ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

На правах рукописи

Спирин Андрей Андреевич

Защита от утечки информации на основе разделения зашифрованных и сжатых данных

05.13.19 – Методы и системы защиты информации,
информационная безопасность

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

г. Орел
2022

Работа выполнена в федеральном государственном казённом военном образовательном учреждении высшего образования «Академия Федеральной службы охраны Российской Федерации», г. Орёл.

Научный руководитель:
доктор технических наук, доцент Козачок Александр Васильевич

Официальные оппоненты:
Оков Игорь Николаевич
заслуженный изобретатель РФ, доктор технических наук, профессор

Петренко Сергей Анатольевич

доктор технических наук, профессор

Ведущая организация:
Краснодарское высшее военное училище им. С.М.Штеменко

Защита состоится _____ 2022 г. в ___ часов
на заседании диссертационного совета У 05.13.19 федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого» (195251, г. Санкт-Петербург, ул. Политехническая, д. 29, корпус ___, аудитория ___).

С диссертацией можно ознакомиться в библиотеке и на сайте www.spbstu.ru федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого».

Автореферат разослан _____.

Ученый секретарь
диссертационного совета У.05.13.19 ,
канд. физ.-мат. наук Шенец Николай Николаевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы.

Информационно-аналитические агентства выявляют устойчивый рост количества утечек конфиденциальных данных. Согласно отчета экспретно-аналитического центра ГК Infowatch в 2020 году в России около 79 % зафиксированных утечек произошли по вине внутреннего нарушителя, 77 % из них являлись умышленными.

Для защиты от утечек информации применяют средства обнаружения и предотвращения утечек информации, использующие поиск цифровых сигнатур и слепков, регулярных выражений, алгоритмы машинного обучения и поведенческие подходы. Однако существующие средства не способны с высокой точностью обнаружить передачу зашифрованных данных в виду их статистической схожести и высокой энтропии, кроме того существующие механизмы классификации используют заголовки контейнеров файлов, содержащих в себе цифровые сигнатуры.

Анализ работ в предметной области исследования позволил выявить практическую проблему - низкую точность классификации зашифрованных и сжатых данных (около 70 %), обладающих высокой энтропией и использование заголовков файлов, содержащих в себе цифровые сигнатуры сжатых последовательностей. По этой причине задача классификации зашифрованных и сжатых данных для защиты от утечек информации в зашифрованном виде является актуальной.

Степень разработанности темы. Проблеме защиты от утечек информации посвятили свои исследования отечественные и зарубежные ученые: Д.П. Зегжда, П.Д. Зегжда, В.П. Лось, А.А. Грушо, А.А. Шелупанов, Е.Ю. Павленко, Е.Б. Александрова и Р. Альшамари, П. Дорфингер, Г. Панхольцер, Ю. Ванг, З. Жанг, К. Пападопулус, Д. Массей, А. Р. Хакпур, А. Х. Лиу. Наибольший вклад в область изучения, разработки и совершенствования методов классификации зашифрованных и сжатых данных внесли следующие ученые: В.С. Матвеева, А.И. Гетьман, М.К. Иконникова, Ф. Казино, К. Р. Чо, К. Патсакис, З. Танг, Х. Зенг, Ю. Шенг, Д. Хахн, Н. Апторп, Н. Фимстер и др. В ходе проведения исследований авторами разработаны методы классификации зашифрованных и сжатых данных, определены их достоинства и недостатки, практически реализованы и апробированы результаты проведенных исследований, установлены проблемные аспекты исследуемой области.

Объектом исследования являются псевдослучайные последовательности, сформированные алгоритмами шифрования и сжатия данных.

Предмет исследования: модели, методы и алгоритмы классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных.

Целью работы является повышение точности классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных.

Научная задача заключается в разработке метода классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде и способа, его реализующего.

Для достижения поставленной цели необходимо решить следующие частные научные **задачи**:

1. Произвести анализ особенностей функционирования современных средств обнаружения и предотвращения утечек информации, выявить существующие ограничения, связанные с обнаружением зашифрованных и сжатых данных и обосновать выбор признакового пространства для моделирования псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных.
2. Разработать модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, отличающуюся учетом их статистических характеристик.
3. Разработать метод классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающий дискриминирующую способность их статистических признаков.
4. Разработать способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде.
5. Реализовать на практике способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных и оценить область его эффективного применения, проанализировать возникающие ошибки классификации.

Научная новизна:

1. Разработана модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, отличающаяся учетом их статистических характеристик.
2. Разработан метод классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающий дискриминирующую способность их статистических признаков.
3. Реализован способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде.

Теоретическая значимость заключается в разработке модели псевдослучайных последовательностей, учитывающей статистические характеристики распределения байт и битовых последовательностей, и метода классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, имеющего точность выше, чем у аналогов, с учетом отсутствия служебных полей в анализируемых данных.

Практическая ценность заключается в повышении точности классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде и отказе от контекстных признаков.

Методология и методы исследования. В ходе проведения диссертационного исследования использованы методы математического моделирования, математической статистики, теории распознавания образов.

Основные положения, выносимые на защиту:

1. Модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, отличающаяся учетом их статистических характеристик.
2. Метод классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающий дискриминирующую способность их статистических признаков.
3. Способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде.

Достоверность полученных в настоящей диссертационной работе результатов подтверждается корректным использованием математического аппарата, применением апробированных математических моделей, результатами экспериментальных исследований, актом о внедрении результатов, положительными результатами обсуждений основных положений работы на научно-технических конференциях.

