УДК 004.85
DOI: 10.31673/2412-9070.2020.062932

А. П. КОЗИРЯЦЬКИЙ, студент;
В. В. ЖЕБКА, канд. техн. наук, доцент;
Л. О. ДЬОМІНА,
Д. О. ТАРАСЕНКО, аспірант,
Державний університет телекомунікацій, Київ

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ЗАСТОСУВАННЯ АЛГОРИТМУ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ІНТЕРНЕТ-ТРАФІКУ

Досліджено ефективність застосування алгоритму машинного навчання для класифікації інтернет-трафіку. Розглянуто алгоритм RF, який діє через побудову безлічі вирішальних дерев. Оцінено ефективність роботи алгоритму RF у задачах класифікації додатків за наявності і відсутності фонового мережного трафіку. Для збору необхідних для аналізу даних було організовано лабораторну мережу з кількох комп'ютерів. Один із комп'ютерів було підімкнено до глобальної мережі Інтернет і на його базі організовано безпроводову точку доступу. На цьому самому комп'ютері здійснювалося захоплення всього трафіку, що проходить через нього, за допомогою програми Wireshark. На інших комп'ютерах, підімкнених до точки доступу, було запущено різні додатки. Здійснювався перегляд веб-сторінок із використанням браузерів Google Chrome і Opera, за допомогою програми Skype проводилися відеодзвінки, виконувалося скачування файлів через торрент клієнта µTorrent, використання сервісу цифрового поширення комп'ютерних ігор Steam тощо. Здобуті дані зберігалися в форматі PCAP. Для приведення отриманих даних у відповідність до вимог розв'язуваного завдання здійснювалося попереднє оброблення даних. В експерименті було проведено побудову випадкового лісу і оцінювання якості класифікації на заданій вибірці. Дослідним шляхом було відібрано найбільш прийнятні параметри алгоритму. Експериментально вибрано, що ліс складається з п'яти дерев із максимально можливою глибиною. Найбільшу ефективність алгоритм має для даних, що належать до DNS трафіку. Крім перевірки роботи алгоритму на тестовій вибірці, що має такий самий класовий склад, як і навчальна, оцінювання його якості проводилося також за наявності фонового трафіку, тобто в разі, коли тестова вибірка містила екземпляри класів, відсутніх у навчальній вибірці.

Ключові слова: машинне навчання; інтернет-трафік; алгоритм RF; програма Wireshark; ефективність; метрики.

Вступ

Класифікація інтернет-трафіку дає можливість виявляти його тип і структуру, що критично важливо для керування такими технологіями, як мережна безпека, диференціація сервісів, керування параметрами трафіку тощо.

Для ефективного керування мережею необхідне точне узгодження застосовних мережних додатків із відповідним трафіком, а також повноцінний контроль над використовуваними додатками. Обмеженість традиційних методів призвела до того, що останніми роками інтенсифікувалися дослідження з пошуку і розвитку альтернативних підходів до ефективної класифікації мережного трафіку.

Для класифікації трафіку особливий інтерес становлять технології машинного навчання (Machine Learning) і інтелектуального аналізу даних (Data Mining), що виявилися найбільш ефективними в різних галузях інформатики, радіотехніки та інших напрямках [1].

Основна частина

Як показали дослідження, одним із найчастіше використовуваних і ефективних для класифікації мережного трафіку із застосуванням машинного навчання є застосування алгоритму RF, який діє завдяки побудові безлічі вирішальних дерев. Оцінимо ефективність роботи алгоритму RF у задачах класифікації додатків за умов наявності і відсутності фонового мережного трафіку. Для збору необхідних для аналізу даних було організовано лабораторну мережу з кількох комп'ютерів. Схематичне зображення використовуваної мережі наведено на рис. 1. Один із комп'ютерів було підімкнено до глобальної мережі Інтернет і на його базі організовано безпроводову точку доступу. На цьому самому комп'ютері здійснювалося захоплення всього трафіку, що проходить через нього, за допомогою програми Wireshark. На інших комп'ютерах, підімкнених до точки доступу було запущено різні додатки [2].

