УДК 004.7: 654.16

Фильтрация нежелательных приложений трафика подвижной радиосвязи для обнаружения угроз информационной безопасности

Шелухин Олег Иванович
доктор технических наук,
профессор, заведующий кафедрой «Информационная безопасность»
Московского технического университета связи и информатики[1].
E-mail: sheluhin@mail.ru.

Смычёк Михаил Александрович
кандидат технических наук,
главный специалист отдела проектирования сетей связи,
АО «Гипрогазцентр»[2].
E-mail: m-smychek@mail.ru.

Симонян Айрапет Генрикович
кандидат технических наук,
доцент кафедры "Информационная безопасность"
Московского технического университета связи и информатики[1].
E-mail: blackman-05@mail.ru.

[1]*Адрес:* 111024,г. Москва, ул. Авиамоторная, д. 8а.
[2]*Адрес:* 603950, г. Нижний Новгород, ул. Алексеевская, 26.

Аннотация: В статье исследуется эффективность алгоритмов классификации машинного обучения: Naive Bayes; C4.5; Random Forests; Support Vector Machine (SVM); One Rule; Adaptive Boost мобильных приложений Google Chrome, Instagram, Facebook, Messenger, Whatsapp трафика подвижной радиосвязи. Для оценки эффективности алгоритмов классификации использовались метрики: Precision, Recall, F-Measure, AUC. Показано, что среди алгоритмов классификации машинного обучения наилучшим является алгоритм Random Forest. Среди анализируемых приложений наиболее эффективно классифицируются приложения Google Chrome и Whatsapp. Наиболее сложными для автоматической классификации методами машинного обучения оказались приложения Facebook и Messenger. Представленные результаты предлагается использовать при выборе наилучших алгоритмов классификации, оценки объёма обучающей и тестирующей выборок для достижения высоких показателей качества классификации в условиях появления неконтролируемого (фонового) трафика; обеспечить возможность создания технологий законного перехвата сетевого трафика по аналогии с телефонными компаниями основываясь на технологиях классификации сетевого трафика.

Ключевые слова: Алгоритмы, классификация, машинное обучение, мобильные приложения, пакет, поток, приложение, протокол, сетевой трафик, сеть, эффективность.

Постановка задачи

Согласно статистике [1], собранной в декабре 2017 года, около 66% всего сетевого трафика генерируется мобильными устройствами (смартфонами и планшетами). Проблема контроля доступа к мобильным приложениям Интернет-ресурсов 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.

Проблема определения сотовым оператором, какими приложениями воспользовался тот или иной пользователь сети актуальна для составления статистики наиболее часто ис-

пользуемых приложений. Подобное определение статистики приложений помогает не только отслеживать состояние сети, выявлять сбои, но и при необходимости ограничивать доступ к сетевым ресурсам, которые с точки зрения информационной безопасности могут нанести вред пользователю.

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 определённой степенью вероятности [2,3,4].

Такие методы позволят разрабатываемой системе легко адаптироваться к постоянно изменяющейся природе Интернет–ресурсов и учитывать специфику анализа сетевого трафика. Одними из наиболее часто используемых и эффективных для классификации сетевого трафика являются методы машинного обучения.

Внедрение предлагаемых подходов позволит производить классификацию, анализ и фильтрацию сетевого трафика вредоносных и нежелательных приложений, с более высокой эффективностью и в соответствии с предложенными показателями, а также в сравнении с другими методами классификации сетевого трафика, такими, как Deep Packet Inspection (DPI) или анализ номеров портов.

Вредоносные приложения могут представлять собой угрозу целостности или доступности данных, а нежелательные – угрозу конфиденциальности.

Целью работы является исследование эффективности алгоритмов классификации нежелательных мобильных приложений трафика подвижной радиосвязи для обнаружения угроз безопасности.

Захват и анализ сетевого трафика мобильных приложений

Для измерения трафика мобильного устройства можно использовать одну из трёх архитектур:

- перехват трафика при помощи прокси-сервера;
- перехват трафика непосредственно с сетевого адаптера мобильного устройства;
- перехват трафика с устройства, которое является для мобильного телефона точкой доступа в сеть Интернет.

