УДК 004.67

ИССЛЕДОВАНИЕ МЕТОДОВ АНАЛИЗА ИНТЕРНЕТ-ТРАФИКА

Вихров М.С.
магистр
МГТУ им. Н.Э. Баумана,
Россия, г. Москва

Аннотация

Анализ интернет-трафика - одна из самых популярных областей изучения и исследований в первую очередь из-за пользы для многих интернет приложений, так как решает задачу классификации интернет-трафика. Целью статьи является обзор технологий классификации и прогнозирования интернет-трафика различными методами машинного обучения. В статье также упоминается о важности использования метода дерева принятия решений в области классификации трафика. Данное направление исследования является крайне актуальным в связи с высоким темпом роста технологии интернет вещей, в связи с чем большими темпами растет число интернет адресов и количество трафика, который требуется идентифицировать. Полученные результаты исследования свидетельствуют о наилучших практиках, применяемых в методах анализа интернет-трафика, что дает направление для разработки методов классификации интернет-трафика.

Ключевые слова: транспортный анализ, выбор признаков, транспортный прогноз, дерево решений, классификация трафика.

RESEARCH OF INTERNET TRAFFIC ANALYSIS METHODS

Vikhrov M.S.
Master's student
Bauman Moscow State Technical University,
Russia, Moscow

Annotation

Internet traffic analysis is one of the most popular areas of study and research, primarily due to its usefulness for many Internet applications, as it solves the problem of classifying Internet traffic. The purpose of this article is to review technologies for classifying and predicting Internet traffic using various machine learning methods. The article also mentions the importance of using tree method of decision-making in the field of traffic classification. This research area is extremely relevant due to the high growth rate of the Internet of things technology, which is why the number of Internet addresses and the amount of traffic that needs to be identified is growing rapidly. The results of the study indicate the best practices used in the methods of Internet traffic analysis, which provides a direction for the development of methods for classifying Internet traffic.

Key words: transport analysis, feature selection, transport forecast, decision tree, traffic classification.

Анализ трафика важен для понимания поведения сетей и правильного использования ресурсов, для оценки QoS и для эффективного устранения неполадок. Актуальность данной темы обусловлена высоким темпом роста технологии интернет вещей и общего количества сетевых устройств, в связи с чем большими темпами растет число интернет адресов и количество трафика, который требуется идентифицировать. Важность анализа трафика также

обусловлена необходимостью обеспечить безопасность в сети. А требования к повышению скорости и точности классификации трафика открывают дорогу для многих исследований в этой области. Целью исследования является анализ различных подходов к идентификации сетевого трафика, с целью определить наиболее перспективные методы машинного обучения для использования при анализе трафика. В начале исследования были поставлены следующие задачи: исследовать характер трафика транспортного уровня, проанализировать сложность транспортного потока, провести сравнение различных функций для классификации трафика, определить методы машинного обучения, применяемые в анализе интернет-трафика. В ходе исследования использовались следующие научные методы: изучение теоретических основ по разработке и проектированию методов классификации, изучение и анализ документации DPI систем, рассмотрение результатов проведенных исследований, сбор и анализ информации о реализованных методах идентификации трафика, проведение сравнительного анализа методов машинного обучения.

Т. Нгуен и Г. Армитаж [10] заявляют в своей работе, что в настоящее время невозможно классифицировать сетевые пакеты на основе номера порта и полезной нагрузки. Статистические особенности сетевых пакетов теперь используются вместе с подходами машинного обучения в классификации интернет-трафика. Ведется большая исследовательская работа в области прогнозирования сетевого трафика, это очень похоже на прогнозирование дорожного трафика. Результаты прогноза могут использоваться, чтобы избегать заторов в пиковое время нагрузки, а также для правильного управления трафиком, проходящим через сеть. Последние работы в данной области рассматриваю технологии машинного обучения. Дерево решений и его различные версии используются для классификации трафика из-за их простой и комплексной архитектуры. В последнее время появились проблемы в области

классификации трафика [1]. Продолжающийся рост числа различных интернет-приложений и растущая мотивация маскировать их, чтобы избежать блокировки, является серьезным препятствием на пути классификации трафика. Большая часть сетевого трафика все еще не классифицируется всей техникой. UDP трафик игнорируется, и многие трассировки трафика не включают двунаправленные потоки. Исследователи используют разные уровни детализации для определения транспортных потоков и классов, что затрудняет сравнение различных методов между собой. Инкапсуляция протоколов и несколько каналов обслуживания ухудшают ситуацию еще больше.

