Academic Research in Educational Sciences VOLUME 2 | ISSUE 12 | 2021
ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

ВЫБОР ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ КЛАССИФИКАЦИИ IP-ТРАФИКА СЕТИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Улугбек Рахимжон угли Охундадаев
Национальный университет Узбекистана
ulugbek_1122@mail.ru

АННОТАЦИЯ

IP-протокол и протоколы транспортного уровня (TCP, UDP) имеют множество различных параметров и характеристик, которые можно получить из как непосредственно заголовков пакетов, так и статистических наблюдений за потоками. Для решения задачи классификации сетевого трафика методами машинного обучения необходимо определить набор данных (признаков), которые целесообразно использовать для решения задачи классификации. Выбор признаков зависит от требований к процессу классификации – скорости и точности классификации. В работах зарубежных авторов показано, что можно выделить до 248 различных атрибутов IP-трафика сети, которые потенциально можно использовать в методах машинного обучения для классификации или кластеризации IP-трафика по приложениям. Анализ показывает, что не все предложенные атрибуты одинаково влияют на точность и скорость классификации.

Ключевые слова: классификация, машинное обучение, объект, характеристика, нейронная сеть, управление данными, атрибут, SVM, IP-трафик, протокол.

ВВЕДЕНИЕ

Классификация трафика актуальна в настоящее время, поскольку полученные результаты могут быть применены к различным приложениям, важным как для сетевого администрирования, так и для конечного пользователя [1, 2]. С точки зрения провайдера идентификация протоколов /приложений/типов приложений по потокам данных в сети может использоваться для:

- контроль сети и трафика в ней (например, для блокировки определенных протоколов, таких как BitTorrent),

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

- обеспечение высокого качества обслуживания клиентов за счет эффективного определения приоритетов потоков и регулировки скорости передачи отдельных пакетов,
- регулирование цен на услуги,
- планирование распределения и использования ресурсов,
- оптимизация предоставляемых услуг и алгоритмов маршрутизации (например, для изменения приоритета передачи разных типов данных в случае высокой загруженности сети).

Оценка текущего использования сети пользователями может дать представление об оптимальном дизайне новых сетей с учетом понимания предпочтений и принципов работы пользователей Интернета и Интернет-сервисов, поскольку можно получить подробную статистику по всем услугам [3].

Поскольку потребности пользователей в использовании сети постоянно меняются, это необходимо знать их и модифицировать сеть для удовлетворения текущих потребностей. Этот требует как умения моделировать структуру сети в текущий момент времени, так и понимания направления ее развития и изменения. Например, сегодня мы можем наблюдать тенденцию к отказу от преобладающего ранее принципа асимметрии сетевого устройства в том смысле, что клиенты загружают гораздо больше информации, чем отправляют в сеть. Появление P2P-приложений, VoIP, видеозвонки, потоковое мультимедиа и другие инновации [4-7] должны побудить УТП отреагировать на реорганизацию своих сетей для удовлетворения новых потребностей клиентов [23-27]. Кроме того, растет число так называемых «умных устройств», которые должны составить основу Интернета вещей в будущем [28]: это также создаст ряд проблем для интернет-провайдеры, чтобы обеспечить максимальную эффективность своей работы. Особо следует отметить мобильные приложения, доля интернет-трафика которых неуклонно растет [8,9].

Использование смартфонов и мобильных приложений может считаться более персонализированными, поэтому получение данных об этом виде трафика позволяет эффективно построить сетевой портрет пользователя. Определение интересов пользователей может служить маркетинговым целям, позволяя проводить более целенаправленные рекламные кампании [10]. С точки зрения безопасности информационных систем, классификация интернет-потоков, может использоваться как важный атрибут при обнаружении кибер-

