УДК 004.67

ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ЗАШИФРОВАННОГО ТРАФИКА

Вихров М.С.

магистр

МГТУ им. Н.Э. Баумана,

Россия, г. Москва

Аннотация

Классификация трафика применяется для широкого спектра от QoS подготовки и выставления счетов в интернет-провайдерах до связанных с безопасностью приложений в брандмауэрах и системах обнаружения вторжений. Целью статьи является исследование методов классификации зашифрованного трафика. Порт-основанная пакетная инспекция, и классические методы машинного обучения широко используются в современных системах, но их точность была снижена в связи с резкими изменениями в интернет-трафике, в частности с увеличением зашифрованного трафика. С распространением методов глубокого обучения, эти методы для задач классификации трафика показали высокую точность. Применения методов глубокого обучения в анализе и классификации трафика является перспективным направлением многих современных исследований в области распознавания интернет трафика. После проведенных исследований, в этой статье обозначаются общие рамки для классификации трафика на основе глубокого обучения. Представляются используемые методы глубокого обучения и их применение в задачах классификации.

Ключевые слова: классификация интернет трафика, машинное обучение, глубокое обучение, пакетная инспекция, временные ряды.

RESEARCH OF METHODS OF CLASSIFICATION OF ENCRYPTED TRAFFIC

Vikhrov M.S.

Master's student

Bauman Moscow State Technical University,

Russia, Moscow

Annotation

Traffic classification is used for a wide range of applications, from QoS preparation and billing in Internet service providers to security-related applications in firewalls and intrusion detection systems. The purpose of this article is to study methods for classifying encrypted traffic. Port-based batch inspection and classic machine learning methods are widely used in modern systems, but their accuracy has been reduced due to dramatic changes in Internet traffic, in particular with the increase in encrypted traffic. With the spread of deep learning methods, these methods for traffic classification problems have shown high accuracy. The use of deep learning methods in traffic analysis and classification is a promising area of many modern research in the field of Internet traffic recognition. After research, this article outlines a General framework for classifying traffic based on deep learning. The methods of deep learning used and their application in classification problems are presented.

Key words: classification of Internet traffic, machine learning, deep learning, batch inspection, time series.

Классификация трафика в соответствующие классы весьма актуальная проблема, так как важна для многих приложений, таких как контроль качества обслуживания (QoS), ценообразование трафика, планирование использования ресурсов, обнаружение вредоносных программ и вторжений. Из-за своей важности, много различных подходов были разработаны для решения этой проблемы в течение последних лет, чтобы приспособить разнообразные изменения потребностей различных сценариев применения. В частности, новые достижения в области связи, включая шифрование и обфускацию портов, поднимают дополнительные проблемы для классификации трафика. Целью исследования является изучение применяемых методов и технологий для классификации зашифрованного трафика. В начале исследования были поставлены следующие задачи: провести обзор проблем методов классификации трафика, исследовать алгоритмы сбора и подготовки данных для применения различных методов машинного обучения, а также провести анализ современных методов предварительной обработки набора данных. При проведении исследования использовались следующие научные методы: изучение теоретических основ по проектированию методов классификации, изучение и анализ документации различных систем классификации трафика, рассмотрение результатов проведенных исследований в разрезе точности классификации трафика, сбор и анализ информации о реализованных методах идентификации трафика, проведение сравнительного анализа методов машинного и глубокого обучения.

Методы классификации трафика претерпели значительные изменения со временем. Первый и самый простой подход к классификации-это использование номеров портов. Однако его точность снижается, потому что более новые приложения либо используют хорошо известные номера портов, чтобы маскировать свой трафик или не используют стандартный

