ПРОГРАММИРОВАНИЕ, 2021, № 4, с. 31–44

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.056.53

МОДЕЛЬ ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, СФОРМИРОВАННЫХ АЛГОРИТМАМИ ШИФРОВАНИЯ И СЖАТИЯ ДАННЫХ

© 2021 г. А. В. Козачок[a],*, А. А. Спирин[a],**

[a] *Академия ФСО, 302034 Орел, ул. Приборостроительная, 35, Россия*

**E-mail: a.kozachok@academ.msk.rsnet.ru*

***E-mail: spirin_aa@bk.ru*

Поступила в редакцию 03.03.2021 г.
После доработки 16.03.2021 г.
Принята к публикации 17.03.2021 г.

Задача классификации источников данных, обладающих высокой энтропией, в области информационной безопасности занимает одну из ключевых позиций. В настоящее время существуют способы классификации зашифрованных и сжатых последовательностей, которые в основном используют цифровые сигнатуры или служебную информацию в случае ее передачи. В работе проведен анализ исследований в области классификации зашифрованных и сжатых данных и разработана модель зашифрованных и сжатых последовательностей. Практические эксперименты свидетельствуют о высокой точности предложенного подхода и позволяют сделать вывод об улучшении существующих методов классификации зашифрованных и сжатых данных. Предложенный способ может быть внедрен в системы защиты данных от утечек либо в корпоративные системы электронной почты для анализа отправляемых за контролируемый периметр организации вложений.

DOI: 10.31857/S0132347421040051

Цель исследования – разработать модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, позволяющую наиболее точно отразить статистические свойства указанных последовательностей.

Метод исследования – статистический анализ данных, математическая статистика, машинное обучение.

Результат исследования – проведен анализ исследований, направленных на решение задачи классификации зашифрованных и сжатых последовательностей в области информационной безопасности. Разработана модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающая их статистические признаки: распределение байт и частоты встречаемости подпоследовательностей ограниченной длины, представляющие собой новое вероятностное пространство. Приведено обоснование выбора статистических признаков, использующихся в модели псевдослучайных последовательностей. Проведены эксперименты по определению гиперпараметров классификатора на сформированном наборе данных из зашифрованных и сжатых файлов без учета их заголовков. Определены ограничения, используемые в модели псевдослучайных последовательностей, заключающиеся в условии равенства длины анализируемых псевдослучайных последовательностей около 600 кбайт. Проведены эксперименты по определению влияния статистических признаков, участвующих в формировании модели псевдослучайных последовательностей, на точность классификации. Полученные практические результаты позволяют классифицировать зашифрованные и сжатые данные с точностью 0.97.

1. ВВЕДЕНИЕ

Задача классификации зашифрованных и сжатых данных занимает в информационной безопасности одну из ключевых позиций: обнаружение зашифрованного вредоносного ПО, расследования в области компьютерной криминалистики, анализ трафика и защита данных от утечек.

Современные методы обнаружения зашифрованных данных опираются на энтропийный подход, который демонстрирует низкую точность классификации высокоэнтропийных данных: зашифрованной и сжатой информации, изображений, выходных последовательностей кодировщиков.

В последнее время особенно остро стоит задача предотвращения утечки конфиденциальных данных из корпоративной сети. Отчет экспертно-

аналитического центра группы компаний InfoWatch [1] свидетельствует о высокой доле (более 79%) внутренних нарушителей как источников инцидентов информационной безопасности в России.

В работе [2] отмечается, что угрозы, вызванные внутренним нарушителем, являются самым опасными и наиболее распространенными для широко круга организаций, в том числе для государственных учреждений. Вредоносные действия в данном случае осуществляются доверенными лицами внутри организаций, что приводит к существенному ущербу.

Решение задачи обнаружения зашифрованных данных является критически важным в различных областях обеспечения информационной безопасности: противодействие утечке конфиденциальных данных и обнаружение вредоносного ПО, детектирование DDoS-атак, решение задач компьютерной криминалистики, расследование инцидентов информационной безопасности. В работе [3] осуществляется обзор методов глубокого анализа данных и обнаружения зашифрованного трафика. Авторы делают вывод о неспособности методов глубокого анализа пакетов производить классификацию зашифрованного трафика, поскольку шифрование изменяет содержание информации, однако, по мнению авторов, оно не изменяет сетевые характеристики потоков передачи данных. Кроме того, отмечается необходимость применения статистических подходов, не использующих содержание передаваемых данных.

Верно классифицируя выходные последовательности алгоритмов шифрования и сжатия данных возможно улучшить существующие методы обнаружения утечки защищаемой информации.

Статья организована следующим образом: в первой части рассмотрены существующие подходы к классификации зашифрованного трафика, обнаружению DDoS-атак, искаженных файлов и файлов вредоносного ПО, сформулирована решаемая задача. Во второй части представлена разработанная модель зашифрованных и сжатых последовательностей на основе частот встречаемости подпоследовательностей длины 9 бит и распределения байт. Приведены результаты практических экспериментов.

2. ПОДХОДЫ К КЛАССИФИКАЦИИ ЗАШИФРОВАННЫХ И СЖАТЫХ ДАННЫХ

Исследования по классификации данных в предметной области информационной безопасности условно возможно разделить на 3 группы: подходы к классификации зашифрованного, сжатого и открытого трафиков данных информационно-телекоммуникационных сетей; обнаружение DDoS-атак; определение типа передаваемых данных, вредоносного программного обеспечения. Обобщенный анализ рассмотренных работ представлен в таблице 1.

2.1. Задача классификации зашифрованных и сжатых данных

Существующие DLP-системы выполняют анализ служебной информации, присущей передаче данных, либо на основе поиска различных сигнатур и регулярных выражений непосредственно в данных. В ряде работ отмечается невозможность обнаружения конфиденциальных данных в зашифрованном или сжатом виде [31, 32].

Ряд исследователей отмечают, что не существует эффективных и точных методов классификации источников с высокой энтропией, например алгоритмов шифрования и сжатия данных. [33, 34].

Внутреннему нарушителю доступны средства шифрования и сжатия, что позволяет сделать вывод об актуальности задачи классификации зашифрованных и сжатых данных. Рассмотренные подходы не могут являться надежным решением для обнаружения передачи зашифрованных данных по ряду причин:

1. Методы анализа трафика не применимы, поскольку нельзя допустить передачу данных за контролируемый периметр организации. Необходимо разработать средства анализа данных до их отправки вовне, например после их загрузки на сервер электронной почты.

2. Подходы с применением нейронных сетей в основном требуют значительно большего времени на обучение классификатора, чем другие алгоритмы машинного обучения. Как правило, в данных подходах применяются заголовки файлов, содержащие “магические” байты, обладающие высокой дискриминирующей способностью.

3. Рассмотренные энтропийные и иные подходы также оперируют заголовками файлов, содержащими “магические” байты.

Таким образом, можно сформулировать требования, предъявляемые к методу классификации зашифрованных и сжатых данных:

1. Применение алгоритмов машинного обучения с минимальным временем обучения классификатора.

2. Использование статистических подходов, не учитывающих заголовки файлов и специальные дискриминирующие сигнатуры.

3. Возможность применения на серверной стороне, до момента передачи за контролируемый периметр организации.

4. Минимальное время выполнения классификации последовательности.