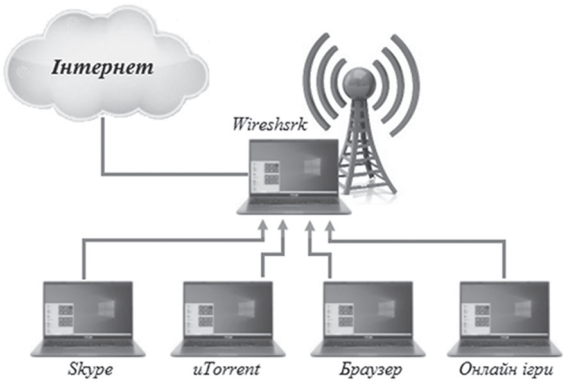

Рис. 1. Схема використовуваної мережі

Далі переглядалися веб-сторінки з використанням браузерів Google Chrome і Opera, через програму Skype проводилися відеодзвінки, здійснювалося скачування файлів за допомогою торрент клієнта µTorrent, використання сервісу цифрового поширення комп'ютерних ігор Steam тощо. Отримані дані зберігалися в форматі PCAP.

Для приведення здобутих даних у відповідність до вимог розв'язуваного завдання було проведено попереднє оброблення даних. Для цього всі пакети було розділено на потоки транспортного рівня та ідентифіковано за п'ятьма значеннями: протокол транспортного рівня (TCP або UDP), IP-адреса джерела, порт джерела, IP-адреса одержувача, порт одержувача. Отримані дані позначалися, тобто кожному потоку було поставлено у відповідність протокол/додаток, до якого цей потік належить [3].

Остаточний набір даних було поділено на дві підвибірки — навчальну і тестову. Склад здобутого датасета наведено в табл. 1.

За допомогою вбудованого в RF алгоритму відбору інформаційних ознак Feature Importance було відібрано такі 11 ознак (табл. 2).

Таблиця 1

Склад отриманої вибірки даних

Протокол	Навчальна вибірка	Тестова вибірка
SSL	1215	295
HTTP	1091	272
DNS	1061	267
BitTorrent	940	232
Steam	775	204
Skype	645	162

Результати класифікації. В експерименті було проведено побудову випадкового лісу і оцінювання якості класифікації на заданій вибірці. Дослідним шляхом відбиралися найбільш прийнятні параметри алгоритму. Так, експериментально вибрано, що ліс складається з п'яти дерев із максимально можливою глибиною.

Матрицю помилок для чистої тестової вибірки наведено в табл. 3. За вертикаллю зазначено реальні значення, за горизонталлю — передбачені навченою моделлю.

Таблиця 3

Матриця помилок для тестової вибірки

Реальні значення	Передбачені значення					
	SSL	HTTP	DNS	BitTorrent	Steam	Skype
SSL	295	0	0	0	0	0
HTTP	0	267	0	4	1	0
DNS	0	0	266	1	0	0
BitTorrent	1	0	0	230	0	1
Steam	0	3	0	0	201	0
Skype	6	0	0	1	0	155

Для визначення ефективності алгоритму використовуються такі метрики: точність, повнота і F-міра, значення яких легко обчислити на підставі матриці помилок класифікації, яка розраховується для кожного класу окремо.

Графічне подання даних метрик, отриманих експериментально, для всіх класів ілюструє рис. 2.

Таблиця 2

Відібрані інформаційні ознаки

Номер з/п	Назва	Опис
1	src_port	Номер порта джерела (джерелом вважається відправник першого пакета)
2	dst_port	Номер порта отримувача
3	max_src-data_port	Максимальний розмір даних у пакеті від джерела
4	min_src_data_ip	Мінімальний розмір даних у пакеті від джерела
5	med_src_data_ip	Медіанний розмір даних у пакеті від джерела
6	prop_src_data_ip	Частка даних, переданих джерелом, у загальній кількості даних потоку
7	max_dst_data_ip	Максимальний розмір даних у пакеті від отримувача
8	mean_dst_data_ip	Середній розмір даних у пакеті від отримувача
9	src_to_dst_ratio_data_ip	Відношення розміру даних, переданих джерелом, до розміру даних, які передані отримувачем
10	min_data_ip	Мінімальне значення розміру даних у потоці
11	var_data_ip	Середньоквадратичне відхилення значення розміру даних у потоці

