Классификация DGA-семейств

Введение

Привет Habr, меня зовут Никита, я младший исследователь компании UDV. Мы занимаемся разработкой решений в сфере информационной безопасности с применением ML.Сегодня расскажу, как мы семейства DGA классифицировали.

С развитием цифровых технологий, которые все глубже проникают в различные сферы жизни, от личной переписки до корпоративных сетей, защита данных становится одной из ключевых задач, обеспечивающих безопасность и стабильность как отдельных пользователей, так и целых организаций.

Одной из серьёзных угроз для информационной безопасности являются алгоритмически сгенерированные домены (DGA), которые позволяют злоумышленникам создавать множество поддельных доменов для обхода защитных систем и затруднения обнаружения вредоносной активности. В данной статье мы рассмотрим, что представляют собой DGA, каким образом они создаются и как методы машинного обучения могут быть применены для их эффективной классификации, с особым акцентом на анализ данных и классификацию наиболее популярных DGA семейств.

DGA и их создание

Domain Generation Algorithm (DGA) — это алгоритмы, встречающиеся в различных семействах вредоносных программ, используемые для генерации большого количества доменных имен, которые помогают обходить традиционные методы защиты и затруднять обнаружение вредоносных коммуникаций. DGA позволяет вредоносным программам генерировать сотни, а иногда и тысячи доменов ежедневно, что значительно усложняет блокировку всех возможных каналов связи между зараженными устройствами и командно-управляющими серверами (C2-серверами) злоумышленников.

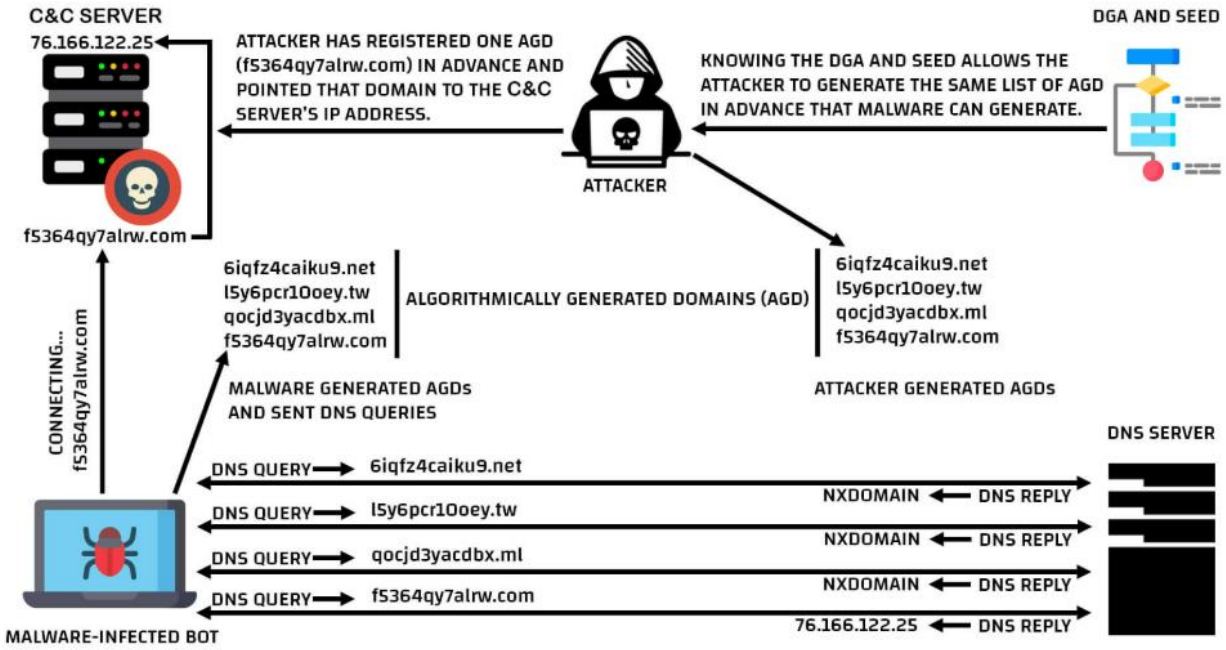

Рисунок 1 Схема работы DGA [1]

Применение DGA наблюдается в ряде известных атак и вредоносных программ, таких как ботнеты Conficker, Zeus, и Necurs. Например, Conficker генерировал до 50 000 доменных имен ежедневно, что сильно затрудняло для защитных систем блокировку всех потенциальных каналов связи. Вредоносное ПО Zeus использовало DGA для связи с C2-серверами, которые обеспечивали управление и контроль за зараженными системами, а также кражу финансовых данных. Necurs, известный как один из крупнейших ботнетов в мире, также использовал DGA для обхода блокировок и устойчивости к попыткам его ликвидации.

