УДК 004.852

АНАЛИЗ МЕТОДОВ СОСТЯЗАТЕЛЬНЫХ АТАК НА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Матузко М.А. (Университет ИТМО)
Научный руководитель – к.т.н. Воробьева А.А.
(Университет ИТМО)

Аннотация. Информационные системы, в основе которых лежат методы машинного обучения, могут иметь различные уязвимости, эксплуатация которых влечет нарушение информационной безопасности. В работе рассматриваются методы состязательных атак на модели машинного обучения, проводится анализ атак и производится их классификация в зависимости от знаний злоумышленника о системе.

Введение. Методы машинного обучения постоянно совершенствуются и применяются в самых разнообразных сферах современной жизни. Однако системы, использующие такие методы, могут содержать уязвимости применения незаметных для человека состязательных атак, которые могут заставить модель машинного обучения работать неправильно. Так, например, в биометрических системах добавление шума или некоторых пикселей на изображение лица человека может заставить систему неверно его идентифицировать. Такие манипуляции ведут к повышению рисков безопасности для информационных систем и могут позволить злоумышленникам получить несанкционированный доступ к информации. Ввиду того, что проблема состязательных атак на модели машинного обучения является довольно новой, на данный момент отсутствует общий анализ методов проведения таких атак. Различные группы ученых приводят разные классификации, но в целом принято разделение на основе знаний злоумышленника о системе.

Основная часть. Состязательные атаки представляют из себя способ воздействия на модель машинного обучения, заставляющий ее работать неправильно. В то время как состязательные примеры являются образцами данных, которые были целенаправленно искажены, что в последствии нарушает правильную работу классификатора.

В реальной жизни сценарий действий нарушителя и применяемые инструменты зависят от всевозможных факторов. Злоумышленники могут иметь различный уровень доступа к системе и разный уровень знаний о ней: ее параметрах и средствах защиты. Также нарушитель может преследовать различные цели, такие как: снижение точности работы модели, которая ведет к снижению достоверности классификации; неправильная классификация, при которой модель будет неверно определять классы; целевая неправильная классификация, которая заставляет модель определять объекты как класс, заранее выбранный злоумышленником; и неправильная классификация источника и цели, при которой все объекты определенного класса будут классифицироваться как другой выбранный злоумышленником класс. Таким образом, в работе будут рассмотрены возможные методы и средства применения атак при различных возможностях нарушителя.

Выводы. Несмотря на высокую точность и производительность, алгоритмы машинного обучения оказались уязвимыми для незначительных искажений, которые могут привести к опасным последствиям в сферах, связанных с безопасностью. Таким образом, все более актуальной становится необходимость разработать надежные методы защиты от состязательных атак. Проведённый анализ атак в дальнейшем может быть применён для определения модели нарушителя, оценки уязвимостей информационных систем, а также для снижения вероятности реализации состязательных атак на системы, основанные на машинном обучении.

Матузко М.А. (автор) Подпись

Воробьева А.А. (научный руководитель) Подпись