УДК 621.8

ФОРМИРОВАНИЕ ПРОСТРАНСТВА ПРИЗНАКОВ ДЛЯ РАСПОЗНАВАНИЯ АНОМАЛИЙ СЕТЕВОГО ТРАФИКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МУЛЬТИФРАКТАЛЬНОГО АНАЛИЗА

А.А. Дидык
Херсонский национальный технический университет,
adidyk@mail.ru

У статті розглянуто використання методів мультифрактального аналізу для вирішення задачі розпізнавання аномалій мережного трафіку. Приведені порівняльні характеристики мультифрактальних спектрів нормального і аномального (що містить деякі види атак) трафіків. Показано, що мультифрактальні спектри двох класів трафіків розрізняються, що надає можливість використовувати мультифрактальний формалізм для формування простору ознак в задачах розпізнавання атак на комп'ютерні системи.
Ключові слова: мультифрактальний аналіз, інформаційна безпека, трафік, задачі розпізнавання

Usage of multifractal formalism for analysis of network traffic structure for the purpose of anomalies revealing are considered in this paper. Multifractal spectrums of normal and abnormal (with presence of some sorts of network attacks) traffics are presented. It's shown, that multifractal spectrums of two sorts of traffic considerably differ and that gives possibility to detect in due time abnormal activity in computer systems.
Keywords: multifractal analysis, information security, traffic, tasks of recognition

В статье рассмотрено использование методов мультифрактального анализа для решения задачи распознавания аномалий сетевого трафика. Приведены сравнительные характеристики мультифрактальных спектров нормального и аномального (содержащего некоторые виды атак) трафиков. Показано, что мультифрактальные спектры двух классов трафиков различаются, что предоставляет возможность использовать мультифрактальный формализм для формирования пространства признаков в задачах распознавания атак на компьютерные системы.
Ключевые слова: мультифрактальный анализ, информационная безопасность, трафик, задачи распознавания

Введение

Современные системы обнаружения вторжения (СОВ) способны в режиме реального времени отслеживать сетевую активность и активность операционной системы с целью обнаружения несанкционированных операций и автоматического реагирования на них фактически в режиме реального времени. Кроме того, СОВ могут анализировать текущие события, учитывая уже произошедшие события, что позволяет идентифицировать распределенные во времени атаки и, таким образом, предсказывать будущие события.

Проблема обнаружения вторжений/аномалий - важная составляющая информационной безопасности. Обнаружение вторжений/аномалий - процесс идентификации вычислительной или сетевой деятельности, которая является злоумышленной или несанкционированной. Большинство систем обнаружения вторжения имеют подобные друг другу структуру и набор компонентов. Каждая СОВ состоит из некоторого набора датчиков или агентов, которые контролируют один или более источников данных, применяя некоторый тип

алгоритма обнаружения, и затем посылают предупреждения или реагируют определенным образом в случае атаки или обнаружения «ненормальной» (аномальной) деятельности.

Современные методы обнаружения аномалий реагируют на изменения в сетевом трафике в момент их возникновения, т.е. когда атака на компьютерную систему уже выполняется. Это зачастую приводит к некоторой задержке в реакции систем обнаружения вторжений на злоумышленную сетевую активность. Такая задержка может привести к потерям в компьютерной системе, что естественно является крайне нежелательным.

Поэтому, одной из важнейших проблем для решения задачи защиты компьютерных систем от несанкционированных вторжений является разработка методов предобработки данных сетевого трафика, которые позволят обнаруживать предвестники аномалий в сетевом трафике. Детектирование таких предвестников позволит формировать «сигналы опасности», реагируя на которые, СОВ сможет заранее перейти в режим повышенной готовности. Это даст возможность избежать задержек в срабатывании систем обнаружения вторжений.

С 1993 г. множество исследований, о которых сообщается в литературе [1,2], показало, что тип трафика данных в широком спектре сетевых ситуаций реального мира хорошо моделируется самоподобными процессами, т.е. имеет фрактальную, а точнее мультифрактальную природу.

Мультифрактал - квазифрактальное множество с переменной фрактальной размерностью. Естественно, что сетевой трафик представляет собой мультифрактал и лучше описывается в терминах именно мультифрактального формализма.

Особое значение фрактального анализа временных рядов в том, что он учитывает поведение системы не только в период измерений, но и его предысторию. Эта особенность фрактального формализма применительно к анализу сетевого трафика позволяет успешно идентифицировать атаки, распределенные во времени. Идентификация этого класса атак является актуальной проблемой для современных систем обнаружения несанкционированных вторжений.

