ФИЛЬТРАЦИЯ НЕЖЕЛАТЕЛЬНЫХ ПРИЛОЖЕНИЙ ИНТЕРНЕТ-ТРАФИКА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА КЛАССИФИКАЦИИ RANDOM FOREST

Шелухин О.И. [1]***, Ванюшина А.В.*** [2]***, Габисова М.Е.*** [3]

Рассмотрена актуальная задача контроля доступа к Интернет–ресурсам имеющая важное прикладное значение: блокирование доступа к нелегальной, экстремистской, антисоциальной информации, предотвращение утечки конфиденциальной информации через Интернет и др. Для решения подобных задач широкое распространение получили методы машинного обучения. Одним из наиболее часто используемых и эффективных для классификации сетевого трафика методов машинного обучения является «случайный лес» (Random Forest), представляющий собой ансамблевый метод, действующий путем построения множества решающих деревьев. Для оценки эффективности работы алгоритма Random Forest (RF) при классификации сетевого трафика по типам прикладных протоколов, работающих в сети Интернет, был осуществлён сбор трафика в сети, состоящей из нескольких компьютеров. Исследовались приложения, генерирующие пакеты, относящиеся к различным протоколам прикладного уровня : BitTorrent, DNS, HTTP, SSL, Skype, Steam. После отбора информационных признаков и предварительной обработки данных сформированы обучающая и две тестовых выборки, одна из которых содержала фоновый трафик. В результате применения алгоритма классификации RF к полученным данным найдены оценки эффективности работы данного алгоритма в условиях наличия и отсутствия фонового сетевого трафика

Ключевые слова*: машинное обучение, решающие деревья, фоновый трафик, протокол, обучающая и тестовые последовательности, метрики, эффективность.*

DOI: 10.21681/2311-3456-2018-2-44-51

Введение

Проблема контроля доступа к Интернет-ресурсам актуальна и имеет важное прикладное значение по следующим основным причинам: блокирование доступа к нелегальной (экстремистской, антисоциальной и т.п.) информации, предотвращение доступа к Интернет-ресурсам в личных целях в учебное или рабочее время, предотвращение утечки конфиденциальной информации через Интернет.

Вредоносные программы и атаки обычно используют непроверяемый канал зашифрованного трафика HTTPS. Не соответствующее политике или нежелательное поведение пользователей.

Первая задача, которая встает перед администраторами, это определить, какой тип сетевого трафика генерируется пользователями. Трафик может быть вредоносным (например, кража данных или разведка сети), неприемлемым и нарушающим политику (например, использование служб обмена файлами) или выходящим за рамки обычных бизнес-процессов (например, генерирование трафика в нерабочее время). Приложения, соответствующие вредоносному трафику, называют нежелательными.

Это могут быть потенциально опасные приложения. У разных сетевых приложений (для использования социальных сетей, служб обмена мгновенными сообщениями, служб обмена файлами, одноранговых служб и др.) разные риски безопасности. Они могут ставить под угрозу данные и системные активы, влиять на производительность труда сотрудников и использовать пропускную способность сети.

Таким образом проблема контроля доступа к Интернет ресурсам актуальна и имеет важное значение по следующим основным причинам:

- блокирование доступа к нелегальной (экстремистской, антисоциальной и другой) информации;
- предотвращение использования Интернет ресурсов не по назначению, в частности, ограничение и контроль доступа к развлекательным и другим ресурсам для личного пользования;
- предотвращение утечки конфиденциальной информации через Интернет.

На сегодняшний день существует множество как коммерческих, так и некоммерческих продуктов, решающих подобные задачи. К наиболее распространённым коммер-

1 Шелухин Олег Иванович, доктор технических наук, профессор, МТУСИ, Москва, Россия. E-mail: sheluhin@mail.ru
2 Ванюшина Анна Вячеславовна, старший преподаватель, МТУСИ, Москва, Россия. E-mail: vanuanna@rambler.ru
3 Габисова Мария Евгеньевна, магистрант, МТУСИ, Москва, Россия. E-mail: mgabisova@yandex.ru

ческим продуктам можно отнести: WebSense (https://www.forcepoint.com), NetNanny (https://www.netnanny.com/) и множество других. Среди open-source решений следует отметить Poesia . Основные количественные показатели при оценке работы систем фильтрации Интернет-трафика следующие: точность анализа – процент верно отфильтрованных Интернет-ресурсов; излишнее блокирование или ложноположительные ошибки – процент «хороших» ресурсов, ошибочно запрещенных системой фильтрации.

