Temporal Trends Analysis for Dengue Outbreak and Network Threats Severity Prediction Accuracy Improvement

Nurfadhlina Mohd Sharef, Nor Azura Husin, Khairul Azhar Kasmiran, Mohd Izuan Ninggal[#]
[#] Department of Computer Science
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
UPM Serdang, Selangor
Malaysia
{nurfadhlina, n_azura, k_azhar, mohdizuan}@upm.edu.my

Journal of Digital
Information Management

ABSTRACT: *Time series analysis is one of the major techniques in capturing trends and pattern of the occurrence for future forecasting. Existing but scarce works have developed temporal-based techniques which target to either predict movement (increase or decrease) or quantify the possibility of the predicted event to happen. Many of these techniques focus on the values of the time series attribute but there is no available work on dengue or intrusion logs that focus on temporal trend analysis-based on temporal relations mining. In this work the proposed technique utilize the temporal trends analysis of the observational attributes towards the pattern of the target's attribute values. In this work, we propose a new temporal trends analysis approach that uses temporal relation mining in forecasting dengue outbreak and cyber intrusion. We leverage the temporal abstractions and temporal logic to define patterns with the goal to optimize prediction accuracy. From the experiment conducted, the results showed that the proposed approach has better prediction as compared to the baseline.*

Subject Categories and Descriptors:
[COMPUTER-COMMUNICATION NETWORKS]: Security and protection

General Terms:
Network Threats, Security Prediction, Dengue outbreak, Temporal Trend Analysis

Keywords: Temporal Trends Analysis, Dengue Outbreak Prediction, Intrusion Severity Prediction

Received: 29 October 2018, Revised 12 January 2019, Accepted 26 January 2019

Review Metrics: Review Scale: 0/6, Review Score: 4.4/6, Inter-reviewer consistency: 89.5%

DOI: 10.6025/jdim/2019/17/3/122-132

1. Introduction

Time series model is a common approach to reasoning about time-based events. Allen's Temporal Logic suggested that it was more common to describe scenarios by time intervals rather than by time points, and listed thirteen relations formulating a temporal logic (before, after, meets, meet-by, overlaps, overlapped-by, starts, started-by, finishes, finished-by, during, contains, equals) [1]. This is in contrast to the Time Series Knowledge Representation (TSKR) approach which expresses the temporal concepts of coincidence and partial order by applying constraints to define the temporal relations mining [2]. Although this method appears feasible and computationally sound, it does not suit our case studies application due to the granularity of the time intervals in our datasets.

This paper reports the updated progress on the development of temporal mining relations-based approach for both the dengue outbreak and intrusion prediction. The basic idea of the proposed approach is that the numerical-based prediction is performed by utilizing the intuitive linguistics-based representation of the univariate and multivariate time series features (through temporal abstractions) in the datasets to increase the accuracy of the regression-based prediction models. In particular, the goal is to identify whether there exist a series of combination of supporting events' occurrence sequence prior to the occurrence of the target. Temporal trend analysis method allows us to order instances chronologically, generating sequential properties where adjacent timestamps usually have a higher similarity than distant ones. On the other hand, temporal data usually has a certain periodic pattern, which repeats with a certain frequency.

Our previous and similar efforts [3]–[7] were tested data within different time series from the ones highlighted in this work. This paper is also distinguished by addressing the temporal trend analysis pattern mining. Different temporal aggregation technique has also applied. We focus on the step to define a language that can adequately represent the temporal dimension of the data. We rely on temporal abstractions (Shahar, 1997) and temporal logic (Allen & Ferguson, 1994) to define patterns able to describe temporal interactions among multiple time series. This allows us to define complex temporal patterns like "the rising rainfall during northeast monsoon precede a decreasing trend in dengue cases", where we model the behaviour of 'before', 'co-occurs' which are among the relations in Allen's temporal logic (Allen & Ferguson, 1994).

After defining temporal patterns, we develop machine learning algorithm for the prediction tasks using the studied datasets. Previous models for the domains of the used datasets (i.e. dengue cases and network threats) have not considered temporal logic as part of their prediction modelling approach. The predictions models proposed utilize trend-based features which allow the monitoring of the direction of the frequency of attacks (either steady, increasing or decreasing, within the s values). Various features combinations denote that the model for the target class is based on its co-occurrence with the chosen feature combination.