Общая постановка задачи обнаружения аномалий

В общем случае процесс обнаружения аномалий можно интерпретировать как отнесение исследуемого объекта, например элемента сетевого трафика, представленного выборкой наблюдений, к одному из двух альтернативных классов – *нормальный* и *аномальный*, *«свой»* и *«чужой»*. Непосредственно процедура соотнесения базируется на существующих различиях некоторой упорядоченной совокупности признаков распознавания, которые традиционно формируют на основе полученных в результате наблюдений параметров трафика: общее количество сетевых пакетов в секунду, количество управляющих ICMP пакетов в секунду, количество обращений к сетевому ресурсу и т.д.

С целью преодоления априорной неопределенности о распознаваемых классах предварительно производится обучение. Для этого подбирают образцы трафика, называемые эталонами, которые наиболее полно характеризуют каждый из классов. Затем по совокупности наблюдений эталонов при различных начальных отсчетах осуществляется набор признаков [3]:

$$X^{(l)} = \begin{bmatrix} x_{11}^{(l)} & x_{12}^{(l)} & \dots & x_{1m}^{(l)} & \dots & x_{1M}^{(l)} \\ x_{21}^{(l)} & x_{22}^{(l)} & \dots & x_{2m}^{(l)} & \dots & x_{2M}^{(l)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{r1}^{(l)} & x_{r2}^{(l)} & \dots & x_{rm}^{(l)} & \dots & x_{rM}^{(l)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{R1}^{(l)} & x_{R2}^{(l)} & \dots & x_{Rm}^{(l)} & \dots & x_{RM}^{(l)} \end{bmatrix},$$

где m — количество обучающих наблюдений по эталону l-го класса, $l = 1,2,\dots,L$, L — число классов (в нашем случае, два), а каждый столбец $X_m^T = (x_{1m}, x_{2m}, \dots, x_{Rm})$, $m = 1,2,\dots,M$ матрицы X есть R-мерный вектор наблюдаемых значений R-признаков $X_1, X_2, \dots, X_R$, отражающих свойства распознаваемых объектов; T – знак транспонирования.

В результате обучения создаются эталонные описания, содержащие признаки $\{X^{(l)}\}$ по всем L распознаваемым классам.

Затем в виде матрицы формируется совокупность полученных наблюдений распознаваемого трафика:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2n} & \cdots & y_{2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{r1} & y_{r2} & \cdots & y_{rn} & \cdots & y_{rN} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{R1} & y_{R2} & \cdots & y_{Rn} & \cdots & y_{RN} \end{bmatrix},$$

где y_{rn} – значение признака r в n-й реализации;

N – количество наблюдений, используемых для распознавания.

В общем случае $n \neq m$, т.е. число наблюдений m в эталонных описаниях может не соответствовать числу наблюдений n в принятых реализациях.

При таком подходе процесс распознавания можно трактовать как последовательное сравнение контрольной выборки признаков Y с L выборками эталонных описаний $X^{(l)}$, которые могут выступать или в качестве матриц размерностью $R \times M$, или вектора длиной R, каждое значение которого усреднено по M реализациям.

Окончательное решение о соотнесении объекта к одному из классов принимается при выборе максимального значения отношения подобия, полученного в результате парного сравнения принятой реализации с эталонными описаниями [4]:

$$\frac{w\left(Y/Y^{(l)}\right)}{w\left(X/X^{(l)}\right)} \geq < C_l,$$

где $Y^{(l)}$, $X^{(l)}$ – R-мерные вектора наблюдаемых значений признаков, характеризующих l-й класс и вектора признаков из подготовленных L эталонных описаний; $w\left(Y/Y^{(l)}\right)$, $w\left(X/X^{(l)}\right)$ – функция подобия; C_l – порог, рассчитанный априорно по эталонным описаниям.

Таким образом, в общем случае задача обнаружения аномалий сводится к принятию решения о принадлежности наблюдаемого объекта к одному из двух классов $s_i^{(1)}, s_i^{(2)}$, описываемых одинаковым набором признаков $Y_1, Y_2, \ldots, Y_R$. Различия между классами будут проявляться только в различии характеристик признаков у разных объектов. Тогда для любого набора признаков $X_1, X_2, \ldots, X_R$ можно задать правило, согласно которому двум классам $s_i^{(1)}$, $s_i^{(2)}$ ставится в соответствие вектор

$$D = \begin{vmatrix} d_1 \\ . \\ . \\ d_q \end{vmatrix},$$

состоящий из q скаляров d, называемых межклассовыми расстояниями. Каждый из скаляров d выражает степень отличия характеристик данных признаков у рассматриваемых классов.

