Modeling Home IoT Traffic using Users' in-Home Activities for Detection of Anomalous Operations

Masaaki Yamauchi*, Masahiro Tanaka*, Yuichi Ohsita*, Masayuki Murata*, Kensuke Ueda†, Yoshiaki Kato‡
* *Graduate School of Information Science and Technology, Osaka University, Suita, Japan.*
Email: {m-yamauchi, m-tanaka, y-ohsita, murata}@ist.osaka-u.ac.jp
† *Advanced Technology R&D Center, Mitsubishi Electric Corporation, Amagasaki, Japan.*
Email: Ueda.Kensuke@ce.MitsubishiElectric.co.jp
‡ *Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Japan*
Email: Kato.Yoshiaki@dh.MitsubishiElectric.co.jp

***Abstract*—In this paper, we modeled home IoT traffic based on users' in-home activities by the sensors and operations of home IoT devices. Then, we applied the model to the detection of anomalous operations. We evaluated our model by using a dataset obtained in an actual home environment. The results demonstrated that the detection using our method achieved 72.3% detections with less than 20.1% misdetections.**

***Index Terms*—Smart home, IoT, Security, Anomaly detection, Situation estimation**

I. Introduction

Cyberattacks targeting IoT devices are increasing. In particular, the operations by attackers become a serious problem. Modeling the legitimate traffic is useful to detect such attacks. We have proposed a method to detect such attacks by modeling the legitimate traffic by the combination of operation sequences and home conditions [1]. However, we did not discuss the definition of the condition in detail.

In this paper, we modeled the legitimate traffic focusing on the home conditions especially on the in-home activities and applied the model to the detection of anomalous operations.

II. Modeling home IoT traffic using users' in-home activities

A. Model of in-home activities

We made a model of users' in-home activities by states of the home, state transition probabilities, and probabilities of operating devices in each state.

1) State of the home: We defined the state of in-home activities by a combination of a state of users and a state of devices. The state of users S_U is defined by the activities of the users in the home. These states are estimated from home IoT sensors. We also defined the state of devices S_O whose operations are targets of anomaly detection. In this paper, we defined four states of devices; s_I: in use, s_X: will be used within T_X minutes, s_Y: within T_Y minutes after the device's operation, and s_N: others, where T_X and T_Y are parameters.

2) State transition probabilities: In our method, we defined state transition probabilities for each pair of states. In this paper, we divided time into time slots and made a transition to the next states at each time slot. The state transition probabilities depend on the time of day. Therefore, our model includes the definition of state transition probabilities for each time of day. The transition probability $a_k(i,j)$ from the state i to the state j at the kth time slot in each day is defined as $P(S_{k-1}=i|S_k=j)$, where S_k is the state at the time slot k.

3) Probabilities of operating device: This model includes the definition of the probabilities of an operating device for each state. The probability of operating device $b(i,n)$ of device n at state i is defined as $P(n \in x_t|S_t=i)$, where x_t is the set of devices operated at time slot t.

B. How to learn the model

In this section, we explained how to learn the model from stored logs of the home activities.

1) Labeling states: We first set a label for each time slot. We divided the log data into time slots and set a label indicating the state of the home S to each time slot. The state of users S_U is set based on the sensor data according to the predefined rules. The state of the device S_O is set based on the time that the device was operated.

2) Calculating state transition probability: The probability of transition from state i to state j at time k is given by the $\frac{N_{k+1,j}}{N_{k,i}}$, where $N_{k,i}$ is the number of time slots in state i at time k in the learning data.

The time of day of the state transition depends on the day, but a similar state transition occurs at a similar time of day. Therefore, we calculated $a_k(i,j)$, considering the data from the time slot $(k-T_Z)$ to $(k+T_Z)$ for each day, where T_Z is a parameter. That is, $a_k(i,j)$ is defined as $\sum_{K-T_Z \le m \le K+T_Z} P(S_m=i|S_{m+1}=j)/D_k$, where D_k is the number of time slots in the training data.

3) Calculating probabilities of operating device: The probability $b(i,n)$ of operating device n at state i is calculated by $\sum_k (N^{(n)}_{k,i}/N_{k,i})$, where $N_{k,i}$ is the number of time slots whose state is i at time slot k in each day, and $N^{(n)}_{k,i}$ is the number of time slots whose state is i and the device n is operated at time slot k.

C. Detecting anomalous operations

The detection is based on the probabilities of the state that the device can be operated. We denoted the probability that

the state at the time slot t is estimated as i by $\alpha_t(i)$, and the observed device operations at the time slot t by x_t.

The detection system updates the probability of each state α at each time slot by the following steps.

Before starting the system, we initialize α_0 as $1/|S|$.