Для измерения трафика был выбран последний метод, позволяющий минимизировать фоновый трафик и не требующий создания VPN-соединений. Общая архитектура изображена на рис. 1. В качестве программного решения для захвата пакетов и их обработки использовался анализатор трафика (сниффер) Wire Shark [5]. Для осуществления классификации, основанной на методах машинного обучения, использован подход, базирующийся на классификации сетевых потоков. Сетевым потоком называется набор сетевых пакетов, передающихся между двумя сетевыми узлами,

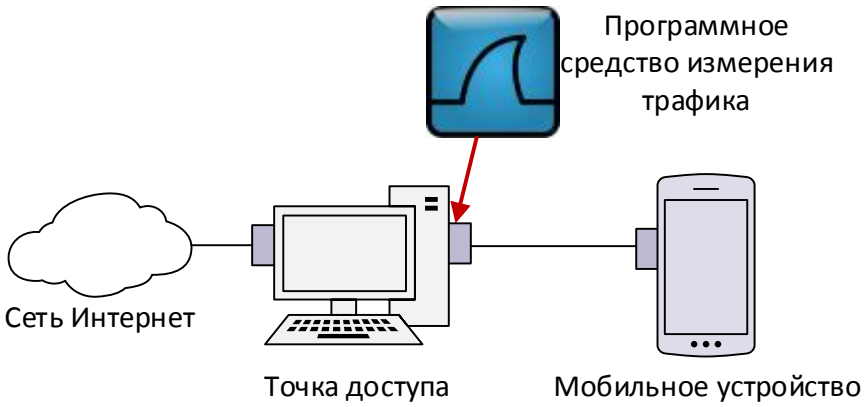
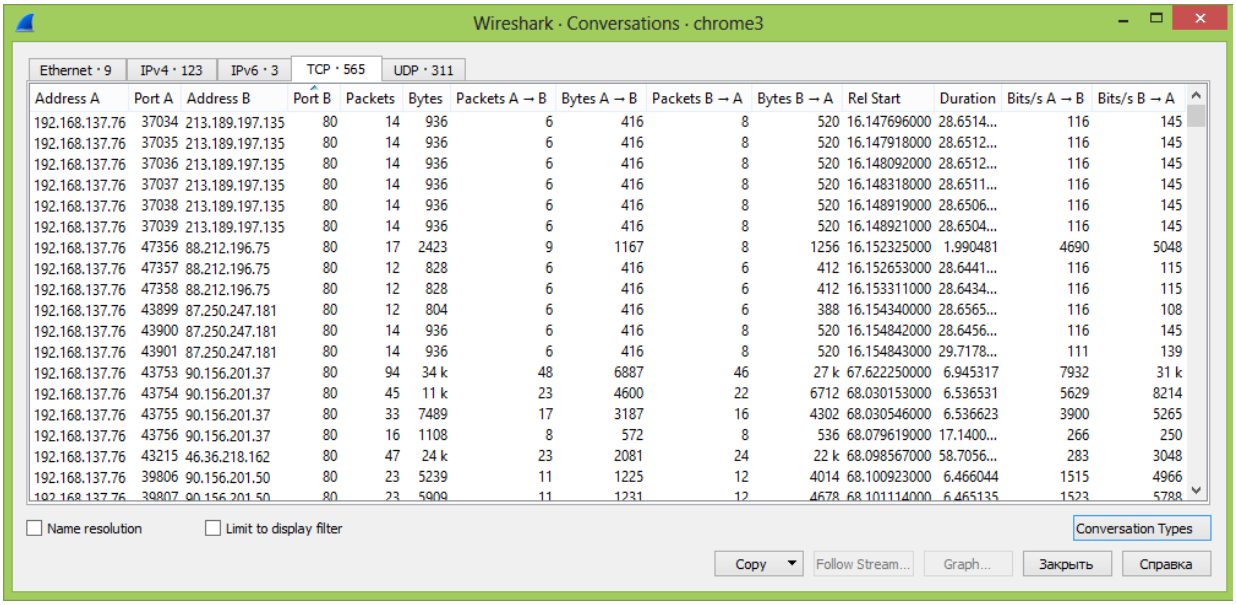


