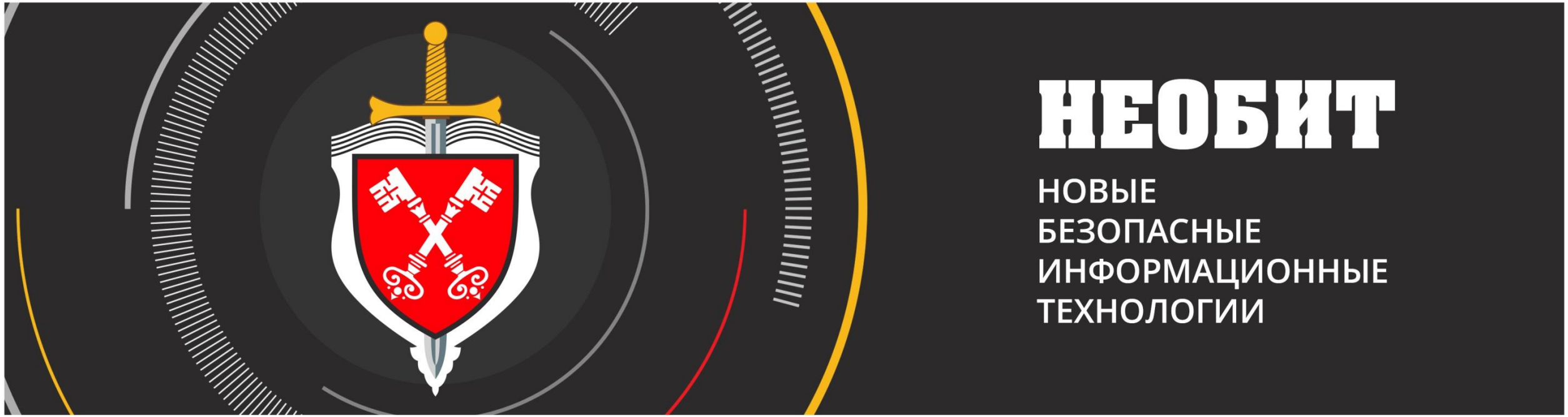


Оценка робастности методов машинного обучения, применимых в средствах защиты цифрового производства

Жуковский Е.В.

Маршев И.И.

Цифровое производство

- Автоматизация процессов производства
- Цифровое моделирование и проектирование
- Использовании искусственного интеллекта
- Сокращение участия человека

Средства защиты цифрового производства: требования

Требования к средствам защиты возрастают:

- автономность
- высокая скорость реакции
- адаптация к новым условиям работы и новым угрозам

Использование машинного обучения в средствах защиты цифрового производства

Применяется в традиционных средствах защиты для выявления:

- спама 
- сетевых атак 
- вредоносных файлов 

Удовлетворяет указанным требованиям:

- автономность (не требуется обновление, участие человека)
- высокая скорость реакции (без участия человека) 
- адаптация (возможность дообучения) 

