УДК 004.89

ИСПОЛЬЗОВАНИЕ КОНТРФАКТИЧЕСКИХ ОБЪЯСНЕНИЙ В ИНТЕРПРЕТАЦИИ МОДЕЛЕЙ

Дубровина Н.П., студент гр. 09-362, IV курс
Казанский федеральный университет, г. Казань

В условиях стремительного роста применения систем машинного обучения в реальных социально и юридически значимых сферах, таких как принятие решений о кредитовании, отбор кандидатов, медицинская диагностика и автоматическая классификация правонарушений, возрастает не только необходимость в интерпретируемости моделей, но и требование к каузальному пониманию их поведения. Пользователю, подвергающемуся воздействию вывода модели, зачастую недостаточно просто узнать, что привело к такому результату – гораздо важнее понять, что можно было бы изменить, чтобы получить иной исход. Именно в этом контексте всё большую значимость приобретают контрфактические объяснения (counterfactual explanations) – подход, позволяющий интерпретировать модель, моделируя гипотетические, но минимально изменённые сценарии, ведущие к альтернативному выводу.

Контрфактическое объяснение отвечает на вопрос: «что должно было бы быть иным во входных данных, чтобы результат модели изменился?»

Например, если система отказала в кредите, объяснение может быть сформулировано как: «если бы ваш доход был на 500 долларов выше и кредитная история содержала на один просроченный платёж меньше, модель одобрила бы заявку». Это объяснение, в отличие от пассивного указания на важность признаков, имеет активную, ориентированную на действия природу: оно не просто описывает, почему решение было принято, но указывает на путь возможной корректировки – будь то поведения пользователя, бизнес-процесса или параметров модели.

Формально контрфактическое объяснение строится путём поиска такого изменения в пространстве входных признаков, которое:

1. Изменяет исход классификации или регрессии модели на целевой (например, с отказа на одобрение);
2. Минимально отклоняется от исходного входа, обеспечивая близость и правдоподобие;
3. Сохраняет реалистичность и допустимость (например, не может требовать отрицательного возраста или недостижимых характеристик).

Для вычисления таких объяснений применяются различные оптимизационные методы: градиентные подходы, методы обратного распространения, эволюционные алгоритмы, рекуррентные корректировки, байесовская оптимизация и методы поиска в ограниченном пространстве. При этом используется модель предсказания как «черный ящик» или в открытом виде – в зависимости от доступности внутренней структуры [1].

Контрфактические объяснения обладают несколькими уникальными свойствами, делающими их особенно ценными в контексте доверенного ИИ: [2]

- Каузальность: они направлены на изменение вывода, а не просто описание существующего состояния. Это делает их ближе к реальной причинной логике и позволяет использовать их в задачах этической и юридической оценки.
- Когнитивная понятность: поскольку они оперируют с альтернативными сценариями, они легко воспринимаются людьми, особенно в форме условных утверждений: «если бы... то тогда...».
- Адаптивность: они могут быть персонализированы под конкретного пользователя, ситуацию или контекст задачи, не требуя общей интерпретации всей модели.
- Ориентация на действия: позволяют пользователю скорректировать своё поведение или данные, чтобы получить желаемый результат, повышая тем самым не только прозрачность, но и воспринимаемую справедливость [3].

Вместе с тем, использование контрфактических объяснений сталкивается с рядом проблем и методологических вызовов. Во-первых, существует проблема реалистичности и достижимости предложенных изменений. Например, рекомендация повысить возраст или изменить пол не может быть реализована, и потому такая интерпретация будет не только бесполезной, но и потенциально дискриминационной. Для преодоления этой проблемы требуется введение ограничений на допустимость признаков: фиксированных, защищённых, недоступных для изменения. Это предполагает либо ручное кодирование допустимых изменений, либо обучение специальных моделей реалистичности [4].

Во-вторых, существует многозначность и нестабильность контрфактических решений: для одного и того же вывода может существовать множество альтернативных путей, и выбор между ними не всегда однозначен. Возникает задача оптимального выбора среди множества объяснений, что может решаться по принципу минимального изменения, максимальной интерпретируемости, вероятности сценария или даже с учётом предпочтений пользователя.

В-третьих, проблема инверсии причинности. Контрфактические объяснения, если их не проверять через каузальные модели, могут предполагать

обратную зависимость – например, что изменение симптома приведёт к исчезновению болезни, хотя, по сути, болезнь является причиной симптома. Это требует интеграции каузальных графов и структурного каузального моделирования в генерацию объяснений [5].

На практике контрфактические объяснения уже находят применение в таких областях как: [6]

- Финансовые технологии: объяснение отказа в кредите, рекомендации по улучшению профиля заёмщика;
- Здравоохранение: альтернативные сценарии прогноза состояния пациента;
- Право и безопасность: объяснение причин классификации действий как подозрительных или отклонённых;
- Образование: интерпретация систем рекомендаций и оценки кандидатов.

Разрабатываются специальные библиотеки и фреймворки, реализующие контрфактические объяснения: DiCE (Diverse Counterfactual Explanations), CARLA, Alibi, WatchFrog, которые позволяют строить объяснения для моделей как в открытом, так и в чёрном ящике. Эти фреймворки поддерживают мультимодальные данные, ограничения на изменяемость признаков, генерацию наборов альтернатив и визуализацию сценариев.

С инженерной точки зрения, контрфактические объяснения становятся частью MLOps-инфраструктуры доверенного ИИ. Они логгируются вместе с предсказаниями, верифицируются по метрикам стабильности и реалистичности, представляются в интерфейсе пользователя и могут использоваться в юридических разбирательствах или нормативной отчётности. Более того, на их основе можно строить обратную связь с моделью – например, включать их в процесс активного обучения или дообучения для коррекции систематических ошибок.

Таким образом, контрфактические объяснения представляют собой не просто способ интерпретации, а инструмент взаимодействия между моделью и человеком, основанный на альтернативных сценариях и направленный на повышение прозрачности, справедливости и адаптивности интеллектуальных систем.

Они открывают путь к таким формам доверия, которые не ограничиваются техническим объяснением, а предполагают участие пользователя в процессе принятия решений и возможность их переосмысления.

В эпоху этически чувствительного ИИ именно контрфактическое мышление может стать основой для перехода от объяснимости к подлинной ответственности.

Список литературы:

1. Wachter, S. Counterfactual Explanations without Opening the Black Box : Automated Decisions and the GDPR / S. Wachter, B. Mittelstadt, C. Russell // Harvard Journal of Law & Technology. – 2018. – Vol. 31, № 2. – P. 841–887.

2. Mothilal, R. K. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations / R. K. Mothilal, A. Sharma, C. Tan // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2020. – P. 607–617. – DOI 10.1145/3351095.3372850.

3. Ustun, B. Actionable Recourse in Linear Classification / B. Ustun, A. Spangher, Y. Liu // Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 10–19. – DOI 10.1145/3287560.3287566.

4. Karimi, A.-H. Algorithmic Recourse : From Counterfactual Explanations to Interventions / A.-H. Karimi, G. Barthe, B. Schölkopf, I. Valera // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2021. – P. 353–362. – DOI 10.1145/3442188.3445899.

5. Verma, S. Counterfactual Explanations for Machine Learning : A Review / S. Verma, J. Dickerson, K. Hines // arXiv. – 2020. – arXiv:2010.10596.

6. Russell, C. Efficient Search for Diverse Coherent Explanations / C. Russell // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 20–28. – DOI 10.1145/3287560.3287569.