Апробация работы. Основные результаты работы докладывались на следующих конференциях:

- XXVIII, XXIX Научно-техническая конференция Методы и технические средства обеспечения безопасности информации (Санкт-Петербург, 2019, 2020);
- VIII Международная научно-техническая и научно-методическая конференция Актуальные проблемы инфотелекоммуникаций в науке и образовании (Санкт-Петербург, 2019);
- X, XI Международная научно-техническая конференция Безопасные информационные технологии (Москва, 2019, 2021);
- XXII Международная научно-практическая конференция РусКрипто 2020 (Москва, 2020);

– Всероссийский конкурс-конференция студентов и аспирантов по информационной безопасности SIBINFO-2020 (Томск, 2020);
– Ежегодная научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского (Москва, 2020, 2021);
– Международная конференция Иванниковские чтения 2020 (Орёл, 2020);
– XIII Всероссийская межведомственная научная конференция Актуальные направления развития систем охраны, специальной связи и информации для нужд органов государственной власти Российской Федерации (Орёл, 2021);

Личный вклад. Все выносимые на защиту научные результаты получены соискателем лично либо при его непосредственном участии. В работах по теме диссертации, опубликованных в соавторстве, соискателем выполнено следующее:

– проведен сравнительный анализ особенностей функционирования современных средств обнаружения и предотвращения утечек информации, выявлены существующие ограничения, связанные с обнаружением зашифрованных и сжатых данных;
– обоснован выбор признакового пространства для моделирования псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных;
– проведен сравнительный анализ алгоритмов машинного обучения в предметной области исследований;
– разработана модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, отличающаяся учетом их статистических характеристик;
– разработан метод классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающий дискриминирующую способность их статистических признаков;
– разработан способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных для защиты от утечки информации в зашифрованном виде;
– выполнена экспериментальная и аналитическая оценка основных параметров разработанного метода защиты от утечки информации.

Публикации. Основные результаты по теме диссертации изложены в 18 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК , 4 — в периодических научных журналах, индексируемых Web of Science и Scopus, 2 — в других научных журналах, 8 — в тезисах докладов, получено 2 свидетельства о государственной регистрации программ для ЭВМ, 1 патент на изобретение.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** раскрыта актуальность научных исследований в области классификации зашифрованных и сжатых данных, определены объект, предмет и цель исследования, сформулирована научная задача, приведены основные положения, выносимые на защиту, представлена структура диссертационного исследования.

В первой главе обоснована актуальность задачи классификации зашифрованных и сжатых данных. Приведены исследования информационно-аналитических агентств о сохраняющемся тренде повышения доли внутренних нарушителей как источника зарегистрированных случаев утечки конфиденциальных данных и их особенно высокой доли в России (рис. 1).

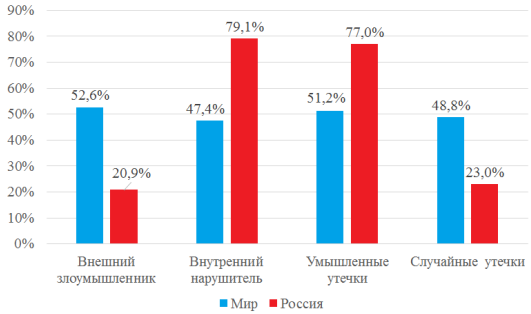

Рис. 1 — Распределение источников зарегистрированных утечек конфиденциальных данных и их характера в 2020 году

Анализ зарегистрированных утечек конфиденциальной информации показал, что в большинстве случаев несанкционированное распространение защищаемой информации осуществлялось внутренними нарушителями, что свидетельствует с одной стороны о высоких результатах, достигнутых при отражении внешних атак, а с другой – недостаточности и устаревании организационных способов борьбы с внутренним нарушителем.

Анализ существующих подходов к классификации зашифрованных и сжатых данных, применяемых в средствах обнаружения и предотвращения утечек информации, позволил выделить две группы методов. Контентные методы позволяют обнаруживать в данных цифровые сигнатуры, регулярные выражения и другие специфические атрибуты конфиденциальной информации. Ряд методов в данном классе относятся к статистическим, выполняющим подсчет энтропии всей последовательности, либо ее части. Контекстные методы выполняют анализ служебной информации, специфичной для конкретного протокола передачи данных или методов хранения, в данный класс также можно отнести поведенческие и эвристические

механизмы. Внутренний нарушитель может преодолеть все существующие средства защиты, поскольку может изменить структуру контейнера передачи данных для обхода контекстных и контентных механизмов анализа данных. Зашифрованные конфиденциальные данные могут быть переданы посредством электронной почты в виде вложения с измененным расширением. Поскольку вложение в зашифрованном виде не содержит цифровых сигнатур, регулярных выражений и других отличительных признаков, то оно не будет распознано как подозрительное и запрещено к передаче.

Используемые в настоящее время подходы для определения типа передаваемых данных по большей части относятся к контекстным, а контентные используют подсчет энтропии, что приводит к низкой точности классификации передаваемых данных в зашифрованном или сжатом виде. Кроме того, используются заголовки файлов, содержащие в себе цифровые сигнатуры сжатых последовательностей, что позволяет нарушителю обойти подобные механизмы защиты данных. Анализ литературных источников также выявляет практическую проблему низкой точности классификации зашифрованных и сжатых данных и неспособноси современных средств обнаружения и предотвращения утечек информации своевременно обнаруживать их передачу.

Выбор статистических методов в качестве средств извлечения признаков ПСП, позволил отказаться от методов классификации зашифрованных и сжатых данных на основе сигнатур и служебной информации, которые может подделать внутренний нарушитель.

Согласно методике оценки угроз безопасности информации[1] были выбраны модель внутреннего нарушителя и угрозы, реализуемые им (рис. 2).