зарегистрированный номер порта. Несмотря на свою неточность, номер порта все еще широко распространен и используется либо отдельно, либо в тандеме с другими функциями. Следующее поколение классификаторов трафика, опирающихся на полезную нагрузку или проверку пакетов данных (DPI), фокусируется на поиске паттернов или ключевых слов в пакетах данных. Эти методы применимы только к незашифрованному трафику и имеют высокие вычислительные издержки. Как результат, появилось новое поколение методов, основанных на потоковой статистике трафика. Эти методы основаны на статистических или функциях временных рядов, которые позволяют им обрабатывать как зашифрованные, так и незашифрованный трафик. Это алгоритмы машинного обучения, такие как случайный лес и к-ближайших соседей. Тем не менее, их производительность сильно зависит от человеко-инженерных особенностей, которые ограничивают их результативность работы. Глубокое обучение устраняет необходимость подбора объектов по предметной области эксперта, потому что он автоматически выбирает объекты через обучение. Эта характеристика делает глубокое обучение весьма желательным подходом к классификации трафика, особенно когда постоянно возникают новые классы, и модели старых классов развиваются. Еще одной важной характеристикой глубокого обучения является то, что оно имеет значительно более высокую способность к обучению по сравнению с традиционными методами машинного обучения, и таким образом может вычислить сильно осложненные угрозы. Глубокое обучение способно к обучению нелинейным связям между необработанными входными данными и соответствующими выходными данными, это убирает необходимость разбивать проблему на мелкие подзадачи отбора и классификации признаков. Недавние работы продемонстрировали эффективность глубокого обучения в методах классификации трафика, в частности, в зашифрованном виде движение. Для

достижения этой цели глубокое обучение требует достаточного количества маркированных данных и достаточной вычислительной мощности. В этой статье проводится обзор общих рамок для (зашифрованного) трафика в задачах классификации. Предоставим общие рекомендации по классификационной задачи, в том числе сбору и очистке данных, функциях выбора и выбору модели. Более того, обсудим глубокие методы обучения и то, как они были применены для задач классификации трафика.

Источник [1] иллюстрирует общую структуру классификации трафика, состоящую из семи этапов. Большинство существующих работ полностью или частично используют эту структуру. Мы обсудим первые четыре шага в этом разделе, а последние три – в следующем, уделяя особое внимание подходам, основанным на глубоком обучении. Первый шаг к построению классификатора сетевого трафика состоит в четком определении цели классификации. Типичные цели включают обеспечение качества обслуживания, планирование использования ресурсов, настройку биллинговой системы, обнаружение вторжений и вредоносных программ. Для достижения этой цели можно классифицировать классы трафика на основе 1) протоколов (например, UDP, TCP, FTP или HTTP), 2) приложений (например, Skype, WeChat или Torrent), 3) типов трафика (например, просмотр, загрузка или видеочат), 4) веб-сайтов, 5) действий пользователя (например, размещение комментария или отправка голосового сообщения), 6) операционных систем, 7) браузеров и так далее. Следовательно, цель состоит в том, чтобы обозначить каждый поток с соответствующими классами трафика. Поток обычно определяется 5-кортежем: исходный IP, конечный IP, исходный порт, порт назначения и протокол. Кроме того, классификация трафика также может быть разделена на два подкласса: онлайн и оффлайн. Онлайновая классификация обычно относится к случаям, когда потоки должны быть классифицированы как можно быстрее, обычно в

пределах первых нескольких десятков пакетов. Например, для обеспечения QoS и маршрутизации классификация должна быть в оперативном режиме, поскольку выходные данные классификации непосредственно используются для принятия решений по текущему потоку. Для других приложений, таких как биллинговые системы, классификация может быть отключена. Хотя классификация трафика применяется к совершенно различным сценариям, большинство исследований объединяет два общих аспекта: (а) исходными данными для классификации являются необработанные пакетные данные, их часть или информация, непосредственно полученная из них, (б) используются аналогичные алгоритмы ML. Основное внимание в статье уделяется классификации типов зашифрованных приложений и трафика. Однако та же методология может быть использована и для других классификационных задач с небольшими модификациями.

