УДК 004.89

МЕТОДЫ СОВМЕСТНОЙ ИНТЕРПРЕТАЦИИ МОДЕЛИ И ДАННЫХ В ИНТЕРАКТИВНОЙ СРЕДЕ

Моисеен Д.И., студент гр. 053, I курс
Томский государственный университет, г. Томск

В условиях стремительного развития методов машинного обучения и повсеместного внедрения искусственного интеллекта в разнообразные сферы деятельности возрастает необходимость не только создавать точные и эффективные модели, но и обеспечивать их прозрачность, объяснимость и возможность глубокой аналитики [1].

Одним из современных и перспективных направлений в области объяснимого ИИ является разработка и внедрение методов совместной интерпретации модели и данных в интерактивной среде. Такой подход позволяет перейти от пассивного ознакомления с результатами предсказаний и статическими объяснениями к активному, исследовательскому взаимодействию с моделью и данными, что существенно расширяет возможности понимания и контроля за поведением интеллектуальных систем [2].

Традиционные методы интерпретации моделей зачастую ориентированы либо на глобальное объяснение – выявление общих закономерностей и факторов, влияющих на решения модели во всей обучающей выборке, либо на локальное объяснение конкретного предсказания. Однако эти подходы редко учитывают глубокую взаимосвязь между особенностями данных и поведением модели в динамическом режиме, что становится особенно критично при работе с большими, высокоразмерными и гетерогенными наборами данных. Совместная интерпретация модели и данных в интерактивной среде обеспечивает синергетический эффект, позволяя пользователю одновременно наблюдать за структурой данных, статистическими характеристиками, аномалиями и при этом понимать, каким образом эти аспекты отражаются на предсказаниях и объяснениях модели [3].

Интерактивность здесь выступает ключевым элементом, позволяя аналитикам и специалистам доменной области не просто получать отчёты и графики, но и активно манипулировать входными параметрами, фильтровать и сегментировать данные, изменять условия анализа и наблюдать, как изменяются интерпретации. Это становится особенно ценным при работе с комплексными задачами, где решение модели зависит от множества факторов и взаимосвязей, которые трудно учесть при статическом анализе. Так, с помощью

интерактивных дашбордов и визуальных инструментов можно выявлять, например, сегменты данных, где модель демонстрирует низкую точность, и одновременно исследовать, какие именно признаки или их сочетания вызывают такие ошибки.

Технологии визуализации занимают центральное место в подобных системах. Для данных различных типов применяются специализированные методы: тепловые карты, диаграммы рассеяния с возможностью кластеризации, графовые представления, временные ряды с интерактивным масштабированием, а также визуализации объяснений – карты важности признаков, attention-механизмы и другие. Совмещение этих видов визуализации в одном интерфейсе даёт комплексное понимание, позволяя связать статистические особенности и динамику данных с причинно-следственными эффектами, выявляемыми моделью. Например, в задачах компьютерного зрения можно одновременно видеть распределение классов, области изображения с высокой важностью для предсказания и визуальные аномалии, что способствует более осознанному принятию решений.

Еще одним значимым компонентом является реализация возможностей «что если» (what-if) анализа, когда пользователь модифицирует входные данные, симулирует сценарии и изучает, как эти изменения отражаются на выводах модели и их интерпретациях. Такой интерактивный анализ помогает выявлять уязвимости модели, зависимость предсказаний от ключевых признаков и потенциальные риски неправильной работы в новых условиях. В совокупности с мониторингом моделей в реальном времени и аналитикой по дрейфу данных, эти методы формируют основу комплексного контроля за жизненным циклом моделей.

Особое внимание в современных решениях уделяется возможностям интеграции с корпоративными системами и платформами для машинного обучения (MLOps), что обеспечивает непрерывный сбор данных, отслеживание изменений, аудит и повторное обучение моделей. Совместная интерпретация в таких средах позволяет не только визуализировать результаты, но и активно управлять моделями, выявлять причины ухудшения качества и принимать решения о корректирующих действиях. Кроме того, она способствует соблюдению нормативных требований, например, в области защиты персональных данных и прозрачности автоматизированных решений, что становится всё более актуальным с ростом регулирования искусственного интеллекта [4].

В перспективе методы совместной интерпретации модели и данных будут развиваться в сторону глубокой интеграции с системами искусственного интеллекта следующего поколения, включая самонастраивающиеся и адаптивные модели, а также системы с элементами коллективного разума и взаимодействия человека и машины. Это позволит создавать среды, в которых

интерпретация и понимание станут органичной частью процесса разработки, внедрения и эксплуатации ИИ, а не отдельным вспомогательным шагом.

Таким образом, методы совместной интерпретации модели и данных в интерактивной среде открывают новые горизонты для повышения доверия, качества и управляемости современных интеллектуальных систем [5].

Они обеспечивают не только прозрачность и объяснимость, но и позволяют выявлять и устранять узкие места, делая ИИ более ответственным, адаптивным и ориентированным на реальные потребности пользователей и бизнеса [6].

Список литературы:

1. Mitchell, M. Model Cards for Model Reporting / M. Mitchell, S. Wu, A. Zaldivar [et al.] // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 220–229. – DOI 10.1145/3287560.3287596.

2. Gebru, T. Datasheets for Datasets / T. Gebru, J. Morgenstern, B. Vecchione [et al.] // Communications of the ACM. – 2021. – Vol. 64, № 12. – P. 86–92. – DOI 10.1145/3458723.

3. Raji, I. D. Closing the AI Accountability Gap : Defining an End-to-End Framework for Internal Algorithmic Auditing / I. D. Raji, A. Smart, R. N. White [et al.] // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2020. – P. 33–44. – DOI 10.1145/3351095.3372873.

4. Amershi, S. Software Engineering for Machine Learning : A Case Study / S. Amershi, A. Begel, C. Bird [et al.] // 2019 IEEE/ACM 41st International Conference on Software Engineering. – Montreal : IEEE, 2019. – P. 291–300. – DOI 10.1109/ICSE-SEIP.2019.00042.

5. Kaur, H. Interpreting Interpretability : Understanding Data Scientists' Use of Interpretability Tools for Machine Learning / H. Kaur, H. Nori, S. Jenkins [et al.] // Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. – New York : ACM, 2020. – P. 1–14. – DOI 10.1145/3313831.3376219.

6. Hohman, F. Visual Analytics in Deep Learning : An Interrogative Survey for the Next Frontiers / F. Hohman, M. Kahng, R. Pienta, D. H. Chau // IEEE Transactions on Visualization and Computer Graphics. – 2019. – Vol. 25, № 8. – P. 2674–2693. – DOI 10.1109/TVCG.2018.2843369.