Уровень опасности угрозы[2] утечки конфиденциальных данных (УБИ.028: Угроза использования альтернативных путей доступа к ресурсам и УБИ.067: Угроза неправомерного ознакомления с защищаемой информацией) является высоким. Он обусловлен рядом причин:

- возможностью реализации утечек конфиденциальных данных в зашифрованном и сжатом виде – достаточно наличия программного обеспечения, свободно распространяемого в сети Интернет (с открытым исходным кодом);
- наличием у внутреннего нарушителя локального доступа к данным;
- значимостью данных (мировая практика имеет множество примеров, когда утечка лишь персональных данных приводила к серьезным судебным искам и в дальнейшем к банкротству компаний, не принимая во внимания коммерческую и другие виды тайн);

[1] Методический документ. Утв. ФСТЭК России 05.02.2021 г., https://fstec.ru/

[2] Банк данных угроз безопасности информации ФСТЭК России, https://bdu.fstec.ru/threat

– доступностью сведений и средств реализации угроз безопасности информации.

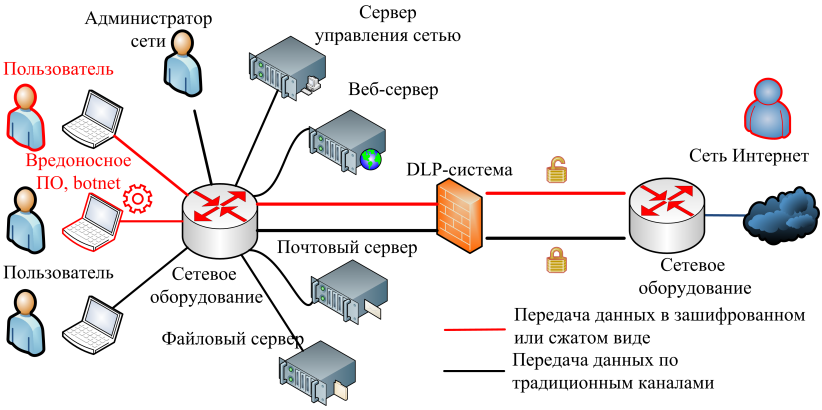

Рис. 2 — Модель угроз и модель нарушителя

Также была осуществлена формальная постановка научной задачи, которая относится к теории распознавания образов, заключающаяся в классификации ПСП на два класса. Введены соответствующие ограничения и допущения:

– размер анализируемой ПСП не менее 600 кбайт;
– при процедуре классификации отбрасываются заголовки файлов;
– проводится бинарная классификация зашифрованных алгоритмами AES, 3DES, Camellia, RC4, ГОСТ 34.12 Кузнечик и сжатых алгоритмами BZ2, ZIP, RAR, 7Z, GZ, XZ ПСП;

Для проведения экспериментов по оценке точности и эффективности разрабатываемого алгоритма обоснован размер выборки ПСП, используемой при проведении исследований. Размер составил 22000 файлов двух классов: сформированных ПО OpenSSL и ПО, реализующим ГОСТ 34.12–Кузнечик и сформированных ПО WinRAR и 7Zip.

Во второй главе представлена модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, отличающаяся учетом их статистических характеристик (ПСП). На основании анализа исследований в предметной области было определено признаковое пространство на основе распределения байт в ПСП. Проведено статистическое тестирование гипотез о равномерности полученных распределений байт для зашифрованных и сжатых данных на основе теста Пирсона Хи-квадрат при уровне значимости $\alpha = 0{,}05$ и 256 степенях свободы. Результаты представлены в таблице 1.

Таблица 1 — Статистическое тестирование гипотез о равномерности распределений байт в ПСП

ПСП	Референсное распределение	Критическое зн. критерия	Расчетное зн. критерия	P-value
Архивы	Равномерное	293,248	50,811	1
Шифры	Равномерное	293,248	0,526	1

Полученные значения не позволяют отвергнуть гипотезу о равномерности распределений байт в ПСП с высоким уровнем статистической значимости.

Для обнаружения статистических признаков иного вероятностного пространства были проведены эксперименты по выделению признаков на основе подсчета энтропии байт (табл. 2) в ПСП и результатах выполнения статистических тестов NIST.

Таблица 2 — Подсчет энтропии распределений байт в ПСП

Признак	Архивы	Шифры
H_{mean}	5,54424	5,54496
H_{sko}	0,00106	0,00002

Полученные значения статистических признаков энтропии байт, вычисленных согласно выражения 1 не позволили использовать их в качестве признакового пространства для построения классификатора, поскольку разница значений энтропии для ПСП незначительна.

$$
\begin{aligned}
H_{mean} &= \frac{\sum_0^{255} H(b_i)}{256}, \\
H_{sko} &= \sqrt{\frac{\sum_0^{255} (H(b_i) - H_{mean})^2}{256}}, \\
H(b_i) &= -\sum_0^{255} p_i \cdot \log_2 p_i,
\end{aligned}
\tag{1}
$$

где $H(b_i)$ – энтропия байта b_i.

Были сформированы признаковые пространства на основе распределения байт и результатов выполнения тестов NIST. Одним из тестов NIST является тест по подсчету количества непересекающихся непериодических шаблонов длиной 9 или 10 бит. На основании данного теста было сформировано другое признаковое пространство, состоящее из частот встречаемости подпоследовательностей ограниченной длины, определяемые выражением 2.

$$
f_j = \frac{n_j}{M - N_j + 1}, \tag{2}
$$

где N_j – длина подпоследовательности j в битах, n_j – количество появлений подпоследовательности j в анализируемой ПСП, M – длина анализируемой ПСП в битах.

