О методах идентификации сетевых трафиков

Рамиз Шыхалиев

Институт Информационных Технологий НАНА, Баку, Азербайджан

ramiz@science.az

***Аннотация* — Точная идентификация сетевых трафиков является очень важным элементом управления и обеспечения безопасности сетей. В литературе имеются множество методов для идентификации сетевых трафиков и в зависимости от типов используемых для идентификации информации точность и полнота этих методов различаются. В статье проводится анализ как традиционных методов идентификации сетевых трафиков, так и методов основанных на машинном обучению.**

Ключевые слова — сетевые трафики, идентификации сетевых трафиков, классификация сетевых трафиков, методы машинного обучения

I. ВВЕДЕНИЕ

Сегодня Интернет развивается очень быстрыми темпами и становится очень масштабным, динамичным, скоростым и мобильным. В Интернете появились множество, так называемых «peer-to-peer (P2P)» совместно используемых приложений, социальных сетей, потоковых видеосервисов, сервисов мгновенных сообщений, онлайн-игр и т.д. Это привело к значительному увеличению количества пользователей и изменению их поведения. В результате, значительно увеличился объем интернет-трафика и изменился его характер. Вместе с тем, в Интернете используются множество различных типов протоколов. Кроме того, сетевые приложения имеют различные функциональные требования, и большинство этих приложений используют номера портов TCP или UDP, которые назначены IANA (Internet Assigned Numbers Authority) [1]. IANA для конкретных сетевых приложений, протоколов и сервисов назначил конкретные номера портов, которые меняются в интервале от 0 до 1023, а также IANA зарегистрированы номера портов, которые меняются в интервале от 1024 до 49151. Однако у большинства приложений нет номеров портов назначенных IANA, но используются номера портов, выбираемые по умолчанию, и часто эти номера совпадают с номерами портов IANA. Поэтому часто невозможно однозначно идентифицировать сетевые приложения с известными или зарегистрированными портами. Следовательно, в таких условиях очень трудно обеспечить требуемый уровень производительности и безопасности сетей, а также QoS (англ. Quality of Service) для приложений, сервисов и т.д.

Вместе с тем, исследования показали, что сетевой трафик представляет собой сложный динамический процесс и является суперпозицией многих потоков с множественными взаимосвязанными характеристиками, которые генерируются различными протоколами. Во-первых, это трафики, связанные с управлением самой сети (например, трафик инициализации клиентов, серверный трафик и т.д.), которые генерируются периодически. Во-вторых, это трафики сетевых сервисов, приложений (например, DNS, FTP, запросы WINS, ARP, сеанс NetBIOS, HTTP, P2P, SMTP, POP3, Telnet и т.д.) и протоколов, которые составляют основную часть сетевого трафика [2].

Исходя из вышесказаных, для обеспечения нормальной и безопасной работы сетей требуются эффективные методы мониторинга, анализа и оценки работы сетей. Для этого, прежде всего необходима точная идентификация сетевых трафиков, что является очень сложной задачей и требуется разработка адекватных методов идентификации сетевых трафиков.

Целью данной статьи является анализ методов идентификации сетевых трафиков имеющиеся в литературе, чтобы оценить их возможности по идентификации сетевых трафиков.

II. ТРАДИЦИОННЫЕ МЕТОДЫ ИДЕНТИФИКАЦИИ СЕТЕВЫХ ТРАФИКОВ

Существующие методы идентификации сетевых трафиков примерно делятся на пять категорий: методы идентификации на основе портов; методы идентификации на основе глубокой инспекции пакетов (Deep packet inspection – DPI), то есть анализ содержимого пакетов; методы идентификации основанные на анализе характеристик сетевых потоков; методы идентификации основанные на анализе поведения хостов; методы идентификации на основе машинного обучения.

