для исходного кадра оказалась минимальной. Причём третий метод даёт лучшие результаты на первой части набора данных, собранный в различных источниках, а первый метод даёт лучшие результаты на результатах, полученных камерой НИЛ-55. Это может быть обусловлено как самой камерой, так и характером съёмки, так как все тестовые видео снимались в помещении, в отличии от найденных наборов данных, в которых преобладают уличные камеры.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ (проекты 19-29-09045, 20-37-70053).

Литература

1. Wu, J. & Qi, F. & Shi, G.. (2011). Unified spatial masking for just-noticeable difference estimation. APSIPA ASC 2011 - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2011. 447-450.
2. J. Wu, W. Lin and G. Shi, "Structural uncertainty based just noticeable difference estimation," 2014 19th International Conference on Digital Signal Processing, Hong Kong, 2014, pp. 768-771.
3. Jinjian, Wu & Qi, Fei & Shi, Guangming. (2012). Self-similarity based structural regularity for just noticeable difference estimation. Journal of Visual Communication and Image Representation. 23. 845-852.

Н.М. Кусакина

ПРИМЕНЕНИЕ АНАЛИЗА БОЛЬШИХ ДАННЫХ В ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

(Самарский государственный технический университет)

Управление сетевой политикой и ресурсами в компьютерной сети организации с каждым днём становится всё более сложной задачей. Этому способствует ежедневный рост сетей, подключение новых устройств, обновление приложений, а также значительное увеличение объемов передаваемого трафика.

В тоже время реагирование на события систем мониторинга находится в зависимости не только от критичности оборудования и размещенных систем, но и от информации о типе и свойствах сетевого трафика. По этой причине анализ сетевого трафика, не только входящего из вне, но и циркулирующего внутри периметра, может представлять проблему для систем управления сетью по причине своей разнородности. Чтобы помочь справиться с данной ситуацией вперед выступает аналитика больших данных. Данные, которые обрабатываются со стороны кибербезопасности, весьма разнообразны и слабоструктурированы. Они содержат в себе и контент социальных сетей, и

журналы истории браузеров, лог файлы серверов и данные потока кликов в Интернете, электронные письма клиентов, спам, трафик взаимодействия серверов внутри АС и сведения о сработке датчиков систем мониторинга.

Целью анализа данных в таком случае становится обнаружение соответствующей структурированной информации: тенденции рынка, сезонность сработок мониторинга, скрытые модели и ранее неизвестные корреляции. Любой бизнес может использовать аналитику больших данных для устранения угроз кибербезопасности.

Анализ больших данных в сочетании с сетевыми потоками и системными событиями мониторинга поможет выявить нарушения и подозрительные действия. Так, используя аналитику больших данных, можно разработать базовые показатели для систем мониторинга, основанные на статистической информации штатного функционирования системы. Пример: мониторинг событий в режиме реального времени. (on-line режим).

Но использование в аналитике больших данных не только статистики, но и связки машинного обучения с аппаратом нейронных сетей позволяет бизнесу провести более тщательный анализ собранной информации. Например, интеллектуальные системы, построенные на основе использования аппарата искусственных нейронных сетей, находят широкое применение в области проектирования новых средств защиты.

Одним из их главных преимуществ искусственных нейронных сетей является возможность анализа неполных входных данных или сигналов с какими-либо помехами, а также проведение нелинейного анализа произошедших событий (в случае распределённого внешнего воздействия на сеть). В этом случае каждое событие в сети будет иметь собственный вес, что важно, так как в реальном сетевом трафике пакет может искажаться как умышленно, так и в результате непреднамеренного сбоя работы системы.

Возможности применения нейронных сетей в Кибербезопасности широко обсуждаются в профессиональном сообществе уже несколько лет. Проведен анализ используемых методов машинного обучения для выявления аномалий сетевого трафика для определения актуального направления развития [3-6].

