Информационные процессы, Том 24, № 1, 2024, стр. 73–83.

ПЕРЕДАЧА ИНФОРМАЦИИ В КОМПЬЮТЕРНЫХ СЕТЯХ

Классификация шифрованного трафика на основе характеристик смежных потоков[1]

А.А. Курапов[*,**], Д.Р. Шамсимухаметов[*], М.В. Любогощев[*], Д.В. Банков[*], Е.М. Хоров[*]

[*] *Институт проблем передачи информации им. А.А. Харкевича Российской академии наук*
[**] *Московский физико-технический институт (национальный исследовательский университет)*

Поступила в редколлегию 05.03.2024 г. Принята 25.04.2024 г.

Аннотация—Для обеспечения требуемого качества обслуживания трафика в современных сетях зачастую необходимо осуществлять классификацию трафика в режиме реального времени. Широкое использование протокола Transport Layer Security (TLS), в особенности в связке с расширением Encrypted ClientHello (ECH), усложняет классификацию трафика за счет шифрования трафика. В данной работе разрабатывается новый алгоритм классификации трафика ECH для обеспечения качества обслуживания (англ.: Quality of Service, QoS). Данный алгоритм анализирует не только классифицируемый (целевой) поток, но и его контекст, т. е. набор потоков, загружаемых одновременно с целевым. Из контекста потока выводятся новые признаки, которые разработанный алгоритм анализирует совместно с признаками целевого потока для решения задачи классификации. Оценка эффективности на большой актуальной базе данных показывает, что разработанный алгоритм классификации трафика значительно повышает качество классификации по сравнению с современными алгоритмами.

КЛЮЧЕВЫЕ СЛОВА: Классификация трафика, алгоритм «случайный лес», TLS, Encrypted ClientHello

DOI: 10.53921/18195822_2024_24_1_73

1. ВВЕДЕНИЕ

Классификация трафика является ключевым механизмом для управления сетью и выполнения требований к качеству обслуживания (англ.: Quality of Service, QoS). Она используется при инжиниринге трафика [1], маршрутизации [2], составлении расписания и распределении ресурсов сети [3]. Большинство задач обеспечения QoS необходимо решать в режиме реального времени, т. е. необходимо осуществлять классификацию потоков данных по их начальным пакетам, в идеале, до прибытия любого пакета с полезной нагрузкой приложения. Такую задачу часто называют ранней классификацией трафика [4, 5], и она исследуется в многочисленных работах [5–14]. Современные алгоритмы ранней классификации трафика используют методы машинного обучения, которые в последнее время получили широкое распространение в различных областях телекоммуникаций [15–17].

Объектом классификации является поток трафика, который представляет собой двунаправленную последовательность пакетов, отправляемых по сети между двумя узлами. Каждый поток однозначно идентифицируется пятью параметрами: протоколом транспортного уровня и парами IP-адресов и портов обменивающихся данными узлов. Поток трафика обычно состоит из пакетов, содержащих один тип данных, например, видео, аудио или веб [16].

[1] Исследование выполнено в ИППИ РАН за счет гранта Российского научного фонда No 21-79-10431, https://rscf.ru/project/21-79-10431/.

Шифрование сетевого трафика в настоящее время распространено повсеместно. Развиваются методы шифрования трафика, в т.ч. основанные на пост-квантовой криптографии [18]. При этом подавляющая доля сетевого трафика приходится на протокол HyperText Transfer Protocol (HTTP), а более 97% HTTP-трафика защищено с помощью протокола TLS [19]. Несмотря на то, что применение данного протокола повышает уровень конфиденциальности и безопасности конечных пользователей, оно, в то же время, затрудняет классификацию трафика.

Для согласования криптографических параметров и аутентификации сторон TLS-соединение начинается с рукопожатия (англ.: TLS-handshake). К сожалению, TLS-рукопожатие частично раскрывает конфиденциальную информацию, позволяющую осуществлять точную классификацию трафика [20]. В частности, TLS 1.3, самая последняя версия протокола, шифрует все сообщения рукопожатия, за исключением ClientHello (CH) и ServerHello (SH). В свою очередь, CH включает расширение Server Name Indication (SNI), которое содержит доменное имя сервера в нешифрованном виде. Используя доменное имя, любой сторонний наблюдатель может классифицировать тип данных, передаваемых в рамках данного потока TLS [21, 22].

Расширение Encrypted ClientHello (ECH) [23] протокола TLS устраняет данную уязвимость: при использовании ECH конфиденциальные метаданные внутри CH, включая SNI, шифруются. Как следствие, помимо повышения безопасности и приватности трафика, расширение ECH также усложняет классификацию трафика, ограничивая доступные для анализа данные. Тем не менее даже при использовании ECH некоторые параметры TLS, содержащиеся в сообщениях CH и SH остаются незашифрованными, например, это необходимо для обеспечения обратной совместимости протокола TLS.

