УДК 004.056.53

А.В. Козачок, А.А. Спирин, О.М. Голембиовская

Алгоритм классификации псевдослучайных последовательностей на основе построения случайного леса

В последнее время увеличилось количество утечек конфиденциальных данных по вине внутренних нарушителей. Так как современные DLP-системы не способны обнаруживать и пресекать каналы утечки информации в зашифрованном или сжатом виде, в работе предлагается алгоритм классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных. Использовался ансамблевый метод машинного обучения – алгоритм построения случайного леса. В качестве признакового пространства был выбран массив частот встречаемости двоичных подпоследовательностей длины 9 бит и статистические характеристики распределения байт в последовательностях. Представленный алгоритм показал точность классификации псевдослучайных последовательностей 0,99.

Ключевые слова: статистический анализ данных, машинное обучение, классификация бинарных последовательностей, системы противодействия утечкам конфиденциальных данных, защита от утечек данных.
doi: 10.21293/1818-0442-2020-23-3-55-60

По результатам исследования экспертно-аналитического центра компании InfoWatch в первом полугодии 2019 г. было скомпрометировано 8,74 млрд. записей персональных данных в результате утечки конфиденциальной информации из коммерческих и некоммерческих организаций [1].

Более чем в 55% случаев утечки конфиденциальной информации произошли по вине внутренних нарушителей, обладавших легитимным доступом к данным [1].

Средняя мощность утечки составила 7,3 млн записей персональных данных, в 2018 г. данный показатель был равен 2 млн записей. Более 6,22 млрд $ составила сумма штрафов и компенсационных выплат, назначенных регуляторами.

Наиболее распространенным программным средством предотвращения утечек информации являются DLP-системы (data leakage prevention), осуществляющие анализ информационных потоков на предмет наличия конфиденциальных данных.

Подходы, реализуемые в DLP-системах, возможно разделить на две группы: контентные и контекстные. Контентные подходы направлены на анализ и обработку содержания передаваемых пакетов или файлов, контекстные оперируют служебной информацией потоков, пакетов или файлов. Контентные подходы подразделяют на сигнатурные (поиск цифровых слепков, регулярных выражений) и статистические, которые включают энтропийные подходы, тесты на случайность и методы подсчета подпоследовательностей.

Анализ литературы в области классификации различных типов данных позволяет выделить генезис методов классификации последовательностей. На начальном этапе производилась классификация сетевых протоколов контекстными методами на основе анализа IP-адресов, времени жизни пакетов, наличия флагов и сигнатур. Далее внимание исследователей привлекли приложения, осуществляющие передачу зашифрованных данных, и появились подходы, позволяющие идентифицировать передачу зашифрованных и открытых сообщений.

В основном подобные подходы базировались на расчёте энтропии блоков данных.

В настоящее время существует множество способов, позволяющих обнаруживать утечки конфиденциальных данных. В работах [2–7] рассматриваются методы глубокого анализа передаваемых данных, анализа контекста передаваемой информации, позволяющие обнаруживать передачу конфиденциальных данных за периметр контролируемой зоны.

Также обнаружение передачи зашифрованных данных необходимо в области защиты от распределенных сетевых атак, в настоящее время для этих целей также применяются сигнатурные методы [8] и методы машинного обучения, использующие в качестве признаков контекстную информацию передаваемых данных [9].

В работе [10] для идентификации приложений в операционной системе Android используются контекстные методы, выполняющие создание цифровых слепков приложений, содержащих информацию об устанавливаемых соединениях (IP-адрес, порт, длина пакета). В работе отмечается, что данный подход позволяет идентифицировать 110 наиболее популярных в сервисе GooglePlay приложений при передаче ими данных в зашифрованном виде посредством протоколов SSL/TLS.

