Published in the Russian Federation
European Journal of Medicine
Has been issued since 2013.
ISSN: 2308-6513
E-ISSN: 2310-3434
Vol. 5, No. 3, pp. 155-165, 2014

DOI: 10.13187/ejm.2014.5.155
www.ejournal5.com

UDC 616

Clustering the Parameters of Rhythmographic Analysis of the Events of the Corporate Network Traffic of the *Cisco MARS* System

[1] Denis V. Lozhkarev
[2] Aleksandr V. Korobeinikov

[1] Izhevsk State Technical University named after MT Kalashnikov, Russian Federation
[2] Kamsky Institute of Humanities and Engineering Technology, Russian Federation
PhD, Associate Professor

Abstract

This article examines the clustering of the parameters of rhythmographic analysis of the events of the corporate network traffic of the *Cisco MARS* system. The author dwells upon classifying clustering methods and illustrates a conceptual clustering algorithm. The author infers that, firstly, the results of processing experimental data derived from network traffic logs substantiate the applicability of the methodology of rhythmographic analysis, which is accepted in cardiology, in the analysis of the rhythm of appearance of events in the *Cisco MARS* system; secondly, the results of clustering the fragments of the rhythm of events substantiate the effectiveness of the approach proposed; thirdly, the practical application of automatic detection of anomalies in network traffic events requires further research.

Keywords: rhythmographic analysis; parameter clustering; corporate network traffic; Cisco MARS system.

Введение

Число злоумышленников, которые наносят вред корпоративным сетям, возрастает с каждым годом. Это связано с развитием новых стандартов передачи данных в среде, которые с каждым годом повышают либо пропускную способность канала, либо надежность передачи путем избыточности. Кроме того число пользователей корпоративных сетей так же стремительно растет и с этим ростом, растет количество вредоносной активности.

Так как в данный момент, широко используемого варианта статического анализа (по шаблонам) аномалий не достаточно, в данной работе исследуется анализ ритмичности появления сетевых событий на основе методики ритмографического анализа, применяемого в кардиологии [1]. Такой подход позволит обнаруживать резкое изменение характера ритма возникновения сетевых событий, связанное с различными сетевыми аномалиями. Кластеризация параметров ритмографического анализа позволит автоматически разделять различные фрагменты журналов отдельных сетевых событий на группы похожих по характеру ритмичности событий.

В работе [2] исследуются устойчивые цепочки (секвенции) сетевых событий корпоративного сетевого трафика системы *Cisco MARS*. Данная работа является продолжением исследований по анализу событий сетевого корпоративного трафика системы *Cisco MARS*.

Материалы

Данные для анализа. Исходными данными для анализа является база данных в формате *dbf*, которая была построена на основе журналов регистрации событий системы управления информационной безопасности (*Security Information Management*) *Cisco Security MARS* (*Monitoring*, *Analysis and Response System*). *Cisco Security MARS* – это система безопасности включающая в себя мониторинг, анализ и реагирование на инциденты информационной безопасности [3, 4]. Источники событий – это всевозможные сетевые устройства и оборудование, такие как маршрутизаторы, коммутаторы, шлюзы безопасности, *VPN* серверы, сетевые и узловые (*host based*) системы обнаружения и предотвращения вторжений, системы *AAA* (*Authentication*, *Authorization*, *Accounting*), антивирусные системы, веб-серверы, серверы баз данных, журналы системных событий (*syslog*) на серверах под управлением операционных систем *Unix* и *Windows*, а так же журналы работы каталога *Microsoft Active Directory* [4].

Каждая запись журнала имела следующие поля: 1) *ID* – номер события в журнале; 2) *ODATE* – дата и время срабатывания события; 3) *OSECOND* – время в секундах со старта системы для сработанного события.

Всего уникальных событий – 484. Анализируемые журналы событий совпадают с данными, которые анализировались в работе [2].

Обоснование применения методика анализа. Одной из основных методик анализа электрокардиограммы (ЭКГ) является анализ вариабельности сердечного ритма (ВСР, другие названия: кардиоинтервалография, КИГ, ритмография) – это исследование изменчивости ритма следования кардиоциклов. Длительность последовательных кардиоциклов нормального ритма меняется с течением времени. Величину и скорость этих изменений определяют значения показателей ВСР. ВСР отражает работу сердечно-сосудистой системы и работу механизмов регуляции организма человека.

При исследовании собранных журналов работы было отмечено, что каждое событие имеет уникальную среди всех типов событий картину ритмичности появления этого события. При нанесении изменения длительности между очередными событиями одного из типов на временную шкалу, получим график ритмограммы (рис. 1). В данной работе анализ ВСР взят за основу, но применяется к исследованию ритма появления сетевых событий. Применение для анализа событий методики анализа принятой для медицинской техники, связан с высокой развитостью аппарата ритмографического анализа в кардиологии. Применительно к анализу сетевого трафика под ВСР в дальнейшем будем понимать вариабельность ритма событий.