Таблица 1. Обобщенный анализ рассмотренных работ

Источник	Год	Объект	Признаки	Мат. аппарат	Точность
Задача классификации трафика					
[4]	2016	Трафик	Энтропия	Дерево решений	0.981
[5]	2017		Служебная информация	Цепи Маркова 2-го порядка	0.9122
[6]	2017		Служебная информация, распределение байт	kNN	0.952
[7]	2018		Служебная информация, распределение байт	Случайные поля Маркова, СНС	0.979
[8]	2019		Служебная информация, трассы DNS, характеристика обмена сертификатами	XGBoost	0.987
[9]	2019		Служебная информация и статистический анализ	СНС, ГСС, автокодировщики	Обзор методов
[10]	2019	Файлы, трафик	Энтропия	МОВ, СЛ	Файлы – 0.72 Трафик – 0.979
[11]	2020	Трафик	Служебная информация	СНС, СЛ, ДР, XGBoost,	0.96
[12]	2020		Служебная информация, распределение байт	Автокорреляция	100%
[13]	2020		Служебная информация	Скрытые цепи Маркова, модель смеси Гауссовых распределений	Трафик сети Tor – 0.999
[14]	2020		Служебная информация, статистические хар-ки	МОВ	0.8
[15]	2020		Служебная информация, статистические хар-ки	СЛ	0.882
Задача обнаружения DDoS-атак					
[16]	2017	DDoS	Служебная информация	Кластерный анализ	0.998
[17]	2017		Служебная информация	Кластерный анализ	0.989
[18, 19]	2018–2019		Служебная информация	Теория вероятностей	0.97
[20]	2020		Служебная информация, статистические признаки	СЛ, ДР	0.98
[21]	2018		Служебная информация	kNN	0.992
Задача классификации файлов					
[22]	2017	Файлы	Статистические признаки	SVM	0.607
[23, 24]	2017		Распределение байт	ГА, НС	98.21–100%
[25]	2018	ВПО	Цепочки команд (n–граммы)	Скрытые цепи Маркова	
[26]	2019	Файлы	Распределение байт	СНС (VGG-16, ResNet)	0.999
[27]	2019		Частота встречаемости слов	XGBoost	98.84%
[28]	2019	Файлы, трафик	Тесты NIST, энтропия блоков данных	HEDGE (High Entropy DistinGuishEr)	0.72
[29]	2020	Файлы	Распределение байт	НС	80–100%
[30]	2020		Энтропия	Скрытые цепи Маркова	52%

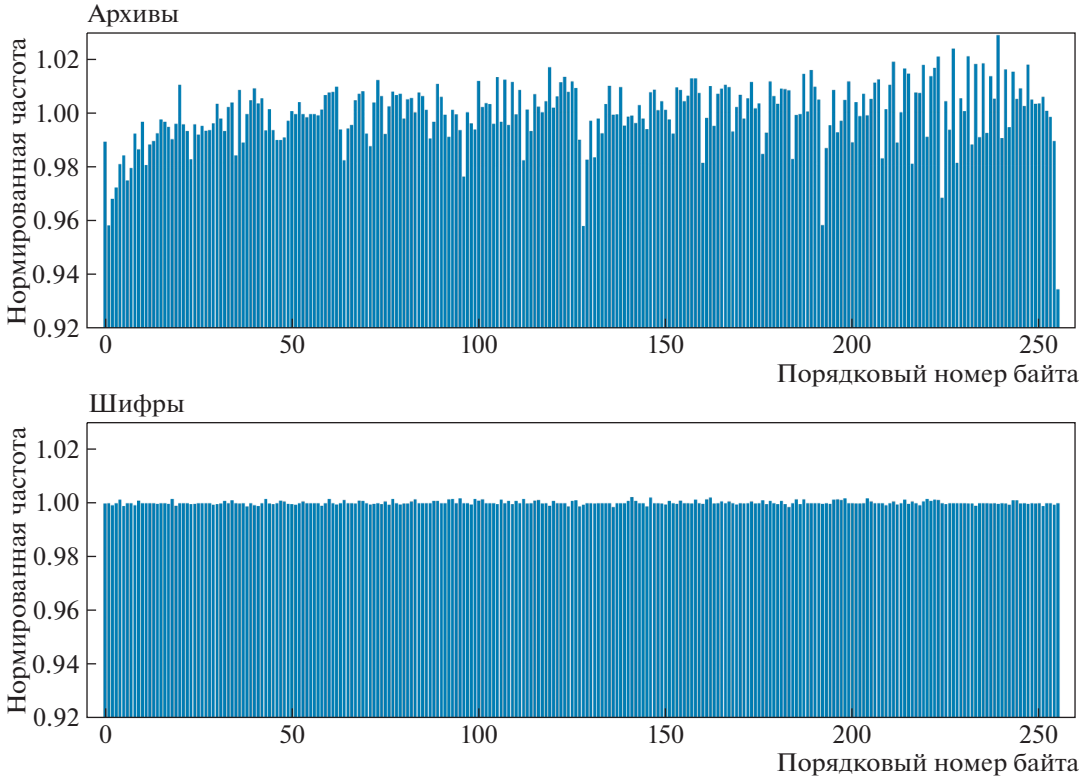

Рис. 1. Распределение байт для зашифрованных и сжатых последовательностей.

В формальном виде задача бинарной классификации псевдослучайных последовательностей (ПСП – последовательностей, обладающих равномерным распределением байт) может быть представлена в выражении (1):

$$F(V(s)) = k, \quad k \in \{0,1\}, \tag{1}$$

где $V(s)$ – модель ПСП, s – анализируемая ПСП, k – класс ПСП, F – функция классификации ПСП.

3. МОДЕЛЬ ПСП

Для построения модели ПСП было сформировано 2000 файлов, содержащих осмысленный текст. Далее каждый из них был преобразован алгоритмами шифрования (OpenSSL[1]: AES, 3DES, Camellia, RC4 и ГОСТ 34.12 “Кузнечик” в режиме простой замены) и сжатия данных (WinRAR[2]: RAR, ZIP и 7Zip[3]: 7Z, XZ, GZ, BZ2). Размер файлов после обработки алгоритмами преобразования данных составил 600 кбайт.

Учитывая, что структура файлов имеет заголовочную часть, которую возможно модифицировать для маскирования информации, содержащейся в ней, данный способ может быть использован злоумышленниками для передачи конфиденциальной информации. В работе [35] авторы сделали вывод о необходимости удаления заголовков файлов, так как они содержат цифровые сигнатуры, позволяющие классифицировать тип данных с высокой точностью. При построении модели ПСП были отброшены первые 10 кбайт файлов, чтобы исключить влияние заголовков файлов и “магических” байт на процедуру классификации.

На первом этапе построения модели ПСП были произведены оценки распределения байт последовательностей двух классов, нормированных по среднему значению частоты встречаемости байт в выборке данных, результаты представлены на рисунке 1.

Визуально, исходя из рисунка 1, возможно сделать вывод о том, что распределение байт зашифрованных ПСП более равномерно, чем распределение байт сжатых ПСП.

Для оценки полученных распределений частот встречаемости байт в ПСП была проведена оценка их соответствия равномерному распределению согласно критерию Хи-квадрат, определяемому выражением (2):

[1] https://www.openssl.org/ (дата обращения 08.02.2021)
[2] https://www.win-rar.com/ (дата обращения 08.02.2021)
[3] https://www.7-zip.org/ (дата обращения 08.02.2021)

Таблица 2. Проверка гипотезы о равномерности распределения байт в ПСП

№ п/п	ПСП	Референсное распределение	Критическое значение критерия	Расчетное значение критерия	P-value
1	Архивы	Равномерное	293.248	50.811	1
2	Шифры	Равномерное	293.248	0.526	1

$$\begin{cases} O_{rc} = \sum\limits_{i=1}^{k} \dfrac{(x_i - m_i)^2}{m_i} \\ X^2 = \sum\limits_{r=1}^{R}\sum\limits_{c=1}^{C} \dfrac{(O_{rc} - E_{rc})^2}{E_{rc}} \end{cases}, \quad (2)$$

где m_i – математическое ожидание частоты появления байта $i \in \{0,\ldots,255\}$ в анализируемой ПСП, для равномерного закона распределения и длины ПСП = 600 кбайт $m_i = \text{const} = 2400$; x_i – значение частоты появления байта $i \in \{0,\ldots,255\}$ в анализируемой ПСП; O_{rc}, E_{rc} – наблюдаемое и ожидаемое значения критерия на выборке данных размером r строк и c столбцов. Результаты представлены в таблице 2.

