УДК 004.491.2

КЛАСТЕРИЗАЦИЯ ЗЛОУМЫШЛЕННЫХ ДЕЙСТВИЙ НА ОСНОВЕ ИНФОРМАЦИИ, ПОЛУЧЕННОЙ ОТ СЕТИ ЛОВУШЕК

А. С. Каштальян[1], О. С. Савенко[2]

[1] *К. т. н., доцент кафедры физики и электротехники Хмельницкого национального университета, Хмельницкий, Украина*
[2] *Начальник отдела информационно-технического обеспечения экономических служб, старший преподаватель кафедры компьютерной инженерии и системного программирования Хмельницкого национального университета, Хмельницкий, Украина*

Реферат

Разработан метод агломеративной кластеризации злоумышленников на основе их динамических характеристик, полученных сетью ловушек, позволяющий определять атаки известных типов и выделять новые типы атак, получены критерии остановки процесса кластеризации.

Ключевые слова: сеть ловушек, кластеризация временных рядов, характеристики атак, информационный критерий

CLUSTERING MALICIOUS ACTIONS BASED ON INFORMATION RECEIVED FROM A HONEYNET

A. S. Kashtalian, O. S. Savenko

Abstract

The method of agglomerative clustering of intruders on the basis of their activity time series with use of information criteria for determination of distance between time series as measure of similarity is proposed, as well as criteria of stopping the clustering process, which allow to obtain the optimal ratio of clusters purity and completeness, and accordingly determining the number of clusters.

Keywords: honeynet, time series clustering, attack characteristics, information criterion.

Введение

Компьютерные сети, подключенные к сети интернет, уязвимы для разного рода угроз. Для обнаружения злоумышленных действий и защиты от них используются разнообразные системы, среди которых межсетевые экраны, антивирусные программы, системы обнаружения атак, системы контроля целостности и т. п. Характерными особенностями этих систем является их периодическое или кратковременное использование для решения определенной проблемы, или постоянное использование со статическими характеристиками. В результате большое количество используемых в современных системах методов анализа направлены на обнаружение известных вторжений. Обнаружение модифицированных известных вторжений и вторжений новых типов часто остается неэффективным. В то же время около 20 % злоумышленных доменов являются новыми и используются в среднем через 1 неделю после регистрации [1]. Поэтому поиск решений, которые позволят более эффективно защитить компьютерные сети, всегда актуален. Важным требованием к таким решениям должна быть возможность обнаружения любых типов вторжений, включая новые и распределенные во времени. Одним из перспективных решений являются ловушки и сети ловушек.

Ловушки в компьютерных сетях выполняют функции сбора и анализа информации относительно злоумышленных действий [2]. Ловушка представляет собой средство, которое позволяет «привлечь» злоумышленника и собрать про него информацию до того, как злоумышленник получит доступ к рабочей компьютерной сети.

Обзор известных работ и постановка задачи

Современные модели ловушек позволяют создать систему ложных объектов атак, интегрированную в общую систему безопасности компьютерных сетей. Система ловушек представляет собой сеть с собственной архитектурой и собственной системой сервисов и фактически встроена в сеть рабочих сервисов, что значительно повышает контролируемость и защищенность системы в целом. Элементами работы сети ловушек являются контроль данных, привлечение данных и сбор данных.

Контроль данных системой ловушек предназначен для уменьшения рисков осуществления сервисов, размещенных в сети. Основной целью контроля действий злоумышленника является быстрое и эффективное реагирование на них. Кроме того, злоумышленнику не должно быть известно, что его действия контролируются. Для уменьшения рисков система ловушек должна соответствовать определенным требованиям: возможность поддерживать входящие и исходящие соединения, возможность контролировать любую несанкционированную деятельность, конфигурация осуществления контроля, наличие методов предупреждения о злоумышленной деятельности, возможность удаленного администрирования контроля данных.

Для эффективной работы сети ловушек при привлечении данных от злоумышленников должны выполняться такие требования: данные, привлеченные ловушкой, не хранятся локально с целью избегания доступа к ним злоумышленников; любое влияние других данных не должно влиять на процесс привлечения данных ловушкой; к активностям, которые отслеживаются, относятся активности сети, системы, приложений и пользователей, привлеченные данные должны автоматически сохраняться для дальнейшего анализа, информация о привлеченных данных должна быть доступна для каждой ловушки в сети, должна обеспечиваться целостность данных, привлеченных сетью ловушек.

doi.org/10.36773/1818-1212-2020-123-5-66-71

Сбор данных сетью ловушек предназначен для дальнейшего их анализа и интеграции с данными о злоумышленных действиях, полученных из других источников. Эта информация используется для текущего и дальнейшего предотвращения атак. Для сбора данных к сети ловушек выдвигается ряд требований: каждая ловушка имеет свой уникальный идентификатор для взаимодействия с сетью; возможность анонимизации данных относительно ловушек в сети; обмен данных должен выполняться с обеспечением конфиденциальности и целостности данных; стандартизация и синхронизация обмена данными, привлеченными сетью ловушек.