Недостатки использования машинного обучения

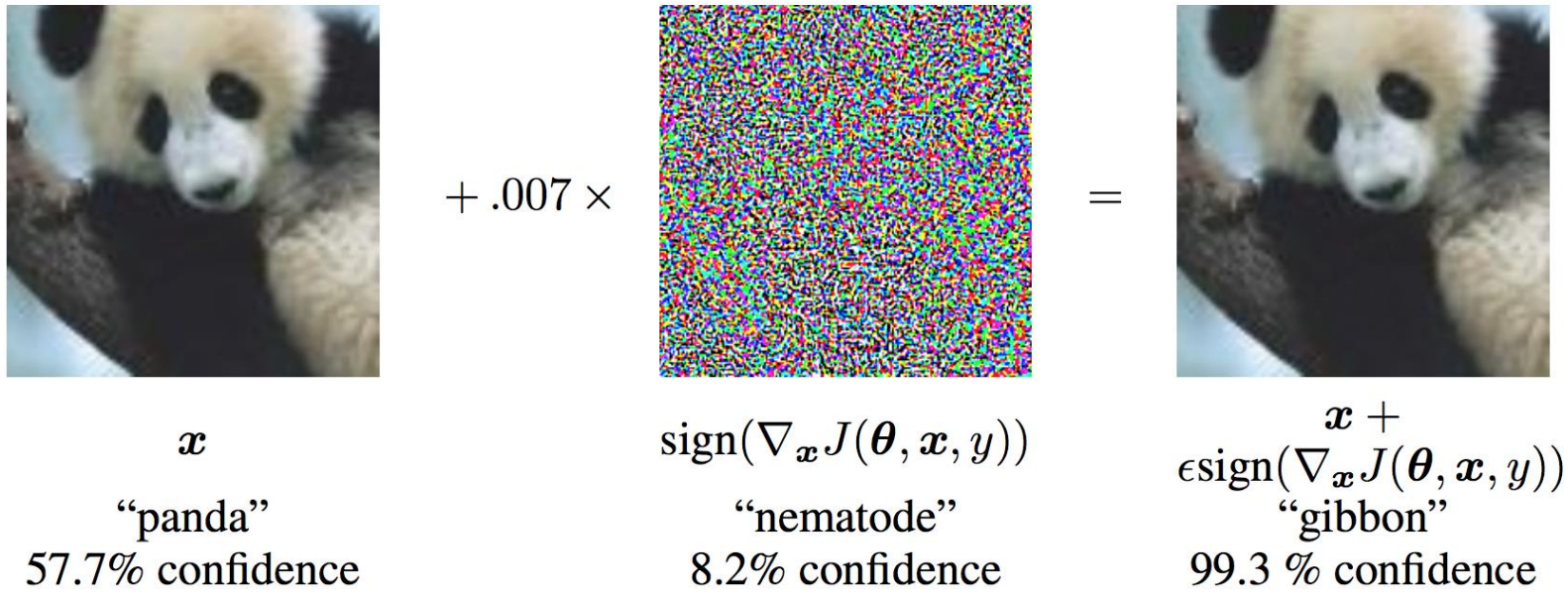
Возможность целенаправленного противодействия алгоритмам машинного обучения:

- воздействие на процесс обучения;
- модификация признаков объекта на этапе классификации.

Робастность – устойчивость статистического метода к выбросам / помехам.

Примеры противодействия машинному обучению (Adversarial attack)

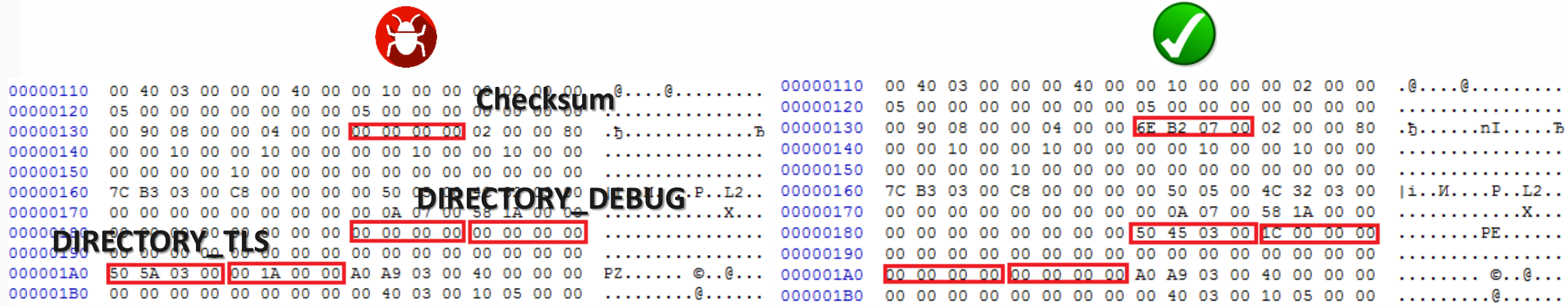
Затруднение распознавания объектов (лица, техника)



Ian J. Goodfellow, Jonathon Shlens, Cristian Szegedy.
Explaining and Harnessing Adversarial Examples. ICLR 2015.