Характеристики потока, используемые алгоритмом классификации трафика для предсказания, называют его признаками. В исследованиях различают потоковые и межпотоковые признаки [24]. Потоковые признаки характеризуют только конкретный поток и не учитывают параметры других активных потоков. Типичные потоковые признаки, рассматриваемые в классификации трафика: нешифрованная полезная нагрузка первых пакетов потока, статистика размеров и времени поступления пакетов [10, 25].

Однако современные сервисы, как правило, одновременно генерируют множество потоков в рамках одной пользовательской сессии: мультимедийные приложения параллельно загружают различные данные: веб-страницы, реклама, аудио, видео и т.д. [12, 26]. Кроме того, массовое использование сетей доставки содержимого (англ.: Content Delivery Networks, CDNs) затрудняет классификацию трафика, поскольку часть контента может загружаться с разных серверов в отдельных потоках [27]. Для повышения качества классификации в литературе рассматриваются подходы, учитывающие характеристики смежных потоков. Например, в подходе Bag-of-Flow (BoF) несколько потоков группируются на основе таких параметров, как используемый протокол транспортного уровня, а также IP-адрес и порт сервера, в предположении, что все потоки, генерируемые одним сервером, относятся к одной категории трафика. Основная идея подхода заключается в том, чтобы уменьшить ошибку классификации потоков, используя результаты предсказания потоков, принадлежащих одной группе BoF. Этот подход приобрел популярность в начале 2010-х годов [28–30]. К сожалению, в современном Интернете множество виртуальных серверов с различными типами контента скрываются под одними и теми же IP-адресами (что является основной причиной использования SNI в TLS) [31, 32] и стандартными портами, например, 443 для HTTPS.

По указанным причинам необходимы новые подходы к классификации трафика. Поэтому в данной статье рассматривается задача классификации трафика ECH с учетом признаков смежных потоков, иными словами, с учетом контекста потока. Для этого, во-первых, в работе разрабатывается новый критерий выбора множества смежных потоков классифицируемого потока. Во-вторых, для заданного критерия выбора множества смежных потоков предлагается

новый набор признаков, расширяющий признаковое пространство лучшего известного из литературы алгоритма классификации шифрованного трафика по полезной нагрузке RB-RF [22]. В-третьих, при помощи экспериментального исследования на обширной базе данных загрузок различных типов трафика демонстрируется, что разработанный алгоритм многократно снижает частоту ошибок при классификации трафика по типу данных по сравнению с лучшими известными из литературы алгоритмами.

Дальнейшее изложение организовано следующим образом. В разделе 2 рассматриваются существующие подходы к классификации трафика. В разделе 3 описан расширенный на статистические признаки целевого и смежных потоков алгоритм RB-RF. Затем, в разделе 4 дано подробное описание используемой базы данных и в разделе 5 обсуждаются результаты предложенного и базовых алгоритмов. Наконец, в разделе 6 приведены основные выводы данной работы.

2. ОБЗОР ЛИТЕРАТУРЫ

В данном разделе приведен обзор современных исследований, посвященных классификации трафика как по признакам целевого потока, так и с использованием межпотоковых признаков.

2.1. Классификация трафика по признакам целевого потока

По используемым для классификации признакам алгоритмы классификации трафика по признакам целевого потока можно разделить на три группы: на основе полезной нагрузки [10–12,33–35], на основе анализа статистики размеров и времени поступления пакетов целевого потока [25,36–41], а также гибридные методы [13,42–44], объединяющие признаки обеих категорий.

Все алгоритмы классификации трафика, основанные на анализе полезной нагрузки, являются алгоритмами ранней классификации трафика. Это достигается благодаря тому, что TLS-рукопожатие остается единственной информативной частью потока, поскольку шифрование полностью скрывает полезную нагрузку всех последующих пакетов данных приложения. В то же время, алгоритмы, основанные анализе статистики целевого потока, редко способны обеспечить раннюю классификацию трафика, поскольку они требуют сбора данных о большом числе пакетов потока, что неприемлемо увеличивает задержку классификации. Зато данные алгоритмы не полагаются на незашифрованные данные и способны классифицировать даже полностью зашифрованный трафик, например, трафик виртуальных частных сетей (англ.: Virtual Private Network, VPN) [45,46]. Наконец, гибридные алгоритмы компенсируют недостатки обоих подходов, обеспечивая высокую точность ранней классификации трафика.

Также алгоритмы классификации трафика можно различать по методологии извлечения признаков. Одна группа алгоритмов основана на нейронных сетях. Данные алгоритмы автономно извлекают признаки из необработанных байтов первых пакетов [10,11,33–35,47–49] или статистических характеристик [50–54] целевого потока. Другая группа алгоритмов основана на классических алгоритмах машинного обучения, которые извлекают признаки потоков некоторым образом. В частности, при классификации по полезной нагрузке потока, в основном, извлекают нешифрованные параметры TLS, такие, как Cipher Suites, Key Share, Protocol Version [12,13,55]. В случае же классификации по статистическим признакам потоков зачастую признаками являются минимальные, максимальные и средние значения, уникальные значения и дисперсия размеров пакетов в потоке и межпакетных интервалов [36,45,56–59].

