ПОЛИТЕХ
Институт кибербезопасности
и защиты информации

Лаборатория кибербезопасности

Выявление состязательных атак на системы обнаружения вторжений с помощью нейронных сетей

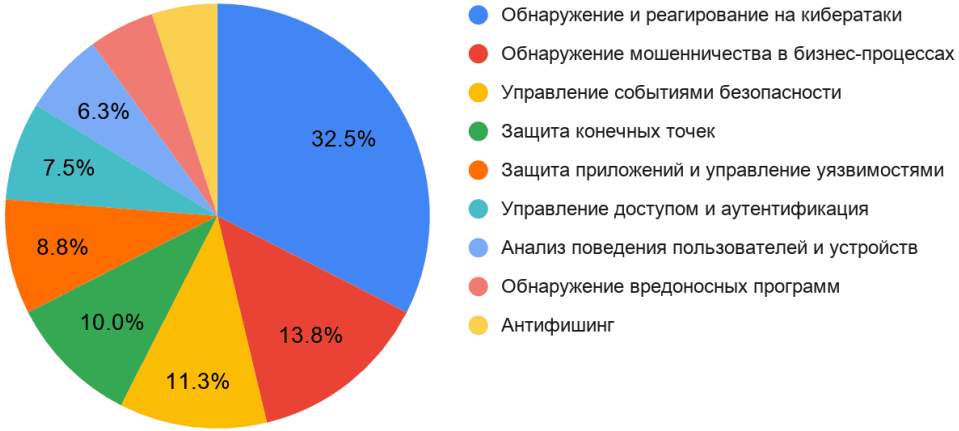
Югай Павел Эдуардович

конференция
РусКрипто

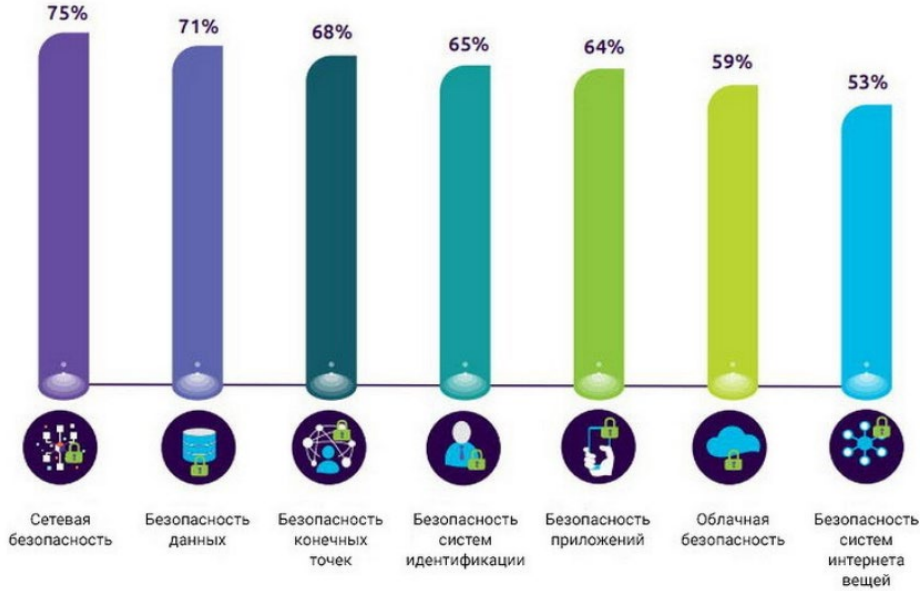
Москва, 2023

Применение технологий искусственного интеллекта в информационной безопасности

Распределение продуктов с применением технологий ИИ по сценариям использования [anti-malware.ru]