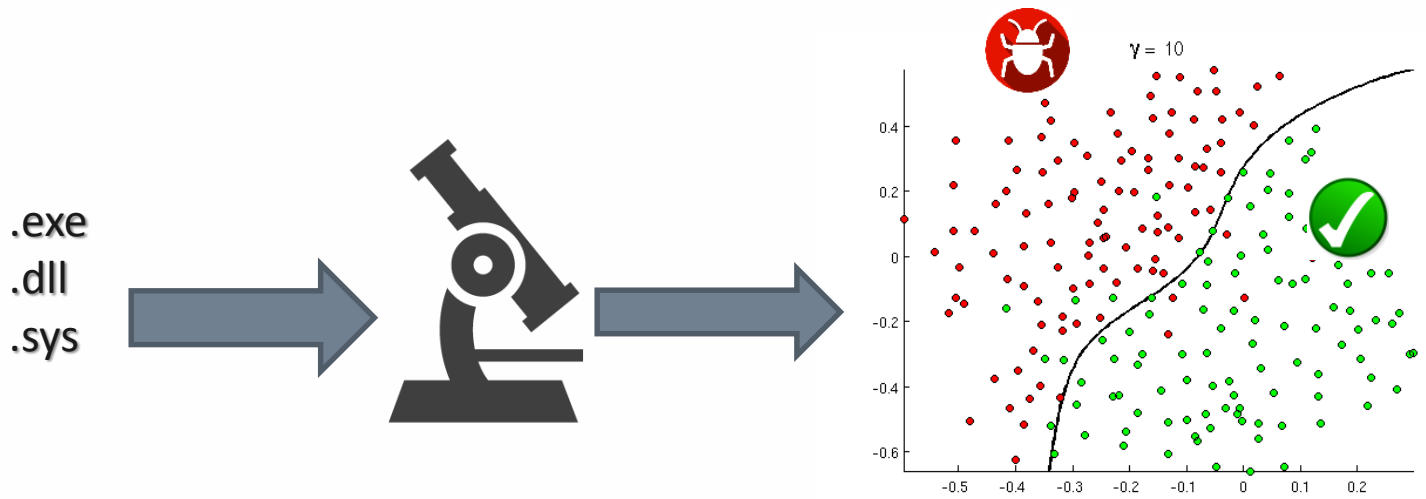
Противодействие машинному обучению, применяемого в средствах защиты

Обход средств антивирусной и сетевой защиты



Задача обнаружения вредоносных исполняемых файлов

- Вход: признаки (информация об исполняемом файле)
- Выход: определение класса, к которому относится файл (вредоносный / легитимный)

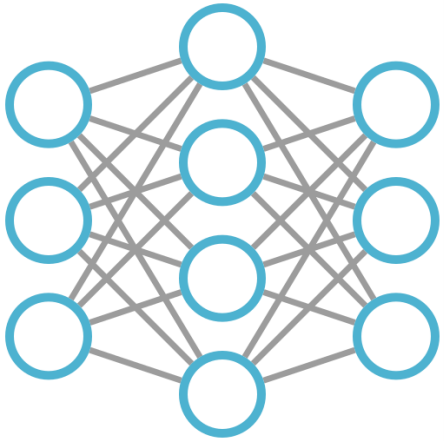


Задачи исследования

- Построить классификаторы на основе наиболее распространенных алгоритмов машинного обучения
- Оценить робастность построенных моделей (устойчивость к статистическим выбросам)

Обнаружение вредоносных исполняемых файлов: алгоритмы машинного обучения

- Случайный лес
- Дерево решений
- Метод ближайших соседей
- Метод опорных векторов



- Нейронные сети
- Байесовские классификаторы
- Градиентный бустинг
- Адаптивный бустинг