DGA	# Domains	Examples
conficker	13,500	akdtfkh.com.bs, flqxktto.co.za, hiroulyv.mu
corebot	13,500	a2ix16edy61f1xg.ddns.net, 1ruvwjy4u850385.ddns.net
cryptolocker	13,500	cryptolocker,ymirohscgjxm.ru, ugjywfellfju.co.uk
emotet	13,500	vtypqtkjnkyfycwj.eu, byoljbjregchfbtd.eu
gozi	13,500	animargumenta.com, surfacepapanobi.com
matsnu	13,500	branch-tower.com, film-water-image.com
murofet	13,500	nirauhusfeormfuwdovrrfinj.biz, jxqtupnmsmknn.info
necurs	13,500	wjilrilcim.ru, gfqcpaifkxd.bit
nymaim	13,500	smilefavorite.uz, soft-professor.com
padcrypt	13,500	adcaeccnacfnccma.com, bnmfladdaccalkff.com
qadars	13,500	8igi8qwu468u.com, e7wx2jkl2zw5.net
suppobox	13,500	groupfirst.net, brooklynnwashington.net
symmi	13,500	qaliesvomo.ddns.net, meakugulu.ddns.net

Рисунок 2 Список семейств DGA

DGA основан на определенном алгоритме, который может использовать различные параметры, такие как текущая дата, время, или случайное значение, для создания

уникальных доменных имен. Эти домены регистрируются злоумышленниками, и когда зараженное устройство пытается связаться с одним из них, оно подключается к действующему командному серверу. Из-за постоянного изменения доменов стандартные методы фильтрации и блокировки становятся неэффективными, что делает DGA мощным инструментом в арсенале киберпреступников

Задача детектирования DGA уже давно успешно решена благодаря развитию методов машинного обучения и различных сигнатурных подходов. Однако, классификация DGA-семейств становится всё более актуальной. Определение конкретного семейства, к которому принадлежит сгенерированный домен, может существенно помочь в прогнозировании дальнейших шагов злоумышленников. Это особенно важно в контексте применения фреймворка MITRE ATT&CK, который предоставляет структурированное описание различных тактик и техник кибератак. Классификация позволяет не только блокировать вредоносные домены, но и строить проактивную защиту, предсказывая возможные направления атаки и подготавливая соответствующие контрмеры.

Обзор существующих решений

Классификация доменов, генерируемых алгоритмами (DGA), является актуальной задачей для исследователей и специалистов в области кибербезопасности. В этой области проводятся исследования и публикуется множество статей, направленные на решение этой задачи. Методы обнаружения варьируются от анализа статистических особенностей сетевого трафика до обучения нейронных сетей, которые позволяют детектировать DGA-активности, находя сложные зависимости в данных. Например, в статье [2] авторы провели обширное измерение и анализ 43 семейств и вариантов вредоносных программ, разработав таксономию для характеристики основных аспектов алгоритмов генерации доменов. В другом исследовании [3] предлагается новый классификатор на основе остаточных нейронных сетей (ResNet), который демонстрирует высокую производительность в многоклассовой классификации DGA, используя реальные данные NX-трафика из DGArchive. В статье [4] авторы применили LightGBM для классификации и метод TextCNN, использующий сверточные нейронные сети для обработки текстовых данных и извлечения признаков для классификации. Датасет исследования состоял из 4 миллионов доменных имен, включая 2.67 миллиона вредоносных и 930,000 нормальных доменов в обучающих данных, а тестовые данные включали 370,000 вредоносных и 30,000 нормальных доменов, тестовые данные состояли из 370,000 вредоносных и 30,000 нормальных доменов.