The essence of this model design is by assuming that for the network threat scenario, there exist a population of network exploiter that generates exploit scripts with various purposes which cause various threat severity levels; therefore, the model design captures this distribution. The approach taken for this model is not by utilizing external factors such as the vulnerability records, incident reports and patches updates records. Meanwhile for the dengue cases prediction, the combination of the features signify the relations of number of cases distribution in the district and their impact to the district population. This paper answers the following questions:

a) What are contributing factors on the prediction model performance for dengue cases and network severity threat?

b) How to utilize temporal trend analysis mining for the prediction of dengue cases and network severity threat?

c) What is the performance of the machine learning models in the prediction of dengue case and network severity threat?

This paper is organised as follows. The first section introduces the idea of the paper followed by related works in section two. The third section describes the proposed approach while section four comprises of the results and discussion. Section five follows with the conclusion and ideas for future works.

2. Related Works

This section highlights the existing approaches related to the problems being presented in this paper namely dengue outbreak prediction and intrusion prediction.

2.1 Pattern Mining Approach for Classifying Multivariate Temporal Data

The problem of complex multivariate temporal data has been studied quite comprehensively in electronic health record systems where the focus was for the learning of classification models. In [8]–[10], temporal abstraction [11], [12] and temporal logics [1] were utilised for predicting patients who are at risk of developing heparin induced thrombocytopenia.

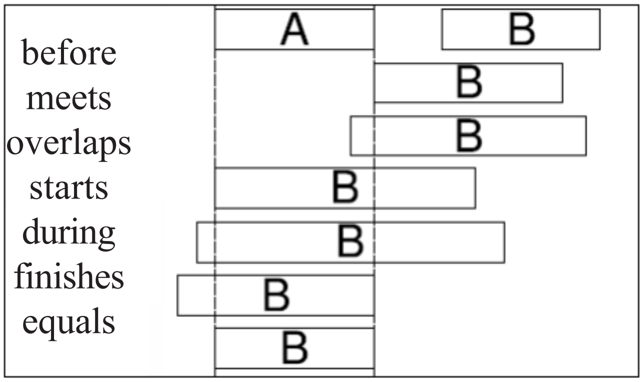

Figure 1. Allen's Temporal Logic

Temporal logic (such as Allen's as shown in Figure 1) has also been useful for the understanding and building prediction model for user action in intelligent environments [13], [14] and part of decision support for oil well drilling [15]. There are various classifiers that could be implemented for multivariate classification, such as the multilayer perceptron and support vector machine [16]. To the best of our knowledge, works that develop temporal trend analysis based models for dengue and intrusion prediction is not available and thus is addressed in this work.

2.2 Dengue Outbreak Prediction

According to the World Health Organisation (WHO), dengue is the most important mosquito-borne viral disease worldwide and is also regarded as a pandemic threat.

The dengue fever brought by Aedes mosquitoes is endemic in Malaysia with a ratio of 328.3 cases per 100,000 population; with increased risk in urban and sub-urban areas. Peak transmission occurs in the late monsoon season (October through February in east peninsular Malaysia, Sabah, and Sarawak; July through August in west peninsular Malaysia).

In-line with WHO recommendation, prevention of dengue in the country rely on integrated vector management approaches which aim at reduction of mosquito breeding sites, environmental management, and the killing of adult and immature mosquitoes. Despite tens of millions ringgit have been spent for activities such as fogging (include the human resource and equipment maintenance), larviciding, awareness and research (including genetically altered male mosquitoes) to control dengue outbreak, the country seem to be no further ahead in combating dengue. Fogging only kills the adult mosquitoes and does not kill the larvae. In addition, fogging sometimes is spared from reaching indoor, which is the favourite spot for the adult mosquitoes.

Predictability of number of dengue cases could assist in treatment planning such as preparing hospital beds and human resource and efficient dosage planning in vector control. Currently, the norms for such decision making is based on two factor namely the number of cases and effected districts locality reported to the health centres. Proactive resource planning is inaccurate if relying on this because these factors are not in linear relationship and the values are highly fluctuates.