Для решения подобных задач в настоящее время широкое распространение получили методы, основанные на технологиях математической статистики и машинного обучения, с помощью которых даже неизвестные вредоносные приложения могут быть детектированы с определенной степенью вероятности [1, 2, 3].

Такие методы позволят разрабатываемой системе легко адаптироваться к постоянно изменяющейся природе Интернет ресурсов и учитывать специфику анализа сетевого трафика.

Одним из наиболее часто используемых и эффективных для классификации сетевого трафика методов машинного обучения является решающее дерево [4, 5].

Случайный лес (Random Forest) представляет собой ансамблевый метод обучения для классификации и регрессии, который действуют путем построения множества решающих деревьев [6].

Целью работы является оценка эффективности работы алгоритма Random Forest (RF) в задачах классификации приложений в условиях наличия и отсутствия фонового сетевого трафика.

Исходные данные

Для сбора необходимых для анализа данных была организована лабораторная сеть из нескольких компьютеров. Один из компьютеров был подключен к глобальной сети интернет и на его базе была организована беспроводная точка доступа. Схематичное изображение используемой сети приведено на рисунке 1.

На этом же компьютере осуществлялся захват всего проходящего через него траффика с помощью программы Wireshark (https://www.wireshark.org). На остальных компьютерах, подключенных к точке доступа были запущены различные приложения. Осуществлялся просмотр веб-страниц с помощью браузеров Google Chrome и Opera. Производились видеозвонки с помощью программы Skype, скачивание файлов с помощью торрент клиента µTorrent, использование сервиса цифрового распространения компьютерных игр Steam и т.д. По завершении захвата, полученные данные были сохранены в формате PCAP.

Для приведения полученных данных в соответствие с требованиями решаемой задачи, была произведена предварительная обработка данных. В результате все пакеты были разделены на потоки транспортного уровня, идентифицируемые пятёркой значений: протокол транспортного уровня (TCP или UDP), IP-адрес источника, порт источника, IP-адрес получателя, порт получателя. Затем данные были помечены, т.е. каждому потоку был поставлен в соответствие протокол/приложение, к которому этот поток относится. После чего полученный набор данных

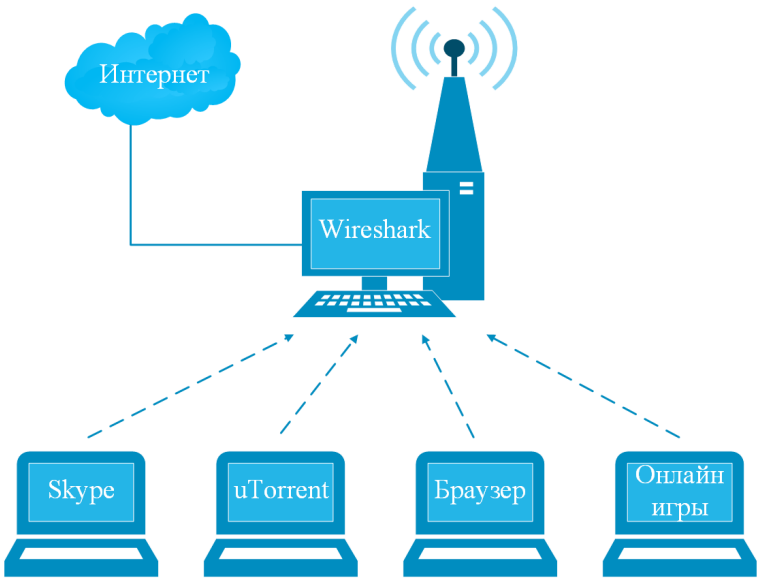

***Рис. 1**. Схема используемой сети*

Таблица 1.
Состав полученной выборки данных

Протокол	Обучающая выборка	Тестовая выборка
SSL	1215	295
HTTP	1091	272
DNS	1061	267
BitTorrent	940	232
Steam	775	204
Skype	645	162

был поделен на две подвыборки - обучающую и тестовую. Состав полученного датасета приведён в таблице 1.

