УДК 004.048

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

ОБ ОДНОМ МЕТОДЕ КЛАССИФИКАЦИИ ТРАФИКА КОМПЬЮТЕРНЫХ СЕТЕЙ

Точная классификация трафика компьютерных сетей (КС) необходима для их эффективного управления, мониторинга и обеспечения безопасности. В статье для классификации трафика КС предлагается использовать алгоритмы машинного обучения с учителем и поиска ассоциативных правил. Предложенный метод классификации трафика позволит повысить производительность и точность классификации даже при небольших обучающих выборках.

Ключевые слова: *компьютерные сети, сетевой трафик, классификация трафика, классификационные признаки трафика, машинное обучение, ассоциативные правила, SVM-метод.*

Введение

Сегодняшние компьютерные сети предоставляют пользователям множество различных интернет-сервисов. В результате компьютеры пользователей КС генерируют множество различных типов трафика. Также постоянно появляются все новые приложения и модели использования интернет-сервиса, что усложняет задачи обеспечения безопасности, мониторинга и управления КС. При этом детальное знание состава общего сетевого трафика является очень важным для решения этих задач [1]. К таким задачам относятся мониторинг безопасности сети, планирование и резервирование сетевых ресурсов, обеспечение QoS (Quality of Service) [2] и т.д. Для решения этих и других задач необходима быстрая и точная классификация трафика КС. Однако точная классификация сетевого трафика в реальном времени является сложной проблемой.

Известно, что трафик КС состоит из трафиков, генерируемых различными приложениями. При этом одной из основных задач классификации трафика КС является определение в общем сетевом трафике доли трафика каждого приложения.

Наиболее простые методы классификации трафика КС основаны на анализе информации, имеющейся на полях заголовка пакетов, например номеров портов, используемых приложениями, или при анализе протоколов прикладного уровня. Эти методы являются достаточно точными, однако имеют некоторые недостатки, так как многие приложения могут использовать номера портов, не назначенных IANA (Internet Assigned Numbers Authority) [3], а для полного анализа протоколов потребуются большие вычислительные ресурсы.

Из соображений безопасности некоторые протоколы шифруются, например SSH [4] и SSL [5], что делает невозможным их анализ. Кроме того, затрудняется анализ некоторых специально разработанных протоколов, которые не имеют никакого открытого описания, например Skype [6], MSN Messenger [7] и др. Вместе с тем, с появлением множеств все новых приложений, снижается точность этих методов.

Для классификации трафика КС также существуют и другие подходы, которые основываются на анализе сигнатуры приложений, поведения хостов, характеристик потоков и т.д. Эти методы также имеют некоторые недостатки. Например, при зашифрованном трафике классификация трафика на основе анализа сигнатуры приложений затрудняется, приложения с одинаковым поведением невозможно различить и т.д.

Для классификации сетевого трафика КС очень важны определение оптимального набора классификационных характеристик потоков, создаваемых приложениями, и влияние этих характеристик на результат классификации. При этом для автоматического обнаружения характеристик определенных видов трафика совместно со статистическими методами используются методы искусственного интеллекта.

Исходя из вышесказанного, а также учитывая сложность и масштабы современных крупных и гетерогенных КС, необходимо автоматизировать процесс классификации сетевого трафика. Для решения этой задачи требуется создавать методы, использующие алгоритмы машинного обучения (МО) с учителем, что позволит обеспечить более высокую точность классификации. Однако основным недостатком классификации на основе алгоритмов МО с учителем является этап обучения. При этом производительность и точность классификации трафика на основе методов МО зависят от размера и достоверности используемых обучающих данных.

В данной статье для классификации трафика предлагается использовать алгоритм МО с учителем и алгоритм поиска ассоциативных правил, что позволит повысить производительность и точность классификации даже при небольших обучающих выборках.

