КЛАССИФИКАЦИЯ КОМПЬЮТЕРНЫХ АТАК С ИСПОЛЬЗОВАНИЕМ МУЛЬТИФРАКТАЛЬНОГО СПЕКТРА ФРАКТАЛЬНОЙ РАЗМЕРНОСТИ

Шелухин О. И.[1], Рыбаков С. Ю.[2], Раковский Д. И.[3]

DOI: 10.21681/2311-3456-2024-2-107-119

Цель исследования: *разработка метода повышения эффективности бинарной и многоклассовой классификации компьютерных атак (КА) путем использования дополнительных информативных признаков, в качестве которых предложно использовать мультифрактальный спектр фрактальной размерности (МСФР) обрабатываемых последовательностей.*

Методы исследования: *дискретный вейвлет анализ, мультифрактальный анализ, машинное обучение, программная реализация комбинированного метода многоклассовой классификации в совокупности с методами фрактального анализа.*

Объектами исследования *являются теоретические и практические вопросы разработки, реализации и визуализации алгоритмов обнаружения и классификации КА в целях информационной безопасности.*

Результаты исследования. *Разработаны метод и алгоритм композиции машинного обучения и методов мультифрактального анализа обрабатываемых процессов с целью повышения эффективности многоклассовой классификации КА. Обоснованы границы изменения входных параметров алгоритма, для корректной многоклассовой классификации компьютерных атак. Показана целесообразность использования при классификации КА характеристик МСФР, что позволяет повысить эффективность классификации атак методами машинного обучения за счет расширения количества атрибутов параметрами МСФР.*

Практическая значимость: *представленный метод является универсальным и может быть применен в различных системах обеспечения информационной безопасности.*

Ключевые слова: *мультифрактальный анализ, показатель Херста, машинное обучение, статистические характеристики, метрики, экспериментальные данные, атрибуты.*

CLASSIFICATION OF COMPUTER ATTACKS USING MULTIFRACTAL SPECTRUM OF FRACTAL DIMENSION

Sheluhin O. I.[4], Rybakov S. Y.[5], Rakovskiy D. I.[6]

The aim of the study: *development of a method to improve the efficiency of binary and multiclass classification of computer attacks (CA) by using additional informative features, as which it is proposed to use the multifractal spectrum of fractal dimension (MSFR) of processed sequences.*

Research methods: *discrete wavelet analysis, multifractal analysis, machine learning, software implementation of the combined method of multiclass classification in conjunction with methods of fractal analysis. The objects of the research are theoretical and practical issues of development, implementation and visualization of algorithms for detection and classification of CA for information security purposes.*

1 Шелухин Олег Иванович, доктор технических наук, профессор Московского технического университета связи и информатики, Москва, Россия. E-mail: sheluhin@mail.ru, ORCID: https://orcid.org/0000-0001-7564-6744

2 Рыбаков Сергей Юрьевич, аспирант кафедры «Информационная безопасность» Московского технического университета связи и информатики, Москва, Россия. E-mail: s.i.rybakov@mtuci.ru, ORCID: https://orcid.org/0000-0002-4593-9009

3 Раковский Дмитрий Игоревич, аспирант кафедры «Информационная безопасность» Московского технического университета связи и информатики, Москва, Россия. E-mail: Prophet_alpha@mail.ru, ORCID: https://orcid.org/0000-0001-7689-4678

4 Oleg I. Sheluhin, Dr. Sc., Full Professor, Moscow Technical University of Communications and Informatics, Moscow, Russia. E-mail: sheluhin@mail.ru; ORCID: https://orcid.org/0000-0001-7564-6744

5 Sergei Y. Rybakov, Postgraduate student, Moscow Technical University of Communication and Informatics, Moscow, Russia. E-mail: s.i.rybakov@mtuci.ru, ORCID: https://orcid.org/0000-0002-4593-9009

6 Dmitry I. Rakovskiy, Postgraduate student, Moscow Technical University of Communication and Informatics, Moscow, Russia. E-mail: Prophet_alpha@mail.ru. ORCID: https://orcid.org/0000-0001-7689-4678

Research results. *The methodology and algorithm of composition of machine learning and methods of multifractal analysis of processed processes to improve the efficiency of multiclass classification of CA are developed. The boundaries of changing the input parameters of the algorithm for correct multiclass classification of computer attacks are substantiated. The feasibility of using the characteristics of MSFR in the classification of CA is shown, which allows to increase the efficiency of classification of attacks by machine learning methods by expanding the number of attributes by the parameters of MSFR.*

Practical significance: *the presented method is universal and can be applied in various systems of information security.*

Keywords: *fractal dimension, Hurst exponent, machine learning, multifractal analysis, statistical characteristics, metrics, spectrum of fractal dimensions.*

Постановка задачи

Многочисленные исследования статистических характеристик сетевого трафика и сетевых компьютерных атак (КА) показывают наличие у них свойств фрактальности или самоподобия, а также изменчивость показателей, характеризующих фрактальные свойства [1–3]. Для оценки степени самоподобия используются понятия фрактальной размерности (ФР) множества (по Хаусдорфу) D и показатель Херста H, характеризующий степень самоподобия процесса, связанные между собой соотношением: $D = 2 - H$.

В подавляющем большинстве работ в области телекоммуникаций[7] [2–4] используется именно показатель Херста H, *отличающийся от* D *на фиксированную величину*. Поэтому в дальнейшем в качестве оценки ФР нормального трафика и КА будем использовать оценки показателя Херста.

Методы фрактального анализа широко используются для обнаружения атак и сетевых аномалий в том числе в режиме реального времени путем мониторинга текущей фрактальной размерности трафика компьютерных сетей [4].

Учитывая, что для оценки ФР трафика требуется как правило много времени и большие объемы данных обнаружение атак с помощью фрактального анализа осуществлялось как правило независимо от других методов, позволяющих определить аномалии во временном ряду в режиме реального времени. Все это послужило поводом для поиска новых методов обнаружения и прогнозирования КА, к числу которых можно отнести комбинацию машинного обучения и фрактальный анализ.

Появились работы, в которых вопросы обнаружения и классификации КА стали интегрироваться с методами машинного обучения [4–7].

В работе [5] на примере базы данных KDD Cup1999 [7,8] показано положительное влияние оценки самоподобных свойств сетевого трафика, характеризуемого средним значением показателя Херста на качество бинарной классификации.

В работах [5,6] на примере набора данных UNSW-NB15 приведены результаты исследования влияния широкого спектра статистических характеристик ФР на качество бинарной классификации. Показано, что параметры ФР могут рассматриваться как дополнительные информационные признаки (атрибуты) КА, учет которых в задачах классификации могут приводить к повышению достоверности обнаружения до 10 %.

В работах [8–10] анализируются вопросы обнаружения кибератак на основе интеграции фрактального анализа и статистических методов.

В работах [8,9] предлагается дополнительно использовать для обнаружения аномальных выбросов в системах передачи данных метод машинного обучения, основанный на применении гибридной искусственной нейронной сети, состоящей из автокодировщика (autoencoder) и классификатора. Проведена экспериментальная оценка предлагаемой методики, показывающая ее достаточно высокую эффективность.

Вместе с тем во всех указанных работах в качестве основного рассматривались традиционные асимптотические методы оценки ФР. Однако используя методы текущей оценки ФР в скользящем окне в реальном масштабе времени можно усовершенствовать рассмотренные выше алгоритмы[8].

Учитывая, что свойство самоподобия наблюдается в широких временных масштабах (например, при различном временном разрешении на уровне бит, пакетов, потоков и т.д.), наличие в сигнале продолжительных атак и аномальной активности изменяет самоподобную природу трафика, приводит

7 Шелухин О. И., Осин А. В., Смольский С. М. Самоподобие и фракталы. Телекоммуникационные приложения. Физматлит, 2008. 362 с. ISBN: 978-5-9221-0949-9
Park. K., Willinger W. Self-similar network traffic and performance evaluation. Self-Similar Network Traffic: An Overview. 2000. С. 1–38. DOI: https://doi.org/10.1002/047120644X.ch1
Sheluhin O. I., Smolskiy S. M., Osin A. V. Self-similar processes in telecommunications. Chichester: John Wiley & Sons, 2007, 334 с. DOI: 10.1002/9780470062098

8 Шелухин О. И., Панкрушин А. В. Обнаружение аномальных выбросов в реальном масштабе времени методами мультифрактального анализа // Нелинейный мир. 2016. Т. 14. № 2. С. 72–82.

