МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра математического моделирования и анализа данных

Аннотация к дипломной работе

ОБНАРУЖЕНИЕ ВРЕДОНОСНЫХ ИСПОЛНЯЕМЫХ ФАЙЛОВ И PDF-ФАЙЛОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Калинчук Мария Валерьевна

Научный руководитель:
кандидат физ.-мат. наук,
заведующий НИЛ статистического анализа
и моделирования
Абрамович Михаил Семёнович

Минск, 2022

РЕФЕРАТ

Дипломная работа,61 страница, 9 рисунков, 16 источников, 5 приложений, 4 таблицы.

Ключевые слова: МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ, МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, АНТИВИРУСНЫЕ СИСТЕМЫ, ANDROID-РАЗРАБОТКА.

Объект исследования — вредоносные PDF и исполняемые файлы, используемые при распространении вредоносного программного обеспечения в пользовательские устройства.

Цель работы — выявление угроз при использовании потенциально заражённых PDF и исполняемых файлов, исследование поведения вредоносных файлов, выработка принципов для обеспечения безопасности пользовательских устройств при помощи методов машинного обучения.

Методы исследования — методы теории вероятности и математической статистики, эмпирические методы: эксперимент, моделирование, наблюдение, сравнение, измерение, классификация.

Результат — проанализированы особенности построения вредоносных файлов разных типов. Выявлены основные признаки для построения анализа методами машинного обучения. Предложены пути исследований при помощи методов машинного обучения. Реализовано Android-приложение для практической реализации результатов работы.

Область применения — статистический анализ вредоносных файлов, классификация, прогнозирование, компьютерная безопасность.

РЭФЕРАТ

Дыпломнаяпраца,61старонка, 9 малюнкаў, 16 крыніц, 5дадаткаў, 4 табліцы.

Ключавыясловы: МЕТАДЫ МАШЫННАГА НАВУЧАННЯ, МАТЭМАТЫЧНАЯ СТАТЫСТЫКА, АНТЫВІРУСНЫЯ СІСТЭМЫ, ANDROID-РАЗРАБОТКА.

Аб'ектдаследавання—шкоднасныяPDF і выкананыя файлы, выкарыстоўваныяпрыраспаўсюджваннішкоднаснагапраграмнагазабеспячэння ў прыстасаваныяпрылады.

Мэта працы—выяўленнепагрозпрывыкарыстанніпатэнцыйназаражаныхPDF і выкананыхфайлаў, даследаваннепаводзіншкоднасныхфайлаў, выпрацоўкапрынцыпаў для забеспячэннябяспекікарыстацкіхпрыладпрыдапамоземетадаўмашыннаганавучання.

Метадыдаследавання—метадытэорыіверагоднасці і матэматычнайстатыстыкі, эмпірычныяметады: эксперымент, мадэляванне, назіранне, параўнанне, вымярэнне, класіфікацыя.

Вынік—прааналізаваныасаблівасціпабудовышкоднасныхфайлаў розных тыпаў. Выяўленыасноўныяпрыкметы для пабудовыаналізуметадамімашыннаганавучання. Прапанаванышляхідаследаванняўпрыдапамоземетадаўмашыннаганавучання. РэалізаванаAndroid-дадатак для практычнайрэалізацыівынікаўпрацы.

Вобласцьужывання—статыстычныаналізшкоднасныхфайлаў, класіфікацыя, прагназаванне, кампутарнаябяспека.

ABSTRACT

Graduate work,61 pages, 9 figures, 16 sources, 5 appendices, 4 tables.

Keywords: MACHINE LEARNING METHODS, MATHEMATICAL STATISTICS, ANTI-VIRUS SYSTEMS, ANDROID DEVELOPMENT.

Object of research—malicious PDFs and executable files used in the distribution of malicious software to user devices.

Purpose of the work — to identify threats when using potentially infected PDF and executable files, study the behavior of malicious files, develop principles for ensuring the security of user devices using machine learning methods.

Research methods — methods of probability theory and mathematicalstatistics, empirical methods: experiment, modeling, observation, comparison, measurement, classification.

Result — the features of the construction of malicious files of different types are analyzed. The main features for constructing analysis using machine learning methods are identified. Ways of research using machine learning methods are proposed. An Android application has been implemented for the practical implementation of the results of the work.

Scope — statistical analysis of malicious files, classification, prediction, computer security.

ВВЕДЕНИЕ

В настоящее время во всём мире уделяется большое внимание развитию программ для обеспечения информационной безопасности. Популярные антивирусные программы (Kaspersky, Eset, Avast и т.д.) считаются одними из самых лучших, так как при появлении новых уязвимостей компании-разработчики достаточно оперативно справляются с их устранением.

Рассматривая тенденцию роста разработки и распространения вредоносного программного обеспечения, можно заметить, что за последний год, во-первых, подешевели хакерские программы (что говорит об увеличении их количества), во-вторых, 77% вредоносных программ было доставлено по электронной почте (это означает, что любой “образовательный” сервис может доставить пользователю якобы PDF-файл для обучения, при этом данный файл будет содержать в себе вирус, однако за счёт доверия пользователя будет скачан и открыт).

Увеличение риска распространения несёт в себе промежуточный, архитектурно независимый код — байт-код. Данный код злоумышленники обычно преобразуют при помощи программ-обфускаторов, которые превращают исходный байт-код в достаточно нечитаемый для человека, но тем не менее исполняемый устройством пользователя.

Для борьбы с обфусцированными файлами существует метод интеллектуального анализа — на наборе данных строятся признаки классификации. В последствии на основе этих признаков определённые алгоритмы машинного обучения могут спрогнозировать вредоносность нового, раннее неизвестного файла.

Важно учесть необходимость оптимизации скорости прогнозирования: при большом объёме данных анализатор может принять запоздалое решение о вредоносности файлов, что означает успех распространения нелегитимного программного обеспечения.

Для достижения цели исследования были поставлены следующие **задачи**:

- проанализировать состояние изученности вопроса и определить методику исследования;
- определить параметр, который позволит применять методы машинного обучения для классификации данных;
- определить выборку методов машинного обучения для получения наиболее точных показателей;
- написать на языке Python классификатор для определения типа фалов, протестировать файлы с различными типами заражения;
- провести сравнительный анализ результатов обнаружения вредоносных файлов, проводимых при помощи методов машинного обучения;
- разработать Android-приложения с использованием методов, показавших наилучшие результаты.

Актуальность данного исследования определяется тем, что при имеющемся достаточном разнообразии исследований, касающихся обеспечения информационной безопасности, в основном все антивирусные программы сравнивают исследуемые файлы с собственными базами сигнатур, которые могут не включать в себя совершенно новые типы вирусов. Очевидно, что существует необходимость в разработке программного обеспечения нового типа для определения вредоносных файлов вне зависимости от того есть ли данный тип сигнатур в базе.