С помощью встроенного в RF алгоритма отбора информационных признаков Feature Importance [7], были отобраны следующие 11 признаков (табл. 2).

Методологи я решения задачи классификации с помощью алгоритма *RF*

Алгоритм *RF* опирается на технику бэггинга - использования композиции независимо обучаемых алгоритмов. В результате, строится множество решающих деревьев, каждое из отдельного случайного подмножества исходной выборки данных, причём размер подвыборок совпадает с размером исходной выборки и имеет повторения. Для k-го дерева генерируется случайный вектор θ_k, который не зависит от сгенерированных ранее векторов $\theta_1, \dots, \theta_{k-1}$, но имеет такое же распределение. Дерево «выращивается» с применением тренировочной выборки и вектора θ_k, в результате чего образуется классификатор $h(x, \theta_k)$, где x - входной вектор.

Деревья строятся с помощью стандартного алгоритма бинарного решающего дерева [7].

Пусть имеется некоторая обучающая выборка $T^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$, где x_i - векторы информационных признаков объекта. Множество всех возможных значений векторов признаков X называется пространством образов (объект - точка в пространстве образов). Состояния зависимой переменной (метки классов) y_i могут принимать только конечное число значений. С каждой вершиной t дерева связаны: некоторое подмножество пространства образов $X_t \subset X$ (с корневой вершиной связывается всё пространство образов X); подвыборка $T_t^m \subset T^m$обучающей выборки (с корневой вершиной связывается вся обучающая выборка T^m); некоторое решающее правило $f_t: X \to \{0,1, \dots, k_t - 1\}$, где $k_t \geq 2$ - количество потомков вершины t (для бинарного дерева $k_t = 2$), это правило определяет разбиение множества

Таблица 2.
Отобранные информационные признаки

No	Название	Описание
1	src_port	номер порта источника (источником считается отправитель первого пакета
2	dst_port	номер порта получателя
3	max_src_data_ip	максимальный размер данных в пакете от источника
4	min_src_data_ip	минимальный размер данных в пакете от источника
5	med_src_data_ip	медианный размер данных в пакете от источника
6	prop_src_data_ip	доля данных, переданных источником в общем количестве данных потока
7	max_dst_data_ip	максимальный размер данных в пакете от получателя
8	mean_dst_data_ip	средний размер данных в пакете от получателя
9	src_to_dst_ratio_data_ip	отношение размера данных, переданных источником к размеру данных, переданных получателем
10	min_data_ip	минимальное значение размера данных в потоке
11	var_data_ip	среднеквадратическое отклонение значения размера данных в потоке

X на k непересекающихся подмножеств (терминальные вершины (листы) такого правила не имеют). Чаще всего в качестве решающего правила берется один из признаков $x_{i(t)}$. Обозначим $t_{i(t)}$ вершину, являющуюся i-тым потомком вершины t. Множество X_t и функция f_t задают множества:

$$X_{t_{i(t)}} = X_t \cap \{x \in X: f_t(x) = i\} \quad (1)$$

В результате каждому внутреннему узлу соответствует один из входных атрибутов, терминальным вершинам соответствуют метки классов. Целью построения решающего дерева является классификация векторов x.

Атрибуты для каждого дерева выбираются из обучающей выборки случайно. Атрибуты, относящиеся к более, чем двум классам, могут быть отобраны больше одного раза для разных узлов.

Расщепление в промежуточном узле при построении дерева необходимо производить так, чтобы загрязнённость была минимальной [9].

Пусть $D(t)$ - некоторая подвыборка, связанная с вершиной t. Загрязнённость вершины $i(t)$ будет равна 0, если подвыборка $D(t)$ содержит экземпляры только одного класса, и будет максимальной при одинаковом количестве экземпляров каждого класса. В результате, количество экземпляров, принадлежащих другим классам (примесей), в каждом классе после разбиения должно стремиться к минимуму. Существует несколько мер загрязнённости вершины. Самыми популярными являются энтропийный критерий и критерий Джини.