В работе [11] отмечается факт отсутствия методов классификации зашифрованных и сжатых данных, что создает угрозу передачи конфиденциальной информации в сжатом виде. Авторы представляют метод классификации зашифрованных и сжатых данных в транспортных пакетах на основе сверточных нейронных сетей (66,9%), алгоритма *k*-ближайших соседей (60,0%) и полносвязных нейронных сетей прямого распространения (54,1%). Для выделения признаков используется подсчет критерия хи-квадрат для каждого квадранта полезной нагрузки пакета. Авторы отмечают, что теоретически суще-

ствуют более эффективные признаки классификации, но их поиск основан на интуиции.

В работе [12] отмечается, что рост интернет-трафика и возрастающее количество формирующих его устройств создают определенную сложность для DLP-систем. Современные системы фильтрации трафика не могут точно и эффективно обнаруживать информацию, обладающую высокой энтропией, например, зашифрованные и сжатые данные, что обусловливает актуальность разрабатываемого алгоритма.

В работе [13] отмечается, что существующие классификаторы с трудом справляются с задачей классификации зашифрованных и сжатых данных. Авторы предлагают алгоритм извлечения признаков, основанный на подсчете энтропии содержимого пакетных данных. Метод основан на увеличении избыточности сообщения путем генерации новых бинарных строк из анализируемых данных. Для формирования признакового пространства авторы предлагают формировать матрицу размером 8×4, строки которой являются значением шага, а столбцы – значениями бинарных подпоследовательностей, для которых выполняется расчет значения энтропии в полученных данных. Сформированное признаковое пространство используется для обучения классификаторов на основе метода опорных векторов или случайного леса.

Полученные авторами результаты свидетельствуют о значительном влиянии типа данных на результаты классификации. Наихудшие значения точности классификации были получены для аудиофайлов (0,65), для видеофайлов значение точности классификации составило менее 0,7, для изображений и текста – примерно 0,72.

Существующие решения обладают схожим слабым местом, они используют сигнатуры и служебную информацию при классификации зашифрованных и открытых данных. Также применяются статистические подходы на основе подсчета энтропии, которые демонстрируют высокую точность при идентификации зашифрованных/открытых или сжатых/открытых данных, однако при использовании энтропийного подхода к задаче классификации зашифрованных и сжатых данных точность окажется невысокой, т.к. шифры обладают рассеивающей способностью, а алгоритмы сжатия устраняют избыточность данных.

Поскольку представленные решения не демонстрируют высокую точность классификации зашифрованных и сжатых последовательностей, была выдвинута гипотеза о наличии у зашифрованных и сжатых данных статистических особенностей, которые не могут обнаружить существующие методы анализа данных.

Подтвердить гипотезу могут алгоритмы машинного обучения, т.к. они способны находить нелинейные зависимости и связи в анализируемых данных, т.е. признаки, которые не очевидны на первый взгляд.

Для проверки гипотезы был сформирован набор данных, содержащий 10 000 зашифрованных файлов размером по 600 Кбайт (AES, 3DES, Camellia, RC4, ГОСТ 34.12-15 «Кузнечик») и 12 000 файлов размером по 600 Кбайт с наиболее часто используемыми расширениями (RAR, ZIP, 7Z, GZIP, XZ, BZIP2). Все файлы были сформированы из осмысленного текста на русском языке. Одинаковых файлов в наборе данных не имелось, при обработке файлов осуществлялось удаление их заголовков, т.е. служебной информации.

Ранее проведенные эксперименты [14] показали применимость предложенного подхода для классификации зашифрованных и сжатых текстовых файлов. Кроме того, зашифрованные и сжатые последовательности обладают свойствами псевдослучайных последовательностей, поскольку они успешно проходят статистические тесты NIST. По этой причине зашифрованные и сжатые данные могут обозначаться как псевдослучайные. Исследовалось несколько алгоритмов машинного обучения, метрикой была выбрана доля правильных ответов (accuracy), т.к. набор данных является сбалансированным [15]. Результаты проведённых экспериментов представлены в табл. 1.