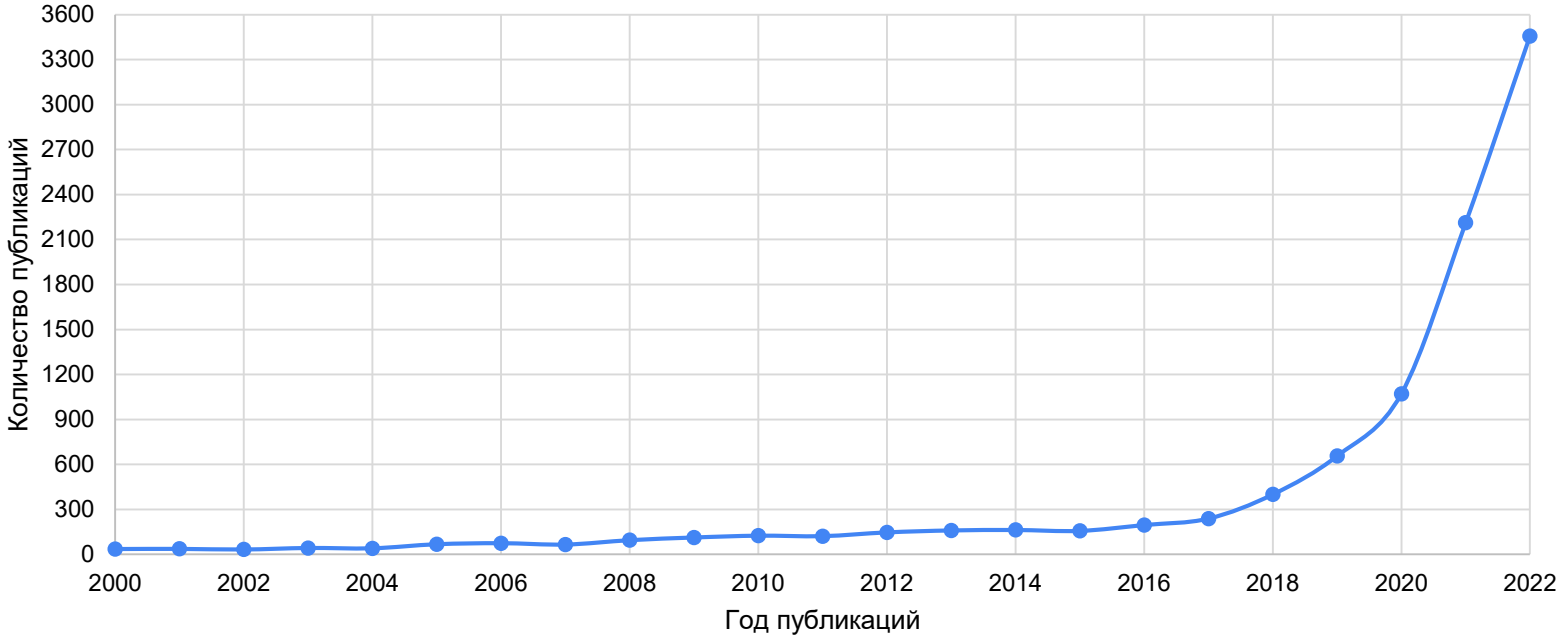


Уровень использования средств ИИ для защиты IT-инфраструктуры [safe.cnews.ru]



Исследования на тему «Состязательные атаки на машинное обучение»

Рост упоминаний темы "Advesarial attacks on machine learning" в журнале Springer



Состязательные атаки на машинное обучение

Атака отравления (Poisoning)

- Создание состязательного примера для обучающего набора данных
- Изменение границы классификации

Атака уклонения (Evasion)

- Манипулирование входными экземплярами
- Неправильная классификация состязательного примера в обученной модели

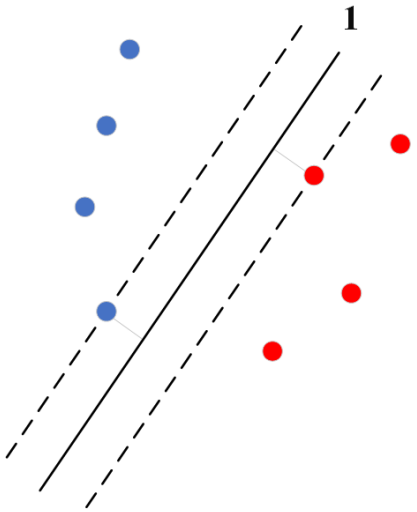
Атака извлечения модели (Model Extraction)