Подходы к классификации трафика, основанные на алгоритмах МО

В последнее десятилетие существенная часть работ по классификации сетевого трафика была основана на применении 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 трафика 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. Для классификации трафика на основе содержащейся в нем информации, такой как зашифрованные приложения и конфиденциальные данные пользователей, в работе [8] авторами были использованы наивные методы Байеса. При этом для классификации сетевого трафика использовались статистические свойства трафика.

В работе [9] 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 [10] авторы предложили подход к классификации трафика на основе анализа потока пакетов в режиме реального времени. В работе [11] для точной классификации трафика применены Байесовские нейронные сети. В [12] для классификации трафика авторами используются однонаправленные статистические функции. В работе [13] для компактного выражения трех статистических характеристик трафика авторами была использована функция плотности вероятности. В работе [14] 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 [15] для классификации трафика авторами были использованы методы ближайших соседей и линейного дискриминантного анализа, при этом для классификации использовали пять статистических характеристик. В работе [16] 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.

С помощью метода классификации трафика 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.

В работе [17] 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 [18] для кластеризации потока трафика был использован алгоритм AutoClass, и для оценки кластеров была предложена метрика внутриклассовой однородности. В работе [19] для кластеризации трафика был использован алгоритм *к*-средних и с помощью анализа полезной информации были промаркированы кластеры для приложений. В работе [20] 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 обнаружения трафиков ранее неизвестных приложений. В работе [21] 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 [22] 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 [23] 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 [24]. 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 [25] авторами было предложено использовать набор обучающих данных в алгоритиме МО без учителя. Однако при малых размерах обучающих данных основную часть отображения составляют «неизвестные» кластеры.

Исходя из вышеприведенного анализа методов классификации сетевого трафика, можно сказать, что точность методов классификации различается в зависимости от типов используемых для классификации информации и приложений. Поэтому для обеспечения точности и полноты классификации трафика необходимо создать комплексный механизм. Для этого лучшим решением является комбинирование существующих механизмов классификации с использованием МО с учителем и без него, чтобы использовать их преимущества.

Метод классификации трафика КС

В основном современные методы классификации сетевого трафика основываются на использовании статистических характеристик потоков и методов МО. Вместе с тем поток определяется множеством (серией) пакетов, которыми обмениваются два хоста и которые классифицируются как однонаправленные, двунаправленные и полные потоки. В свою очередь пакеты определяются кортежем, состоящим из адреса отправителя, порта отправителя, адреса получателя, порта получателя и типа транспортного протокола.

Статистическая характеристика потока определяется характеристикой множества пакетов, составляющих поток. Также могут быть использованы некоторые статистические данные, такие как длительность потока, количество переданных пакетов в байтах, среднее время между пакетами и средний размер пакетов.

В этом разделе для классификации трафика КС предлагается модель, которая может повысить производительность и точность классификации. Для этого в классификационный процесс предлагается включить модуль поиска ассоциативных правил. Предлагаемая модель состоит из нескольких этапов, и процесс классификации ориентирован на классификацию трафика на уровне потока (рис. 1).

Рис.1. Модель классификации сетевого трафика

На этапе предварительной обработки система захватывает IP-пакеты из сети и, осматривая заголовки пакетов, создает потоки трафика. Поток состоит из последовательности IP-пакетов, имеющих в заголовке одну и ту же информацию:

- исходный IP-адрес;
- номер исходного порта;
- IP-адрес назначения;
- номер порта назначения и тип протокола.

На следующем этапе для представления каждого потока извлекаются множества статистических характеристик трафика. А целью выбора характеристик является выбор подмножества характеристик, релевантных для осуществления точной классификации. Поиск ассоциативных правил используется для сокращения размерности пространства характеристик сетевого трафика и выявления корреляций между характеристиками сетевого трафика.

На последном этапе классификатор на основе информации о статистических характеристиках трафика и ассоциативных правилах классифицирует потоки трафика по приложениям.