Рис. 1. Измерение трафика через точку доступа в сеть Интернет



Address A	Port A	Address B	Port B	Packets	Bytes	Packets A → B	Bytes A → B	Packets B → A	Bytes B → A	Rel Start	Duration	Bits/s A → B	Bits/s B → A
192.168.137.76	37034	213.189.197.135	80	14	936	6	416	8	520	16.147696000	28.6514...	116	145
192.168.137.76	37035	213.189.197.135	80	14	936	6	416	8	520	16.147918000	28.6512...	116	145
192.168.137.76	37036	213.189.197.135	80	14	936	6	416	8	520	16.148092000	28.6512...	116	145
192.168.137.76	37037	213.189.197.135	80	14	936	6	416	8	520	16.148318000	28.6511...	116	145
192.168.137.76	37038	213.189.197.135	80	14	936	6	416	8	520	16.148919000	28.6506...	116	145
192.168.137.76	37039	213.189.197.135	80	14	936	6	416	8	520	16.148921000	28.6504...	116	145
192.168.137.76	47356	88.212.196.75	80	17	2423	9	1167	8	1256	16.152325000	1.990481	4690	5048
192.168.137.76	47357	88.212.196.75	80	12	828	6	416	6	412	16.152653000	28.6441...	116	115
192.168.137.76	47358	88.212.196.75	80	12	828	6	416	6	412	16.153311000	28.6434...	116	115
192.168.137.76	43899	87.250.247.181	80	12	804	6	416	6	388	16.154340000	28.6565...	116	108
192.168.137.76	43900	87.250.247.181	80	14	936	6	416	8	520	16.154842000	28.6456...	116	145
192.168.137.76	43901	87.250.247.181	80	14	936	6	416	8	520	16.154843000	29.7178...	111	139
192.168.137.76	43753	90.156.201.37	80	94	34 k	48	6887	46	27 k	67.622250000	6.945317	7932	31 k
192.168.137.76	43754	90.156.201.37	80	45	11 k	23	4600	22	6712	68.030153000	6.536531	5629	8214
192.168.137.76	43755	90.156.201.37	80	33	7489	17	3187	16	4302	68.030546000	6.536623	3900	5265
192.168.137.76	43756	90.156.201.37	80	16	1108	8	572	8	536	68.079619000	17.1400...	266	250
192.168.137.76	43215	46.36.218.162	80	47	24 k	23	2081	24	22 k	68.098567000	58.7056...	283	3048
192.168.137.76	39806	90.156.201.50	80	23	5239	11	1225	12	4014	68.100923000	6.466044	1515	4966
192.168.137.76	39807	90.156.201.50	80	23	5909	11	1231	12	4678	68.101114000	6.465135	1523	5788

Рис. 2. Атрибуты потоков приложения Google Chrome

имеющими одинаковый транспортный протокол и межпакетный интервал у которых не превышает заданного времени.

В качестве примера на рис. 2 в программе Wireshark визуализирован трафик, собранный из приложения Google Chrome.

Полученный набор данных был разделён на две отдельные выборки: обучающую и тестовую. Распределение количества потоков в каждой выборке отражено в таблице 1.

Выбор атрибутов классификации

Выделение атрибутов – процесс определения оптимального набора атрибутов, который требуется для решения задачи классификации или кластеризации. Слишком большое количество атрибутов усложняет анализ данных, а также увеличивает время обучения и создания модели классификатора. Поэтому одним из наиболее важных параметров при создании модели классификатора является качество и количество используемых атрибутов. Наибольшее распространение в задачах классификации получил алгоритм InfoGain [3], заключающийся в выборе признаков (атрибутов) на основе их информационного выигрыша. С помощью алгоритма InfoGain, были отобраны 14 атрибутов потоков, представленные в таблице 2.

Алгоритмы и критерии оценки эффективности алгоритмов классификации

Для классификации приложений использовались следующие алгоритмы машинного обучения: Naïve Bayes; C4.5 [7]; Random Forests [8]; Support Vector Machine (SVM) [9]; One Rule [10]; Adaptive Boost [11].

Для оценки эффективности алгоритмов классификации использовались следующие метрики информационного поиска [2,3,4,12]:

- Precision (Точность);
- Recall (Полнота);
- F-Measure (F-мера);

Таблица 1 – Распределение потоков приложений по выборкам

	Количество потоков	
Тип приложения	Обучающая выборка	Тестовая выборка
Google Chrome	3830	366
Instagram	1087	309
Facebook	1056	321
Messenger	1630	226
Whatsapp	1245	419
Всего	8848	1641

Таблица 2 – Атрибуты классификации

№	Атрибут	Содержание
1	Address A	IP-адрес источника
2	Port A	номер порта источника
3	Address B	IP-адрес получателя
4	Port B	номер порта получателя
5	Packets	количество переданных пакетов в потоке
6	Bytes	количество переданных байт в потоке
7	Packets AB	количество пакетов, переданных в направлении от источника к получателю
8	Bytes AB	количество байт, переданных в направлении от источника к получателю
9	Packets BA	количество пакетов, переданных в обратном направлении
10	Bytes BA	количество байт, переданных в обратном направлении
11	RelStart	начало потока относительно момента начала захвата в секундах
12	Duration	продолжительность данного потока в секундах
13	Bits/s AB	средняя скорость передачи в направлении от источника к получателю
14	Bits/s BA	средняя скорость передачи в обратном направлении

- ROC кривые (Receiver Operating Characteristic Curve);
- AUC (Area Under Curve) – площадь под ROC-кривой.