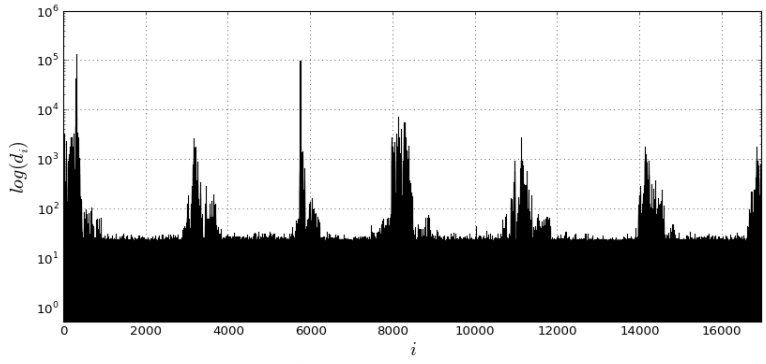


Рис. 1. Ритмограмма (событие – 19, период – все время)

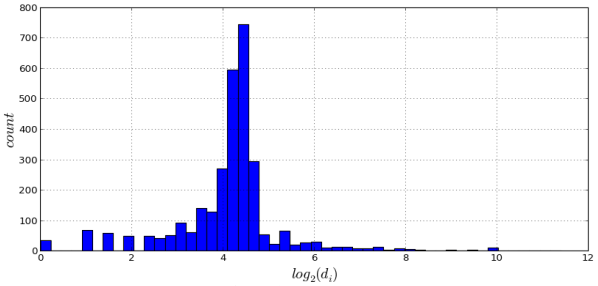


Рис. 2. Гистограмма (событие – 19, период – 1 день)

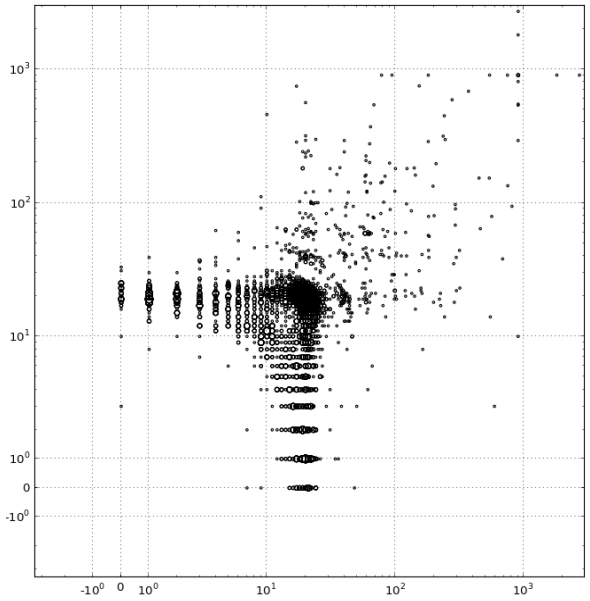


Рис. 3. Скатерограмма (событие – 19, период – 1 день)

Для анализа вариабельности сердечного ритма человека применяют разделение всей ритмограммы на короткие (5 минут) и длинные временные периоды анализа (24 часа), которые позволяет учитывать суточные колебания биологических ритмов человека и менее подвержены влиянию случайных факторов [1].

Для исследования сетевых событий в данной работе было выбрано 5 временных периодов: 1 час, 6 часов, 12 часов, 1 сутки, 1 неделя и все время (22 дня).

Построение диаграмм. При анализе ВСР используют диаграммы: ритмограмму (рис. 1), гистограмму (рис. 2) и скатерограмму (рис. 3) [1].

Ритмограмма (рис. 1) – график вариационного ряда интервалов, у которого по оси y отложены значения интервалов d_i, а по оси x порядковые номера интервала i или время появления интервала t_i, что более предпочтительно, потому что в этом случае выдерживается временной масштаб графика. Ритмограмма является основным графиком и на её основе строятся остальные диаграммы. Значение интервала приведено в логарифмическом масштабе: $log_{10}(d_i)$.

Гистограмма (рис. 2) – график сгруппированных значений интервалов, где по одной оси откладывается их длительность, по другой – количество или процент от общего числа. Значение длительности между события приведено в логарифмическом масштабе: $log_2(d_i)$.

Скатерограмма (*Lorenz plot*) (рис. 3) – это графическое отображение соответствия (корреляции) соседних интервалов на двумерной координатной плоскости, по осям которой отложены временные значения интервалов d_{i-1} и d_i. Значения длительности приведены в логарифмическом масштабе: $log_{10}(d_i)$ и $log_{10}(d_{i-1})$.

Отметим что, скатерограмма приведена в модифицированном виде. Радиус окружности пропорционален количеству соседних интервалов попавших в точку диаграммы с данными координатами.

Результаты

Вычисление параметров. По построенным диаграммам вычисляется ряд числовых показателей.

Анализ гистограммы относят к геометрическим методам. В кардиологии выполняется расчет параметров по 2 методикам: по отечественному стандарту (по Баевскому) и по Европейскому стандарту.

