

**Доверенный искусственный интеллект**

**Обнаружение аномалий**

А.Е. Самотуга, П.С. Ложников

**ОБНАРУЖЕНИЕ АНОМАЛИЙ**

После осуществления разработки модели ИИ, ее валидации, определения характеризующих ее метрик, сравнения ее с другими моделями и наконец выбора наилучшей из них, происходит внедрение модели в некоторый бизнес-процесс. Мы уже упоминали то, что сейчас все больше следует говорить не просто о машинном обучении, а скорее о направлении под название ModelOps или даже о его дальнейшем развитии в виде MLOps, которые аналогично DevOps (Методология автоматизации технологических процессов сборки, настройки и развёртывания программного обеспечения) подходу предполагают непрерывную интеграцию непрерывное развертывание (CI/CD) но применительно к моделям. Данный подход предполагает, что помимо прочего, будет осуществляться непрерывное наблюдения за поведением, характеристиками, метриками модели, с целью выявления проблем в ее работе, выявлении ухудшения ее характеристик. В случае, если такие проблемы будут обнаружены, модель должна быть отправлена на анализ и переобучение, либо, при необходимости, может быть осуществлена ее замена. Подход такого рода предполагает наблюдение за рядом проблем такого рода, как дрейф данных, дрейф концепций и ряда других.

## Определение дрейфа данных и других проблем

Изменение независимых переменных (ковариат или входных признаков) модели потенциально вызывает изменение в совместном распределении входных и выходных переменных. Компоненты, содержащие технологию искусственного интеллекта, можно проверять на наличие источников дрейфа данных в контексте анализа рисков безопасности, и при необходимости можно планировать адекватные меры. Дрейф данных часто связан с неполным представлением входной области во время обучения. Это может быть связано, например, с неспособностью учитывать сезонные изменения входных данных, непредвиденный ввод данных операторами или добавление новых датчиков, которые становятся доступными в качестве входных функций. Естественно, дрейф данных становится проблемой, как только модель выходит из строя из-за изменения границ решения модели.

Некоторые примеры дрейфа данных можно отнести к неспособности применить передовой опыт в разработке моделей. Типичные примеры включают выбор неподходящих обучающих данных, данных, распределение которых не отражает фактическое распределение, встречающееся в контексте приложения, или пропуск важных примеров в обучающих данных. Эти проблем могут быть устранены путем улучшения моделирования и переобучения.

Дрейф данных также может быть вызван внешними факторами, такими как сезонные изменения или изменения в процессе, вызывающие дрейф данных. Примеры включают замену датчика новым вариантом с другими показателями точности измерения или с другими условиями между обучением и ранее невиданными данными. Модель может столкнуться с дрейфом данных уже в развернутом состоянии, когда переобучение может оказаться невозможным. В этих случаях модель может быть построена для оценки поправочных коэффициентов на основе особенностей входных данных или допускает контролируемую коррекцию. Ожидается, что конструкция модели обеспечит безопасные выходные данные даже при наличии ранее неизвестных входных данных. Важно понимать, что даже соблюдение надлежащих методов разработки моделей, таких как создание достаточно разнообразного обучающего набора данных, не устраняет необходимость тщательного анализа того, можно ли результирующую модель обобщить на производственные данные.

На данный момент выявлены наиболее вероятные источники дрейфа данных, как и направления по улучшению моделей таким образом, чтобы они были эффективны даже в случаях простого ковариативного без существенного влияния на результаты классификации.

Дрейф концепции относится к изменению взаимосвязи между входными переменными и выходными данными модели и может сопровождаться изменением распределения входных данных. Например, выходные данные модели могут использоваться для измерения приемлемого минимального расстояния оператора во время выполнения на основе измерений расстояния, полученных датчиком времени полета (входные данные). Если принятые пределы безопасности изменяются из-за внешних факторов (например, увеличение скорости машины, не учтенное в модели), происходит дрейф концепции, в то время как и процесс, и входные данные остаются прежними.