к мультифрактальной структуре обрабатываемых процессов [4,10–11], а также см.[9]

Информация о различии ФР обрабатываемых процессов (если они доступны для обработки) при разном разрешении по времени может быть использована для модификации рассмотренных алгоритмов обнаружения/классификации КА и может привести к улучшению показателей классификации методами машинного обучения.

Целью работы является повышение эффективности обнаружения и классификации компьютерных атак на основе использования композиции параметров мультифрактального спектра фрактальной размерности (МСФР) обрабатываемых процессов и методов машинного обучения.

Структура экспериментальных данных

В качестве примера, на котором иллюстрируется влияние мультифрактальных характеристик анализируемого трафика на эффективность классификации КА, рассмотрена база Kitsune (2019) [18,19], в которой собран набор данных сетевого трафика от устройств Интернета вещей (IoT). Целью создания базы Kitsune являлось предоставление исследователям большого набор данных о реальных и маркированных вредоносных программах, и безопасном трафике Интернета вещей для разработки алгоритмов машинного обучения. Особенностью набора данных является отсутствие специализации устройств IoT, что позволяет проводить исследования трафика без уточнения дополнительной информации [16].

На (рис. 1) изображена топология компьютерной сети IoT для сбора данных, а также векторы, поясняющие происхождение атак. Захват сетевого трафика производился на маршрутизаторе в точках, указанных на рисунке цифрами. В каждом наборе данных первый миллион пакетов представлял собой чистый сетевой трафик, пакеты с номером миллион и выше содержали определенную компьютерную атаку.

В данном наборе содержится информация о четырех типах атак: разведка (Recon), человек посередине (MitM), отказ в обслуживании (DoS) и вредоносное ПО для ботнетов (Botnet Malware) Mirai. Mirai — это вредоносное ПО, которое заражает IoT-устройства (умные бытовые приборы с доступом в интернет), работающие на процессорах ARC, и превращает их в сеть дистанционно управляемых ботов, которых также называют «зомби». Этот ботнет часто используется для запуска DDoS-атак.

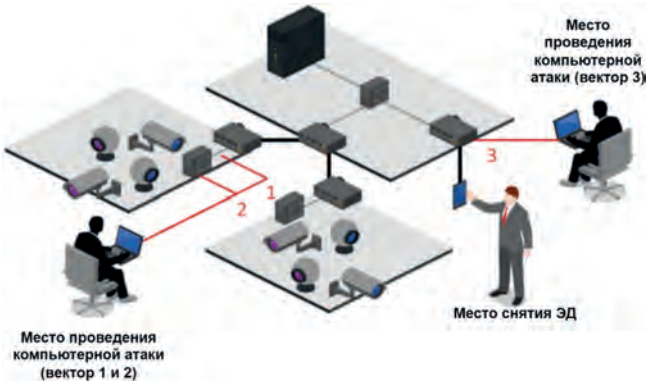

Рис. 1. Топология исследуемой сетевой инфраструктуры[10]

Данные об атаках были получены из коммерческой IP-системы наблюдения и сети, включающей в себя устройства интернета вещей (IoT). Каждый набор данных содержит миллионы сетевых пакетов и различные кибератаки. Для каждого типа атак имеется следующий набор данных:

- Предварительно обработанный набор данных, который готов для применения алгоритмов машинного обучения в формате .csv;
- Файл с метками, также в формате .csv.

Набор данных также содержит метки, описывающие взаимосвязь между потоками, связанными со вредоносными или возможными вредоносными действиями.

Для обнаружения вредоносных потоков на основе ручного анализа сети использовались следующие метки:

- Метка «**Атака**» указывает на то, что с зараженного устройства на другой хост произошла какая-то атака. Атакой будем называть любой поток, который, анализируя свою полезную нагрузку и поведение, пытается воспользоваться каким-либо уязвимым сервисом. Например, перебор какого-нибудь логина по телнету, внедрение команды в заголовок GET-запроса и т.п.
- Метка «**Доброкачественный**» указывает на то, что в соединениях не обнаружено никаких подозрительных или вредоносных действий.
- Метка **Mirai** указывает на то, что соединения имеют характеристики ботнета Mirai и добавляется, когда потоки имеют схожие шаблоны с наиболее распространенными известными атаками Mirai.
- Метка **C&C** указывает на то, что зараженное устройство было подключено к серверу CC. Эта активность была обнаружена при анализе захвата сетевых вредоносных программ, поскольку подключения к подозрительному серверу носят

9 Sheluhin O. I., Garmashev A. B., Aderemi A. A. Detection of teletraffic anomalies using multifractal analysis // International Journal of Advancements in Computing Technology. 2011. Т. 3. № 4. С. 174–182. DOI: 10.4156/ijact.vol3.issue4.19
Шелухин, О. И. Мультифракталы: инфокоммуникационные приложения. Москва: Научно-техническое издательство «Горячая линия-Телеком», 2011. 576 с. ISBN 978-5-9912-0142-1.

10 Mirsky, Y., Doitshman, T., Elovici, Y., Shabtai, A. Kitsune: an ensemble of autoencoders for online network intrusion detection // Arxiv, 2018. С. 1–15. DOI: 10.48550/arXiv.1802.09089

периодический характер, либо наше зараженное устройство загружает с него какие-то двоичные файлы, либо от него приходят и уходят какие-то IRC-подобные или декодированные заказы.

- Метка **DDoS** указывает на то, что зараженное устройство выполняет распределенную атаку типа «отказ в обслуживании». Эти потоки трафика обнаруживаются как часть DDoS-атаки из-за количества потоков, направленных на один и тот же IP-адрес.

При сборе информации с компьютерной сети необработанные, «сырые» данные, поступающие с перечня устройств, захватывались в виде пакетов. Каждый пакет ассоциировался с временной меткой и рядом категориальных атрибутов таких как: MAC-адрес, IP-адрес, порты назначения и отправки и т.д. Преобразование пакетов в многомерные метрические векторы осуществлялось с использованием метода демпфированной инкрементной статистики (ДИС; от англ. – Damped Incremental Statistics, DIS).

ДИС ассоциировалось с параметром $\lambda > 0$, а также с кортежем $IS := (N, L_{S_i}, SS_i)$, где N количество, $L_{S_i} = \Sigma_{i=1}^{N} x_i$ линейная сумма, а $SS_i = \Sigma_{i=1}^{N} x_i^2$ – квадрат суммы наблюдаемых на текущий момент экземпляров занесенных в инкрементную статистику. Каждая инкрементная статистика связана с потоками данных, определяемыми связкой MAC-адреса, IP-адресами, портами стека протоколов TCP/IP и четырьмя типами данных:

- IP отправителя (srcIP);
- MAC-адрес отправителя (srcMAC), включая пару (srcMAC, srcIP), ассоциированную с отправителем;
- Информация, ассоциированная с каналом передачи данных – пара IP-адресов отправителя – получателя (srcIP, dstIP);
- Сокет, ассоциированный с каналом передачи данных – в виде четверки «IP-адрес отправителя, порт отправителя, IP-адрес получателя, порт получателя (srcIP, srcPort, dstIP, dstPort).

Каждый новый пакет, поступающий на вход ДИС, обновлял статистику по правилам: $\gamma = 2^{-\lambda(t - t_{last})}$; $\Delta IS_\lambda = (\gamma\omega + 1, \gamma LS + x, \gamma SS + x^2)$, где t_{last} – отметка времени поступившего пакета, ассоциированного с потоком статистики; ΔIS_λ – приращение инкрементной статистики. Параметр λ определяет интенсивность затухания статистики во времени.