Полученные значения частот были приведены к логарифмическому масштабу для повышения точности классификации алгоритмами машинного обучения. Таким образом, было получено четыре признаковых пространства, определяемые выражением 3.

$$
\begin{aligned}
&V_{bytes} = \langle F(b_i), B_{mean}, B_{sko}, b_{min}, b_{max} \rangle, \\
&V_{NIST} = \langle p_{freq}, p_{Bfreq}, p_{CS1}, p_{CS2}, p_{Runs}, p_{LRuns}, p_{Rank}, \\
&p_{FFT}, p_{NOT}, p_{OT}, p_{Aent} \rangle, \\
&V_{Sub}^{ln} = \langle f_j \dots f_{2^N} \rangle, \\
&V_{Stat} = (\langle f_j \dots f_{2^N} \rangle, \langle b_0 \dots b_{255}, B_{mean}, B_{sko}, b_{min}, b_{max} \rangle),
\end{aligned}
\tag{3}
$$

где $V_{Stat} = V_{bytes} \parallel V_{Sub}^{ln}$ – признаковое пространство на основе распределения байт и частот встречаемости подпоследовательностей ограниченной длины.

Результаты проведенных экспериментов по классификации ПСП на основе полученных признаковых пространств представлены в таблице 3.

Таблица 3 — Точность классификации ПСП в зависимости от признаковых пространств

Алгоритм	V_{bytes}	V_{NIST}	V_{Sub}^{ln}	V_{stat}
RF	0,9	0,64	0,86	0,89
DT	0,88	0,58	0,86	0,89
kNN	0,89	0,55	0,9	0,91

Для определения оптимальной длины подпоследовательностей были проведены эксперименты по определению зависимости точности классификации от длины подпоследовательностей, результаты представлены на рисунке 3. При увеличении длины подпоследовательностей N в битах увеличивается точность классификации ПСП, однако после значения в 9 бит рост становится незначительным, порядка тысячных долей, но время, затрачиваемое на извлечение признаков начинает увеличиваться экспоненциально. По этой причине было выбрано значение в 9 бит, как наиболее рациональное с точки зрения времени извлечения признаков и точности классификации.

Распределение нормированных по среднему значению частот подпоследовательностей длины 9 бит представлено на рисунке 4.

Полученное распределение частот визуально отличается от равномерного распределения байт, что удалось подтвердить экспериментально и сделать выбор другого вероятностного пространства, описываемого распределением частот подпоследовательностей длины 9 бит. Таким образом

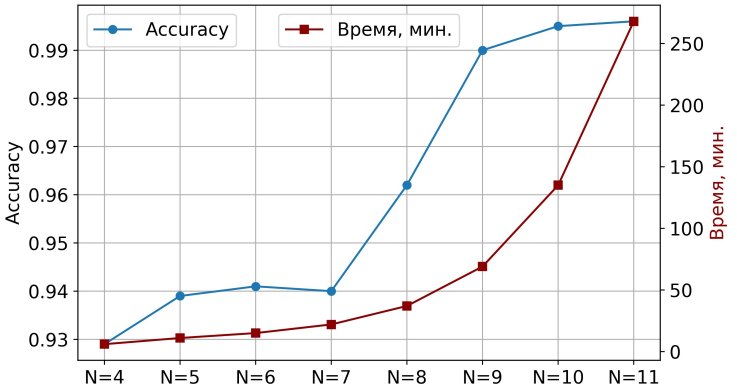

Рис. 3 — Зависимость точности классификации ПСП и времени извлечения признаков от длины подпоследовательности N в битах

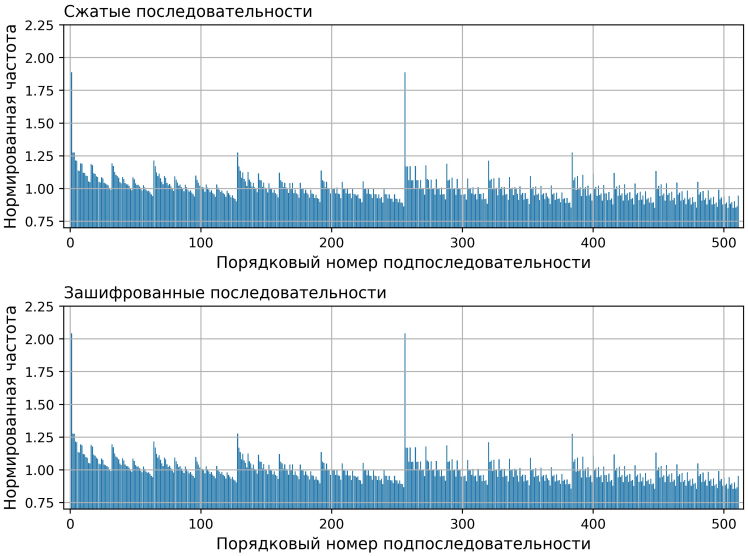

Рис. 4 — Распределение нормированных частот подпоследовательностей длины 9 бит

моделью ПСП будем считать вектор статистических характеристик, полученных из распределения байт и частот встречаемости подпоследовательностей длины 9 бит согласно выражению 4.

$$V_{Stat} = (\langle f_j \dots f_{2^N} \rangle, \langle b_0 \dots b_{255}, B_{mean}, B_{sko}, b_{min}, b_{max} \rangle) = \langle v_1 \dots v_\Phi \rangle \qquad (4)$$

Полученное признаковое пространство содержит 772 статистических признака: 256 значений частот встречаемости байт и 4 характеристики их распределения, 512 значений частот появлений подпоследовательностей длины 9 бит. Большое количество признаков теоретически позволяет наиболее точно моделировать псевдослучайные последовательности, сформированные алгоритмами шифрования и сжатия данных, однако из теории распознавания образов известно, что необходимо оптимизировать количество признаков для преодоления эффекта переобучения классификатора, который приводит к снижению точности классификации.