В качестве классификатора предлагается использовать метод опорных векторов (МОВ) – Support Vector Machine (SVM) [26]. МОВ – это известный метод машинного обучения, который основывается на теории статистического обучения и структурной минимизации риска.

По сравнению с другими методами МО, МОВ имеет некоторые преимущества, такие как использование небольшого набора обучающих выборок, высокая точность, высокая производительность классификации и т.д.

Метод МОВ широко используется для классификации сетевого трафика [27, 28, 29] и был выбран в качестве алгоритма классификации благодаря своей способности одновременно минимизировать эмпирические ошибки классификации и максимизации геометрической полосы классификационного пространства. Эти свойства позволяют снизить структурный риск при обучении с ограниченным количеством выборок.

Задача классификации сетевого трафика на основе МО определяется следующим образом. Классификатор, основанный на МО, использует множество обучающих данных, которое содержит N кортежей (x, y) и функцию обучения $f(x) \to y$, где x является множеством классификационных характеристик потоков, создаваемых приложениями, а y является множеством классов трафика приложений. Пусть $X = \{x_1, x_2, \dots, x_n\}$ является множеством потоков. Причем поток x_i характеризуется вектором значений атрибутов $x_i = \{x_{ij} | 1 \le j \le m\}$, где m является числом атрибутов и x_{ij} является значением -го атрибута i-го потока. Также пусть $Y = \{y_1, y, \dots, y_k\}$ является множеством классов трафика, где k является числом интересующих классов и y_i может быть такими классами, как P2P-, WWW-, FTP-трафики и т.д.

Поиск ассоциативных правил

Для решения задачи сокращения размерности пространства характеристик сетевого трафика КС и выявления корреляций между ними используем алгоритм нахождения ассоциативных правил [30].

На основании ассоциативных правил может быть определена часто встречаемая устойчивая комбинация характеристик сетевого трафика, которая определяется значением поддержки правила.

Задача сокращения размерности пространства характеристик сетевого трафика состоит в следующем. Пусть $X = \{x_1, x_2, ..., x_n\}$ – множество характеристик, которые характеризуют сетевой трафик КС. Пусть $T = \{t_1, t_2, ..., t_m\}$ – сетевой трафик КС, который состоит из трафиков m субъектов КС (например пользователей, серверов, приложений), где каждый t_i-й трафик содержит набор характеристик, входящих во множество X, т.е. $T \subseteq X$.

Предполагается, что сетевой трафик T имеет множество характеристик A, входящих во множество X, если $A \subseteq X$. Тогда ассоциативным правилом характеристик сетевого трафика КС является импликация $A \Rightarrow B$, где $A \subset X$, $B \subset X$ и $A \cap B = \varnothing$. В сетевом трафике КС правило $A \Rightarrow B$ выполняется с достоверностью c, если c% трафиков субъектов КС, входящих в T, содержит множество признаков A, а также множество признаков B. При этом достоверность правила рассчитывается по формуле:

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)}.$$

Правило $A \Rightarrow B$ имеет поддержку s, если s% трафиков субъектов КС, входящих в сетевой трафик T КС, содержит $A \cup B$, причем $s(A \Rightarrow B) = s(A \cup B)$.

Для нахождения в больших объемах данных ассоциативных правил наиболее широко используется алгоритм APriori, основным достоинством которого является его гибкость [31].

При этом имеется возможность задавать и *min_sup,* и *min_conf* правила, что позволяет получать множество разных групп правил. Однако генерация большого

количества ассоциативных правил создает серьезную проблему для их анализа. Поэтому для нахождения ассоциаций в характеристиках трафика недостаточно использования только алгоритма APriori. Исходя из этого, предлагается обобщить «подобные» правила, т.е. определить обобщенные ассоциативные правила, в которых получаемые правила включают характеристики, являющиеся предками характеристик, входящих в трафики субъектов КС. В результате можно выявить ассоциативные правила не только между отдельными признаками сетевого трафика КС, но и между трафиками субъектов КС.

