DOI: https://doi.org/10.15407/usim.2018.06.074
УДК 681.513.7

YA. M. ANTONYUK, researcher,
International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03680, Ukraine,
ant@noc.irtc.org.ua

T. N. OLEKSYUK, engineer,
International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03680, Ukraine,
prus@noc.irtc.org.ua

YA.O. KOVALENKO, engineer,
International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03680, Ukraine,
kovalenko@noc.irtc.org.ua

B.A. SHIYAK, researcher,
International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03680, Ukraine,
bosh@noc.irtc.org.ua

THE PRINCIPLES OF THE MACHINE LEARNING APPLICATION IN CLASSIFICATION OF NETWORK TRAFFIC

Approaches to classification of network computing traffic on the basis of division of DPI and methods of structural analysis are systematized. The illustration of one of methods of structural analysis is developed. The algorithm which is possible for implementing in vitro is given. Perspectives of use of the given systematization are planned.

Keywords: *DPI, analysis of network traffic, network safety, classification of network traffic, machine learning.*

Introduction

The problem of the network traffic classification is solved on nodes of regional providers distribution, the corporate network centers, campus control nodes. Historically this task is most relevant in the field of traffic management for increasing efficiency of the existing communication channels and quality of the provided services for ultimate users. However, nowadays, the relevance of this task considerably increased. One more trend stimulating development of the network traffic classification techniques is significant increase in a share of the encoded traffic that results in inapplicability of approaches based on the contents analysis.

Problem definition

In a general view the problem of the network traffic classification can be formulated as follows: receiving on an input of some network traffic characteristics with delivery at the exit of a class to which this type of traffic belongs [1].

Network traffic classification is the first step which helps to identify the different applications and protocols transferred on network. The second step is management of this traffic, its optimization and prioritization. After classification all packets become noted on belonging to a certain protocol or the application that allows network devices to apply maintenance policies (QoS), leaning on these tags and flags.

Solution methods

There are two main methods of traffic classification:

1. Classification based on data units (Payload-Based Classification). It is based on fields with data units, such as ports (Layer 4) of OSI (the sender and the receiver or both). This method is the most widespread, but does not work with the ciphered and tunnelled traffic.

2. Classification based on the statistical technique. It is based on the analysis of traffic behavior (time between packets, session time, etc.).

It is necessary to take into account that global hasty growthover of the transferrable traffic amount and carrying communication channels capacity is brought to the necessity of the algorithms search with the lowered calculable complication.

Universal approach to traffic classification is based on information in IP packet heading — as a rule, it is the IP address (Layer 3), the MAC address (Layer 2), the used protocol. This approach has limit possibilities, as information is taken only from IP-heading, the methods of Layer 4 is limited. In fact far not all applications use the standard ports.

The deep analysis of packets (DPI) allows to carry out more perfect classification. (deep packet inspection — the deep analysis of packets). DPI is a basis of the majority network attacks detection, the systems of ensuring security policies of corporate networks, shaping and blocking of the user traffic by telecom operators. Now at the market there are several vendors making the solutions DPI engaged in their integration — Procera, Allot, Sandvine, Cisco.

One of the classification approach options which is used in the Cisco [1] company is given in Fig. 1.

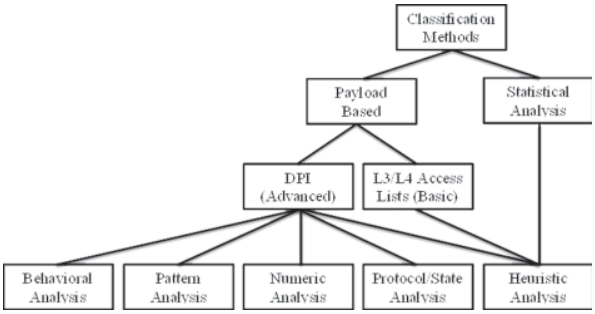

Fig. 1. Classification option illustrating general principles

The main lack of DPI is the mechanism of the analysis which needs to see the payload of the analyzed packets. Thus, DPI is not applicable if the client uses encryption or, for example, if we have no means of DPI at the time of traffic passing. If in the long term it is required to carry out some analysis of the traffic flowing on network — then there is a problem of saving of all payload for the subsequent analysis that does the general task is impracticable bulky.

Method for solving the classification problem based on an analysis of a set of statistical flow metrics

Thus, the alternative solution one of the main tasks of DPI is a definition of the application layer protocol — based on very small amount of information, without verification with the list of widely known ports (well-known ports) and without analysis of payload is considered here.

