для тестирования работы моделей использовалась встроенная функция score пакета sklearn.metrics.mean_squared_error [8]. Полученные результаты обучения можно увидеть в табл. 2.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ОБУЧЕНИЯ МОДЕЛЕЙ

№	Параметры моделей			
	Модель	***Score_train***	***Score_test***	***Score_diff***
1.	ExtraTreesClassifier	71.68	71.82	0.14
2.	Explainable Boosting Machine	73.46	73.62	0.16
3.	Linear SVC	72.55	72.98	0.43
4.	RidgeClassifier	72.43	72.87	0.44
5.	Naive Bayes	71.11	71.56	0.45
6.	Logistic Regression	71.24	71.73	0.49
7.	Stochastic Gradient Decent (SGD)	72.14	72.64	0.50
8.	Support Vector Machines (SVC)	72.01	72.64	0.63
9.	AdaBoostClassifier	71.03	71.88	0.85
10.	XGBClassifier	75.79	73.40	2.39
11.	k-Nearest Neighbors (KNN)	81.13	67.06	14.07
12.	Random Forest	97.58	70.77	26.81
13.	BaggingClassifier	95.80	68.05	27.75
14.	Decision Tree Classifier	97.58	63.40	34.18

Критерии сравнения результатов обучения: (1) Score_train – проверка точности на выборке, которая участвовала в обучении, (2) Score_test – на тестовой выборке, (3) Score_diff – разница этих показателей.

Расчет этих показателей был необходим для выявления переобучения, когда модель показывает низкую точность прогнозирования на новых данных.

III. МЕТОДЫ ОБЪЯСНЕНИЯ И ИХ ПРИМЕНЕНИЕ

A. Обзор методов объяснения

Основополагающих методов объяснения всего выделяют всего два:

- Векторы Шепли (SHAP – Additive Explanation Values) [9]

В теории игр существует такое понятие, как векторы Шепли, позволяющие численно оценить вклад каждого игрока для достижения общего результата. Векторы Шепли можно применить в машинном обучении, если игроками считать наличие отдельных признаков, а результатом игры – ответ модели на конкретном примере. Они позволяют оценить вклад каждого признака в ответ обученной модели. Таким образом, рассматривается вклад каждого признака в величину предсказания модели на конкретном тестовом примере, что помогает интерпретировать это предсказание.

- LIME (Local Interpretable Model-agnostic Explanations) [10]

Метод для модель-агностических локально интерпретируемых объяснений.

Подход основывается на идее интерпретации ответа модели на одном конкретном прогнозе. Для него вычисляется локальная линейная аппроксимация, то есть на вход модели подаются данные с варьирующимися малыми изменениями. На этих входных данных модель обучается и взвешивается по мере близости выбранных экземпляров к интересующему экземпляру.

Данный подход может быть применен к широкому спектру алгоритмов машинного обучения и не зависит от внутреннего устройства интерпретируемой моделей.

Модификаций LIME и SHAP методов существует большое количество. Достаточно крепко укоренилась практика агрегировать различные методы в более крупные библиотеки, среди которых можно выделить несколько наиболее основных.

Shap [11]. Библиотека Shap позволяет работать с большим количество различных моделей машинного обучения и интерпретировать их с помощью векторов Шепли, описанных ранее. Для наглядной визуализации работы алгоритмов машинного обучения Shap использует различные графики, а сфера применения включает в себя табличные, текстовые данные и изображения.

Lime [12] – пакет, позволяющий получить объяснения локального линейного приближения модели.

explainerDashboard [13] – библиотека для построения различных графиков, информационных интерактивных панелей и других визуальных инструментов, визуализирующих объяснения результатов машинного обучения. Функционал библиотеки содержит такие интерпретационные инструменты, как метод определения важностей признаков, метод объяснимости локального прогноза, Что-Если анализ, деревья решений, зависимость признаков и другие.

