МАТЕМАТИКА

А. В. Копыльцов, И. В. Сорокин

ВЕЙВЛЕТ-АНАЛИЗ СТРУКТУРНОЙ ЭНТРОПИИ ФАЙЛОВ

Предлагается алгоритм сегментации файла на однородные участки по уровню их информационной энтропии. Проведено сравнение непрерывного и дискретного вейвлет-преобразования, а в качестве базисных вейвлетов рассмотрены WAVE-вейвлет и вейвлет Хаара. Предлагаемое решение представляет интерес в области информационной безопасности, поскольку позволяет достаточно эффективно анализировать энтропию файла, которая может быть использована как один из важных признаков в системах выявления и классификации вредоносного программного обеспечения.

Ключевые слова: вейвлеты, кратномасштабный анализ, сегментация временных рядов, структурная энтропия файлов.

A. Kopyltsov, I. Sorokin

WAVELET ANALYSIS OF STRUCTURAL ENTROPY OF FILES

An algorithm for splitting a file into homogeneous segments according to their level of information entropy is suggested. The algorithm allows to describe any file by a certain sequence of constituent elements, in other words, by its structural entropy. In addition to describing the algorithm, the article briefly compares the continuous and discrete wavelet transforms, in particular, the WAVE and Haar wavelets. The proposed solution is primarily of interest for information security, because it allows to analyze effectively the structural entropy of any file. This can be used as an important indicator in the systems for detection and classification of malicious software.

Key words: wavelets, multi-resolution analysis, segmentation of time series, structural entropy of files.

Введение

В настоящее время существуют десятки разновидностей программ-упаковщиков, предназначенных для скрытия вредоносного функционала от стандартных антивирусных методик детектирования (сигнатуры, логи эмуляции и т. п.). В большинстве случаев каждое семейство таких программ обладает характерными признаками. Одним из таких признаков является структурная энтропия файла. Под этим подразумевается мера неупорядоченности в расположении составляющих элементов, в данном случае характерных участков файла [9].

В некоторых случаях, просмотрев файл, можно с большой вероятностью судить о его принадлежности к упомянутым программам-упаковщикам. Поэтому анализ структурной энтропии файлов представляет интерес для выявления и классификации вредоносных программ.

Постановка задачи

В работе [7] было предложено решение, позволяющее сравнивать файлы по их структурной энтропии [9], а именно, используя метод скользящего окна, файл представляется в виде временного ряда. Затем полученный временной ряд разбивается на участки различные по значению энтропии. В итоге для сравнения двух файлов используются последовательности их участков с определенным значением энтропии. Однако предложенный алгоритм выявления стационарных участков в файле обладает рядом недостатков. Во-первых, количество полученных участков зависит от длины файла. Во-вторых, алгоритм не учитывает закономерности в расположении повторяющихся участков и к тому же обладает медленной скоростью работы. Для устранения этих недостатков предлагается новый алгоритм сегментации файла с использованием вейвлет-анализа.

Описание модели

В качестве исходных данных рассматривается временной ряд, характеризующий уровень энтропии в файле на всем его протяжении. Значения этого ряда определяются методом скользящего окна с постоянным смещением Δt:

$$f_k = f(t_k),\ t_k = \Delta t k,\ k = 0,\ 1, ..., N-1,$$

где f_r — уровень информационной энтропии в k-м окне, N — общее количество окон в файле [7].

Для анализа полученных временных рядов предлагается использовать вейвлет-анализ, суть которого заключается в следующем. Во-первых, выбирается базисный вейвлет, от свойств которого будет зависеть характер сегментации. Во-вторых, строится вейвлет-преобразование на нескольких масштабах, при этом полученные коэффициенты будут содержать информацию о взаимосвязи используемого вейвлета и анализируемого временного ряда. В итоге, определение сегментов будет основываться на анализе значимых вейвлет-коэффициентов.