Классификаторы на основе нейронных сетей не нуждаются в сторонних методах извлечения признаков. Однако эти классификаторы требуют значительных вычислительных ресурсов для обучения и предсказания, что может снижать их применимость в сценарии классификации

трафика в режиме реального времени на менее мощных устройствах, таких как домашние точки доступа Wi-Fi. Кроме того, в процессе предобработки байтов полезной нагрузки потоков, нейросети могут непреднамеренно интерпретировать шифрованные байты в качестве полезной информации, что может приводить к снижению эффективности по сравнению с прямым игнорированием шифрованных байтов алгоритмами на основе классических методов машинного обучения. Наконец, нейросетевые алгоритмы классификации плохо интерпретируемы, что может оказаться существенным недостатком во многих задачах, в частности, при детектировании вредоносного трафика [43].

2.2. Классификация трафика с учетом признаков смежных потоков

В работах [24, 28–30, 60] предлагается учитывать паттерны смежных потоков (т. е. потоков, возникающих почти одновременно с целевым), для повышения качества классификации трафика. Основным способом определения группы смежных потоков в данных работах является упомянутый в разделе 1 подход Bag of Flows (BoF). Он вводит одноименное понятие BoF. BoF — это группа потоков с совпадающей тройкой параметров: используемый протокол транспортного уровня, IP-адрес и порт сервера [29]. Основное предположение, лежащее в основе данных работ, заключается в том, что все потоки от одного сервера имеют одинаковый тип. Тогда все потоки внутри BoF необходимо классифицировать одинаково.

Процесс классификации происходит в две стадии. Сначала каждый поток из BoF классифицируется независимо с помощью традиционных алгоритмов классификации трафика, которые могут быть основаны на анализе полезной нагрузки, статистических признаках целевого потока или являться гибридными. Затем единое решение выносится по всей группе BoF по правилу большинства.

Данный подход в основном рассматривается в сценариях классификации трафика по протоколам. Например, авторы работы [30] показали, что подход, основанный на комбинации классификатора k-Nearest Neighbors (k-NN), анализирующего статистические признаки целевого потока, и концепции BoF, превосходит стандартный k-NN в классификации таких протоколов, как HTTP, BitTorrent, SMTP, SSH, IMAP и др.

Однако группировка потоков в BoF с помощью протокола транспортного уровня и IP-адреса и порта сервера — не единственная стратегия в классификации трафика по признакам смежных потоков. По мере распространения сервисов CDN эффективность BoF стремительно снижается, поскольку один сервер с одним IP-адресом может независимо генерировать трафик разного типа. Поэтому в исследовании [61] предлагается объединить потоки в одну группу на основе методов кластеризации таких как k-means или DBSCAN (Density-Based Spatial Clustering of Applications with Noise), анализирующих, помимо IP-адресов и портов, характеристики потоков, такие как интервалы между приходами пакетов и длины пакетов в байтах. Авторы называют каждый кластер потоков «пучком» потоков. Для определения класса всех потоков «пучка» используется схема голосования по d кандидатам, где значения d может варьироваться в диапазоне от единицы до размера «пучка». В этой схеме голосование проводится не по всем потокам из «пучка», а только по специально отобранным d из них. Величина d определяет компромисс между точностью классификации и сложностью вычислений. В работе продемонстрировано высокого качество классификации на базе данных трафика восьми популярных приложений, однако эффективность такого подхода на более разнообразном наборе данных остается под вопросом.

Итак, подходы классификации трафика, основанные на анализе паттернов смежных потоков, представляют ряд возможностей для улучшения качества классификации. К сожалению, агрегирование потоков в одну группу BoF по протоколу транспортного уровня, IP-адресу и порту сервера неэффективно в современной сети Интернет из-за распространения сервисов

CDN. Поэтому в данной работе предложен новый способ учета данных о потоках, загружающихся параллельно с целевым.

3. РАСШИРЕННЫЙ АЛГОРИТМ RB-RF

В основе разработанного алгоритма лежит алгоритм RB-RF, анализирующий полезную нагрузку TLS и предложенный в работе [12]. Он работает в два этапа. На первом этапе производится переупаковка нешифрованных в сценарии ECH байтов полезной нагрузки первых TLS-сообщений CH и SH в единый вектор. Затем, во время второго этапа, данный вектор подается на вход алгоритма «случайный лес» (англ.: Random Forest). В работе [12] показано, что данный алгоритм имеет высокую точность даже в сценарии повышенной безопасности ECH. Однако в той работе RB-RF исследовался на базе данных TLS-потоков мощностью всего 3547 потоков. Кроме того, потоки одних классов слабо разнесены по времени сбора. В статьях [62,63] показано, что при достаточном разнообразии базы данных в терминах мощности выборки (более 120 000 потоков) и времени сбора (более 4 месяцев) RB-RF имеет точность менее 50% в метрике F-Score.

