# **МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное бюджетное образовательное учреждение высшего образования

«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА ПО ДИСЦИПЛИНЕ

«ПРОЕКТНАЯ ДЕЯТЕЛЬНОСТЬ»

на тему:

**«Анализ трафика устройств интернета вещей с помощью алгоритмов машинного обучения»**

Куратор проекта: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,\_\_\_\_\_\_\_\_\_\_\_\_ /

*подпись*  *ФИО, уч. звание и степень*

Студенты: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_ /

*подпись*  *ФИО, группа*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_ /

*подпись*  *ФИО, группа*

Москва, 2020

**Аннотация**

В настоящей статье приводится описание интернета вещей, обозначение его актуальности в данной области, выделение проблем, связанные с угрозами сетевых атак, способы решения данных проблем методами классификации, а также выявление лучшего из представленных классификаторов путем сравнения практических расчетов и полученных результатов.

**Партнёры**

Внешние партнёры отсутствуют в данном проекте.

**Общее задание**

Рассматривается классификация атак на устройства интернет вещей (IoT) с помощью классических алгоритмов машинного обучения.

* актуальность темы работы или исследования

На данный момент рассматриваемая область плохо изучена и в результате исследования не удалось найти исследования или научные разработки, которые затрагивали эффективные методы выявления аномалий в сетях интернета вещей. В связи с быстрым развитием интернета вещей и его внедрением во все большее количество сфер подобные исследования являются актуальными как на сегодняшний день, так и в будущем, т.к. они позволяют определить возможные проблемы, связанные как с развитием новой технологии, так и с повсеместным ее внедрением.

* цели и задачи

Оценка эффективности многоклассовой классификации классическими алгоритмами машинного обучения на данных современного интернет трафика устройств типа интернет вещей. Выделение главных признаков сетевого трафика, позволяющих определить тип трафика классическим алгоритмам машинного обучения.

* научная новизна

На данный момент опубликованы исследования и научные разработки, посвященные исследованию интернет трафика в компьютерных сетях. Используемые в них данные не являются современными и из-за быстрого развития технологий теряют свою актуальность.  
Наша работа отличается абсолютной новизной данных, используемых в анализе. Так же на данный момент не опубликовано работ, посвященных анализу трафика именно интернета вещей т.к. данная технологическая область является достаточно новой и неизученной.

* теоретическая и практическая значимость работы

Данная работа показывает специфику современного трафика интернета вещей, его основные свойства и опорные признаки. Результаты данной работы можно использовать в дальнейших исследованиях с более сложными алгоритмами машинного обучения.  
Обученные модели машинного обучения можно использовать для анализа любого трафика интернета вещей при условии выделения в нем признаков аналогичных приведенным в работе.  
Программный код из данной работы можно использовать для обучения моделей машинного обучения, направленных на анализ трафика интернета вещей даже при несовпадении признаков с приведенными в работе.

* методология и методы исследования

Для оценки качества результатов использовались основные метрики качества принятые в машинном обучении: точность (precision), полнота (recall), F-мера (F-score), ROC-кривые (Receiver Operating Characteristic curve - кривая ошибок), AUC-ROC и AUC-PR(Area Under Curve - площадь под кривой ошибок и площадь под кривой precision-recall). Исследования проводились основными инструментами принятыми в машинном обучении на языке программирования Python из бесплатного программного комплекса Anaconda.

* степень достоверности и апробация результатов:

Достоверность результатов работы обусловлена системной проработкой проблемы, достоверностью применяемых методов. Для экспериментальных работ показана воспроизводимость результатов исследования. Теория построена на известных, проверяемых данных, согласуется с опубликованными экспериментальными данными по смежным темам.

**План работы**

* Разработка механизма эффективной проверки алгоритма на всей совокупности данных
* Добавление математического аппарата
* Выбор или создание оптимальной метрики для оценивания работы алгоритма
* Создать обученные модели для классификации
* Оптимизируя модели, улучшить показатели классификаторов по выбранной метрике
* Провести сравнительный анализ между затронутыми классическими алгоритмами машинного обучения
* На их основе определиться с выбором наиболее подходящего алгоритма для классификации подобных данных.
* Оформить работу в научно-техническом стиле

**Команда проекта**

|  |  |  |  |
| --- | --- | --- | --- |
| Куратор | | | |
| ФИО (полностью) | | | e-mail |
| Харченко Елена Алексеевна | | | elenakhaa@yandex.ru |
| Студенты | | | |
| ФИО | № группы | Роль | | |
| Маковей Сергей Олегович | 171-341 | Участник занимается анализом полученных данных на основе рассмотренных классических алгоритмов машинного обучения(классификаторов) и документированием полученных результатов в научно-техническом стиле | | |
| Сигида Максим Павлович | 171-341 | Участник занимается построением классификаторов (наиболее распространенных классических алгоритмов машинного обучения) и исследованием эффективности данных алгоритмов. | | |