Основной задачей при сегментации является возможность выявлять такие места в файле, в которых происходит изменение среднего значения энтропии. Для извлечения подобной информации из анализируемого временного ряда в качестве базисных вейвлетов нами использовались WAVE-вейвлет при непрерывном вейвлет-преобразовании, определенный как первая производная функции Гаусса [5]:

$$\psi_{WAVE}(t) = (-1)^m \partial_t^m \left[\exp(-\frac{t^2}{2}) \right],\ m = 1, \qquad (1)$$

и вейвлет Хаара при дискретном вейвлет-преобразовании:

$$\psi_{HAAR}(t) = \begin{cases} 1, & 0 \le t < 1/2, \\ -1, & 1/2 \le t < 1, \\ 0, & t < 0,\ t \ge 1. \end{cases} \tag{2}$$

Выбор этих вейвлетов обусловлен тем, что они обладают схожими свойствами: несимметричны и имеют один нулевой момент (рис. 1). Единственное отличие между ними в том, что WAVE-вейвлет обладает большей гладкостью, чем вейвлет Хаара, благодаря чему достигается более точное описание локальных свойств исходных данных.

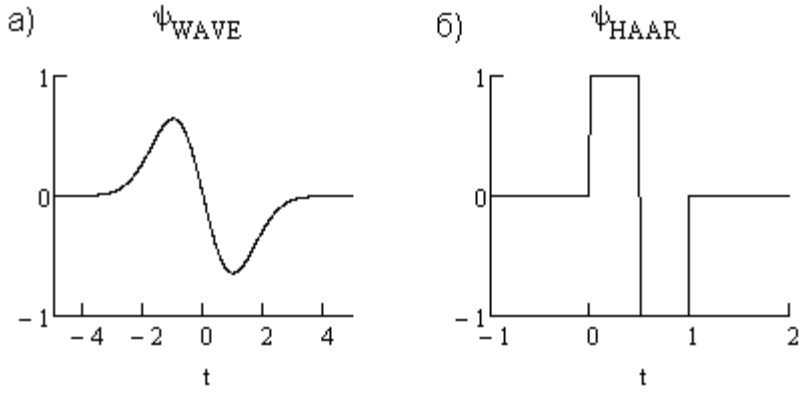

Рис. 1. Графики базисных вейвлетов: а — WAVE-вейвлет; б — вейвлет Хаара

В силу того что при анализе временного ряда необходимо достаточно точно определять характерные места изменения среднего уровня энтропии, использование непрерывного вейвлет-преобразования (НВП) является более предпочтительным, чем дискретное вейвлет-преобразование (ДВП). В частности, благодаря избыточности НВП, связанной с непрерывным изменением масштабного коэффициента и параметра сдвига, оно позволяет более полно и четко анализировать исходные данные [1]. Однако при этом появляются затраты на дополнительные расчеты.

В общем виде интегральное вейвлет-преобразование для функции $f(t) \in L^2(R)$ задается следующим образом:

$$W(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt, \tag{3}$$

где ψ^* — комплексное сопряжение базисного вейвлета (WAVE-вейвлет или вейвлет Хаара), a — параметр масштаба, b — параметр сдвига базисного вейвлета.

Основываясь на том, что функция $f(t) \in L^2(R)$ упакованности файла задана дискретными значениями через равные интервалы, для вычисления коэффициентов $W(a, b)$ используется следующая оценка:

$$W(a,b) = \frac{1}{|a|^{1/2}} \sum_{k=0}^{N-1} f_k \psi\left(\frac{t_k - b}{a}\right), \tag{4}$$

где f_k — уровень информационной энтропии в k-м окне, N — общее количество окон в файле.

Существуют и другие оценки для вычисления коэффициентов, упрощающие процедуру расчетов. В работе [6] предложено решение, основанное на свойстве локальности вейвлет-функции, благодаря которому можно производить суммирование только в некоторой окрестности от данного смещения, а не по всем значениям временного ряда.

Для вычисления ДВП также используется формула (4), но только теперь параметр масштаба изменяется в соотвествии со степенью двойки. Во-первых, благодаря этому появляется возможность использовать кратномасштабный анализ [3], при применении которого на каждом следующем масштабе преобразования используются значения, вычисленные на предыдущем, за счет чего уменьшается количество выполняемых операций сложения. Во-вторых, из-за этого на исходные данные накладывается ограничение на количество отсчетов. Их должно быть кратно степени двойки: $a_n = 2^n$, где n — наибольший масштаб. Поэтому временной ряд с каждой стороны добавляется новыми значениями. Для этого используется метод наименьших квадратов, при этом аппроксимирующий полином первой степени строится только по крайним значениям временного ряда.

