Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.852

Вешторт
Алесь Витальевич

# Средства автоматического анализа вредоносного программного обеспечения

**АВТОРЕФЕРАТ**
на соискание степени магистра технических наук

по специальности 1-31 80 10 – «Теоретические основы информатики»

Научный руководитель
М.Д. Степанова
кандидат технических наук, доцент



Минск 2020

# ВВЕДЕНИЕ

В настоящее время проблема автоматизированного анализа вредоносного программного обеспечения стоит необычайно остро из-за огромных темпов создания новых вредоносных образцов. По данным Virustotal.com, только на этот ресурс еженедельно загружается до 400 тыс. образцов вредоносного программного обеспечения.

Очевидно, что ручной анализ такого количества образцов требует больших затрат человеческих и временных ресурсов.

Существующие программные решения лишь частично справляются с задачей анализа вредоносного программного обеспечения, оставляя значительный пласт работы аналитику вредоносных образцов.

**Целью** данной работы является исследование возможностей применения методов машинного обучения при обработке данных, полученных с помощью существующих программных решений в области анализа вредоносного программного обеспечения.

Акцент в работе делается на обнаружении нетипичных программ для уменьшения размера и сокрытия содержимого файла: пакеров, крипторов и протекторов. Нетипичные пакеры и протекторы с достаточно высокой вероятностью свидетельствуют о вредоносности файла, позволяя априори относить образцы программного обеспечения, обработанные такими средствами, к числу вредоносных.

Предлагаемый метод автоматизированного анализа объединяет в себе достоинства статического и динамического методов и значительно уменьшает долю ручного анализа в процессе исследования.

Такой подход позволит не только получить гораздо более подробную и релевантную информацию об образце по сравнению существующими решениями в области как статического, так и динамического анализа, но и позволит существенно снизить затраты времени и усилий аналитика по обработке этой информации.

Кроме того, в данной работе предлагается метод кластеризации вредоносных образцов с применением самоорганизующейся карты Кохонена. Полученные классы легко интерпретируются аналитиком и в дальнейшем могут применяться при решении задачи классификации вредоносного программного обеспечения.

Таким образом, основными **задачами** данной работы являются:

- разработка набора средств обнаружения нетипичных пакеров и протекторов программного обеспечения;
- разработка средства кластеризации вредоносных образцов.

**Объектом исследования** в данной работе являются образцы вредоносного программного обеспечения. **Предметом исследования** являются средства и методы автоматического анализа образов вредоносного обеспечения.

Работа разделена на 8 разделов.

1-й раздел содержит обзор литературы по теме магистерской диссертации, во 2-м разделе представлены общие теоретические сведения о характере анализируемых данных, разделы 3-5 содержат теоретическую информацию о предлагаемых методах анализа вредоносного программного обеспечения, в разделах 6-8 представлена практическая реализация этих методов в виде программных средств.

Общий объем работы составляет 77 страниц, в том числе 1 приложение. Количество использованных библиографических источников: 35.

Результаты исследований, приведенные в данной работе, были представлены на 54-й научной конференция аспирантов, магистрантов и студентов БГУИР и международной научной конференции «Информационные технологии и системы 2018 (ИТС 2018)».

Кроме того, в сборнике материалов международной научной конференции «Информационные технологии и системы 2018 (ИТС 2018)» по теме диссертации была опубликована научная статья «Автоматизированный анализ вредоносного программного обеспечения с применением рекуррентных нейронных сетей».

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе представлено описание трёх основных подходов к автоматизированному анализу вредоносного программного обеспечения с их преимуществами и недостатками, а именно: сопоставление с образцом, статический анализ и динамический анализ. Приведены примеры программных средств, реализующих эти подходы на практике, в частности описаны достоинства и недостатки следующих программных продуктов: средства сопоставления с образцом YARA, интерактивного мультипроцессного дизассемблера Hex-Rays IDA и системы автоматического анализа вредоносного программного обеспечения Cuckoo Sandbox.

Во второй главе представлена общая характеристика анализируемых данных. В этой главе производится анализ переносимого исполняемого файла с одной стороны, как бинарной последовательности, и с другой стороны – как последовательности вызовов функций интерфейса прикладного программирования операционной системы Windows, которая формируется в процессе исполнения файла операционной системой.