- Дублирование целевой модели машинного обучения

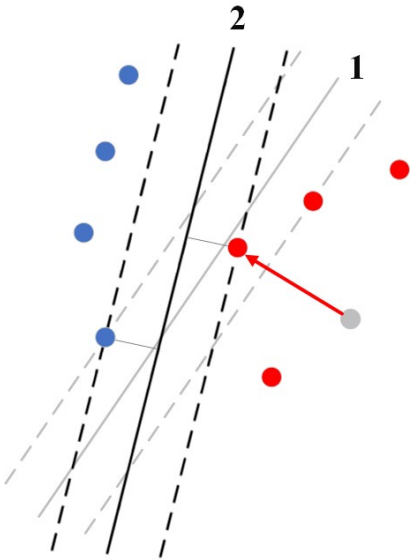
Атака отравления

Идея: Внедрение зловредных (состязательных) примеров в обучающий набор данных

Исходная граница принятия решений классификатора **до атаки**



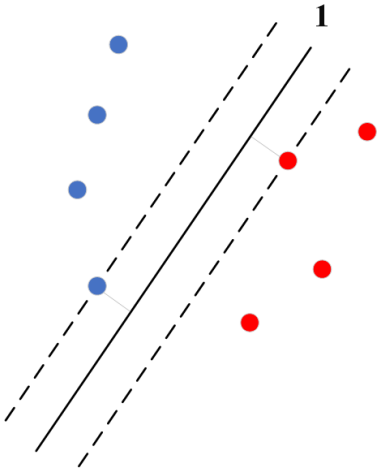
Измененная граница принятия решений классификатора **после атаки**



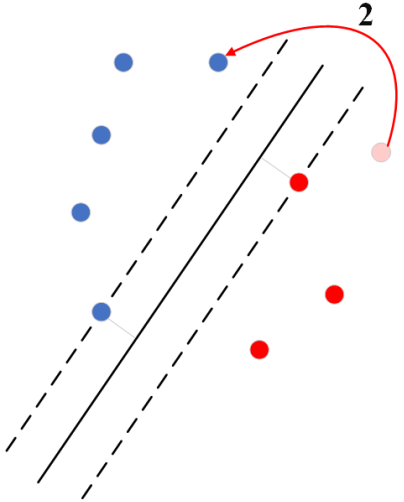
Атака уклонения

Идея: Манипулирование входными данными во время тестирования обученной модели машинного обучения

Обученная модель машинного обучения



Воздействие на входной образец и как следствие его неправильная классификация

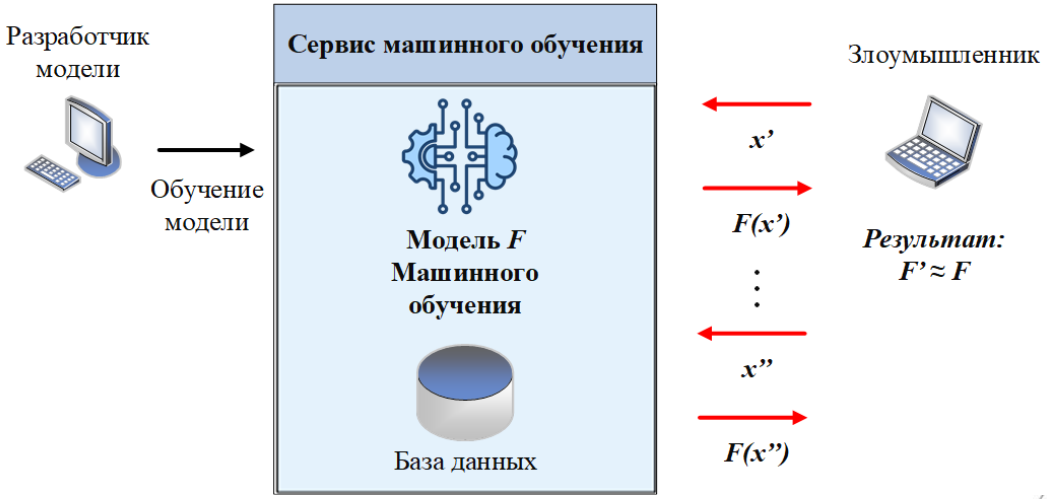


Атака извлечения модели

Идея:

- Дублирование используемой модели машинного обучения
- Принцип работы модели неизвестен (т.н. «черный ящик»)

Принцип атаки извлечения модели



Использование машинного обучения в системах обнаружения вторжений (СОВ)

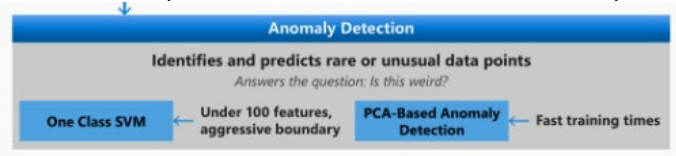
Примеры СОВ с машинным обучением

- OSSEC+
- Security Onion
- Microsoft ATA

Распространенные алгоритмы МО в СОВ:

- Метод k-ближайших соседей (KNN)
- Метод опорных векторов (SVM)
- Метод k-средних
- Логистическая регрессия
- Ансамбль методов
- Глубокое обучение

Примеры используемых ML-алгоритмов **Microsoft ATA** для обнаружения аномалий – одноклассовый метод опорных векторов (One Class SVM) и метод главных компонент (PCA)



Пример используемого ML-модуля «Logscan» **SecurityOnion** для обнаружения аномалий с помощью сканирования файлов логирования

Модели

Logscan использует следующие модели для обнаружения аномальной активности при входе в Security Onion Console:

- **K1** : поиск большого количества попыток входа с одного IP-адреса в течение 1 минуты.
- **K5** : поиск высокого процента неудачных попыток входа с одного IP-адреса в 5-минутном окне.
- **K60** : поиск аномальных моделей неудачных входов со всех IP-адресов, обнаруженных в течение 1 часа.