Одним из наиболее важных требований для обучения модели глубокого обучения является наличие большого и репрезентативного набора данных. Хотя имеется несколько общедоступных и недавних наборов данных, доступных для целей исследования, для большинства проблем классификации, связанных с трафиком, не существует единого согласованного набора данных. Возможные причины включают в себя: 1) количество возможных классов трафика огромно, и практически невозможно, чтобы один набор данных содержал все типы трафика; 2) нет общепринятых методов сбора данных и маркировки; 3) различные методы сбора и сценарии приводят к различной доступности и распределению объектов. На практике исследователи часто собирают набор данных, специфичный для их цели классификации. Для этого первым шагом является определение места сбора данных. Сбор данных может происходить на стороне клиента или сервера канала связи, на краю сети, в ядре сети или в любом промежуточном месте. Точка сбора может кардинально повлиять на

доступные характеристики, надежную маркировку и обобщение, о которых пойдет речь далее. Правильная маркировка имеет решающее значение для эффективности методов классификации трафика. Однако маркировка данных не всегда является тривиальной. Некоторые исследования использовали свободные модули DPI, такие как nDPI и libprotoident, для маркировки захваченных данных. В таких случаях точность меток и, следовательно, любых соответствующих алгоритмов классификации ограничивается точностью методов DPI. Кроме того, такие методы обычно не работают для зашифрованного трафика. Контролируемая среда на стороне клиента связи была бы самым простым местом для маркировки данных. Это решение практично только тогда, когда точка захвата находится достаточно близко к источнику данных, чтобы убедиться, что нет никакого другого источника, который мог бы повлиять на маркировку. Кроме того, даже в полностью контролируемой среде нелегко полностью распознать и удалить фоновый трафик. Было показано, что 70% трафика смартфонов является фоновым трафиком и только 30% напрямую связано с пользовательскими взаимодействиями [1]. Несмотря на ограничения, наиболее распространенной стратегией на практике является сбор данных каждого класса в контролируемой среде.

Полезная информация в пакетах не всегда доступна. Пакеты, захваченные на беспроводных линиях связи или сотовой связи, шифруются на уровне 2, и поэтому полезные поля заголовка верхнего уровня не являются открытым текстом. Кроме того, в некоторых точках захвата, таких как маршрутизатор в центре интернет-провайдера, можно захватить только одно направление потока из-за асимметричной природы маршрутизации в Интернете. Кроме того, время может искажаться при агрегировании трафика, что является более серьезной задачей в ядре интернет-провайдеров. Это явление трансформирует

распределение межприватного времени и сильно зависит от состояния сети, нагрузки на трафик и времени. Длина пакета также может изменяться, когда трафик проходит через туннель, прокси и т. д. Наконец, все эти изменения также влияют на статистические характеристики, полученные из всего потока. Следовательно, модель, обученная на наборе данных, захваченном в одной точке захвата, может быть не столь точной при использовании в другой точке захвата.

Репрезентативный набор данных должен содержать разнообразные и обильные выборки из каждого класса, чтобы избежать перенасыщения. Было показано, что точность падает на целых 26%, когда OS / vendor отличается в обучающем и тестовом наборе [2]. Кроме того, модель может быть приспособлена к особенностям пользователей, а не к особенностям трафика, если набор данных содержит взаимодействия только одного или нескольких пользователей. Это также большое ограничение на исследования, которые захватили трафик, генерируемый сценарием [3], который, вероятно, имеет более детерминированное поведение. Как правило, набор данных, собранный дальше от клиентской стороны связи, например, в ядре интернет-провайдера, где наблюдается разнообразный трафик, менее подвержен этой проблеме. Лучший способ гарантировать, что обученная модель в наборе данных является репрезентативной, протестировать модель в тестовом наборе, который поставляется из другой конфигурации устройства/пользователя, чем обучающий набор.

Очистка и предварительная обработка данных существенно влияют на производительность алгоритмов ML. В сетевой среде некоторые относительно распространенные события могут изменить распределение функций на уровне пакетов. Например, ретрансляция пакетов, дублирование ACK и неупорядоченные пакеты могут изменить трафик. В некоторых исследованиях

сообщалось об улучшении после удаления таких пакетов [4], а в некоторых об отсутствии различий [2]. Это происходит потому, что для классификации используются различные наборы данных и объекты. Например, методы, которые используют статистические характеристики всего потока, вероятно, невосприимчивы к нескольким несвязанным пакетам. С другой стороны, методы, которые используют первые несколько пакетов для классификации, могут быть затронуты больше. Обратите внимание, что этот этап предварительной обработки иногда игнорируется из-за его вычислительной сложности. Еще одним этапом предварительной обработки, который имеет решающее значение для эффективности методов глубокого обучения, является нормализация данных. На этом шаге все входные объекты масштабируются, чтобы иметь значение в диапазоне [1; +1] (или [0; 1]). Это позволяет градиентным методам быстрее сходиться и выравнивает важность всех объектов при вычислении расстояния между точками данных.