Also, as in the DPI means, as a rule, the object of classification is the traffic flow of the transport layer - this is a set of IP packets that have the transport layer protocol, as well as an unordered pair of endpoints: <(ip source , source port), (ip assignment, port of destination)>.

The idea of a method [6], is that the different applications using different protocols also generate flows of the transport layer with different statistical characteristics. At adequate definition of a set of statistical flow metrics, on values of these metrics it is possible to predict with a high accuracy what application generated this flow, and,

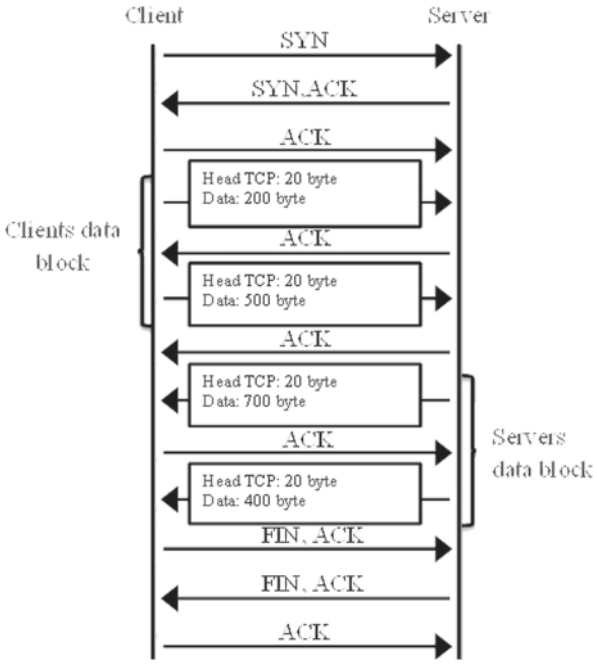

Fig. 2. An illustration of basic networking concepts

respectively, what application layer protocol is transferred by this flow.

For convenience we will define the basic concepts of the network interaction (Fig. 2):

▪ The client — the initiator of TCP-connection or the sender of the first UDP-datagram of a flow, depending on a transport layer protocol.

▪ The server — host TCP-connection or the addressee of the first UDP-datagram of a flow, depending on a transport layer protocol.

▪ A chunk of data is the application-level payloads collection that were transferred from one side to the other (from the client to the server or vice versa), without being interrupted by the payload from the other side.

In the set pattern after TCP handshake a client begins to pass an actual load - consequently, the aggregate of data begins from the client side. While a server sends no actual load in reply, and ACK sends just, the aggregate data proceeds from the client side. When a server feels a necessity to pass some application layer loading, the aggregate data of client closes, and the aggregate data of server begins. Thus, all transmission of actual load is alternation of aggregates data then with one, then on the other hand. With very intensive data exchange on both sides, data can be degenerated into separate IP packets.

We define the statistical flow metrics: all statistical characteristics of the flow will take values from four rows of numbers:

▪ The sequence of the sizes of the transport layer segments (TCP or UDP) sent from a client side.

▪ The sequence of the sizes of the transport layer segments sent from a server side.

▪ The sequence of the sizes of the data portions sent from a client side.

▪ The sequence of the sizes of the data portions sent from a server side.

For the short example shown in the figure above, this series will have the following meanings:

▪ Client side segment sizes: [220, 520]

▪ Server-side segment sizes: [720, 420]

▪ The size of the data portions on the client side: [700]

▪ The size of the chunks of data from the server: [1100]

These 4 rows of numbers fairly accurately characterize the data flow, and based on them we can determine the application layer protocol.

We formulate the statistical characteristics of the data stream, starting from these 4 rows of numbers:

1. Client-side average packet size.
2. Standard deviation of packet size on the client side.
3. Server-side average packet size.
4. Standard deviation server packet size.
5. Average client data size.
6. Standard deviation of the data portion from the client.
7. Average server chunk size.
8. Standard deviation of the data portion size on the server side.
9. Average number of packets per client data portion.
10. Average number of packets per server data portion.
11. Client Efficiency — the amount of application load transferred, superimposed on the total amount of application and transport load transferred.