Для обнаружения атаки и определения ее типа необходимо определить характерные признаки атаки. Ловушка должна обеспечивать порты и сервисы, которые подвергаются атакам, и организацию сбора, хранения и обработки данных сетевого трафика. Информация, которую собирает ловушка, позволяет обнаружить существующие и новые типы атак и определить их типовые характеристики.

Реализация атаки включает несколько этапов, в том числе сбор информации, непосредственно реализацию атаки, возможное уничтожение следов атаки. Существующие системы обнаружения вторжений во многом ориентированы на обнаружение атак уже в процессе их реализации. Более эффективным является обнаружение атак на первом этапе сбора информации, так как позволяет минимизировать потери от возможных злоумышленных действий. Такое раннее обнаружение могут обеспечить ловушки и их сети. Сбор информации злоумышленником предусматривает: изучение окружения, идентификацию топологии сети, идентификацию узлов, идентификацию сервисов или сканирование портов, идентификацию операционной системы, определение уязвимостей.

Трафик, который фиксируется на сети ловушек, отображает деятельность только злоумышленников. Злоумышленники совершают атаки разных типов с разной частотой, поэтому несут опасность разной степени. Несмотря на существенное количество атак и их источников, действия злоумышленников имеют схожие цели и способы осуществления атак. Поэтому важной задачей является нахождение подобных характеристик злоумышленников и объединение их в группы для использования этой информации в системах обнаружения и предупреждения атак.

В большом количестве систем обнаружения вторжений используются традиционные и современные методы кластеризации на основе злоумышленных действий. В работе [3] используется традиционный метод кластеризации k-средних. В современных системах используются нейросетевые методы [4] и генетические алгоритмы [5]. Также используются подходы к кластеризации с использованием метода опорных векторов [6] и на основе деревьев решений [7]. Во многих работах характеристики атак рассматриваются как статические и вопрос количества кластеров остается вне рассмотрения. В данной работе предлагается использовать для кластеризации динамические характеристики, а также рассматривается критерий определения числа кластеров.

Характеристики атак

В зависимости от характеристик среди атак можно выделить определенные типы, к которым относятся удаленное и локальное проникновение, удаленный и локальный отказ в обслуживании (DoS). Рассмотрим более детально характеристики некоторых типов атак.

Значительный объем почтового трафика является характерным признаком атаки «почтовая бомба». Этот почтовый трафик может быть создан как одно сообщение большого объема или значительное количество сообщений небольшого объема. Целью такой атаки может быть один электронный почтовый ящик или разные электронные почтовые ящики, расположенные на одном сервере. Источник атаки может быть одним, или атака может быть распределенной. Для выявления атаки «почтовая бомба» ловушке необходимо собрать информацию: источник сообщения, объем сообщения. На основе этих данных необходимо провести анализ активности во времени на SMTP сервисе ловушки, определить пороговые значения объема сообщений и скорости их получения.

Атака «neptune» может рассматриваться как полуоткрытая TCP-SYN атака. Для успешной реализации этого типа атаки используется особенность предварительной установки соединения в TCP-протоколе. Злоумышленник присылает большое количество SYN-пакетов, запросов на установку соединения, на TCP-сервер. Сервер в ответ отправляет SYN/ACK пакет и ждет ACK-пакет от клиента, который злоумышленником не отправляется, таким образом соединение остается полуоткрытым. Из-за ограниченного числа соединений на TCP-сервере происходит переполнение и блокировка работы сервера. Для выявления атаки «neptune» ловушка собирает данные об источнике запросов и количестве полуоткрытых запросов. Из этих запросов формируется временной ряд активности злоумышленника или злоумышленников, по которому можно обнаружить атаку.

Атака «portsweep» предполагает сканирование злоумышленников значительного количества портов целевого сервера для определения сервисов, которые на нем работают. На ловушке собирается информации об источнике атаки и количестве обращений к портам. Количество обращений к портам представляет собой временной ряд активности злоумышленника.

Таким образом, временной ряд активности злоумышленника относительно ловушки несет важную информацию о наличии атаки и ее типе, а также об уровне опасности самого злоумышленника для компьютерной сети.