Таблица 1

Сравнение алгоритмов машинного обучения

Классификатор	Accuracy
RandomForest	0,94
DecisionTree	0,87
K-neighbors	0,88
GradientBoosting	0,89

На рис. 1 представлен алгоритм классификации псевдослучайных последовательностей.

Data: $P : |P| = Q$, $S : |S| = 512$, $B : |B| = 256$, $E = S + B$, $InspectedData = I$

Result: $ImportanceFeatures, Y$

$F_{P,E} \leftarrow <>$
for $p \in P$ **do**
  for $s \in S$ **do**
    $n_s \leftarrow$ **Count**(p,s)
    $f_{p,s} \leftarrow \frac{n_s}{M_p - N_s + 1}$
    $F_{P,E} \leftarrow f_{p,s}$
  for $b \in B$ **do**
    $n_b \leftarrow$ **Count**(b,s)
    $bytes_p \leftarrow < b, n_b >$
    $F_{P,E} \leftarrow bytes_p$
  $F_{P,E} \leftarrow$ **Std**$(bytes_p)$
  $F_{P,E} \leftarrow$ **Min**$(bytes_p)$
  $F_{P,E} \leftarrow$ **Max**$(bytes_p)$
  $F_{P,E} \leftarrow$ **Delta**(max_b, min_b)
  $ImportanceFeatures \leftarrow$ **GetWeight**$(F_{P,E}, RandomForest)$
  for $i \in I$ **do**
    $y \leftarrow$ **GetClass**$(i, ImportanceFeatures, DecisionTree)$
return Y

Рис. 1. Алгоритм классификации псевдослучайных последовательностей

Исходными данными являются: набор зашифрованных и сжатых данных *P*, множество бинарных подпоследовательностей *S* длины 9 бит, множество байт *B* и данные для проверки *I*. Результатом работы алгоритма является массив определенных классов *Y* для файлов из множества *I* и множество наиболее значимых признаков.

Для каждой последовательности *p* из множества *P* выполняется определение частот встречаемости подпоследовательностей длины 9 бит из множества *S* и подсчет встречаемости каждого байта *b* множества *B*. Также в классификатор заносятся статистические значения полученного распределения байт: среднее значение частоты встречаемости байт, минимальное и максимальное значение какого-либо байта и их разница. Далее посредством функции GetWeight осуществляется определение наиболее значимых признаков, т.е. признаков, обладающих наиболее выраженными дискриминирующими способностями для разделения зашифрованных и сжатых данных. Для данной цели применяется алгоритм построения случайного леса, способ получения значений гиперпараметров классификатора представлен далее.

Для поиска лучших параметров были проведены эксперименты по оценке классификаторов в зависимости от числа параметров, учитываемых ими. Результаты представлены на рис. 2, определены 6 признаков, необходимых для достижения наибольшей точности классификации.

Рис. 2. Определение оптимального числа признаков классификатора

Для снижения времени, затрачиваемого на обучение классификатора и проведение процедуры классификации псевдослучайных последовательностей, были проведены эксперименты по поиску оптимальной глубины формируемых деревьев. Результаты представлены в табл. 2.

Таблица 2

Определение оптимальный глубины деревьев в случайном лесе

Максимальная глубина деревьев в случайном лесе	Метрика accuracy
5	0,861
10	0,902
15	0,918
25	0,937
35	0,94

Так как алгоритм построения случайного леса является ансамблевым методом, то были проведены эксперименты по определению оптимального количества деревьев в ансамбле. Для достижения высокой точности классификации достаточно 61 дерева, более высокие значения точности классификатор достигает при 151 и 172 деревьях в ансамбле, однако прирост точности незначителен, порядка 0,001, а время, затрачиваемое на обучение и дальнейшую классификацию, увеличивается примерно в 2 раза.

Далее выполняется классификация анализируемых файлов из множества *I* посредством наиболее значимых признаков Importance Features и алгоритма построения дерева решений. Для каждого анализируемого файла *i* осуществляется подсчет встречаемости подпоследовательностей длины 9 бит и некоторых байт, представленных в табл. 3.