Критерий основан на понятии количества информации, которую содержит расщепление. Энтропию i-го узла можно вычислить воспользовавшисьсоотношением$i(t) = -\sum_{i=1}^{c} P(\omega_i) log_2 P(\omega_i)$, где $P(\omega_i)$ - доля экземпляров класса ω_i в $D(t)$, c - количество классов.

Критерий Джини определяемый по формуле $i(t) = 1 - \sum_{i=1}^{c} P^2(\omega_i)$, оценивает «расстояние» между распределениями классов.

Ещё одна, менее используемая, мера загрязнённости основана на частоте ошибок классификации, определяемая как минимальная вероятность того, что экземпляр будет классифицирован неверно: $i(t) = 1 - \max_j P(\omega_j)$.

Чем меньше критерий расщепления, тем лучше качество расщепления.

На практике важно, чтобы загрязнённость уменьшалась при переходе от узла к его потомкам [7]. Определим уменьшение загрязнённости на узле t как

$\Delta i(t) = i(t) - P_L i(t_L) - P_R i(t_R)$, где P_L и P_R - доли экземпляров левого (t_L и правого (t_R) потомков узла t соответственно. Заметим, что формула справедлива только для случая бинарного дерева. Наилучшим расщеплением считается то, при котором величина $\Delta i(t)$ максимальна. Как правило при построении дерева понятие максимальной $\Delta i(t)$ не является точным, поскольку при выборе оптимального расщепления не осуществляется полный перебор всех возможных вариантов, а лишь производится «сужение» набора до нескольких вариантов, из которых затем и выбирается тот, при котором $\Delta i(t)$ принимает наибольшее значение.

Если выращивать полное дерево, пока в листовых вершинах не будет достигнута минимально возможная загрязнённость, произойдёт переобучение модели, т.е. она просто «запомнит» все варианты классификации для тренировочной выборки и не будет способна к работе на тестовых данных. В пределе каждая листовая вершина будет отвечать за один классифицируемый экземпляр. Если же остановить расщепление слишком рано, окажется довольно высокой ошибка и будет страдать эффективность. Поэтому важной задачей является построение сбалансированного дерева, для чего необходим выбор правильного критерия остановки расщепления. Таких критериев существует несколько. Один из них - задание минимального порога на число экземпляров в листовых вершинах. Как только количество экземпляров в данной вершине становится меньше этого порога, расщепление останавливается, и вершина считается терминальной. Ещё один метод - выбор предельно допустимого значения параметра $\Delta i(t)$. Расщепление вершины не происходит, если после него уменьшение загрязненности меньше заданного порога. Можно также просто ограничивать глубину дерева [8,9]. В этом случае построение дерева заканчивается, если достигнута заданная глубина. Весьма распространённым методом является кросс-валидация , при которой происходит оценка ошибки классификации на тренировочном множестве и тестовом. Расщепление происходит до тех пор, пока ошибка минимальна.

Следует обратить внимание на то, что использование ранней остановки расщепления обладает важным недостатком: решение об остановке расщепления принимается без учёта случая, при котором оно могло бы быть продолжено. То есть, даже если продолжение расщепления сильно повысило бы эффективность классификации, такая ситуация не рассматривается. Существует иной подход

Таблица 3.
Матрица ошибок для тестовой выборки

предсказ. реальн.	SSL	HTTP	DNS	BitTorrent	Steam	Skype
SSL	295	0	0	0	0	0
HTTP	0	267	0	4	1	0
DNS	0	0	266	1	0	0
BitTorrent	1	0	0	230	0	1
Steam	0	3	0	0	201	0
Skype	6	0	0	1	0	155

к уменьшению деревьев – прунинг (pruning, отсечение ветвей) полных решающих деревьев, когда промежуточные узлы заменяются на терминальные и относятся к превалирующему в поддереве классу. В RF прунинг не используется, так как имеет высокую вычислительную сложность.