Dalex [14] – пакет для объяснения моделей и исследования их поведения. Вокруг обученной модели создается оболочка, в которой может быть выполнено сопоставление с набором глобальных и локальных методов интерпретации.

interpretML [15] – это пакет с открытым исходным кодом от компании Microsoft, который предоставляет методы для интерпретации результатов работы алгоритмов машинного обучения. Техники интерпретации данного фреймворка делятся на два типа: Glassbox и Blackbox. Подразумевается, что первые генерируют объяснения понятные для человека в то время, как модели Blackbox (черный ящик) предоставляют приблизительные. Стоит отметить, что в состав Glassbox моделей входит Explainable Boosting Machine (EBM). Эта аддитивная модель, основанная на дереве с автоматическим определением взаимодействия.

A. Применение методов объяснения

К обученным ранее моделям были применены вышеупомянутые методы интерпретации. Часть графиков, построенных в качестве результата, представлены на рис. 1–2. Данные объясняющие технологии позволяют наглядно оценить модель машинного обучения, сравнить важность признаков, обнаружить аномалии и закономерности в данных. Кроме этого, такие инструменты как InterpretML и ExplainerDashboard позволяют манипулировать входными данными и самому оценивать работу алгоритмов машинного обучения. В табл. III показаны результаты работы объясняющих методов.

ТАБЛИЦА II РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ ОБЪЯСНЯЮЩИХ МЕТОДОВ

Оценка скорости объяснения				
Model	***Shap***	***Lime***	***ExplainerDashboard***	***Dalex***
Logistic Regression	+	+	+	+
SVM	-	+	t+*	t+
Linear SVC	-	+	t+*	t+
Random Forest	+	+	+	+
KNN	-	+	t+*	t+
Naive Bayes	-	+	t+*	+
Decision Tree	+	+	+	+
XGB	+	+	+	+
Ridge	+	+	t+*	+
Bagging	-	+	t+*	+
Extra Trees	+	+	+	+
AdaBoost	-	+	t+*	+
SGD	+	-	t+*	+
Logistic Regression	+	+	+	+

Большинство из них успешно выводят объяснения, но часть из них, отмеченные “-“ требуют дополнительной настройки для правильного выполнения. Также стоит отметить, что объяснение, выполняемое на модели помеченное “t+” требует большого количества времени для работы и для датасета, используемого в работе, может достигать нескольких десятков часов, “t+*” – время, в котором учитывается вычисление shap value. В тоже время остальные модели интерпретируются относительно быстро (в пределах одной минуты).

По результатам работы можно сделать вывод, что для задачи обучения точной и объяснимой модели машинного обучения рекомендуется использовать XGB классификатор, Explainable Boosting Machine, логистическую регрессию и Extra Trees классификатор, если основываться на точности и скорости интерпретации модели.

IV. ПРИЧИННО-СЛЕДСТВЕННЫЕ СВЯЗИ

При всех достоинствах системы искусственного интеллекта, их легко обмануть или сбить с толку незнакомыми ситуациями, которые человек может легко решать ввиду своего жизненного опыта. Например, можно обучить нейронную сеть определять пойдет ли дождь в случае, когда небо затянуто тучами. Но сеть никогда не “догадается”, что именно из туч дождь и идет сеть никогда не “догадается”, что именно из туч дождь и идет. Понимание следствий, что из одних вещей вытекают другие, позволило бы избежать обучения систем определять каждый новый случай и позволило бы значительно сократить объемы хранимых данных.

При постановке задачи определение причинно-следственных связей, в первую очередь, следует обращаться к трудам Джуда Перла [17], который является основоположником теории причинности. В одной из первых работ Д. Перл показал, что корреляцию нельзя считать причинно-следственной связью. Проблема, о которой упоминал автор еще в 2020 году лежит даже в области глубокого обучения, которое

способно определить взаимозависимость, но не способно определить причинно-следственную связь. Ученик Д. Перла, Элиас Барейнбойм [18], создатель каузальной Байесовской машины по поиску переменных, которые оказывают влияние на другие переменные [19], считает, что работу над задачами поиска причин того или иного явления надо начинать не со сбора данных, а с использования причинно-следственной логики Перла и каузальной Байесовской машины.