Полученные вейвлет-коэффициенты (с использованием либо НВП, либо ДВП) можно преобразовать следующим образом:

$$S(a_i, b_j) = \left| W(a_i, b_j \right|^2. \qquad (5)$$

Из всех коэффициентов можно выделить только значимые, а именно локальные экстремумы, т. е. те которые имеют максимум по переменным a и b. Линии определяемые такими локальными экстремумами принято называть скелетоном [4].

При увеличении масштабного коэффициента локальных экстремумов становится все меньше, в силу того что базисный вейвлет накладывается на все больший диапазон исходного временного ряда. Другими словами, при больших масштабах не учитываются незначительные изменения в анализируемых данных. Благодаря этому появляется возможность задавать желаемый уровень детализации.

С другой стороны, чем больше масштаб преобразования, тем менее четче получаются границы искомых переходов. Это зависит от того, насколько близко они расположены друг к другу. Поэтому предлагаемый алгоритм сегментации определяет значимые вейвлет-коэффициенты, начиная с наименьшего и до заданного максимального масштаба преобразования. Тем самым с каждым новым масштабом происходит корректировка вновь полученных значимых коэффициентов.

Таким образом, итоговое количество сегментов определяется значимыми вейвлет-коэффициентами на заданном максимальном масштабе, а их границы определяются точностью при минимальном масштабе преобразования.

Результаты и их обсуждение

В качестве демонстрационного примера рассмотрим вредоносный файл, который по классификации Dr.Web определяется как Trojan.PWS.Ibank.53. На рис. 2 показан график упакованности файла, построенный с использованием метода скользящего окна. В каждой точке этого графика отображается уровень информационной энтропии в определенном окне, вычисленный на основе частоты появления байт из этого окна [7].

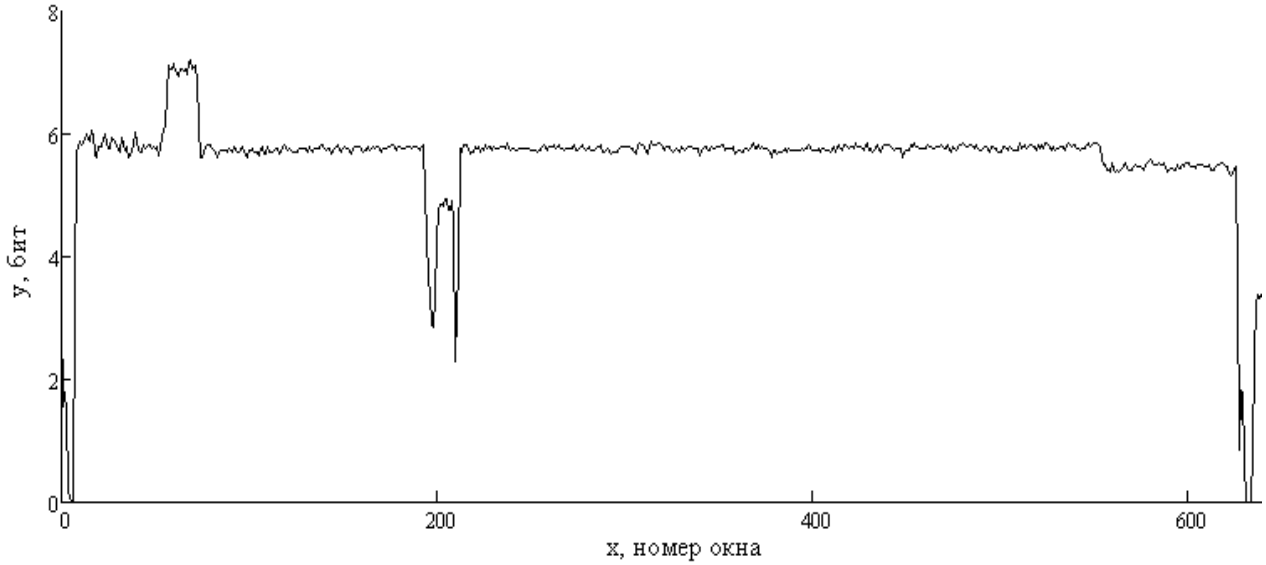

Рис. 2. График упакованности файла, построенный с использованием скользящего окна размером 256 байт и смещением в 128 байт (по оси абсцисс — порядковый номер окна, по оси ординат — уровень информационной энтропии в битах)