Таблица 3

Наиболее значимые признаки при проведении классификации

Классификационный признак	
b0	Функция от значения частоты нулевого байта
max $F(x)\|x=0\ldots255$	Функция от максимального значения частоты какого-либо байта
1 0000 0000	Частоты подпоследовательностей длины 9 бит
0 0000 0000	
0 0000 0001	
0 1111 1111	

Далее осуществляется итерационный проход по сформированному дереву решений на основе признаков из Importance Features и вычисленных признаков анализируемого файла *i*, в конце каждой ветви дерева находится один из классов псевдослучайных последовательностей, который присваивается файлу *i*.

В табл. 4 представлены значения площади под кривой ошибок (AUC-ROC-кривая) классификации зашифрованных/сжатых данных, позволяющие оценить полученный классификатор.

Одной из метрик оценки классификатора является площадь под кривой ошибок AUC-ROC (Area Under Curve Receiver Operating Characteristic). Данная кривая представляет собой линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR). TPR – это полнота (отражает, какое количество из всех возможных релевантных

элементов выбрано, т.е. характеризует способность классификатора обнаруживать класс объектов в целом), а FPR отражает, какую долю из объектов отрицательного класса алгоритм предсказал неверно. В идеальном случае, когда классификатор не делает ошибок (FPR = 0, TPR = 1), мы получим площадь под кривой, равную единице; в противном случае, когда классификатор случайно выдает вероятности классов, AUC-ROC будет стремиться к 0,5, так как классификатор будет выдавать одинаковое количество TPR и FPR. Данный случай обозначен пунктирной линией на графике [16]. Площадь под кривой ошибок полученного классификатора составила в среднем 0,99.

Также были проведены эксперименты по исследованию зависимости точности классификации от длины анализируемых файлов, результаты представлены на рис. 3. Направление для анализа файла (с начала файла или конца) не влияет на полученную зависимость.

Для оценки точности были выбраны различные метрики, представленные в табл. 5. Наибольшую точность удалось достичь при использовании метрики площадь под кривой ошибок.

Таблица 4

Значения метрики AUC-ROC полученного классификатора

Итерация кроссвалидации	Метрика AUC-ROC
1	0,99
2	0,99
3	0,99
4	0,99
5	0,99
Среднее значение	0,99

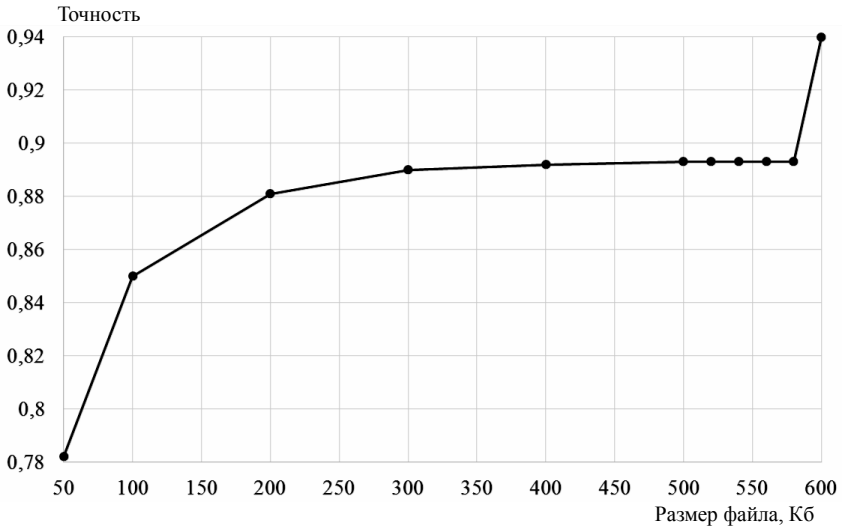

Рис. 3. Зависимость точности бинарной классификации зашифрованных и сжатых последовательностей в зависимости от размера файлов

Таблица 5

Оценка различных метрик при бинарной классификации зашифрованных и сжатых последовательностей

Метрика	Значение
AUC-ROC	0,99
F1-score	0,95
Accuracy	0,95

Полученные в ходе экспериментов значения гиперпараметров классификатора представлены в табл. 6.