Обнаружение вредоносных исполняемых файлов: признаки

Статические

- Заголовки PE-файла
- Строки
- Таблица импорта/экспорта
- Энтропия
- N-граммы

Динамические

- Трасса выполнения
- Строки
- Названия функций

Обучающая и тестовая выборки

- Более 100 тыс. анализируемых экземпляров легитимных и вредоносных файлов
- Формировалась из различных источников

Методы оценки:

- перекрестная валидация
- случайное сэмплирование
- обучение на старых /проверка на новых экземплярах

Анализ выборки антивирусными средствами

№	Название антивируса	Количество обнаруженных файлов (шт.)	Количество обнаруженных файлов (%)
1	McAfee-GW-Edition	25413	95,96
2	McAfee	25316	95,59
3	ESET-NOD32	25109	94,81
4	AVware	24996	94,39
5	VIPRE	24983	94,34
6	NANO-Antivirus	24847	93,82
7	GData	24676	93,18
8	Symantec	24525	92,61
9	Avira (no cloud)	24431	92,25
10	Sophos AV	24325	91,85

Эффективность обнаружения ВПО

Алгоритм	Точность обнаружения, %	Ошибки первого рода, %	Ошибки второго рода, %
Случайный лес. Мера энтропии	99,267	0,339	0,394
Случайный лес. Критерий Джини	99,257	0,356	0,387
Дерево решений. Мера энтропии	98,940	0,524	0,536
Дерево решений. Критерий Джини	98,829	0,572	0,599
Метод k-ближайших соседей	97,827	1,157	1,016
Адаптивный бустинг	97,815	0,846	1,340
Наивный Байес. Распределение Бернулли	89,895	6,363	3,742
Нейронные сети	82,942	6,664	10,394
Наивный Байес. Распределение мультиномиальное	61,681	2,218	36,101
Наивный Байес. Распределение Гаусса	66,180	1,057	32,763

Эффективность работы классификаторов

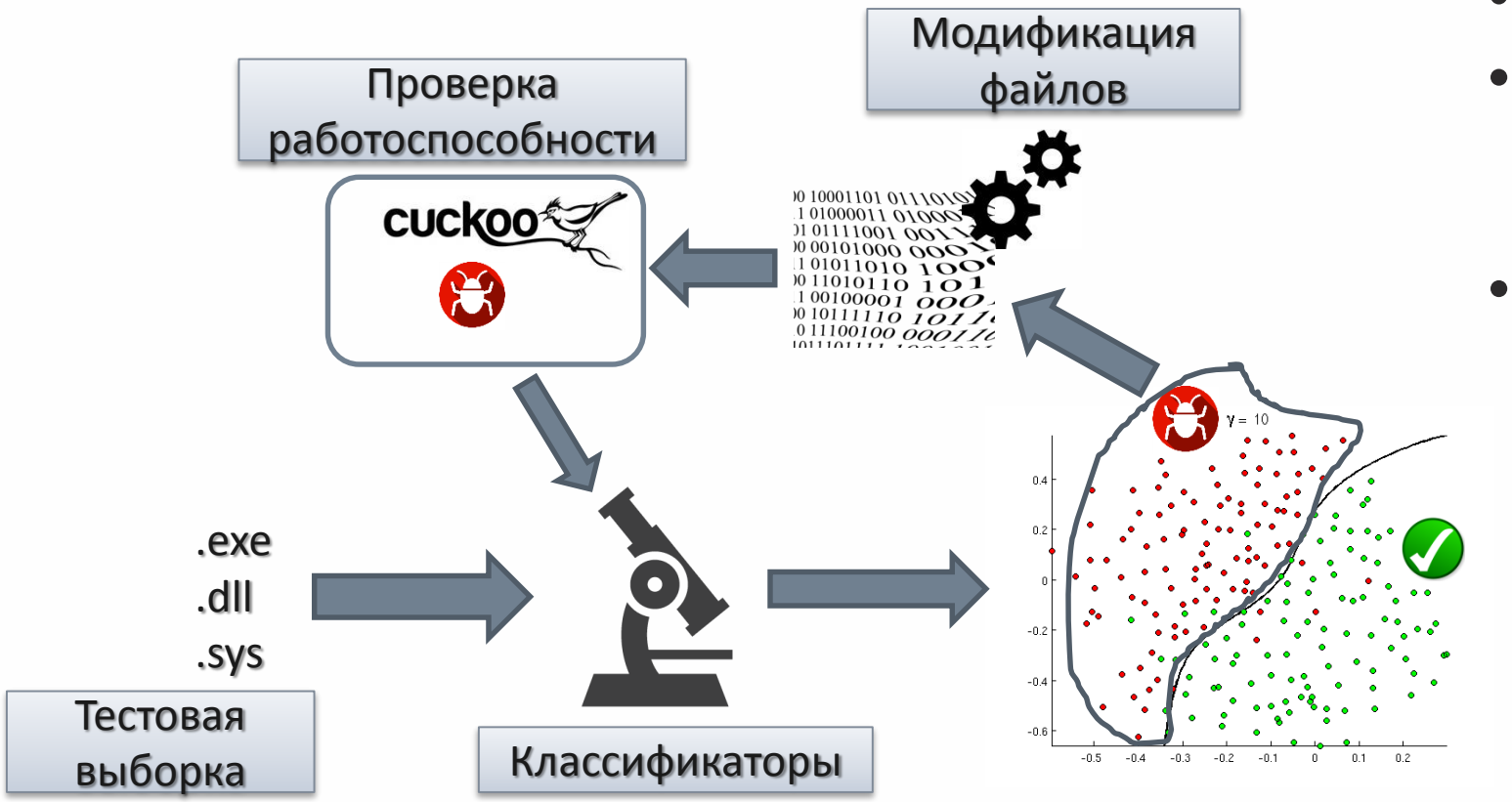
- Лучший результат:

Случайный лес. Мера энтропии 99,267% 0,339% 0,394%

(перекрестная проверка)

- Реализованные классификаторы будут доступны в Интернете для возможности проверки исполняемого файла

Противодействие машинному обучению



- **«Черный ящик»**
- Модифицировались 43 поля исполняемых файлов
- Не более 10 мутаций на один файл

Противодействие машинному обучению

Алгоритм	Точность обнаружения, %	Ошибки первого рода, %	Ошибки второго рода, %
Случайный лес. Мера энтропии	96,461 / 99,267	2,904 / 0,339	0,635 / 0,394
Случайный лес. Критерий Джини	93,587 / 99,257	5,293 / 0,356	1,119 / 0,387
Дерево решений. Мера энтропии	85,027 / 98,940	7,895 / 0,524	7,078 / 0,536
Метод k-ближайших соседей	77,616 / 97,827	19,359 / 1,157	3,025 / 1,016
Адаптивный бустинг	73,775 / 97,815	19,540 / 0,846	6,685 / 1,340
Дерево решений. Критерий Джини	67,695 / 98,829	16,969 / 0,572	15,336 / 0,599
Наивный Байес. Распределение Гаусса	55,596 / 66,180	44,343 / 1,057	0,061 / 32,763
Наивный Байес. Распределение мультиномиальное	47,066 / 61,681	0,635 / 2,218	52,299 / 36,101
Нейронные сети	44,344 / 82,942	0 / 6,664	55,656 / 10,394
Наивный Байес. Распределение Бернулли	7,804 / 89,895	36,540 / 6,363	55,656 /3,742

Повышение устойчивости к противодействию

- Настройка порогового значения
- Повторное обучение -> увеличение кол-ва ложноположительных срабатываний
- Повторное обучение с добавлением в обучающую выборку наименее схожих с уже присутствующими в ней вредоносными файлами
- Сокрытие алгоритма классификации

Результаты и выводы

- Машинное обучение крайне востребовано в средствах защиты цифрового производства
- Разработаны классификаторы, эффективно справляющиеся с задачей выявления ВПО
- Разработанные классификаторы будут доступны для свободного использования
- Самые простые способы противодействия могут значительно понизить эффективность применения большинства методов машинного обучения
- Требуется совершенствование методов для повышения их устойчивости к целевым атакам