Традиционно, для идентификации сетевых трафиков применялись простые методы основанные на анализе характеристик сетевых трафиков. В качестве этих характеристик использовались характеристики пакетов, такие как номера портов, IP-адреса отправителей и получателей, типы приложений и протоколов, а также содержимое пакетов, статистические характеристики трафика и т.д. Некоторые из этих методов рассмотрены в [3, 4]. Однако сегодня идентификация сетевых трафиков на основе номеров портов является малоэффективной [5, 6]. Это, в основном, связано с появлением большего количества сетевых приложений и сервисов, использующих нестандартные TCP-порты, а также приложений, туннелирующих HTTP и широкое

DOI: 10.25045/NCInfoSec.2017.24

использование в Интернете P2P приложений. В результате некоторые приложения не могут быть идентифицированы вовсе. Выходом из этой ситуации могут быть анализ содержимого пакетов и создание для каждого приложения сигнатуры, но при этом появляются как минимум две проблемы: юридическая, которая связана с частной жизнью пользователя, и невозможность идентификации зашифрованных сетевых трафиков.

Идея использования статистических характеристик сетевых трафиков для их идентификации или для описания их свойств не новая. В работах [7, 8] впервые рассматривались вопросы по определению характеристик интернет-трафика и в основном определялась взаимосвязь между характеристиками потоков и прикладными протоколами, генерирующими их. Эти работы показывают, что аналитические модели случайных переменных могут быть использованы для описания свойств нескольких протоколов.

Несмотря на то, что идентификация сетевых трафиков является довольно определенной областью исследования, цели имеющихся в этой области работ не идентичны. Целью некоторых работ является только идентификация P2P трафика, целью других – детальная классификация сетевого трафика, то есть точная идентификация приложения, генерирующего конкретный трафик. К тому же с появлением новых сетевых приложений может изменяться характер существующих сетевых характеристик и для идентификации сетевых трафиков могут использоваться иные идентификационные характеристики. Например, появление некоторых новых приложений, таких как BitTorrent, PPStream, PPLive и т.д., привело к широкому использованию протокола UDP.

В работах [9, 10] были предложены методы идентификации сетевых трафиков с детальным анализом содержимого пакетов. Главным недостатком этих методов является то, что они требуют очень больших вычислительных ресурсов. В то же время точность идентификации сетевых трафиков в основном зависит от моделей, построенных на основе выявленных закономерностей и отражающих основные особенности сетевого трафика. Однако, несмотря на достаточно высокую точность идентификации, полученную в работе [10], для обучения наивного алгоритма Байеса в качестве входных данных были использованы трафики, классифицированные вручную.

В работе [11] предлагается метод классификации сетевых трафиков, основанный на статистическом анализе активности хостов. При этом не анализируется содержимое пакетов, и для классификации сетевого трафика шаблоны поведения хостов сопоставляются с одним или несколькими приложениями.

Исследование недостатков методов идентификации сетевых трафиков, основанных на анализе номеров портов и содержимого пакетов, показало, что для идентификации сетевого трафика а более подходящими являются методы машинного обучения (МО) [12].

III. Идентификация сетевых трафиков на основе методов машинного обучения

При идентификаци сетевых трафиков, одним из важных областей исследования является классификация. Цель классификации состоит в построении классификационных моделей для прогнозирования неизвесного образца на основании изучения набора обучающих данных.

В последнее десятилетие существенная часть работ по идентификации сетевых трафиков была основана на их классификации с применением методов МО. Эти работы могут быть классифицированы как работы, использующие методы МО с учителем, без учителя и так называемые полуобучаемые (гибридные) методы.

В методах классификации сетевых трафиков с использованием методов МО с учителем анализируются обучающие данные и выводится предполагаемая функция, которая может прогнозировать выходные классы из любого тестового потока. При этом очень важным является выбор достаточно обоснованных обучающих данных. К методам МО с учителем относятся следующие: дерева решений (англ. Decision Trees – DT); наивный Байесовский классификатор (англ. Naive Bayes Classification – NBC); обычная регрессия наименьших квадратов (англ. Ordinary Least Squares Regression – OLSR); логическая регрессия (англ. Logistic Regression – LR); метод опорных векторов (англ. Support Vector Machine – SVM) т.д.

