УДК 004.048

Шыхалиев Р.Г.
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

АНАЛИЗ И КЛАССИФИКАЦИЯ СЕТЕВОГО ТРАФИКА КОМПЬЮТЕРНЫХ СЕТЕЙ

Данная статья посвящена анализу и моделированию классификации сетевого трафика компьютерных сетей (КС), которые очень важны при их мониторинге. Для моделирования классификации сетевого трафика предложено использование методов машинного обучения без учителя. В качестве метода обучения без учителя использован алгоритм кластеризации k-средних.

Ключевые слова: *сетевой трафик, кластеризация, алгоритм k-средних.*

Введение

Использование в современных КС большего количества сетевых сервисов и приложений, аппаратного и программного обеспечения приводит к появлению в сети большого разнообразия трафиков. При этом для проведения эффективного мониторинга и управления КС решение задачи точной идентификации и классификации трафиков относительно сетевых сервисов, приложений и протоколов является очень важной. Потому, что сетевой трафик является одним из важнейших фактических показателей работы КС. Сетевой трафик является носителем информации о поведении пользователей и функционировании КС. На основе статистического анализа сетевого трафика можно косвенно определить статистические характеристики поведения КС.

Идентификация и классификация сетевого трафика особенно важна для решения таких задач, как определение приоритетов при формировании полосы пропускания для отдельных трафиков, установление правил по управлению сети, обеспечение безопасности сети, диагностический мониторинг КС и т.д. [1]. Например, для того, чтобы обеспечить нормальную работу приложений, важных для корпорации, администратор сети должен идентифицировать и ограничивать (или заблокировать) P2P (peer-to-peer) трафик. Кроме того, эффективное решение большинства технических задач, таких как определение параметров и моделирование рабочей нагрузки каналов связи, планирование загрузки сетевых оборудований, инициализация маршрутов и т.д., также зависит от точной идентификации и классификации сетевого трафика.

Прежде чем классифицировать сетевой трафик, очень важно определить их классификационные характеристики. Эти характеристики могут быть определены в результате анализа свойств, описывающих сетевой трафик, к которым могут относиться различные особенности общего сетевого трафика КС [2, 3]. Классификацию сетевого трафика можно определить как анализ трафиков, созданных различными сетевыми приложениями. Другими словами, цель классификации состоит в определении того, какие типы трафика передаются по КС [4, 5].

Для классификации сетевого трафика обычно применялись простые методы, основанные на анализе информации, характеризующей пакеты (номера портов, IP-

адреса отправителей и получателей, типы приложений и протоколов и т.д.). Некоторые из этих методов рассмотрены в [6, 7]. Однако сегодня классификация сетевого трафика на основе номеров портов является малоэффективной [8, 9]. Это, в основном, связано с появлением большего количества сетевых приложений и сервисов, использующих нестандартные TCP-порты, а также приложений, туннелирующих HTTP и широкое 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 трафиков.

Несмотря на то, что классификация сетевого трафика является довольно определенной областью исследования, цели имеющихся в этой области работ не идентичны. Целью некоторых работ является только классификация 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 классификации сетевого трафика могут использоваться иные классификационные характеристики. Например, появление некоторых новых приложений, таких как BitTorrent, PPStream, PPLive и т.д., привело к широкому использованию протокола UDP.

В работах [10, 11] были предложены методы классификации сетевого трафика с детальным анализом содержимого пакетов. Главным недостатком этих методов является то, что они требуют очень больших вычислительных ресурсов. В то же время точность классификации сетевого трафика в основном зависит от моделей, построенных на основе выявленных закономерностей и отражающих основные особенности сетевого трафика. Однако, несмотря на достаточно высокую точность классификации, полученную в работе [11], для обучения наивного алгоритма Байеса в качестве входных данных были использованы трафики, классифицированные вручную.

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

Данная статья посвящена анализу и моделированию классификации сетевого трафика КС на основе МО, а именно метода обучения без учителя – методу кластерного анализа, а в качестве алгоритма кластеризации предлагается использовать алгоритм k-средних.

Анализ сетевого трафика

Исследования сетевого трафика показали, что он 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 КС (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 КС.