По отечественному стандарту вычисляются следующие параметры [1]:

1) M_0 – мода, наиболее частое значение среди интервалов; 2) AM_0 – амплитуда моды, доля интервалов, соответствующая моде; 3) ΔX – вариационный размах, разность между длительностью наибольшего и наименьшего интервала; 4) ИВР – индекс вегетативного равновесия; 5) ВПР – вегетативный показатель ритма; 6) ПАПР – показатель адекватности процессов регуляции; 7) ИН – индекс напряжения регуляторных систем.

$$ИВР = \frac{AM_0}{\Delta X};\quad ВПР = \frac{1}{M_0 \Delta X};\quad ПАПР = \frac{AM_0}{M_0};\quad ИН = \frac{AM_0}{2M_0 \Delta X}. \quad (1)$$

Для анализа по европейскому стандарту вычисляется ряд параметров, формулы для вычисления которых приведены в работе [1]. Перечисленные параметры используются для коротких и суточных записей ритма:

1) *SDNN* – стандартное отклонение нормальных интервалов; 2) *SDANN* – стандартное отклонение средних значений интервалов, вычисленных по коротким периодам анализа за все время анализа; 3) *SDNN index* – среднее значение стандартных отклонений нормальных интервалов, вычисленных по коротким периодам анализа за все время анализа; 4) *RMSSD* – квадратный корень из средней суммы квадратов разностей между соседними нормальными интервалами;

Для анализа суточного ритма дополнительно вычисляют [1]:

7) *SDSD* – стандартное отклонение разницы между соседними нормальными интервалами; 8) *M* – среднее значение нормальных интервалов; 9) *Min* – минимальное значение длительности интервала; 10) *Max* – максимальное значение длительности интервала; 11) *M Dif.* – средняя абсолютная разница соседних интервалов; 12) *CVr* – коэффициент вариабельности; 13) *N* – общее количество анализируемых интервалов; 14) *D* – дисперсия длительности интервалов.

Кластеризация показателей ВСР. При интерпретации показателей ВСР полученные значения параметров следует разбить на группы похожих по характеру ритма событий. В дальнейшем эксперт обозначит каждую такую группу характера ритмичности события как норму или аномалию. Это позволит в дальнейшем анализировать текущее ритмографические параметры событий и автоматически относить характер ритма к норме или аномалии.

Для обучения системы анализа ВСР в части автоматического определения характера ритма целесообразно использовать методы кластерного анализа данных. Кластеризация [5] – это задача машинного обучения, в которой требуется разбить заданную выборку объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Входные данные алгоритма кластеризации – это обучающая выборка, состоящая из m образцов: $A = \{a_1, ..., a_m\}$. Для группировки образцов используется некая функция расстояния между ними $\rho(a_i, a_j)$. Необходимо произвести разбиение исходной выборки на непересекающиеся подмножества (кластеры), с таким условием, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $a_i \in A$ приписывается метка (номер) кластера b_i.

Задача кластеризации – это задача раздела искусственного интеллекта, который изучает методы построения систем, способных обучаться. Эта задача относится к классу задач: обучение без учителя. Обучение без учителя отличается от обучения с учителем (классификации) тем, что метки кластеров b_i исходных образцов a_i изначально не заданы. Задача классификации решается на этапе применения результатов кластеризации.

Для решения задачи кластеризации (*clustering problem*) необходим набор неклассифицированных объектов и средства измерения подобия объектов. Целью кластеризации является организация объектов в классы, удовлетворяющие некоторому стандарту качества, например на основе максимального сходства объектов каждого класса.

Числовая таксономия (*numeric taxonomy*) – один из первых подходов к решению задач кластеризации. Числовые методы основываются на представлении объектов с помощью набора свойств, каждое из которых может принимать некоторое числовое значение. При наличии корректной метрики подобия каждый объект (вектор из n значений признаков) можно рассматривать как точку в n-мерном пространстве. Мерой сходства двух объектов можно считать расстояние между ними в этом пространстве. На рис. 4 представлена классификация известных методов кластеризации [6].

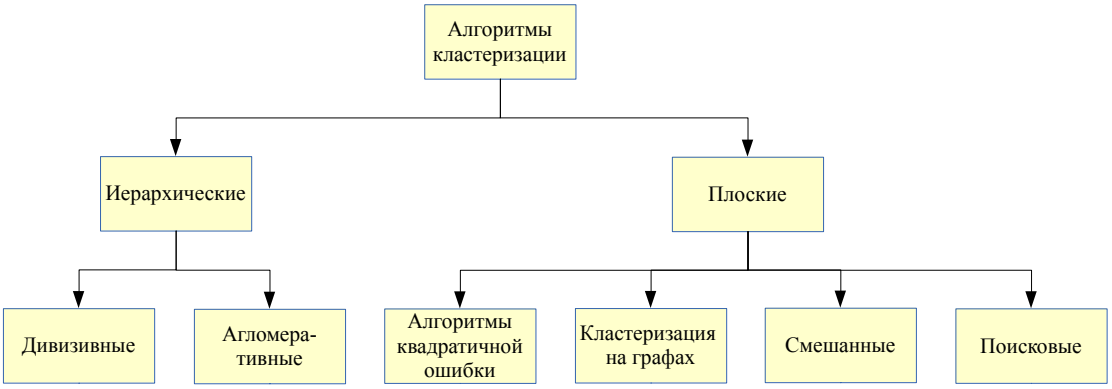


Рис. 4. Классификация методов кластеризации

Алгоритм концептуальной кластеризации. В отличие от традиционной кластеризации, которая обнаруживает группы схожих объектов на основе меры сходства между ними, концептуальная кластеризация определяет кластеры как группы объектов, относящейся к одному классу или концепту – определённому набору пар атрибут-значение.