Алгоритм RB-RF улучшен двумя способами.

Во-первых, чтобы обеспечить высокое качество классификации на разнообразной базе данных, в данной работе признаковое пространство RB-RF расширено на лучшие из известных в литературе [13, 14, 43] статистические признаки целевого потока. Таким образом, помимо полезной нагрузки, расширенный RB-RF анализирует длины и интервалы между поступлениями пакетов в буфер, а также средние значения этих параметров, их дисперсию и квантили.

Во-вторых, добавлен учет контекста. По причинам, изложенным в разделе 2, концепция BoF в текущем виде неприменима в современных сетях. Тем не менее учет признаков смежных потоков в дополнение к классическим признакам полезной нагрузки TLS и статистическим признакам целевого потока может повысить качество классификации. Поэтому в данной статье RB-RF расширяется для анализа признаков потоков, составляющих контекст целевого потока. Контекст целевого потока определяется как совокупность всех потоков, загружаемых тем же клиентом, что и целевой поток, и которые начали загружаться за менее чем 0,5 секунды до начала целевого потока. Для каждого потока из контекста подсчитываются признаки, аналогичные признакам полезной нагрузки и статистическим признакам целевого потока, но исключительно по пакетам, загруженным в интервале от начала потока из контекста до первого нисходящего пакета с данными уровня приложения целевого потока.

Итак, признаковое пространство расширенного алгоритма RB-RF состоит из признаков нешифрованной полезной нагрузки первых TLS-сообщений, статистических признаков целевого потока, а также признаков смежных потоков. Описанные признаки, на втором этапе расширенного алгоритма RB-RF, подаются на вход классификатора «случаный лес», который осуществляет предсказание для каждого целевого потока.

4. БАЗА ДАННЫХ

Для исследования эффективности расширенного алгоритма RB-RF используется база данных из работы [62], расширенная на потоки буферизованного аудиотрафика сервиса Apple Music (см. Таблицу 1). В отличие от оригинальной статьи, веб-трафик разных сервисов в экспериментах объединен в единый класс «веб». Кроме того, сервисы Spotify, VkMusic, YandexMusic и AppleMusic также составляют единый класс «буферизованный аудиотрафик», а потоки видеотрафика сервисов коротких видеоклипов и буферизованного потокового видеотрафика отнесены к классу «буферизованный видеотрафик». Наконец, класс «трафик прямого видеотранслирования» содержит потоки LiveVideo-YouTube. Аналогично работе [64], база данных

Таблица 1. База данных.

No	**Тип Трафика**	**Сервис**	**Число потоков**	**SNI паттерн**
1	буферизованный аудиотрафик	Spotify	668	*audio*spotify*akamai* *audio*scdn*
2		VkMusic	199	*vkuseraudio*
3		YandexMusic	111	*storage*yandex.*
4		AppleMusic	86	*od*itunes.apple.* *audio*itunes.apple.*
5	видеотрафик сервисов коротких видеоклипов	TikTok	210	*tiktokcdn*
6		VkClips	1020	*vkvd*
7		YouTubeShorts	1976	r*-*googlevideo.*
8	буферизованный потоковый видеотрафик	Kinopoisk	192	*strm*yandex.*
9		Vimeo	638	*vod-adaptive*akamai*
10		VkVideo	180	*vkvd*
11		YouTube	278	r*-*googlevideo*
12	трафик прямого видеотранслирования	YouTubeLive	911	*rtmps.youtube* *upload.youtube*
13	веб-трафик	различные	147847	другие

размечена по SNI. Паттерны SNI, использованные для разметки каждого из типов трафика представлены в таблице 1.

Из предложенной базы потоков извлечены признаки, описанные в разделе 3. Предобработанная база данных содержит признаки полезной нагрузки с учетом шифрования ECH, статистические признаки целевого и смежных потоков, а также метку одного из четырех QoS-классов каждого потока.

5. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

В экспериментах база данных разбита на обучающую и тестовую выборки в соотношении 8:2 соответственно. В качестве базовых алгоритмов рассматриваются алгоритм классификации трафика по полезной нагрузке целевого потока RB-RF, описанный в работе [12], и гибридный алгоритм классификации трафика по признакам целевого потока hC4.5, описанный в работе [13]. Решающие деревья классификаторов «случайный лес» и «C4.5» были построены для 20 различных значений параметра random_state библиотеки sklearn [65]. Результаты обученных алгоритмов усреднены.

