Исследование и оценка методов противодействия атакам, реализуемым на этапе проектирования систем, основанных на искусственном интеллекте

Ерыпалов К.И (Национальный Исследовательский Университет ИТМО)

Воробьева Алиса Андреевна, кандидат технических наук, факультет информационных технологий, доцент
(Национальный Исследовательский Университет ИТМО)

В представленной работе рассмотрены и проанализированы основные виды атак на обучающие наборы данных систем, основанные на искусственном интеллекте. Изучены и применены на практике методы защиты от данных атак. Сформированы выводы о наиболее опасных видах атак и наиболее эффективных методах противодействия.

Введение.

В настоящее время активное развитие получили информационные технологии, основанные на методах искусственного интеллекта (ИИ). Методы машинного обучения применяются во многих сферах современной жизни: в банках для оценки кредитных рисков, в промышленности для управления производством и минимизации потерь, в медицине для улучшения качества диагностирования, в маркетинге для предсказания действий покупателей, в биометрических технологиях. В 2020-2021 годах атаки злоумышленников были направлены в том числе и на такие системы. В аналитических отчетах отмечают, что один из трендов развития систем информационной безопасности направлен на развитие технологий защиты, основанных на ИИ, а также обеспечение безопасности информационных систем, основанных на искусственном интеллекте. Данные факты обуславливают актуальность данного исследования.

Основная часть.

Исследование заключается в поиске оптимальных методов защиты от атак на модель, основанную на ИИ. В работе рассмотрены три основных типа атак, реализуемых на этапе проектирования систем: внедрение данных (у злоумышленника отсутствует доступ к алгоритму, происходит внедрение состязательных примеров в обучающий набор данных с целью компрометации целевой модели), модификация данных (отсутствует доступ к самому алгоритму, но присутствует к обучающим данным, что позволяет модифицировать данные обучающей выборки для изменение целевой модели до использования данных в процессе обучения) и логическое искажение (злоумышленные может вмешиваться в алгоритм обучения, получая полный контроль над моделью). Проанализированы методы противодействия указанным атакам: создание генеративно-состязательной сети, скрытие градиента и сжатие признаков. Проведены эксперименты по оценке точности модели до реализации каждого из типов атак, после реализации атак и после применения указанных методов противодействия. Проведено сравнение полученных данных, результаты экспериментов проанализированы, сделаны выводы о том, какой вред могут нанести атаки и какие способы противодействия наиболее эффективны и рациональны.

Вывод

Проведены теоретические и экспериментальные исследования возможных атак на системы, основанные на искусственном интеллекте, и методов противодействия. По результатам исследования атаки проранжированы по степени опасности для системы, определены наиболее эффективные методы борьбы.