В алгоритме *Cobweb* [7] реализован инкрементальный алгоритм обучения, не требующий получения входных обучающих примеров единовременно до начала обучения. Решена проблема определения необходимого числа кластеров для разбиения входных данных – для определения количества кластеров, глубины иерархии и принадлежности категории новых экземпляров используется глобальная метрика качества. При предъявлении нового экземпляра алгоритм *Cobweb* оценивает качество отнесения этого примера к существующей категории и модификации иерархии категорий в соответствии с новым представителем. Критерием оценки качества классификации является полезность категории (*category utility*). Критерий полезности категории был определён при исследовании человеческой категоризации. Он учитывает влияние категорий базового уровня и другие аспекты структуры человеческих категорий.

Критерий полезности категории максимизирует вероятность того, что два объекта, отнесённые к одной категории, имеют одинаковые значения свойств и значения свойств для объектов из различных категорий отличаются. Полезность категории определяется формулой:

$$CU = \sum_{k=1}^{n} \sum_{j} \sum_{i} P(A_j = v_{ij})P(C_k \mid A_j = v_{ij})P(A_j = v_{ij} \mid C_k). \quad (2)$$

Значения суммируются по всем категориям C_k, всем свойствам A_j и всем значениям свойств v_{ij}. $P(A_j = v_{ij} \mid C_k)$ – предсказуемость, то есть вероятность того, что объект, для которого свойство A_j – принимает значение v_{ij}, относится к категории C_k. Величина $P(C_k \mid A_j = v_{ij})$ – предиктивность, то есть вероятность того, что для объектов из категории C_k свойство A_j принимает значение v_{ij}. Значение $P(A_j = v_{ij})$ – это весовой коэффициент, усиливающий влияние наиболее распространенных свойств. Благодаря совместному учету этих значений высокая полезность категории означает высокую вероятность того, что объекты из одной категории обладают одинаковыми свойствами, и низкую вероятность наличия этих свойств у объектов из других категорий.

Описанный выше вариант алгоритма *Cobweb* имеет недостаток, заключающийся в возможности работы только с качественными показателями. Данный недостаток устраняется модификацией алгоритма для работы с количественными показателями [8].

Результаты экспериментов. Для оценки применимости предлагаемого в данной

работе подхода были использованы описанные выше экспериментальные данные журналов. Для проведения экспериментов было разработано программное обеспечение на языке *Pyphon*.

Были построены диаграммы и вычислены статистические параметры для каждого фрагмента для каждого временного периода всех событий. В результате каждый временной промежуток (фрагмент) события характеризуется набором ритмографических параметров.

Кластеризация образцов ритма, представленных параметрами выполнялась на основе модифицированного алгоритма *Cobweb* для количественно заданных показателей [8]. В результате кластеризации было получено дерево кластеров для различных временных периодов фрагментации ритмограмм. Пример дерева кластеров представлен (рис. 5).

Рассмотрим дерево кластеров для события 19 с периодом 12 часов (удалены кластеры, содержащие один образец):

Кластер 0 "037a" (22 образцов)
Кластер 2 "6b0b" (8 образцов)
    Кластер 2.1 "b742" (3 образца)
    Кластер 2.4 "1eb0" (3 образца)
Кластер 3 " 82be" (13 образца)
    Кластер 3.1 "dc45" (3 образца)
    Кластер 3.2 "ede2" (3 образца)

На рис. 6 представлены некоторые гистограммы образцов (фрагментов) ритма событий, а в табл. 1 представлены ритмографические параметры образцов, входящих в различные кластеры.

Параметры образцов вычислялись только по графикам гистограммы и ритмограммы, однако, скатерограммы образцов в разных кластерах характерно отличаются и совпадают в одном кластере.

Таблица 1.

Пример параметров (событие – 19, период – 12 часов)

	14	**12**	**1**	**11**	**17**	**10**	**8**	**6**
Образец	**1 "63*a*3"**	**2 "379*c*"**	**3 "*a*7b8"**	**4 "2*b*6*b*"**	**5 "76*c*3"**	**6 "723*f*"**	**7 "2704"**	**8 "*b*672"**
Кластер	**2.1 "*b*742"**	**2.1 "*b*742"**	**2.2 "1*eb*0"**	**2.2 "1*eb*0"**	**3.1 "*dc*45"**	**3.1 "*dc*45"**	**3.2 "*ede*2"**	**3.2 "*ede*2"**
M_0	9.0	9.0	5.0	4.0	4.0	4.0	4.0	4.0
AM_0	92	115	208	192	6100	7275	11820	12582
ΔX	8.3	7.4	11.7	11.7	11.4	11.4	9.9	10.6
ИВР	11	16	18	16	535	638	1193	1191
ВПР	0.013	0.015	0.017	0.021	0.022	0.022	0.025	0.024
ПАПР	10	13	42	48	1525	1819	2955	3145
ИН	0.6	0.9	1.8	2.1	67	80	149	149
SDNN	2.3	1.8	2.4	2.7	1.3	1.2	1.1	1.1
SDANN	3555	3555	3555	3555	3555	3555	3555	3555
SDNN index	1508	1508	1508	1508	1508	1508	1508	1508
RMSSD	2.4	1.9	2.9	2.6	1.6	1.5	1.6	1.6
SDSD	1.4	1.2	1.8	1.7	1.1	1.1	1.1	1.1
M Dif.	2.0	1.4	2.3	2.0	1.1	1.0	1.1	1.1
Max	12.4	12.8	11.7	11.7	11.4	11.4	9.9	10.6
Min	4.1	5.4	0.0	0.0	0.0	0.0	0.0	0.0
M	8.7	9.0	6.5	6.7	4.3	4.0	3.9	3.9
D	5.3	3.1	5.9	7.4	1.8	1.4	1.2	1.2
CVr	27	20	37	41	31	30	28	28
N	41	44	145	114	1233	1362	1906	2132