Для уменьшения количества анализируемых классификатором параметров был разработан алгоритм редуцирования признакового пространства модели псевдослучайных последовательностей, представленный на рис. 5

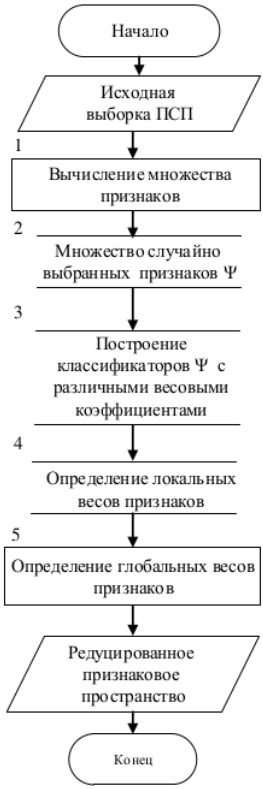

Рис. 5 — Алгоритм редуцирования признакового пространства

Редуцирование признакового пространства позволило повысить точность классификации ПСП, результаты экспериментов представлены в таблице 4.

Таблица 4 — Точность классификации ПСП в зависимости от признаковых пространств

Алгоритм	V_{bytes}	V_{NIST}	V_{Sub}^{ln}	V_{stat}	$V_{stat}^{reduced}$
RF	0,9	0,64	0,86	0,89	0,97
DT	0,88	0,58	0,86	0,89	0,93
kNN	0,89	0,55	0,9	0,91	0,88

Третья глава посвящена разработке метода классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающему дискриминирующую способность их статистических признаков и способу, его реализующему.

На первоначальном этапе определяются файлы, обладающие высокой энтропией и имеющие равномерное распределение байт, т.е. зашифрованные и сжатые. В блоках 1-7 происходит классификация потенциально опасных данных от легитимно используемых в корпоративных сетях.

При превышении значения энтропии файла значения 6,5 данный файл потенциально может содержать в себе зашифрованные или сжатые последовательности. Таким образом в блоке 1 происходит вычислении энтропии анализируемых данных. При превышении указанного порогового значения файл передается в блок 3, в обратном случае процесс анализа завершается.

В блоке 3 файл анализируется обученным классификатором на основе алгоритма экстремально рандомизированных деревьев. Данный этап подразумевает обучение классификатора, настройку его гиперпараметров и определение наиболее значимых признаков из модели ПСП, позволяющих наиболее точно классифицировать зашифрованные/сжатые данные от остальных классов. Определенные в блоке 3 признаки передаются в блок 4, где осуществляется вычисление значений признаков анализируемого файла.

Далее, на этапах 5 и 6 выполняется итерационное движение по узлам деревьев. Процесс останавливается в случае достижения терминального узла дерева, содержащего в себе соответствующую метку класса, присваиваемую анализируемому файлу. Если файл распознан как псевдослучайная последовательность, то работа алгоритма продолжается, иначе завершается.

В блоке 8 анализируемый файл передается на вход классификатора на основе случайного леса. На данном этапе происходит обучение классификатора, определение его гиперпараметров и вычисление наиболее значимых признаков, позволяющих классифицировать зашифрованные и сжатые последовательности между собой, т.е. выполняется бинарная классификация. Полученные признаки передаются в блок 9, где происходит вычисление их значений в анализируемом файле.

В блоках 10-12 выполняется процедура классификации анализируемого файла, в результате которого ему присваивается метка зашифро-

ванных или сжатых данных. Выбор алгоритма дополнительных деревьев выполнен на основе проведенных экспериментов. Результаты оценки точности мульти классовой классификации данных 4 классов: зашифрованных/сжатых (aes, camellia, des, rc4, ГОСТ 34.12 «Кузнечик», zip, rar, 7z, gz, xz, bz2); изображений (jpg); текста (txt); табличных данных MS office (xls) представлены в таблице 5.

Таблица 5 — Оценка точности классификации 4 классов данных различными алгоритмами машинного обучения

Алгоритм	*Acc*	*AUC*	*Rec*	*Prec*	*F*1	*Kappa*	*MCC*	*TT,sec*
RF	0,9998	1,0	0,9992	0,9998	0,9998	0,9994	0,9994	41,62
LGBM	0,9998	1,0	0,9992	0,9998	0,9998	0,9992	0,9992	76,88
ETC	0,9998	1,0	0,9989	0,9998	0,9998	0,9992	0,9992	6,84
DT	0,9996	0,9995	0,9991	0,9996	0,9996	0,9986	0,9986	14,37
LR	0,9994	0,9994	0,9972	0,9994	0,9994	0,9977	0,9977	53,38
LDA	0,9978	0,9983	0,9921	0,9978	0,9978	0,9922	0,9922	14,62
RC	0,9977	0,0	0,9916	0,9977	0,9976	0,9918	0,9918	2,634

Исходя из полученных результатов, можно сделать вывод о том, что все протестированные алгоритмы имеют высокую точность как по метрике Accuracy, так и по остальным метрикам. Поскольку на основе метрик достаточно затруднительно сделать выбор в пользу конкретного алгоритма, была рассмотрена временная характеристика работы данных алгоритмов. Наименьшем временем обучения обладает классификатор на основе дополнительных деревьев. Поскольку время обучения прямо пропорционально связано с временем классификации файлов, данный алгоритм был выбран в качестве наилучшего.

Блок схема разработанного способа классификации ПСП представлена на рисунке 6.

Для определения эффективности разработанного способа классификации псевдослучайных последовательностей были проведены эксперименты по классификации зашифрованных, сжатых и открытых данных 2 классов: офисных документов MS Office (doc, docx, xls, xlsx, ppt, pptx) и изображений формата jpeg. Анализируемая выборка состояла из 8000 файлов 4-х классов. Результаты определения значимости признаков при мультиклассовой классификации в блоке 4 представлены на рисунке 7.