Примеры состязательных атак

Существуют следующие основные алгоритмы состязательных атак:

Тип атаки	Название атаки
Атака уклонения	**FGSM**
	DeepFool
	JSMA
	PGD
	BIM
	Carlini & Wagner
Атака отравления	**Feature Collision**
	SVM Poisoining
	Backdoor Attack
Извлечение модели	**Knock off Nets**
	MiFace
	Copycat CNN

Атака Fast Sign Gradient Method (FGSM)

Идея: Добавление шума, направление которого совпадает с градиентом функции стоимости по отношению к данным.

Описывается формулой:

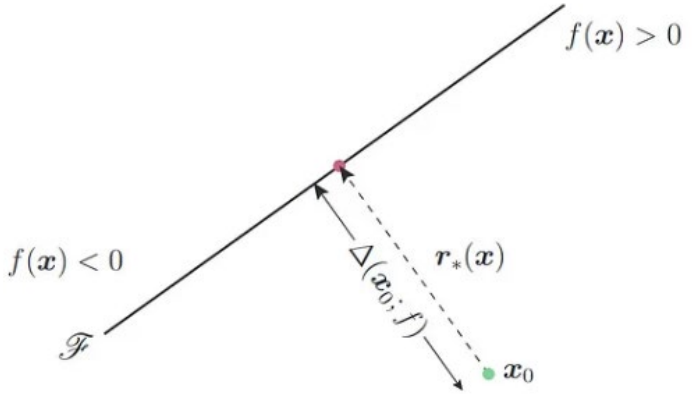
$$\boldsymbol{x' = x + k * sign(\nabla_x J(x, y))}$$

- x' – состязательный пример;
- x – входной экземпляр;
- y – результат классификации;
- $J(x, y)$ – функция потерь;
- $\nabla_x J(x, y)$ – градиент функции потерь;
- k – коэффициент возмущения.

Атака DeepFool

Идея: Найти минимальное возмущение для изменения решения классификатора

Поиск минимального смещения от линии принятия решения бинарного классификатора



Атака Jacobian Saliency Map Attack (JSMA)

Идея:

- Использование градиента
- Вычисление оценки значимости
- Циклический поиск
- Наилучший образец
- Описывается формулой:

$$label(x) = \arg\max F(x)$$

Достоинства и недостатки рассмотренных атак

Название атаки	Достоинства	Недостатки
Атака FGSM	Сравнительно малое время нахождения состязательного примера	Большое количество возмущений и небольшой показатель ошибочной классификаии
Атака DeepFool	Эффективен при создании состязательных примеров с меньшим количеством возмущений и более высоким показателем ошибочной классификации	Требует больше вычислительных ресурсов относительно FGSM и JSMA
Атака JSMA	Получение показателей влияния каждого входного параметра на результат классификации	Перебор всех возможных входных параметров, большое время вычислений

Требования к обеспечению кибербезопасности модели машинного обучения, используемой в СОВ

СОВ в облачном сервисе

- Принцип «черного ящика»