Similarly to the approaches in other countries, the existing studies in Malaysia have focused on the week, month and district information of the dengue cases besides weather attributes such as rainfall, humidity, precipitation and wind speed (Azam, Yeasmin, Ahmed, & Chakraborty, 2016; Husam I.S., Bakar, Zainudin, Sahani, & Ali, 2017; Husin, Salim, & Ahmad, 2008; Paul & Tham, 2015; Pham, Nellis, Sadanand, Jamil, & Khoo, 2016; Review, 2013; Wongkoon, Jaroensutasinee, & Jaroensutasinee, 2012). Although warmer temperatures contribute to increased adult mosquito survival, there is no clear pattern or similarities emerged in on the effect of relative humidity on dengue rates. Other attributes such as month, age, sex, race, work, address, district office in charge, district and outbreak (Husam I.S. et al., 2017) and population (Jaafar, Abidin, & Jamil, 2016) have also been considered.

Approaches being developed in these works are time series using ARIMA and SARIMA (Review, 2013; Wongkoon et al., 2012), system dynamics (Jaafar et al., 2016) and neural network (Husin et al., 2008). The existing studies typically model specifically for the selected districts rather than providing a representative model for all districts. For example, a dengue model for Subang was developed by (Dom, Hassan, Latif, & Ismail, 2013), while (Husam I.S. et al., 2017) focused on Seremban area.

Related previous studies by (Husin et al., 2016, 2008) on the districts in Selangor namely Klang, Hulu Selangor, Sepang, Hulu Langat and Kuala Selangor based on dengue cases for the year 2004 to 2005. While agencies related to the vector control focus on the physical-based activities to eliminate the mosquitoes breeding spots, there is no available method that study the temporal pattern and the impact of the knowledge on the number of recent cases as a guide to identify areas that have higher breeding possibility.

2.3 Network Threats Severity Prediction

Attacks against computer network infrastructure are a major and persistent threat in this current globalized world. While attacks in the past have been more towards proving technical competency and one-upmanship, today's attacks trend more towards attack monetization and cybercrime. Systems such as Intrusion Detection Systems, Intrusion Prevention Systems (IPS) and Intrusion Response Systems can be used to mitigate this threat [17]. However, accurate and timely threat detection must be performed by these systems for them to be effective. This is no small feat considering the high volume of legitimate network traffic and that highly sophisticated attackers may actively attempt to obfuscate attack traffic from detectors. Using timeline data for early detection of attacks appears promising and research on this approach should be further explored.

Our previous works on threat factor profiling [3], [4] has been focusing on the sources integration, fuzzifying the threat severity, computing asset (ATL) and organisational (OTL) based threat level and developing model to predict the ATL and OTL. Presentation and visualization of the results are in multiple granularities such as hourly, daily, weekly, and so on. However, we did not explore on the prediction of threat occurrence according to the sequence of other threats' temporal trend analysis occurrence pattern.

Existing approaches that utilize time series for intrusion prediction can be categorized into binary based forecasting (i.e., increase or decrease number of attacks), and numerical based prediction and many of the models are developed by regarding the problem as anomaly detection [16]–[19]. Various machine learning-based approaches are proposed for threat prediction which utilizes factors such as causal networks [20], attacker IP [21], and patch levels [22]. In contrast, this paper focuses on the effectiveness of utilizing temporal trend analysis of the threat severity distributions to model the threat severity prediction.

3. Pattern Mining Approach For Predicting Multivariate Temporal Data

Let $D = \{<x_i, y_i>\}$ be a dataset such that $x_i \in X$ is the event records for object i up to time t_i, and $y_i \in Y$ is a numerical value (or class label) associated with attributes that relate to the event at time t_i. Our objective is to learn a function f: $X \rightarrow Y$ that can predict accurately the value (or class label) for future event. Learning f directly from X is very difficult because the instances consist of multiple irregularly sampled time series of different length. Therefore, we want to learn a space transformation ψ: $X \rightarrow X'$ that maps each instance of the event x_i to a fixed-size feature vector that preserves the predictive temporal characteristics of x_i as much as possible.