Признаки представляют собой инкрементальные (пошаговые) статистики поступающих данных. Так если $S = \{x1, x2, \ldots\}$ представляет собой неограниченный поток данных, где $x_i \in R$, последовательность наблюдаемых размеров пакетов, то процедура обновления кортежа для вставки x_i в IS имеет вид $IS \leftarrow (N + 1, L_{S_i} + x_i, SS_i + x_i^2)$, а текущие статистики в любой момент времени имеют вид

$$\mu_{S_i} = \frac{1}{N}\Sigma_{i=1}^{N} x_i;\ \sigma_{S_i}^2 = \frac{1}{N}[\Sigma_{i=1}^{N} x_i^2 - \mu_S^2] \text{ и } \sigma_S = \sqrt{\sigma_S^2}.$$

Помимо перечисленных, при формировании атрибутов КА и нормального трафика список статистик вычисляемых из инкрементальной статистики IS_i,λ, включал также коэффициенты ковариации $cov(x_i,x_j)$ и корреляции R_{ij}, дополнительные двумерные статистики

$$M_{ij} = \sqrt{\mu_{S_i}^2 + \mu_{S_j}^2} \text{ и } Q_{ij} = \sqrt{(\sigma_{S_i}^2)^2 + (\sigma_{S_j}^2)^2}\ .$$

После предобработки, используя перечисленные статистики для пяти значений коэффициента «старения» данных λ: 5,3,1,0.1,0.01 сформировано 115 атрибутов [17].

Поскольку наборы данных, сформированные для каждой из компьютерных атак, различны между собой по количеству пакетов, то каждой атаке ставится в соответствие две .csv таблицы: таблица, ассоциированная с обезличенными экспериментальными данными (ЭД) размерностью 115 атрибутов и таблица, ассоциированная с целевым столбцом – бинарной классовой меткой о проведении (отсутствии) компьютерной атаки.

Размерность ЭД для каждого типа атаки существенно различаются. Наименьший объем данных ассоциировался с набором данных типа «*Mirai*» (компьютерная атака, направленная на заражения сети интернета вещей вредоносным программным обеспечением) – всего 750 тысяч записей. Наибольший объем ЭД был зафиксирован у атаки типа отказ в обслуживании, «SSL Renegotiation» – более 6 миллионов пактов.

Для апробации разработанного алгоритма из всего множества наборов данных, наряду с КА типа «*Mirai*» был выбран набор, соответствующий атаке «OS Scan» содержащий ~1,6 млн записей ЭД.

Необработанные данные о сетевом трафике собирались (в формате pcap) с помощью зеркалирования портов на коммутаторе, через который обычно проходит трафик. Всякий раз, при поступлении пакета формировался *поведенческий снимок* хостов и протоколов, передавших этот пакет, который представляет собой набор из 23 признаков в пяти временных окнах L = 5: 100 мс; 500 мс; 1,5 с; 10 с и 1 мин с учетом коэффициента «старения» λ = 5; 3; 1; 0.1; 0.01.

В результате набор атрибутов (признаков), характеризующий перечисленные выше КА, формировался путем извлечения 115 статистических данных о трафике за указанные пять временных интервалов. Набор этих атрибутов представлен в (табл. 1).

Мультифрактальный спектр фрактальной размерности (МСФР)

Мультифракталы – это неоднородные фрактальные объекты, для полного описания которых, в отличие от обычных фракталов, недостаточно введения всего лишь одной величины – его фрактальной

Таблица 1.

Перечень атрибутов в наборе данных Kitsune в разных временных окнах

	№ атрибутов в разных	Атрибут
1	1, 24, 47, 70, 93	Длина комбинации MAC-IP в битах, (μ)
2	2, 25, 48, 71, 94	Длина SrcIP в битах, (μ)
3	3, 26, 49, 72, 95	Длина Channel в битах, (μ)
4	4, 27, 50, 73, 96	Длина Socket в битах, (μ)
5	5, 28, 51, 74, 97	Длина комбинации MAC-IP в битах, (σ)
6	6, 29, 52, 75, 98	Длина SrcIP в битах, (σ)
7	7, 30, 53, 76, 99	Длина Channel в битах, (σ)
8	8, 31, 54, 77, 100	Длина Socket в битах, (σ)
9	9, 32, 55, 78, 101	Длина Channel в битах, (M_{ij})
10	10, 33, 56, 79, 102	Длина Socket в битах, (M_{ij})
11	11, 34, 57, 80, 103	Длина Channel в битах,(Q_{ij})
12	12, 35, 58, 81, 104	Длина Socket в битах,(Q_{ij})
13	13, 36, 59, 82, 105	Длина Channel в битах, ($Cov_{i,j}$)
14	14, 37, 60, 83, 106	Длина Socket в битах, ($Cov_{i,j}$)
15	15, 38, 61, 84, 107	Длина Channel в битах,(R_{ij})
16	16, 39, 62, 85, 108	Длина Socket в битах, (R_{ij})
17	17, 40, 63, 86, 109	Количество пакетов MAC-IP,(N)
18	18, 41, 64, 87, 110	Количество пакетов SrcIP в битах,(N)
19	19, 42, 65, 88, 111	Количество пакетов Channel в битах, (N)
20	20, 43, 66, 89, 112	Количество пакетов Socket в битах, (N)
21	21, 44, 67, 90, 113	Межпакетные задержки исходящего трафика, (N)
22	22, 45, 68, 91, 114	Межпакетные задержки исходящего трафика, Channel, (μ)
23	23, 46, 69, 92, 115	Межпакетные задержки исходящего трафика, Channel, (σ)
24	116,117,118,119,120	МСФР данных в окне разрешения $\{\hat{H}_{t_{\Delta i}}, i = \overline{1,5}\}$

размерности (D или H). Для описания мультифракталов необходим целый спектр таких размерностей, число которых не ограничено.

Такая ситуация заставляет заняться поиском новых количественных характеристик подобных процессов.

Определение 1.[11] *Стохастический процесс $X(t)$ называется мультифрактальным, если он обладает стационарными приращениями и удовлетворяет неравенству $M[|X(t)|^q] = C(q)\ t^{\tau(q)+1}$ для некоторого положительного $q \in Q$,$[0,1] \subset Q$, где $\tau(q)$ — масштабная или скейлинговая функция (функция разбиения) и моментный коэффициент $C(q)$ не зависит от t.*

Для характеристики мультифрактального множества часто используют функцию мультифрактального спектра $f(\alpha)$, характеризующую спектр сингулярностей мультифрактала для переменной α, которая является показателем Липшица-Гельдера и имеет смысл меры сингулярности.

В результате мультифрактальный спектр $f_L(a)$ находится преобразованием Лежандра от функции разбиения $\tau(q)$:

$$f_L(a) = \inf_{q \in R} (aq - \tau(q)).$$

Таким образом, мультифрактальный спектр $f_L(a)$ представляет собой меру «частоты» показателя сингулярности $\alpha(t)$ к моменту времени t и показывает вероятность определенного значения показателя сингулярности $\tau(q) \leq \inf_a (qa - f_L(a))$.

Изменение спектра сингулярности может свидетельствовать об изменении характера исследуемых процессов, которое может быть невидимо как для традиционных методов, так и для фрактального анализа, основанного на расчете только показателя Херста.

11 Riedi, R.H., Crouse, M.S., Ribeiro, V.J., Baraniuk, R. A. Multifractal Wavelet Model with Application to Network Traffic // EEE Transactions on Information Theory. 1999. Т. 45. № 3. С. 992-1018
Taqqu M. S., Teverovsky V., Willinger W. Is network traffic self-similar or multifractal? // Fractals. 1997. Т. 5. №1. с. 63-73.

Сам по себе расчет спектра сингулярности дает исследователю сравнительно мало информации об исследуемом временном ряде. Гораздо больше информации можно получить при изучении динамики его спектра сингулярности с помощью скользящего окна при различном разрешении по времени.

Поэтому далее исследуемые процессы будут рассматриваться именно с точки зрения динамики мультифрактальных характеристик при нормальном функционировании КС и при возникновении аномалий или сетевых атак.

Чтобы обнаружить связь между поведением исследуемого процесса и его мультифрактальными характеристиками, введем иное понятие мультифрактального спектра фрактальной размерности (МСФР) исследуемого процесса.