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

атак, аномалий в работе сети [12] незаконные или необычные действия пользователя и другие нарушения, что может улучшить общую безопасность Интернета. Методы, используемые для классификации интернет-трафик меняется вместе с глобальными изменениями в структуре трафика [13]. В внедрение новых технологий начинает негативно сказываться на качестве исполнения ранее применявшихся методов, что приводит к необходимости создания и разработки новых подходы. К таким глобальным изменениям, влияющим на проблему классификации трафика, можно отнести [14, 29-30]:

- отказ от утвержденного списка портов для протокола / приложения (преднамеренный или из-за устаревания этого списка);
- обфускация протоколов, чтобы скрыть те, которые заблокированы / подавлены провайдером;
- все более распространенное шифрование трафика, которое не позволяет содержимое полезная нагрузка пакета, которая будет использоваться для классификации;
- постоянное появление новых протоколов и приложений и т. Д. По указанным выше причинам проблема классификации интернет-трафика сегодня не может считаться решенной, и исследовательские группы продолжают предлагать новые решения, позволяющие показывать эффективные результаты в изменяющейся реальности.

Эволюция методов классификации трафика Методы классификации трафика развивались и изменялись с течением времени. Это в первую очередь связано с требованиями и ограничениями, налагаемыми сетью. Изменения в структуре сетевого трафика а особенности его передачи приводят к тому, что старые методы классификации становятся неэффективными или просто непригодными [15-19]. С другой стороны, развитие методов классификации и оборудования, на котором может работать система, позволяет использовать больше функций и более совершенные способы их применения при принятии решений. Важный характеристики методов классификации сетевого трафика включают:

- гранулярность: с каким уровнем точности система производит классификацию: семейство протоколов / класс приложений или конкретные протоколы, конкретные приложения.

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

- скорость ответа: способна ли система быстро классифицировать (через несколько пакетов), который подходит для анализа в реальном времени, или для классификации полностью требуются данные потока.
- вычислительные затраты: вычислительная сложность и затраты на использование памяти для классификации пакета или потока.

Первые системы классификации трафика основывались на извлечении номеров портов из пакеты и сопоставление их с IANA (Internet Assigned Numbers Authority) список. IANA выделяет и регистрирует номера портов, используемые для определенных конкретных [20-21] в целях, например, порт 80 выделен для протокола HTTP. Информация о протоколе уже можно использовать для приблизительного определения типа активности пользователя. Этот метод классификации работает очень быстро и не требует хранения потока и прост в вычислительном отношении. Это делает его удобным, например, для фильтрации трафика в межсетевых экранах. Однако у него есть ряд существенных недостатков, что по мере развития устройства сети отрицательно сказывается на ее результатах.

Постановка задачи

Одним из первых шагов для реализации любого метода машинного обучения является необходимость выбора определенных параметров (или признаков), на основе которых будет решаться задача классификации или кластеризации данных. В качестве признаков (признаков) классификации IP-трафика могут выступать либо данные, содержащиеся в заголовках пакетов сетевого и транспортного протоколов, либо данные, полученные статистическим путем. К первому типу данных относятся IP-адреса источника и назначения, номера портов протоколов транспортного уровня источников и назначения, поля TTL и другие данные, формируемые источником и получателем пакетов. Ко второму типу данных относятся межпакетный интервал, размер пакетов и т.д., то есть данные, рассчитываемые на основе статистики обработки пакетов.

Результатом обучения должно стать построение модели классификации на основе анализа и обобщения представленных признаков (образцов).

Целью работы является выбор признаков для оптимальной реализации классификации IP-трафика методами машинного обучения.

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

Модель классификации

В [1] предложено 248 различных признаков, которые могут характеризовать IP-трафик. Безусловно, не все признаки одинаково влияют на процесс классификации, поэтому на практике классификаторы выбирают наименьшее множество признаков, которые позволяют классифицировать.

Рисунок 1 иллюстрирует последовательность событий, связанных с обучением классификатора с учителем. На фазе обучения используем обучающую выборку, которая формирует модель классификации, на фазе тестирования используем тестирующую выборку и формируем результаты классификации.