Значимость признаков была определена согласно значений индекса Шепли, определяемого выражением 5

$$\omega_i(p) = \sum_{S \in N/i} \frac{|S|!(n-|S|-1)!}{n!}(p(S \bigcup i) - p(S)), \qquad (5)$$

где $p(S \bigcup i)$ – предсказание модели на основе i-го признака; $p(S)$ – предсказание модели без i- го признака; n – количество признаков; S – набор признаков без i – го признака. Зависимость точности классификации ПСП

Рис. 6 — Способ классификации ПСП

от количества признаков, используемых в модели ПСП представлена на рисунке 8.

Наиболее рациональным числом признаков относительно времени выполнения классификации ПСП и достигаемой точности является значение 100.

Результаты оценки точности классификации в зависимости от максимальной глубины деревьев в лесу представлены на рисунке 9.

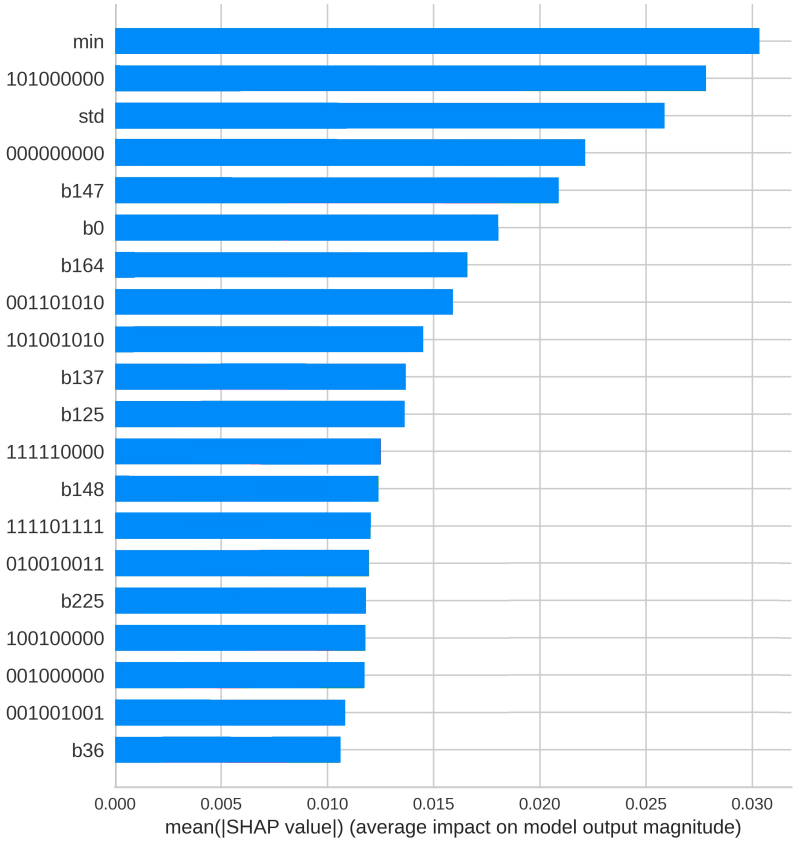

Рис. 7 — Значимость признаков согласно значения индекса Шепли

Максимальная точность достигается при глубине деревьев в случайном лесу равной 7 узлам.

Результаты оценки точности классификации в зависимости от количества деревьев в лесу, т.е. количества базовых классификаторов в ансамбле случайного леса представлены на рисунке 10.

Наибольшая точность классификации ПСП была достигнута при использовании 20 базовых классификаторов.

Для оценки адекватности разработанной модели ПСП и применимости алгоритма классификации были проведены эксперименты по определению точности классификации зашифрованных и сжатых ПСП, сформированных из различных источников: бинарных, текстовых, аудио файлов и изображений.

Результаты проведенных экспериментов при использовании метрики площадь под кривой ошибок ($AUC - ROC$) представлены на рисунке 11.

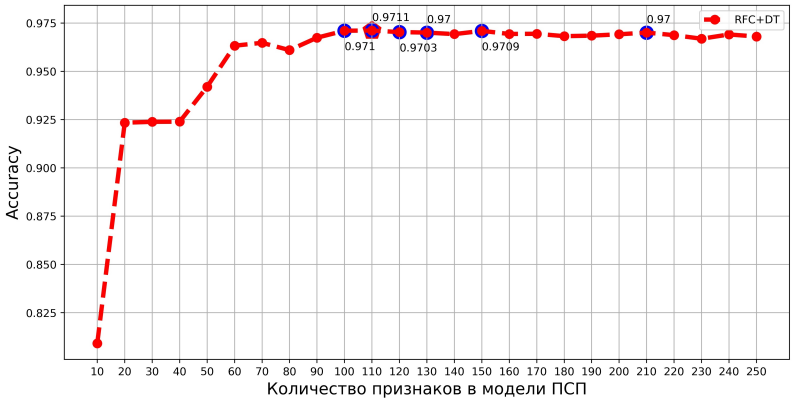

Рис. 8 — Зависимость точности классификации ПСП от числа признаков в модели ПСП

Рис. 9 — Зависимость точности классификации ПСП от максимальной глубины деревьев в лесу

Обоснован выбор размера сканирующего окна классификатора равный 600 кбайт.

При малом размере сканирующего окна, менее 100 килобайт, наблюдается точность классификации равная 0,88, далее при увеличении размера происходит плавное увеличение точности классификации, достигающая значения в 0,971 при размере окна 600 килобайт.