Определение 2. *Под мультифрактальным спектром фрактальной размерности (МСФР) будем понимать последовательность текущих оценок ФР $\hat{H}_{t_{Д_i}}$ в окне анализа Δ фиксированной длины в зависимости от интервала разрешения (времени дискретизации $t_Д$):*

$$\{\hat{H}_{t_{Д_i}} = f(t_{Д_i}); \mathrm{i} = \overline{1,L}; t_{Д_i} \in \Delta; \Delta = const\}. \quad (1)$$

Анализируемый случайный процесс можно считать мультифрактальным поскольку при разных временных шкалах (при разном временном разрешении) величина ФР изменяется. В общем случае оценка ФР является случайной величиной $\hat{H} \in N(m_{\hat{H}}, \sigma^2_{\hat{H}})$ и полно характеризуется моментами распределения – средним значением $m_{\hat{H}}$ и дисперсией $\sigma^2_{\hat{H}}$ оценки.

Для оценки МСФР рассматриваемых КА в виде (1) были выбраны следующие параметры : окно оценки ФР Δ = 2000 отсчетов ; количество окон L = 5, так что $i = \overline{1,5}$; время дискретизации наблюдаемых процессов в анализируемых пяти окнах: $t_{Д_1}$ = 100 мс; $t_{Д_2}$ = 500 мс; $t_{Д_3}$ = 1 сек; $t_{Д_4}$ = 2 сек; $t_{Д_5}$ = 10 сек соответственно.

Для оценки текущих значений ФР в режиме реального времени предлагается использовать оценки фрактальной размерности (показателя $\hat{H}$) в скользящем окне методами дискретного вейвлет-анализа[12].

Рассмотрим процесс формирования оценки ФР на примере трафика IoT при воздействии атаки Mirai в скользящем окне размером Δ = 2000 отсчетов представленный на (рис. 2).

Пусть $\{X(ti),(i = \overline{1,I})\}$ будет дискретным случайным процессом, определенным на интервале $i = 1...I$ и пусть разложение трафика по вейвлет коэффициентам осуществляется в скользящем окне размера Δ. Смещение окна анализа осуществляется с шагом $K \le P$. В результате при смещении окна анализа слева направо положение окна пробежит m положений $M = \frac{P}{K}$, $m = \overline{1,M}$. Тогда вейвлет-коэффициенты детализации при m-ом положении окна $d^m_{j,k}$ могут быть найдены в конце анализируемого интервала.

На практике при использовании оценки показателя Херста в скользящем окне Δ, оценка формируется с высокой дисперсией и резкими скачками показателя Херста, как это можно заметить на (рис. 2б). Для нейтрализации резких выбросов и уменьшения дисперсии в [10] предлагается воспользоваться процедурой трешолдинга (thresholding) – фильтрацией оценки.

Под трешолдингом (thresholding) понимают метод пороговой очистки сигналов от шумов, основанный на вейвлет преобразовании.

В результате использования трешолдинга формула для текущей оценки $\hat{H}$ с использованием дискретного вейвлет преобразования (ДВП) приобретает следующий вид [9,10]:

$$\hat{H}(t_m) = \sum_{l=1}^{L_0} a^{(H)}_l \varphi^{(H)}_l(t_m) + \sum_{j=1}^{J} \sum_{l=1}^{L_j} T(d^{(H)}_{j,l}) \psi^{(H)}_{j,l}(t_m), \quad (2)$$

где $a^{(H)}_{j_0,l}$, $d^{(H)}_{j,l}$ – аппроксимирующие и детализирующие коэффициенты оценки показателя Херста при m-м положении окна фильтрации; $T(d^{(H)}_{j,l})$ – отфильтрованные с помощью преобразования трешолдинга детализирующие вейвлет-коэффициенты; $a^{(H)}_{j_0,l} = \langle \hat{H}(t_m), \varphi(d^{(H)}_{j,l}) \rangle$ – масштабный коэффициент аппроксимации, равный скалярному произведению оценки показателя Херста $\hat{H}(t_m)$ и масштабной функции «самого грубого» масштаба J, смещенной на l единиц масштаба вправо от начала координат; $d^{(H)}_{j,l} = \langle \hat{H}(t_m), \psi^{(H)}_{j,l} \rangle$ – вейвлет-коэффициент детализации масштаба j, равный скалярному произведению оценки показателя Херста $\hat{H}(t_m)$ и вейвлета масштаба j, смещенного на l единиц масштаба вправо от начала координат. Здесь $L_0 = 2^{Jmax}$, $L_0 \le L$, а $J_{max} = [\log_2 L]$ – максимальное число масштабов разложения; $[\log_2 L]$ – целая часть числа.

При гауссовских и квазидекоррелированных вейвлет коэффициентах, дисперсия оценки $\hat{H}$ может оценена соотношением [7]:

$$\sigma_{\hat{H}}^2 = var\hat{H}(j_1,j_2) = \frac{2}{n_{j_1} \ln^2 2} \frac{1 - 2^J}{1 - 2^{-(J+1)}} (J^2+4) + 2^{-2J}, \quad (3)$$

где $J = j_2 - j_1$ число октав, вовлеченных в линейное сглаживание и $n_{j_1} = 2^{-j_1} N_0$ число доступных коэффициентов в рамке j_1.

В гауссовском и асимптотическом приближении можно получить доверительный интервал $\hat{H} - \sigma_{\hat{H}} z_\beta \le H \le \hat{H} + \sigma_{\hat{H}} z_\beta$, где z_β представляет $1 - \beta$ квантиль стандартного Гауссовского распределения, то есть $P(z \ge z_\beta) = \beta$. Все результаты, представленные ниже, и при числовом моделировании, и на фактическом анализе данных, были подсчитаны при β = 0.025 (т.е. 95 % доверительный интервал).

12 Sheluhin O. I., Lukin I. Y. Network traffic anomalies detection using a fixing method of multifractal dimension jumps in a real-time mode // Automatic Control and Computer Sciences. 2018. Т. 52. № 5. С. 421–430. DOI 10.3103/S0146411618050115

Воспользовавшись (2) для обработки экспериментальных данных трафика IoT, были получены статистические характеристики показателя Херста нормального трафика и атаки типа Mirai.

Алгоритм формирования фильтрованной оценки, имеет вид:

1) Фильтрация производится в окне размером $L = 500$.
2) Производится 6-уровневое ДВП накопленной оценки показателей Херста $\hat{H}$.
3) Происходит удаление всех детализирующих вейвлет коэффициентов.
4) Применяется обратное ДВП.

На выходе после фильтрации получается отфильтрованная оценка без аномальных выбросов.

Предложенная модификация алгоритма оценки ФР основана на использовании дополнительной фильтрации показателя Херста $\hat{H}$ внутри скользящего окна. Для получения достоверной оценки значения показателя Херста необходимо использовать вейвлеты типа Хаар поскольку при их использовании наблюдается самая низкая дисперсия оценки ФР [6].

По итогам исследования были получены значения МСФР для нормального трафика в разных точках описанной топологии сети IoT и разных типов КА. В (табл. 2) приведены статистические характеристики оценок показателя Херста $\hat{H}$ в скользящем окне с применением процедуры трешолдинга для пяти окон оценивания размером 100 мс, 500 мс, 1,5 с, 10 с и 1 мин соответственно с учетом коэффициентов «старения» λ (λ = 5, 3, 1, 0.1, 0.01).

Количественный анализ полученных результатов показывает, что в отсутствии КА трафик IoT характеризуется оценками среднего значения $m_{\hat{H}}$ в интервале {0...0,5}, для интервалов дискретизации $t_{Д_1}$ = 100 мс; $t_{Д_4}$ = 2 сек и $t_{Д_5}$ = 10 сек, что означает, то анализируемый случайный процесс не обладает самоподобием. При $t_{Д_2}$ = 500 мс и $t_{Д_3}$ = 1 сек значение $m_{\hat{H}}$ лежит в диапазоне {0,5...1,0}, что свидетельствует о наличии фрактальных свойств у нормального трафика при этом временном разрешении.