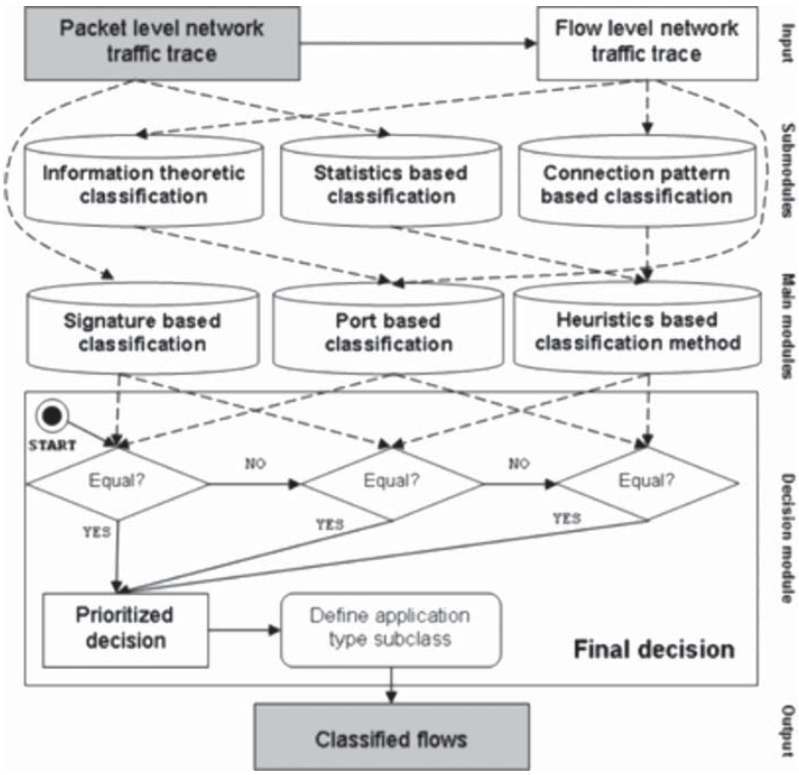

Fig. 3. Scheme of interaction between different components of traffic classification

12. Server efficiency.

13. Ratio of bytes — how many times the client transmitted more bytes than the server.

14. Payload Ratio — how many times the client transmitted more bytes than the server.

15. Package Ratio — how many times the client has transmitted more packets than the server.

16. The total number of bytes transferred by the client.

17. The total amount of application load transferred from the client.

18. The total number of transferred segments of the transport layer from the client.

19. The total number of transferred data pieces from the client.

20. Total number of bytes transferred by the server.

21. The total amount of application traffic transferred from the server side.

22. The total number of transmitted segments of the transport layer from the server.

23. The total amount of data transferred by the server.

24. The size of the first segment of the transport level from the client

25. Client's second transport segment size.

26. The size of the first segment of the transport layer on the server side.

27. The size of the second segment of the transport layer on the server side.

28. The size of the first data portion from the client

29. The size of the second data portion from the client

30. The size of the first data portion from the server

31. The size of the second data portion from the server

32. Protocol type of the transport layer (0 — UDP, 1 — TCP)

For definition of the application layer protocol of a specific flow, at these calculated statistical metrics we will lead this task to a problem of machine learning. The considered task is in the classical formulation a problem of the objects classification on several classes.

At each object 32 characteristics, and relevance each of them to the class tag which is available for an object at this investigation phase remains undecided. Therefore for machine learning it is offered and selected a popular algorithm "Random Forest" [7] as it is poorly sensitive to noise and correlation of signs.

This algorithm works on the principle of "learning with the teacher." This means that the algorithm requires a certain selection of objects, for which class labels are already known. This sample can be divided in a ratio of 1 to 2 for training and testing. On a training set, a training model is conducted (in our case, training consists in building a set of decision trees), and on a test sample, it evaluates how well the model copes with the task.

Conclusions

Based on the review of the traffic classification approaches, we can draw the following conclusions.

- There are a large number of algorithms and approaches with different advantages, disadvantages, processing speed, area of applicability and accuracy of the results obtained.
- Different algorithms comparison is considerably complicated due to the lack of public base of the full-fledged marked network routes on which it would be possible to carry it out. The lack of such base is caused by the objective reasons, such as need of information security support and privacy of net surfers. Available sets of routes, for example, in base of the CAIDA organization are «anonymized», they do not contain level data of the appendix in packets. It allows to apply the statistical approaches and approaches using headings 3 and 4 of levels, but excludes application of approaches on the basis of DPI.
- One of the directions, which are being developed nowadays, is application of different algorithms of machine learning, graph and statistical analysis, because of their applicability, including the encoded traffic (unlike DPI approaches) which share quickly grows. This direction, however, is also not saved from shortcomings. In particular, the algorithms accuracy can decrease if at the analyzed flow there is application traffic which were not used in training activity. The occasional retraining at characteristics change of the known protocols and emergence new is needed.
- Other developing direction is development of the combined approaches and the classification systems. One of the reasons for their development is the attempt of overcoming the shortcomings of the separate approaches (for example, the low accuracy or processing speed) and use of their advantages.
- Perspective of using the classification is application in a problem of prioritizing of traffic classes. It is offered to define the priorities of the traffic classes based on a multicriteria task solution of the theory of usefulness [4].