Временной ряд количества запросов от *i*-го источника к *j*-й ловушке за единицу времени:

$$X^{ij} = \left\{ x_1^{ij}, x_2^{ij}, x_3^{ij}, \ldots, x_k^{ij} \right\},$$

где x_k^{ij} – количество запросов от *i*-го источника к *j*-й ловушке за 1 сек.; k – продолжительность исследуемого временного окна; $i=1,2,3,\ldots m$, m – количество источников атак; $j=1,2,3,\ldots n$, n – количество ловушек.

Информационные критерии, используемые для определения расстояния между временными рядами

При кластеризации временных рядов, которые отображают активность злоумышленников, необходимо учитывать, что признаки изменяются во времени. Для кластеризации временных рядов эффективными являются методы, которые основаны на использовании моделей и предусматривают определение параметров модели. Параметры моделей представляют собой пространство признаков, в котором выполняется кластеризация. Для определения расстояний между параметрами моделей временных рядов используется ряд информационных критериев, среди которых байесовский информационный критерий, обобщенный байесовский информационный критерий, информационный критерий Байеса-Айкаике.

doi.org/10.36773/1818-1212-2020-123-5-66-71

Байесовский информационный критерий (BIC) представляет собой оценку моделей, определенный через апостериорную вероятность [8]. Пусть M_1, M_2, …, M_r – r-моделей-кандидатов, предположим, что каждая модель M_i описывается параметрическим распределением $f_i\left(x\middle|\boldsymbol{\theta}_i\right)\left(\boldsymbol{\theta}_i \in \theta_i \subset R^{k_i}\right)$ и априорное распределение $\pi_i\left(\boldsymbol{\theta}_i\right)$ k_i-размерного параметрического вектора $\boldsymbol{\theta}_i$. Для заданных n наблюдений $\mathbf{x}_n = \{x_1,\ldots,x_n\}$ для i-й модели M_i частное распределение или вероятность $\mathbf{x}_n$ задается выражением

$$p_i\left(\mathbf{x}_n\right) = \int f_i\left(\mathbf{x}_n\middle|\boldsymbol{\theta}_i\right)\pi_i\left(\boldsymbol{\theta}_i\right)d\boldsymbol{\theta}_i\,. \qquad (1)$$

Если предположить, что априорная вероятность i-й модели, то апостериорная вероятность i-й модели:

$$P\left(M_i\middle|\mathbf{x}_n\right) = \frac{p_i\left(\mathbf{x}_n\right)P\left(M_i\right)}{\sum_{j=1}^{r} p_j\left(\mathbf{x}_n\right)P\left(M_j\right)},\ i = 1,2,\ldots,r\,.$$

Апостериорная вероятность отображает вероятность того, что полученные наблюдения $\mathbf{x}_n$ сгенерированы i-й моделью. Таким образом, если из множества r-моделей выбирается одна модель, целесообразно использовать модель с максимальной апостериорной вероятностью. Это означает, что выбирается модель з наибольшим значением $p_i\left(\mathbf{x}_n\right)P\left(M_i\right)$, так как знаменатель у всех моделей одинаковый. Если предположить, что априорные вероятности $P(M_i)$ всех моделей одинаковы, то выбирается модель, которая обеспечивает наибольшее частное распределение $p_i(\mathbf{x}_n)$ данных.

Байесовский информационный критерий определяется как натуральный логарифм (1), умноженный на -2:

$$-2\log p_i\left(\mathbf{x}_n\right) = -2\log\left\{\int f_i\left(\mathbf{x}_n\middle|\boldsymbol{\theta}_i\right)\pi_i\left(\boldsymbol{\theta}_i\right)d\boldsymbol{\theta}\right\} \approx$$

$$\approx -2\log f_i\left(\mathbf{x}_n\middle|\hat{\boldsymbol{\theta}}_i\right) + k_i\log n,$$

где $\hat{\boldsymbol{\theta}}_i$ -оценка k_i-размерного вектора параметров $\boldsymbol{\theta}_i$ -модели $f_i\left(x\middle|\boldsymbol{\theta}_i\right)$, полученная методом максимального правдоподобия. То есть из r-моделей, оцененных методом максимального правдоподобия, выбирается модель, которой соответствует наименьшее значение БИК.

Таким образом, даже если исходить из предположения, что все модели имеют одинаковую априорную вероятность, апостериорная вероятность, полученная на основе наблюдений, позволяет выбрать модель, которая максимально соответствует этим наблюдениям.