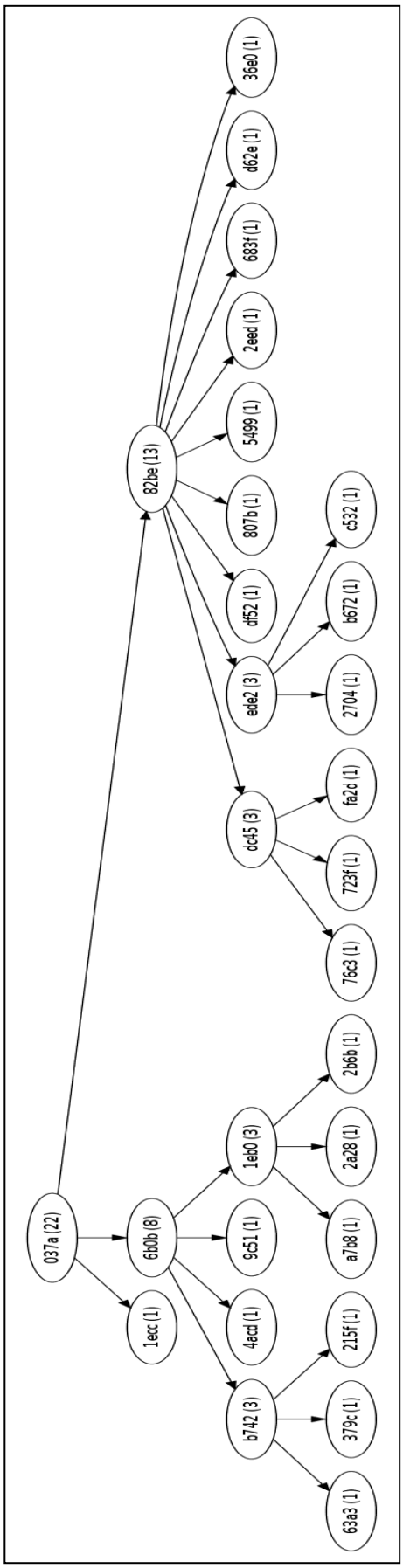
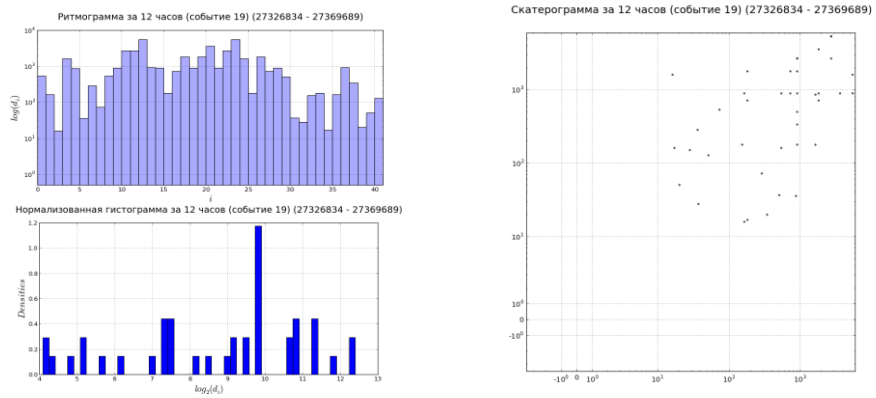
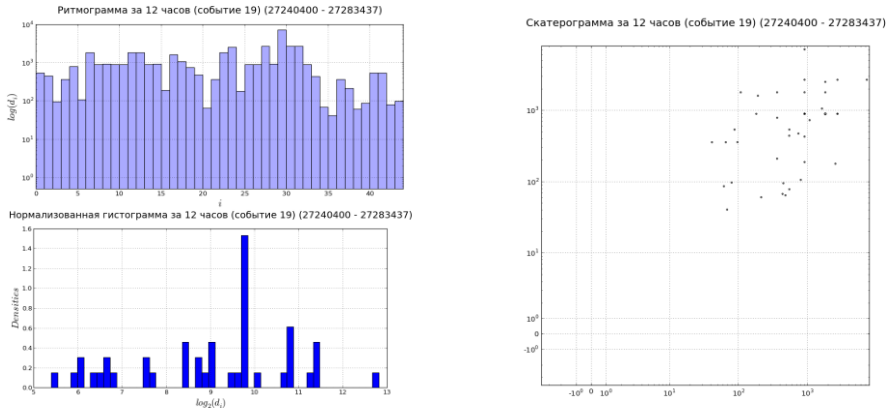