Сперва сравним между собой непрерывное и дискретное вейвлет-преобразование. Для того чтобы нагляднее увидеть их различия, рассмотрим преобразования только на первых 250 отсчетах.

На рис. 3 изображена поверхность полученных вейвлет-коэффициентов *W(a*, *b)* при НВП, где в качестве базисного вейвлета использовался WAVE-вейвлет. Каждая точка на такой поверхности показывает, насколько исходные данные согласуются с выбранным базисным вейвлетом.

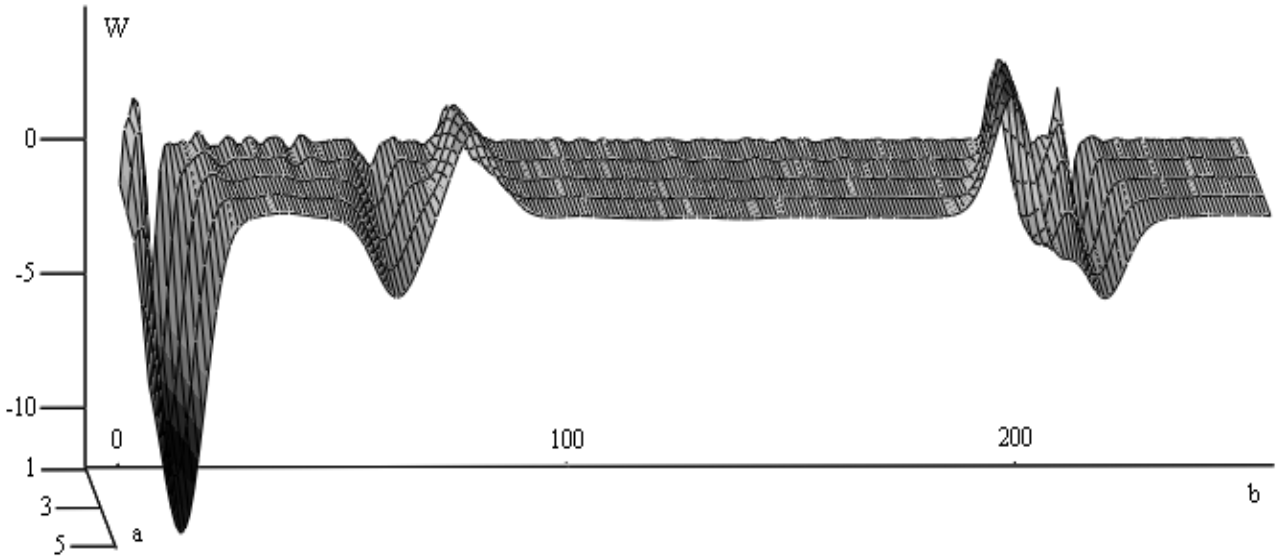

Рис. 3. Спектр вейвлет-коэффициентов при непрерывном вейвлет-преобразовании с использованием WAVE-вейвлета на пяти масштабах преобразования (по оси a — масштаб измеряемый в количестве отсчетов временного ряда, по оси b — сдвиг базисного вейвлета во временном ряду, по оси W — вейвлет-коэффициенты)

Можно заметить, что с увеличением масштабного параметра *a* происходит сглаживание всплесков. Другими словами, на все больших масштабах рассматриваются все более значимые изменения в исходных данных.