Данная особенность объясняется используемыми признаками, которые являются статистическими. При увеличении размера анализируемой ПСП возрастает и значимость значений частот подпоследовательностей длины 9 бит, что позволяет классификатору с более высокой точностью разделять зашифрованные и сжатые данные.

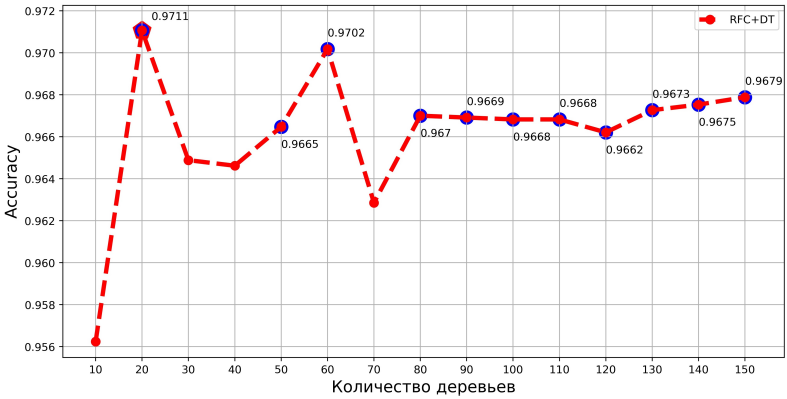

Рис. 10 — Зависимость точности классификации ПСП от числа деревьев в лесу

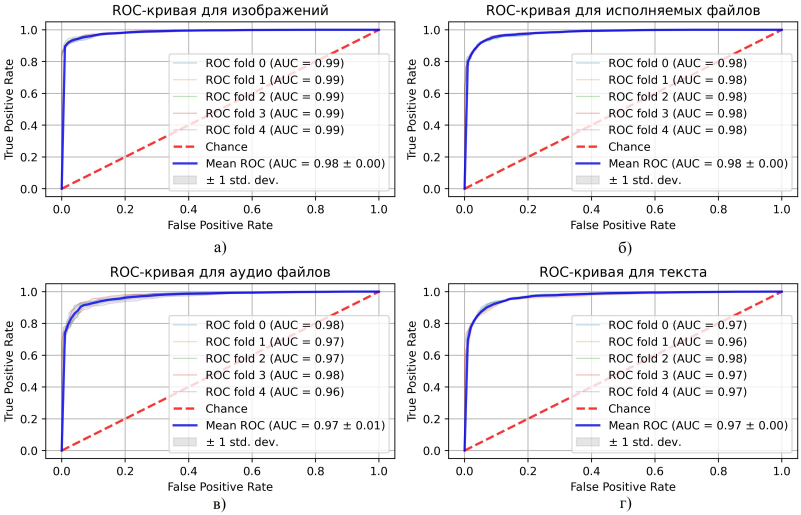

Рис. 11 — Значение метрики $AUC - ROC$ при классификации зашифрованных и сжатых ПСП, сформированных из файлов различных типов данных: а) изображения; б) исполняемые файлы; в) аудио файлы; г) текст

Разработанный способ классификации ПСП может быть внедрен в существующие средства защиты информации от утечек, функциональная схема средств защиты информации от утечек с подсистемой анализа зашифрованных данных представлена на рисунке 12.

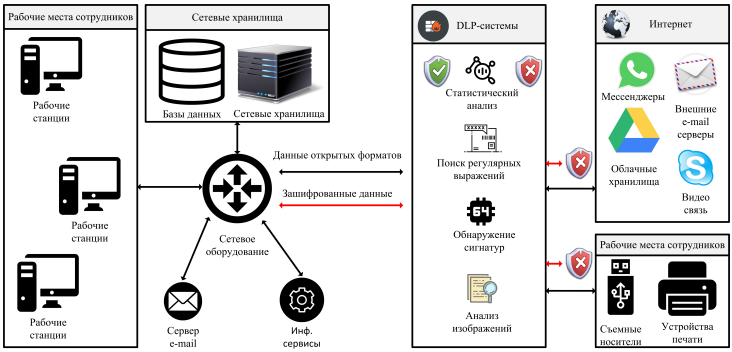

Рис. 12 — Функциональная схема средства защиты информации от утечек с подсистемой анализа зашифрованных данных

ЗАКЛЮЧЕНИЕ

Основные результаты работы:

1. На основе проведенного анализа особенностей функционирования современных средств обнаружения и предотвращения утечек информации, выявлены ограничения существующих подходов классификации зашифрованных и сжатых данных, связанные с использование заголовков файлов и низкой точностью классификации без их применения.
2. В ходе проведенных экспериментов определена рациональная длина подпоследовательностей равная 9 битам, участвующих в формировании признакового пространства.
3. Модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, позволила учесть особенности сжатых и зашифрованных ПСП при их представлении в бинарном виде, как в виде подпоследовательностей длиной 9 бит, так и виде распределения байт.
4. Разработан метод классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных, учитывающий дискриминирующую способность их статистических признаков и демонстрирующий более высокую точность классификации относительно существующих аналогов.
5. Произведена оценка эффективности разработанных решений. Удалось достичь точности классификации зашифрованных и сжатых ПСП в 0,971 при выбранных гиперпараметрах: количество признаков модели ПСП – 100; максимальная глубина деревьев – 7 узлов; количество деревьев в ансамбле – 20; размер ПСП не менее 60 кбайт. Полученное значение точности классификации ПСП превосходит существующие аналоги.

6. Реализован на практике способ классификации псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных в прототипе подсистемы анализа зашифрованных данных в корпоративных сетях и средствах обнаружения сетевых атак для защиты от Botnet атак.