Таблица 6

Полученные значения параметров классификатора

Гиперпараметр классификатора	Значение
Длина подпоследовательности	9 бит
Количество признаков	6
Максимальная глубина леса	35
Количество деревьев	61

Сокращенное количество признаков, используемых классификатором, как и ограниченное число деревьев и максимальная глубина леса, существенно снижают время обучения классификатора и время, затрачиваемое на классификацию последовательностей.

Для формирования классификатора наибольший вес имеют статистические признаки, получаемые из подпоследовательностей длины 9 бит и байтового распределения анализируемых данных.

Заключение

Исходя из анализа отчетов информационно-аналитических агентств, занимающихся вопросами информационной безопасности, была рассмотрена проблема наличия канала утечки конфиденциальных данных за счет внутренних нарушителей. Одной из возможных причин подобных утечек может являться наличие высоких полномочий у нарушителей за счет их легитимного нахождения в информационной системе и наличия у них средств шифрования или сжатия данных. Для совершенствования существующих систем противодействия утечкам конфиденциальных данных был предложен алгоритм классификации псевдослучайных последовательностей на основе построения случайного леса, выдвинута и подтверждена гипотеза о наличии у них статистических особенностей, позволяющих построить классификатор с точностью 0,99. Предложенный алгоритм позволит улучшить существующие DLP-

системы за счет увеличения точности классификации зашифрованных и сжатых данных.

Исследование выполнено при финансовой поддержке Минобрнауки России (грант ИБ) в рамках научного проекта № 18/2020.

Литература

1. Глобальное исследование утечек конфиденциальной информации в первом полугодии 2019 года. – URL: https://www.infowatch.ru/analytics/reports/27614 (дата обращения: 23.09.2020).

2. Cheng L. Enterprise data breach: causes, challenges, prevention, and future directions / L. Cheng, F. Liu, D. Yao // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2017. – Vol. 7, No. 5. – P. e1211.

3. Shu X. Privacy-Preserving Detection of Sensitive Data Exposure / X. Shu, D. Yao, E. Bertino // IEEE Transactions on Information Forensics and Security. – 2015. – Vol. 10, No. 5. – P. 1092–1103.

4. Privacy-preserving scanning of big content for sensitive data exposure with MapReduce / F. Liu, X. Shu, D. Yao, A.R. Butt // Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. – 2015. – P. 195–206.

5. Rapid and parallel content screening for detecting transformed data exposure / X. Shu, J. Zhang, D. Yao, W. Feng // Proceedings of the Third International Workshop on Security and Privacy in Big Data. – 2015. – P. 191–196.

6. Fast Detection of Transformed Data Leaks / X. Shu, J. Zhang, D. Yao, W. Feng // IEEE Transactions on Information Forensics and Security. – 2016. – Vol. 11, No. 3. – P. 528–542.

7. A data leakage prevention method based on the reduction of confidential and context terms for smart mobile devices / X. Yu, Z. Tian, J. Qiu, F. Jiang // Wireless Communications and Mobile Computing. – 2018. – Vol. 2018. – P. 1–11.

8. Добрышин М.М. Предложение по совершенствованию систем противодействия DDoS-атакам // Телекоммуникации. – 2018. – № 10. – С. 32–38.

9. VACCINE: Using Contextual Integrity For Data Leakage Detection / Y. Shvartzshnaider, Z. Pavlinovic, A. Balashankar, T. Wies, L. Subramanian, H. Nissenbaum, P. Mittal // The World Wide Web Conference. – 2019. – P. 1702–1712.