Как известно, IP-протокол является универсальным протоколом для любого типа приложений, используемых в КС, и вся нагрузка по транспорту трафиков ложится на него. Также известно, что основными транспортными протоколами, которые работают на IP, являются TCP и UDP, и в основном сетевой трафик КС состоит из TCP-трафиков, что является основной особенностью IP-трафика. Согласно работе [13], приблизительно 98,2% IP-трафика содержит информацию о TCP-протоколе. Вместе с тем широкое использование в КС мультимедийных и интерактивных приложений, а также новых расширенных протоколов, таких как RTP и RSVP, приводит к росту объема UDP-трафика.

В литературе имеется обширное исследование по моделированию IP-трафика КС. В работах [14, 15] рассматриваются ключевые вопросы определения характеристик сетевого трафика и схемы сбора образцов трафиков различных приложений. Из-за неоднородного характера IP-трафика при определении характеристик необходима его детализация, так как трафики различных приложений, например, HTTP, FTP, IP-voice, DNS, NTP и т.д., имеют отличительные характеристики. Поэтому при классификации сетевого трафика важным являются определение ключевых особенностей трафика различных приложений и их интеграция.

Трудность моделирования IP-трафика связана с тем, что в IP-сетях используются многоуровневый стек протоколов TCP/IP и множество приложений, и для каждого уровня имеется свое предназначение, поэтому сетевой трафик на каждом уровне имеет различные характеристики. Также известно, что характеристики TCP-трафика очень отличаются от характеристик UDP-трафика. 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) [16]. 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.

Для определения классификационных характеристик сетевого трафика в основном используются статистические характеристики (а не содержимое), описывающие передаваемый по сети трафик. Идея использования статистических характеристик сетевого трафика для их классификации или для описания их свойств не новая. В работах [17, 18] 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.

В работе [19] предлагается метод классификации сетевого трафика, 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.

Основным этапом определения классификационных характеристик сетевого трафика является процесс измерения сетевого трафика. Измерение сети [20] является основой исследования поведения сети, которое включает активное и пассивное измерения. На основании измерения исследователь может получить важную информацию о свойствах сетевого трафика. При этом измерения могут быть проведены разными способами, в разных местах сети и в различные периоды времени и длительности. Место измерения указывает на то, какая часть или элемент КС, а также какая величина измеряется. При этом очень важно различать измерения сетевого трафика от идентификации приложений, так как в первом случае осуществляются сбор и обработка данных, а во втором случае – распознавание и классификация некоторых характеристик сетевого трафика. В свою очередь идентификация сетевого трафика является неотъемлемой частью классификации, так как классификация невозможна без его идентификации.

Активное измерение [21] осуществляется путем ввода стандартных тестовых пакетов (например, пакетов ping, traceroute, pathchar, ICMP-протоколов (Internet Control Message Protocol)), чтобы измерить их поведение. Однако активное измерение имеет некоторые недостатки; во-первых, ввод тестовых пакетов может воздействовать на сетевой трафик и нарушить нормальное функционирование, что в свою очередь повлияет на результат мониторинга; во-вторых, одновременный мониторинг большого количества узлов, т.е. генерация большого количества тестовых пакетов, может нарушить нормальное функционирование. В отличие от активного измерения, пассивное измерение [22] осуществляется на основе анализа данных, собранных с одного или нескольких узлов. При пассивном измерении сети нет необходимости посылки тестовых пакетов, и это означает, что сетевой трафик не нарушается и на результат мониторинга ничего не влияет. Обычно пассивное измерение используется для мониторинга основных сетевых устройств. Недостаток пассивного измерения сети связан с трудностью обработки огромного объема данных, собранных при измерении.

Анализ сетевого трафика может быть осуществлен на нескольких абстрактных уровнях: на уровне номера портов, содержимого пакета, потока, заголовка пакета и на уровне бита (т.е. объема трафика). При этом характеристики сетевого трафика на каждом уровне отличаются, например, на уровне пакета, сетевой трафик характеризуется размером пакета и временным интервалом между пакетами. А анализ на уровне бита в основном касается количественных характеристик сетевой сети, таких как интенсивность передачи и пропускная способность обмена в каналах сети. На уровне пакета рассматривается процедура прибытия IP-пакетов, т.е. интенсивность их задержки и потери пакетов.

