Applying Correlation and Regression Analysis to Detect Security Incidents in the Internet of Things

Daria Lavrova¹ and Alexander Pechenkin²

¹Peter the Great Saint-Petersburg Polytechnic University, Russian Federation
²Peter the Great Saint-Petersburg Polytechnic University, Russian Federation,
lavrova.daria@gmail.com, alexander.pechenkin@iibks.fkt.spbstu.ru

Abstract: In this paper, authors propose a complex of correlation-based methods for security incidents detection and investigation in large-scale networks of heterogeneous devices such as Internet of Things. Proposed methods aim to detect both known and unknown attacks in the Internet of Things.

Keywords: Internet of Things, event correlation, security incident, regression analysis, correlation analysis, rule-based correlation.

1. Introduction

The Internet of Things (IoT) have been a current trend in information security, and the development of the IoT and its introduction in all spheres of human activity caused a large number of new security threats [1].

In accordance with [2], unknown attacks have a leading position among other for last months (since December 2014 until June 2015). The complexity of security incidents detection in the IoT is based on a high heterogeneity of the IoT and on a low capacity of many smart devices, which does not allow integration with third-party protection means.

To overcome these problems one need to analyze huge amounts of heterogeneous data from the IoT devices. We propose a complex of methods, which is partly, solves a problem of security incidents detection in the IoT.

Proposed approach is inspired by an approach which is often used in security information and event analysis and management, such systems are called SIEM (security information and event management) [3]. SIEM systems primarily work with security events generated by various network security tools (such as firewalls, IDS, IPS etc), and the decision about security incident existence or absence is taken in accordance with the results of event correlation.

Event correlation allows detecting interconnections between events from different devices. There are many different ways for event correlation such as rule-based correlation, statistical-based correlation, model-based correlation etc [4, 5]. For providing cyber security in the IoT, it is necessary to have methods for:

- Both known and unknown attacks detection
- Security incidents investigation

Proposed complex of methods aims to detect and to investigate security incidents and using this methods for detection both known and unknown attacks is possible.

2. Related Work

Proposed SIEM-based approach is also inspired by the work [6]. In this paper, authors propose an event-based concept for an autonomous adaptive risk management solution for the Internet of Things. Authors also propose to use event correlation for monitoring and detecting threats. In this paper authors describe several ways for event correlation, such as rule-base correlation, state machine automata based correlation, the codebook/correlation matrix techniques, anomaly-based approach, statistical correlation and probabilistic modeling. Authors primarily use rule-base correlation. The main difference and the novelty of our approach lies in implementation also a method for finding implicit and hidden relationships between events using correlation and regression analysis. This allows detecting new functional relationships between devices in the Internet of Things, which may display relationships which should not be. Such “bad” relationships characterize security incidents in the Internet of Things. So this approach allows to detect not only known attacks but also unknown attacks because such relationships may be caused by zero-day vulnerabilities, security leakages etc.

Also in our approach, as compared with the work [6], event correlation is investigated more deeply, by building event chains. That helps us to detect security incidents in the Internet of Things at an early stage.

Proposed complex of methods contains rule-base event correlation method that is a common approach used in a lot of research papers. In [7] authors propose correlation rules to define the relationship between physical and logical security events caused by abnormal behavior activities, that will helps to detect the multi-stage attacks. In [8] authors note that rule-based analysis is a classical approach to event correlation. That is, a correlation system constantly uses a set of predefined rules to evaluate incoming observations until a conclusion is reached. Therefore the correlation ability depends solely on the depth and capability of the rule set. Rule-based correlation method of detection priori insecure events proposed in our paper, works similarly, based on known attacks vectors in the Internet of Things and their manifestations.

In accordance with [9], anomaly detection is a common practice to ensure information security. In our work, we also use statistical-based correlation for monitoring anomalies in the Internet of Things events and their parameters. Statistical-based correlation does not employ any preexisting knowledge of the malicious activity, but instead relies upon the knowledge (and recognition) of normal activities, which has been accumulated over time. Ongoing events are then rated by a built-in algorithm and may be compared to the accumulated activity patterns, to distinguish normal from abnormal. In [10] authors proposed an approach which is based on a historical analysis of abnormality events with their probabilistic correlations. In [11] authors analyze the
use of different types of statistical tests for the correlation of anomaly detection alerts. Our statistical and rule-based correlation method of detection potential security events works similarly, but it also uses a rule-based approach for detecting security incidents.