**Индивидуальные планы реализации проекта**

Маковей Сергей

|  |  |  |
| --- | --- | --- |
| Задачи | Время | Завершено |
| 1. Изучение предметной области | 5 часов | + |
| 2. Разработка требований к данным | 2 часа | + |
| 3. Поиск источников данных | 4 часа | + |
| 4. Первичный анализ данных из разных источников | 5 часов | + |
| 5. Выбор оптимального набора данных | 1 час | + |
| 6. Разработка отчета по проделанной работе | 3 часа | + |
| 7. Анализ классов в наборе данных | 5 часов | + |
| 8. Изучение математического аппарата | 5 часов | + |
| 9. Оформление введения в научно-техническом стиле | 2 часа | + |
| 10. Разработка второй и третьей главы | 6 часов | + |
| 11. Разработка сайта | 6 часов | + |
| 12. Доработка работоспособности сайта и дизайна | 5 часов | + |
| 13. Оформление второй главы в научно-техническом стиле | 3 часа | + |
| 14. Оформление третьей главы в научно-техническом стиле | 3 часа | + |
| 15. Исправление структуры работы | 1 час | + |
| 16. Проведение сравнительного анализа | 5 часов | + |
| 17. Поиск других исследований по данной теме | 4 часа | + |
| 18. Оформление раздела “Заключение” | 4 часа | + |
| 19. Общие исправления в работе | 5 часов | + |

Сигида Максим

|  |  |  |
| --- | --- | --- |
| Задачи | Время | Завершено |
| 1. Изучение предметной области | 2 часа | + |
| 2. Первичный анализ данных из разных источников | 6 часа | + |
| 3. Выбор оптимального набора данных | 1 час | + |
| 4. Выбор инструментов для работы с данными | 3 часа | + |
| 5. Разработка удобного механизма для данных | 5 часов | + |
| 6. Анализ распределений по захватам | 3 часа | + |
| 7. Перепроверка классов в данных | 2 часа | + |
| 8. Анализ классов в наборе данных | 3 часа | + |
| 9. Анализ временных рядов в данных | 3 часа | + |
| 10. Анализ признаков в наборе данных | 5 часов | + |
| 11. Предобработка признаков | 2 часа | + |
| 12. Преобразование признаков | 4 часа | + |
| 13. Разбор полезности признаков | 4 часа | + |
| 14. Работа над созданием обучающей выборки | 6 часов | + |
| 15. Обучение моделей классификаторов | 3 часа | + |
| 16. Тестовое использование моделей | 6 часов | + |
| 17. Разработка матрицы ошибок | 4 часа | + |
| 18. Сравнительный анализ полученных моделей | 4 часа | + |
| 19. Подбор гиперпараметров для улучшения моделей | 4 часа | + |
| 20. Оформление первой главы в научно-техническом стиле | 3 часа | + |

**Содержательные главы.**

Введение.

Глава 1. Обзор классификаторов для выявления сетевых аномалий.

Раздел 1. Обзор специализированного ПО для выявления сетевых атак.

Раздел 2. Математический аппарат для выявления сетевых атак.

Раздел 3. Цели и задачи работы.

Глава 2. Реализация классификации для выявления атак на устройства Интернет вещей.

Раздел 1. Описание данных

Раздел 2. Первичная обработка данных

Раздел 3. Методы

2.3.1 Метод KNN (k ближайших соседей).

2.3.2 Метод наивного Байеса.

2.3.3 Метод дерево решений.

Глава 3.Сравнительный анализ или испытание классификаторов.

Заключение.

**Итоговый результат**

В результате проведенного исследования эффективности методов классификации по выявлению сетевых атак и анализа представленных данных можно сделать вывод, что из-за несбалансированности обучающих данных алгоритмы машинного обучения показали себя достаточно плохо на исследуемом наборе. Для решения этой проблемы требуется сбалансировать классы, например, путем дублирования или перейти к бинарной классификации – «нормальный» vs «не нормальный».

Однако если рассмотреть определение наиболее сбалансированных классов вредоносного трафика (PartOfAHorizontalPortScan, Okiru и DDoS), то лучше всего себя показал алгоритм дерева решений. Алгоритм с методом ближайших соседей так же показал хороший результат, однако все же уступает алгоритму дерева решений. Алгоритм наивного Байеса показал себя хуже всего из-за того, что он слишком часто ошибается, ложно относя трафик к одному из двух классов (PartOfAHorizontalPortScan и C&C -HeartBeat).

Таким образом, для многоклаcсовой классификации вредоносного трафика из исследованных алгоритмов машинного обучения лучше всего подходит алгоритм дерева решений.

**Заключение**

Угрозы сетевых атак неизбежны. А с ростом развития Интернета вещей их эффективное выявление становится основной задачей для любой крупной сети. Рассмотрев методы классификации и их способности выявлять аномалии трафика, встает вопрос о возможности использования методов машинного обучения для решения проблем с угрозами сети. На данный момент недостатки исследованных методов не позволяют использовать их в качестве эффективного решения проблемы. Однако если выделить все найденные недостатки, подойти серьезно к данному вопросу, то с помощью использования сильных сторон данных методов в совокупности позволит создать оптимальное решение для эффективного выявления угроз сетевого трафика на устройства Интернета вещей (IoT).