Полученные значения позволяют сделать вывод о равномерном характере распределений байт в рассматриваемых последовательностях и считать их псевдослучайными. Признаковое пространство на основе распределения байт может быть представлено выражением (3):

$$V_{Bytes} = \langle F(b_i)\rangle, \quad i \in \langle 0,\ldots,255\rangle, \quad (3)$$

где $F(b_i)$ – значение частоты появления байта i в анализируемой ПСП.

Результаты экспериментов по определению точности классификации зашифрованных и сжатых последовательностей различными алгоритмами машинного обучения при использовании признаков, являющихся значениями частот распределение байт представлены в выражении (4).

$$\begin{cases} Acc_{RF}(V_{Bytes}) = F(V_{Bytes}) = 0.9 \\ Acc_{DT}(V_{Bytes}) = F(V_{Bytes}) = 0.88 \\ Acc_{kNN}(V_{Bytes}) = F(V_{Bytes}) = 0.89 \end{cases}, \quad (4)$$

где V_{Bytes} – признаковое пространство, определяемое выражением (3), $Acc_{RF,DT,kNN}$ – метрики "точность классификации" ПСП соответствующим алгоритмом машинного обучения.

Полученные результаты позволяют построить классификатор зашифрованных и сжатых последовательностей, однако точность классификации недостаточно высока. Распределение байт для данных, обладающих высокой энтропией, например для зашифрованных и сжатых данных, подчиняется равномерному закону распределения. В литературе встречаются исследования, приводящие к выводу о невозможности классификации зашифрованных и сжатых данных с высокой точностью при использовании распределения байт [36], что объясняется другой предметной областью исследования, связанной с обнаружением вредоносного ПО, где размеры анализируемых данных недостаточно велики.

Кроме того, для построения модели ПСП недостаточно лишь распределения байт, так как они стремятся к равномерному распределению. Необходимо обнаружить новое вероятностное пространство признаков, позволяющее создать наиболее точную модель ПСП, вследствие чего были проведены расчёты средних значений энтропии Шеннона для зашифрованных и сжатых последовательностей согласно выражению (5):

$$H = -\sum_{i=0}^{255} p_i * \log_2 p_i, \quad (5)$$

где p_i – вероятность появления байта $i \in \{0,\ldots,255\}$ в анализируемой последовательности.

Были получены средние значения энтропии и значения некоторых статистических параметров распределений энтропии байт в анализируемых ПСП, результаты представлены в таблице 3, полученные согласно выражению 6:

$$\begin{cases} E_{mean} = \dfrac{\Sigma E(b_i)}{256} \\ E_{sko} = \sqrt{\dfrac{\Sigma(E(b_i) - E_{mean})^2}{256}} \\ E_{median} = X_{median} + i_{median} * \dfrac{\dfrac{\Sigma E(b_i)}{2} - Sum_{median-1}}{E(b_i)} \end{cases}, \quad (6)$$

где $E(b_i)$ – значение энтропии байта i в анализируемой ПСП; i_{median} – размер медианного интервала; X_{median} – нижняя граница медианного интервала, $Sum_{median\ -\ 1}$ – сумма значений энтропии байт, предшествующих медианному интервалу.

Проведены эксперименты по поиску более дискриминирующих статистических признаков, чем распределение и значения энтропии байт. В некоторых работах [28] применяются статистические те-

Таблица 3. Статистические признаки распределения значений энтропии байт в ПСП

№ п/п	Признак	Архивы	Шифры
1	E_{mean}	5.54424	5.54496
2	E_{sko}	0.00106	0.00002
3	E_{median}	5.54468	5.54496

Таблица 4. Оценка точности классификации алгоритмами машинного обучения при использовании модели ПСП на основе тестов NIST

№ п/п	Алгоритм	Accuracy
1	Random Forest	0.643
2	Decision Tree	0.575
3	kNN	0.684

сты NIST[4] для извлечения признаков из анализируемых данных.

Далее для построения модели ПСП были использованы статистические тесты NIST. Элементами модели являлись значения p-value, полученные в результате прохождения статистических тестов, диаграмма размаха полученных средних значений p-value представлена на рисунке 2.

В результате было установлено, при уровне значимости $\alpha = 0.05 \times 100\%$ зашифрованных последовательностей прошли все тесты на случайность, а сжатые последовательности – 98%. На основании выявленных закономерностей (схожие значения энтропии, прохождение тестов NIST на случайность) было принято решение обозначить зашифрованные и сжатые последовательности как псевдослучайные.

Таким образом, модель ПСП на основе распределений значений p-value тестов NIST представляет собой усредненные значения p-value, полученные в результате выполнения 11 статистических тестов: частотный побитовый (p_{freq}), блочный частотный (p_{Bfreq}), тесты кумулятивных сумм (p_{CS1}, p_{CS2}, тест на последовательность одинаковых бит (p_{Runs}), тест на самую длинную последовательность единиц в блоке (p_{LRuns}), тест рангов бинарных матриц (p_{Rank}), спектральный тест (p_{FFT}), тест на обнаружение неперекрывающихся шаблонов (p_{NOT}), тест на обнаружение перекрывающихся шаблонов (p_{OT}), тест приблизительной энтропии шаблонов (p_{AEnt}). Таким образом, признаковое пространство на основе тестов NIST является вектором, определяемым выражением (7):

$$V_{NIST} = \langle p_{freq}, p_{Bfreq}, p_{CS1}, p_{CS2}, p_{Runs}, p_{LRuns}, p_{Rank}, p_{FFT}, p_{NOT}, p_{OT}, p_{AEnt} \rangle, \quad (7)$$

где p_i – средние значения p-value по выборке данных для соответствующих тестов NIST.

Для оценки признакового пространства были проведены эксперименты, результаты представлены в таблице 4.

При использовании модели на основе тестов NIST точность классификации ПСП по сравнению с моделью на основе распределения байт значительно ухудшилась. Данный факт объясняется тем, что тесты NIST направлены на обнаружение закономерностей в анализируемых последовательностях, проверку их на случайность возникновения символов. Кроме того, процедура тестирования предполагает получение результатов в 10 возможных интервалах и вычислении p-value на основе табличных значений (критерий Хи-квадрат и равномерное распределение), что стирает статистические различия между рассматриваемыми ПСП, но позволяет оценить гипотезу о случайности анализируемых данных.

В статистических тестах NIST присутствует тест на проверку непересекающихся шаблонов. Методика тестирования предполагает разбиение анализируемой последовательности на N блоков данных длиной M бит и поиска в них бинарных шаблонов длиной m бит (подпоследовательностей). В случае обнаружения подпоследовательности его частота увеличивается на 1, и окно поиска сдвигается на следующий бит после последнего бита найденного шаблона. Для подтверждения гипотезы о случайности последовательности вычисляются математическое ожидание и дисперсия частоты встречаемости шаблонов в теоретически случайном распределении согласно выражениям (8) и (9):

$$\mu = \frac{M - m + 1}{2^m}, \quad (8)$$

$$\sigma^2 = M * \left(\frac{1}{2^m} - \frac{2m-1}{2^{2m}}\right)\frac{M - m + 1}{2^m}, \quad (9)$$

где μ – математическое ожидание частоты встречаемости шаблона длиной m бит в теоретически случайной последовательности, разделенной на M блоков даных.