Основными составляющими потока являются адрес и протокол. Например, в работе [13] поток определяется как последовательность обмена пакетами между

двумя хостами, которые определяются как кортеж, состоящий из пяти элементов (IP-адрес источника, номер порта источника, IP-адрес назначения, номер порта назначения, тип протокола). На этом уровне рассматриваются главным образом вопросы прибытия потока, интервал между поступлениями пакетов и т.д.

В общем, целью классификации сетевого трафика является отображение потока сетевых данных в определенные типы приложений или классы трафиков. Для классификации трафиков отдельных приложений, сервисов и протоколов также могут быть использованы такие индикаторы передачи сетевого трафика, как значения интенсивности передачи пакетов, их размеры и типы, а также распределение IP- и MAC-адресов источника и назначение в передаваемых пакетах.

Классификация сетевого трафика

Формально задача классификации сетевого трафика определяется следующим образом. Пусть дано множество потоков сетевых данных $X = \{f_1, f_2, \ldots, f_n\}$, где каждый поток сетевых данных f_i характеризуется p множеством атрибутов $\{x_{i_1}, x_{i_2}, ..., x_{i_p}\}$ и множеством классов трафика $C = \{C_1, C_2, ..., C_k\}$. Требуется определить такое отображение $f : X \to C$, чтобы каждый поток f_i соответствовал только одному классу трафика [23]. В качестве атрибутов потоков сетевых данных могут использоваться средняя длина пакета, средняя продолжительность, размер потока и т.д., а в качестве классов трафика Web, Peer-to-Peer, FTP и т.д.

Методы МО являются очень важной частью дисциплины искусственного интеллекта. Способность методов МО непрерывно получить новое знание или преобразовать структуры знания, которые облегчают их использование, позволила широко использовать эти методы для классификации сетевого трафика. При создании модели классификации сетевого трафика обучающие данные (например, набор потока сетевых данных) используются для изучения особенностей классов, что является основанием создания классификатора.

Процедура МО может быть разделена на две части: создание модели классификации и собственно классификация. Методы МО делятся на методы обучения с учителем и обучение без учителя [12, 23]. Обучение с учителем создает структуры знаний, которые используются для отнесения новых образцов в заранее определенные классы. Обучение машины проводится с представлением к ее входу наборов типовых примеров, которые принадлежат заранее определенным классам. Результатом процесса обучения является построение модели классификации на основе анализа и обобщения представленных образцов. На самом деле обучение с учителем создает модель взаимосвязи входа и выхода, т.е. осуществляет отображение набора входных атрибутов на выходные классы.

Однако использование в модели классификации сетевого трафика метода обучения без учителя может создать определенные преимущества. Основным преимуществом является то, что модель позволит идентифицировать новые приложения и группировать их в новый кластер, тогда как модели, использующие методы обучения с учителем, могут идентифицировать трафики, для которых созданы обучающие примеры, и не могут обнаружить новых приложений [24]. При классификации сетевого трафика методы без учителя не нуждаются в начальной

ручной разметке входных данных, они только основываются на подобии между классифицируемыми объектами и в качестве входных данных используются статистические характеристики потока сетевых данных.

Для создания модели классификации сетевого трафика в качестве метода обучения без учителя предлагается использовать кластеризацию [12, 23].

Формально кластеризация обучающих потоков может быть описана следующим образом Пусть даны множество потоков сетевых данных $X = \{f_t, f_2, ..., f_n\}$ и желаемое количество кластеров k. Требуется определить такое отображение $f : X \to \{C_1, C_2, ..., C_k\}$, чтобы каждый поток был отнесен только к одному кластеру C_i, $1 \le i \le k$, при условии, что $D = \bigcup_{j=1}^{k} C_j$ и $C_i \bigcap C_j = \varnothing$, $\forall i \ne j$ [23, 25, 26].

Цель кластеризации заключается в построении оптимального разбиения объектов на группы, т.е. в разбиении n сетевого потока на k кластеров. При этом для кластеризации объектов выбирается метрика подобия. Оптимальность кластеризации может быть определена как требование минимизации среднеквадратичной ошибки разбиения.