Важнейшей задачей процесса обнаружения аномалий является определение набора признаков $Y_1, Y_2, \ldots, Y_R$, т.е. формирование признакового пространства таким образом, чтобы при минимально возможной размерности R обеспечить требуемую достоверность классификации.

Таким образом, **цель работы** — *исследовать применение методов мультифрактального анализа для формирования пространства признаков при решении задач обнаружения аномалий в сетевом трафике.*

Мультифрактальный формализм

Само понятие фрактал было введено Бенуа Мандельбротом в семидесятые годы. Термин происходит от латинского fractus, прилагательного от глагола «frangere» – ломать, разбивать на части. То есть, как гласит одно из определений фракталов, фрактал – это множество, части которого подобны целому.

Основной характеристикой фрактального объекта является его размерность [5]. Пусть множество G в пространстве R^n. Разобьем пространство R^n на n-мерные кубы с длиной ребра δ и обозначим число кубов, необходимых для покрытия ими множества G, через $N(\delta)$. Тогда

величина размерности Хаусдорфа-Безиковича D должна удовлетворять следующему условию:

$$\lim_{\delta \to 0} N(\delta)\delta^d = \begin{cases} 0, & d > D \\ \infty, & d < D \end{cases}$$

Данное определение можно упростить, сделав его более удобным для практического применения. Видно, что при $\delta \approx 0$, оно эквивалентно

$$D \approx -\ln N(\delta)/\ln \delta$$

или

$$D = \lim_{\delta \to 0} \frac{\ln N(\delta)}{\ln\left(\frac{1}{\delta}\right)}.$$

Фрактальная размерность, как правило, является дробной величиной, отражающей, некоторым образом, геометрическую сложность объекта.

Реальные физические объекты и сигналы, даже обладающие признаками самоподобия, очень редко могут быть описаны с помощью лишь одной величины фрактальной размерности. Именно поэтому в последнее время получил большое распространение анализ, основанный на теории мультифракталов - неоднородных фрактальных объектов. Понятие мультифрактала предоставляет новые обширные возможности фрактального анализа сложных стохастических процессов. Для характеристики мультифрактала недостаточно одной величины, его фрактальной размерности, а необходим бесконечный спектр таких размерностей. Идея мультифрактального анализа состоит в разложении исследуемого множества со сложной статистикой по множествам однородных фракталов с четко выраженной фрактальной размерностью. При этом мультифрактальный анализ может привести к нетривиальным результатам в применении не только к самоподобным объектам с фрактальной геометрией.

Стандартный метод определения фрактальной размерности не позволяет обнаруживать различия между однородными фрактальными объектами и неоднородными (мультифрактальными). Данный метод, называемый клеточным методом, заключается в том, что исходное множество покрывается областями (ячейками или клетками) меньшего масштаба ε. Чем меньше размер ячейки, тем меньше величина ее заселенности. Затем подсчитывается число ячеек, в которых содержится хотя бы одна точка изучаемого множества, и уже эти значения используются в дальнейшем анализе. Поэтому данный метод нечувствителен к неоднородности распределения точек между ячейками.

Таким образом, использование обычной фрактальной размерности не позволяет отличить монофрактальные множества от мультифрактальных. Для этого вводятся новые характеристики. При этом процедура определения размерности несколько усложняется: для каждой заполненной ячейки подсчитывается число содержащихся в ней точек $N_i(\varepsilon)$, которое затем преобразуется в долю:

$$p_i(\varepsilon) = \frac{N_i(\varepsilon)}{N},$$

где N – общее число точек множества. Далее используется полученный набор $\{p_i\}$. Этот набор вероятностей является одной из основных характеристик мультифрактала.

Для самоподобных множеств зависимость p_i от размера ячейки ε имеет степенной характер:

$$p_i(\varepsilon) \approx \varepsilon^{\alpha_i},$$

где α_i представляет собой некоторый показатель степени (разный, вообще говоря, для разных ячеек i). Известно, что для регулярного (однородного) фрактала все показатели степени α_i одинаковы и равны фрактальной размерности D:

$$p_i = \frac{1}{N(\varepsilon)} \approx \varepsilon^D.$$

Таким образом, мультифрактал - некое объединение различных однородных фрактальных подмножеств исходного множества, каждое из которых имеет свое собственное значение фрактальной размерности.