Отношение апостериорных вероятностей двух моделей M_1 и M_2 (апостериорная вероятность модели $P\left(M_i\middle|\mathbf{x}_n\right)$ (i=1,2)):

$$\frac{P\left(M_1\middle|\mathbf{x}_n\right)}{P\left(M_2\middle|\mathbf{x}_n\right)} = \frac{p_1\left(\mathbf{x}_n\right)}{p_2\left(\mathbf{x}_n\right)}\frac{P\left(M_1\right)}{P\left(M_2\right)}.$$

Отношение

$$B_{12} = \frac{p_1\left(\mathbf{x}_n\right)}{p_2\left(\mathbf{x}_n\right)}\frac{\int f_1\left(\mathbf{x}_n\middle|\boldsymbol{\theta}_1\right)\pi_1\left(\boldsymbol{\theta}_1\right)d\boldsymbol{\theta}_1}{\int f_2\left(\mathbf{x}_n\middle|\boldsymbol{\theta}_2\right)\pi_2\left(\boldsymbol{\theta}_2\right)d\boldsymbol{\theta}_2}$$

представляет собой коэффициент Байеса.

Байесовский информационный критерий с учетом статистической модели $f_i\left(\mathbf{x}_n\middle|\hat{\boldsymbol{\theta}}_i\right)$, полученной методом максимального правдоподобия:

$$BIC = -2\log f\left(x_n\middle|\hat{\boldsymbol{\theta}}\right) + p\log n\,.$$

Число наблюдений n должно быть достаточно большим.

Обобщенный байесовский информационный критерий (GBIC). Предположим, модель $f\left(\mathbf{x}_n\middle|\boldsymbol{\theta}_P\right)$ строится путем максимизации нелинезированной логарифмической функции правдоподобия. Тогда оценка модели на основе байесовского подхода:

$$GBIC = -2\log f\left(\mathbf{x}_n\middle|\hat{\boldsymbol{\theta}}_P\right) + n\lambda\hat{\boldsymbol{\theta}}_P^T K\hat{\boldsymbol{\theta}}_P + (p-d)\log n +$$

$$+\log\left|J_\lambda\left(\hat{\boldsymbol{\theta}}_P\right)\right| - d\log\lambda - \log|K|_+ - (p-d)\log(2\pi)\,,$$

где λ – гиперпараметр, который выбирается с учетом минимизации GBIC; K – $p \times p$ определенная матрица ранга d; $|K|_+$ – произведение d ненулевых собственных векторов K; и

$$J_\lambda\left(\hat{\boldsymbol{\theta}}_P\right) = -\frac{1}{n}\frac{\partial^2 \log f\left(x_n\middle|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}_P} + \lambda K\,.$$

Плотность вероятности $\pi\left(\boldsymbol{\theta}\middle|x_n;\lambda\right)$ определяется как априорная вероятность p-размерного параметра $\boldsymbol{\theta}$ для модели $f\left(x_n\middle|\boldsymbol{\theta}\right)$:

$$\pi\left(\boldsymbol{\theta}\middle|x_n;\lambda\right) = \frac{f\left(x_n\middle|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\middle|\lambda\right)}{\int f\left(x_n\middle|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}}\,.$$

Информационный критерий Байеса-Айкаике (ABIC). Если рассмотреть частное распределение $p\left(x_n\middle|\boldsymbol{\lambda}\right)$ байесовской модели как параметрическую модель с гиперпараметром $\boldsymbol{\lambda}$, тогда оценка модели может быть рассмотрена как:

$$ABIC = -2\log\left\{\max_{\boldsymbol{\lambda}} p\left(x_n\middle|\boldsymbol{\lambda}\right)\right\} + 2q =$$

$$= -2\max\log\left\{\int f\left(x_n\middle|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\middle|\boldsymbol{\lambda}\right)d\boldsymbol{\theta}\right\} + 2q.$$

Согласно байесовскому подходу на основе ABIC величина гиперпараметра λ байесовской модели может быть оценена либо максимизацией частного распределения $p\left(x_n\middle|\boldsymbol{\lambda}\right)$, либо логарифмического частного распределения $\log p\left(x_n\middle|\boldsymbol{\lambda}\right)$. Из двух или более байесовских моделей, которые характеризируются гиперпараметрами, необходимо выбирать модель, которая минимизирует ABIC.

Информационный критерий для оценки меры подобия двух временных рядов на основе байесовского информационного критерия. Пусть заданы два временных ряда, представленные векторами $X = \left\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n_x}\right\}$ и $Y = \left\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{n_y}\right\}$. Необходимо дать оценку двум следующим гипотезам [9]:

$$H_1 : \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n_x},\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{n_y} - N\left(\mu,\Sigma\right),$$

$$H_1 :\ \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n_x} - N\left(\mu_X,\Sigma_X\right);$$

$$\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{n_y} - N\left(\mu_Y,\Sigma_Y\right).$$