При нахождении обобщенных ассоциативных правил важным элементом является таксономия (иерархия) характеристик сетевого трафика КС. Под таксономией здесь понимается лес направленных деревьев, листьями которых являются характеристики сетевого трафика КС, а внутренними узлами – их группы.

Пример иерархии некоторых протоколов и групп протоколов, которые являются характеристиками сетевого трафика КС, приведен на рис. 2. В получаемых на основании такой таксономии правилах, как в предпосылке, так и впоследствии, могут присутствовать элементы, находящиеся на разных уровнях таксономии. Например, «если в трафике пользователя присутствует HTTP-протокол, то, скорее всего, будет присутствовать и DNS-протокол».

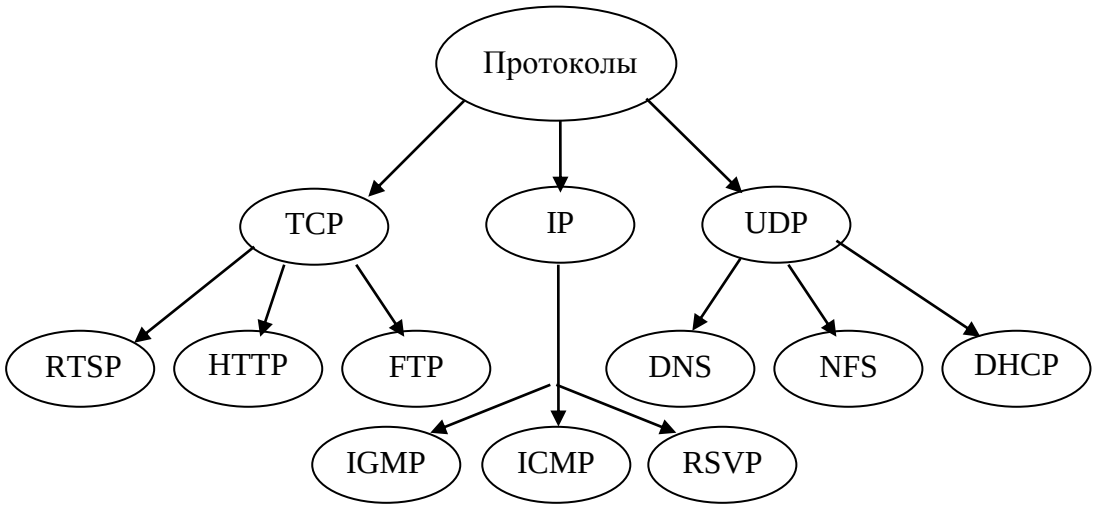

Рис.2. Иерархия некоторых протоколов и групп протоколов

Введение дополнительной информации о группировке характеристик сетевого трафика КС в виде иерархии может дать следующие преимущества:

1. Могут быть выявлены ассоциативные правила не только между отдельными характеристиками сетевого трафика КС, но и между различными группами характеристик.

2. В некоторых случаях отдельные характеристики сетевого трафика КС могут иметь очень маленькую поддержку, однако значение поддержки всей группы, в которую входит эта характеристика, может быть больше порога *min_sup*.

3. Введение информации о группировке характеристик сетевого трафика КС может использоваться для отсечения неинформативных правил.

Таким образом обобщенным ассоциативным правилом называется импликация $A \Rightarrow B$, где $A \subset X$, $B \subset X$ и $A \cap B = \varnothing$, и где ни один из элементов, входящих в набор B, не является предком ни одного элемента, входящего в A. Поддержка и достоверность подсчитываются так же, как и в случае ассоциативных правил.

Для нахождения обобщенных ассоциативных правил целесообразно использовать специализированный алгоритм [32], который является более эффективным, чем стандартный алгоритм APriori.