Дальнейшая работа в предметной области исследования может быть направлена на поиск новых статистических признаков, позволяющих понизить размерность признаковых пространств и уменьшить размер сканирующего окна. Также перспективным направлением дальнейших исследований можно признать поиск более эффективных алгоритмов машинного обучения, позволяющих увеличить точность классификации зашифрованных и сжатых данных.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

1. *Спирин, А. А.* Алгоритм классификации псевдослучайных последовательностей / А. А. Спирин, А. В. Козачок // Вестник воронежского государственного университета. Серия: системный анализ и информационные технологии. — 2020. — 1. — с. 87—98.
2. *Спирин, А. А.* Алгоритм классификации псевдослучайных последовательностей на основе построения случайного леса / А. А. Спирин, А. В. Козачок, О. М. Голембиовская // Доклады ТУСУР. — 2020. — т. 23, № 3. — с. 55—60.
3. *Спирин, А. А.* Предложения по раннему обнаружению деструктивных воздействий Botnet на компьютерные сети связи / А. А. Спирин, М. М. Добрышин, А. Д. Лактионов // Телекоммуникации. — 2020. — 12. — с. 25—29.
4. *Козачок, А.* Модель псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных / А. Козачок, А. Спирин // Программирование. — 2021. — 4. — с. 31—44.

В изданиях, входящих в международную базу цитирования Scopus

5. Classification of pseudo-random sequences based on the random forest algorithm / A. A. Spirin [et al.] // Ivannikov Memorial Workshop Proceedings. — 2020. — P. 55—58.
6. Modeling of Pseudo-Random Sequences Generated by Data Encryption and Compression Algorithms / A. V. Kozachok [et al.] // CEUR Workshop proceedings. Vol. 3035. — Bauman Moscow State Technical University. 2021. — P. 98—106.

7. *Kozachok, A. V.* Model of Pseudo-Random Sequences Generated by En-cryption and Compression Algorithms / A. V. Kozachok, A. A. Spirin // Programming and Computer Software. — 2021. — Vol. 47, no. 4. — P. 249—260.

8. *Kozachok, A. V.* An Encrypted File Detection Algorithm / A. V. Kozachok, A. A. Spirin, V. I. Kozachok // Automatic Control and Computer Sciences. — 2021. — Vol. 55, no. 8. — P. 1121—1128.

В сборниках трудов конференций

9. *Спирин, А. А.* Исследование статистических свойств псевдослучайных последовательностей / А. А. Спирин, А. В. Козачок // Методы и технические средства обеспечения безопасности информации. XXVIII Научно-техническая конференция. т. 28. — Санкт-Петербургский политехнический университет Петра Великого. 2019. — с. 67—68. — URL: https://www.elibrary.ru/item.asp?id=38251162.

10. *Спирин, А. А.* О статистических свойствах алгоритмов сжатия и шифрования / А. А. Спирин, А. В. Козачок // Актуальные проблемы инфотелекоммуникаций в науке и образовании. VIII Международная научно-техническая и научно-методическая конференция. т. 2. — СПбГУТ. 2019. — с. 350—353. — URL: http://www.sut.ru/doci/nauka/1AEA/APINO/8-APINO%202019.%20%D0%A2.2.pdf.

11. *Спирин, А. А.* Подход к классификации последовательностей, сформированных алгоритмами сжатия и шифрования / А. А. Спирин, А. В. Козачок // XXII конференция «РусКрипто 2020». — 2020. — URL: https://www.ruscrypto.ru/resource/archive/rc2020/files/05_kozachok_spirin.pdf (дата обр. 12.09.2020).

12. *Спирин, А. А.* Подходы к классификации псевдослучайных последовательностей / А. А. Спирин, А. В. Козачок // Безопасные информационные технологии. X Международная научно-техническая конференция. — 2019. — с. 191—195. — URL: https://www.elibrary.ru/item.asp?id=42476608 (дата обр. 09.12.2020).

13. *Спирин, А. А.* Классификация последовательностей, сформированных алгоритмами сжатия и шифрования / А. А. Спирин, А. В. Козачок // Методы и технические средства обеспечения безопасности информации. XXIX научно-техническая конференция. — 2020. — с. 105—107. — URL: https://www.elibrary.ru/item.asp?id=44017301 (дата обр. 09.12.2020).

14. *Спирин, А. А.* Подход к классификации псевдослучайных последовательностей / А. А. Спирин // Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов им. Е.В. Арменского. — НИУ ВШЭ. 2020. — с. 19. — URL: https://miem.hse.ru/mirror/pubs/share/344671975.pdf (дата обр. 09.12.2020).

15. *Спирин, А. А.* Бинарная классификации псевдослучайных последовательностей / А. А. Спирин // Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов им. Е.В. Арменского. — НИУ ВШЭ. 2021. — с. 5. — URL: https://miem.hse.ru/data/2021/03/10/1397674215/%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0%20%20%D0%BA%D0%BE%D0%BD%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%86%D0%B8%D0%B8%202021.pdf (дата обр. 09.12.2020).

16. *Спирин, А. А.* Моделирование псевдослучайных последовательностей, сформированных алгоритмами шифрования и сжатия данных / А. А. Спирин, А. В. Козачок. — 2021. — URL: https://www.elibrary.ru/item.asp?id=46399220 (дата обр. 07.04.2021).

В прочих изданиях

17. *Spirin, A. A.* Classification of sequences generated by compression and encryption algorithms / A. A. Spirin, A. V. Kozachok // Journal of Science and Technology on Information Security. — 2019. — Vol. 10, no. 2. — P. 3—8.

18. *Spirin, A. A.* Pseudo random sequences classification algorithm / A. A. Spirin, A. V. Kozachok // Journal of Science and Technology on Information Security. — 2020. — Vol. 7, no. 1. — P. 1—13.