Далее вычисляется значение критерия Хи-квадрат по полученным значениям частот встречаемости шаблонов подпоследовательностей согласно выражению (2).

Итоговое решение о прохождении теста принимается на основе вычисленного значения p-value согласно неполной гаммы-функции, выражение (10):

[4] A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, National Institute of Standards and Technology (NIST) Special Publication 800-22, Rev. 1a, Apr. 2010. https://csrc.nist.gov/publications/detail/sp/800-22/rev-1a/final (дата обращения 08.02.2021)

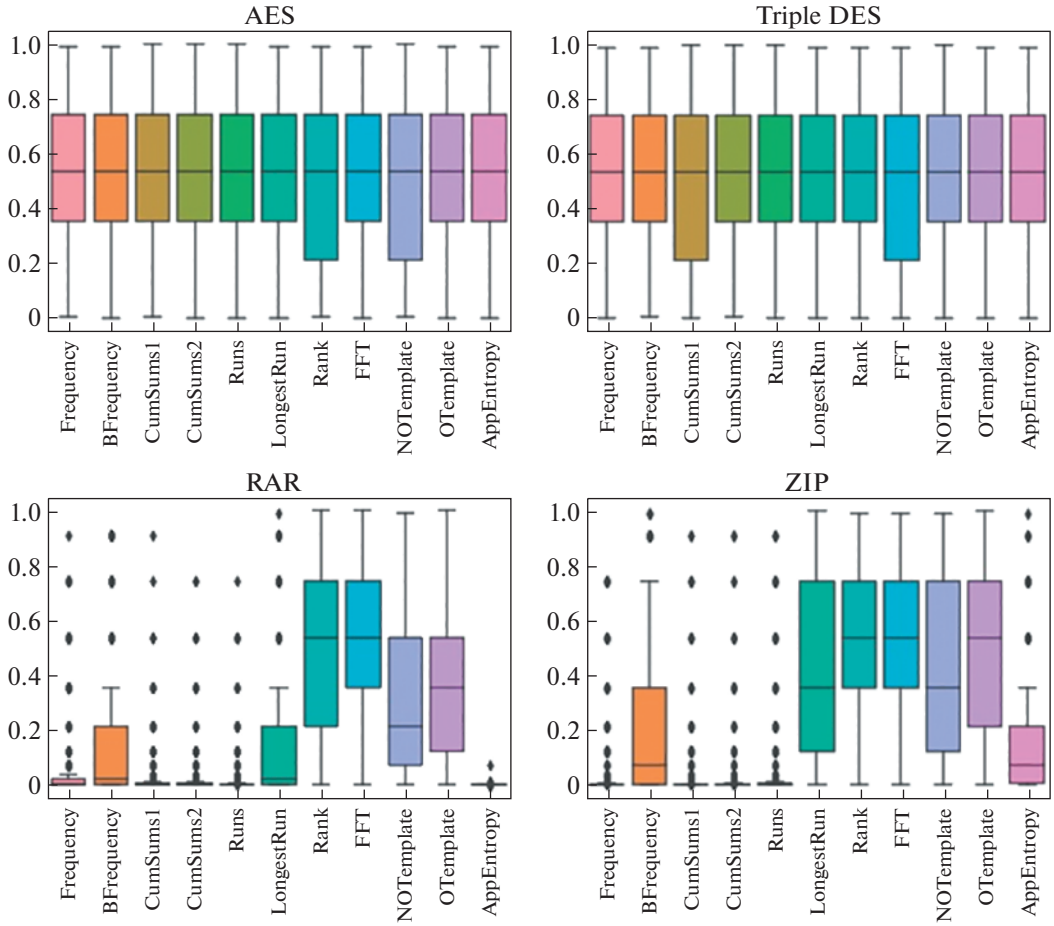

Рис. 2. Диаграммы размаха значений p-value тестов NIST для различных типов ПСП.

$$p - value = igamc\left(\frac{N}{2}, \frac{\chi^2_{obs}}{2}\right), \quad (10)$$

где *igamc* – неполная гамма-функция.

Неполная гамма-функция в общем виде определяется выражением (11):

$$Q(\alpha, x) = 1 - P(\alpha, x) = \frac{\Gamma(\alpha, x)}{\Gamma(\alpha)} = \frac{1}{\Gamma(\alpha)}\int_x^\infty e^{-t}t^{\alpha-1}dt, \quad (11)$$

где $Q(\alpha, 0) = 1$ и $Q(\alpha, \infty) = 0$.

Если значение p-value больше, чем выбранный порог α, то принимается решение о подтверждении нулевой гипотезы и анализируемая последовательность считается случайной с уровнем доверия, определяемым выбранным пороговым значением α. В противном случае нулевая гипотеза отвергается и последовательность считается неслучайной. Тест предназначен для тестирования последовательностей длиной не менее 10^6 бит, что составляет примерно 122 кбайт. В описании теста указано, что длина шаблонов должна быть 9 или 10 бит, количество блоков длины M бит $N \le 100$. Длина каждого блока задана в процедуре тестирования и равняется значению 131072. При использовании другого значения должны соблюдаться условия $M \ge 0.01 * n$ и $N = \left\lfloor \frac{n}{M} \right\rfloor$. Данные условия необходимы для обеспечения статистической значимости получаемых значений p-value.

В описании теста не указана причина использования подпоследовательностей длины 9 и 10 бит, кроме того, на их использование наложены достаточно жесткие ограничения в виде разделения подпоследовательности на определенное количество блоков определенной длины.

На основании данного теста была выдвинута гипотеза о возможности построения модели ПСП с использованием частот встречаемости подпоследовательностей ограниченной длины N бит.

Для проверки гипотезы и выявления статистических особенностей в анализируемых ПСП были проведены эксперименты по подсчету частот встречаемости подпоследовательностей различной длины $N = [4, \ldots, 11]$ бит без перекрытия. При их обнаружении дальнейший подсчёт начинается

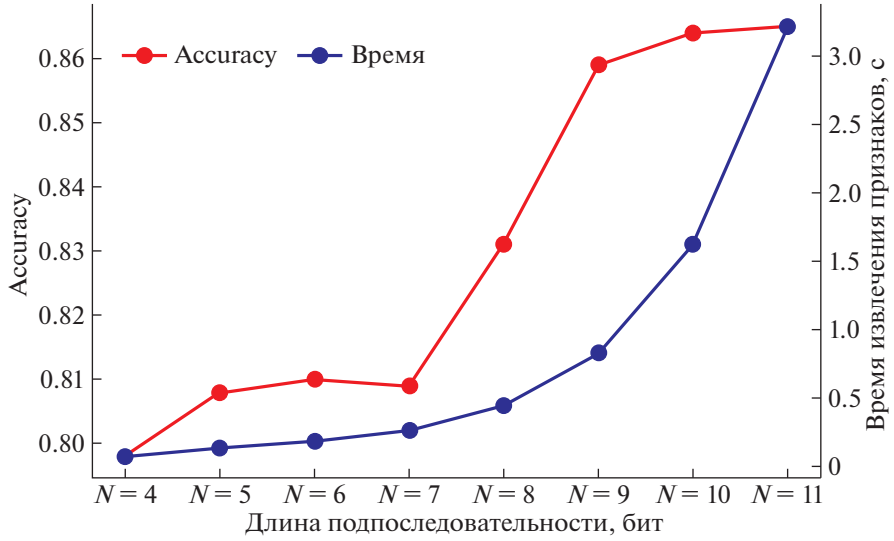

Рис. 3. Зависимость точности классификации ПСП алгоритмом построения случайного леса от длины подпоследовательностей, используемой моделью ПСП и времени извлечения признаков для одной последовательности.