Результаты классификации

В таблице 3 отражены значения параметра Precision (точность) для рассмотренных алгоритмов классификации по каждому из классов.

Для наглядности значения метрики можно представить в виде сравнительных гистограмм (рис. 3).

Из рисунка 3 можно сделать вывод о том, что алгоритм One Rule для текущей выборки существенно проигрывает всем другим алгоритмам в точности. Методы Naïve Bayes, C4.5, AdaBoost M1 и Random Forest наиболее эффективны при фильтрации приложения WhatsApp,

Таблица 3 – Сравнение критерия точности для разных алгоритмов

Класс / Алгоритм	Naïve Bayes	One Rule	Random Forest	AdaBoost M1	C4.5
Facebook	0,421	0,321	0,701	0	0,433
Google Chrome	0,35	0,323	0,58	0,303	0,615
Instagram	0,783	0,219	0,764	0	0,575
Messenger	0,191	0,072	0,66	0	0,365
Whatsapp	0,993	0,221	0,909	0,775	0,846

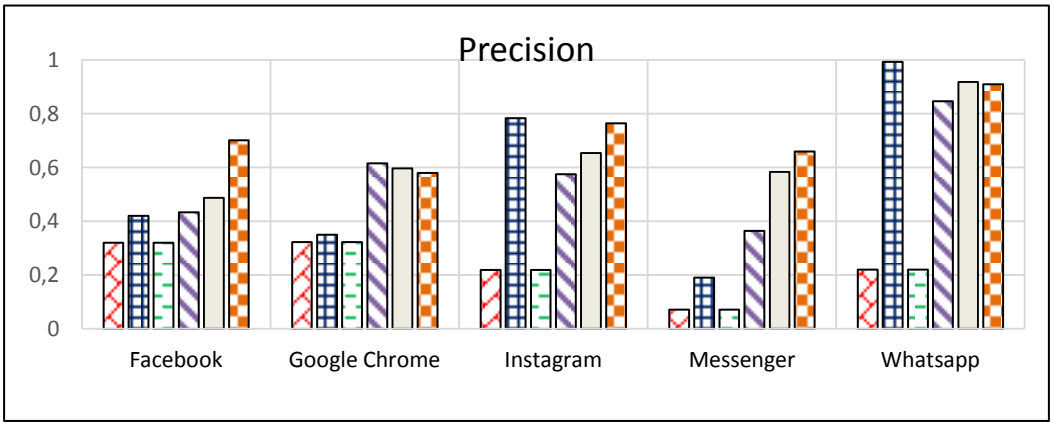


Рис. 3. Сравнительные гистограммы по критерию точности

Таблица 4 – Сравнение критерия полноты для разных алгоритмов

Класс \ Алгоритм	Naïve Bayes	One Rule	Random Forest	AdaBoost M1	C4.5
Facebook	0,05	0,196	0,343	0	0,361
GoogleChrome	0,885	0,566	0,923	0,989	0,893
Instagram	0,117	0,152	0,702	0	0,411
Messenger	0,416	0,093	0,611	0	0,549
Whatsapp	0,325	0,158	0,885	0,823	0,566