Заключение

Статья посвящена вопросу классификации трафика компьютерных сетей. Для решения этого вопроса предлагается использовать вместе алгоритмы машинного обучения с учителем и поиска ассоциативных правил. Предлагаемая модель состоит из нескольких этапов, и процесс классификации ориентирован на классификацию трафика на уровне потока. На этапе предварительной обработки система, осматривая заголовки IP-пакетов, взятых из сети, создает потоки трафика. После этого для представления каждого потока извлекаются множества статистических характеристик трафика. При этом поиск ассоциативных правил используется для сокращения размерности пространства характеристик сетевого трафика и выявления корреляций между характеристиками сетевого трафика. Наконец, классификатор на основе информации о статистических характеристиках трафика и ассоциативных правилах классифицирует потоки трафика по приложениям. В качестве классификатора предложено использовать метод опорных векторов (Support Vector Machine (SVM)). Предложенный метод классификации трафика позволит даже при небольших обучающих выборках повысить производительность и точность результата классификации.

Благодарность

Данная работа выполнена при финансовой поддержке Фонда Развития Науки при Президенте Азербайджанской Республики – грант № EİF-RİTN-MQM-2/İKT-2-2013-7(13)-29/27/1

Литература

1. Nguyen T.T., Armitage G. A survey of techniques for internet traffic classification using machine learning // IEEE Commun. Surveys & Tutorials, 2008, vol.10, no.4, pp.56–76,
2. Roughan M., Sen S., Spatscheck O., Duffield N. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification / Proceedings of the ACM SIGCOMM Conference on Internet Measurement, 2004, pp.135–148.
3. IANA, http://www.iana.org/assignments/port-numbers.
4. RFC 4251. http://www.ietf.org/rfc/rfc4251.txt.
5. RFC 2246. http://www.ietf.org/rfc/rfc2246.txt.
6. Skype. http://www.skype.com.
7. MSN Messenger. http://join.msn.com/messenger/overview2000.
8. Moore A.W., Zuev D. 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, 2005, pp.50–60.
9. Williams N., Zander S., Armitage G. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification // ACM SIGCOMM Computer Communication Review, 2006, vol.36, no.5, pp.5–16.
10. Nguyen T., Armitage G. 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, 2006, pp.369–376.
11. Auld T., Moore A.W., S.F.Gull. Bayesian neural networks for internet traffic classification // IEEE Trans. Neural Networks, January 2007, vol.18, no.1, pp.223–239.