После построения всех деревьев лес организуется как самый простой ансамблевый классификатор. Каждое дерево голосует за ожидаемый класс и экземпляр определяется в класс, набравший наибольшее количество голосов по всем деревьям в лесу.

RF имеет ряд преимуществ: низкое число управляющих параметров и параметров модели; устойчивость к переобучению; не требуется отбор признаков, потому что он могут использовать большое количество потенциальных атрибутов. Одним из важных преимуществ RF является то, что дисперсия модели уменьшается с увеличением количества деревьев в лесу, в то время как смещение остается тем же самым. Алгоритм RF также имеет некоторые недостатки, такие как низкая интерпретируемость, потери производительности из-за коррелированности переменных, и зависимость от генератора случайных чисел.

Результаты исследования

В эксперименте было проведено построение случайного леса и оценка качества классификации на заданной выборке. Опытным путём были подобраны наиболее приемлемые параметры алгоритма. Лес состоит из 5 деревьев с максимально возможной глубиной.

В таблице 3 представлена матрица ошибок для чистой тестовой выборки. По вертикали указаны реальные значения, по горизонтали - предсказанные обученной моделью.

Для определения эффективности алгоритма используются следующие метрики: точность, полнота, F1-мера, значения которых легко рассчитать на основании матрицы ошибок классификации (табл. 4), составляемой для каждого класса отдельно.

В матрице отображается количество правильных и не правильных решений по заданного класса. TP (*True Positive*) обозначает истинно-положительное решение, TN (*True Negative*) - истинно-отрицательное решение, FP (*False Positive*) - ложно-положительное, а FN (*False Negative*) - ложно-отрицательное решение.

Точность $Precision = \frac{TP}{TP+FP}$ - доля правильно классифицированных единиц данного класса относительно всех экземпляров, которые алгоритм отнёс к данному классу. Полнота $Recall = \frac{TP}{TP+FN}$ - доля правильно классифицированных единиц относительно всех экземпляров, относящихся к данному классу и

F1-мера вычисляемая по формуле $F1 = \frac{Precision \cdot Recall}{Precision+Recall}$.

Графическое представление данных метрик полученных экспериментально для всех анализируемых классов приведено на рисунке 2.

Видно, что наибольшую эффективность алгоритм имеет место для данных, относящихся к DNS трафику.

Таблица 4.
Матрица ошибок классификации

	Реальный класс: X	Реальный класс: не X
Предсказанный класс: X	TP	FP
Предсказанный класс: не X	FN	TN