Модель классификации - это набор алгоритмов, приложений, выбор атрибутов и алгоритмы оценки качества классификации.

Набор данных поступает на вход обучаемому алгоритму, а на выходе получается классификатор.

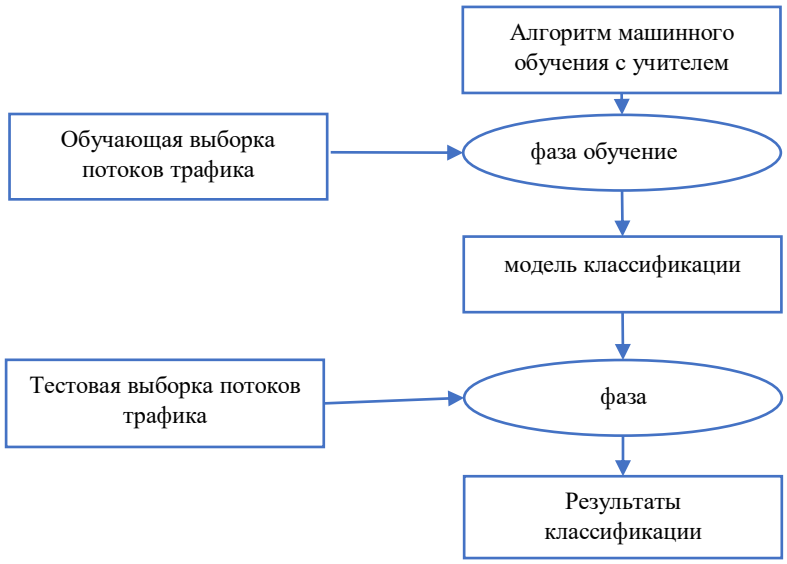

Рис. 1. Обучение классификатора с учителем

На «фазе обучения» используется фильтр выбора признаков, который позволяет ограничивать число признаков, действительно используемых при обучении классификатора, и, таким образом, формируется модель классификации.

При статистической классификации сетевого трафика, объекты потоков описываются измеренными значениями определенного набора атрибутов, а затем применяются для обучения и классификации. Другими словами, каждый

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

объект представляет собой вектор признаков $X = (x_1, x_2, ..., x_d)$, который может считаться точкой данных в d-мерном пространстве признаков, где d количество признаков.

Набор признаков, как правило, состоит из некоторых наблюдаемых характеристик пакетного уровня и уровня потоков трафика, показывающие отличительное поведение и внутреннюю природу сетевых приложений. В других аналогичных данных определяется, как правило, меньший по размеру набор признаков. Измеряется набор пакетов и байтов, передаваемых в потоке, а также максимальное, минимальное, среднее значение и стандартное отклонение длины пакета и межпакетного интервала. Эти характеристики вычисляются в ранних под-потоках (первые десять пакетов) и отдельно в каждом направлении. Всего в сумме можно выделить двадцать признаков, которые показаны в табл. 1.

Таблица 1

Простые признаки трафика

Что наблюдается	Статистика	Кол-во признаков
Пакеты	Кол-во пакетов	2
Байты	Объём байтов	2
Размер пакета	Мин., макс, среднее значение стандартного отклонения	8
Межпакетный интервал	Мин., макс, среднее значение стандартного отклонения	8
Всего		20

В качестве первоначальных атрибутов выбраны параметры сетевых потоков протоколов сетевого (TCP, UDP) и транспортного уровней (IP), представленных в табл. 2.

Имеются две причины для использования только простых признаков. С одной стороны, как показано в [2], простые атрибуты имеют наибольшее значение. С другой стороны, такое небольшое количество атрибутов требует намного меньших вычислительных затрат, по сравнению с большим набором признаков.

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

МЕТОДЫ ОТБОРА ПРИЗНАКОВ

Методы отбора признаков можно разделить на две категории: скалярный отбор, отбирающий признаки по отдельности и векторный отбор, выбирающей признаки, основываясь на взаимной корреляции между ними. Скалярный отбор имеет преимущество в упрощении вычислений, однако может быть неэффективным для набора данных с взаимно коррелированными признаками. С помощью методов векторного отбора удается выбирать оптимальные комбинации признаков.