со следующего бита после последнего бита подпоследовательности, если же совпадение не обнаружено, то происходит смещение на 1 бит. Подсчет частот встречаемости подпоследовательностей был выполнен согласно выражению (12):

$$f_j = F(j) = \frac{n(j)}{(M - N(j) + 1)}, \quad j \in \{0,\ldots,2^N\}, \quad (12)$$

где f_j – частота встречаемости подпоследовательности j в анализируемой ПСП; $n(j)$ – количество вхождений подпоследовательности j в анализируемую ПСП; M – длина анализируемой ПСП в битах; $N(j)$ – длина подпоследовательности j в битах.

Данный подход, в отличие от статистических тестов NIST, где используются только непериодические шаблоны определенной длины, позволит проверить все возможные подпоследовательности и определить их дискриминирующую способность. Внедрение в разрабатываемую модель периодических последовательностей не достаточно обосновано с точки зрения статистических критериев на проверку случайности, однако имеет рациональное объяснение с точки зрения получения признаков, характеризующих конкретный класс ПСП.

Данный способ отличается от подсчета подпоследовательностей длиной 8 бит, так как получаемое распределение подпоследовательностей вычисляется по формуле (1), для распределения байт выполняется подсчет количества байт, и оно стремится к равномерному распределению.

Таким образом, модель ПСП представляет собой вектор статистических характеристик, определяемый выражением (13):

$$V_{sub} = (f_j,\ldots,f_{2^N}) \quad (13)$$

Для определения оптимальной длины подпоследовательности были проведены эксперименты, исходная выборка данных была преобразована в 8 наборов данных $\{V\} = \{V_4,\ldots,V_{11}\}$, содержащих в себе векторы ПСП, определяемые выражением (4), и состоящие из значения частот встречаемости подпоследовательностей длины 4–11 бит. Поскольку полученные значения частот встречаемости подпоследовательностей являлись малыми величинами (порядка 1×10^{-5}), то они были приведены к логарифмическому масштабу согласно выражению (4), что дает прирост точности классификации ПСП алгоритмами машинного обучения [37]:

$$V_{Sub}^{\ln} = \ln(V_{Sub}) = (\ln(f_j),\ldots,\ln(f_{2^N})). \quad (14)$$

Полученные признаковые пространства подавались на вход алгоритма построения случайного леса, результаты представлены на рисунке (3).

Наиболее рациональным значением длины подпоследовательностей, в зависимости от точности классификации и времени, затрачиваемого на извлечение частот подпоследовательностей, является значение 9 бит.

Отличительной особенностью полученного распределения 9-битных подпоследовательностей (рис. 4), нормированного по среднему значению частоты, является отсутствие его равномерности, что вводит новое вероятностное пространство, позволяющее моделировать ПСП.

Для проверки разработанной модели на адекватность также были проведены эксперименты по классификации ПСП различными алгоритма-

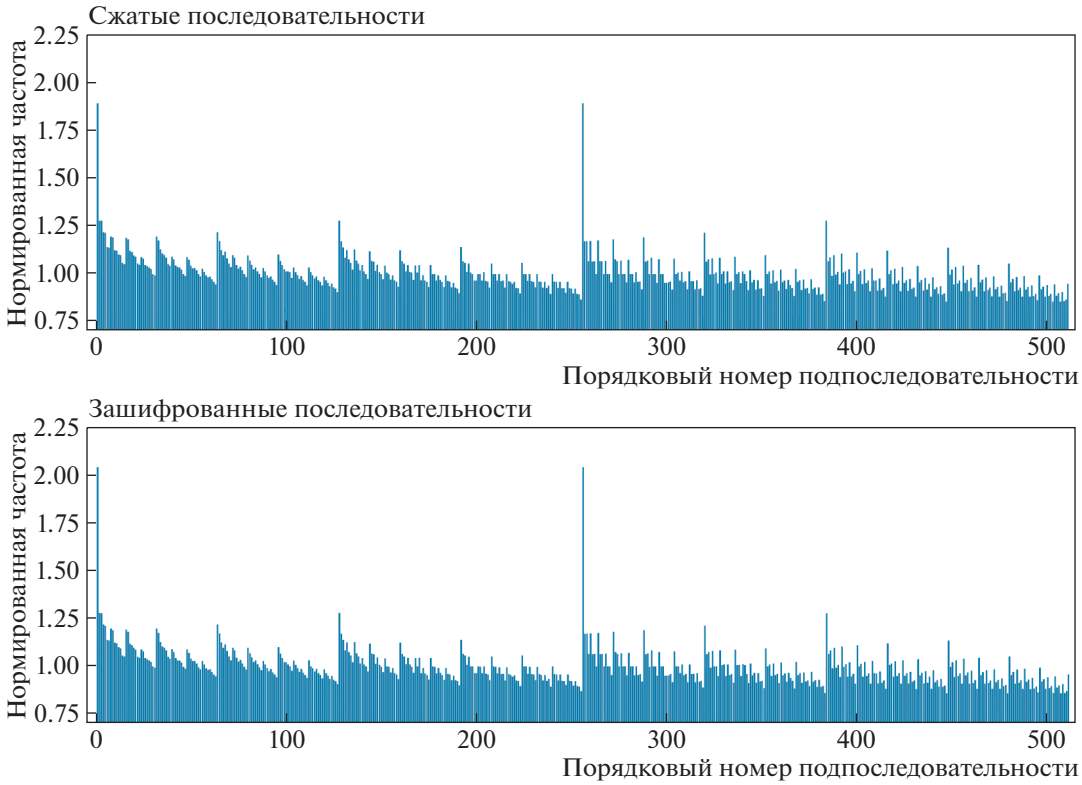

Рис. 4. Распределение 9-битных подпоследовательностей в ПСП.

ми машинного обучения, результаты представлены в таблице (5).

Полученные результаты свидетельствуют об ухудшении точности классификации алгоритмами построения случайного леса и дерева решений при использовании модели ПСП на основе учета частот встречаемости подпоследовательностей длины 9 бит, однако точность классификации алгоритмом k-ближайших соседей увеличилась на значение 0,012. Данный факт объясняется тем, что новая модель содержит в себе 512 значений вместо 256 при учете распределения байт. Алгоритмы построения случайного леса и дерева решений учитывают все имеющиеся в модели признаки, однако вероятно, не все из них имеют высокую дискриминирующую способность, из-за чего и происходит снижение точности классификации. Для алгоритма k-ближайших соседей наблюдается хоть и не значительная, но противоположная тенденция, так как данный алгоритм относится к метрическим и его точность возрастает с увеличением числа признаков.

Поскольку распределение байт и частоты подпоследовательностей длины 9 бит имеют различные распределения и вероятностные пространства, то наиболее рациональным шагом в проведении дальнейших исследований было построение синтезированной на их основе модели ПСП. К модели также были добавлены статистические признаки распределения байт: среднее значение (B_{mean}), среднеквадратическое отклонение (B_{sko}), минимальное (b_{min}) и максимальные (b_{max}) значения количества байт в ПСП, определяемые согласно выражению (15):

$$\begin{cases} B_{mean} = \dfrac{\sum n(b_i)}{256}, \\ B_{sko} = \sqrt{\dfrac{\sum (n(b_i) - B_{mean})^2}{256}}, \\ b_{min} = Min(n(b_i)), \\ b_{max} = Max(n(b_i)), \end{cases} \tag{15}$$

где $n(b_i)$ – количество появлений байта i в анализируемой ПСП.