Базовые алгоритмы используют оптимальные гиперпараметры, определенные в соответствующих оригинальных исследованиях. Для расширенного RB-RF в качестве гиперпараметров взяты максимальная доля используемых признаков равная 0,7 и 150 решающих деревьев. Для остальных гиперпараметров использованы стандартные значения из библиотеки sklearn.

Для оценки их эффективности используются стандартная метрика F-score [66], а также частота ошибок, т. е. (1 – F-score). Кроме того, в данном разделе исследуется разделяющая способность признаков предложенного алгоритма с помощью метрики Gini-Impurity (GI) [67].

В таблице 2 приведены результаты базовых и разработанного в данной работе алгоритмов для задачи классификации трафика на рассматриваемой базе данных. Результаты показывают, что предложенный алгоритм достигает усредненного по всем классам значения F-score 98,9%, что соответствует снижению частоты ошибок более чем в 4 раза по сравнению с hC4.5. Более того, расширенный RB-RF превосходит hC4.5 по точности классификации каждого из классов в отдельности. Наконец, результаты алгоритма RB-RF показывают, что при доста-

точном разнообразии данных точная классификация трафика исключительно по полезной нагрузке невозможна: несмотря на сравнительно небольшое число целевых классов, алгоритм RB-RF достигает лишь 66,2% точности в метрике F-score.

Таблица 2. Эффективность алгоритмов в задаче с разметкой по категориям QoS

No	**класс**	**F-score, [%]**		
		расширенный RB-RF	**hC4.5**	**RB-RF**
1	Буферизованный аудиотрафик	98.3	88,8	14,8
2	Буферизованный видеотрафик	97,6	92,4	51,4
3	Трафик прямого видеотранслирования	100.0	99,9	100,0
4	Веб	99,9	99,7	98,4
	Среднее	98,9	95,2	66,2

Таблица 3. Интерпретация результатов расширенного RB-RF

Нормированная сумма GI признаков «расширенного RB-RF»			
ClientHello	ServerHello	целевой поток	смежные потоки
22,7%	13,8%	44,0%	20%

В таблице 3 приведены нормированные суммы значений GI признаков каждой из четырех групп, соответствующих признакам полезной нагрузки и статистическим признакам целевого и смежных потоков. Из представленных результатов можно видеть, что все группы признаков имеют высокую разделяющую способность. При этом наибольшую разделяющую способность, 44%, имеют статистические признаки целевого потока. Признаки полезной нагрузки TLS-сообщений CH, SH имеют суммарную разделяющую способность 35% относительно всех признаков CH, SH и статистических признаков целевого и смежных потоков. Наконец, признаки смежных потоков также имеют высокую разделяющую способность — 20%. Таким образом, признаки каждой из четырех групп оказываются полезны для решения задачи классификации.

6. ЗАКЛЮЧЕНИЕ

В данной статье исследуется задача классификации TLS-трафика в сценарии расширенной безопасности ECH. В ней представлен новый подход, предполагающий анализ не только свойств классифицируемого (целевого) потока, но и паттернов смежных потоков, таких как распределения времени между пакетами и размеры пакетов в смежных потоках. Оценка эффективности предложенного алгоритма на разнообразной базе данных трафика TLS показывает, что учет этих особенностей позволяет в четыре раза снизить среднюю частоту ошибок по сравнению с современными алгоритмами классификации шифрованного трафика. Также в статье исследуется разделяющая способность групп признаков по полезной нагрузке TLS-сообщений CH и SH, статистических признаков целевого потока, а также признаков смежных потоков, и демонстрируется, что все рассматриваемые группы признаков оказываются полезны для решения задачи классификации шифрованного трафика.

СПИСОК ЛИТЕРАТУРЫ

1. Huang Huawei, Guo Song, Li Peng et al. Joint optimization of rule placement and traffic engineering for QoS provisioning in software defined network // IEEE Transactions on Computers. 2015. Vol. 64, no. 12. Pp. 3488–3499.

2. Lin Chienhung, Wang Kuochen, Deng Guocin. A QoS-aware routing in SDN hybrid networks // Procedia Computer Science. 2017. Vol. 110. Pp. 242–249.

3. Liubogoshchev Mikhail, Zudin Dmitry, Krasilov Artem et al. DeSlice: An Architecture for QoE-Aware and Isolated RAN Slicing // Sensors. 2023. Vol. 23, no. 9. P. 4351.

4. Rezaei Shahbaz, Kroencke Bryce, Liu Xin. Large-scale mobile app identification using deep learning // IEEE Access. 2019. Vol. 8. Pp. 348–362.

5. Dainotti Alberto, Pescapé Antonio, Sansone Carlo. Early classification of network traffic through multi-classification // International Workshop on Traffic Monitoring and Analysis / Springer. 2011. Pp. 122–135.