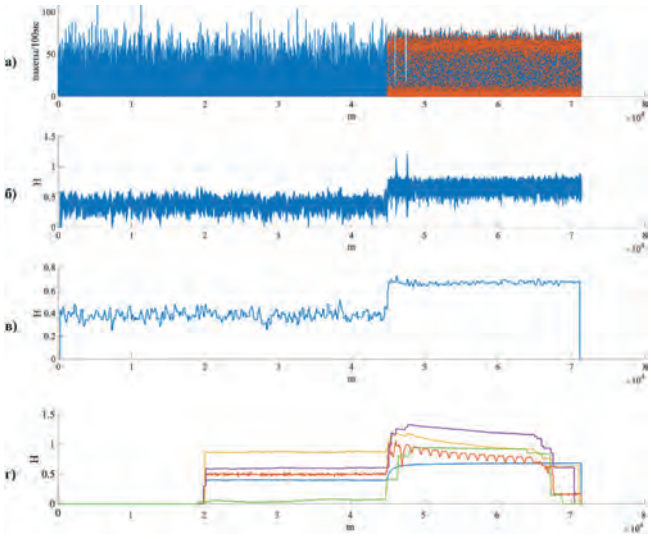

Рис. 2. Оценка ФР трафика IoT при воздействии атаки Mirai в скользящем окне размером а) фрагмент трафика с атакой Mirai, б) текущая оценка $\hat{H}$ в скользящем окне без фильтрации; в) оценка $\hat{H}$ в скользящем окне с фильтрацией; г) оценка параметров МСФР в скользящем окне при $i = \overline{1,5}$.

Для КА типа Mirai фрактальными свойствами атака обладает при $t_{Д_1}$ = **100** мс; $t_{Д_2}$ = **500** мс и $t_{Д_5}$ = **10** сек. В этом случае значение $m_{\hat{H}}$ лежит в диапазоне {0,5..1,0}, что свидетельствует о наличии фрактальных свойств у КА при этом временном разрешении.

При $t_{Д_3}$ = 1 сек; $t_{Д_4}$ = 2 сек параметр $m_{\hat{H}} > 1$, что указывает на наличие аномалий или на нестационарность обрабатываемого процесса.

Указанные значения ФР могут быть использованы в качестве дополнительных атрибутов алгоритма обнаружения атак в сетях IoT для атаки Mirai. На (рис. 3) показаны оценки $m_{\hat{H}}$ в скользящем окне Δ при различном временном разрешении.

Величина $\hat{H}$ обычно характеризует степень самоподобия процесса следующим образом. Случай $0{,}5<H<1{,}0$ характеризует трендоустойчивый процесс, обладающий длительной памятью и является самоподобным. Случай $0<H<0{,}5$ характерен для случайного процесса, не обладающего самоподобием. Случай $H>1{,}0$ соответствует аномалии (нестационарности) анализируемого процесса.

Таблица 2

Статистические характеристики оценки трафика IoT с трешолдингом

$t_{Д_i}$	$m_{\hat{H}}$ **нормального трафика**	$\sigma^2_{\hat{H}}$ **нормального трафика**	$\sigma_{\hat{H}}$ **нормального трафика**	$m_{\hat{H}}$ **атаки**	$\sigma^2_{\hat{H}}$ **атаки**	$\sigma_{\hat{H}}$ **атаки**
100 мс	0.3983	0.00003	0.0019	0.6745	0.000075	0.0087
500 мс	0.5285	0.00096	0.0098	0.8089	0.0065	0.0804
1 сек	0.6987	0.000069	0.0084	1.073	0.0054	0.0732
2 сек	0.4080	0.0027	0.0522	1.1703	0.0027	0.052
10 сек	0.0646	0.0002	0.0143	0.9303	0.0004	0.007

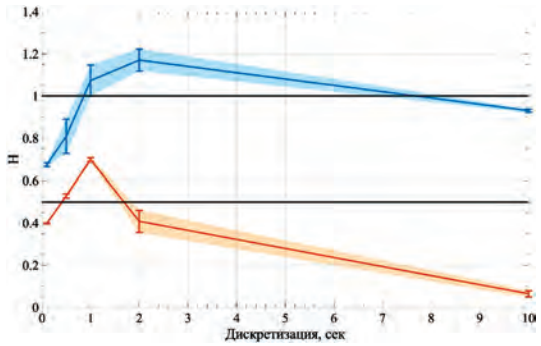

Рис. 3. Оценка показателя Херста на атаке Mirai и нормального трафика отфильтрованная оценка без аномальных выбросов при разном временном разрешении.

Используя численные значения среднего $m_{\hat{H}}$ и СКО фрактальной размерности $\sigma_{\hat{H}}$ нормального трафика и КА представленные в (табл. 2) предлагается добавить к уже имеющимся атрибутам значения МСФР нормального трафика и КА типа атаки Mirai, пяти элементов характеризующих $\{\hat{H}_{t_{Дi}}, i = \overline{1,5}\}$ как это показано в строке №24 (табл. 1). В результате количество атрибутов для нормального трафика и КА типа Mirai, увеличивается до 120.

Заметим, что для проведения сравнительного анализа в других типах атак (например, атаки *OS Scan*) такого расширения количества атрибутов не производилось, поскольку малая длительность атак такого типа не позволила получить информацию о МСФР.

Алгоритмы и метрики классификации

Оценим эффективность добавления МСФР к исходным данным на примере набора данных Kitsune содержащем 115 атрибутов метрического типа. Для оценки эффективности проведем два эксперимента:

1) бинарная классификация КА типа «Mirai Botnet» для двух случаев: без модификаций и с добавлением МСФР;
2) многоклассовая классификация КА типов «Mirai Botnet» и «OS Scan» для двух случаев: без модификаций и с добавлением МСФР.

В эксперимент многозначная [20] классификация не включалась, поскольку структура данных Kitsune исключает возникновение многозначных классовых меток [21].

Для решения задачи классификации выбран алгоритм типа «Случайный лес» (RF, Random Forest) в стандартной реализации библиотекой scikit-learn[13]. Выбор алгоритма обусловлен широтой применения указанного алгоритма для решения задач однозначной классификации [22–25].

Эксперименты проводились для двух наборов гиперпараметров RF: «глубина решающего дерева» = {2, 5}. Под «глубиной» дерева решений понимается гиперпараметр, который определяет количество уровней или узлов от корня до любого листа и определяется количеством уровней, не включая корневой узел. Параметры экспериментов сведены в (табл. 3).

Объем набора данных для КА «Mirai Botnet» составляет 764 136 шт. записей, из которых 121 620 шт. (16%) относятся к КА. Объем набора данных для КА «OS Scan» составляет 1 697 850 шт. записей, из которых 65 700 шт. (4%) относятся к КА.

Для проведения эксперимента №2 наборы данных для КА «Mirai Botnet» и «OS Scan» объединялись посредством операции конкатенации. При проведении эксперимента с МСФР, для всех записей, не относящихся к исходному набору КА «Mirai Botnet», атрибуты 116 … 120, связанные с МСФР, считались равными 0.

Эффективность классификации оценивалась по следующим метрикам: точность (*precision*), полнота (*recall*), F-мера (*F-score*), ROC-кривые (*Receiver Operating Characteristic curve – кривая ошибок*), AUC-ROC (*Area Under Curve – площадь под кривой ошибок*):

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \quad (4)$$

13 sklearn.ensemble.RandomForestClassifier [Электронный ресурс] // scikit-learn. URL: https://scikitlearn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата обращения: 05.02.2024).

Таблица 3

Параметры проводимых экспериментов для оценки эффективности добавления МСФР

№ эксперимента	Тип задачи	«Глубина решающего дерева»	Добавление МСФР для КА «Mirai Botnet»
1	Бинарная классификация: Mirai Botnet и Normal	2	нет
		5	нет
		2	да
		5	да
2	Многоклассовая классификация: Mirai Botnet, OS scan и Normal	2	нет
		5	нет
		2	да
		5	да

где TP (англ. True Positive) – истинно положительный исход классификации; TN (англ. True Negative) – истинно отрицательный исход классификации, FP (англ. False Positive) – ложно положительный исход классификации, FN (англ. False Negative) – ложно отрицательный исход классификации.

$$precision = \frac{TP}{TP + FP}, \quad (5)$$

$$recall = \frac{TP}{TP + FN}, \quad (6)$$

$$F_{score} = \frac{2TP}{(2TP + FN + FP)}, \quad (7)$$

$$AUC = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dt = \langle TPR \rangle. \quad (8)$$

Дополнительно оценим влияние добавления МСФР по информативности атрибутов в каждом из двух экспериментов оценивалось индексом Джини [23]. Индекс Джини, обычно используемый в деревьях решений и других алгоритмах машинного обучения, является показателем того, как часто случайно выбранный элемент будет неправильно идентифицирован. Индекс Джини – рассчитывается путем суммирования квадратов вероятностей каждого результата в распределении и вычитания результата из 1:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2, \quad (9)$$

где T – множество объектов обучающей выборки; n – количество классов; p_i – вероятность встречаемости класса i в множестве T.