Firstly, we update $\hat{\alpha}_t(i)$ by using the learned state transition probability $a_k(i,j)$. The estimated current state $\hat{\alpha}_t(i)$ is obtained by

$$\hat{\alpha}_t(i) = \sum_c \alpha_{t-1}(c) a_{T(t-1)}(c,i) \tag{1}$$

where $T(t)$ is a function for obtaining the time of day corresponding to the time slot t.

Secondary, we correct $\hat{\alpha}_t(i)$ by using observed value x_t, where x_t is a set of operated devices at time slot t. By this definition, the probability $P(x_t|S_t = i)$ is obtained by

$$P(x_t|S_t = i) = \prod_n \beta(x_t, i, n) \tag{2}$$

where

$$\beta(x_t, i, n) = \begin{cases} b(i,n) & n \in x_t \\ 1 & n \notin x_t \end{cases} \tag{3}$$

Then, $\alpha_t(i)$ can be estimated by

$$\alpha_t(i) = \frac{P(x_t|S_t = i)\hat{\alpha}_t(i)}{\sum_j P(x_t|S_t = j)\hat{\alpha}_t(i)} \tag{4}$$

In our method, an operation at a time slot t satisfying $\alpha_t(s_X) + \alpha_t(s_I) > \theta$, where θ is a threshold, is regarded as legitimate.

III. Evaluation

A. Data collection

To evaluate our method, we collected data in a real home for four months from December 2018 to March 2019. We set buttons to each home appliance to record the operating time and asked the subjects living in the homes to push the button when they used it. We also deployed environmental sensors that collect temperature, humidity, air pressure, CO2 concentration, and noise value in the home and collected the sensed values.

B. Settings

In this evaluation, we focused on the cooking stoves; anomalous operations on the cooking stoves are the targets of detection. The cooking stoves are frequently used in a real home. Moreover, the anomalous operations on cooking stoves cause serious problems such as a fire. Besides, we set the length of the time slot to one minute.

1) Labeling rules: In this evaluation, we set three states of the home S_U.

- Out of home: all subjects have left
- Sleeping: the noise sensor value is less than a threshold and no devices operated
- Active: other than the above

We defined the states of devices S_O with some cooking appliances because it is better to estimate the situation at home. Cooking appliances include cooking stoves, a microwave oven, an oven toaster, and a rice cooker.

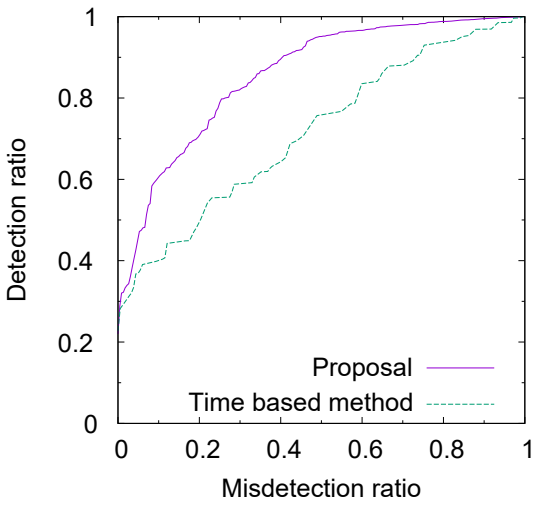

Fig. 1. Comparison with time of day method

C. Metrics

In this evaluation, we divided the data into small datasets so that each dataset includes data for each day. We set one of the small datasets as the test data, and the others as the training data. Then, we summarized the all results of each test data.

To evaluate the accuracy, we added anomalous operations of the cooking stoves in each time slot and counted anomalous operations detected by our model. We defined the detection ratio by a ratio of detected anomalous operations to added anomalous operations, and the misdetection ratio by a ratio of misdetected legitimate operations to legitimate operations.

D. Result

Figure 1 compared our method that we set parameters as $\{T_X, T_Y, T_Z\} = \{30, 30, 30\}$ with a comparing method that using only the time of day information.

The result indicates that our method achieves higher detection ratio than the compared method using only the time of day information. This is because our method accurately estimates the states that cooking stoves tend to be used.

IV. Conclusion and Future Work

In this paper, we modeled home IoT traffic based on users' in-home activities. We evaluated the detection using the model and demonstrated that our model accurately estimates the states that home IoT devices tend to be used. However, we may achieve more accurate detection by combining our model and the method using the operation sequences, which is one of our future research topics.

References

[1] M. Yamauchi, Y. Ohsita, M. Murata, K. Ueda, and Y. Kato, "Anomaly detection in smart home operation from user behaviors and home conditions," *IEEE Trans. on Consum. Electronics*, vol. 66, no. 2, pp. 183–192, May 2020.