Далее объясняется характер трафика, генерируемого одним источником, а также представлены различные методологии анализа интернет-трафика. В работе Д. Гомеса, П. Инасио «Анализ исходного трафика, транзакции ACM для мультимедийных вычислений, коммуникаций и приложений» [6] рассматривается трафик, генерируемый одним источником. Авторы изучают статистические свойства трафика и анализируют существующую корреляционную структуру среди сетевых пакетов. Цель исследования состоит в том, чтобы выделить наиболее подходящее распределение, которое соответствует времени поступления, размеру пакета и размеру пакета в байтах. Так же в своей работе Джао Гомес и соавторы [6] собирают и анализируют логи, содержащие трафик от различных классов, а затем применяют несколько известных алгоритмов распределения, таких как pareto, log-normal, weibull, rayleigh, для экспериментальных вариаций различных распределений трафика. Несмотря на то, что исследование полностью фокусируется на трафике, генерируемым одним источником, это не может дать нам прогноз относительно генерируемого сетевого трафика разными пользователями в сети или в нескольких сетях. Более того, статья учитывает только однонаправленные потоки, тогда как двунаправленные потоки полностью игнорируются для определения окончательных результатов.

В свою очередь, Волтер Виллингер и другие ученые. [11] утверждают, что время между сеансами является независимым и соответствует идентичному распределению, то есть следует пуассоновскому процессу. Количество байтов в пакете показывает явление долгосрочной зависимости.

Измерения трафика, выполненные путем настройки консервативных факторов выборки на маршрутизаторах с очень ограниченным локальным хранилищем, были освещены в статье китайских ученых, рассматривающих достаточность выбранных данных для обнаружения аномалий [5]. В рамках данного исследования было определено, что парадигма итеративного измерения имеет следующие преимущества:

- Интересные схемы движения могут быть обнаружены или изучены на лету;
- Онлайн анализ собранной информации;
- Итеративная эволюция последующих правил для более точной проверки трафика;
- Сокращение избыточных измерений;
- Более точный ответ на запрос пользователя.

Так же в статье Ф. Кхана [4] описывается «Мозаика с несколькими решениями», в которой правила набора настраиваются путем выполнения итеративного анализа собранных данных. Эта статья предлагает три алгоритма анализа потокового трафика:

- Равновесный откат (ER);
- Импульс потока (FM);
- Направленный импульс (DM).

Эти алгоритмы изменяют автономный характер традиционных измерений.

Существуют две категории моделей прогнозирования сетевого трафика, основанные на характеристиках сетевого трафика:

- Модель единого прогноза: для прогнозирования сети используется математическая модель движения трафика. Несколько примеров: авторегрессионная модель, серая модель, модель подобия и модель хаоса. Но эти модели страдают от недостатка точности;
- Модель комбинированного прогноза: прогнозируется путем объединения нескольких моделей. Примеры включают комбинированную модель нейронной сети Пуассона, ARIMA (авторегрессивное интегрированное скользящее среднее) и SNF (Neuro Fuzzy) комбинированная модель, ковариантная и нейронная сетевая комбинированная модель, QPSO (оптимизация квантового поведение частиц в рое) и BPNN (нейронная сеть обратного распространения) в сочетании с вейвлет-анализом. Эти модели страдают от ошибок в прогнозировании динамического обновления сети в реальном времени.

Анализ сложности транспортного потока рассматривался учеными с нескольких сторон. В рамках своей работы «Метод энтропии для анализа сложности транспортных потоков» Ж. Сян и другие ученые [12] проанализировали сложность временных рядов потока трафика, генерируемых моделью Нагель Шрекенберга (NS), исследование показало приемлемую сложность трафика для анализа методами машинного обучения.