6. Huang N-F, Jai G-Y, Chao H-C. Early identifying application traffic with application characteristics // 2008 IEEE International Conference on Communications / IEEE. 2008. Pp. 5788–5792.

7. Huang Nen-Fu, Jai Gin-Yuan, Chao Han-Chieh et al. Application traffic classification at the early stage by characterizing application rounds // Information Sciences. 2013. Vol. 232. Pp. 130–142.

8. Gómez Sena Gabriel, Belzarena Pablo. Early traffic classification using support vector machines // Proceedings of the 5th International Latin American Networking Conference. 2009. Pp. 60–66.

9. Qu Buyu, Zhang Zhibin, Guo Li, Meng Dan. On accuracy of early traffic classification // 2012 IEEE Seventh International Conference on Networking, Architecture, and Storage / IEEE. 2012. Pp. 348–354.

10. Cheng Jin, Wu Yulei, Yuepeng E et al. MATEC: A lightweight neural network for online encrypted traffic classification // Computer Networks. 2021. Vol. 199. P. 108472.

11. Liu Xun, You Junling, Wu Yulei et al. Attention-based bidirectional GRU networks for efficient HTTPS traffic classification // Information Sciences. 2020. Vol. 541. Pp. 297–315.

12. Shamsimukhametov Danil, Kurapov Anton, Liubogoshchev Mikhail, Khorov Evgeny. Is encrypted clienthello a challenge for traffic classification? // IEEE Access. 2022. Vol. 10. Pp. 77883–77897.

13. Shbair Wazen M, Cholez Thibault, François Jérôme, Chrisment Isabelle. Early Identification of Services in HTTPS Traffic // arXiv preprint arXiv:2008.08350. 2020.

14. Shamsimukhametov Danil, Kurapov Anton, Liubogoshchev Mikhail, Khorov Evgeny. Early Traffic Classification with Encrypted ClientHello: a World-Wide Study // IEEE Access. 2024. Manuscript ID Access-2024-07906 submitted on 28-Feb-2024.

15. Glinskiy Kirill, Kureev Aleksey, Khorov Evgeny. ALPACA: An Asymmetric Loss Prediction Algorithm for Channel Adaptation Based on a Convolutional-Recurrent Neural Network in URLLC Systems // IEEE Access. 2023.

16. Uddin Mostafa, Nadeem Tamer. TrafficVision: A case for pushing software defined networks to wireless edges // 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS) / IEEE. 2016. Pp. 37–46.

17. Kurapov Anton, Shamsimukhametov Danil, Liubogoshchev Mikhail, Khorov Evgeny. CloudETC: a Privacy-Preserving Encrypted Traffic Classification Platform for QoS in Wi-Fi // 2023 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom) / IEEE. 2023. Pp. 244–246.

18. Zyablov, V., Ivanov, F., Krouk, E. et al. On New Problems in Asymmetric Cryptography Based on Error-Resistant Coding. Probl Inf Transm 58, 2022. Pp. 184–201.

19. HTTParchive. `https://httparchive.org/reports/state-of-the-web#pctHttps`. Online: accessed 22 March 2024.

20. Rescorla Eric. The Transport Layer Security (TLS) Protocol Version 1.3. RFC 8446. 2018. (Accessed on 1/02/2023). https://tools.ietf.org/html/rfc8446.

21. Bermudez Ignacio N, Mellia Marco, Munafo Maurizio M et al. Dns to the rescue: Discerning content and services in a tangled web // Proceedings of the 2012 Internet Measurement Conference. 2012. Pp. 413–426.

22. Shamsimukhametov D, Liubogoshchev M, Khorov E, Akyldiz IF. Are Neural Networks the Best Way for Encrypted Traffic Classification? // IEEE En&T. 2021. Pp. 1–5.

23. Rescorla Eric, Oku Kazuho, Sullivan Nick, Wood Christopher A. TLS Encrypted Client Hello: Internet-Draft draft-ietf-tls-esni-18: Internet Engineering Task Force, 2024. Work in Progress. https://datatracker.ietf.org/doc/draft-ietf-tls-esni/18/.

24. Zhao Jingjing, Jing Xuyang, Yan Zheng, Pedrycz Witold. Network traffic classification for data fusion: A survey // Information Fusion. 2021. Vol. 72. Pp. 22–47.

25. Rezaei Shahbaz, Liu Xin. Deep learning for encrypted traffic classification: An overview // IEEE communications magazine. 2019. Vol. 57, no. 5. Pp. 76–81.

26. Mansy Ahmed, Fayed Marwan, Ammar Mostafa. Network-layer fairness for adaptive video streams // 2015 IFIP Networking Conference (IFIP Networking) / IEEE. 2015. Pp. 1–9.

27. Ma Xiaoteng, Li Qing, Zou Longhao et al. QAVA: QoE-aware adaptive video bitrate aggregation for HTTP live streaming based on smart edge computing // IEEE Transactions on Broadcasting. 2022. Vol. 68, no. 3. Pp. 661–676.