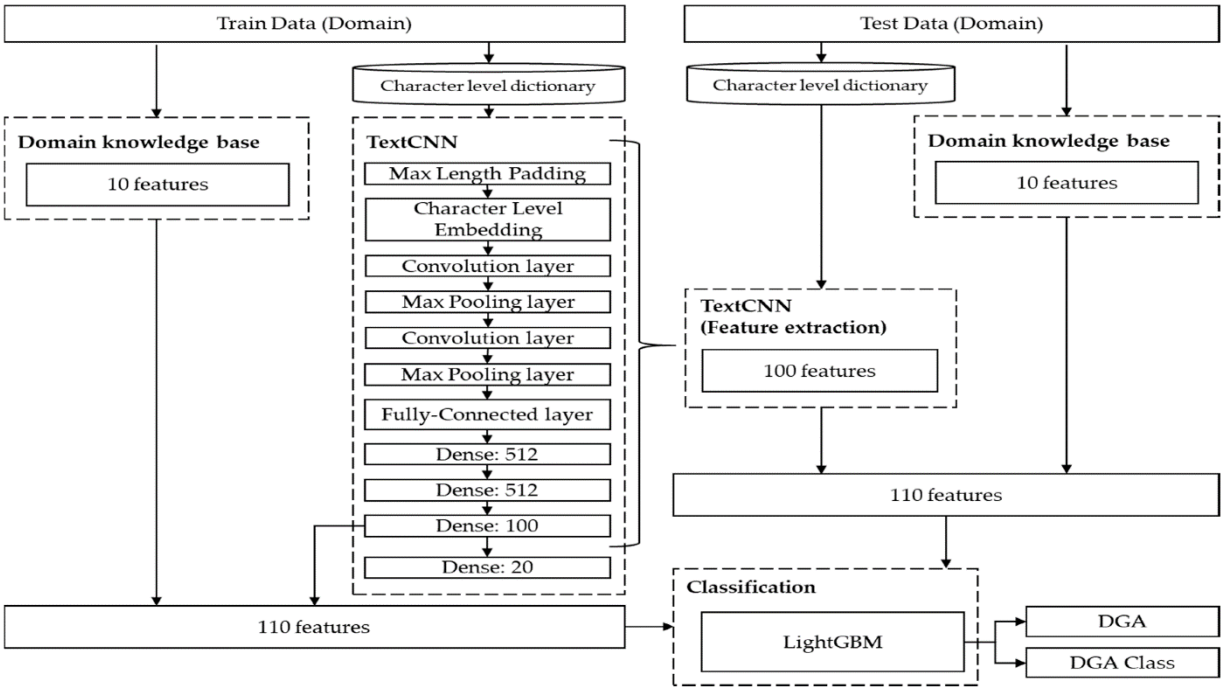

Рисунок 3 Схема работы алгоритма с использованием LightGBM [4]

Обзор датасета

В данном исследовании мы опирались на данные из открытого датасета, доступного по ссылке Zenodo. Этот датасет содержит большое количество доменных имен, сгенерированных различными DGA, а также легитимные доменные имена.

Каждая запись в датасете включает следующие ключевые элементы:

- **Доменные имена**: текстовые строки, представляющие как легитимные, так и сгенерированные DGA домены.
- **Метки**: каждый домен помечен либо как легитимный, либо как принадлежащий к определенному DGA семейству.
- **Признаки**: различные характеристики доменных имен, такие как длина, частота встречаемости символов и т.д.

Из всего множества признаков мы выбрали наиболее важные, используя SHAP value [5] каждого признака.