10. Robust smartphone app identification via encrypted network traffic analysis / V.F. Taylor, R. Spolaor, M. Conti, I. Martinovic // IEEE Transactions on Information Forensics and Security. – 2017. – Vol. 13, No. 1. – P. 63–78.

11. Hahn D. Detecting compressed cleartext traffic from consumer internet of things devices / D. Hahn, N. Apthorpe, N. Feamster // arXiv preprint. – 2018. – URL: https://arxiv.org/pdf/1805.02722.pdf (дата обращения: 23.09.2020).

12. Casino F. HEDGE: efficient traffic classification of encrypted and compressed packets / F. Casino, K.K.R. Choo, C. Patsakis // IEEE Transactions on Information Forensics and Security. – 2019. – Vol. 14, No. 11. – P. 2916–2926.

13. Tang Z. Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification / Z. Tang, X. Zeng, Y. Sheng // International Journal of ICIC. – 2019. – Vol. 15, No. 3. – P. 845–860.

14. Козачок А.В. Алгоритм классификации псевдослучайных последовательностей / А.В. Козачок, А.А. Спирин // Вестник Воронеж. гос. ун-та. Сер.: Системный анализ и информационные технологии. – 2020. – № 1. – С. 87–98.

15. Breiman L. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen. – London: Chapman & Hall/CRC, 2017. – 358 p.

16. Muschelli J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric // Journal of Classification. – 2019. – URL: https://doi.org/10.1007/s00357-019-09345-1 (дата обращения: 25.10.2020).

Козачок Александр Васильевич
Д-р техн. наук, сотрудник Академии Федеральной службы охраны Российской Федерации (Академия ФСО России)
Приборостроительная ул., д. 35, г. Орел, Россия, 302034
ORCID 0000-0002-6501-2008
Тел.: +7 (486-2) 54-13-57
Эл. почта: a.kozachok@academ.msk.rsnet.ru

Спирин Андрей Андреевич
Сотрудник Академии ФСО России
Приборостроительная ул., д. 35, г. Орел, Россия, 302034
ORCID 0000-0002-7231-5728
Тел.: ++7 (486-2) 54-13-57
Эл. почта: spirin_aa@bk.ru

Голембиовская Оксана Михайловна
Канд. техн. наук, сотрудник Брянского государственного технического университета (БГТУ)
Бульвар 50-лет Октября, д. 7, г. Брянск, Россия, 241035
ORCID 0000-0002-6433-3133
Тел.: +7(483-2) 58-83-55
Эл. почта: bryansk-tu@yandex.ru

Kozachok A.V., Spirin A.A., Golembiovskaya O.M.
Random forest based pseudorandom sequences classification algorithm

Recently, the number of confidential data leaks caused by internal violators has increased. Since modern DLP-systems cannot detect and prevent information leakage channels in encrypted or compressed form, an algorithm was proposed to classify pseudo-random sequences formed by data encryption and compression algorithms. Algorithm for constructing a random forest was used. An array of the frequency of occurrence of binary subsequences of 9-bit length and statistical characteristics of the byte distribution of sequences was chosen as the feature space. The presented algorithm showed the accuracy of 0,99 for classification of pseudorandom sequences. The proposed algorithm will improve the existing DLP-systems by increasing the accuracy of classification of encrypted and compressed data.

Keywords: statistical analysis of data, machine learning, classification of binary sequences, DLP systems, protection against leakage of information.
doi: 10.21293/1818-0442-2020-23-3-55-60

References

1. Global Confidential Information Leak Survey in the first half of 2019. Available at: https://www.infowatch.ru/analytics/reports/27614 (Accessed: September 23, 2020) (in Russ.).

2. Cheng L., Liu F, Yao D. Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2017, vol. 7, no. 5, pp. e1211.

3. Shu X., Yao D., Bertino E. Privacy-preserving detection of sensitive data exposure. *IEEE Transactions on Information Forensics and Security*, 2015, vol. 10, no. 5, pp. 1092–1103.