Результаты классификации

Бинарная классификация КА «Mirai Botnet»

Классификация проводилась для двух глубин решающего дерева depth в ансамбле Random Forest – depth1 = 2 и depth2 = 5. Результаты бинарной классификации приведены в (табл. 4), которая разделена на две части:

1) результаты классификации для экспериментальных данных без добавления в число атрибутов МСФР при двух глубинах решающего дерева;
2) результаты классификации для экспериментальных данных с добавлением в число атрибутов МСФР при тех же двух глубинах решающего дерева.

Как видно из (табл. 4), добавление МСФР КА «Mirai Botnet» в атрибутное пространство позволяет однозначно классифицировать указанную КА (без ложноположительных и ложноотрицательных результатов классификации).

Возможность однозначной классификации обусловлена свойствами атрибутов № 116 ... 120 в (табл. 2), характеризующих спектр МСФР в пяти анализируемых окнах.

Однако поскольку ансамблевый классификатор RF при построении деревьев формирует решающие правила, исходя из информативности атрибутов, более актуально исследование результатов классификации для многоклассового случая.

Многоклассовая классификация КА «Mirai Botnet» и «OS Scan»

В случае многоклассовой классификации КА «*Mirai Botnet*», «*OS Scan*» и нормального трафика «*Normal*» каждая запись маркируется одним классом из заранее определенного множества $Label = (label_k, \xi = \overline{1,\Xi})$, $\Xi = 3$. Известно несколько методов оценки эффективности многоклассовой классификации, большинство из которых выполняют преобразование стандартное представление результата многоклассовой классификации в виде одного отдельного столбца $L = (l_1, l_2, ..., l_N)$, где N – количество записей в данных[14].

Известно несколько методов преобразования данных в бинарное представление. Как правило это осуществляется преобразованием в набор из Ξ столбцов, где каждый столбец показывает, маркирована ли n-я запись в наборе данных ξ-й классовой меткой – или нет.

14 Gibaja E., Ventura S. A Tutorial on Multilabel Learning // ACM Computing Surveys. 2015. Т. 47, № 3, С. 1–38. DOI: 10.1145/2716262.

Таблица 4

Оценки эффективности бинарной классификации алгоритмом RF для атаки «Mirai Botnet»

Метрика	Экспериментальные данные			
	без добавления МСФР		с добавлением МСФР	
	depth1	depth2	depth1	depth2
$accuracy$	0.921	0.991	1	1
$precision$	0.997	0.999	1	1
$recall$	0.908	0.990	1	1
f_{score}	0.951	0.995	1	1
$ROC\ AUC_{OVR}$	0.949	0.994	1	1

В методе One-vs-Rest (один против всех, OVR; также используется обозначение One-vs-All, OVA) преобразование классовых меток производится по правилу:

$$l_{n\xi} = \begin{cases} 1, l_n = label_\xi \\ 0, l_n = label_\xi \end{cases}; \xi = \overline{1,\Xi}, n = \overline{1,N}, \quad (10)$$

где $label_\xi$ – класс, сопоставляемый всем остальным, l_n – метка, присвоенная многоклассовым классификатором для n-й записи данных; $l_{n\xi}$ – результат преобразования для ξ-го класса.

После преобразования методом OVR (10), каждый столбец $L_\xi = (l_{1\xi}, l_{2\xi}, ..., l_{N\xi})$ может быть оценен как результат бинарной классификации по формулам (4) ... (8).

Для оценки влияния добавления МСФР в число атрибутов одной из атак, был проведен эксперимент с КА «*Mirai botnet*», аналогичный случаю бинарной классификации, с добавлением атаки типа «OS scan».

Для оценки выигрыша в классификации по каждой классовой метке выбран метод OVR. Количество возможных классовых меток $Label_{experiment} = (normal, miraibotnet, OSscan)$ равно трем. Соответственно количество столбцов с бинарными классовыми метками также равнялось трем Ξ = 3.

Из (рис. 4) видно, что распределение атрибутного пространства «сдвинуто» в сторону более широкого временного окна: 500 мс (атрибуты с 24 по 46) и 1,5 с (атрибуты с 47 по 69).

Основная концентрация атрибутов приходится на интервал 500 мс. – 1,5 с (67%), остальные распределены в диапазоне 10 с и 1 мин. «Сдвиг» информационной значимости атрибутов в область более широких временных окон обусловлен наличием атаки второго типа – «OS Scan», а также большого количества данных о нормальном функционировании КС. С учетом объединения наборов данных (для КА «Mirai Botnet», «OS Scan») доля «нормальных» записей в итоговом наборе составляет 92% (2 274 666 шт.) против 84% (642 516 шт.) в исходном наборе данных «Mirai Botnet».

Анализ 15 наиболее значимых атрибутов с учетом добавления МСФР позволил сделать выводы, что 5 атрибутов, связанных с МСФР – уникальны.

В результате проведенного эксперимента по многоклассовой классификации получены результаты, представленные в (табл. 5) и (табл. 6). В (табл. 5) приведены результаты оценки эффективности классификации при глубине решающих деревьев *RF depth* 1, а в (табл. 6) при глубине решающих деревьев *RF depth* 2.

Обе таблицы разделены на две части:

В первой части приведены результаты классификации для экспериментальных данных *без добавления* спектра МСФР для КА «*Mirai botnet*» для классовых меток $Label_{experiment} = (normal, miraibotnet, OSscan)$, оцененных методом OVR;

Во второй части приведены результаты классификации для экспериментальных данных *с добавлением* спектра МСФР для КА «*Mirai botnet*» для классовых меток $Label_{experiment} = (normal, miraibotnet, OSscan)$, оцененных методом OVR.

Как видно без добавления МСФР для «Mirai botnet», эффективность классификации КА «*OS scan*» относительно «*Mirai botnet*» и «*Normal*» близка к идеальной ($AUC_{OVR\ «OS\ Scan»}$ = 0.997). Эффективность классификации «*Mirai botnet*» относительно «*OS scan*» и «*Normal*» ниже и составляет $AUC_{OVR\ «Mirai»}$ = 0.937. Эффективность «*Normal*» относительно «*Mirai botnet*» и «*OS scan*» определяется ошибками классификации для КА.

После добавления МСФР для КА «*Mirai botnet*», эффективность классификации «*Mirai botnet*» относительно «*OS scan*» и «*Normal*» возрастает до значений, близких к 1 (выигрыш 7,6% по $AUC_{OVR\ «Mirai»}$). Очевидно, что наблюдаемые ошибки вызваны недостаточной глубиной решающих деревьев.

Анализ эффективности классификации «*OS Scan*» относительно «*Mirai botnet*» и «*Normal*» выявил незначительное ухудшение показателей эффективности. Анализ структуры алгоритма построения решающих деревьев *RF* показывает, что деревья формируют решающие правила на основании энтропии по иерархическому принципу. Наибольшая энтропия «сконцентрирована» в атрибутах №116...120, и как минимум одно решающее правило каждого дерева (из двух возможных при глубине depth 1) относится к данному множеству. Поскольку атрибуты №116...120 не информативны для КА «*OS scan*», решающие деревья классифицируют данную КА хуже, чем «*Mirai botnet*».