В свою очередь, Мадлена Коста, Э. Голдбергер и К. Пэнг [7] предлагают многомасштабную энтропию. NS модель является одномерной моделью трафика на основе сотовых автоматов. Параметры для модели NS:

- Длина полосы круга L ≈ размер данных;
- Плотность сети - количество пользователей в сети;

Результаты утверждают следующее:

- Вероятность рандомизации и плотность сети влияют на сложность и время обработки;
- На больших временных масштабах, с одинаковыми вероятностями рандомизации, сложность находится на стабильном и высоком уровне при более высокой плотности сети;
- При увеличении масштаба времени с одинаковыми вероятностями рандомизации сложность уменьшается в сценариях с более низкой плотностью сети;

Модели, представленные для прогнозирования сетевого трафика, могут оказаться очень полезными для реального управления полосой пропускания. Прогнозирование характера трафика, генерируемого определенной организацией в разное время дня или года может помочь заинтересованным органам власти реагировать на заторы проактивным образом, а не реактивным. Это также может помочь в определении стоимости полосы пропускания, используемой этой организацией.

Выбор функций является очень важным шагом, который выполняется перед классификацией трафика. Это играет важную роль в повышении производительности классификатора. Авторы Д. Зуев и М. Кроган в своей работе [2] отмечают, что, если 248 статистических признаков используются для характеристики потока сетевого трафика, то стоимость вычислений классификатора будет очень высокой. Соответственно, существует три категории методов выбора объектов:

- Методы фильтрации: выбираются атрибуты «k», имеющие более высокую корреляцию с выходным классом. RELIEF и FOCUS - наиболее часто используемые фильтрующие методы. Нигель Уильямс в своей публикации отмечал, что он и его коллеги [8] обнаружили, что выбор

фильтрующего метода может значительно улучшить производительность вычислений и незначительно снизить точность классификации.

- Методы обертки: это поисковая стратегия, которая находит оптимальное подмножество путем добавления или удаления функций из набора входных функций.
- Встроенные методы выбора: они выполняют выбор подмножества объектов, вызывая логические описания. ID3 (итеративный дихотомайзер), C4.5 и CART (дерево классификации и регрессии) - несколько примеров этого подхода.

В свою очередь, Т. Эн-Наджары и его соавторы в своей работе [9] построили модель логистической регрессии для обработки класса проблем дисбаланса путем перевода мультиклассовой классификации в классовую классификацию. В рамках статьи «Предварительное сравнение производительности пяти алгоритмов машинного обучения для практической классификации потоков трафика IP» [8] изучаются два алгоритма сокращения набора функций и пять классификаторов измерения производительности на основе классификации скорости и времени наращивания. Выделяются следующие алгоритмы сокращения набора функций: выбор характеристик на основе согласованности и выбор признаков на основе корреляции. Так же отмечаются следующие классификаторы: BayesNet; наивный байесовский с использованием диктретизации; наивный байесовский с использованием оценки плотности ядра (NBK); наивное байесовское дерево; C4.5.

В статье авторами перечислены двадцать две особенности потока сетевого трафика, которые считаются полным набор функций. Так же в рамках данной публикации делается вывод, что C4.5 является самым быстрым алгоритмом при использовании любого из наборов функций, а NBK создает самый быстрый классификатор.

По сравнению со всеми алгоритмами сокращения признаков, метод «дерево решений» показал наиболее многообещающие результаты, по сравнению с другими методами, такими как «выбор функции фильтра» и метода обертки. Несмотря на то, что они просты, всеобъемлющи и лаконичны, время и сложность применения метода «дерево решений» существенно меньше.

Существуют различные механизмы управления пропускной способностью:

- Формирование трафика;
- Алгоритмы планирования;
- Предотвращение перегрузки;
- Протоколы / алгоритмы резервирования полосы пропускания;
- Классификация трафика.

Методы предотвращения перегрузки контролируют нагрузку сетевого трафика в попытках предвидеть и избегать перегрузки в общих узких местах сети. В своем руководстве по настройке решения для обеспечения качества обслуживания компания «Cisco» [3], специализирующаяся в области высоких технологий, перечислила следующие механизмы предотвращения заторов:

- Падение хвоста;
- Взвешенное ранее обнаружение (WRED) и распределенное WRED;
- DiffServ-совместимый WRED.