В свою очередь методы векторного отбора можно, разделить на обёрточные методы и методы фильтрации. Обёрточные методы используют алгоритмы МО, в формате «черный ящик», и выбирают наиболее подходящие признаки таким образом, чтобы алгоритм обучения был оптимальным.

Оберточные алгоритмы производят оценку атрибутов используя значения показателя точности при работе целевого алгоритма МО.

Таблица 2

Первоначально выбранные атрибуты классификации

Атрибут	Описание
classname	Название укрупненного класса протокола (WEB, MAIL, FTP и т.д.) классифицированного трафика, которое будет использоваться при создании модели классификатора.
tot_pkts_qty	Общее количество пакетов в данном потоке в обоих направлениях.
tot_pkts_bytes	Общий размер в бантах всех пакетов в данном потоке в обоих направлениях.
rev_pkts_qty	Количество пакетов потока в обратном направлении в случае, если поток двунаправленный.
rev_pkts_bytes	Размер в байтах всех пакетов потока в обратном направлении.
fw_pkts_qty	Количество пакетов потока в прямом направлении.
fw_pkts_bytes	Размер в байтах всех пакетов потока в прямом направлении
is_reversable	Булева переменная, отражающая является ли данный поток двунаправленным.
Transport_protocol	Протокол транспортного уровня (TCP-. или UDP-)

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

src_port	Порт транспортного уровня источника (как для TCP, так и для UDP-)
dst_port	Порт адресата
Wirelen	Исходная длина всех пакетов потока в физическом канале, деленная на общее количество пакетов
header_count	Количество всех заголовков всех пакетов деленное на количество пакетов
tcp_syn	Процент пакетов с флагом SYN протокола транспорта уровня TCP. В случае, если используется UDP его значение равно GAP OFFSET - расстоянию в байтах между заголовков и полезной нагрузкой пакета, деленное на количество пакетов.
tcp_ack	Процент пакетов с флагом ACK TCP - протокола, для UDP берется GAP OFFSET - расстояние от начала пакета до конца заголовков, деленное на количество пакетов.
Flags	Среднее количество флагов TCP-протокола, для UDP берется среднее количество заголовков.
Pay-load_length	Средний размер полезной нагрузки протокола транспортного уровня в потоке.
is_fragment	Процент фрагментированных потоков.
Hlen	Количество заголовков протокола IP.
payload_offset	Среднее расстояние от начала пакета до полезной нагрузки.

Оберточный алгоритм Wrapper

Оберточные алгоритмы выделения атрибутов используют целевой алгоритм для оценки каждого подмножества признаков [3]. При оценке точности на классификаторе используется перекрестная проверка с настраиваемым числом «сверток». Перекрестная проверка может быть прекращена досрочно, если стандартная девиация результатов не превышает заданного порога, который обычно равен 0,01. Так же возможно использование отдельного, тестового набора данных, значения которого для получения более реальных результатов независимы от обучающей выборки.

Выбор подмножества атрибутов в оберточном алгоритме осуществляется методом прямого поиска. Начиная поиск с пустого множества, поочередно проводится оценка каждого из атрибутов на целевом классификаторе. После

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

выбора лучшего атрибута он добавляется в подмножество. Для оставшихся атрибутов алгоритм повторяется до тех пор, пока каждый из них не будет добавлен в подмножество. В результате прямого поиска получаем атрибуты классификации, отсортированные от лучшего к худшему.

Как правило, оберточные алгоритмы показывают лучшие, по сравнению с фильтрующими результаты, при этом эти атрибуты классификации оптимизированы под сам классификатор. Однако с увеличением исходного набора данных и количества первоначальных атрибутов оберточный алгоритм будет значительно более трудоемким.