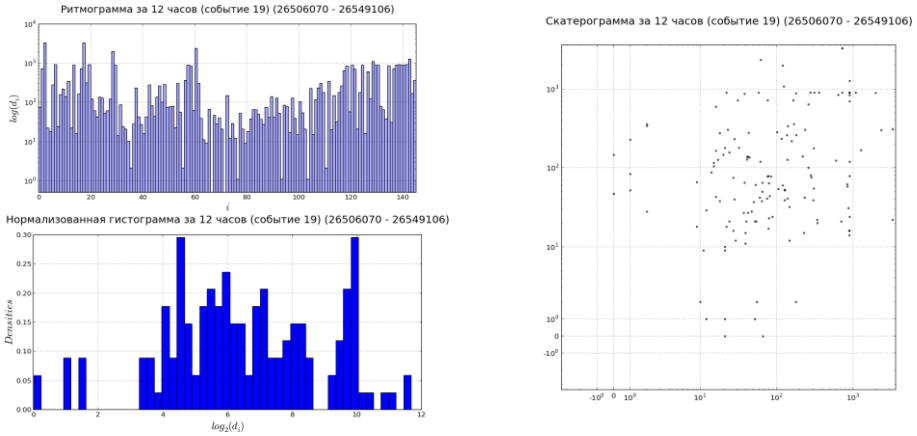
Рис. 5. Дерево кластеров (событие – 19, период – 12 часов)



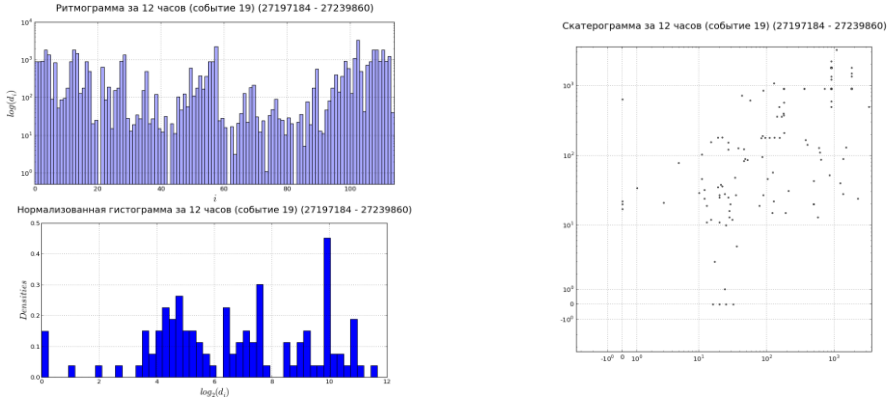
а) образец "63*a*3" кластер 2.1 "*b*742";



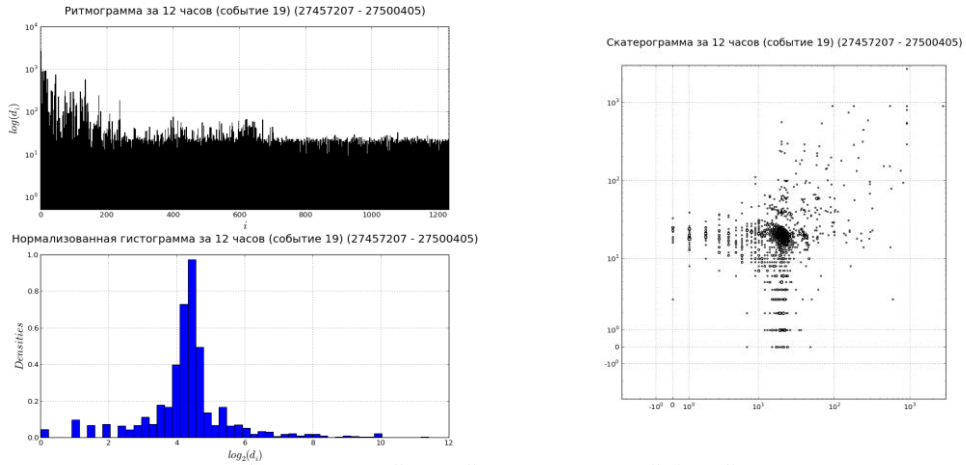
б) образец "379*c*", кластер 2.1 "*b*742";