WRED, основанный на потоках, заставляет обеспечивать большую справедливость для всех потоков на интерфейсах в отношении того, как отбрасываются пакеты.

В результате проведенного исследования и анализа выполненных работ, было выявлено, что характер трафика в общем случае самоподобен, сложность транспортного потока приемлема для применения методов машинного

обучения, сравнение различных функций для классификации трафика показало лучшие алгоритмы в методах фильтрации, обертки и встроенных методах выбора, а также были выделены различные методы машинного обучения, применяемые в анализе трафика и показывающие наилучшие результаты.

В этой статье рассматриваются различные работы, выполненные в разных тематиках анализа интернет-трафика. Характер трафика самоподобен, и есть много доступных потоковых решений для повышения точности классификации пакетов, проходящих через программное и аппаратное обеспечение. Большая работа также была проделана исследователями в области прогнозирования трафика, что отражается в данной статье. В работах были выявлены подмножества функций из полного набора, чтобы улучшить точность и скорость классификации. Управление пропускной способностью это очень перспективная область, которая использует классификацию трафика в качестве основной задачи, что подтверждает высокую актуальность исследования данной проблемы.

Библиографический список:

1. Dainotti A. Issues and Future Directions in Traffic Classification/ A.Dainotti, A. Pescape // IEEE Network. 2012. – pp. 35.
2. Moore A. Discriminators for use in flow-based classification / A. Moore, D. Zuev, M. Crogan // Queen Mary and Westfield College, Dept. of Computer Science. 2005. – pp. 10.
3. Congestion Avoidance Overview Cisco IOS Quality of Service Solutions Configuration Guide, Release 12.2 [Электронный ресурс]. // CISCO URL:https://www.cisco.com/c/en/us/td/docs/ios/qos/configuration/guide/12_2sr/qos122sr_book.pdf (Дата обращения 20.02.2020)
4. Khan F. Streaming Solutions for Fine-Grained Network Traffic Measurements and Analysis/ F. Khan, N. Hosein, S. Ghiasi, C. Chuah, P. Sharma // IEEE/ACM Transactions On Networking. 2014. – pp. 377.

5. Mai J. Is sampled data sufficient for anomaly detection/ J. Mai, C. Chuah, A. Sridharan, T. Ye, H. Zang // Proc. 6th ACM SIGCOMM IMC .2006. – pp. 165.
6. Joao V. P. Gomes Source Traffic Analysis, ACM Transactions on Multimedia Computing/ J. V. P. Gomes, P. R. M. Inacio, B. Lakic, M. M. Freire, H. J. A. Silva, P. P. Monteiro // Communications and Applications. 2010. - №6 – pp. 2.
7. Costa M. Multiscale entropy analysis of biological signals/ M. Costa, A. L. Goldberger, C. K. Peng // Physical Review. 2005. – pp. 75.
8. Williams N. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification/ N. Williams, S. Zander, G. Armitage // ACM SIGCOMM Computer Communication Review. 2006. -№ 36 – pp. 5-7.
9. En-Najjary T. Application-based feature selection for internet traffic classification / T. En-Najjary, G. Urvoy-Keller, M. Pietrzyk, J. Costeux // Proceedings of 22nd International Teletraffic Congress. 2010. – pp. 10.
10. Nguyen T. T. A Survey of Techniques for Internet Traffic Classification using Machine Learning/ T. T. Nguyen, G. Armitage // IEEE Communications Surveys & Tutorials. 2008. - № 10 - pp. 56.
11. Willinger W. Self-similarity and heavy tails: Structural modeling of network traffic / W. Willinger, V. Paxson, M. S. Taqqu // A Practical Guide to Heavy Tails: Statistical Techniques and Applications. 1998. - pp. 27.
12. Xiang Z. Using Multiscale Entropy Method to Analyze the Complexity of Traffic Flow / Z. Xiang, Y. Li, L. Xiong // Second International Conference on Business Computing and Global Informatization. 2012. – pp. 785.

Оригинальность 90%