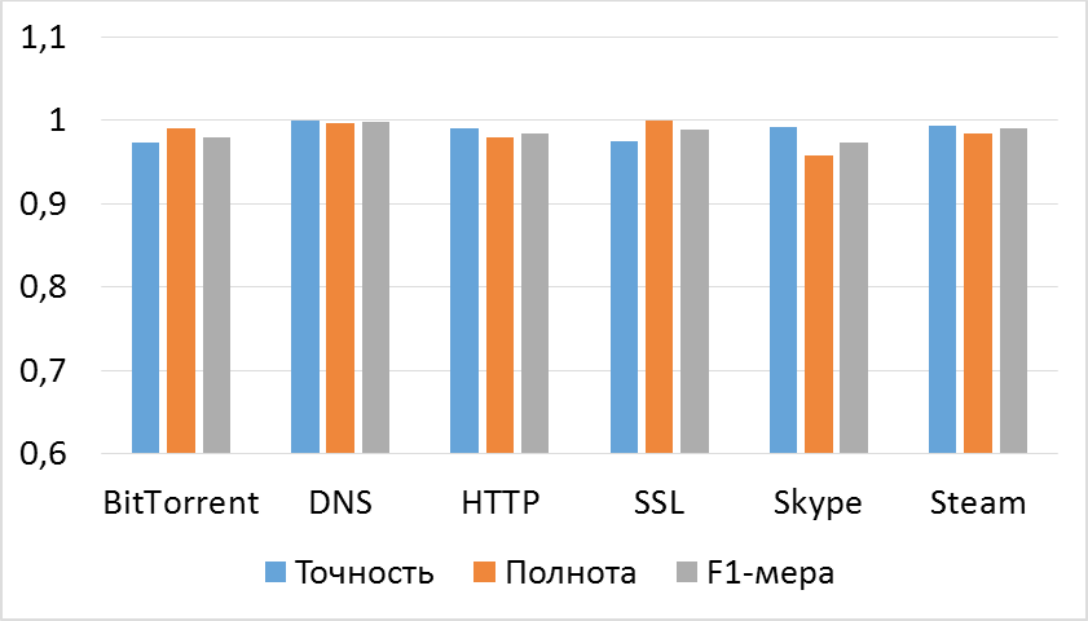

***Рис. 2**. Точность, полнота, F1-мера для тестовой выборки*

Помимо проверки работы алгоритма на тестовой выборке, имеющей такой же классовый состав, как и обучающая, оценка его качества производилась также в условиях присутствия фонового трафика, т.е. в случае когда в тестовой выборке присутствовали экземпляры классов, отсутствующих в обучающей выборке .

Состав этой выборки приведён в таблице 5.

Такая ситуация, когда в классифицируемых данных присутствует фоновый трафик, более приближена к действительности, ведь используемые в сети Интернет протоколы очень многообразны [8,9]. Такой *DataSet* позволяет получить оценку работы алгоритма в реальных условиях. В таблице 6 представлена матрица ошибок для данного случая.

Как видно из таблицы, все экземпляры, относящиеся к классу LLMNR модель классифицировала как SSL, все экземпляры RTP были отнесены

Таблица 5.
Состав тестовой выборки данных с примесями

Протокол	Количество потоков
SSL	295
HTTP	272
DNS	267
BitTorrent	232
Steam	204
Skype	162
LLMNR	169
Quic	95
RTP	19

Таблица 6.
Матрица ошибок для тестовой выборки при наличии фонового трафика

предсказ. реальн.	SSL	HTTP	DNS	BitTorrent	Steam	Skype	LLMNR	Quic	RTP
SSL	295	0	0	0	0	0	0	0	0
HTTP	0	267	0	4	1	0	0	0	0
DNS	0	0	266	1	0	0	0	0	0
BitTorrent	1	0	0	230	0	1	0	0	0
Steam	0	3	0	0	201	0	0	0	0
Skype	6	0	0	1	0	155	0	0	0
LLMNR	169	0	0	0	0	0	0	0	0
Quic	0	0	25	9	0	61	0	0	0
RTP	0	0	0	0	0	19	0	0	0

