УДК 004.89

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ МОДЕЛЕЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ РЕДКИХ СОБЫТИЙ

Самаков А.П., студент гр. 24226, II курс
Новосибирский государственный университет, г. Новосибирск

Задачи классификации редких событий (rare event classification) занимают особое место в прикладном машинном обучении, поскольку именно они часто связаны с высокими социальными, финансовыми или инфраструктурными рисками [1].

Предсказание эпидемических вспышек, выявление случаев мошенничества, обнаружение отказов в системах жизнеобеспечения, диагностика редких заболеваний, предсказание аварий или кибератак – всё это примеры, где доля положительных исходов чрезвычайно мала по отношению к общему числу наблюдений.

Несмотря на то, что современные модели достигают высокой точности, интерпретация их предсказаний в таких задачах остаётся крайне трудной и методологически неочевидной. Это связано не только с асимметрией данных, но и с необходимостью соблюдения строгих требований к достоверности, прозрачности и воспроизводимости интерпретаций, особенно если от них зависят решения с реальными последствиями.

Одна из фундаментальных проблем в задачах редких событий заключается в том, что модель обучается на крайне несбалансированном наборе. Даже при применении методов балансировки классов, генерации синтетических данных или переопределения функции потерь, сама структура модели остаётся смещённой в сторону преобладающего класса. Это приводит к тому, что объяснение, сгенерированное для примера положительного класса (редкого события), может базироваться на признаках, которые в общей статистике модели считаются второстепенными. В результате интерпретация становится хрупкой: она зависит от контекста конкретного случая и может не воспроизводиться на других примерах. Это делает интерпретации в задачах классификации редких событий особенно чувствительными к методологическим ошибкам, и требует дополнительных механизмов контроля доверия и верификации объяснений [2].

Кроме того, редкие события, как правило, не обладают общей предсказательной структурой, доступной через глобальные правила.

Например, в задачах финансового мошенничества или медицинской диагностики одного и того же синдрома, разные случаи могут иметь абсолютно

разные причины и проявления. Поэтому глобальные интерпретации, усредняющие поведение модели, могут не отражать суть отдельных решений. Более релевантным становится локальный анализ: интерпретация предсказания на уровне конкретного примера, через такие методы как LIME, SHAP, контрфактические объяснения или attention-визуализация. Эти методы позволяют изучать, какие именно признаки в конкретном случае повлияли на классификацию как редкое событие. Однако и здесь возникает сложность: если модель «никогда» не видела подобных объектов, то и объяснение будет строиться на крайне малых различиях, зачастую в пределах шума или коррелирующих признаков.

Сложность добавляется тем, что влияние одного и того же признака на классификацию может быть радикально разным в зависимости от конкретного подмножества входов. В задачах с редкими событиями важны не просто сильные признаки, а слабые, но специфичные – те, что срабатывают лишь в особом контексте. Такие признаки часто оказываются в тени доминирующих факторов и не попадают в глобальные объяснения.

Именно поэтому возрастают требования к семантической точности интерпретации: необходимо не просто указать, какие признаки важны, но и в каком взаимодействии они проявляют значимость. Это ведёт к интересу к мультиатрибутивным объяснениям, логическим правилам с условиями взаимодействия, и к моделям, где интерпретация учитывает взаимозависимости признаков, а не их индивидуальные веса.

Контрфактические объяснения становятся особенно важными в этом контексте. Вместо того чтобы просто указывать значимые признаки, они формулируют ответ на вопрос: «что нужно изменить, чтобы пример перестал считаться редким событием?» Это не только даёт пользователю возможность осмыслить границы решений модели, но и предоставляет возможность для формального тестирования её логики. Если такие изменения невозможны или нелогичны (например, модель предлагает изменить пол или генетические данные), это указывает на потенциальную предвзятость или ошибку в структуре модели. В задачах диагностики и безопасности контрфактические объяснения часто являются единственным способом установить причинную цепочку между входом и предсказанием.

Одной из перспективных практик становится использование аудитных фреймворков интерпретации, где каждое решение модели, относящееся к редкому классу, сопровождается пакетом интерпретационных артефактов: визуализацией вклада признаков, сравнением с ближайшими случаями, описанием возможных контрфактических сценариев, выводом о доверительном интервале. Это позволяет создать контекстуализированную интерпретацию, учитывающую не только сам вход, но и поведение модели в его локальной окрестности. Особенно полезными такие подходы становятся в системах, где решение

человека зависит от объяснения модели, например в врачебных консилиумах или при вынесении решений по безопасной эксплуатации оборудования.

Визуализация в задачах классификации редких событий также требует особого подхода. Отображение всех признаков или использование тепловых карт может быть неинформативным из-за высокой разреженности значимых компонентов. Поэтому эффективными оказываются интерактивные визуализации, где пользователь может фокусироваться только на отклонениях от нормы, на аномальных активациях, на признаках с высоким отношением правдоподобия в положительном классе. Некоторые системы представляют предсказание как отклонение от базовой модели нормы, используя графы, временные шкалы или мультимодальные слои визуального объяснения [3].

Таким образом, интерпретация моделей в задачах классификации редких событий требует выхода за рамки стандартной логики объяснения. Здесь недостаточно указать важные признаки – требуется создание контекстуального, локального, устойчивого и верифицируемого объяснения, согласующегося с внешними знаниями и возможными действиями пользователя. Интерпретации должны учитывать специфику несбалансированных данных, отсутствие универсальных закономерностей и высокую стоимость ошибок [4].

Разработка соответствующих инструментов и методик интерпретации не только повышает доверие к ИИ-системам, но и делает возможным их применение в чувствительных и регламентируемых сценариях, где простая точность предсказания уже недостаточна без объяснения, заслуживающего доверия [5, 6].

Список литературы:

1. King, G. Logistic Regression in Rare Events Data / G. King, L. Zeng // Political Analysis. – 2001. – Vol. 9, № 2. – P. 137–163. – DOI 10.1093/oxfordjournals.pan.a004868.

2. He, H. Learning from Imbalanced Data / H. He, E. A. Garcia // IEEE Transactions on Knowledge and Data Engineering. – 2009. – Vol. 21, № 9. – P. 1263–1284. – DOI 10.1109/TKDE.2008.239.

3. Chawla, N. V. SMOTE : Synthetic Minority Over-Sampling Technique / N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. – Vol. 16. – P. 321–357. – DOI 10.1613/jair.953.

4. Lin, T.-Y. Focal Loss for Dense Object Detection / T.-Y. Lin, P. Goyal, R. Girshick [et al.] // 2017 IEEE International Conference on Computer Vision. – Venice : IEEE, 2017. – P. 2999–3007. – DOI 10.1109/ICCV.2017.324.

5. Davis, J. The Relationship between Precision-Recall and ROC Curves / J. Davis, M. Goadrich // Proceedings of the 23rd International Conference on Machine Learning. – New York : ACM, 2006. – P. 233–240. – DOI 10.1145/1143844.1143874.

6. Saito, T. The Precision-Recall Plot is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets / T. Saito, M. Rehmsmeier // PLoS ONE. – 2015. – Vol. 10, № 3. – Article e0118432. – DOI 10.1371/journal.pone.0118432.