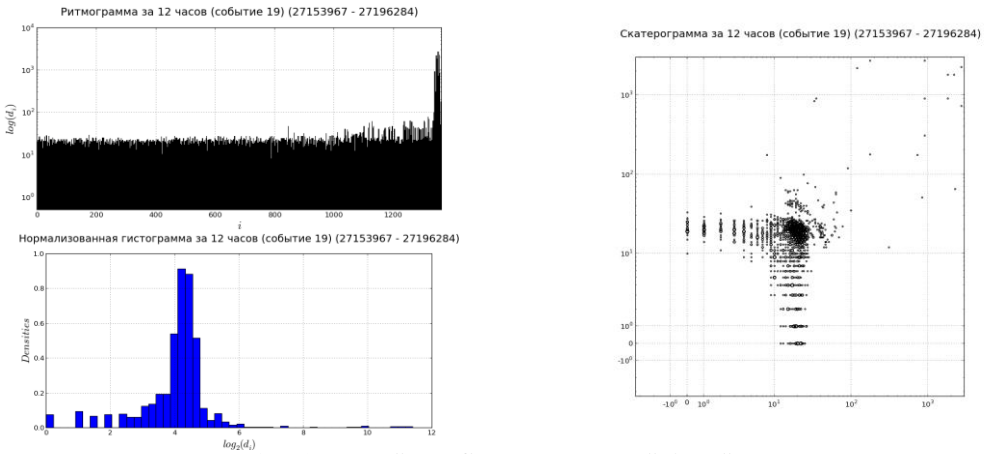
в) образец "*a*7b8", кластер 2.2 "1*eb*0";



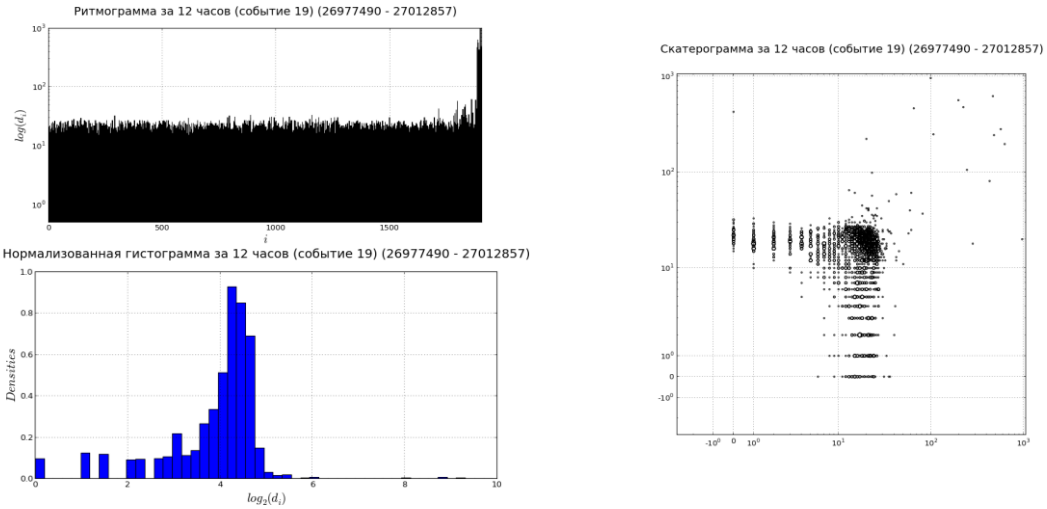
г) образец "2*b*6*b*", кластер 2.2 "1*eb*0";



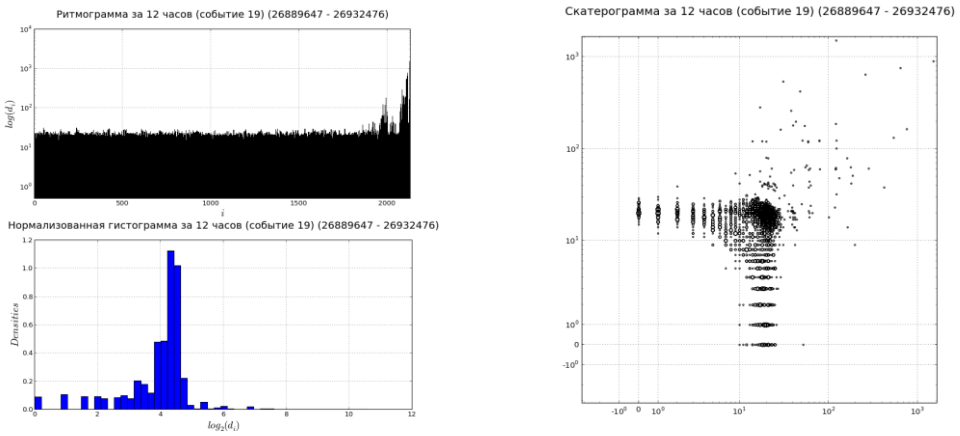
д) образец "76*c*3", кластер 3.1 "*dc*45";



е) образец "723*f*", кластер 3.1 "*dc*45";



ё) образец "2704", кластер 3.2 "*ede*2";



з) образец "*b*672", кластер 3.2 "*ede*2";

Рис. 6. Результат кластеризации гистограмм (событие – 19, период – 12 часов)

Выводы:

1. Результаты обработки экспериментальных данных журналов сетевого трафика подтверждают применимость методики ритмографического анализа принятого в кардиологии для анализа ритма появления событий в системе *Cisco MARS*;
2. Результаты кластеризации фрагментов ритма событий подтверждают эффективность предложенного подхода;
3. Практическое применение автоматического обнаружения аномалий событий сетевого трафика требует дальнейших исследований.

Примечания:

1. Коробейников А.В. Алгоритмы и комплексы программ мониторно-компьютерных систем для анализа морфологии и ритма электрокардиограмм: диссертация канд. техн. наук: 05.13.18, 05.11.16. Ижевск, 2004. 170 с.