В результате анализа исследований по применению нейронных сетей для выявления сетевых аномалий сделан вывод, что гибридные нейронные сети весьма эффективны для поиска оптимального решения при ограниченном объеме данных. На основе проведенного анализа используемых архитектур искуственных нейронных сетей принято решение об использовании гибридной нейронной сети на CNN и RNN. Созданная модель изображена на рисунке 1, она реализована нами на языке Python с использованием Keras и бекендом Tensorflow.

К сбору датасета и предпроцессингу данных решено перейти, исходя из следующей идеи: сетевые взаимодействия различных служб и приложений можно представить в виде набора потоков транспортного уровня, каждый со своими особенными характеристиками. Аналогично, каждый поток относится к какому-либо сервису/приложению, таблица соотношения, которых применяется

для обучения модели нейронной сети. С математической точки зрения, определение типа потока будет решением задачи мульти-классовой классификации.

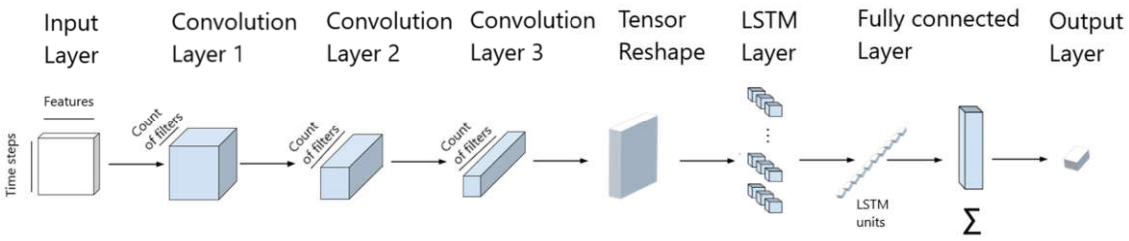

Рис. 1. Модель гибридной сети CNN+ RNN

Сбора трафика для составления дата сета производился в два этапа:

в течение суток с пограничного МСЭ собирался трафик, соответствующий, режиму штатного функционирования сети;

в течение выделенного периода времени проведены атаки на ресурсы сети, и собран соответствующий трафик.

На основе этих двух дампов собран входной датасет, разделенный в последующем на тренировочный (обучающий) набор – 70% данных, и тестовый набор – 30%. Такое разделение является стандартной практикой для оценки работы модели.

Датасет состоит из, примерно, 1 000 000 сетевых потоков, каждый из которых состоит из двадцати пакетов, совместно использующих уникальную двунаправленную комбинацию IP-адресов источника и назначения, их номеров портов, а также транспортного протокола. Любой последующий пакет текущего потока отбрасывается. Данное ограничение принято нами, поскольку этого количества пакетов в потоке достаточно для идентификации отдельного соединения, а большее число пакетов снизит производительность разрабатываемого классификатора.

Аналогично, каждый поток относится к какому-либо сервису/приложению, таблица соотношения которых необходима для обучения модели нейронной сети. То есть под определением типа потока мы понимаем определение приложения, которое использует данный поток, протокол прикладного уровня.

Составление надежного и полного набора признаков является важной частью процесса создания модели классификатора на основе нейронных искусственных сетей, так как результате мы должны получить в простой компактной форме описание всех необходимых признаков для обучающей выборки. Датасет для обучения нейронной сети оказался несбалансированным, так как распределение частот появления потоков той или иной службы получилось неравномерным. Данная закономерность характерна для разнородной компьютерной сети, содержащей множество различных информационных систем и устройств.

Разнородность входного датасета была принята во внимание, поскольку его несбалансированность влияет на итоговую точность работы алгоритма. Так как объекты самого многочисленного класса могут вносить больший вклад в метрику точности. В дальнейшем, при расчете итоговых показателей метрик алгоритма было использовано макро-усреднение, чтобы выровнять вносимое классами влияние.

Основными метриками при оценке алгоритмов machine learning являются: точность, полнота и F-мера [1-2]. В нашем случае мульти-классовой классификации точность (Precision) и полнота (Recall) рассчитываются с использованием матрицы неточностей (confusion matrix), размерность которой N на N, где N = 9 — количеству выходных классов.