Applying both correlation and regression analysis for providing cyber security in the Internet of Things is not widely covered. There are a sufficiently large number of works devoted to correlation analysis for providing information security in complex information systems including the Internet of Things. In [12] authors propose an assessment standard of how to calculating the threat factor and its weight value by analyzing the relationship between the utilization of the various factors, and decomposing the path on Internet of things system using the graph theory, probability theory and the correlation function.

Applying regression analysis methods for the IoT analytics is represented in works [13-15]. In [13] a multiple regression analysis is proposed to test whether the evaluations of the smart fridge after the scenario are related to their counterparts after interaction. In [14] authors propose a novel logit regression-based trust model called LogitTrust to model dynamic trust for service-oriented mobile ad hoc networks (MANETs) wherein a node can be a service requester or a service provider. MANETs are part of the Internet of Things implementation. In [15] regression analysis is applied to estimate the relationships between trust and a set of variables characterizing the behavior of a node.

The novelty of our method for finding implicit relationships of devices using correlation and regression analysis lies in applying a special coefficient that reflects the similarity in dynamics of two data sets. Data sets are represented as event parameters, collected for a certain period. Proposed method is based on an assumption that the series of interconnected data are changing similarly. Thus, interconnected events and interconnected devices could be obtained. When it becomes clear how devices are interconnected will be much easier to detect security incidents by regular monitoring of preserving or violation of relationships.

Proposed method for finding implicit relationships of devices using correlation and regression analysis is inspired by work [16]. In [16] authors proposed a security event correlation algorithm based on attribute similarity of mixed data type, which can refine the security event information and improve the availability of the security event information. This method is adopted to event characteristics of a proposed formal model of the Internet of Things event. For building event chains, we partly use an approach, proposed in [17]. This approach is based on event’s clustering and correlation analysis between groups. In our approach, we add correlation analysis not only between security events but also between standard events for finding implicit correlations between clusters.

3. Features and Formal Models of IoT

Nowadays evolution of the IoT follows the trend of increasing intelligence of devices that are part of the IoT. This means that there is an increasing number of projects in which devices communicate with each other without human involvement. So the extent of human influence on the device functioning is minimized and eventually could be reduced to zero. Thus, the data for analysis is fully represents the data from the devices, each of which operates in accordance with a certain algorithm. For this reason, device’s data analytics will give better results than data analysis from users whose behavior is difficult to predict.

Proposed complex of correlation-based methods consists of 4 methods:

- Rule-based correlation method of detection prior insecure events
- Both statistical and rule-based correlation method of detection potential security events
- Correlation and regression method of detection potential security events
- Event correlation method for investigation security incidents

Rule-based correlation method of detection prior insecure events aims to detect known attacks in the IoT. Statistical and rule-based correlation method aims to detect abnormal event characteristics by evaluation of statistical parameters, such as maximum, minimum, standard deviation etc. This method also uses rule-based correlation for detecting security events, but it is able to detect anomalies in data that could be an unknown attack’s manifestation. Correlation and regression method of detection potential security events aims to find implicit interconnections in data from IoT devices and to reveal an analytical form of interconnections. Violations of these interconnections may indicate potential security incidents.

The last event correlation method for security incidents investigation aims to analyze interconnections between both security and standard events. On the basis of event correlation results event chaining building becomes possible, thus allows detect security incidents in the IoT at early stage. For representation data from the IoT devices as an event there is one need to implement a formal model of an event, according to which large amounts of heterogeneous data from device will be aggregated and lead to a common sight due to the normalization. A formal model of an event is also needed for development the complex of security incidents detection methods.