Поверхность на рис. 4 в сравнении с рис. 3 наглядно демонстрирует различие в свойствах базисного вейвлета. На нем изображены вейвлет-коэффициенты, полученные при ДВП с использованием вейвлета Хаара. Стоит заметить, что здесь масштабный параметр изменяется в соответствии со степенью двойки, поэтому по оси ординат указывается номер степени, т. е. максимальный масштаб равен $a = 2^4$.

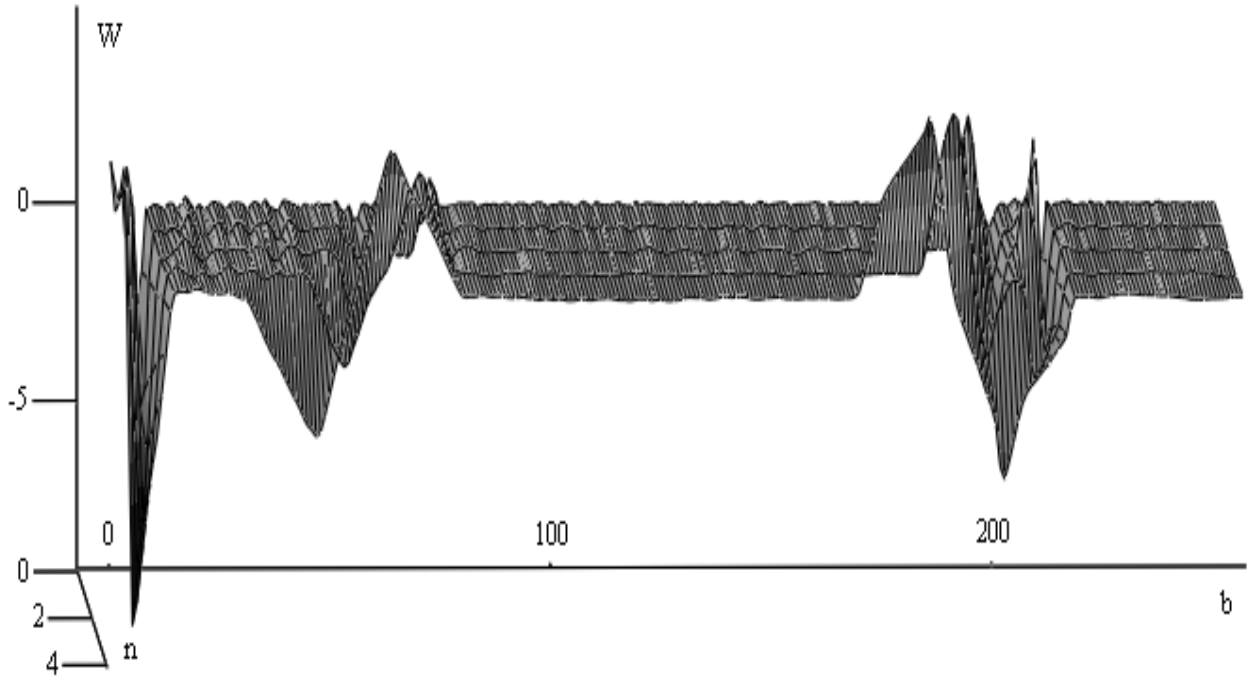

Рис. 4. Спектр вейвлет-коэффициентов при дискретном вейвлет-преобразовании с использованием вейвлета Хаара на пяти масштабах преобразования (по оси n — масштаб заданный степенью двойки и измеряемый в количестве отсчетов временного ряда, по оси b — сдвиг базисного вейвлета во временном ряду, по оси W — вейвлет-коэффициенты)

Теперь рассмотрим преобразованные вейвлет-коэффициенты $S(a, b)$ и их локальные экстремумы. Для этого используется улучшенный способ визуализации, так называемые спектрограмма и скелетон [4; 13].

На рис. 4 представлен вейвлет-анализ демонстрационного примера на первых 250 отсчетах при ДВП с использованием вейвлета Хаара на 7 масштабах преобразования. При построении линий локальных экстремумов (рис. 5в) задавался минимальный порог, для того чтобы отфильтровать незначительные всплески на спектрограмме (рис. 5б). Как можно заметить, полученные значимые вейвлет-коэффициенты очень хорошо отражают структуру исходных данных. Именно на основании этих коэффициентов и строится алгоритм сегментации.