Точность каждому классу равняется отношению соответствующего диагонального элемента матрицы и суммы всей строки класса c:

$$\mathrm{Precisionc} = \frac{\mathrm{Ac,c}}{\sum_{\mathrm{i=1}}^{9} \mathrm{Ac,i}}$$

Полнота класса c вычисляется как отношение диагонального элемента матрицы и суммы всех элементов столбца класса c:

$$\mathrm{Recallc} = \frac{\mathrm{Ac,c}}{\sum_{\mathrm{i=1}}^{9} \mathrm{Ai,c}}$$

Дальнейшее вычисление данных метрик для алгоритмов мульти-классовой классификации, как правило, идет по одному из вариантов: микро- или макро-усреднение.

В первом случае для каждой подзадачи бинарной классификации из рассматриваемого множества задач вычисляются показатели точности, полноты. Далее их величина усредняется по всем задачам, и по усредненным показателям вычисляется итоговая метрика. В случае макро-усреднения сначала для каждого класса вычисляется итоговая метрика (точность/полнота), а затем результаты усредняются по каждому из классов. Результирующая точность, как и полнота, является средним арифметическим точности (полноты) по всем классам.

Как выше описано, решено использовать именно макро-усреднение, так как в этом случае каждый класс будет вносить равный вклад, независимо от своего размера. Далее произведен расчет F-меры (Fβ). Которая представляет собой гармоническое среднее между точностью и полнотой: она стремится к нулю, если точность или полнота стремятся к нулю; достигает максимума в единицы при стремлении точности и полноты к единице. В формуле ниже β принимает значения в диапазоне $0<\beta<1$ в случае, если точность имеет больший приоритет.

При итоговой точности создаваемой модели более 95% возможен переход от «пороговых» правил в сторону поведенческого анализа потоков в промышленном использовании предлагаемого классификатора. Дальнейшее применение классификатора рассматривается в связке с IDS/IPS системами.

Литература

1. Колесниченко Д., “Машинное обучение на практике” [Электронный ресурс], 2018 - URL Режим доступа: https://xakep.ru/2018/08/01/rts-tender/ (дата обращения 15.03.2019).

2. Полякова Е.В., “ Исследование методов машинного обучения для анализа и принятия решений на основе данных Интернета вещей” [Электронный ресурс], 2018 - URL Режим доступа: https://publications.hse.ru/chapters/204754963 (дата обращения 15.03.2019).

3. C.-Y. Lee, P. W. Gallagher, and Z. Tu., «Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree», [Электронный ресурс], 2015 - URL Режим доступа: https://arxiv.org/abs/1509.08985 (дата обращения 15.03.2019).

4. Z. C. Lipton, J. Berkowitz, and C. Elkan, «A critical review of recurrent neural networks for sequence learning», Электронный ресурс], 2015 - URL Режим доступа: https://arxiv.org/abs/1506.00019 (дата обращения 15.03.2019).

5. Middlemiss M., Dick G., «Feature Selection of Intrusion detection data using a hybrid genetic of hybrid Intelligent systems», IOS Press Amsterdam, 2018.

6. F. Pierazzi, G.Apruzzese, M. Golajanni. A. Guido, «Scalable architecture for onlineprioritization of cyber threats», International Conference on Cyber Conflict, 2017.

К.Ф. Родичев, Ф.А. Дмитриев

РАЗРАБОТКА АЛГОРИТМА ШИФРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ГРАФОВ

(Самарский университет)

Цель: создание алгоритма шифрования и исследование свойств шифра.

Задачи:

- Ввести основные понятия для описания алгоритма.
- Описать идею представления числа в виде графа.
- Разработать алгоритм шифрования на основе изложенной идеи.
- Написать программную реализацию алгоритма.
- Обосновать соответствие алгоритма изложенным требованиям.
- Проанализировать криптостойкость алгоритма.