С помощью методов классификации сетевых трафиков с использованием алгоритмов МО без учителя (то есть алгоритмов кластеризации) в немаркированных данных трафика находят кластеры и определяют вхождение данных в те или иные кластеры. К методам МО без учителя относятся следующие: алгоритмы кластеризации (англ. Clustering Algorithms); анализ главных компонент (аргл. Principal Component Analysis – PCA); независимый компонентный анализ (Independent Component Analysis); сингуля́рное разложе́ние (англ. Singular Value Decomposition – SVD); «случайный лес» (англ. Random Forest – RF); самоорганизующаяся карта Коохонена (англ. Self-organizing map – SOM) и т.д.

В работе [13] авторами были оценены алгоритмы с учителем, в том числе наивный Байесовский алгоритм с дискретизацией, наивный Байесовский алгоритм с оценкой ядра плотности, дерева решений C4.5, сети и дерева Байеса. В работе [14] авторы предложили подход к классификации трафика на основе анализа потока пакетов в режиме реального времени. В работе [15] для точной классификации трафика применены Байесовские нейронные сети. В [16] для классификации трафика авторами используются однонаправленные статистические функции. В работе [17] для компактного выражения трех статистических характеристик трафика авторами была использована функция плотности вероятности. В работе [18] для классификации трафика авторы предложили использовать одноклассные SVM (one

class support vector machines) и для каждого набора рабочих параметров SVM был предложен простой алгоритм оптимизации.

Все эти работы используют параметрические алгоритмы МО, которые для параметров классификатора требуют процедуры интенсивного обучения и нуждаются в повторном обучении при обнаружении новых приложений.

Имеется несколько работ, основанных на непараметрических алгоритмах МО. В работе [19] для классификации трафика авторами были использованы методы ближайших соседей и линейного дискриминантного анализа, при этом для классификации использовали пять статистических характеристик. В работе [20] для классификации трафика предлагается так называемый BLINC-метод, который использует поведение хостов. Несмотря на то, что непараметрические методы имеют некоторые преимущества, нежели параметрические, они не так широко используются для классификации трафика.

В работе [21] авторами было предложено с помощью EM-алгоритма (Expectation maximization (EM) algorithm) группировать потоки трафиков в небольшом количестве кластеров, причем каждый кластер маркируется вручную. В работе [22] для кластеризации потока трафика был использован алгоритм AutoClass, и для оценки кластеров была предложена метрика внутриклассовой однородности. В работе [23] для кластеризации трафика был использован алгоритм к-средних и с помощью анализа полезной информации были промаркированы кластеры для приложений. В работе [24] для кластеризации трафика на основе двух наборов эмпирически собранных данных авторами были оценены к-средних, DBSCAN и AutoClass-алгоритмы.

В общем, эти методы кластеризации могут быть использованы для идентификации трафиков ранее неизвестных приложений. В работе [25] авторы предложили интегрировать кластеризацию, основанную на статистических характеристиках потока, с методом сравнения сигнатуры полезной информации, что исключает необходимость использования обучающих наборов данных. А в работе [26] авторы предложили комбинировать кластеризацию, основанную на статистических характеристиках потока, и кластеризацию, основанную на статистических характеристиках полезной информации для обнаружения неизвестного трафика.

Однако методы кластеризации имеют проблему отображения большого количества кластеров к реальным приложениям. Эта проблему очень трудно решить, если нет информации о реальных приложениях. Для решения этой проблемы в работе [27] предложен новый непараметрический подход, который заключается во включении корреляционной информации потоков в процесс классификации.

Полуобучаемые или гибридные методы МО классификации сетевого трафика используют как маркированные, так и немаркированные статистические характеристики потока [28]. Из-за такого подхода эти методы обеспечивают более точную и быструю классификацию трафика, а также позволяют идентифицировать неизвестные приложения и приложения с измененным поведением. В работе [29] авторами было предложено использовать набор обучающих данных в алгоритиме МО без учителя. Однако при малых размерах обучающих данных основную часть отображения составляют «неизвестные» кластеры.

В работе [30] для идентификации трафиков протоколов TCP и UDP автор предложил метод классификации основанные на использования метода опорных векторов (англ. Support Vector Machine – SVM). В этом подходе для выбора подмножества наилучших характеристик используется генетический алгоритм (англ. Genetic Algorithm), а для вычисления весов каждой характеристики используется метод роя частиц (англ. Particle Swarm Optimization – PSO). Вместе с этим, традиционный алгоритм SVM используется для классификации различных потоков трафика и оптимизируется с помощью алгоритма PSO, который может эффективно улучшить производительность алгоритма SVM. Предложенный подход позволяет классифицировать трафики Интернета на основе статистических характеристик потоков трафика без использования информации о порте или хосте и нет необходимости проверки сигнатуры приложений.