REFERENCES

1. Getman, A.I., Markin, Yu.V., Evstropov, E.F., Obydenkov, D.O., 2017. "Obzor zadach i metodov ikh resheniya v oblasti klassifikatsii setevogo trafika". Trudy ISP RAN, 29 (3), pp. 117—150. (In Russian). DOI: 10.15514/ISPRAS-2017-29(3)-8.
2. Yeon-sup Lim, Hyun-chul Kim, Jiwoong Jeong. Internet Traffic Classification Demystified: On the Sources of the Discriminative Power. [online] Available at: <http://conferences.sigcomm.org/co-next/2010/CoNEXT_papers/09-Lim.pdf> [Accessed 11 Oct. 2018].
3. Kuzmin, V.V., 2014. Klassifikatsiya i identifikatsiya trafika v multiservisnoy seti operatora svyazi. Sovremennyye problemy nauki i obrazovaniya, 5. [online] Available at: <https://www.science-education.ru/ru/article/view?id=15039> [Accessed 11 Oct. 2018].
4. Shelukhin O.I., Erokhin S.D., Vanyushina A.V. Klassifikatsiya IP-trafika metodami mashinnogo obucheniya Izdatelstvo: M.: Goryachaya liniya — Telekom 2018 stranits: 283
5. Mashinnoye obucheniye vmesto DPI. Stroim klassifikator trafika. [online] Available at: <https://habr.com/post/304926/> [Accessed 11 Oct. 2018].
6. Random_forest. [online] Available at: <https://ru.wikipedia.org/wiki/Random_forest1> [Accessed 11 Oct. 2018].

Received 04.12.18

Я.М. Антонюк, наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій і систем
НАН і МОН України, просп. Глушкова, 40, Київ 03187, Україна
ant@noc.irtc.org.ua

Я.А. Коваленко, пров. інж.-прогр.,
Міжнародний науково-навчальний центр інформаційних технологій і систем
НАН і МОН України, просп. Глушкова, 40, Київ 03187, Україна,
kovalenko@noc.irtc.org.ua

Т.Н. Олексюк, інж.-прогр.,
Міжнародний науково-навчальний центр інформаційних технологій і систем
НАН і МОН України, просп. Глушкова, 40, Київ 03187, Україна,
prus@noc.irtc.org.ua

Б.А. Шияк, мол. наук. співроб.,
Міжнародний науково-навчальний центр інформаційних технологій і систем
НАН і МОН України, просп. Глушкова, 40, Київ 03187, Україна,
bosh@noc.irtc.org.ua

ПРИНЦИПИ ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ В КЛАСИФІКАЦІЇ МЕРЕЖЕВОГО ТРАФІКУ

Вступ. Завдання класифікації мережевого трафіку вирішується на вузлах розподілу регіональних провайдерів, корпоративних мережевих центрах, кампусних вузлах управління. Історично ця задача найбільш актуальна в галузі управління трафіком для підвищення ефективності використання існуючих каналів зв'язку і якості послуг, що надаються для кінцевих користувачів.

Мета. Метою дослідження є розробка підходу до вирішення у загальному вигляді задачі класифікації мережевого трафіку, а саме, отримання на вхід деяких характеристик мережевого трафіку з видачею на виході класу, до якого даний вид трафіку відноситься.

Методи рішення. Розглянуто два основні методи класифікації трафіку:

1. Класифікація на основі блоків даних (*Payload-Based Classification*). що грунтується на аналізі полів з блоками даних, таких як порти (*Layer* 4) *OSI* (відправник і одержувач чи обидва). Даний метод є найбільш поширеним, але не працює з зашифрованим і тунельованним трафіком.

2.Классификация на основі статистичного методу. Грунтується на аналізі поведінки трафіку (час між пакетами, час сеансу і т. п.) та аналізі службових полів .

Результати. Розроблено рекомендації щодо застосування методу рішення задачі класифікації на основі аналізу набору статистичних метрик потоку. Розглянуто альтернативний спосіб вирішення однієї з головних завдань *DPI* — визначення протоколу прикладного рівня — на основі дуже невеликої кількості інформації, без звірки зі списком широко відомих портів (*well-known ports*) і без аналізу корисного навантаження.