В третьей главе представлен набор бинарных статических признаков переносимого исполняемого файла, позволяющих определить, был ли данный экземпляр исполняемого файла обработан нетипичными пакерами программного обеспечения. Помимо этого, в главе представлена модель искусственной нейронной сети, позволяющая решить задачу классификации на основе выделенных признаков.

В четвертой главе представлен альтернативный метод определения факта обработки исполняемого файла нетипичными пакерами, основанный на способности рекуррентных нейронных сетей запоминать последовательности входных данных для автоматического выделения последовательностей вызовов функций интерфейса прикладного программирования операционной системы Windows. Кроме того, в этой главе представлен метод кодирования интерфейсов функций Windows API для представления их в виде входных данных рекуррентной нейронной сети.

В пятой главе представлен метод формирования кластеров вредоносного программного обеспечения на основе поведенческих признаков с применением самоорганизующейся карты Кохонена.

В шестой главе представлена реализация средства обнаружения нетипичных пакеров на основе методов, описанных в третьей главе. Средство представляет собой искусственную нейронную сеть с одним скрытым слоем и дополнительным слоем softmax. Сеть была реализована на языке Python с использованием модулей NumPy, TensorFlow и Keras, и обучена методом обратного распространения ошибки. Здесь же приведены результаты валидации разработанного средства на тестовой выборке.

В седьмой главе представлена реализация средства обнаружения нетипичных пакеров на основе методов, описанных в четвертой главе. Средство представляет собой искусственную нейронную сеть с двумя скрытыми слоями (LSTM, ReLU). Сеть была реализована на языке Python с использованием модулей NumPy, TensorFlow и Keras, и обучена методом обратного распространения ошибки. Здесь же приведены результаты валидации разработанного средства на тестовой выборке.

В восьмой главе описана реализация автоматизированного средства кластеризации вредоносного программного обеспечения на основе самоорганизующейся карты Кохонена и поведенческих признаков, полученных с помощью системы автоматического анализа вредоносного программного обеспечения Cuckoo Sandbox. Приложение было реализовано на языке Python с использованием модулей NumPy, и Kivy

## ЗАКЛЮЧЕНИЕ

В процессе выполнения данной работы был разработан прототип средства обнаружения нетипичных пакеров на основе статических признаков переносимых исполняемых файлов. Точность классификации на тестовой выборке для него составила: для чистых файлов – 98 %, для запакованных файлов – 84 %, для вредоносных файлов (обработанных нетипичными пакерами) – 86 %. Общая точность классификации составила 89,4 %, процент ложно отрицательных вредоносных экземпляров – 14 %.

Таким образом, была подтверждена гипотеза о возможности автоматического обнаружения нетипичных пакеров программного обеспечения путем обработки статических бинарных признаков алгоритмами машинного обучения. Такой подход позволит автоматически определять файлы, обработанные такими пакерами как вредоносные, таким образом снизив нагрузку на вирусного аналитика.

Кроме того, был разработан прототип средства обнаружения нетипичных пакеров на основе рекуррентных нейронных сетей и динамического анализа переносимых исполняемых файлов. Точность классификации на тестовой выборке составила: для чистых файлов – 74,1 %, для запакованных файлов – 73,3 %, для вредоносных файлов (обработанных нетипичными пакерами) – 70 %. Общая точность классификации составила 72,6 %, процент ложно отрицательных вредоносных экземпляров – 29,3 %.

Хотя применительно к обнаружению пакеров подход, основанный на динамическом анализе файлов, оказался менее эффективен, в своей сущности он является гораздо более универсальным и может быть использован для обнаружения других видов вредоносного поведения. Комплекс средств, основанных на данном подходе, может в перспективе заменить традиционные сигнатуры, реагирующие только лишь на явно заданное вредоносное поведение, и значительно увеличить эффективность динамических систем автоматического анализа вредоносного программного обеспечения.

В рамках данной работы также было разработано средство формирования классов вредоносного программного обеспечения на основе поведенческих признаков, соответствующих сигнатурам системы автоматического анализа Cuckoo Sandbox. Указанное средство представляет собой графическое приложение, комбинирующее ручную кластеризацию и возможности самоорганизующихся карт Кохонена.

В результате его применения на основе 600 образцов было выделено шесть классов вредоносного программного обеспечения: adware, filecoder, trojan-hosts, keylogger, password stealer, virus. Разбитая на указанные классы выборка может быть впоследствии использована для обучения модели классификатора.