Основное преимущество подходов на основе причинности в том, что причинно-следственные связи можно переносить в другие предметные области и использовать там, как человек использует опыт. Классические подходы к обучению, которые используются сейчас, не способны на такое.

Еще одна важная деталь, которая обосновывает необходимость разработки инструментов определение причинно-следственных связей, это место человека при взаимодействии с системой искусственного интеллекта. Когда у второй появится способность определять причинность, человек сможет делиться своим опытом, обогащая систему, а та, в свою очередь, сможет начать обогащать человека (коэволюционирующий гибридный интеллект [4]).

V. Заключение

Классические методы объяснимого искусственного интеллекта (LIME, SHAP и их модификации) продолжают достаточно эффективно решать задачи объяснения конкретных областей, например, при анализе сердечно-сосудистых заболеваний. Однако, определяемая ими взаимосвязь конкретных параметров данных с конечным результатом модели не отвечает на вопрос о причинах заболевания. Для ответа на вопрос о причинах и причинно-следственных связях следует прибегать к теории причинности Д. Перла, для которой уже есть прецеденты практических реализаций – каузальная Байесовская машина, исследование которой будет проведено в дальнейших работах.

Список литературы

[1] Шевская Н.В. Объяснимый искусственный интеллект и методы интерпретации результатов. Моделирование, оптимизация и информационные технологии. 2021;9(2). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1005 DOI: 10.26102/2310-6018/2021.33.2.024

[2] Shevskaya N.V. Explainable Artificial Intelligence Approaches: Challenges and Perspectives //2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS). IEEE, 2021. С. 540-543.

[3] Popov N.V., Shevskaya N.V. Explainable Artificial Intelligence Methods Based on Feature Space Analysis //2021 IV International Conference on Control in Technical Systems (CTS). IEEE, 2021. С. 242-245.

[4] Krinkin, K., Shichkina, Y., & Ignatyev, A. (2021, September). Co-evolutionary hybrid intelligence. In 2021 5th Scientific School Dynamics of Complex Networks and their Applications (DCNA) (pp. 112-115). IEEE.

[5] Cardiovascular Disease dataset | Kaggle .[электронный ресурс] URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset (дата обращения 2022-04-04)

[6] 5 способов обнаружить выбросы. [электронный ресурс] URL: https://www.machinelearningmastery.ru/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623/ (дата обращения 2022-04-04)

[7] scikit-learn. Machine Learning in Python. [электронный ресурс] URL https://scikit-learn.org/stable/(дата обращения 2022-04-04)

[8] 3.3. Metrics and scoring: quantifying the quality of predictions. [электронный ресурс] URL: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения 2022-04-04)

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

[10] 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.

[11] SHAP documentation. [электронный ресурс] URL: https://shap.readthedocs.io/en/latest/ (дата обращения 2022-04-04)

[12] Local Interpretable Model-Agnostic Explanations (lime) [электронный ресурс] URL: https://lime-ml.readthedocs.io/en/latest/# (дата обращения 2022-04-04)

[13] Explainerdashboard. [электронный ресурс] URL: https://explainerdashboard.readthedocs.io/en/latest/ (дата обращения 2022-04-04)

[14] Dalex. [электронный ресурс] URL: https://dalex.drwhy.ai (дата обращения 2022-04-04)

[15] Understand Models. Build Responsibly [электронный ресурс] URL: https://interpret.ml/ (дата обращения 2022-04-04)

[16] B. Bergsteinarchive // What AI still can’t do. February 19, 2020. [электронный ресурс] URL: https://www.technologyreview.com/2020/02/19/868178/what-ai-still-cant-do/ (дата обращения 2022-04-04)

[17] Pearl J., Mackenzie D. The book of why: the new science of cause and effect. Basic books, 2018.

[18] Elias Bareinboim Personal Page. [электронный ресурс] URL: https://causalai.net/ (дата обращения 2022-04-04)

[19] Bareinboim E., Brito C., Pearl J. Local characterizations of causal Bayesian networks //Graph Structures for Knowledge Representation and Reasoning. – Springer, Berlin, Heidelberg, 2012. С. 1-17.