Гипотеза H_1 предполагает, что X и Y получены из одной модели, гипотеза H_2 означает, что X и Y формируются двумя моделями. Пусть $Z = X \cup Y$ и $n = n_x + n_y$. Тогда разница между величинами байесовского критерия для двух гипотез может быть определена таким образом:

doi.org/10.36773/1818-1212-2020-123-5-66-71

$$\Delta BIC(X,Y) = BIC(H_1,Z) - BIC(H_2,Z) =$$
$$= log\, p\left(X \left| \hat{\boldsymbol{\mu}}_X, \hat{\boldsymbol{\Sigma}}_X \right.\right) + log\, p\left(Y \left| \hat{\boldsymbol{\mu}}_Y, \hat{\boldsymbol{\Sigma}}_Y \right.\right) - log\, p\left(Z \left| \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \right.\right) -$$
$$-\frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right) log\, n = \frac{n}{2} log \left|\hat{\boldsymbol{\Sigma}}\right| - \frac{n_x}{2} log \left|\hat{\boldsymbol{\Sigma}}_X\right| -$$
$$-\frac{n_y}{2} log \left|\hat{\boldsymbol{\Sigma}}_Y\right| - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right) log\, n,$$

где $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}_X$ и $\hat{\boldsymbol{\mu}}_Y$ – векторы средних значений выборок *Z*, *X* и *Y* соответственно; $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}}_X$ и $\hat{\boldsymbol{\Sigma}}_Y$ – ковариационные матрицы выборок; *d* – размерность вектора признаков [10]. Чем меньше величина Δ*BIC*, тем более подобными являются два временных ряда. При условии $\lambda = 0$ расстояние Δ*BIC* является эквивалентным отношению правдоподобия.

Агломеративная кластеризация злоумышленников по временным характеристикам

Процесс агломеративной кластеризации начинается с *L* кластеров, каждый из которых содержит по одному объекту. На каждом шаге агломеративной кластеризации две наиболее подобных группы объединяются в одну.

Процедура кластеризации временных рядов, которые отображают характеристики профилей злоумышленников, основана на агломеративной кластеризации. Процедура включает следующие шаги:

1. Определение начальных кластеров. На первом этапе кластеризации каждый часовый ряд представляет собой кластер. Соответственно множество кластеров:

$$C = \{C_1, C_2, C_3, \ldots, C_l\},$$

где $C_1 = \{X^{11}\}$, $C_2 = \{X^{12}\}$, $C_3 = \{X^{13}\}$, и так далее, $C_l = \{x^{mn}\}$; $l = m \times n$ – количество начальных кластеров.

2. Определение расстояний между парами кластеров, которые определяют на основе одного из информационных критериев, приведенных выше. Если в качестве такого информационного критерия используется байесовский информационный критерий, то расстояние между двумя кластерами $C_i = \{X_1, X_2, X_3, \ldots, X_m\}$ и $C_j = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$ определяется как

$$D_{ij} = \Delta BIC\left(C_i, C_j\right) = \frac{\sum_{a=1}^{m}\sum_{b=1}^{n} \Delta BIC(X_a, Y_b)}{\left|C_i \times C_j\right|},\ a \neq b,$$

где *m* – количество объектов в кластере C_i; *n* – количество объектов в вкластере C_j; *i,j*=1,2,3…*l* – количество кластеров текущей итерации процесса кластеризации.

3. Определение из множества расстояний $D = \{D_{12}, D_{13}, \ldots, D_{ij}, \ldots, D_{l-1,l}\}$, минимального для текущей итерации, $D_{min\,iter} = min(D) = D_{ij}$, выбор соответствующей пары кластеров C_i и C_j и объединение множества объектов этих двух кластеров в один кластер $C_{join(ij)} = C_i \cup C_j$.

4. Определение интегрального критерия расстояний между объектами внутри кластера. Интегральный критерий определяется на основе одного из информационных критериев. Если критерий определяется на основе байесовского информационного критерия, то он имеет вид:

$$J = \frac{\sum_{i=1}^{l} \Delta BIC(C_i)}{l},$$

где $\Delta BIC(C_i)$ – расстояние между объектами внутри одного кластера; *i*=1,2,3,…,*l*; *l* – количество кластеров. Это расстояние для кластера $C_i = \{X_1, X_2, X_3, \ldots, X_m\}$ определяется соотношением:

$$\Delta BIC(C_i) = \frac{\sum_{a=1}^{m}\sum_{b=a}^{m} \Delta BIC(X_a, X_b)}{m}.$$

5. Проверка условия остановки процесса объединения кластеров. Если условие выполняется, получают окончательное множество кластеров $C_{res} = \{C_1, C_2, C_3, \ldots, C_{lres}\}$, где l_{res} – окончательное число кластеров. Выполнение пунктов 2–5 повторяется до тех пор, пока условие остановки не будет выполняться.