В литературе определено несколько метрик подобия. Метрика выбирается в зависимости от пространства, где расположены объекты, или от неявных характеристик кластеров. Расстояние $r(f_i, f_j)$ между потоками f_i и f_j определяется в результате применения выбранной метрики в пространстве характеристик. Для определения расстояния подобия между потоками f_i и f_j мы используем метрику Евклида:

$$r(f_i, f_j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2} .$$

Как видно, чем меньше Евклидово расстояние между двумя векторами потока, тем больше подобие между ними.

В литературе существует много различных алгоритмов кластеризации [12, 23, 25, 26]. В качестве алгоритма кластеризации предлагается использовать алгоритм *k*-средних. Выбор алгоритма *k*-средних обосновывается его быстротой и простотой и пригоден для кластеризации сетевого трафика и состоит из следующих шагов:

1. Случайный выбор центров кластеров χ_i, $i = 1.2, ..., k$ из обучающих потоков сетевых данных.
2. В кластер добавляется наиболее подобный поток сетевых данных.
3. Пересчет центров кластеров согласно текущему членству и затем перераспределение потоков сетевых данных на основе новых центров.
4. Если критерий остановки алгоритма не удовлетворен, вернуться к шагу 2.

В качестве критерия остановки выбирается минимальное изменение среднеквадратической ошибки разбиения:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} \left| r(f_j, \chi_i) \right|^2 .$$

Алгоритм *k*-средних имеет некоторые недостатки. Например, алгоритм чувствителен к начальному выбору центров кластеров, и он часто вместо глобального оптимума находит локальный оптимум. Это делает необходимым

многократное повторное выполнение алгоритма *k*-средних, чтобы получить разумное разделение потоков.

Заключение

Данная статья посвящена анализу и моделированию классификации сетевого трафика. Сетевой трафик является одним из важнейших фактических показателей работы КС и является носителем информации о поведении пользователей. На основе статистического анализа сетевого трафика можно косвенно определить статистические характеристики функционирования КС.

Для классификации сетевого трафика предложено использование методов МО без учителя. Использование в модели классификации сетевого трафика метода обучения без учителя позволит идентифицировать новые приложения и группировать их в новый кластер. В качестве метода обучения без учителя использован алгоритм кластеризации *k*-средних. Выбор алгоритма обосновывается его быстротой и простотой.

Предложенная в работе модель классификации позволит решить такие задачи, как определение приоритетов при формировании полосы пропускания для отдельных трафиков, установление правил по управлению трафиком, обеспечение безопасности сети, диагностический мониторинг и т.д.

Литература

1. H.Kim, M.Fomenkov, D.Barman, M.Faloutsos, and K.Lee, Internet traffic classification demystified: myths, Caveats, and the Best Practices // Proceedings of the 4th Conference on Emerging Network Experiment and Technology, December 09–12, 2008, pp.112–124.
2. S.Sen, O.Spatscheck, and D.Wang, Accurate, scalable In-network identification of P2P traffic using application signatures // Proceedings of the 13th International conference on World Wide Web. New York, USA, May 17–20, 2004. pp.512–521.
3. L7-filter, Application layer packet classifier for linux, 2009 - http://l7-filter.sourceforge.net/ (accessed 2009-04-02).
4. B.C.Park, Y.J.Win, M.S.Kim, and J.W.Hong, Towards automated application signature generation for traffic identification //NOMS: Network operations and management symposium, Salvador, Bahia, Brazil,7–11 April 2008, pp.160–167.
5. G.Szabo, D.Orincsay, S.Malomsoky, and I.Szabo, On the validation of traffic classification algorithms // Proceedings of the 9th International Passive and Active Measurement conference, April 29–30, 2008, pp.72–81.
6. P. Gupta and N.McKeown, Algorithms for packet classification // IEEE Network Magazine. 2001, vol. 15, no. 2, pp.24–32.
7. 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, November 1994. pp.115–123.
8. 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.