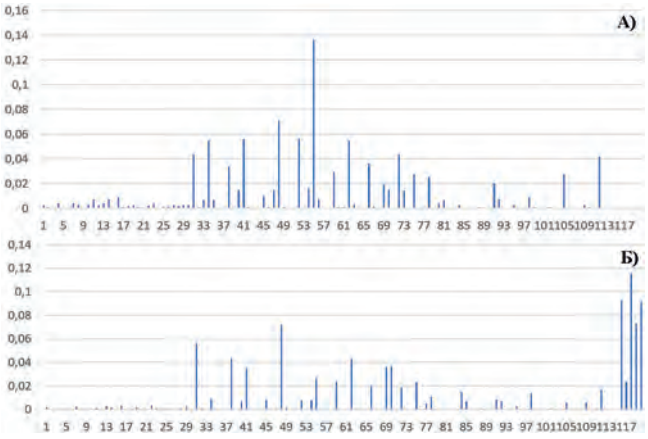

Рис. 4. Оценка информативности (9) атрибутов в задаче многоклассовой классификации КА «Mirai Botnet» и «OS Scan» по критерию Джини для двух случаев: а) без добавления спектра МСФР (атрибуты 116–120 приравнены 0); б) с добавлением спектра МСФР

Таблица 5.

Оценки эффективности многоклассовой классификации алгоритмом RF с depth 1 методом OVR для каждой классовой метки для двух наборов экспериментальных данных для КА «Mirai botnet»

Метрики	Экспериментальные данные					
	Без добавления МСФР для КА «Mirai botnet»			После добавления МСФР для КА «Mirai botnet»		
	«*Mirai botnet*» относительно «*OS scan*» и «*Normal*»	«*OS scan*» относительно «*Mirai botnet*» и «*Normal*»	«*Normal*» относительно «*Mirai botnet*» и «*OS scan*»	«*Mirai botnet*» относительно «*OS scan*» и «*Normal*»	«*OS scan*» относительно «*Mirai botnet*» и «*Normal*»	«*Normal*» относительно «*Mirai botnet*» и «*OS scan*»
accuracy	0,966	0,999	0,966	0,999	0,999	0,999
precision	0,993	1	0,956	1	1	0,999
recall	0,876	0,993	0,998	0,999	0,993	1
f_{score}	0,931	0,997	0,977	0,999	0,996	0,999
$ROC\ AUC_{OVR}$	0,937	0,997	0,942	0,999	0,996	0,999

Содержимое (табл. 5) визуализировано на гистограмме (рис. 5). Построено 5 групп гистограмм по каждой метрике оценки эффективности классификации. В каждой группе значения упорядочены по двум «тройкам»:

- Экспериментальные данные без добавления спектра МСФР для «*Mirai botnet*»: «*Mirai botnet*» относительно «*OS scan*» и «*Normal*»; «*OS scan*» относительно «*Mirai botnet*» и «*Normal*»; «*OS scan*» относительно «*Mirai botnet*» и «*Normal*».
- Экспериментальные данные с добавлением спектра МСФР для «*Mirai botnet*»: «*Mirai botnet*» относительно «*OS scan*» и «*Normal*»; «*OS scan*» относительно «*Mirai botnet*» и «*Normal*»; «*OS scan*» относительно «*Mirai botnet*» и «*Normal*».

Численные значения оценок эффективности многоклассовой классификации алгоритмом RF с глубиной

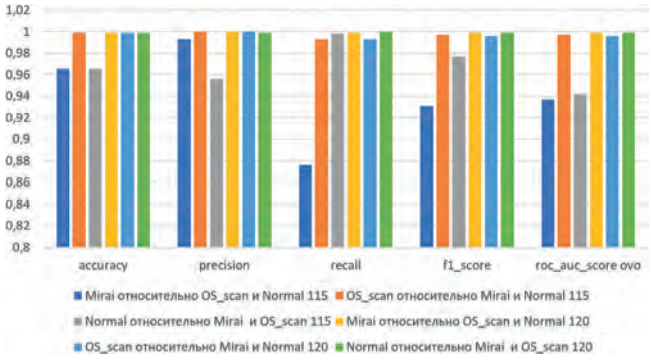

Рис. 5. Визуализация эффективности многоклассовой классификации алгоритмом RF с глубиной решающего дерева depth 1 OVR для каждой классовой метки для двух наборов атрибутов экспериментальных данных.

Таблица 5.

Оценки эффективности многоклассовой классификации алгоритмом RF с depth 1 методом OVR для каждой классовой метки для двух наборов экспериментальных данных для КА «Mirai botnet»

Метрики	Экспериментальные данные					
	без добавления спектра МСФР в КА «Mirai botnet»			после добавления спектра МСФР в КА «Mirai botnet»		
	«*Mirai botnet*» относительно «*OS scan*» и «*Normal*»	«*OS scan*» относительно «*Mirai botnet*» и «*Normal*»	«*Normal*» относительно «*Mirai botnet*» и «*OS scan*»	«*Mirai botnet*» относительно «*OS scan*» и «*Normal*»	«*OS scan*» относительно «*Mirai botnet*» и «*Normal*»	«*Normal*» относительно «*Mirai botnet*» и «*OS scan*»
accuracy	0.999	0.999	0.999	1	0.999	0.999
precision	1.0	1.0	0.999	1	1	0.999
recall	0.999	0.991	1.0	1	0.998	1.0
f_{score}	0.999	0.995	0.999	1	0.999	0.999
$ROC\ AUC_{OVR}$	0.999	0.995	0.999	1	0.999	0.999

решающего дерева *depth2* методом OVR для каждой классовой метки для двух наборов экспериментальных данных приведены в (табл. 6).

В сравнении данными (табл. 4), видно, что при глубине решающего дерева *depth* 5, каждая из двух КА классифицируется с эффективностью, близкой к идеальной.

Содержимое (табл. 6) визуализировано на гистограмме (рис. 6). Для каждой метрики оценки эффективности классификации построено пять групп гистограмм.

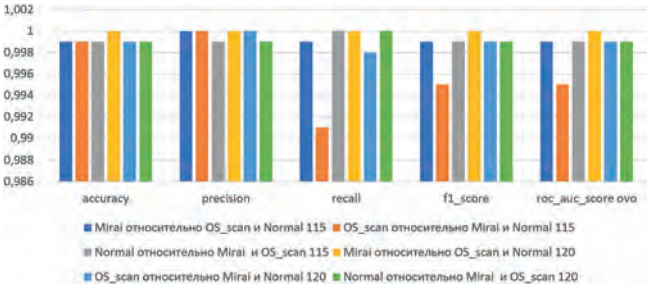

Рис. 6. Визуализация оценки эффективности многоклассовой классификации алгоритмом RF с глубиной решающего дерева depth 2 OVR для каждой классовой метки для двух наборов атрибутов экспериментальных данных

Как видно добавление в качестве дополнительных атрибутов МСФР для КА «*Mirai botnet*», повышает эффективность классификации «*Mirai botnet*» относительно «*OS scan*» и «*Normal*» до 1. Это означает, что все записи, связанные с КА «*Mirai botnet*» в тестовой выборке, классифицированы корректно.

Сравнительный анализ эффективности классификации «*OS scan*» относительно «*Mirai botnet*» и «*Normal*» представленных в таблице 4 выявил незначительное увеличение качества классификации, связанное с перераспределением информативной значимости атрибутного пространства. В отличие от эксперимента с глубиной решающего дерева *depth* = 2, глубины *depth* = 5 «хватает» для построения эффективных решающих правил.

Заключение

Введено понятие *мультифрактального спектра фрактальной размерности* (*МСФР*) в виде последовательности текущих оценок ФР $\hat{H}_{t_{Дi}}$ в окне анализа Δ фиксированной длины Δ = const в зависимости от интервала разрешения.

Проведена оценка эффективности добавления МСФР к исходным данным.

Показано, что в случае бинарной классификации добавление МСФР КА «*Mirai Botnet*» в атрибутное пространство позволяет однозначно классифицировать указанную КА без ложноположительных и ложноотрицательных результатов классификации.