28. Zhang Jun, Chen Chao, Xiang Yang et al. Internet traffic classification by aggregating correlated naive bayes predictions // IEEE transactions on information forensics and security. 2012. Vol. 8, no. 1. Pp. 5–15.

29. Zhang Jun, Xiang Yang, Wang Yu et al. Network traffic classification using correlation information // IEEE Transactions on Parallel and Distributed systems. 2012. Vol. 24, no. 1. Pp. 104–117.

30. Divakaran Dinil Mon, Su Le, Liau Yung Siang, Thing Vrizlynn LL. SLIC: Self-learning intelligent classifier for network traffic // Computer Networks. 2015. Vol. 91. Pp. 283–297.

31. Yamauchi Hiroaki, Nakao Akihiro, Oguchi Masato et al. A Study on Service Identification Based on Server Name Indication Analysis. 2019. Pp. 470–474.

32. Yamauchi Hiroaki, Nakao Akihiro, Oguchi Masato et al. Service Identification Based on SNI Analysis. 2020. Pp. 1–6.

33. Zheng Weiping, Zhong Jianhao, Zhang Qizhi, Zhao Gansen. MTT: an efficient model for encrypted network traffic classification using multi-task transformer // Applied Intelligence. 2022. Vol. 52, no. 9. Pp. 10741–10756.

34. Xu Yuwei, Cao Jie, Song Kehui et al. FastTraffic: A lightweight method for encrypted traffic fast classification // Computer Networks. 2023. Vol. 235. P. 109965.

35. Malekghaini Navid, Akbari Elham, Salahuddin Mohammad A et al. AutoML4ETC: Automated neural architecture search for real-world encrypted traffic classification // IEEE Transactions on Network and Service Management. 2023.

36. Wang Chao, Xu Tongge, Qin Xi. Network traffic classification with improved random forest // CIS2015 / IEEE. 2015. Pp. 78–81.

37. Taylor Vincent F, Spolaor Riccardo, Conti Mauro, Martinovic Ivan. Appscanner: Automatic fingerprinting of smartphone apps from encrypted network traffic // 2016 IEEE European Symposium on Security and Privacy (EuroS&P) / IEEE. 2016. Pp. 439–454.

38. Shen Meng, Liu Yiting, Zhu Liehuang et al. Optimizing feature selection for efficient encrypted traffic classification: A systematic approach // IEEE Network. 2020. Vol. 34, no. 4. Pp. 20–27.

39. Zaki Faiz, Afifi Firdaus, Abd Razak Shukor et al. GRAIN: Granular multi-label encrypted traffic classification using classifier chain // Computer Networks. 2022. Vol. 213. P. 109084.

40. Shen Meng et al. Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey // IEEE Communications Surveys & Tutorials. 2023. Vol. 25, no. 1. Pp. 791–824.

41. Dener Murat, Al Samed, Ok Gökçe. RFSE-GRU: Data Balanced Classification Model for Mobile Encrypted Traffic in Big Data Environment // IEEE Access. 2023. Vol. 11. Pp. 21831–21847.

42. Malekghaini Navid, Akbari Elham, Salahuddin Mohammad A et al. Deep learning for encrypted traffic classification in the face of data drift: An empirical study // Computer Networks. 2023. Vol. 225. P. 109648.

43. Anderson Blake, McGrew David. Machine learning for encrypted malware traffic classification: accounting for noisy labels and non-stationarity // Proceedings of the 23rd ACM SIGKDD. 2017. Pp. 1723–1732.

44. Lin Peng, Ye Kejiang, Hu Yishen et al. A novel multimodal deep learning framework for encrypted traffic classification // IEEE/ACM Transactions on Networking. 2022.

45. Ma Chencheng, Du Xuehui, Cao Lifeng. Improved KNN Algorithm for Fine-Grained Classification of Encrypted Network Flow // Electronics. 2020. https://www.mdpi.com/2079-9292/9/2/324. https://www.mdpi.com/2079-9292/9/2/324.

46. Draper-Gil Gerard, Lashkari Arash Habibi, Mamun Mohammad Saiful Islam, Ghorbani Ali A. Characterization of encrypted and vpn traffic using time-related // Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016. Pp. 407–414.

47. Wang Wei, Zhu Ming, Wang Jinlin et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks // 2017 IEEE International Conference on Intelligence and Security Informatics (ISI) / IEEE. 2017. Pp. 43–48.

48. Wang Wei, Zhu Ming, Zeng Xuewen et al. Malware traffic classification using convolutional neural network for representation learning // 2017 International Conference on Information Networking (ICOIN) / IEEE. 2017. Pp. 712–717.

49. Wang Zhanyi. The applications of deep learning on traffic identification // BlackHat USA. 2015. Vol. 24, no. 11. Pp. 1–10.