Applied to the IoT an event represents in a form of tuple called event, describes the following information fields:

- Source
- Destination
- Action
- Date
- Time
- Functions

Event’s parameter Source is a set of characteristics to uniquely identify the device that generates the event. Source is described by IP address and ports of a device, and also by device ID. So, Source = \{Source_IP, Source_id, Source_ports\}. Destination is a set of characteristics to uniquely identify the device that receives the event, Destination = \{Destination_IP, Destination_id, Destination_ports\}. Action characterizes an event type, Action = \{Message, Command, Unknown\}. Date consists of three characteristics:
Date = \{Date\_start, Date\_stop, Date\_abs\}, where Date\_start is a date from which event searching starts, Date\_stop is a date until which event searching is performed, Date\_abs is a date of logging an event in a system. Analogically, time characterizes time of event generation.

Time = \{Time\_start, Time\_stop, Time\_abs, Time\_critical\}. Here Time\_start is a date from which event searching starts, Time\_stop is a date until which event searching is performed, Time\_abs is a date at which the event was logged in the system, Time\_critical is a critical period.

There is also a set of functions for event processing:

- Get\_Num
- Get\_Value
- Find\_Event

Functions set:

Functions = \{Get\_Num, Get\_Value, Find\_Event\}.

Get\_Num function determines the number of events, Get\_Value determines the value of an event’s characteristic, Find\_Event function finds in logfile an event with appropriate characteristics. Thus an event is represented as a tuple:

Event = \{Source, Destination, Action, Time, Functions\}.

4. Rule-Based Correlation Method of Detection Prior Insecure Events

Under a prior insecure events to be understood events whose presence in the indicates a violation of the correct functioning of the IoT device / segment. Using rules-based correlation implies the description of a particular sequence of logical operations that characterize the actions of the attacker. With respect to the IoT, the use of rule-based correlation approach is offered subject to the availability of knowledge about how an attacker’s action may affect the data from the IoT devices. To do this, it is advisable to study the existing classification of network attacks and attacks on the IoT. Knowledge from known attacks’ classifications expands the list of developed rules and allows to improve an accuracy of detected attacks. Summarizing existing security incidents for the IoT, we reduce a prior insecure events to the following:

- Data absence from the device
- Connection with an IP-address from the black-list
- Error message
- Unknown type event

A data absence from the device can mean that the device is turned off or broken, that may characterize a potential security incident.

Rules are described using a formal language developed for submission to the correlation rules. This language combines the pseudo-code and mathematical and logical operators. Rules represent an algorithmic scheme, according to which developed rules are functioning, operators reflect the logic of finding the values of the parameters, or the number of investigated events described in accordance with the formal model of the event on the IoT. The main elements of a formal language are:

- Entities: event and the tuple’s elements, Event\_in\_List, that means a set of events represented in a log file, blackIPList etc
- Events’ characteristics: Values (Value\_threshold, Value\_start, Value\_stop, Value\_current, Value\_error etc), type
- Mathematical operators: =, ≠, +, −, <, >, ≤, ≥, U, ∩, ∅, ∈;
- Logical operators: →, &,
- Conditional operators: if, for, then
- Functions: Get\_Num, Get\_Value, return
- Alert signal: alert!

An address to a tuple’s element is realized with the usage of black square brackets, in which an event is written. For example, getting the type of event is written as Event[Event].

An address to an element with certain type is realized with the usage of index. For example, if we want to sort only events with Action = Message, we use Action\_type [Event] = Message.

Indexing elements is also used to detail the object, assign it a label. For example, to mark a specific value from the device as erroneous, we use the error designation (error) as the index for the value: Value\_error.

The rule for detecting events indicating the absence of data on the device is as follows:

\begin{align*}
\text{for:} & \text{Event\_in\_List} \\
\text{if} & \text{Source[Event] = Source\_current} \\
\text{& Time\_abs[Event] < Time\_abs[Event]} \\
\text{then} & n = n + 1
\end{align*}

\begin{align*}
\text{if} & (n = 0) \\
\text{then} & \text{alert!}
\end{align*}

where:

- Event\_in\_List – list of events
- Source\_current – current value of event’s source
- Time\_abs[Event] – time at which the event was logged in a system
- Time\_threshold – a threshold time value, in this case a response time
- n – is a variable counter
- alert! – is an alert signal