Для остановки процесса объединения кластеров используется один из трёх критериев:

1. Количество кластеров. При использовании этого критерия требуется задать окончательное число кластеров, для чего необходима предварительная оценка. Этот критерий рекомендуется использовать, когда число кластеров злоумышленников известно, а также для определения уровня опасности злоумышленника.

2. Абсолютный минимум функции зависимости интегрального критерия от числа итераций объединения кластеров $min(J(k))$, где *k* – число итераций. При использовании этого критерия в качестве окончательного выбирается множество кластеров, которое соответствует итерации, на которой наблюдается минимум *J(k)*. Этот критерий остановки обеспечивает максимальную чистоту полученных кластеров. Недостатком остановки по этому критерию является тот факт, что часть кластеров может быть неполной, соответственно, злоумышленники одного типа могут находиться в разных кластерах.

3. Градиентный критерий. В этом варианте определяется скачок функции зависимости относительного градиента $\nabla J(k)/J(k)$ от числа итераций, выбирается множество кластеров, которые соответствуют этой итерации (рис. 1). Использование этого критерия обеспечивает максимальную полноту полученных кластеров.

Результаты экспериментальных исследований

При проведении экспериментальных исследований использовались данные собранные сетью ловушек [11]. Для анализа использовались временные ряды количества запросов за одну секунду от одного источника к одной ловушке, которые сформировали исходные данные для кластеризации. Для кластеризации использовалось несколько информационных критериев, в том числе BIC, GBIC и ABIC. Эксперименты проводились с учетом предварительно известного числа кластеров, с использованием этого числа в качестве критерия остановки кластеризации. Также проводилась кластеризация без предварительной информации о количестве кластеров с использованием двух критериев остановки: абсолютного минимума и относительного градиента. Эксперименты подтвердили, что использование абсолютного минимума обеспечивает максимальную чистоту кластеров, но кластеры могут оставаться неполными, и вычисленное количество кластеров превышает их действительное

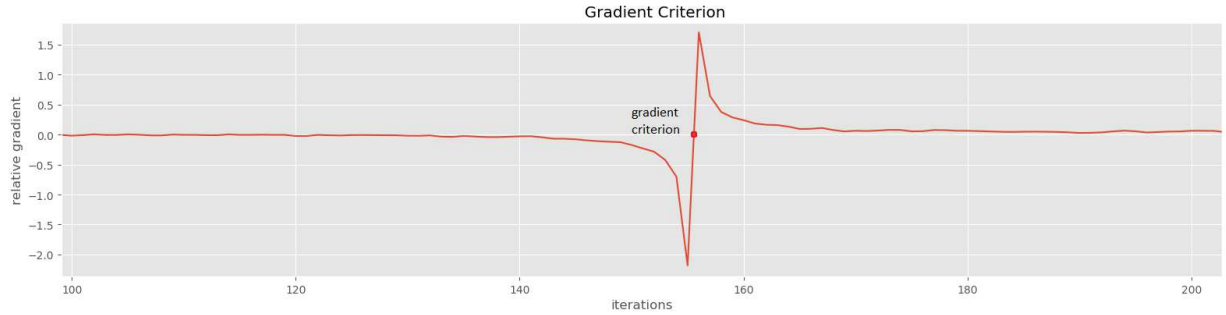

Рисунок 1 – Градиентный критерий остановки процесса кластеризации – зависимость относительного градиента функции $\nabla J(k)/J(k)$ от количества кластеров