В случае многоклассовой классификации, при добавлении МСФР, эффективность классификации «*Mirai botnet*» алгоритмом Random Forest при глубине решающего дерева depth = 2 относительно «*OS scan*» и «*Normal*» возрастает до значений, близких к 1 (выигрыш 7,6% по $AUC_{OVR\ «Mirai»}$). Наблюдаемые ошибки вызваны недостаточной глубиной решающих деревьев. Добавление в качестве дополнительного параметра фрактальной размерности для КА «Mirai botnet» при глубине решающего дерева depth = 5, повышает эффективность классификации «*Mirai botnet*» относительно «*OS scan*» и «*Normal*» до 1. Добавление МСФР при глубине решающего дерева depth = 5, позволяет достичь идеальной классификации КА ($AUC_{OVR\ «Mirai»}$ = 1).

Литература

1. *Akopov A., Beklaryan L. Traffic Improvement in Manhattan Road Networks with the Use of Parallel Hybrid Biobjective Genetic Algorithm // IEEE Access. 2024. № 12. С. 19532-19552. DOI: 10.1109/ACCESS.2024.3361399.*
2. *Xing Z., Huang M., Li W., Peng D. Spatial linear transformer and temporal convolution network for traffic flow prediction. Scientific Reports. 2024. № 14. С. 1–14. DOI: 10.1038/s41598-024-54114-9.*
3. *Sankaranarayanan, M., Mala, C., Jain, S. Traffic Density Estimation for Traffic Management Applications Using Neural Networks. International Journal of Intelligent Information Technologies. 2024. № 20. С. 1–19. DOI: 10.4018/IJIIT.335494.*
4. *Шелухин О. И. Сетевые аномалии. Обнаружение, локализация, прогнозирование. М: Горячая линия – Телеком, 2019. 448 с. ISBN: 978-5-9912-0756-0*
5. *Sheluhin, O. Kazhemskiy M. Influence Of Fractal Dimension Statistical Charachteristics On Quality Of Network Attacks Binary Classification // Conference of Open Innovations Association, FRUCT. Helsinki: FRUCT Association, 2021. № 28. С. 407–413.*
6. *Sheluhin O. I., Rybakov S. Y., Vanyushina A. V. Detection of network anomalies with the method of fixing jumps of the fractal dimension in the online mode // Wave Electronics and Its Application in Information and Telecommunication Systems. 2022. Т. 5. № 1. С. 430–435.*
7. *Шелухин О. И., Рыбаков С. Ю., Ванюшина А. В. Влияние фрактальной размерности на качество классификации компьютерных атак методами машинного обучения // Наукоемкие технологии в космических исследованиях Земли. 2023. Т. 15. № 1. С. 57–64. DOI 10.36724/2409-5419-2023-15-1-57-64*
8. *Котенко И. В., Саенко И. Б., Лаута О. С., Крибель А. М. Метод раннего обнаружения кибератак на основе интеграции фрактального анализа и статистических методов // Первая миля. 2021. № 6 (98). С. 64–71. DOI: 10.22184/2070-8963.2021.98.6.64.70*
9. *Котенко И. В., Саенко И. Б., Лаута О. С., Крибель А. М. Методика обнаружения аномалий и кибератак на основе интеграции методов фрактального анализа и машинного обучения // Информатика и автоматизация. 2022. Т. 21. № 6. С. 1328–1358. DOI: 10.15622/ia.21.6.9*
10. *Перов Р. А., Лаута О. С., Крибель А. М., Федулов Ю. В. Метод выявления аномалий в сетевом трафике // Наукоемкие технологии в космических исследованиях Земли. 2022. Т. 14. № 3. С. 25–31. DOI: 10.36724/2409-5419-2022-14-3-25-31*

11. *Carvalho G., Woungang I., Anpalagan, A. Cloud Firewall Under Bursty and Correlated Data Traffic: A Theoretical Analysis // IEEE Transactions on Cloud Computing. 2020. Т. 20. №3. С. 1620–1633. DOI: 10.1109/TCC.2020.3000674.*
12. *Liu Y., Tang J., Wang J., Wu H., Chen Y. Fractional analytics hidden in complex industrial time series data: a case study on supermarket energy use // В сборнике «2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China», 23–27 July 2019. 2019. С. 1–6. DOI: 10.1109/ICIAI.2019.8850769.*
13. *Di Mauro M., Liotta A. An Experimental Evaluation and Characterization of VoIP Over an LTE-A Network // IEEE Transactions on Network and Service Management. 2020. С. 1626–1639. DOI: 10.1109/TNSM.2020.2995505.*
14. *Poltavtseva M., Andreeva T. Multi-Dimensional Data Aggregation in the Analysis of Self-Similar Processes // Nonlinear Phenomena in Complex Systems. 2020. Т. 23. С. 262–269. DOI: 10.33581/1561-4085-2020-23-3-262-269.*
15. *Butakova, M. A., Chernov, A. V., Kovalev, S. M., Sukhanov, A. V., Zajaczek, S. Network Traffic Anomaly Detection in Railway Intelligent Control Systems Using Nonlinear Dynamics Approach. В сборнике «Zelinka, I., Brandstetter, P., Trong Dao, T., Hoang Duy, V., Kim, S. (eds) AETA 2018 – Recent Advances in Electrical Engineering and Related Sciences: Theory and Application. AETA 2018. Lecture Notes in Electrical Engineering, vol 554. Springer, Cham». ISBN: 978-3-030-14906-2. DOI: 10.1007/978-3-030-14907-9_46*
16. *Dadkhah S., Carlos Pinto Neto C., Ferreira R., Chukwuka Molokwu R., Sadeghi S., Ghorbani, A. CICIoMT2024: Attack Vectors in Healthcare devices-A Multi-Protocol Dataset for Assessing IoMT Device Security // Preprints. 2024. С. 1–30. DOI: 10.20944/preprints202402.0898.v1*
17. *Aksoy A. & Valle L., Kar G. Automated Network Incident Identification through Genetic Algorithm-Driven Feature Selection // Electronics. 2024. №. 13. Т. 293. С. 1–25. DOI: 10.3390/electronics13020293.*
18. *Miyamoto, K., Goto, H., Ishibashi, R., Han, C., Ban, T., Takahashi, T., Takeuchi, J. Malicious Packet Classification Based on Neural Network Using Kitsune Features. // Intelligent Systems and Pattern Recognition – 2nd International Conference, ISPR 2022, Revised Selected Papers. 2022. С. 306–314. DOI: 10.1007/978-3-031-08277-1_25*
19. *Alabdulatif A., Rizvi S. Machine Learning Approach for Improvement in Kitsune NID // Intelligent Automation & Soft Computing. 2022. Т. 32. С. 827–840. DOI: 10.32604/iasc.2022.021879.*
20. *Шелухин О. И., Раковский Д. И. Многозначная классификация компьютерных атак с использованием искусственных нейронных сетей с множественным выходом // Труды учебных заведений связи. 2023. Т. 9. № 4. С. 97-113. DOI: 10.31854/1813-324X-2023-9-4-97-113*
21. *Valverde-Albacete, Francisco J. & Peláez-Moreno, Carmen. (2024). A Formalization of Multilabel Classification in Terms of Lattice Theory and Information Theory: Concerning Datasets. Mathematics. №12. Т. 346. С. 1–31. DOI: 10.3390/math12020346.*
22. *Veeramsetty V., Reddy K. R., Santhosh M., Mohnot A., Singal G. Short-term electric power load forecasting using random forest and gated recurrent unit // Electrical Engineering. 2022. Т. 104. С. 307–329. DOI: 10.1007/s00202-021-01376-5.*
23. *Rao R. S., Umarekar A., Pais A. R. Application of word embedding and machine learning in detecting phishing websites // Telecommunication Systems. 2022. Т. 79, № 1, С. 33–45. DOI: 10.1007/s11235-021-00850-6.*
24. *Vijayakumar D. S., Ganapathy S. Multistage Ensembled Classifier for Wireless Intrusion Detection System // Wireless Personal Communications. 2022. Т. 122, № 1, С. 645–668. DOI: 10.1007/s11277-021-08917-y.*
25. *Behdani Z., Darehmiraki M. An Alternative Approach to Rank Efficient DMUs in DEA via Cross-Efficiency Evaluation, Gini Coefficient, and Bonferroni Mean // Journal of the Operations Research Society of China. 2022. Т. 10, № 4. С. 763–783. DOI: 10.1007/s40305-019-00264-x.*