Регулярное дообучение модели МО

- Повышение вероятности обнаружения новых угроз

Многоклассовая классификация

- Повышение сложности создания состязательного примера

Формирование собственного обучающего набора данных

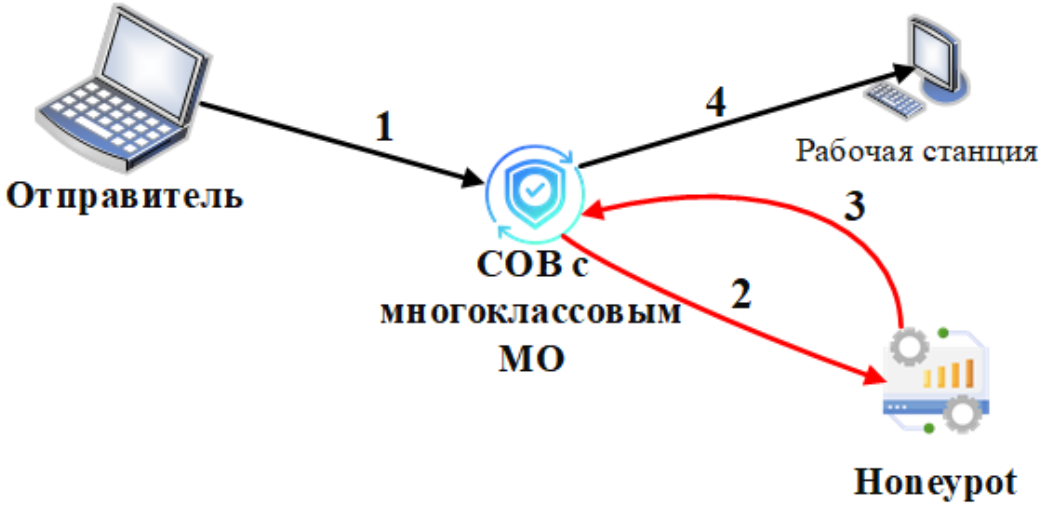
- Повышение точности благодаря уникальности обучающего набора в зависимости от потребностей и условий предприятия

Отказ от распространенных обучающих наборов данных

- Отсутствие возможности использования уже существующих состязательных примеров для известных обучающих наборов данных

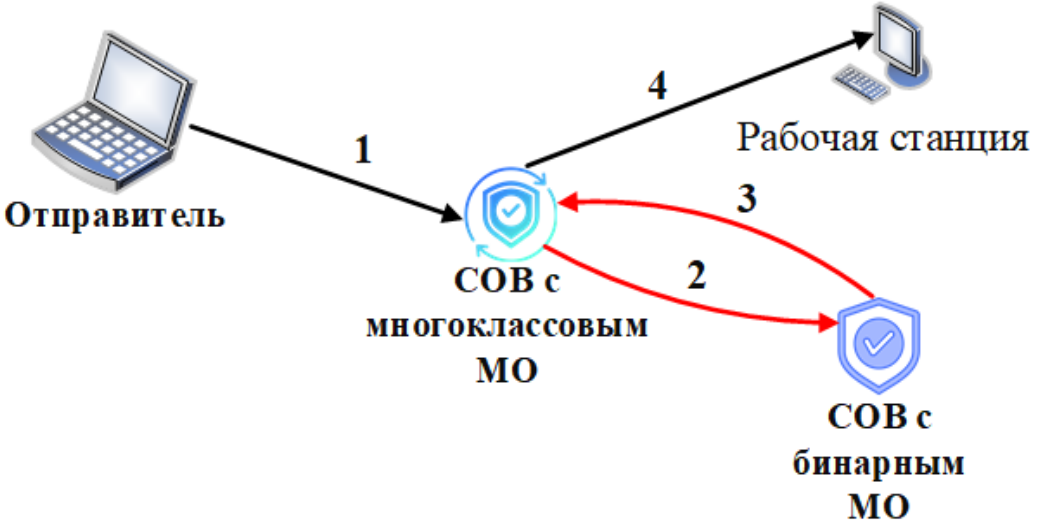
Предлагаемый подход к выявлению состязательных атак на СОВ с помощью многоклассового машинного обучения (МО) и Honeypot-систем

1. Классификация экземпляра СОВ с многоклассовым МО.
2. Если экземпляр направлен в Honeypot, то экземпляр – вредоносный, иначе шаг 4.
3. Блокировка экземпляра. Дообучение многоклассовой модели.
4. Отправка образца рабочей станции



Предлагаемый подход к выявлению состязательных атак на СОВ с помощью многоклассового МО и бинарного МО

1. Классификация экземпляра СОВ с многоклассовым МО (ММО)
2. Отправка экземпляра СОВ с бинарным МО (БМО).
3. Классификация экземпляра СОВ БМО. Отправка результата СОВ ММО.
4. Если экземпляр вредоносный, то дообучение СОВ ММО и СОВ БМО, иначе экземпляр отправляется цели.



Заключение

Наблюдается высокий рост использования алгоритмов нейронных сетей в области ИБ

Наблюдается увеличение применения состязательных атак на машинное обучение

FGSM, DeepFool и JSMA являются наиболее распространенными состязательными атаками

Сформулированы требования к модели машинного обучения

Предложены два подхода к выявлению состязательных атак на СОВ:

- **Многоклассовый МО и Honeypot-системы**
- **Многоклассовый МО и бинарный МО**