Таблица 5. Оценка точности классификации алгоритмами машинного обучения при использовании разработанной модели ПСП

№ п/п	Алгоритм	Accuracy
1	Random Forest	0.859
2	Decision Tree	0.858
3	kNN	0.902

Таблица 6. Оценка точности классификации алгоритмами машинного обучения при использовании разработанной модели ПСП

№ п/п	Алгоритм	Accuracy
1	RF	0.894
2	DT	0.891
3	kNN	0.91

Таблица 7. Оценка точности классификации алгоритмами машинного обучения при использовании разработанной модели ПСП и алгоритма классификации ПСП

№ п/п	Алгоритм	Accuracy
1	RF+DT	0.97
2	DT	0.92
3	kNN	0.91

Таким образом, модель ПСП представляет собой вектор статистических характеристик, определяемый выражением (16):

$$V = (f_j, \ldots, f_{2^N}, b_0, \ldots, b_{255}, B_{mean}, B_{sko}, b_{min}, b_{max}), \quad (16)$$

Для оценки адекватности разработанной модели были проведены эксперименты по определению точности классификации ПСП алгоритмами машинного обучения, результаты представлены в таблице (6).

С помощью модели ПСП, учитывающей как равномерное распределение байт, так и новое распределение подпоследовательностей длины 9 бит, точность классификации ПСП увеличилась при использовании всех рассмотренных алгоритмов машинного обучения. Наивысшая точность наблюдалась у алгоритма kNN.

Однако полученные значения точности классификации ПСП не превышают значений при использовании модели ПСП на основе распределения байт. Кроме того, в настоящий момент модель ПСП представляет собой вектор статистических признаков длиной 772, что линейно влияет на время их извлечения и время выполнения классификации. Для преодоления указанных недостатков был разработан алгоритм классификации ПСП [38], учитывающий веса признаков, используемых в модели.

Оценка точности разработанного подхода представлена в таблице 7.

Для оценки влияния количества признаков на точность классификации были проведены эксперименты, результаты которых представлены на рисунке 5.

Наибольшая точность классификации была достигнута при использовании 110 признаков, однако точность классификации при 100 признаках меньше на 0.0001 пункт, а с ростом их количества в модели линейно увеличивается время выполнения процедуры извлечения признаков и классификации. Таким образом, на основе применения разработанного алгоритма классификации ПСП, учитывающего веса признаков и редуцирующего наименее значимые, был получен классификатор ПСП, достигающий точности классификации ПСП в 0.97.

На основании изложенного модель ПСП представляет собой вектор статистических характеристик согласно выражению (14).

Исследования в области информационной безопасности, связанные с применением алгоритмов машинного обучения также направлены

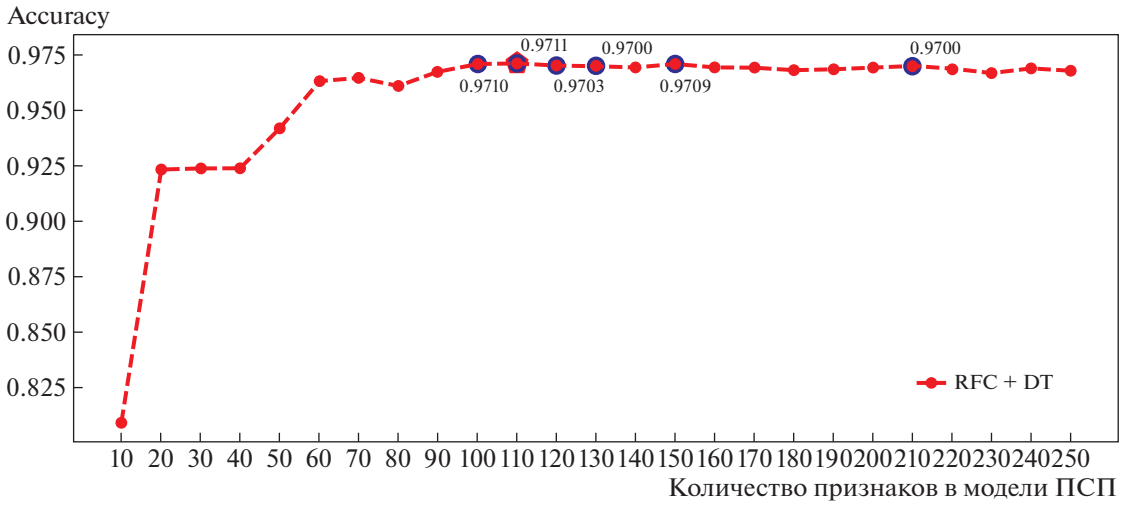

Рис. 5. Оценка влияния количества признаков в модели ПСП на точность классификации.

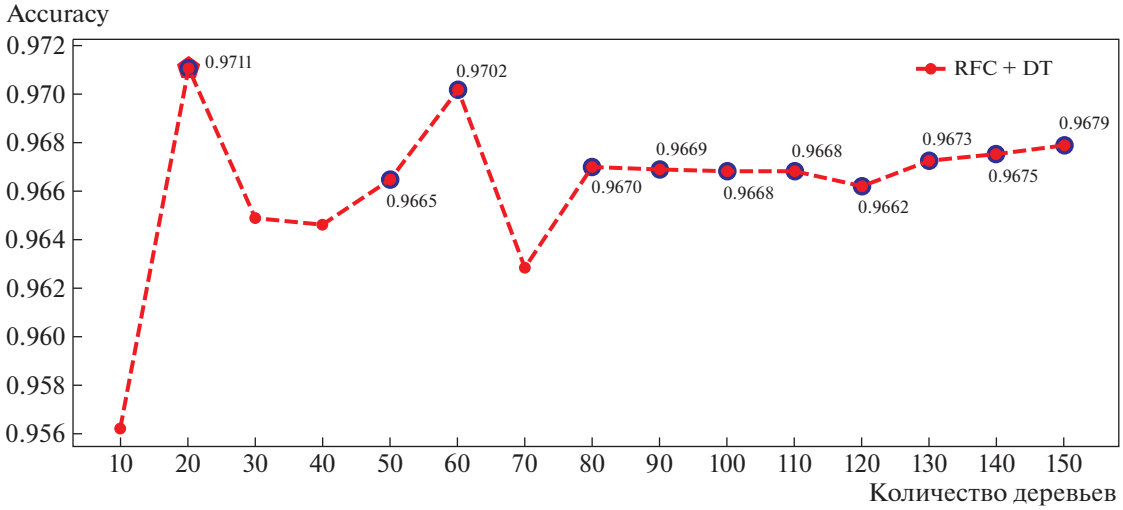

Рис. 6. Оценка влияния количества деревьев на точность классификации.

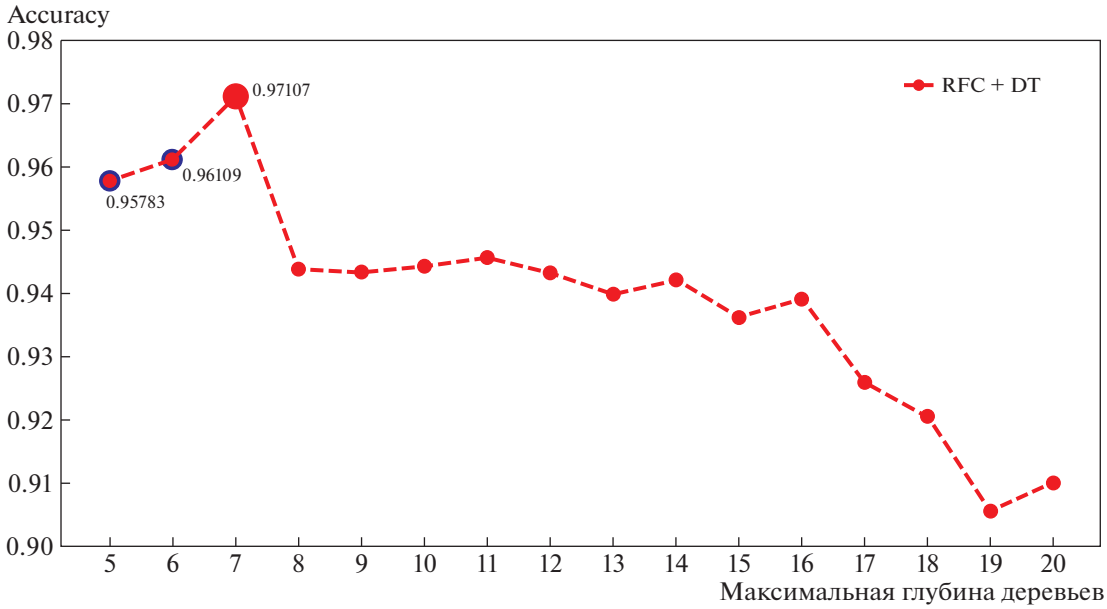

Рис. 7. Оценка влияния глубины деревьев на точность классификации.