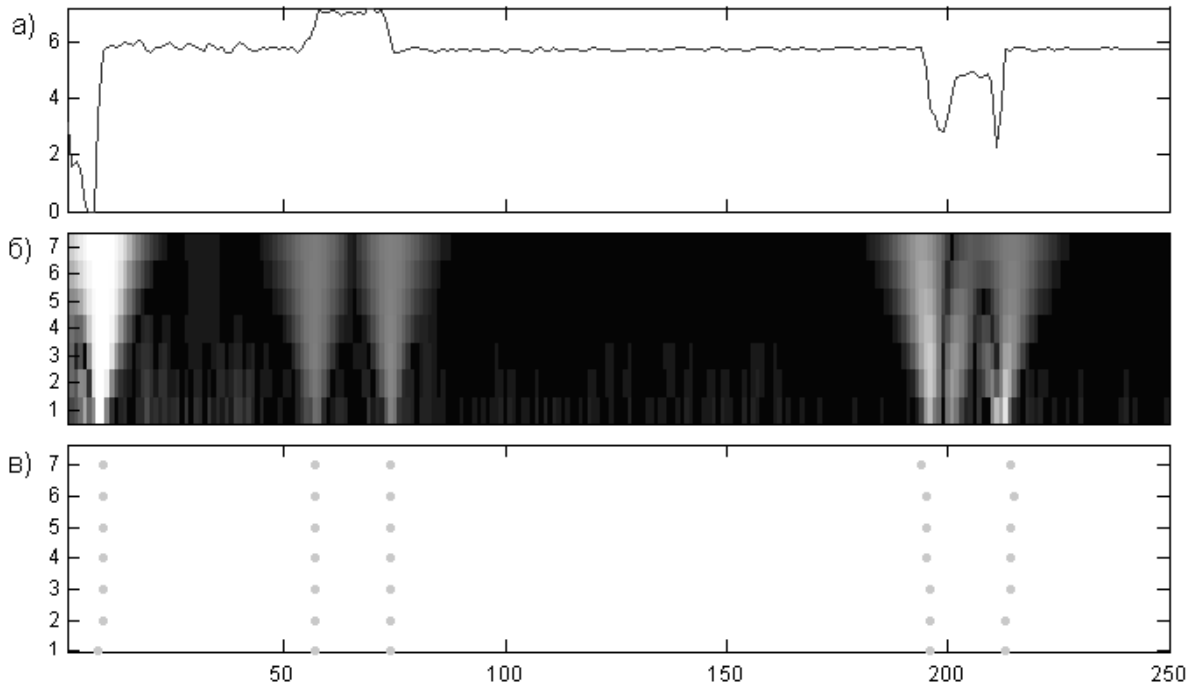

Рис. 5. Вейвлет анализ участка файла на первых 250 отсчетах: а) — график упакованности файла (по оси абсцисс — порядковый номер окна, по оси ординат — уровень информационной энтропии в битах); б) — спектрограмма вейвлет-коэффициентов (по оси абсцисс — сдвиг базисного вейвлета во временном ряду, по оси ординат — масштаб преобразования заданный степенью двойки и измеряемый в количестве отсчетов временного ряда, яркость определяется значением вейвлет-коэффициента); в) — линии локальных экстремумов при заданном минимальном пороге (аналогично рисунку сверху: по оси абсцисс — параметр сдвига, по оси ординат — параметр масштаба)

В заключение рассмотрим результат сегментации всего файла. Таблица 1 содержит список выделенных участков на основании значимых вейвлет-коэффициентов, полученных при ДВП с использованием вейвлета Хаара. При этом использовались следующие параметры. Максимальный масштаб преобразования был задан как $a_4 = 2^4 = 16$; пороговый предел для определения значимых вейвлет-коэффициентов равен 0,5.