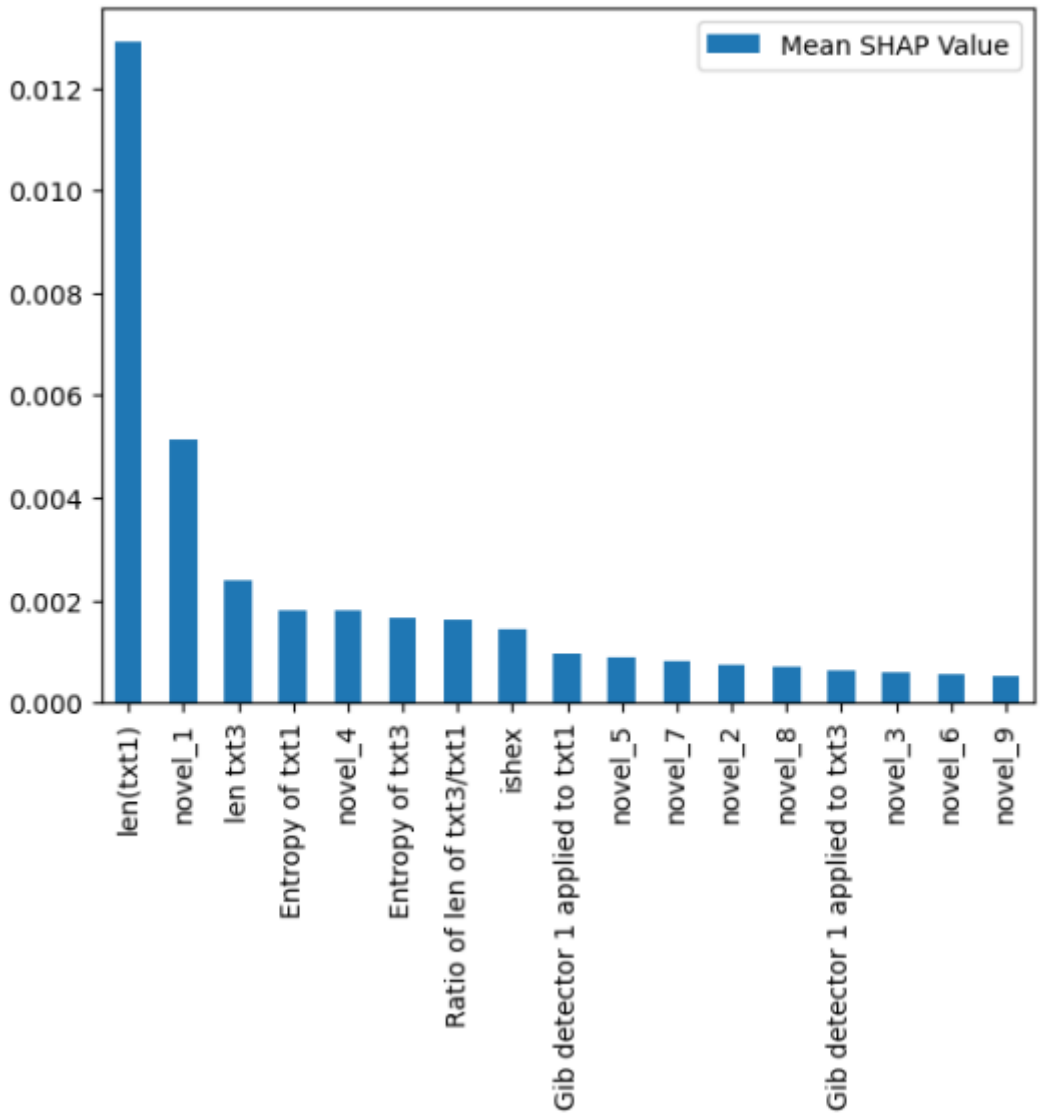

Рисунок 4 График средних значений SHAP value

По итогу мы получили вот такой список признаков из датасета:

'ishex' - Доменное имя представлено/не представлено шестнадцатеричными символами,

'len(txt1)' - Длина доменного имени,

'len txt3' - Количество цифр в доменном имени,

'novel_1' - Максимальное количество последовательных согласных,

'novel_2' - Отношение количества согласных к длине доменного имени,

'novel_3' - Максимальное количество последовательных гласных,

'novel_4' - Отношение доброкачественных 3-граммов (набор 3-грамм, сформированный из легитимных доменов) в наборе 3-грамм доменного имени ко всему набору 3-грамм,

'novel_5' - Отношение доброкачественных 4-граммов (набор 4-грамм, сформированный из легитимных доменов) в наборе 4-грамм доменного имени ко всему набору 4-грамм,

'novel_6' - Отношение доброкачественных 5-граммов (набор 5-грамм, сформированный из легитимных доменов) в наборе 5-грамм доменного имени ко всему набору 5-грамм,

'novel_7' - Отношение 3-граммов, содержащих гласные, ко всему набору 3-грамм доменного имени,

'novel_8' - Отношение 4-граммов, содержащих гласные, ко всему набору 4-грамм доменного имени

'novel_9' - Отношение 5-граммов, содержащих гласные, ко всему набору 5-грамм доменного имени

'Ratio of len of txt3/txt1' - Отношение количества символов доменного имени, не являющихся цифрами, к общему количеству символов в нём,

'Entropy of txt1' - Энтропия доменного имени,

'Entropy of txt3' - Энтропия доменного имени без цифр.

Также мы добавили ещё 2 признака, используя детектор случайного набора символов [6]:

'Gib detector 1 applied to txt1' - Вероятность не читаемости доменного имени, полученная от детектор случайного набора символов,

'Gib detector 1 applied to txt3' - Вероятность не читаемости доменного имени без цифр, полученная от детектор случайного набора символов.