Object i, O_i is represented by a series of instances sorted according to the state sequence. We define a state to be an abstraction for a specific attribute. For example, state E: $A_i = D$ represents a decreasing trend in the values of temporal variable A_i. We define a state interval to be a state that holds during an interval, that is, state interval (E, b_i, e_i) is a realization of state E in a data instance and has specific start time (b_i) and end time (e_i).

A state sequence is a series of state intervals, where the state intervals are ordered according to their start times. After abstracting all temporal variables, we represent every instance (i.e., dengue case, attack log) in the database D as a state sequence. As a result, D can be viewed as a set of state sequences. The steps to prepare the Oi involve the transformation of the instance into a fixed time series length as follow.

(a) Sort the data according to sequence of occurrence. Define the temporal window size, s based on the temporal interval length, l in the dataset such as hourly, daily, weekly, monthly and annually. This determination could be performed based on manual observation of the time series of the data. Several combinations of temporal patterns prepared separately can be considered for experimentation purpose to identify the best interval.

(b) Determine aggregation operator, op such as summation or frequency of occurrence. This depends on the dataset. For example, in the cyber threat dataset we may be interested to know the total occurrence of each threat according to severity level, or threat category.

(c) Determine b_i and e_i for s as part of the window sliding process. Again, several combinations could be considered. The data recorded within the window size would be aggregated according to suitable operator. For example, if the $l = \text{hour}$, $s = 2$ and assuming there is no missing data in the dataset, in the first setting, $b_i = 1$ (e.g., between time 12.00 AM until 1.59 AM) and $e_i = b_i + 2 = 3$, for the first instance and the values are the aggregated values as explained in step (b). The second instance would be comprised of data between $\text{hour} = 3$ and $\text{hour} = 5$, and so on. In the second setting, the $b_i = 2$ and $e_i = 4$. Note that some data may be eliminated according to the window sliding. If the dataset contains missing data, decision and action on whether to replace the missing data with some values based on techniques for missing data treatment (e.g most frequent values, most recent, mean or median). In this work the dataset is processed as is and without assuming any missing data since the data that we are handling are provided by suitable authorities and for the purpose of reporting occurrence. Therefore, for example if $s = 3$ and the first instance has value $\text{hour} = 10$, the second instance has $\text{hour} = 13$, the value of hour in the third until twentieth instance is 14, while the value of hour in the twenty first instance is 15, and the operator is frequency of occurrence, the transformed value for the first time series would comprise of the total frequency of data from the first until the twentieth instance.

(d) Then, time series variables are converted into representation of temporal trend according to a higher level description using temporal abstractions. Typically, the temporal trend is identified based on the numerical values between one temporal series with another. The time lab series is segmented by comparing the value of the current instance with the most recent value for the considered attribute. In this work the following abstractions are used: Decreasing (Desc), Steady (Stea) and Increasing (Inc), i.e., $\Sigma = \{Desc, Stea, Inc\}$. For example, the (aggregated value) for attribute number of dengue case in the first instance is 10 and 12 in the second instance. Therefore, the temporal abstractions to be placed in the second instance is 'Inc'.

According to the instances of O_i, the temporal trend analysis can be represented according to before (b) and co-occurs (c), where

Given two state intervals E_i and E_j, (E_i, b_j, e_j)
before (E_i, b_j, e_j) if $e_i < b_j$ and (E_i, b_i, e_j)
co-occurs with (E_i, b_j, e_j), if $b_i \leq b_j \leq e_i$, i.e. E_i starts before E_j and there is a nonempty time period where both E_i and E_j occur.

Then, a standard machine learning method can be used to learn function f.

4. Results and Discussion

This section provides the discussion of the prediction task for the chosen datasets. The experiment using the first dataset focuses on the effectiveness of the temporal-based pattern mining by comparing between the Root Mean Squared Error (RMSE) of the Multilayer Perceptron (MLP) and support vector machine for regression (SMOReg [23]) algorithms. For the MLP, the learning rate is 0.3 and momentum is 0.2. For SMOReg, the Polynomial kernel is used where $K(x, y) = <x, y>$^p or $K(x, y) = (<x, y>+1)$^p and $exponent = 1.0$, while the RegSMOImproved is used as regular optimizer with $epsilon = 1.0$ E-12, $epsilon$ $parameter$ of the $epsilon$ $insensitive$ $loss$ $function = 0.01$, $seed = 1$, $tolerance = 0.001$ and $variant = variant1$ is used. The numerical data is normalized between the scales 0 to 1 for result optimization. The Microsoft Excel 2010 and Weka tool version 3.9 are used for the data analysis and model development and testing activity. The setting for validation data is as shown in the tabulated result. The values of root mean squared errors are used as accuracy evaluation of the model's performance.