Таблица 1

Результат сегментации демонстрационного файла

Смещение участка от начала файла	Количество окон в участке	Средний уровень энтропии
0	8	1,4394
8	48	5,8268
56	17	7,0360
73	122	5,7472
195	15	4,2480
210	417	5,6986
627	15	1,9222

Сопоставляя данные таблицы с графиком упакованности файла (рис. 1), можно увидеть, насколько точно произошло выделение участков. Следует также обратить внимание, насколько гибким оказывается использование вейвлет-анализа. Варьируя максимальный масштаб преобразования, можно по-разному задавать чувствительность к изменениям в анализируемых данных. Так, при заданных выше параметрах алгоритма в демонстрационном примере не учитывается изменение энтропии на 540-м отсчете.

Заключение

Вейвлет-анализ, благодаря своим возможностям выявлять локальные особенности исходных данных, может быть применен для их сегментации. При этом сегментация будет зависеть от нескольких важных вещей. Во-первых, от формы базисного вейвлета, а именно от него будет зависеть способность выявлять те или иные характеристики. Во-вторых, от типа преобразования, которое будет влиять как на точность, так и на его быстроту. В-третьих, от определения значимых коэффициентов.

Алгоритмы, подобные данному, используются во многих прикладных областях. Например, в медицине для анализа электрофизиологических сигналов [11], в системах автоматического распознования речи [12], при обработке различных изображений [8] и т. п. [2; 10].

Использование описанного алгоритма в рамках информационной безопасности позволяет достаточно эффективно анализировать структурную энтропию файла, которая, в свою очередь, может быть использована как один из важных признаков в системах выявления и классификации вредоносного программного обеспечения.

СПИСОК ЛИТЕРАТУРЫ

1. *Астафьева Н. М.* Вейвлет-анализ: основы применения и примеры применения. УФН. 1996. Т. 166. № 11. С. 1145–1170.

2. *Бурнаев Е. В., Оленев Н. Н.* Меры близости на основе вейвлет коэффициентов для сравнения статистических и расчетных временных рядов: Межвуз. сб. научн. и научно-метод. трудов за 2005 г. Киров: Изд-во ВятГУ, 2006. Вып. 10. С. 41–51.

3. *Добеши И.* Десять лекций по вейвлетам. Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. 464 с.

4. *Витязев В. В.* Вейвлет анализ солнечной активности за 300 лет. Постерный доклад на Всероссийской астрономической конференции ВАК-2001. http://www.astro.spbu.ru/astro /win/personal /staff/vityazev.html

5. *Жарких А. А., Квашенко В. А.* Сравнение точности представления гауссовых вейвлетов различных порядков // Вестник Мурманского государственного технического университета. Мурманск, 2009. С. 218–223.

6. *Карпенко С. В.* Повышение скорости вычисления непрерывного вейвлет-преобразования. Медленные колебательные процессы в организме человека. Теоретические и прикладные аспекты нелинейной динамики в физиологии и медицине: Материалы IV всероссийского симпозиума и школы-семинара. Новокузнецк, Россия, 24–27 мая 2005. С. 78–80.

7. *Копыльцов А. В., Сорокин И. В.* Энтропийный анализ файлов для выявления и классификации вредоносного программного обеспечения. ИБРР-2009, октябрь 2009. С. 28–29.

8. *Панин С. В., Шакиров И. В., Сырямкин В. И., Светлаков А. А.* Применение вейвлет-анализа изображений поверхности для изучения процессов пластической деформации и разрушения на мезомасштабном уровне. Российская академия наук, Сибирское отделение. Автометрия. 2003. Т. 39. № 1. С. 37–53.

9. *Прангишвили И. В.* Энтропийные и другие системные закономерности // Вопросы управления сложными системами. М.: Наука, 2003. 432 с.

10. *Привалов М. В., Скобцов Ю. А., Кудряшов А. Г.* Сегментация компьютерных томограмм на основе вейвлет-преобразования // Вестник Херсонского национального технического университета. 2009. Вып. 1(34). С. 31–36.

11. *Синютин С. А.* Технология структурного анализа электрофизиологических сигналов с использованием вейвлет-преобразования // Программные продукты и системы. ЗАО НИИ «Центрпрограммсистем», 2002. 15 с.