Кроме того, при использовании обёрточных методов, отобранные атрибуты подвержены переобучению.

Фильтрующие методы, напротив, используют основные характеристики данных для оценки атрибутов - действуя тем самым независимо от целевого алгоритма. В методах отбора атрибутов на основе фильтрации происходит сопоставление с классом и соответствующим подмножеством признаков.

Типичным методом отбора атрибутов на основе фильтрации является sequential forward floating selection (SFFS), который находит наилучшее аппроксимирующее решение по количеству отобранных функций. SFFS начинается с пустого пула атрибутов и, используя локальный оптимальный отбор признаков в два этапа увеличивает пул, включая этап включения и этап условного исключения. Эвристическая основа алгоритма SFFS заключается в предположении, что критерий отбора является монотонным с изменением размера и информацией набор. SFFS аппроксимирует оптимальное решение при доступной вычислительной стоимости.

Алгоритм InfoGain

Алгоритм выбора признаков на основе информационного выигрыша InfoGain является одним из самых простых и быстрых алгоритмов выделения признаков [4]. Алгоритм часто используется при решении задачи категоризации текста, где размерность данных не позволяет использовать более сложные методы выделения признаков. Работа метода основана на вычислении энтропии рассматриваемого класса до и после применения атрибута.

Так, если A - это признак, а C - рассматриваемый класс, то энтропия до наблюдения признака оценивается выражением:

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c) \quad (1)$$

а после наблюдения:

$$H(C \mid A) = -\sum_{c \in C} p(a) \sum_{c \in C} p(c \mid a) \log_2 p(c \mid a). \quad (2)$$

Изменение энтропии за счет применения атрибута характеризует информационный выигрыш [4]. Каждому атрибуту a из множества атрибутов A присваивается оценка, основанная на информационном выигрыше между ним самим и классом:

$$IG_i = H(C) - H(C \mid A_i) = H(A_i) - H(A_i \mid C) = H(A_i) + H(C) - H(A_i \mid C) \quad (3)$$

Результатом работы алгоритма InfoGain является ранжирование признаков по их значимости.

Использование корреляционной меры позволяет оптимизировать отбор признаков. В результате, метод фокусируется на двух проблемах: критерии корреляционной меры и алгоритме отбора признаков. В качестве критерия корреляционной меры могут использоваться коэффициент корреляции Пирсона, критерий взаимной информации и другие соответствующие критерии.

Типичным алгоритмом выбора признаков с использованием корреляционной меры является алгоритм CFS (Correlation-based Future Selection) - алгоритм выбора признаков на основе корреляции [6].

Алгоритм CFS [5] является одним из первых, который производит оценку множества признаков, а не каждого признака по отдельности. В основе алгоритма лежит оценка множества атрибутов, учитывающая полезность каждого независимого признака в определении класса, и корреляцию между ними:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \quad (4)$$

где $Merit_s$ - оценка качества подмножества S содержащего k признаков; $\overline{rf_{cf}}$ - средняя корреляция «признак-класс»; а $\overline{rf_{ff}}$ - средняя корреляция между признаками данного подмножества.

Числитель выражения (4) представляет собой метрику качества данного подмножества признаков, а знаменатель то, насколько излишняя информация в нем содержится.

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

В результате, «плохие», или не имеющие ценности признаки будут отброшены за счет плохой оценки качества в данном подмножестве, а избыточные признаки из-за высокой корреляции с одним или более признаком в подмножестве.

Для применения оценки (4) следует произвести вычисления корреляции, или зависимости между атрибутами:

$$SU = 2.0 \times \left[\frac{H(X) + H(Y) + H(X,Y)}{H(X) + H(Y)} \right] \quad (5)$$

После вычисления матрицы корреляции CFS использует эвристический поиск для нахождения хорошего подмножества признаков.