В работе [31] для идетификации сетевых трафиков авторы предложили гибридную модель, в которой используется алоритм Apriori для атоматической генерации ассоциативных правил и самоорганизующаяся карта Коохонена (англ. Self-organizing map – SOM). Предложенный подход позволяет идентицировать сетевые трафики без использования контента и номеров портов, а также генерировать ассоциативные правила для идентификации неизвестных приложений. При этом, алгоритм Apriori позволяет выбирать наиболее типичные правила для каждого типа трафика, в то время алгоритм основанный на SOM алгоритм позволяет группировать трафики схожых протоколов и приложений.

Автор в работе [32] предложил подход к идентификации P2P трафиков основанный на алгоритме случайного леса (англ. Random forest algorithm). Алгоритм случайного леса это комбинация деревьев решений. Построение случайного леса позволяет повысить точность и эффективность идентификации P2P трафиков.

Заключение

Исходя из проведенного в статье анализа методов идентификации сетевых трафиков, можно сказать, что точность методов различается в зависимости от типов используемых для классификации информации и методов. Поэтому для обеспечения точности и полноты идентификации сетевых трафиков необходимо создать комплексный механизм.

Для этого лучшим решением является комбинирование существующих механизмов классификации с

использованием МО с учителем и без него, а также использование ансамбля классификаторов. Это позволить намного повысить точность и полноту идентификации сетевых трафиков.

ЛИТЕРАТУРА

[1] IANA, http://www.iana.org/assignments/port-numbers (August 2005).

[2] Р.Г. Шыхалиев Анализ и классификация сетевого трафика компьютерных сетей, Проблемы Информационных Технологий, №2, с. 15-23, 2010.

[3] P. Gupta and N.McKeown, Algorithms for packet classification, IEEE Network Magazine. vol.15, no.2, pp. 24-32, 2001.

[4] M.L. Bailey, B. Gopal, M.A. Pagels, L.L. Peterson, and P. Sarkar, PathFinder: A pattern-based packet classifier, Proceedings of the First Symposium on Operating Systems Design and Implementation, pp. 115-123, 1994.

[5] C. Logg and L. Cottrell Characterization of the Traffic between SLAC and the Internet, July 2003. http://www.slac.stanford.edu/comp/net/slac-netflow/html/SLAC-netflow.html.

[6] W. Moore and D. Papagiannaki Toward the Accurate Identification of Network Applications, In Proceedings of the Sixth Passive and Active Measurement Workshop, pp. 41-54, 2005.

[7] V. Paxson Empirically derived analytic models of wide-area TCP connections, IEEE/ACM Trans. Netw., vol.2, no.4, pp. 316-336, 1994.

[8] V. Paxson and S. Floyd Wide area traffic: the failure of Poisson modeling, IEEE/ACM Trans. Netw., vol.3, no.3, pp. 226-244, 1995.

[9] W. Li, M. Canini, A.W. Moore and R. Bolla Efficient application identification and the temporal and spatial stability of classification schema, Computer Networks, vol.53, no.6, pp. 790-809, 2009.

[10] A.W. Moore, D. Zuev Internet traffic classification using Bayesian analysis techniques, Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems, vol.33, no.1, pp. 50-60, 2005.

[11] T. Karagiannis, K. Papagiannaki, and M. Faloutsos BLINC: multilevel traffic classification in the dark, Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Pomputer Communications, pp. 229-240, 2005.

[12] N. J. Nilsson Introduction to Machine Learning http://robotics.stanford.edu/people/nilsson/MLDraftBook/MLBOOK.pdf

[13] N. Williams, S. Zander, G. Armitage A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification, ACM SIGCOMM Computer Communication Review, vol.36, no.5, pp. 5-16, 2006.

[14] T. Nguyen, G. Armitage Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks, Proceedings of the 31st IEEE Conference on Local Computer Networks, pp. 369-376, 2006.