на определение гиперпараметров используемых классификаторов [39]. Далее для достижения максимально возможной точности классификации ПСП необходимо установить параметры случайного леса и дерева решений.

С целью определения наиболее оптимальных параметров классификатора были проведены эксперименты по определению количества деревьев и их максимальной глубины: полученные результаты представлены на рисунках 6–7 соответственно.

Наиболее рациональным количеством деревьев в случайном лесу оказалось значение 20.

Наиболее рациональной глубиной деревьев в случайном лесу и дереве решений оказалось значение 7. Оно является локальным максимумом при незначительных колебаниях значений точности классификации и времени построения классификатора.

Также были проведены эксперименты по определению зависимости точности классификации от длины анализируемой последовательности. Результаты представлены на рисунке 8.

Высокой точности классификации удалось достичь при длине анализируемой последовательности в 600 кбайт, при ее увеличении с 50 кбайт

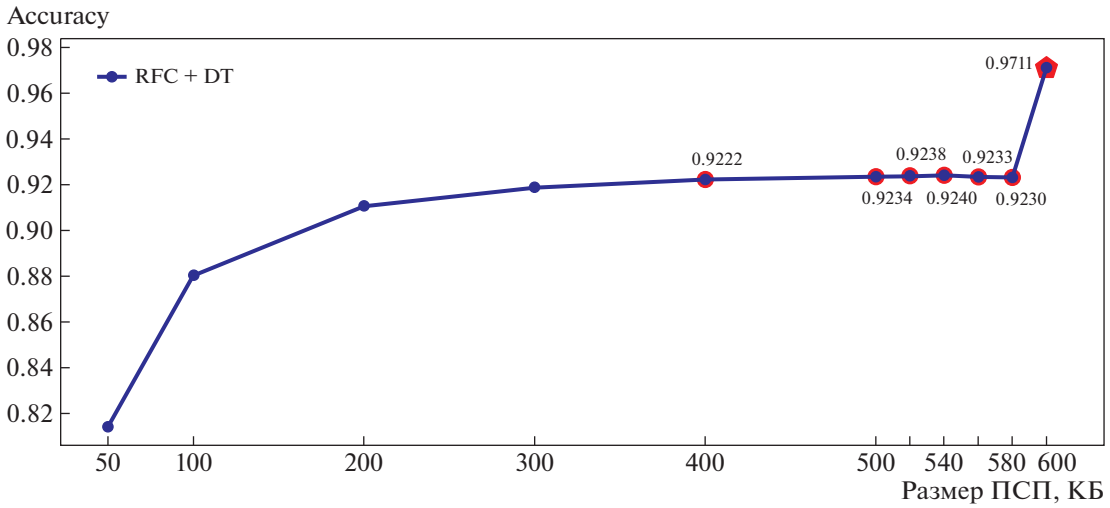

Рис. 8. Оценка точности классификации ПСП в зависимости от их длины.

увеличивается точность классификации. Данный факт объясняется возрастающими значениями частот подпоследовательностей и увеличением изменений в распределениях этих частот для зашифрованных и сжатых данных, что позволяет модели ПСП отражать статистические различия между ними.

4. ВЫВОДЫ

Основной вклад работы заключается в следующем:

1. Проведен анализ работ в области информационной безопасности на предмет использования методов классификации данных и алгоритмов машинного обучения. Сделан вывод о недостатках существующих подходов и предложены требования к разрабатываемому подходу классификации зашифрованных и сжатых данных до их передачи во внешнюю сеть.

2. Предложена модель ПСП, сформированных алгоритмами шифрования и сжатия данных (псевдослучайных последовательностей), отличающаяся от аналогов учетом распределения бинарных подпоследовательностей длины N бит.

3. Сформированы ограничения, используемые в работе: для достижения максимальной точности классификации ПСП необходимы относительно большие фрагменты данных – не менее 600 кбайт; при использовании фрагментов около 50 кбайт точность по метрике, доля правильных ответов составляет 0.81. К достоинствам предложенного способа классификации ПСП следует отнести отсутствие учета в модели ПСП заголовков файлов и “магических” байт сжатых ПСП.

Разработанный подход показал высокую точность классификации зашифрованных и сжатых последовательностей, равную 0.97, и может быть применен для улучшения существующих DLP-систем или внедрен в сервер электронной почты для проведения процедуры анализа почтовых вложений перед их отправкой за периметр организации.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Le D.C., Zincir-Heywood N., Heywood M.I.* Analyzing data granularity levels for insider threat detection using machine learning // IEEE Transactions on Network and Service Management, 2020. V. 17. № 1. P. 30–44.
2. *Bhatiaa A., Bahugunaa A.A., Tiwaria K., Haribabua K., Vishwakarmab D.* A Survey on Analyzing Encrypted Network Traffic of Mobile Devices // arXiv preprint arXiv:2006.12352 [cs.CR]. 2020.
3. *Mamun M.S.I., Ghorbani A.A., Stakhanova N.* (2016) An Entropy Based Encrypted Traffic Classifier // In: Qing S., Okamoto E., Kim K., Liu D. (eds) Information and Communications Security. ICICS 2015. Lecture Notes in Computer Science. V. 9543. Springer, Cham. https://doi.org/10.1007/978-3-319-29814-6_23.
4. *Shen M., Wei M., Zhu L., Wang M.* Classification of encrypted traffic with second-order markov chains and application attribute bigrams // IEEE Transactions on Information Forensics and Security. 2017. V. 12. № 8. P. 1830–1843. https://doi.org/10.1109/TIFS.2017.2692682
5. *Zhang Z., Kang C., Fu P., Cao Z., Li Z., Xiong G.* Metric learning with statistical features for network traffic classification // IEEE 36th International Performance Computing and Communications Conference (IPCCC), San Diego, CA. 2017. P. 1–7. https://doi.org/10.1109/PCCC.2017.8280467.