Современные методы классификации трафика используют одну или несколько категорий функций. Временные ряды – функции временных рядов включают длину пакета, время между прибытиями и направление последовательных пакетов. Во многих исследованиях, где эти особенности были репрезентативными, было показано, что первых нескольких пакетов до первых 20 пакетов достаточно для разумной точности даже для зашифрованного трафика [12]. Недавно было также показано, что характеристики временных рядов набора выборочных пакетов достигают хорошей точности [3]. Заголовок – это включает в себя все полезные поля заголовка в пакете, обычно информацию уровня 3 и уровня 4, когда они не зашифрованы. В эпоху, предшествующую глубокому обучению, такие поля, как номер порта, протокол и длина пакета, были тщательно выбраны экспертами в качестве репрезентативных функций. В некоторых современных

подходах, особенно основанных на глубоком обучении, в качестве входных данных берутся целые пакеты [11]. Обратите внимание, что IP-адреса серверов могут использоваться для ограничения диапазона классов трафика для повышения точности в операционных сетях. Например, можно использовать IP-адреса Google для ограничения классов трафика для приложений Google. Тем не менее, IP-адреса должны использоваться разумно из-за широкого распространения использования CDNs и динамического выделения IP-адресов. Данные полезной нагрузки – даже для зашифрованного трафика существует информация над заголовком уровня 4, которую можно использовать для классификации. Например, некоторые исследования достигли высокой точности, используя пакеты рукопожатия TLS 1.2, которые содержат простые текстовые данные. Статистические характеристики – существует множество статистических характеристик, которые могут быть получены из всего потока, таких как средняя длина пакета, максимальная длина пакета и минимальное межвариантное время. Большое количество работ использовало эти особенности и продемонстрировало высокую точность классификации трафика. Однако, для получения статистических характеристик классификатор необходим для наблюдения всего потока или большой его части и поэтому пригоден только для автономной классификации. Кроме того, в некоторых случаях, таких как классификация приложений, статистические характеристики могут зависеть от специфики поведения пользователя, особенностей ОС, сетевых условий и т. д. Следовательно, набор данных следует собирать с большей осторожностью. Хотя временные ряды и статистические характеристики могут немного отличаться для незашифрованного трафика и зашифрованной версии одного и того же трафика, они доступны независимо от шифрования. Следовательно, методы, зависящие от этих функций для незашифрованного трафика, также могут работать с зашифрованным трафиком.

С другой стороны, данные полезной нагрузки и некоторая информация заголовка, например, информация уровня 4 о трафике, зашифрованном IPsec, могут не существовать в открытом тексте для зашифрованного трафика. Однако в этих случаях все еще существуют незашифрованные поля, доступные во время рукопожатия, которые можно использовать для классификации. Стоит отметить, что в некоторых случаях политика конфиденциальности и законы запрещают доступ или хранение содержимого пакетов, что ограничивает использование полезных функций.

В результате проведенного исследования и анализа выполненных работ, было выявлено, что общая структура классификации трафика состоит из семи этапов, применяемых в большинстве подходов для классификации трафика, также было выявлено, что репрезентативный набор данных должен содержать разнообразные и обильные выборки из каждого класса, для того, чтобы избежать перенасыщения выборки и обеспечить правильное обучение методов классификации для последующей их точной работы, а также было выявлено, что одним из ключевых этапов предварительной обработки, который имеет решающее значение для эффективности методов глубокого обучения, является нормализация данных. На этом шаге все входные объекты масштабируются, чтобы иметь значение в диапазоне [1; +1] (или [0; 1]). Это позволяет градиентным методам быстрее сходиться и выравнивает важность всех объектов при вычислении расстояния между точками данных, что обеспечивает наиболее точную работу методов классификации трафика.