50. Chen Zhitang, He Ke, Li Jian, Geng Yanhui. Seq2img: A sequence-to-image based approach towards ip traffic classification using convolutional neural networks // 2017 IEEE International conference on big data (big data) / IEEE. 2017. Pp. 1271–1276.

51. Pathmaperuma Madushi H, Rahulamathavan Yogachandran, Dogan Safak, Kondoz Ahmet M. Deep Learning for Encrypted Traffic Classification and Unknown Data Detection // Sensors. 2022. Vol. 22, no. 19. P. 7643.

52. Rasteh Ali, Delpech Florian, Aguilar-Melchor Carlos et al. Encrypted internet traffic classification using a supervised spiking neural network // Neurocomputing. 2022. Vol. 503. Pp. 272–282.

53. Roy Sangita, Shapira Tal, Shavitt Yuval. Fast and lean encrypted Internet traffic classification // Computer Communications. 2022. Vol. 186. Pp. 166–173.

54. Luxemburk Jan, Čejka Tomáš. Fine-grained TLS services classification with reject option // Computer Networks. 2023. Vol. 220. P. 109467.

55. Anderson Blake, McGrew David. Identifying encrypted malware traffic with contextual flow data // Proceedings of the 2016 ACM workshop on artificial intelligence and security. 2016. Pp. 35–46.

56. Jaber Mohamad, Cascella Roberto G, Barakat Chadi. Can we trust the inter-packet time for traffic classification? // 2011 IEEE international conference on communications (ICC) / IEEE. 2011. Pp. 1–5.

57. Kumano Yuichi, Ata Shingo, Nakamura Nobuyuki et al. Towards real-time processing for application identification of encrypted traffic // 2014 International Conference on Computing, Networking and Communications (ICNC) / IEEE. 2014. Pp. 136–140.

58. Chari Madhusoodhana, Srinidhi H, Somu Tamil Esai. Network traffic classification by packet length signature extraction // 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) / IEEE. 2019. Pp. 1–4.

59. Shbair Wazen M, Cholez Thibault, Francois Jerome, Chrisment Isabelle. A multi-level framework to identify https services // NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium / IEEE. 2016. Pp. 240–248.

60. Ding Lei, Liu Jun, Qin Tao, Li Haifei. Internet traffic classification based on expanding vector of flow // Computer networks. 2017. Vol. 129. Pp. 178–192.

61. Chen Zihan, Cheng Guang, Wei Zijun et al. Classify Traffic Rather Than Flow: Versatile Multi-Flow Encrypted Traffic Classification With Flow Clustering // IEEE Transactions on Network and Service Management. 2023.

62. Shamsimukhametov D. R., Kurapov A. A., Liubogoshchev M. V., Khorov E. M. Indistinguishability of Traffic by Open TLS Parameters with Encrypted ClientHello // Journal of Communications Technology and Electronics. 2023. Vol. 68, no. 12.

63. Курапов АА, Шамсимухаметов ДР, Любогощев МВ, Хоров ЕМ. О классификации трафика по метаданным протокола HTTP/3 // Информационные процессы. 2023. Vol. 23, no. 4. Pp. 568–578.

64. Шамсимухаметов ДР, Курапов АА, Любогощев МВ, Хоров ЕМ. Неразличимость трафика по открытым параметрам TLS при использовании Encrypted ClientHello // Информационные процессы. 2023. Vol. 23, no. 2. Pp. 231–240.

65. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. Pp. 2825–2830.

66. Powers David. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation // Mach. Learn. Technol. 2008. — 01. Vol. 2.

67. Rokach Lior, Maimon Oded. Top-down induction of decision trees classifiers-a survey // IEEE Transactions on Systems, Man, and Cybernetics, Part C. 2005. Vol. 35, no. 4. Pp. 476–487.

Traffic Classification Based on Neighbour Flows Patterns

A.A. Kurapov, D.R. Shamsimukhametov, M.V. Liubogoshchev, D.V. Bankov, E.M. Khorov

Real-time traffic classification is necessary to ensure the required quality of service (QoS) in modern networks. The widespread use of the Transport Layer Security (TLS) protocol, especially in conjunction with the Encrypted ClientHello (ECH) extension, complicates traffic classification by encrypting traffic. This paper develops a new ECH traffic classification algorithm for QoS provisioning. The algorithm analyzes not only the flow to be classified (the target one), but also the neighboring flows, i.e. the set of flows loaded simultaneously with the target flow. New features are derived from the flow neighbors, which the developed algorithm analyzes along with the features of the target flow to solve the classification problem. Performance evaluation on a large heterogeneous up-to-date database shows that the developed traffic classification algorithm significantly improves the classification quality compared with the state-of-the-art algorithms.

KEYWORDS: Traffic Classification, Random Forest, TLS, Encrypted ClientHello