6. *Yang Y., Kang C., Gou G., Li Z. Xiong G.*, TLS/SSL Encrypted Traffic Classification with Autoencoder and Convolutional Neural Network // IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, United Kingdom, 2018. P. 362–369. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00079.
7. *Chen Y., Zang T., Zhang Y., Zhouz Y., Wang Y.* Rethinking Encrypted Traffic Classification: A Multi-Attribute Associated Fingerprint Approach // IEEE 27th International Conference on Network Protocols (ICNP), Chicago, IL, USA, 2019. P. 1–11. https://doi.org/10.1109/ICNP.2019.8888043.
8. *Wang P., Chen X., Ye F., Sun Z.* A survey of techniques for mobile service encrypted traffic classification using deep learning // IEEE Acces., 2019. V. 7. P. 54024–54033. https://doi.org/10.1109/ACCESS.2019.2912896
9. *Tang Z., Zeng X., Sheng Y.* Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification // International Journal of ICIC. 2019. V. 15. № 3. P. 845–860. https://doi.org/10.24507/ijicic.15.03.845
10. *Obasi T.C.* Encrypted Network Traffic Classification using Ensemble Learning Techniques // Doctoral dissertation, Carleton University, 2020. https://doi.org/10.22215/etd/2020-14171.
11. *Choudhury P., Kumar K.P., Nandi S., Athithan G.* An empirical approach towards characterization of encrypted and unencrypted VoIP traffic // Multimedia Tools and Applications. 2020. V. 79. № 1–2. P. 603–631. https://doi.org/10.1007/s11042-019-08088-w
12. *Yao Z., Ge J., Wu Y., Lin X., He R., Ma Y.* Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models // Journal of Network and Computer Applications. 2020. V. 166. P. 102711. https://doi.org/10.1016/j.jnca.2020.102711
13. *Baldini G., Hernandez-Ramos J.L., Nowak S., Neisse R., Nowak M.* Mitigation of Privacy Threats due to Encrypted Traffic Analysis through a Policy-Based Framework and MUD Profiles // Symmetry. 2020. V. 12. № 9. P. 1576. https://doi.org/10.3390/sym12091576
14. *Shen M., Liu Y., Zhu L., Xu K., Du X., Guizani N.* Optimizing Feature Selection for Efficient Encrypted Traffic Classification: A Systematic Approach // IEEE Network. 2020. V. 34. № 4. P. 20–27. https://doi.org/10.1109/MNET.011.1900366
15. *Panchenko A., Lanze F., Pennekamp J., Engel T., Zinnen A., Henze M., Wehrle K.* Website Fingerprinting at Internet Scale // Network and Distributed System Security Symp. 2016. P. 21–24. https://doi.org/10.14722/ndss.2016.23477
16. *Wei S., Ding Y., Han X.* TDSC: Two-stage DDoS detection and defense system based on clustering // In 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). 2017. P. 101–102. https://doi.org/10.1109/DSN-W.2017.11.
17. *Sahoo K.S., Tripathy B.K., Naik K., Ramasubbareddy S., Balusamy B., Khari M., Burgos D.* An evolutionary SVM model for DDOS attack detection in software defined networks // IEEE Access. 2020. V. 8. P. 132502–132513. https://doi.org/10.1109/ACCESS.2020.3009733
18. *Grechishnikov E.V., Dobryshin M.M., Kochedykov S.S., Novoselcev V.I.* Algorithmic model of functioning of the system to detect and counter cyber attacks on virtual private network // Journal of Physics: Conference Series. 2019. V. 1203. № 1. P. 012064. https://doi.org/10.1088/1742-6596/1203/1/012064
19. *Добрышин М.М.* Предложение по совершенствованию систем противодействия DDoS-атакам // Телекоммуникации. 2018. № 10. С. 32–38. eLIBRARY ID: 36284311.
20. *Добрышин М.М., Спирин А.А., Лактионов А.Д.* Предложения по раннему обнаружению деструктивных воздействий Botnet на компьютерные сети связи. // Телекоммуникации. 2020. №. 12. С. 25–29. eLIBRARY ID: 44404522
21. *Zhu L., Tang X., Shen M., Du X., Guizani M.* Privacy-preserving DDoS attack detection using cross-domain traffic in software defined networks // IEEE Journal on Selected Areas in Communications. 2018. V. 36. № 3. P. 628–643. https://doi.org/10.1109/JSAC.2018.2815442
22. *Wang F., Quach T.T., Wheeler J., Aimone J.B., James, C.D.* Sparse coding for n-gram feature extraction and training for file fragment classification // IEEE Transactions on Information Forensics and Security. 2018. V. 13. № 10. P. 2553–2562. https://doi.org/10.1109/TIFS.2018.2823697
23. *Karampidis K., Papadourakis G.* File type identification-computational intelligence for digital forensics // Journal of Digital Forensics, Security and Law. 2017. V. 12. № 2. P. 6. https://doi.org/10.15394/jdfsl.2017.1472
24. *Karampidis K., Kavallieratou E., Papadourakis G.* Comparison of Classification Algorithms for File Type Detection. A Digital Forensics Perspective // Polybits. 2017. V. 56. P. 15–20. https://doi.org/10.17562/PB-56-2
25. *Kozachok A.V.* Development of a Heuristic Mechanism for Detection of Malware Programs Based on Hidden Markov Models // Automatic Control and Computer Sciences. 2018. V. 52. № 8. P. 1117–1123. https://doi.org/10.3103/S0146411618080345
26. *Srinivas M., Nayak A., Bhatt A.* Forged File Detection and Steganographic content Identification (FFDASCI) using Deep Learning Techniques // In CLEF (Working Notes). 2019. http://ceur-ws.org/Vol-2380/paper_142.pdf
27. *Konaray S.K., Toprak A., Pek G.M., Akçekoce H., Kılınç D.* Detecting File Types Using Machine Learning Algorithms // 2019 Innovations in Intelligent Systems and Applications Conference (ASYU). 2019. P. 1–4. https://doi.org/10.1109/ASYU48272.2019.8946393
28. *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. V. 14. № 11. P. 2916–2926. https://doi.org/10.1109/TIFS.2019.2911156

29. *De Gaspari F., Hitaj D., Pagnotta G., De Carli L., Mancini L.V.* EnCoD: Distinguishing Compressed and Encrypted File Fragments // International Conference on Network and System Security, Springer, Cham. 2020. P. 42–62. https://doi.org/10.1007/978-3-030-65745-1_3.

30. *Mousavi S.S.* Detecting Disk Sectors Data Types Using Hidden Markov Model // 17th International ISC Conference on Information Security and Cryptology (ISCISC). 2020. P. 60–64. https://doi.org/10.1109/ISCISC51277.2020.9261906.

31. *Cheng L., Liu F., Yao D.* Enterprise data breach: causes, challenges, prevention, and future directions // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2017. V. 7. № 5. C. e1211.

32. *Doroud H.* et al. Speeding-up dpi traffic classification with chaining // IEEE Global Communications Conference (GLOBECOM). IEEE. 2018. C. 1–6.

33. *Hahn D., Apthorpe N., Feamster N.* Detecting compressed cleartext traffic from consumer internet of things devices // arXiv preprint arXiv:1805.02722. 2018.

34. *Wood D., Apthorpe N., Feamster N.* Cleartext data transmissions in consumer iot medical devices // Proceedings of the 2017 Workshop on Internet of Things Security and Privacy. 2017. C. 7–12.

35. *Scaife N., Carter H., Traynor P., Butler K. R.* Cryptolock (and drop it): stopping ransomware attacks on user data // IEEE 36th International Conference on Distributed Computing Systems (ICDCS). 2016. P. 303–312. https://doi.org/10.1109/ICDCS.2016.46.

36. *Raff E., Zak R., Cox R., Sylvester J., Yacci P., Ward R., Nicholas C.* An investigation of byte n-gram features for malware classification // Journal of Computer Virology and Hacking Techniques. 2018. V. 14. № 1. P. 1–20. https://doi.org/10.1007/s11416-016-0283-1

37. *Козачок А.В., Спирин А.А.* Алгоритм классификации псевдослучайных последовательностей // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2020. № 1. С. 87–98. https://doi.org/10.17308/sait.2020.1/2595

38. *Козачок А.В., Спирин А.А., Голембиовская О.М.* Алгоритм классификации псевдослучайных последовательностей на основе построения случайного леса // Доклады Томского государственного университета систем управления и радиоэлектроники. 2020. Т. 23. № 3. С. 55–60.

39. *Kozachok A.V., Kozachok V.I.* Construction and evaluation of the new heuristic malware detection mechanism based on executable files static analysis // Journal of Computer Virology and Hacking Techniques, 2018. V. 14. № 3. P. 225–231. https://doi.org/10.1007/s11416-017-0309-3