Стандартный метод мультифрактального анализа основан на рассмотрении обобщенной статистической суммы $Z(q,\varepsilon)$, в которой показатель степени q может принимать любые значения в интервале $-\infty < q < +\infty$:

$$Z(q,\varepsilon) = \sum_{i=1}^{N(\varepsilon)} p_i^q(\varepsilon).$$

Спектр обобщенных фрактальных размерностей D_q (размерностей Реньи), характеризующих данное распределение точек в области ζ, определяется с помощью соотношения:

$$D_q = \frac{\tau(q)}{q-1},$$

где нелинейная функция $\tau(q)$ (в научной литературе она называется скейлинговой экспонентой) имеет вид:

$$\tau(q) = \lim_{\varepsilon \to 0} \frac{\ln(Z(q,\varepsilon))}{\ln \varepsilon}.$$

Если $D_q = D = const$, т.е. не зависит от q, то данное множество точек представляет собой обычный фрактал, который характеризуется всего лишь одной величиной – фрактальной размерностью D. Напротив, если функция D_q как-то меняется с q, то рассматриваемое множество точек является мультифракталом.

Таким образом, мультифрактал в общем случае характеризуется скейлинговой экспонентой $\tau(q)$, определяющей поведение статистической суммы $Z(q,\varepsilon)$ при $\varepsilon \to 0$:

$$Z(q,\varepsilon)=\sum_{i=1}^{N(\varepsilon)} p_i^q(\varepsilon)\approx \varepsilon^{\tau(q)}.$$

Еще одной важной характеристикой мультифрактала является мультифрактальный спектр (спектр сингулярностей мультифрактала) $f(\alpha)$. Физический смысл функции $f(\alpha)$ заключается в том, что она представляет собой хаусдорфову размерность некоего однородного фрактального подмножества L_α из исходного множества L, характеризуемого одинаковыми вероятностями заполнения ячеек $p_i \approx \varepsilon^\alpha$. Таким образом, набор различных значений функции $f(\alpha)$ (при разных α) представляют собой спектр фрактальных размерностей однородных подмножеств L_α, на которые можно разбить исходное множество L.

α_i представляет собой некоторый показатель степени, вообще говоря разный, для разных ячеек i (в специальной литературе по отношению к нему используются термины “показатель сингулярности” или “экспонента сингулярности”). Чем меньше значение α_i, тем более сингулярной является мера. Известно, что для регулярного (однородного) фрактала все показатели степени α_i одинаковы и равны фрактальной размерности D:

$$p_i = 1/N(\varepsilon)\approx \varepsilon^D .$$

Однако для такого более сложного объекта, как мультифрактал, вследствие его неоднородности, вероятности заполнения ячеек p_i в общем случае неодинаковы, и показатель степени α_i для разных ячеек может принимать различные значения. Достаточно типичной является ситуация, когда эти значения непрерывно заполняют некоторый закрытый интервал $(\alpha_{\min},\alpha_{\max})$, причем:

$$p_{\min}\approx \varepsilon^{\alpha_{\max}},\ \text{а}\ p_{\max}\approx \varepsilon^{\alpha_{\min}}.$$

Несложно показать, что

$$\left.\frac{d\tau}{dq}\right|_{q\to+\infty} = D_\infty = \alpha_{\min}, \left.\frac{d\tau}{dq}\right|_{q\to-\infty} = D_{-\infty} = \alpha_{\max},$$

т.е. интервал возможных значений α определяется предельными значениями (при $q\to\pm\infty$) обобщенных фрактальных размерностей D_q.

Формально переход от переменных $\{q,\tau(q)\}$ к переменным $\{\alpha, f(\alpha)\}$, задаваемый вышеприведенными соотношениями, может быть осуществлен при помощи следующих преобразований Лежандра:

$$\alpha = \frac{d\tau}{dq},$$

$$f(\alpha) = q\frac{d\tau}{dq} - \tau .$$

Обратное преобразование Лежандра определяется формулами:

$$q = \frac{df}{d\alpha},$$
$$\tau(q) = \alpha \frac{df}{d\alpha} - f\ .$$

Мультифрактальный спектр позволяет получить общую характеристику мультифрактального множества.