Власне, для машинного навчання запропоновано і вибрано популярний алгоритм «*Random Forest*», оскільки він слабо чутливий до шумів і кореляції ознак.

Висновки. На підставі огляду підходів до класифікації трафіку робляться висновки з існування великої кількості алгоритмів і підходів з різними перевагами, недоліками, що відрізняються за швидкістю обробки, області застосування і точності результатів, порівняння яких значно ускладнено через відсутність загальнодоступної бази повноцінних розмічених мережевих трас, на яких було б можливо проводити порівняння.

Напрямком, що розвивається є комбінування підходів і систем класифікації в ході спроб подолання недоліків окремих підходів і використання їх переваг.

Перспективою використання рішення задачі класифікації є застосування в задачі пріоритетності класів трафіку. Запропоновано визначати пріоритети класів трафіків на основі рішення багатокритеріальної задачі теорії корисності.

Ключові слова: аналіз мережевого трафіку, мережева безпека, класифікація мережевого трафіку, машинне навчання, *DPI*

Я.М. Антонюк, науч. сотр.,
Международный научно-учебный центр информационных технологий и систем
НАН и МОНУкраины, просп. Глушкова, 40, Киев 03187, Украина,
ant@noc.irtc.org.ua

Я.А. Коваленко, ведущ. инж.-прогр.,
Международный научно-учебный центр информационных технологий и систем
НАН и МОНУкраины, просп. Глушкова, 40, Киев 03187, Украина,
kovalenko@noc.irtc.org.ua

Т.Н. Олексюк, инж.-прогр.,
Международный научно-учебный центр информационных технологий и систем
НАН и МОНУкраины, просп. Глушкова, 40, Киев 03187, Украина,
prus@noc.irtc.org.ua

Б.А. Шияк, мл. науч. сотр.,
Международный научно-учебный центр информационных технологий и систем
НАН и МОНУкраины, просп. Глушкова, 40, Киев 03187, Украина,
bosh@noc.irtc.org.ua

ПРИНЦИПЫ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ В КЛАССИФИКАЦИИ СЕТЕВОГО ТРАФИКА

Вступление. Задача классификации сетевого трафика решается на узлах распределения региональных провайдеров, корпоративных сетевых центрах, кампусных узлах управления. Исторически эта задача наиболее актуальна в области управления трафиком для повышения эффективности использования существующих каналов связи и качества предоставляемых услуг для конечных пользователей.

Цель. Целью исследования является разработка подхода к решению в общем виде задачи классификации сетевого трафика, а именно, получение на вход некоторых характеристик сетевого трафика с выдачей на выходе класса, к которому данный вид трафика относится.

Методы решения. Рассмотрены два основных метода классификации трафика:

1. Классификация на основе блоков данных (*Payload-Based Classification*). Основывается на полях с блоками данных, таких как порты (*Layer* 4) *OSI* (отправитель и получатель или оба). Данный метод является наиболее распространенным, но не работает с зашифрованным и туннелированным трафиком.

2.Классификация на основе статистического метода. Основывается на анализе поведения трафика (время между пакетами, время сеанса и т. п.).

Результаты. Разработаны рекомендации по применению метода решения задачи классификации на основе анализа набора статистических метрик потока. Рассмотрен альтернативный способ решения одной из главных задач *DPI* — определение протокола прикладного уровня — на основе ограниченного количества информации, без сверки со списком известных портов (*well-known ports*) и без анализа полезной нагрузки.

Собственно, для машинного обучения предложено и выбрано популярный алгоритм «*Random Forest*», поскольку он слабо чувствителен к шумам и корреляции признаков.

Выводы. На основании осмотра подходов к классификации трафика делаются выводы о существовании большого количества алгоритмов и подходов с различными преимуществами, недостатками, отличающиеся по скорости обработки, области применения и точности результатов, сравнение которых значительно затруднено из-за отсутствия общедоступной базы полноценных размеченных сетевых трасс, на которых было бы возможно проводить сравнения.

Развивающимся направлением является разработка комбинированных подходов и систем классификации в ходе попыток преодоления недостатков отдельных подходов и использование их преимуществ.

Перспективой использования решения задачи классификации является применение в задаче приоритезации классов трафика. Предложено определять приоритеты классов трафиков на основе решения многокритериальной задачи теории полезности.

Ключевые слова: анализ сетевого трафика, сетевая безопасность, классификация сетевого трафика, машинное обучение, *DPI*