Similar to this rule, the rest two rules are written. The rule for detecting events indicating an error message is as follows:

\begin{align*}
\text{for:} & \text{Event\_in\_List} \\
\text{if} & \text{Action\_type[Event] = Message} \\
\text{& Action\_value[Event] = Value\_error} \\
\text{then} & \text{alert!}
\end{align*}

The rule for detecting events indicating a connection with IP address from blacklist is as follows:

\begin{align*}
\text{for:} & \text{Event\_in\_List} \\
\text{if} & \text{Action\_type[Event] = Command} \\
\text{& Source\_IP[Event] ∈ blackIPList}
\end{align*}
& Destination_IP[Event] ∈ [blackIPList])
  then alert!
}

Advantage of this rule-based correlation method is its high accuracy of security incidents detection. Disadvantages are need to regularly update the rules, inability to respond to situations that are not described in rules, therefore, inability to detect unknown attacks.


Potential security events could be caused by different event characteristics:
- Number of events, general or for a certain time period
- Values of the event tuple’s elements

Potential security events caused by number of events are:
- Too many events from a device
- Too few events from a device

Potential security events caused by values of the event tuple’s elements are:
- Element value is too big
- Element value is too little
- Element value significantly changes over time
- Violation of event’s periodicity

First two types of events could be easily calculated. Special attention should be paid to an event, which characterizes significantly changes of an element value over time. The need to detect such type of events lies in a gradual increase of an attack intensity. Therefore, if event monitoring is carried out taking into account the recent changes, then calculation of new threshold values will include values, which are modified by an intruder. Due to this, the boundary values will gradually shift without an alert signal.

Calculation of threshold values is based on the standard deviation’s calculation. Using a standard deviation, we can calculate an allowable threshold value for every event’s characteristic. If the boundary is violated it will be an evidence of attack. Since in various times the load on a device can also be different, for the early detection of attack requires constant monitoring and counting of boundaries for each time step. This approach is inspired by works [18], in which authors analyze a problem of early DDoS attacks detection. In accordance with the results of experiments described in [18] the most perspective analysis method is the method which considering the seasonality. Detailed study of different kind of DDoS attacks and existing defense strategies is also represented in [19].

Let $x_i = Get\_Num(\text{Event\_in\_List})$, where $\text{Event\_in\_List}$ contains events from the device for the certain period. Denote the number of daily periods as $n$. Then we can write all events from the device as a matrix:

$$
\begin{pmatrix}
X_{11} & X_{12} & X_{13} & \cdots & X_{124} \\
X_{21} & X_{22} & X_{23} & \cdots & X_{224} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & X_{n3} & \cdots & X_{n24}
\end{pmatrix}
$$

Each matrix row includes daily data about number of events. The first row represents the data of a current day, in this regard, it can be not completed until the end. The calculation is made on the columns of the matrix: $x_{11}, x_{21}, x_{12}, \ldots, x_{24}$.

If the device is using in accordance with weekly or diurnal cycle, it is necessary to exclude lines that match the festive and weekends.

Rules for detecting events which threshold values are calculated using this statistical-based approach are also described by correlation rules in the same formal language. Advantage of this statistical and rule-based correlation method is its ability to detect anomalies in data and, consequently, to detect security incidents in the IoT in early stages. The main difficulty of this approach lies in a nicety of the right choice of periods, similar to each other.

6. Correlation and Regression Method of Detection Potential Security Events

6.1 Correlation analysis

Both previously presented methods are aimed at detection of security incidents, which manifested in the events from a single device. However, in case of providing security of the IoT it is necessary to use a “sensor fusion” principle. That means that the resulting information from combining of sensory data or data derived from disparate sources such has less uncertainty than would be possible when these sources were used individually.

Knowledge about interconnectedness of devices in the analyzed segment of the IoT will enable to detect unknown attacks because violations of interconnections may characterize a potential security incident. In addition, this knowledge helps in investigating security incidents.

Proposed correlation and regression method of detection potential security events includes two steps:
- Finding interconnections in data from the IoT devices
- Determination of an analytical form for identified interconnections

Proposed method is based on the assumption that in the analyzed segment of the IoT, where devices control each other without operator’s influence (or with minimal influence), data sets generated by interconnected devices vary in concert and their dynamics is very similar. For this method uses correlation analysis for determination the existence of dependence, her strength and direction.