4. Liu F, Shu X., Yao D., Butt A.R. Privacy-preserving scanning of big content for sensitive data exposure with MapReduce. *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, 2015, pp. 195–206.

5. Shu X., Zhang J., Yao D., Feng W. Rapid and parallel content screening for detecting transformed data exposure. *Proceedings of the Third International Workshop on Security and Privacy in Big Data*, 2015, pp. 191–196.

6. Shu X, Zhang J., Yao D., Feng W. Fast Detection of Transformed Data Leaks. *IEEE Transactions on Information Forensics and Security*, 2016, vol. 11, no. 3, pp. 528–542.

7. Yu X, Tian Z., Qiu J., Jiang F. A data leakage prevention method based on the reduction of confidential and context terms for smart mobile devices. *Wireless Communications and Mobile Computing*, 2018, vol. 2018, pp. 1–11.

8. Dobryshin M.M. *Predlozenie po sovershenstvovanij system protivodejstvij DDoS-atakam* [Proposal for improving systems to counter DDoS attacks]. *Telecommunications*, 2018, vol. 10, pp. 32–38 (in Russ.).

9. Shvartzshnaider Y., Pavlinovic Z., Balashankar A., Wies T., Subramanian L., Nissenbaum H., Mittal P. VACCINE: Using Contextual Integrity For Data Leakage Detection. *The World Wide Web Conference*, 2019, pp. 1702–1712.

10. Taylor V.F., Spolaor R., Conti M., Martinovic I. Robust smartphone app identification via encrypted network traffic analysis. *IEEE Transactions on Information Forensics and Security*, 2017, vol. 13, no. 1, pp. 63–78.

11. Hahn D., Apthorpe N., Feamster N. Detecting compressed cleartext traffic from consumer internet of things devices. *ArXiv preprint:1805.02722*, 2018. Available at: https://arxiv.org/pdf/1805.02722.pdf (Accessed: September 23, 2020).

12. Casino F., Choo K.K.R., Patsakis C. HEDGE: efficient traffic classification of encrypted and compressed packets. *IEEE Transactions on Information Forensics and Security*, 2019, vol. 14, no. 11, pp. 2916–2926.

13. Tang Z., Zeng X., Sheng Y. Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification. *International Journal of ICIC*, 2019, vol. 15, no. 3, pp. 845–860.

14. Kozachok A.V., Spirin A.A. *Algoritm klassificazii psevdosluchainih posledovaltesnostei* [Pseudo random sequences classification algorithm]. *Proceedings of the Voronezh state university. Series: system analysis and information technology*, 2020, vol. 2020, no. 1, pp. 87–98 (in Russ.).

15. Breiman L., Friedman J.H., Olshen R.A., Stone J.G. *Classification and regression trees*. London: Chapman & Hall/CRC, 2017, 358 p.

16. Muschelli J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. *Journal of Classification*, 2019. Available at: https://doi.org/10.1007/s00357-019-09345-1 (Accessed: October 25, 2020).

Alexander V. Kozachok
Doctor of Engineering Sciences,
Employee, Academy of the Federal Guard Service of the Russian Federation (Academy of the FGS of the Russia),
35, Priborostroitelnaya st., Orel, Russia, 302034
ORCID 0000-0002-6501-2008
Phone: +7 (486-2) 54-13-57
Email: a.kozachok@academ.msk.rsnet.ru

Andrey A. Spirin
Employee, Academy of the FGS of the Russia
35, Priborostroitelnaya st., Orel, Russia, 302034
ORCID 0000-0002-7231-5728
Phone: ++7 (486-2) 54-13-57
Email: spirin_aa@bk.ru

Oksana M. Golembiovskaya
Candidate of Engineering Sciences, Employee,
Bryansk State Technical University
7, 50th anniversary of October Boulevard, Bryansk, Russia, 241035
ORCID 0000-0002-6433-3133
Phone: +7 (483-2) 58-83-55
Email: bryansk-tu@yandex.ru