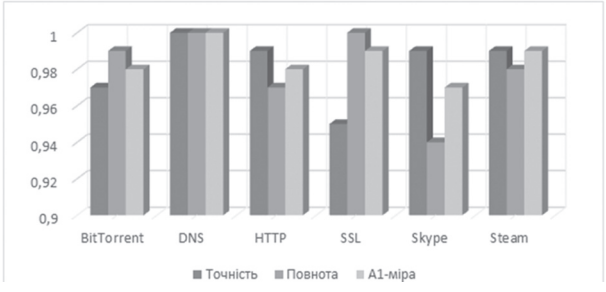

Рис. 2. Точність, повнота, F-міра для тестової вибірки

Як випливає з рис. 2, найбільшу ефективність алгоритм має для даних, що належать до DNS трафіку. Крім перевірки роботи алгоритму на тестовій вибірці, що має такий самий класовий склад, як і навчальна, оцінювання його якості відбувалося також за наявності фонового трафіку, тобто в разі, коли в тестовій вибірці були екземпляри класів, відсутніх у навчальній вибірці.

Наведемо склад тестової вибірки даних із домішками:

Протокол	SSL	HTTP	DNS	BitTorrent	Steam	Skype	LLMNR	Quic	RTP
Якість протоколу	295	272	267	232	204	162	169	95	19

Ситуація, коли в класифікованих даних наявний фоновий трафік, є більш наближеною до дійсності через різноманіття використовуваних у мережі Інтернет протоколів. Такий датасет дає можливість оцінити роботу алгоритму в реальних умовах. Матрицю помилок для цього випадку наведено в табл. 4.

Отже, табл. 4 унаочнює, що всі екземпляри, які належать до класу LLMNR, модель класифікувала як SSL, усі екземпляри RTP було віднесено до Skype, а екземпляри класу Quic здебільшого було поділено між класами DNS і Skype [4; 5].

Таблиця 4

Матриця помилок для тестової вибірки з фоновим трафіком

Реальні значення	Передбачені значення								
	SSL	HTTP	DNS	BitTorrent	Steam	Skype	LLMNR	Quic	RTP
SSL	295	0	0	0	0	0	0	0	0
HTTP	0	267	0	4	1	0	0	0	0
DNS	0	0	266	1	0	0	0	0	0
BitTorrent	1	0	0	230	0	1	0	0	0
Steam	0	3	0	0	201	0	0	0	0
Skype	6	0	0	1	0	155	0	0	0
LLMNR	169	0	0	0	0	0	0	0	0
Quic	0	0	25	9	0	61	0	0	0
RTP	0	0	0	0	0	19	0	0	0

Розглянемо, як змінилися показники якості класифікації (рис. 3).

Як бачимо, наявність фонового трафіку практично не вплинула на значення повноти, але значно погіршила значення точності класифікації, оскільки збільшилася кількість False Positive примірників, поява яких спричинена наявністю фонового трафіку, що належить до класів, які в навчанні не брали участь.

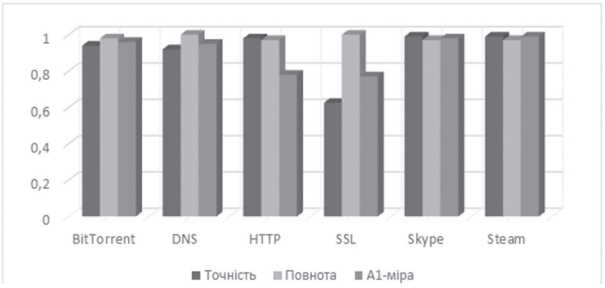

Рис. 3. Точність, повнота, F-міра за наявності фонового трафіку

Висновки

Аналіз показав, що алгоритм RF продемонстрував високу ефективність у режимі off-line, про що, зокрема, свідчить F-міра, яка становить відповідно 0,987 і 0,759 за відсутності і наявності фонового трафіку.

Наявність фонового трафіку належить до класів, що не брали участь у навчанні алгоритму, значно погіршуючи точність класифікації.

Таким чином, алгоритм RF мало придатний для класифікації в режимі реального часу через часову складність оброблення, що оцінюється як $Mmn \log [(n)]$, де n — кількість примірників, m — кількість інформаційних ознак, а M — кількість дерев.