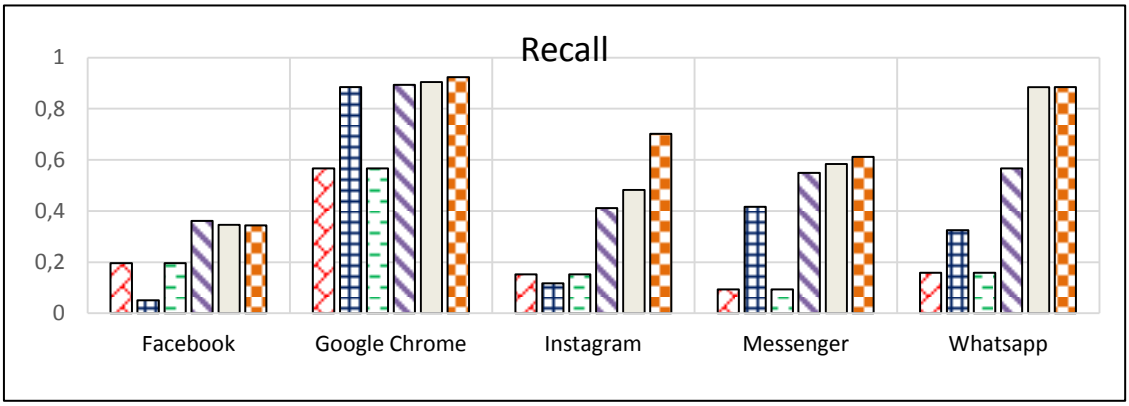


Рис. 4. Сравнительные гистограммы по критерию полноты

а для Messenger наименее эффективны алгоритмы AdaBoostM1 и Random Forest.

В таблице 4 приведены значения параметра Recall (полнота) для рассмотренных методов классификации для каждого класса.

Для наглядности значения метрик представлены в виде сравнительных гистограмм на рис. 4.

На рис. 5 представлены универсальные характеристики классификатора в виде F-меры, объединяющей характеристики точности и полноты. Видно, что наилучшим алгоритмом классификации мобильных приложений является алгоритм Random Forest, позволяющий обеспечить достоверность более 90%.

Таким образом, исследование эффективности алгоритмов классификации нежелательных мобильных приложений трафика подвижной радиосвязи для обнаружения угроз безопасности показало, что наилучшим является алгоритм Random Forest.

Среди анализируемых приложений наибо-

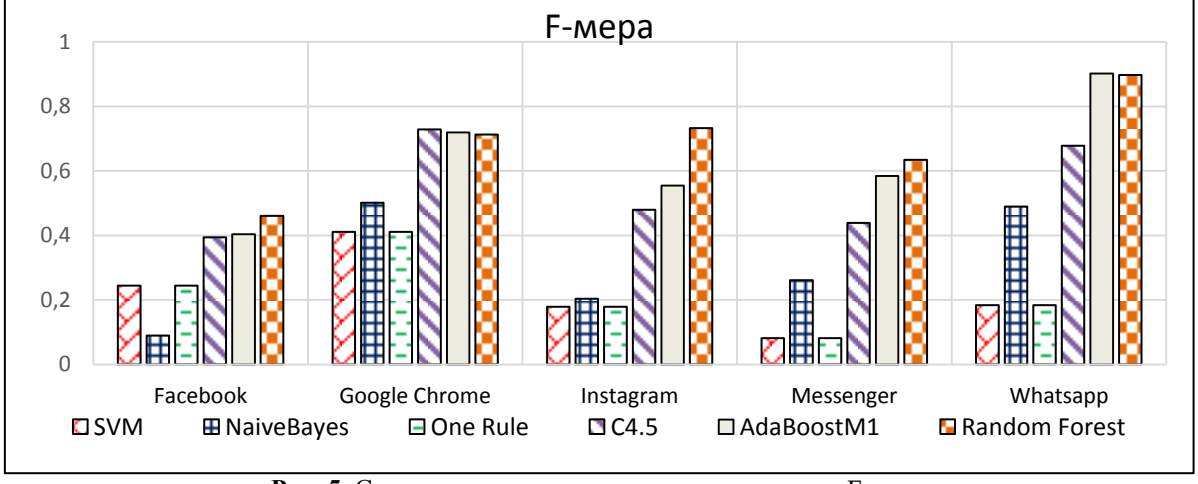


Рис. 5. Сравнительные гистограммы по критерию F-меры

лее эффективно классифицируются приложения Google Chrome и Whatsapp. Наиболее сложными для автоматической классификации оказались приложения Facebook и Messenger.

Литература

1. Статистика сайта «Сайты Рунета» // Web: http://www.liveinternet.ru/stat/ru/oses.html?slice=rus;id=2;id=15;id=12;id=4;id=11;id=checked;period=month

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. С. 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. Том 11. №2. 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. С. 46-52.

5.Wireshark // URL: http://www.howtogeek.com/ /104278/how-to-use-wireshark-to-capture-filter-and-inspect-packets/.

6. WEKA. The university of Wairato. – URL: http://weka.wikispaces.com/.

7. *Quinlan, J.R.* C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

8. *Ho, Tin Kam.* Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August, 1995.

9. *Cortes, C., Vapnik, V.* Support-vector networks / Machine Learning. 20, 1995. Pp. 293-297.