В данной работе мы сфокусировались на классификации топ-10 DGA семейств по присутствию в датасете. Эти семейства являются наиболее часто встречающимися и представляют наибольший интерес для нас. Вот список этих семейств: chinad, murofetweekly, rovnix, pitou, corebot, sisron, tofsee, ebury, xxhex, qadars.

Использование LightGBM

Для обнаружения доменов, сгенерированных алгоритмами (DGA), в DNS-трафике был применен LightGBM (Light Gradient Boosting Machine) [7]. Это алгоритм машинного обучения, который основан на методе градиентного бустинга и работает по следующему принципу: он строит ансамбль слабых моделей, обычно деревьев решений, которые последовательно улучшают качество предсказаний. В процессе обучения LightGBM создает новые деревья, каждое из которых исправляет ошибки предыдущих деревьев, путем минимизации функции потерь.

Основным преимуществом LightGBM является его способность обрабатывать данные с высокой энтропией и различными статистическими характеристиками, такими как длина

доменного имени, количество цифр и распределение символов. Эти признаки, будучи важными для классификации DGA-доменов, позволяют модели строить точные предсказания относительно принадлежности домена к тому или иному семейству.

Для подготовки данных использовались стандартные процедуры обработки, включая удаление пропусков и дублирующихся записей. После этого были сгенерированы признаки, которые включали такие параметры, как длина доменных имен, количество цифр, частота встречаемости символов и другие важные характеристики. Поскольку данные по DGA-семействам могут быть неравномерно распределены, для балансировки классов применялся метод undersampling, что позволило избежать смещения модели в сторону более представленных классов.

Обучение модели LightGBM проводилось с использованием метода подбора гиперпараметров через grid search. Это позволило оптимизировать такие параметры, как количество деревьев, глубина деревьев и скорость обучения, что значительно улучшило производительность модели. Важно отметить, что модель обучалась на 11 классах: 10 целевых семейств и один дополнительный класс, предназначенный для обработки доменов, которые могут принадлежать неизвестным DGA-семействам. Этот 11-й класс играет важную роль, так как позволяет модели эффективно справляться с новыми, ранее не встречавшимися угрозами.

Экспериментальные результаты

	precision	**recall**	**F1-score**	**support**
0	**0.98**	**0.94**	**0.96**	**176**
1	**0.88**	**0.89**	**0.89**	**163**
2	**0.89**	**0.95**	**0.92**	**190**
3	**1.00**	**1.00**	**1.00**	**202**
4	**0.92**	**0.93**	**0.92**	**174**
5	**0.94**	**0.96**	**0.95**	**187**
6	**0.98**	**0.96**	**0.97**	**198**
7	**0.98**	**1.00**	**0.99**	**198**
8	**0.96**	**0.98**	**0.97**	**210**
9	**0.82**	**0.99**	**0.90**	**197**
10	**0.99**	**0.99**	**0.99**	**13105**

Из таблицы видно, что все 10 семейств (0-9) детектируются с точностью> = 89%, а также модель помечает 10 классом все доменные имена, принадлежащие другим семействам.

Высокие значения точности (precision) и полноты (recall) указывают на то, что модель хорошо различает доменные имена, принадлежащие к различным семействам DGA. Это критически важно для специалиста по безопасности, так как позволяет более уверенно идентифицировать, какое именно семейство DGA может быть задействовано в атаках. Это, в свою очередь, позволяет выстраивать более целенаправленную защиту и планировать контрмеры, такие как блокировка или мониторинг определённых доменов, прогнозирование новых доменных имён, которые могут появиться в будущем, и разработка стратегий для предотвращения последующих этапов атак.

Автор статьи: Никита Быков, младший исследователь исследовательского центра UDV Group

Полезные ссылки:

1. https://hackersterminal.com/domain-generation-algorithm-dga-in-malware/
2. https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_plohmann.pdf
3. https://arxiv.org/abs/2006.11103
4. https://www.researchgate.net/publication/342581282_Effective_DGA-Domain_Detection_and_Classification_with_TextCNN_and_Additional_Features
5. https://shap.readthedocs.io/en/latest/
6. https://github.com/domanchi/gibberish-detector
7. https://lightgbm.readthedocs.io/en/stable/

Теги: DGA, dns-трафик, lightgbm, датасет

Хабы: DNS, GitHub