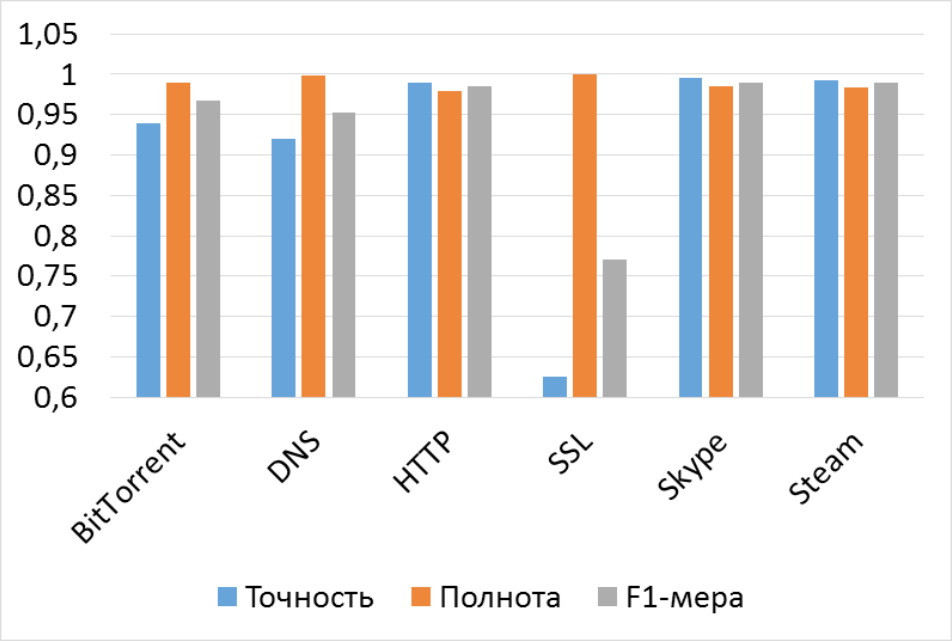

***Рис. 3**. Точность, полнота, F1-мера при наличии фонового трафика*

к Skype, а экземпляры класса Quic в основном были разделены между классами DNS и Skype. Рассмотрим, как изменились показатели качества классификации представленные на рисунке 3.

Как видно, наличие фонового трафика практически не повлияло на значение полноты, но значительно ухудшило точность классификации, поскольку увеличилось количество False Positive экземпляров вызванных наличием фонового трафика, принадлежащего к классам, которые в обучении не участвовали.

Выводы

Для анализа эффективности алгоритма RF в задачах классификации сетевого трафика был собран *DataSet*, содержащий потоки, относящиеся к различным протоколам прикладного уровня: *BitTorrent, DNS, HTTP, SSL, Skype, Steam*. Произведена оценка работы алгоритма RF как на «чистой» тестовой выборке, так и на выборке, содержащей фоновый трафик в виде экземпляров, относящихся к классам неизвестным обученному алгоритму. Показано, что наличие примесей существенно влияет на точность классификации выполненной с помощью алгоритма RF . Алгоритм RF продемонстрировал высокую эффективность в режиме off-line, о чем в частности свидетельствует F1-мера равная соответственно 0,987 и 0,759 при отсутствии и наличии фонового трафика.

Наличие фонового трафика принадлежащего к классам, которые в не участвовали в обучении алгоритма, значительно ухудшает точность классификации. Алгоритм RF мало пригоден для классификации в режиме реального времени из-за временной сложности обработки, оцениваемой соотношением $Mmnlog_{0}^{n}$, где n - количество экземпляров, m - количество информационных признаков, а M - количество деревьев.

***Рецензент**: Басараб Михаил Алексеевич, доктор физико-математических наук, профессор, МГТУ им. Н.Э. Баумана, Москва, Россия. E-mail: bmic@mail.ru*

Литература:

1. Щербакова Н.Г. Анализ IP-трафика методами Data Mining. Проблема классификации // Проблемы информатики. 2012. № 4 . С. 30–46.
2. Шелухин О. И., Симонян А.Г., Ванюшина А.В. Эффективность алгоритмов выделения атрибутов в задачах классификации приложений при интеллектуальном анализе трафика. Электросвязь, 2016. №11. Стр. 79-85.
3. Шелухин О.И., Симонян А.Г., Ванюшина А.В. Влияние структуры обучающей выборки на эффективность классификации приложений трафика методами машинного обучения // T-Comm: Телекоммуникации и транспорт. 2017. №2 (11). С. 25-31.
4. Костин Д.В., Шелухин О.И., Сравнительный анализ алгоритмов машинного обучения для проведения классификации сетевого зашифрованного трафика //T-Comm: Телекоммуникации и транспорт. 2016. № 9 (10) С.46-52
5. Sheluhin O.I., Simonyan A.G., Vanyushina A.V. (2017). Benchmark data formation and software analysis for classification of traffic applications using machine learning methods. T-Comm, vol. 11, no.1, pp. 67-72.

6. Глухова А.И. Сущность метода принятия управленческих решений «дерево решений»// Master's Journal. 2014. № 2. С. 316-321.
7. Witten, I. H. (Ian H.) Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. – 2nd ed. p. cm. – (Morgan Kaufmann series in data management systems), San Francisco 2005. P. 525. ISBN: 0-12-088407-0
8. En-Najjary T, Urvoy-Keller G., Pietrzyk M., and Costeux J.-L.Application-based feature selection for internet traffic classification. In Teletraffic Congress (ITC), 2010, 22nd International, pages 1 - 8, 2010.
9. Pietrzyk M., En-Najjary T., Urvoy-Keller G., and Costeux J.-L.. Hybrid traffic identification. Technical Report EURECOM+3075, Institut Eurecom, France, 04 2010.