Linear interconnections could be found by using Pearson’s correlation coefficient (or the correlation coefficient):

$$
r = \frac{cov(X,Y)}{\sigma_x \sigma_y}
$$

where:
- $X$ is a variable #1
- $Y$ is a variable #2
- $cov$ is the covariance
- $\sigma_x$ is the standard deviation of $X$
- $\sigma_y$ is the standard deviation of $Y$

However, this factor can not reflect the existence of interconnection, if it is non-linear. For detecting non-linear interconnection, we use a special coefficient proposed in [20]. It is proposed to call the coefficient of similarity in dynamics. The degree of similarity in dynamics is investigated by estimation of the coefficients of the polynomial, which describes the initial set of observations, using the mathematical apparatus of finite differences. Formula of the coefficient of similarity in dynamics is as follows:
\[ k_s = \frac{\sum \Delta y \Delta x}{\left( \sum (\Delta y) \right)^2 \sum (\Delta x)^2} \] (2)

where:

- \( x \) is a set of values of an estimated event’s element for device #1
- \( y \) is a set of values of an estimated event’s element for device #1
- \( \Delta^i x \) is a \( i^{th} \) order difference for \( x \)
- \( \Delta^i y \) is a \( i^{th} \) order difference for \( y \)

Derivation of the coefficient’s formula is represented in detail in work [20].

In our research, we conducted an experiment. We took a dataset that characterizes the intensity of the automated blast furnace; it represents a nonlinear dependence between the hot blast temperature and the output per a single production mechanism. This dependence is not linear, because the rise of temperature can increase output only up to certain limits, beyond a some threshold intensity of output decreases.

<table>
<thead>
<tr>
<th>#</th>
<th>Hot blast temperature denoted as ( x )</th>
<th>Output denoted as ( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.01</td>
<td>8.80</td>
</tr>
<tr>
<td>2</td>
<td>1.15</td>
<td>9.20</td>
</tr>
<tr>
<td>3</td>
<td>1.91</td>
<td>8.70</td>
</tr>
<tr>
<td>4</td>
<td>2.47</td>
<td>10.20</td>
</tr>
<tr>
<td>5</td>
<td>2.66</td>
<td>9.30</td>
</tr>
<tr>
<td>6</td>
<td>2.74</td>
<td>9.40</td>
</tr>
<tr>
<td>7</td>
<td>2.93</td>
<td>10.70</td>
</tr>
<tr>
<td>8</td>
<td>4.04</td>
<td>8.50</td>
</tr>
<tr>
<td>9</td>
<td>4.50</td>
<td>8.90</td>
</tr>
<tr>
<td>10</td>
<td>4.64</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Table 1. The normalized initial data

Analyzing the dynamics of the data series in accordance with Figure 1, it can be concluded that the connection is not obvious.

The need to calculate the correlation coefficient in addition to the coefficient of similarity in dynamics is based on the fact that the proximity of coefficient of similarity in dynamics to 1 may indicate the existence of a strong nonlinear interconnection, while the value of correlation coefficient is close to 1 only in case of a strong linear relationship. Thus, the value of the correlation coefficient is a complementary factor for the second step of the method – an establishment of forms of analytic relationship.

Determination of an analytical form for identified interconnections based on the following approaches:

- Linear regression
- Nonlinear regression
- Extrapolation

Assuming a linear interconnection, we write the line equation in normal form and find its coefficients using the method of least squares. In the absence of a pronounced linear interconnection, experimental data is represented a curve line in accordance with a set of nonlinear functions, which are standard when trying to identify an analytical form of interconnection. If identifying an analytical form of interconnection using described regression approaches is impossible then we propose to use extrapolation for prediction of a set of future values.