10. *Breiman, Leo*, Bagging predictors / Machine Learning. 24, 1996. Pp. 123-140

11. *Schapire R.* (2003) The Boosting Approach to Machine Learning: An Overview, MSRI Workshop on Nonliear Estimation and Classification, Springer, New York. – URL: https://doi.org/10/1007/978-0-387-21579-2_9.

Поступила 24 января 2018 г.

English

Filtering unwanted mobile radio traffic applications to detect information security threats

Oleg Ivanovich Shelukhin – Professor, Doctor of Engineering, Head of Information Security Department, Moscow Technical University of Communication and Information Science[1].

E-mail: sheluhin@mail.ru.

Mikhail Alexandrovich Smychyok – Candidate of Technical Sciences, Principal Discipline Engineer, Communication Networks Design Department JSC Giprogaztsentr[2].

E-mail: m-smychek@mail.ru.

Ayrapet Genrikhovich Simonyan – Candidate of Technical Sciences, Associate Professor, Department of Information Security, Moscow Technical University of Communication and Information Science[1].

E-mail: blackman-05@mail.ru.

[1]*Address:* 111024, Moscow, Aviamotornaya str., 8A.

[2]*Address:* 603950, Nizhny Novgorod, Alekseevskaya str. 26.

Abstract: To identify the kind of applications the network user utilizes is the mobile operator's problem as it is needed for statistics maintenance of the most frequently used applications. This application statistics identification aids not only to monitor network status but to detect failures as well, and if necessary to restrict access to network resources that may inflict harm to the user from the point of view of information security. The introduction of machine learning methods enables automatic classification, analysis and filtering of malicious and unwanted mobile applications of network traffic. Malicious mobile applications can pose a threat to data access or its integrity, and unwanted applications can pose a threat to privacy.

This article examines the efficiency of the following machine learning classification algorithms: Naive Bayes; C4.5; Random Forests; Support Vector Machine (SVM); One Rule; Google chrome Adaptive Boost mobile applications, Instagram, Facebook, Messenger, Whatsapp, mobile radio traffic. Such metrics as Precision, Recall, F-Measure, AUC were used to evaluate the classification algorithms efficiency. Random Forest algorithm is shown as the best one among machine learning classification algorithms. Google chrome and Whatsapp applications are most efficiently classified among the analyzed ones. Facebook and Messenger applications have proved to be the most difficult to automatically classify via machine learning methods. The presented results are proposed to be used in the selection of the best classification algorithms, evaluation of learning and testing samplings scope

to achieve high classification quality indicators in the environment of uncontrolled (background) traffic; to ensure the possibility of developing technologies for legitimate interception of network traffic in a manner similar to telephone companies based on the technology of network traffic classification.

Keywords: algorithms, classification, machine learning, mobile applications, packet, stream, application, protocol, network traffic, network, efficiency.

References

1. Site statistics "Runet Sites" / / Web: http://www.liveinternet.ru/stat/ru/oses.html?slice=rus;id=2;id=15;id=12;id=4;id=11;id=checked;period=month.

2. Shelukhin O. I., Simonyan, A. G., Vanyushina A. V. Algorithms efficiency for attributes isolation in applications classification problem with intelligent traffic analysis. // Elektrosvyaz. - 2016. - No. 11. – P. 79-85.

3. Shelukhin O. I., Simonyan, A. G., Vanyushina A. V. The effect of the learning sampling structure on the efficiency of application traffic classification using machine learning // T-Comm: Telekommunikatsii i transport. 2017. Vol. 11. No. 2. – P. 25-31.

4. Kostin D. V., Shelukhin O. I., Comparative analysis of machine learning algorithms for the encrypted network traffic classification // T - Comm: Telekommunikatsii i transport. 2016 no. 9. – P. 46-52.

5. Wireshark // URL: http://www.howtogeek.com/
/104278/how-to-use-wireshark-to-capture-filter-and-inspect-packets/.

6. WEKA. The university of Wairato. – URL: http://weka.wikispaces.com/.

7. Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

8. Ho, Tin Kam. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August, 1995.

9. Cortes, C., Vapnik, V. Support-vector networks / Machine Learning. 20, 1995. Pp. 293-297.

10. Breiman, Leo, Bagging predictors / Machine Learning. 24, 1996. Pp. 123-140

11. Schapire R. (2003) The Boosting Approach to Machine Learning: An Overview, MSRI Workshop on Nonliear Estimation and Classification, Springer, New York. – URL: https://doi.org/10/1007/978-0-387-21579-2_9.