Результаты эксперимента

В качестве исходных данных для проведения эксперимента использовался тестовый трафик для оценки систем обнаружения вторжений, произведенный в MIT Lincoln Labs [6]. Эти данные представляют как нормальный, так и аномальный трафик, собранный в тестовой сети, в которой выполнялось моделирование различных сетевых атак. Эти данные предназначены для тестирования систем обнаружения вторжений. Набор данных содержит трафик, собранный в течение нескольких недель. Трафик 1-ой недели является «нормальным», в то время как трафик 2-ой недели содержит аномалии, т.е. атаки различных классов.

На рисунке 2(а) представлены мультифрактальные спектры, полученные преобразованием Лежандра, для 1-го дня 1-ой недели (без атак) и 1-го дня 2-ой недели (с атаками). На рисунке 2(б) представлены такие же спектры для 2-го дня 1-ой недели (без атак) и 2-го дня 2-ой недели (с атаками).

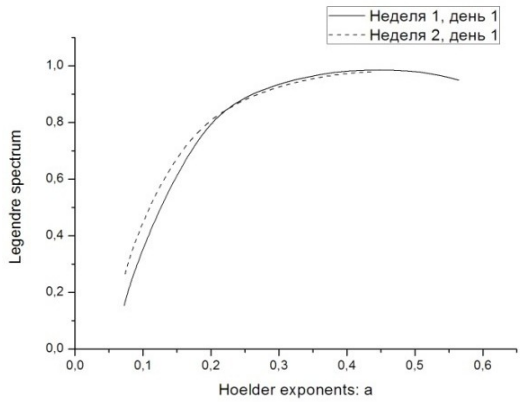

Рис. 2(а)

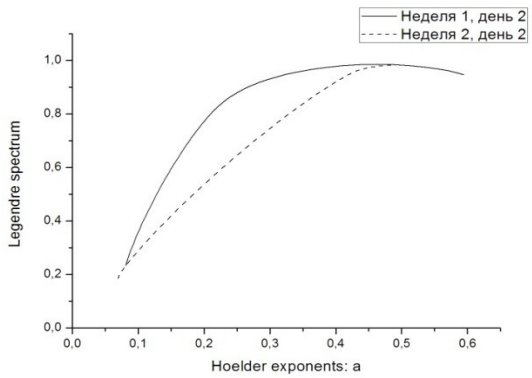

Рис. 2(б)

Мультифрактальные спектры сетевых трафиков.

В таблице 1 приведены описательные статистики мультифрактальных спектров.

Таблица 1.

Описательные статистики мультифрактальных спектров.

	N общее	Среднее	Медиана	Размах	Стандартное отклонение	Ассиметрия	Эксцесс	Коэффициент вариации
Неделя 1, день 1	109	0,561	0,586	0,833	0,302	-0,018	-1,602	0,538
Неделя 2, день 1	101	0,554	0,516	0,714	0,236	0,322	-1,327	0,426
Неделя 1, день 2	109	0,560	0,502	0,752	0,286	0,207	-1,657	0,510
Неделя 2, день 2	101	0,368	0,198	0,797	0,298	1,337	-0,029	0,810

На рисунках 3 и 4 представлены статистические распределения спектров и их коробчатые диаграммы.