ЗАКЛЮЧЕНИЕ

Таким образом, выбор атрибутов для классификации трафика компьютерной сети - это сложная задача, которая зависит от цели классификации. Если, например, классификация совершается в режиме реального времени для мониторинга уровня QoS, можно проверить только часть некоторых отдельных потоков из сети, что дает больше возможностей, чем классификация трафика для учета данных. Тем не менее, в обоих случаях можно использовать параметры, основывающиеся на размерах пакетов. В тоже время следует избегать использования каких-либо параметров, основывающихся на времени (продолжительность потока и другие параметры). В дальнейшем, как правило, полагается, что характеристики потоков, основывающихся на размере пакетов в сети, не зависят от текущих условий. Другие атрибуты содержат имя протокола транспортного уровня, удаленные номера портов, количество пакетов, имеющих набор TCP флагов или номера. Количество атрибутов для классификации трафика не так важно, как для техника кластеризации. Тем не менее, важно избегать атрибутов, которые содержат конкретные значения только для небольшого количества случаев, принадлежащих к конкретному классу. Это может привести к переобучению классификатора и, также, не будет возможности идентифицировать неизвестные случаи.

ISSN: 2181-1385
Scientific Journal Impact Factor (SJIF) 2021: 5.723
Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89
DOI: 10.24412/2181-1385-2021-12-886-897

REFERENCES

1. Aamir, M., Zaidi, M.A., 2013. A survey on DDoS attack and defense strategies: from traditional schemes to current techniques. Interdisciplinary Inf. Sci. 19(2), 173-200.
2. Aamir, M., Zaidi, S.M.A., 2015. Denial-of-service in content centric (named data) networking: a tutorial and state-of-the-art survey. Security Commun. Networks 8 (11), 2037-2059.
3. Beitollahi, H., Deconinck, G., 2012. Analyzing well-known countermeasures against distributed denial of service attacks. Comput. Commun. 35 (11), 1312-1332.
4. Berkhin, P., 2006. "A survey of clustering data mining techniques". In: Grouping Multidimensional Data. Springer, pp. 25-71.
5. Boroujerdi, A.S., Ayat, S., 2013. "A, robust ensemble of neuro-fuzzy classifiers for DDoS attack detection", in. In: Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, pp. 484-487.
6. Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5-32. "Content Delivery Network (CDN) and Cloud Computing Services | Akamai." [Online].
7. Available: https://content.akamai.com/us-en-PG11224-summer-2018- sotiweb-attack-report.html. Data Mining. [Online]. Available: http://www.stat.cmu.edu/ ryantibs/datamining/.
8. Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 2, 224-227.
9. Du, Q., Fowler, J.E., 2008. Low-complexity principal component analysis for hyperspectral image compression. Int. J. High Perf. Comput. Appl. 22 (4), 438-448. 8 Bulletin of National University of Uzbekistan: Mathematics and Natural Sciences
10. Fitriani S., Mandala S., Murti M.A. "Review of semi-supervised method for Intrusion Detection System" in Multimedia and Broadcasting (APMediaCast) Asia Pacific Conference on, 2016, 2016, 36-41.
11. Gao Y., Feng Y., Kawamoto J., and Sakurai K., "A machine learning based approach for detecting DRDoS attacks and its performance evaluation," in Information Security (AsiaJCIS), 2016 11th Asia Joint Conference on, 2016, pp. 8086.
12. Gu, Y., Wang, Y., Yang, Z., Xiong, F., Gao, Y., 2017. Multiple-features-based semisupervised clustering DDoS detection method. Math. Problems Eng. 2017.
13. Hu, X., Knysz, M., Shin, K.G., 2011. "Measurement and analysis of global IPusage patterns of fast-flux botnets", in INFOCOM. Proc. IEEE 2011, 26332641.
14. Idhammad, M., Afdel, K., Belouch, M., 2018. "Semi-supervised machine learning approach for DDoS detection,". Appl. Intell., 1-16
15. Jollife, I., 2011. "Principal component analysis". In: International Encyclopedia of Statistical Science. Springer, pp. 1094-1096