ПРОБЛЕМА КАЧЕСТВА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

А.А. Халин

ФГБОУ ВО «Курский государственный университет», Курск, Российская Федерация
E-mail: khalin_andrey@rambler.ru

В работе проводится обсуждение проблемы качества наборов данных, необходимых для обучения или дообучения моделей машинного обучения. Рассматриваются способы намеренного изменения данных с целью воздействия на модель машинного обучения заранее заданным образом. Приводятся способы защиты от проводимых атак.

Ключевые слова: машинное обучение; наборы данных; качество данных; атака отравления данных.

В последние годы продолжает наблюдаться высокий интерес как к искусственному интеллекту (ИИ) в целом, так и к машинному обучению, являющемуся одним из направлений в ИИ, в частности. Это подтверждается, например, ежегодными отчетами, выпускаемыми Стэнфордским университетом [1], анализами публикационной активности [2] и т.д.

При этом важным вопросом при развитии методов машинного обучения продолжает оставаться подбор качественных наборов данных («набор данных (dataset): совокупность данных, в том числе соответствующих им метаданных, организованных по определенным правилам и принципам описания» [3]), на которых будет производиться обучение модели. Например, в [4] качество данных (наборов данных) определяется, как понимание или оценка их пригодности для достижения цели, поставленной в рамках предметной области. В [5] же под качеством понимается «степень соответствия совокупности присущих характеристик объекта требованиям». Там же можно найти, что «данные – факты об объекте». Можно привести еще ряд качества данных, но все они в общих чертах будут схожи.

Проблема качества данных появилась не вчера. Во все времена от этого зависело очень много. Можно привести большое количество примеров из истории, но данная работа не целена на это, поэтому не будем здесь на этом останавливаться на этом. Отметим лишь, что попытки разработки «широкой» концепция качества данных предпринимаются не так давно. Здесь в качестве примера можно привести, например, работу [6], в которой авторы выявили общие закономерности, связанные с качеством данных, а также привели рекомендации для специалистов по информационной безопасности.

Несмотря на сказанное выше, именно сейчас проблема качества данных стоит особо остро. В том же отчете [1], говорится, что медицина является областью, в которой искусственный интеллект и машинное обучение используются наиболее активно. При этом растет количество дезинформации и подделок. И если речь идет о жизни и здоровье людей, то нельзя считаться с такими подделками, какими бы они ни были. Поэтому перейдем теперь к обсуждению вопроса о намеренном изменении данных.

В проводимых исследованиях мы можем использовать различные модели машинного обучения: разработанные непосредственно нами или же разработанные и предобученные другими специалистами, но в любом случае при работе с этими моделями мы будем использовать некоторые данные. Чаще всего эти данные берутся из открытых источников, например, распространенным хранилищем различных наборов данных является система Kaggle. При этом у нас нет никакой уверенности в качестве этих данных, которые могут быть модифицированы (испорчены) как умышленно, так и неумышленно. Здесь стоит отметить, что нами никак не рассматривается тот факт, что воздействие может происходить непосредственно на модель машинного обучения (МО) с целью ее реакции заданным образом. Также мы не касаемся того факта, что предобученная модель может давать некорректные результаты, так как предобучение проходило на испорченных данных. И получение некорректных результатов будет сохраняться в такой модели даже если мы потом используем качественные данные. Если мы говорим о неумышленной «порче» данных, то она может быть связана с ошибками, допущенными при проведении разметки, и связана во многом с уровнем подготовки специалистов, занимающихся разметкой.

Говоря же о преднамеренной модификации данных, можно выделить следующие способы такого воздействия [7-8]:

- Изменение разметки данных;
- p-фальсификация;
- Использование чистых меток (атака столкновением признаков, атаки уклонением и др. [9 - 10]);
- Непосредственная подготовка некачественных данных и т.д.

При наличии такого числа различных способов проведения атак на модели МО естественным является вопрос противостояния атакам. Выделяют несколько способов противостояния атакам на модели МО [11]:

- Методы маскировки градиента;
- Оборонительная дистилляция;
- Состязательное обучение;
- Реконструкция входных данных и т.д.

При этом нужно отметить, что несмотря на наличие способ противостояния нужно быть аккуратными при их использовании, так как имеются способы их обхода, например, в работах [12 - 13] говорится о таких обходах. Таким образом, как бы нам не хотелось положиться на те или иные методы защиты, необходимо тщательно подбирать набор данных и/или модель МО для работы, чтобы не получить результат, который принесет дополнительные проблемы в последующих исследованиях или при внедрении их, например, в здравоохранение, образование и т.д.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. 2023 AI Index Report [Электронный ресурс]. URL: https://aiindex.stanford.edu/report/ (дата обращения: 20.13.2024).
2. *Hao, K.* We analyzed 16,625 papers to figure out where AI is headed next // MIT Technol. Rev. 2019. С. 1-19.
3. ГОСТ Р 59898– 2021. Национальный стандарт российской федерации. ОЦЕНКА КАЧЕСТВА СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 26 ноября 2021 г. 1620-ст. Москва: Российский институт станд, 2021. С. 24.
4. *Rupa Mahanti.* Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. Quality Press. 2019.
5. ISO 9000:2015(ru) Системы менеджмента качества. Основные положения и словарь [Электронный ресурс]. URL: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:ru (дата обращения 20.13.2024).
6. *Strong D. M., Lee Y. W., Wang R. Y.* Data quality in context //Communications of the ACM. Т. 40. №. 5. 1997. С. 103-110.
7. *Намиот Д. Е.* Введение в атаки отравлением на модели машинного обучения //International Journal of Open Information Technologies. Т. 11. №. 3. 2023. С. 58-68.
8. *Anley C.* Practical attacks on machine learning systems. 2022.
9. *Костюмов В. В.* Обзор и систематизация атак уклонением на модели компьютерного зрения //International Journal of Open Information Technologies. Т. 10. №. 10. 2022. С. 11-20.
10. *Пришлецов Д. Е., Пришлецов С. Е., Намиот Д. Е.* Камуфляж как состязательные атаки на модели машинного обучения //International Journal of Open Information Technologies. Т. 11. №. 9. 2023. С. 41-49.
11. Атаки и методы защиты в системах машинного обучения: анализ современных исследований / Котенко И. В. [и др.] //Вопросы кибербезопасности. №. 1. 2024. С. 59.
12. *Athalye A., Carlini N., Wagner D.* Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples //International conference on machine learning. PMLR. 2018. С. 274-283.
13. *Carlini N., Wagner D.* Towards evaluating the robustness of neural networks //2017 IEEE symposium on security and privacy (sp). IEEE. 2017. С. 39-57.