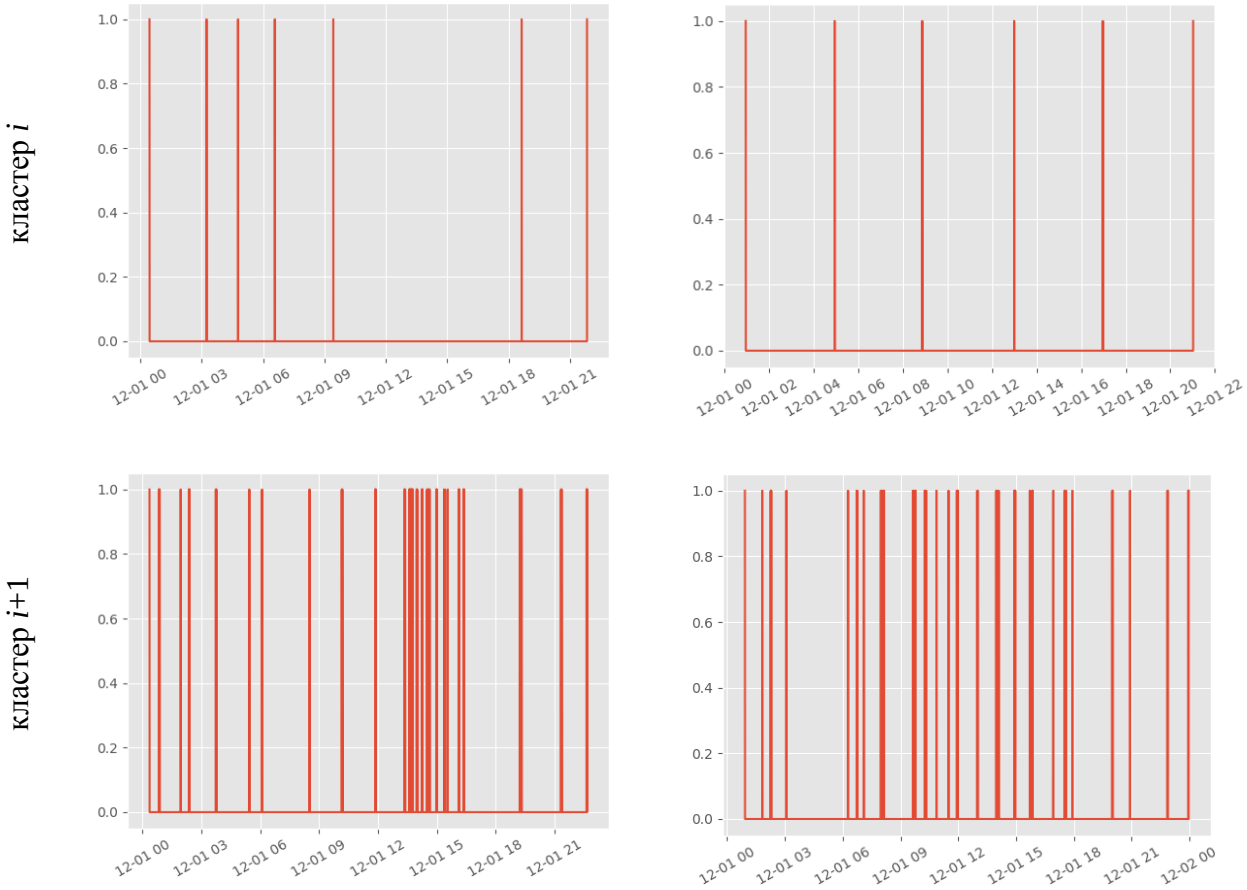

Рисунок 2 – Образцы временных рядов активностей злоумышленников, относящиеся к разным кластерам

количество. Использование в качестве критерия остановки относительного градиента обеспечивает максимальную полноту кластеров, но чистота кластеров ниже, чем при использовании абсолютного минимума. В идеальном случае кластеризации итерация, на которой останавливается процесс кластеризации, имеет один и тот же номер и при использовании критерия абсолютного минимума, и при использовании относительного градиента. В этом случае обеспечивается максимальная чистота и полнота кластеров. Наивысшие показатели точности в процессе эксперимента показала кластеризация с использованием байесовского информационного критерия (BIC). На рис. 2 показаны временные ряды активности злоумышленников из двух разных кластеров.

Заключение

Использование ловушек и сетей ловушек для сбора и анализа характеристик злоумышленников позволяет обнаруживать кроме известных атак также новые атаки, находить среди них атаки со сходными характеристиками, и, соответственно, обнаруживать новые типы атак. Сеть ловушек позволяет отслеживать злоумышленные и подозрительные действия во всех системах компьютерной сети. Информация о новых типах атак необходима для эффективной защиты компьютерных сетей в режиме реального времени. Описанный в статье подход агломеративной кластеризации злоумышленников позволят непрерывно отслеживать злоумышленников с подобными динамическими характеристиками, что позволяет повысить эффективность работы систем обнаружения вторжений.