В заключении следует отметить, что методы классификации трафика претерпели значительные изменения со временем. И как результат, появилось новое поколение методов, основанных на потоковой статистике трафика. Эти методы основаны на статистических или функциях временных рядов, которые позволяют им обрабатывать как зашифрованные, так и незашифрованный

трафик. Это алгоритмы машинного обучения, такие как случайный лес и к-ближайших соседей, а также глубокое обучение, которое является весьма желательным подходом к классификации трафика, особенно когда постоянно возникают новые классы, и модели старых классов развиваются.

Библиографический список:

1. Stber, T. Who do you sync you are?: smartphone fingerprinting via application behavior / T. Stber, M. Frank, J. Schmitt, I. Martinovic // Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks, ACM. 2013. – pp. 7-12.
2. Aceto G. Traffic Classification of Mobile Apps through Multi-classification / G. Aceto, D. Ciuonzo, A. Montieri, A. Pescap // GLOBECOM IEEE Global Communications Conference. 2017. – pp. 1-6.
3. Rezaei S. How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets / S. Rezaei, X. Liu // GLOBECOM IEEE Global Communications Conference. 2018. – pp. 12-20.
4. Dubin R. I Know What You Saw Last MinuteEncrypted HTTP Adaptive Video Streaming Title Classification / R. Dubin, A. Dvir, O. Pele, O. Hadar, // IEEE Transactions on Information Forensics and Security. 2019. – Vol. 12. –pp .3039-3049.
5. Taylor V. F. Robust smartphone app identification via encrypted network traffic analysis/ V. F. Taylor, R. Spolaor, M. Conti, I. Martinovic // IEEE Transactions on Information Forensics and Security. 2018. – Vol. 13. – pp. 63-78.

6. Aceto G. Mobile encrypted traffic classification using deep learning / G. Aceto, C. Domenico, M. Antonio, P. Antonio // Network Traffic Measurement and Analysis Conference (TMA).2018. – pp. 1-8.
7. Wang W. End-to-end encrypted traffic classification with one-dimensional convolution neural networks/ W. Wang, M. Zhu, J. Wang, X. Zeng, Z. Yang // In Intelligence and Security Informatics (ISI), IEEE International Conference on. IEEE. 2017. – pp. 43-48.
8. Chen Z. Seq2Img: A sequence-to-image based approach towards IP traffic classification using convolutional neural networks / Z. Chen, K. He, J. Li, Y. Geng // Big Data, IEEE International Conference on. IEEE.2017. – pp. 1271-1276.
9. Wang W. HAST-IDS: learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection / W. Wang // IEEE Access. 2018. – Vol. 6. – pp. 1792-1806.
10. Vu L. A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification / L. Vu, C.T. Bui, Q.U. Nguyen // Proceedings of the Eighth International Symposium on Information and Communication Technology. 2017. – pp. 333-339.
11. Lotfollahi M. Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning / M. Lotfollahi, R. Shirali, M.J. Siavoshani, M. Saberian // IEEE Access. 2018. – Vol. 6. – pp. 1652-1676.
12. Lopez-Martin M. Network traffic classifier with convolutional and recurrent neural networks for Internet of Things / M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, J. Lloret // IEEE Access. 2017. – Vol. 5. – pp.18042-18050.

13. Hochst J. Unsupervised Traffic Flow Classification Using a Neural Autoencoder / J. Hochst, L. Baumgartner, M. Hollick, B. Freisleben // IEEE 42nd Conference on Local Computer Networks (LCN). 2017. – pp. 523-526.
14. Zhang J. Robust network traffic classification / J. Zhang // IEEE/ACM Transactions on Networking (TON). 2015. – Vol. 23. – pp. 1257-1270.
15. Woodward M. Active one-shot learning / M. Woodward, C. Finn // NIPS Deep Reinforcement Learning Workshop. 2018. – pp. 8-15.

Оригинальность 87%