В идеале системы будут включать в себя формы обнаружения дрейфа, отличать дрейф от шума, присутствующего в системе, и адаптироваться к изменениям с течением времени. Потенциальные подходы включают такие модели, как метод раннего обнаружения дрейфа (EDDM), обнаружение дрейфа с помощью опорных векторных машин или наблюдение ошибки вывода во время обучения, чтобы позволить обнаружить дрейф и потенциальную адаптацию.

## Утечка данных

В некоторых ML-системах злоумышленник, отправляя в систему специально подобранные объекты, способен получить информацию о том, какие объекты использовались в обучающей выборке. Это может представлять угрозу, если объекты содержат конфиденциальную информацию (например, персональные медицинские данные) либо если сам факт использования объекта в обучении не должен разглашаться (например, преступник может обнаружить, что его фотография использовалась для обучения полицейской системы распознавания).

При этом, злоумышленнику требуется полный доступ к ML-модели, чтобы многократно тестировать её, прежде чем он сможет добыть нужную информацию. Поэтому один из методов защиты – использовать многоуровневые облачные ML-модели вместо моделей в пользовательских продуктах. Другая хорошая идея – анонимизировать данные, которые применяются в обучении. Наконец еще одно направление, это использование моделей.

1. **Обнаружение заражения данных**

Механизмы обнаружения аномального нейронного поведения или аномальных входных данных могут быть использованы для выявления ситуаций возможного отказа. Например, мониторинг может обнаружить, когда технология ИИ производит потенциально небезопасные действия, либо из-за функциональных недостатков, либо из-за неисправности, после чего могут быть приняты некоторые меры для поддержания системы в безопасном состоянии. Мониторинг обычно работает параллельно с технологией ИИ, которая используется для обеспечения функциональной безопасности. Мониторинг может разрабатываться либо с использованием методов, отличных от ИИ, либо с использованием ИИ. В последнем случае для обоснования такого подхода можно было бы использовать соображения, касающиеся уровня независимости между мониторингом и первичной системой ИИ.

Для разработки приложений мониторинга рассматриваются четыре момента:

— тип неисправностей технологии ИИ, которые могут быть обнаружены;

— способы выявления дефектов технологии ИИ во время его выполнения;

— контрольные показатели эффективности различных мониторингов времени выполнения;

— виды вмешательства, которые могут быть использованы для обхода неисправности после обнаружения.

1. **Программные продукты для обнаружения аномалий в данных.**

Для обнаружения аномалий в данных в том числе, их отравления, можно воспользоваться обширным опытом, который накоплен в направлении поиска аномалий данных при их анализе. В частности, принято выделять следующие типы аномалий данных - точечные, контекстуальные и коллективные. В зависимости от принадлежности к тому или иному классу могут применяться разные методы и режимы обнаружения аномалий. Есть три режима обнаружения: режим распознавания с учителем, режим распознавания частично с учителем, режим распознавания без учителя. Что касается методов, то используются: классификация, кластеризация, статистический анализ, алгоритм ближайшего соседа, спектральные методы, гибридные методы.

Также упомянем, что использование рекуррентных нейронных сетей и автокодировщиков, механизма минимизации ложных срабатываний. Например, при помощи нейросети RNN LSTM по показателям CPU удалось обнаружить обычный выброс, который бы привел к отказу системы.

**Список источников**

1. ISO/IEC CD 24029-2 [Электронный ресурс]. URL:<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/98/79804.html> (дата обращения: 12.04.2022).

2. ISO/IEC TR 24029-1:2021 [Электронный ресурс]. URL:<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77609.html> (дата обращения: 12.04.2022).

3. Ищем аномалии и предсказываем сбои с помощью нейросетей [Электронный ресурс]. URL: https://habr.com/ru/company/krista/blog/478392/ (дата обращения: 28.04.2022).