12. *Федоров В. М., Юрков П. Ю.* Сегментация сигналов на основе дискретного вейвлет-преобразования: Материалы IX Международной научно-практической конференции «Информационная безопасность». Таганрог: Изд-во ТТИ ЮФУ, 2007. Ч. 1.С. 179–184.

13. *Яковлев А. Н.* Введение в вейвлет-преобразования. МОРФ. Новосибирск: Новосибирский государственный технический университет, 2003. 100 с.

REFERENCES

1. *Astaf'eva N. M.* Vejvlet-analiz: osnovy primenenija i primery primenenija. UFN. 1996. T. 166. № 11. S. 1145–1170.

2. *Burnaev E. V., Olenev N. N.* Mery blizosti na osnove vejvlet kojefficientov dlja sravnenija statisticheskih i raschetnyh vremennyh rjadov. Mezhvuz. sb. nauchn. i nauchno-metod. trudov za 2005 god. Kirov: Izd-vo VjatGU, 2006. Vyp. 10. C. 41–51.

3. *Dobeshi I.* Desjat' lekcij po vejvletam. Izhevsk, NIC «Reguljarnaja i haoticheskaja dinamika», 2001. 464 s.

4. *Vitjazev V. V.* Vejvlet-analiz solnechnoj aktivnosti za 300 let. Posternyj doklad na Vserossijskoj astronomicheskoj konferencii VAK-2001. http://www.astro.spbu.ru/astro /win/personal /staff/vityazev.html

5. *ZHarkih A. A., Kvashenko V. A.* Sravnenie tochnosti predstavlenija gaussovyh vejvletov razlichnyh porjadkov // Vestnik Murmanskogo gosudarstvennogo tehnicheskogo universiteta. Murmansk, 2009. S. 218–223.

6. *Karpenko S. V.* Povyshenie skorosti vychislenija nepreryvnogo vejvlet-preobrazovanija. Medlennye kolebatel'nye processy v organizme cheloveka. Teoreticheskie i prikladnye aspekty nelinejnoj dinamiki v fiziologii i medicine: Materialy IV vserossijskogo simpoziuma i shkoly-seminara. Novokuzneck, Rossija, 24–27 Maja 2005. S. 78–80.

7. *Kopyl'cov A. V., Sorokin I. V.* JEntropijnyj analiz fajlov dlja vyjavlenija i klassifikacii vredonosnogo programmnogo obespechenija. IBRR-2009, oktjabr' 2009. S. 28–29.

8. *Panin S. V., Shakirov I. V., Syrjamkin V. I., Svetlakov A. A.* Primenenie vejvlet-analiza izobrazhenij poverhnosti dlja izuchenija processov plasticheskoj deformacii i razrushenija na mezomasshtabnom urovne. Rossijskaja akademija nauk, Sibirskoe otdelenie. Avtometrija. 2003. T. 39. № 1. S. 37–53.

9. *Prangishvili I. V.* Jentropijnye i drugie sistemnye zakonomernosti. Voprosy upravlenija slozhnymi sistemami. M.: Nauka, 2003. 432 s.

10. *Privalov M. V., Skobcov Ju. A., Kudrjashov A. G.* Segmentacija komp'juternyh tomogramm na osnove vejvlet-preobrazovanija // Vestnik Hersonskogo nacional'nogo tehnicheskogo universiteta, 2009. Vyp. 1(34). S. 31–36.

11. *Sinjutin S. A.* Tehnologija strukturnogo analiza jelektrofiziologicheskih signalov s ispol'zovaniem vejvlet-preobrazovanija. «Programmnye produkty i sistemy». ZAO NII «Centrprogrammsistem», 2002. 15 s.

12. *Fedorov V. M., Jurkov P. JU.* Segmentacija signalov na osnove diskretnogo vejvlet-preobrazovanija: Materialy IX Mezhdunarodnoj nauchno-prakticheskoj konferencii «Informacionnaja bezopasnost'». Taganrog: Izd-vo TTI JUFU, 2007. CH. 1. S. 179–184.

13. *Jakovlev A. N.* Vvedenie v vejvlet-preobrazovanija. MORF. Novosibirsk: Novosibirskij gosudarstvennyj tehnicheskij universitet, 2003. 100 s.