THE FILTERING OF UNWANTED APPLICATIONS IN INTERNET TRAFFIC USING RANDOM FOREST CLASSIFICATION ALGORITHM

Sheluhin O.[4] , Vanyushina A.[5], Gabisova M.[6]

An actual task of a significant application value, namely, access control to the Internet resources has been examined: access blocking to illegal, extremist, antisocial information, confidential information leaks through the Internet prevention, etc. To solve the above mentioned problems, machine learning methods have prevailed. One of the most frequently used and effective methods of machine learning for network traf–fic classification is a Random Forest method, that is ensemble and operates by constructing a multitude of decision trees. In order to estimate the efficiency rate at which Random forest algorithm classifies network traffic based on types of application protocols functioning in the Internet, a dump of a small network traffic has been created. The applications, generating packets, related to different application level protocols (Bit–Torrent, DNS, HTTP, SSL, Skype, Steam) have been investigated. After pinpointing informational signs and preliminary data processing, learning and two test sets, one of which contained background traffic, have been formed. After applying Random forest classification algorithm to the resulting data, estimations of efficiency of this algorithm both with and without background traffic have been performed. It has been shown, that the algorithm efficiency significantly decreases when background traffic is present.

Keywords*: machine learning, decision trees, background traffic, protocol, training and test sequences, metrics, efficiency.*

References

1. Sherbakova N.G. Analiz IP-trafic metodami Data Mining. Problema klassifikatzii // Problemi informatiki. 2012. № 4. p. 30–46
2. Sheluhin O.I., Simonyan A.G., Vanyushina A.V. Effectivnost algoritmov videlenia atributov v zadathax klassifikatzii prilogeniy pri intellectualnom analize trafica. // Electrosviaz. 2016. №11. pp. 79-85
3. Sheluhin O.I., Simonyan A.G., Vanyushina A.V. Vliyanie structure obuthayshei viborki na effectivnost klassificatii prilozeniy trafika metodami mashinnogo obuthenia // T-Comm: Telekommunikation I transport. 2017. №2 (11). pp. 25-31.
4. Kostin D.V. Sheluhin O.I. Sravnitelniy analiz algoritmov mashinnogo jbuthenia dlya provedenia rlassifikatziy setevogo zashifrovannogo traffika // T- Comm : Telekommunikation I transport. 2016. № 9 (10), pp. 46-52
5. Sheluhin O.I., Simonyan A.G., Vanyushina A.V. (2017). Benchmark data formation and software analysis for classification of traffic applications using machine learning methods. // T-Comm, vol. 11, no.1, pp. 67-72.
6. Glukhova A.I. Suchnost metoda prinyatia upravlentheskix resheniy «derevo resheniy» // Master's Journal. 2014. № 2. С. 316-321.
7. Witten, I. H. (Ian H.) Data mining : practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. – 2nd ed. p. cm. – (Morgan Kaufmann series in data management systems), San Francisco 2005. P.525. ISBN: 0-12-088407-0.
8. En-Najjary T, Urvoy-Keller G., Pietrzyk M., and Costeux J.-L.Application-based feature selection for internet traffic classification. In Teletraffic Congress (ITC), 2010 22nd International, pages 1 - 8, 2010.
9. Pietrzyk M., En-Najjary T., Urvoy-Keller G., and Costeux J.-L.. Hybrid traffic identify cation. Technical Report EURECOM+3075, Institut Eurecom, France, 04 2010.

4 Oleg Sheluhin, Doctor of Technical Sciences, Head of department «Information security», Professor Moscow Technical University of Communication and Informatics, Moscow, Russia. E-mail: sheluhin@mail.ru

5 Anna Vanyushina, Moscow Technical University of Communication and Informatics, Senior lecturer of «Information security» department, Moscow, Russia. E-mail: vanuanna@rambler.ru

6 Mariya Gabisova, Master's degree student of «Information security» department Moscow Technical University of Communication and Informatics, , Moscow, Russia. E-mail: mgabisova@yandex.ru