List of standard using functions when trying to identify an analytical form of interconnection is as follows:

- A linear function \( y = ax + b \)
- An exponential function \( y = ab^x \)
- A rational function of the form \#1 \( y = \frac{1}{ax+b} \)
- A logarithmic function \( y = a \ln x + b \)
- A power function \( y = ax^b \)
- A hyperbolic function \( y = a + \frac{b}{x} \)
- A rational function of the form \#2 \( y = \frac{x}{ax+b} \)

In accordance with instructions proposed in [21], a regression form of the equation is:

\[ \tilde{y} = f(x) + \varepsilon \] (3)

Carrying out the calculations specified in [21] we could identify an analytical form of nonlinear interconnection between event’s characteristics and, as a consequence, between devices.

The scheme of a correlation and regression method of detection potential security events is represented by Figure 2.
Proposed method includes two steps:

- Correlation between two events
- Building an event-correlated chain

In the first step, we propose to correlate two events:

- Security event with security event,
- Standard event with security event,
- Standard event with standard event.

The second and the third correlation types are also advisable if an attack was skipped at an early stage so that an intruder penetrated in a protected segment of the IoT. In addition, there could be an insider in the protected segment and seemingly, legitimate actions may entail an attack characterized by security events.

Event correlation could be seen from calculation of similarities between them. In this method, similarity, and hence, correlation of events could be evaluated according to three parameters:

- Analysis of character parameters’ correlation
- Analysis of numerical parameters’ correlation
- Analysis of both character and numerical parameters’ correlation

A general scheme for this method is described as follows.

- Depending on the type of security incident, which is supposed to detect, assign weight to each type of event’s characteristics,
- Specify a correlation threshold, which would mean that if the correlation function takes a value greater than this threshold, then the events with sufficient force are correlated,
- Calculate the correlation between events using the correlation function,
- Compare an obtained value with a threshold correlation value,
- Group correlated events.

Further, it is advisable to calculate the correlation between the groups of events to build an event-correlated chain. For calculation of the similarity between events, we use functions proposed in [22, 23].

The similarity function for character parameters:

\[
Sim_{cha}(event_i, event_j) = \frac{\sum_{k=1}^{q} \varphi(value_{ik}, value_{jk})}{p},
\]

where:

- \(\varphi(value_{ik}, value_{jk})\) is a number of character parameters of an event
- \(value_{ik}, value_{jk}\) are values of character parameters of \(event_i\) and \(event_j\) respectively

The similarity function for numerical parameters:

\[
Sim_{num}(event_i, event_j) = \frac{\sum_{f=1}^{\omega_f} \sum_{i=1}^{\omega_f} Sim_f(event_i, event_j)}{\sum_{f=1}^{\omega_f} \omega_f},
\]

where:

- \(\omega_f\) is a weight of event’s parameter designated as \(f\)
- \(Sim_f(event_i, event_j)\) is a similarity of parameter \(f\) between \(event_i\) and \(event_j\)

The similarity function for both character and numerical parameters:

\[
Sim(event_i, event_j) = \mu Sim_{cha}(event_i, event_j) + 1 - \mu Sim_{num}(event_i, event_j),
\]

where \(\mu\) is a weight coefficient that can be used to correct the numeric and character parameters in the contribution to the function of correlation between the events.

This method requires the development of efficient event clustering algorithms and their functioning in a high performance cluster. The complexity of development of this method could be in a reduction of accuracy caused by the incorrect assignment of weighting coefficients.

8. Conclusions

In this paper, we propose a complex of correlation-based methods for security incidents detection in the IoT. Providing cyber security in the IoT is a task, which topicality and complexity increases along with an evolution of the IoT. Proposed complex of methods aims to detect and to investigate security incidents and using this methods for detection both known and unknown attacks is possible.

The main contributions of this paper includes a novel approach for security incidents detection based on a finding implicit interconnections between data for different devices in the IoT. This approach is based on a calculation of the coefficient of similarity in dynamics. This coefficient indicates how similar are changes in time the values of two datasets. Such similarity may be caused by interconnection of these two data sets, and, consequently, may be caused by a functional interconnection between devices that generate analyzed data. We conducted a small experiment using the machine data of the blast furnace, and found that even a
small volume of a dataset is able to detect an interconnection. Direction for our further works is to develop methods to collect and process data from the IoT in real time for rapid response to high critically security incidents.

9. Acknowledgement


References