9. W.Moore and D.Papagiannaki, Toward the Accurate Identification of Network Applications // In Proceedings of the Sixth Passive and Active Measurement Workshop, March 31 – April 1, 2005, pp.41–54.
10. W.Li, M.Canini, A.W.Moore and R.Bolla, Efficient application identification and the temporal and spatial stability of classification schema // Computer Networks, 2009, vol. 53, # 6, pp.790–809.
11. A.W.Moore and D.Zuev, Internet traffic classification using bayesian analysis techniques // Proceeding of the Conference on Measurement and Modeling of Computer Systems, Banff, Alberta, Canada, June 06-10, 2005, pp.50-60.
12. N. J. Nilsson, Introduction to Machine Learning http://robotics.stanford.edu/people/nilsson/MLDraftBook/MLBOOK.pdf, accessed September 2009.
13. L.Zhanh and J.Tang, Characterization and performance study of IP traffic in WDM networks // Computer communications, 2001, No.24, pp.1702–1713.
14. A.Feldmann, Characteristics of TCP connection arrivals Technical report of the AT&T Labs Research, 1998.
15. R.Caceras, P.Danzig, S.Jamin, and D.Mitzel, Characteristics of Wide-Area TCP/IP Conversations, ACM SIGCOMM, 1991.
16. IANA, http://www.iana.org/assignments/port-numbers (as of August 2005).
17. V.Paxson, Empirically derived analytic models of wide-area TCP connections, IEEE/ACM Trans. Netw., 1994, vol. 2, no. 4, pp.316–336.
18. V.Paxson and S.Floyd, Wide area traffic: the failure of Poisson modeling, IEEE/ACM Trans. Netw., 1995, vol. 3, no. 3, pp.226–244.
19. 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, New York, USA, 2005, pp.229–240.
20. CAIDA Homepage, http://www.caida.org.
21. I.D.Graham, S.F.Donnelly, S.Artin, J.Martens, and J.G.Cleary, Non intrusive and Accurate Measurement of Unidirectional Delay and Delay Variation on the Internet // Proceedings of the Internet Society's 8th Annual Networking Conference, Geneva, July 21–24, 1998,
22. B.Huffaker, M.Fomenkov, D.Moore, E.Nemeth, and K.Claffy, Measurements of the Internet topology in Asia-pacific Region, 2000, http://www.caida.org/outreach/papers/asia_paper/
23. M.Dunham, Data Mining: Introductory and Advance Topics. Prentice Hall, New Jersey, 1st edition, 2003.
24. J.Erman, A.Mahanti, and M.Arlitt, Internet Traffic Identification using Machine Learning. In GLOBECOM'06, San Francisco, USA, November, 2006.
25. A.K.Jain, M.N.Murty, and P.J.Flynn, Data Clustering: A Review // ACM Computing Surveys, 1999, vol.31, # 3, pp.254–323.
26. D.Jiang, C.Tang, and A.Zhang, Cluster Analysis for Gene Expression Data: A Survey // IEEE Transactions On Knowledge And Data Engineering, vol. 16, #12, December 2004, pp.1370–1386.

UOT 004.048
Şıxəliyev R.H.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
ramiz@science.az

Kompyuter şəbəkələrinin şəbəkə trafikinin analizi və təsnifatı

Məqalə kompyuter şəbəkələrinin monitorinqi üçün vacib olan, şəbəkə trafiki analizinə və təsnifatının modelləşdirilməsinə həsr edilmişdir. Şəbəkə trafikinin modelləşdirilməsi üçün müəllimsiz maşın təlimi üsullarının tətbiqi təklif edilmişdir. Müəllimsiz təlim üsulunda k-means klasterləşdirmə alqoritmindən istifadə edilmişdir.
Açar sözləri: *şəbəkə trafiki, klasterizasiya, k-means alqoritmi.*

Shikhaliyev R.H.
Institute of Information Technology ANAS, Baku, Azerbaijan
ramiz@science.az

The analysis and classification of the computer networks traffic

The paper is devoted to analysis of network traffic and modeling of its classification which are importent for computer networks monitoring. For modeling of network traffic classification, an unsupervised machine training method is proposed where k-means clusterization algorithm is used.
Keywords: *network traffic, clusterization, k-means algorithm.*