Список використаної літератури

1. ***Weyrich M., Ebert C.*** *Reference architectures for the internet of things // IEEE Software. 2018. Vol. 33, № 1. P. 112–116.*

2. ***Lightweight,*** *payload-based traffic classification: An experimental evaluation / F. Risso, M. Baldi, O. Morandi [et al.] // Proc. IEEE ICC, 2018. P. 5869–5875.*

3. ***Sen S., Spatscheck O., Wang D.*** *Accurate Scalable In-Network Identification of P2P Traffic Using Application Signatures // Proc. of the 13th international conference on World (WWW'04). New York, NY, USA, 2016. P. 512–521.*

4. ***ICAP*** *[Електронний ресурс]: [Інтернет-портал]. URL: https://tools.ietf.org/html/rfc3507 (дата звернення 20.10.2020). Internet Content Adaptation Protocol (ICAP)*

*5. **QUIC** [Електронний ресурс]: [Інтернет-портал]. URL: https://tools.ietf.org/html/draft-tsvwg-quic-protocol-00 (дата звернення 25.10.2020). QUIC: A UDP-Based Secure and Reliable Transport for HTTP/2 draft-tsvwg-quic-protocol-00*

А. П. Козыряцкий, В. В. Жебка, Л. А. Демина, Д. А. Тарасенко

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ИНТЕРНЕТ-ТРАФИКА

Исследована эффективность применения алгоритма машинного обучения для классификации интернет-трафика. Рассмотрен алгоритм RF, который действует путем построения множества решающих деревьев. Оценена эффективность работы алгоритма RF в задачах классификации приложений в условиях наличия и отсутствия фонового сетевого трафика. Для сбора необходимых для анализа данных была организована лабораторная сеть из нескольких компьютеров. Один из компьютеров был подключен к глобальной сети Интернет и на его базе была организована беспроводная точка доступа. На этом же компьютере осуществлялся захват всего трафика, проходящего через него с помощью программы Wireshark. На других компьютерах, подключенных к точке доступа, были запущены различные приложения. Осуществлялся просмотр веб-страниц с помощью браузеров Google Chrome и Opera, с помощью программы Skype проводились видеозвонки, выполнялось скачивание файлов с помощью торрент клиента μTorrent, использование сервиса цифрового распространения компьютерных игр Steam и т. д. Полученные данные хранились в формате PCAP. Для приведения полученных данных в соответствие с требованиями решаемой задачи, осуществлялась предварительная обработка данных. В эксперименте было проведено построение случайного леса и оценено качество классификации на заданной выборке. Опытным путем были подобраны наиболее приемлемые параметры алгоритма. Экспериментально выбрано, что лес состоит из пяти деревьев с максимально возможной глубиной. Наибольшую эффективность метод имеет для данных, относящихся к DNS трафику. Кроме проверки работы алгоритма на тестовой выборке, которая имеет такой же классовый состав, как и учебная, оценка его качества проводилась также в условиях присутствия фонового трафика, т. е. в случае, если в тестовой выборке присутствовали экземпляры классов, отсутствующих в обучающей выборке.

Ключевые слова: машинное обучение; интернет-трафик; алгоритм RF; программа Wireshark; эффективность; метрики.

A. Kozyriatskyi, V. Zhebka, L. Domina, D. Tarasenko

INVESTIGATION OF EFFICIENCY OF APPLICATION OF MACHINE LEARNING ALGORITHM FOR CLASSIFICATION OF INTERNET TRAFFIC

The article investigates the effectiveness of the machine learning algorithm for the classification of Internet traffic. The RF algorithm, which works by constructing many decision trees, is considered. The efficiency of the RF algorithm in the problems of application classification in the presence and absence of background network traffic is evaluated. A laboratory network of several computers was set up to collect the data needed for analysis. One of the computers was connected to the World Wide Web and a wireless access point was set up on its base. On the same computer, all the traffic passing through it was captured using Wireshark. Various applications were running on other computers connected to the access point. Web pages were viewed using Google Chrome and Opera browsers, using Skype, video calls were made, files were downloaded using the μTorrent torrent client, the Steam digital game distribution service was used, etc. The obtained data were stored in the PCAP format. To bring the obtained data in line with the requirements of the problem, the data was pre-processed. In the experiment, a random forest was constructed and the quality of classification on a given sample was assessed. The most acceptable parameters of the algorithm were selected experimentally. It is experimentally chosen that the forest consists of 5 trees with the maximum possible depth. The algorithm is most effective for data related to DNS traffic. In addition to checking the operation of the algorithm on the test sample, which has the same class composition as the training, the assessment of its quality was also carried out in the presence of background traffic, i.e. in the test sample there were copies of classes absent in the training sample.

Keywords: machine learning; Internet traffic; RF algorithm; Wireshark program; efficiency; metrics.