2. Коробейников А.В., Конин А.В., Менлитдинов А.С. Стохастический подход к секвенциальному анализу событий корпоративного сетевого трафика системы *Cisco MARS* // Вестник КИГИТ. 2012. № 7 (25). С. 060-070.

3. *Gary Halleen, Greg Kellogg* Security Monitoring with Cisco Security MARS – Cisco Press 800 East 96th Street Indianapolis, IN 46240 USA, 2007.

4. *John Jarocki* Configuring and Tuning Cisco CS-MARS – SANS Institute InfoSec Reading Room, 2007.

5. Кошелева В.А. Концептуальная кластеризация как метод извлечения знаний из баз данных // IV международная научная конференция студентов, аспирантов и молодых ученых «Компьютерный мониторинг и информационные технологии». 13-14 мая 2008 г. – URL: www.ami.nstu.ru/~vms/lecture/data_mining/kurs_klaster.htm (дата обращения 01.05.2014).

6. Филиппова Т.П., Коробейников А.В. Разработка методов кластеризации и классификации на основе алгоритма *Cobweb* // Информационные технологии в науке, промышленности и образовании: сб. тр. региональной научно-технической очно-заочной конференции. Ижевск: ИжГТУ, 2013. С. 110-115.

7. Люгер Д. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. 4-е изд. М.: Вильямс, 2003. 864 с.

8. Коробейников А.В., Исламгалиев И.И. Модификация алгоритма концептуальной кластеризации *Cobweb* для количественных данных с использованием нечеткой функции принадлежности // Приволжский научный вестник. Ижевск: Самохвалов Антон Витальевич, 2013. № 3. С. 9-14.

References:

1. Korobeinikov A. V. Algoritmy i kompleksy programm monitorno-komp'yuternykh sistem dlya analiza morfologii i ritma elektrokardiogramm: dissertatsiya kand. tekhn. nauk: 05.13.18, 05.11.16. Izhevsk, 2004. 170 s.

2. Korobeinikov A.V., Konin A.V., Menlitdinov A.S. Stokhasticheskii podkhod k sekventsial'nomu analizu sobytii korporativnogo setevogo trafika sistemy Cisco MARS // Vestnik KIGIT. 2012. № 7 (25). S. 060-070.

3. Gary Halleen, Greg Kellogg Security Monitoring with Cisco Security MARS – Cisco Press 800 East 96th Street Indianapolis, IN 46240 USA, 2007.

4. John Jarocki Configuring and Tuning Cisco CS-MARS – SANS Institute InfoSec Reading Room, 2007.

5. Kosheleva V. A. Kontseptual'naya klasterizatsiya kak metod izvlecheniya znanii iz baz dannykh // IV mezhdunarodnaya nauchnaya konferentsiya studentov, aspirantov i molodykh uchenykh «Komp'yuternyi monitoring i informatsionnye tekhnologii». 13-14 maya 2008 g. – URL: www.ami.nstu.ru/~vms/lecture/data_mining/kurs_klaster.htm (data obrashcheniya 01.05.2014).

6. Filippova T. P., Korobeinikov A. V. Razrabotka metodov klasterizatsii i klassifikatsii na osnove algoritma Cobweb // Informatsionnye tekhnologii v nauke, promyshlennosti i obrazovanii: sb. tr. regional'noi nauchno-tekhnicheskoi ochno-zaochnoi konferentsii. Izhevsk: IzhGTU, 2013. S. 110-115.

7. Lyuger D. F. Iskusstvennyi intellekt: strategii i metody resheniya slozhnykh problem. 4-e izd. M.: Vil'yams, 2003. 864 s.

8. Korobeinikov A. V., Islamgaliev I. I. Modifikatsiya algoritma kontseptual'noi klasterizatsii Cobweb dlya kolichestvennykh dannykh s ispol'zovaniem nechetkoi funktsii prinadlezhnosti // Privolzhskii nauchnyi vestnik. Izhevsk: Samokhvalov Anton Vital'evich, 2013. № 3. S. 9-14.

УДК 616

Кластеризация параметров ритмографического анализа событий корпоративного сетевого трафика системы *CISCO MARS*

[1] Денис Вячеславович Ложкарев
[2] Александр Васильевич Коробейников

[1] Ижевский государственный технический университет имени М.Т. Калашникова, Российская Федерация
[2] Камский институт гуманитарных и инженерных технологий, Российская Федерация
Кандидат технических наук, доцент

Аннотация. В статье рассматривается кластеризация параметров ритмографического анализа событий корпоративного сетевого трафика системы *CISCO MARS*. Уделено внимание классификации методов кластеризации, а также показан алгоритм концептуальной кластеризации. В выводах отмечается, что, во-первых, результаты обработки экспериментальных данных журналов сетевого трафика подтверждают применимость методики ритмографического анализа принятого в кардиологии для анализа ритма появления событий в системе *Cisco MARS*; во-вторых, результаты кластеризации фрагментов ритма событий подтверждают эффективность предложенного подхода; в-третьих, практическое применение автоматического обнаружения аномалий событий сетевого трафика требует дальнейших исследований.

Ключевые слова: ритмографический анализ; кластеризация параметров; корпоративный сетевой трафик; система CISCO MARS.