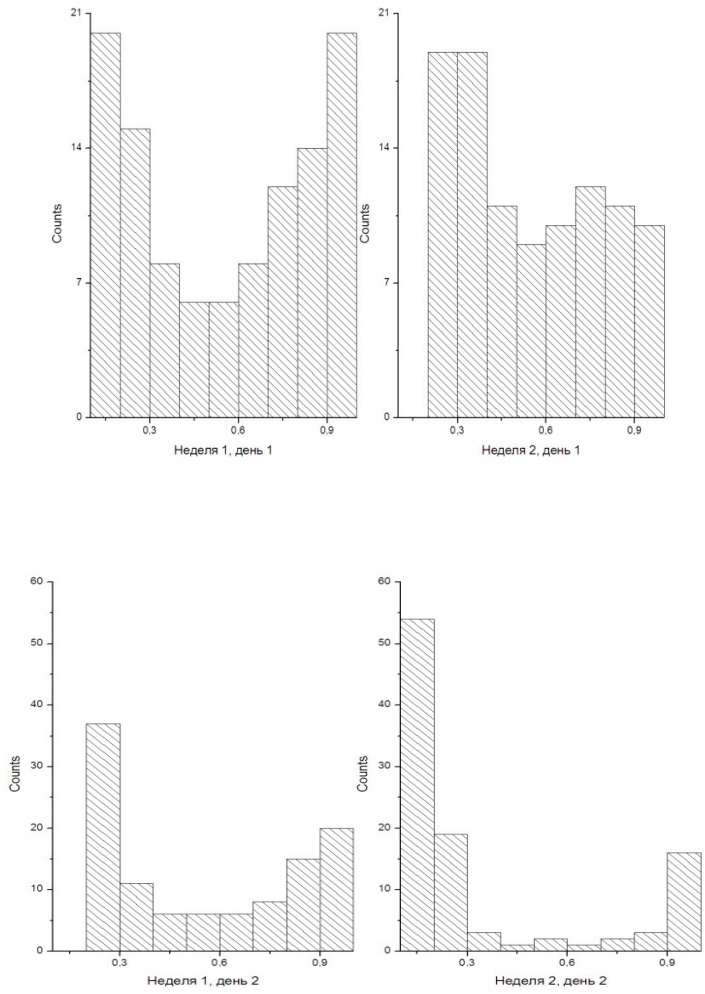

Рис. 3. Статистические распределения мультифрактальных спектров трафиков 1-го дня (вверху) и 2-го дня (внизу)

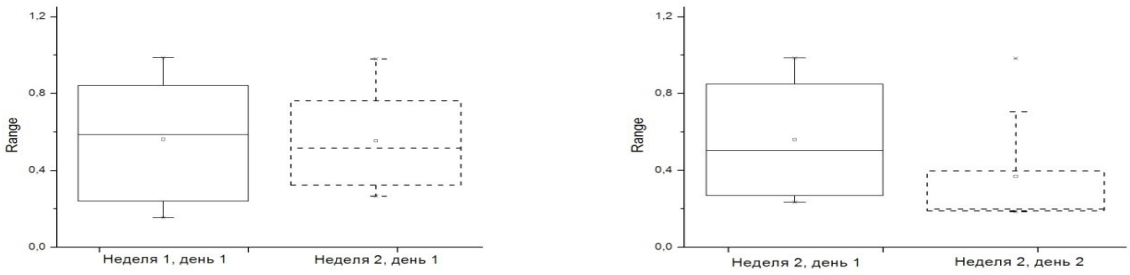

Рис. 4. Коробчатые диаграммы мультифрактальных спектров трафиков 1-го дня (вверху) и 2-го дня (внизу).

Как видно из рисунка 2, спектры трафиков без атак и трафиков, содержащих атаки, отличаются даже визуально, что говорит о различной мультифрактальной структуре двух видов трафиков. По парное сравнение описательных статистик, статистических распределений и коробчатых диаграмм спектров сингулярностей трафиков без атак и с атаками ясно показывает, что наличие атак вносит изменения в статистическую структуру функции мультифрактального спектра.

Выводы.

Атаки, или аномальный трафик, вносят заметные изменения в мультифрактальную структуру сетевого трафика, что видно по изменению значений фрактальных характеристик отдельных частей трафика. Таким образом, возможно эффективное использование методов мультифрактального анализа для формирования пространств признаков при решении задач обнаружения аномалий в сетевом трафике. Анализируя изменения во времени такого показателя, как фрактальная размерность отдельных участков трафика, и того, как на трафик воздействуют внешние и внутренние факторы, можно научиться предсказывать поведение процесса, а также диагностировать и предсказывать аномальные состояния.

Литература

1. *Leland. W.; Taqqu, M.; Willinger, W.; and Wilson, D.* «On the Self-Similar Nature of Ethernet Traffic (Extended Version). IEEE/ACM Transactions on Networking, February 1994.

2. *Crovella, M., and Bestavros, A.* «Self-Similarity in World-Wide Web Traffic: Evidence and Possible Causes». *Proceedings, ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, May 1996.

3. *Дворников С.В., Желнин С.Р., Медведев М.В.* Метод формирования признаков распознавания сигналов диапазона декаметровых волн по их вейвлет-коэффициентам, рассчитанным на основе лифтинговой схемы. // «Информация и космос», № 2, 2006.

4. *Фукунага К.* Введение в статистическую теорию распознавания образов / Пер. с англ. – М.: Наука., 1979.

5. *Федер Е.* Фракталы. – М.: Мир, 1991.–254 с.

6. 1999 DARPA intrusion detection evaluation, MIT Lincoln Labs, 1999. [Online]. *http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html*