12. Erman J., Mahanti A., Arlitt M., Williamson C. Identifying and discriminating between web and peer-to-peer traffic in the network core / Proceedings of the 16th international conference on World Wide Web, 2007, pp.883–892.
13. Crotti M., Dusi M., Gringoli F., Salgarelli L. Traffic classification through simple statistical fingerprinting // ACM SIGCOMM Computer Communication Review, 2007, vol.37, no.1, pp.5–16,
14. Este A., Gringoli F., Salgarelli L. Support vector machines for tcp traffic classification // Computer Networks, 2009, vol.53, no.14, pp.2476–2490.
15. Roughan M., Sen S., Spatscheck O., Duffield N. 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, 2004, pp.135–148.
16. Kim H., Claffy K., Fomenkov M., Barman D., Faloutsos M., Lee K. Internet traffic classification demystified: myths, caveats, and the best practices / Proceedings of the ACM CoNEXT Conference, 2008, pp.1–12.
17. McGregor A., Hall M., Lorier P., Brunskill J. Flow clustering using machine learning techniques / Proceedings of Passive and Active Measurement Workshop, 2004, pp.205–214.
18. Zander S., Nguyen T., Armitage G. Automated traffic classification and application identification using machine learning / Annual IEEE Conference on Local Computer Networks, 2005, pp.250–257.
19. Bernaille L., Teixeira R., Akodkenou I., Soule A., Salamatian K. Traffic classification on the fly // ACM SIGCOMM Computer Communication Review, 2006, vol.36, no.2, pp.23–26.
20. Erman J., Arlitt M., Mahanti A. Traffic classification using clustering algorithms / Proceedings of the SIGCOMM workshop on Mining network data, 2006, pp.281–286.
21. Wang Y., Xiang Y. and S.-Z. Yu. An automatic application signature construction system for unknown traffic // Concurrency Computations: Pract. Exper., 2010, vol.22, pp.1927–1944.
22. Finamore A., Mellia M., Meo M.. Mining unclassified traffic using automatic clustering techniques // TMA International Workshop on Traffic Monitoring and Analysis, 2011, pp. 150–163.
23. Zhang J., Xiang Y., Wang Y., Zhou W., Xiang Y., Guan Y. Network traffic classification using correlation information // IEEE Transactions on Parallel and Distributed Systems, 2012, vol.24, no.1, pp.1–15.
24. Gu1 C., Zhang S., Chen X., Du A. Realtime traffic classification based on semi-supervised learning // Journal of Computational Information Systems 2011, no.7, pp.2347–2355.
25. Erman J., Mahanti A., Arlitt M., Cohen I., Williamson C. Offline/realtime traffic classification using semi-supervised learning // Performance Evaluation, October 2007, vol.64, no.9-12, pp.1194–1213,
26. http://cs229.stanford.edu/notes/cs229-notes3.pdf
27. Este A., Gringoli F., Salgarelli L. Support Vector Machines for TCP Traffic Classification // The International Journal of Computer and Telecommunications Networking, 2009, vol.53, no.14, pp.2476–2490.
28. Yang A., Jiang S., Deng H. A P2P Network Traffic Classification Method Using SVM / The 9th International Conference for Young Computer Scientists, 2008, pp.398–403.
29. Sena G.G., Belzarena P. Early Traffic Classification Using Support Vector Machines / Proceedings of the 5th International Latin American Networking Conference, LANC ’09. 2009, pp.60–66.
30. Шихалиев Р.Г. Об одном методе сокращения размерности анализируемых признаков сетевых трафиков, используемых для мониторинга компьютерных сетей // Телекоммуникации, 2011, №06, с.44–48.

31. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules in Large Databases / Proc. Conf. Very Large Databases, 1994, pp.487–499.
32. Srikant R., Agrawal R. Mining Generalized Association Rules / Proceedings of the 21th International Conference on Very Large Data Bases, 1995, pp.407–419.

UOT 004.048
Şıxəliyev Ramiz H.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
ramiz@science.az
Kompüter şəbəkələrinin trafikinin klassifikasiyasının bir modeli haqqında
Kompüter şəbəkələrinin (KŞ) trafikinin dəqiq klassifikasiyası onların effektiv idarə edilməsi, monitorinqi və təhlükəsizliyinin təmin edilməsi üçün vacibdir. Məqalədə KŞ-nin trafikinin klassifikasiyası üçün maşın təlimi və assosiativ qaydaların tapılması alqoritmlərinin birgə istifadəsi təklif edilir. Təklif edilmiş metod hətta kiçik təlim verilənləri istifadə edildikdə belə klassifikasiyanın məhsuldarlığını və dəqiqliyini artırmağa imkan verir.
Ключевые слова: *kompüter şəbəkələri, şəbəkə trafiki, trafikin klassifikasiyası, trafikin klassifikasiyası əlamətləri, maşın təlimi, assosiativ qaydalar, SVM-metodu.*

Ramiz H. Shikhaliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
ramiz@science.az
One method for computer networks traffic classification
Precise traffic classification of computer networks (CN) is necessary for their effective management, monitoring and security. Article proposes sharing use of machine learning and associative rules mining algorithms for CN traffic classification. The proposed method of classifying traffic will improve performance and classification accuracy with small training datasets.
Keywords*: computer networks, network traffic, traffic classification, traffic classification features, machine learning, associative rules, SVM-method.*