doi.org/10.36773/1818-1212-2020-123-5-66-71

Список цитированных источников

1. Cisco 2019. Annual Report. Defining the Future of the Internet. – https://www.cisco.com/c/dam/en_us/about/annual-report/cisco-annual-report-2019.pdf. – Access : 29.09.2020.
2. Sochor, Tomas. Attractiveness Study of Honeypots and Honeynets in Internet Threat Detection / Tomas Sochor, Matey Zuzcak – Springer International Publishing Switzerland 2015, P.Gaj at al. (Eds.): CN 2015, CCIS 522. – P. 69–81. – 2015. – DOI : 10.1007/978-3-319-19419-6 7.
3. Varaprasad, Rao M. Algorithm for Clustering with Intrusion Detection using Modified & Hashed K-Means Algorithms / M. Varaprasad Rao, A. Damodaram, N. Ch. Bhatra Charyulu // Proceedings of 2nd International Conference on Computer Science Engineering Applications. – New Delhi, India, Springer Publications. – Vol 2. – May 25-27. – 2012. – P. 737–744.
4. Debar, H. A neural network component for an intrusion detection system / H. Debar, M. Becker, and D. Siboni // In Proceedings 1992 IEEE Computer Society Symposium on Research in Security and Privacy, 1992. – P. 240–250.
5. Bhattacharjee, P. S. Intrusion detection system for nsl-kdd data set using vectorised fitness function in genetic algorithm / P. S. Bhattacharjee, A. K. M. Fujail, and S. A. Begum // Adv. Comput. Sci. Technol. – Vol. 10. – No. 2. – 2017. – P. 235–246.
6. Puri, A. A novel technique for intrusion detection system for network security using hybrid svm-cart / A. Puri and N. Sharma // IJEDR. – Vol. 5. – No. 2. – 2017. – P. 155–161.
7. Sung, S. Jo H. comparative study on the performance of intrusion detection using decision tree and artificial neural network models / S. Jo, H. Sung, and B. Ahn // Journal of the Korea Society of Digital Industry and Information Management. – Vol. 11. – No. 4. – 2015. – P. 33–45.
8. Konishi, S. Information Criteria and Statistical Modeling. Springer Series in Statistics / S. Konishi, G. Kitagawa // Springer Science+Business Media, LLC, 2008. – 287 p.
9. Chen, S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion / S. Chen and P. S. Gopalakrishnan // In Proc. DARPA Broadcast News Transcription and Understanding Workshop. – Lansdowne, VA, Feb. – 1998. – P. 127–132.
10. Cettolo, M. Evaluation of BIC-based algorithms for audio segmentation / M. Cettolo, M. Vescovi, and R. Rizzi // Computer Speech and Language. – Vol. 19. – 2005. – P. 147–170.
11. Traffic Data from Kyoto University's Honeypots. – https://www.takakura.com/Kyoto_data/. – Access : 01.09.2020.

References

1. Cisco 2019. Annual Report. Defining the Future of the Internet. – https://www.cisco.com/c/dam/en_us/about/annual-report/cisco-annual-report-2019.pdf. – Access : 29.09.2020.
2. Sochor, Tomas. Attractiveness Study of Honeypots and Honeynets in Internet Threat Detection / Tomas Sochor, Matey Zuzcak – Springer International Publishing Switzerland 2015, P.Gaj at al. (Eds.): CN 2015, CCIS 522. – P. 69–81. – 2015. – DOI : 10.1007/978-3-319-19419-6 7.
3. Varaprasad, Rao M. Algorithm for Clustering with Intrusion Detection using Modified & Hashed K-Means Algorithms / M. Varaprasad Rao, A. Damodaram, N. Ch. Bhatra Charyulu // Proceedings of 2nd International Conference on Computer Science Engineering Applications. – New Delhi, India, Springer Publications. – Vol 2. – May 25-27. – 2012. – P. 737–744.
4. Debar, H. A neural network component for an intrusion detection system / H. Debar, M. Becker, and D. Siboni // In Proceedings 1992 IEEE Computer Society Symposium on Research in Security and Privacy, 1992. – P. 240–250.
5. Bhattacharjee, P. S. Intrusion detection system for nsl-kdd data set using vectorised fitness function in genetic algorithm / P. S. Bhattacharjee, A. K. M. Fujail, and S. A. Begum // Adv. Comput. Sci. Technol. – Vol. 10. – No. 2. – 2017. – P. 235–246.
6. Puri, A. A novel technique for intrusion detection system for network security using hybrid svm-cart / A. Puri and N. Sharma // IJEDR. – Vol. 5. – No. 2. – 2017. – P. 155–161.
7. Sung, S. Jo H. comparative study on the performance of intrusion detection using decision tree and artificial neural network models / S. Jo, H. Sung, and B. Ahn // Journal of the Korea Society of Digital Industry and Information Management. – Vol. 11. – No. 4. – 2015. – P. 33–45.
8. Konishi, S. Information Criteria and Statistical Modeling. Springer Series in Statistics / S. Konishi, G. Kitagawa // Springer Sci-ence+Business Media, LLC, 2008. – 287 p.
9. Chen, S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion / S. Chen and P. S. Gopalakrishnan // In Proc. DARPA Broadcast News Transcription and Understanding Workshop. – Lansdowne, VA, Feb. – 1998. – P. 127–132.
10. Cettolo, M. Evaluation of BIC-based algorithms for audio segmentation / M. Cettolo, M. Vescovi, and R. Rizzi // Computer Speech and Language. – Vol. 19. – 2005. – P. 147–170.
11. Traffic Data from Kyoto University's Honeypots. – https://www.takakura.com/Kyoto_data/. – Access : 01.09.2020.

Материал поступил в редакцию 12.01.2021