[15] T. Auld, A.W. Moore, S.F. Gull Bayesian neural networks for internet traffic classification, IEEE Trans. Neural Networks, vol. 18, no. 1, pp. 223-239, 2007.

[16] J. Erman, A. Mahanti, M. Arlitt, C. Williamson Identifying and discriminating between web and peer-to-peer traffic in the network core, Proceedings of the 16th international conference on World Wide Web, pp. 883-892, 2007.

[17] M. Crotti, M. Dusi, F. Gringoli, L. Salgarelli Traffic classification through simple statistical fingerprinting, ACM SIGCOMM Computer Communication Review, vol.37, no.1, pp. 5-16, 2007.

[18] A. Este, F. Gringoli, L. Salgarelli Support vector machines for tcp traffic classification, Computer Networks, vol.53, no.14, pp. 2476-2490, 2009.

[19] M. Roughan, S. Sen, O. Spatscheck, N. Duffield Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification, Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pp. 135-148, 2004.

[20] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, Lee K. Internet traffic classification demystified: myths, caveats, and the best practices, Proceedings of the ACM CoNEXT Conference, pp. 1-12, 2008.

[21] A. McGregor, M. Hall, P. Lorier, J. Brunskill Flow clustering using machine learning techniques, Proceedings of Passive and Active Measurement Workshop, pp. 205-214, 2004.

[22] S. Zander, T. Nguyen, G. Armitage Automated traffic classification and application identification using machine learning, Annual IEEE Conference on Local Computer Networks, pp. 250-257, 2005.

[23] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, K. Salamatian Traffic classification on the fly, ACM SIGCOMM Computer Communication Review, vol.36, no.2, pp. 23-26, 2006.

[24] J. Erman, M. Arlitt, A. Mahanti Traffic classification using clustering algorithms, Proceedings of the SIGCOMM workshop on Mining network data, pp. 281-286, 2006.

[25] Y. Wang, Y. Xiang and S.-Z. Yu. An automatic application signature construction system for unknown traffic, Concurrency Computations: Pract. Exper., vol.22, pp. 1927-1944, 2010.

[26] A. Finamore , M. Mellia , M. Meo Mining unclassified traffic using automatic clustering techniques, TMA International Workshop on Traffic Monitoring and Analysis, pp. 150–163, 2011.

[27] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, Y. Guan Network traffic classification using correlation information, IEEE Transactions on Parallel and Distributed Systems, vol. 24, no.1, pp. 1-15, 2012.

[28] C. Gu1, S. Zhang, X. Chen, A. Du Realtime traffic classification based on semi-supervised learning, Journal of Computational Information Systems, no.7, pp. 2347-2355, 2011.

[29] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, C. Williamson Offline/realtime traffic classification using semi-supervised learning, Performance Evaluation, vol.64, no.9-12, pp.1194-1213, 2007.

[30] J. Tan An Internet Traffic Identification Approach Based on GA and PSO-SVM, Journal of computers, vol. 7, no. 1, pp. 19-29, 2012.

[31] Z. Nascimento, D. Sadok, S. Fernandes A Hybrid Model for Network Traffic Identification Based on Association Rules and Self-Organizing Maps (SOM), The Ninth International Conference on Networking and Services, pp. 213-219, 2013.

[32] H. Yajun P2P Network Traffic Identification Based on Random Forest Algorithm, Journal of networks, vol. 9, no. 9, pp. 2456-2461, 2014.

ABOUT NETWORK TRAFFIC IDENTIFICATION METHODS

Ramiz Shikhaliyev
Institute Information Technology ANAS, Baku, Azerbaijan
ramiz@science.az

Abstract – Accurate identification of network traffic is a very important element of network management and security. In the literature, there are many methods for identifying network traffic and, depending on the types of information used to identify information, the accuracy and completeness of these methods differ. To increase the accuracy and completeness of the identification of network traffic, in recent years more and more studies have been devoted to the creation of effective identification methods based on machine learning. The article analyzes both traditional methods for identifying network traffic and methods based on machine learning.

Keywords – network traffic, network traffic identification, network traffic classification, machine learning methods