4.1 Dengue Outbreak Prediction

The dengue dataset is originally prepared by the Ministry of Health (MoH), Malaysia which was published in the open data portal to provide information about the statistics of dengue cases in the districts of each state in Malaysia. The dataset contains four attributes namely Year, Week (Wk), number of locality with dengue cases in that district (NL) and number of case (NC). The models are developed to predict NL and NC.

This dataset is chosen to investigate the performance of the proposed approach in univariate time series data and investigate the performance of the machine learning model in various experiment settings. For this purpose, the records of dengue cases in 2010 and 2011 for Hulu Langat, which is in Selangor, Malaysia is chosen as a continuation to the previous studies [7], [24] and therefore the rainfall (RF) and temperature (TMP) attributes are included. Since Year, Week and NL are released by MoH, these attributes are considered as the compulsory attributes in our experiment setting. The examples of data is as shown in TABLE 1. The value of l = week, and *s* = 1 are used so that comparison against our previous works could be made.

The performances of new attributes being investigated are (i) monsoon (MSN) with values i.e., northeast, southeast, inter1, inter2, and (ii) Number of recent cases (NRC), which captures the number of the most recent case in that district. TR denotes the trend of rainfall while TT denotes the trend of the temperature. The performances of the combination of attributes are being investigated with their setting as shown in TABLE 2.

An experiment that utilizes RMSE to compare the performance of the regression-based prediction models through the combination of various features are performed using 2 datasets. The first dataset is the Hulu Langat's dengue records collected in 2010 and 2011 while the second dataset is the records in 2013 and 2014. A 70:30 split ratio for training:test is prepared. The experiment is also conducted to compare the performance of the models in the datasets.

TABLE 3 shows the results of the NL and NC for dengue outbreak prediction for Hulu Langat's records in 2010 and 2011. For NL prediction using MLP the best setting is Setting 10 (which combines all features) while for SMOReg, the best setting is setting 5 (which is just the combination of week number, TT and TR). Meanwhile, for NC, the Setting 4 is the best for MLP and Setting 8 is the best for SMOReg. However, considering the prediction score of both NL and NC, Setting 10 is the best for MLP and Setting 8 is the best for SMOReg. These results indicate that the temporal trend analysis mining approach contributes towards the improvement of the prediction score.

Week	NL	NC	NRC	RF	TMP	monsoon
10	7	106	103	182	26	northeast
11	3	32	106	182	26	northeast
12	6	82	32	182	26	northeast

Table 1. Dengue Cases Records In 2010 For Hulu Langat

Setting	Wk	NRC	RF	TMP	MSN	TR	TT
1	Y						
2	Y	Y					
3	Y		Y	Y			
4	Y		Y	Y	Y		
5	Y					Y	Y
6	Y				Y	Y	Y
7	Y	Y	Y	Y	Y		
8	Y	Y				Y	Y
9	Y	Y			Y	Y	Y
10	Y	Y	Y	Y	Y	Y	Y

Table 2. Attributes Combination for Dengue Outbreak Prediction

Setting	MLP		SMOReg	
	NL	NC	NL	NC
1	.2831	.4335	.3826	.47
2	.2733	.4064	.3488	.4295
3	.3363	.4668	.4583	.5354
4	.2976	.318	.3756	.565
5	.2969	.3429	.3819	.4737
6	.593	.6484	.3568	.4244
7	.3436	.4192	.3214	.4958
8	.4235	.3741	.2207	.2704
9	.2678	.4001	.3604	.5098
10	.2523*	.3349*	.3319	.4895

Table 3. Results of Dengue Outbreak Prediction Model for Hulu Langat for Year 2010 and 2011 Based on 70% Training and 30% Testing Split

Referring to TABLE 4 for NL prediction using MLP the best setting is Setting 7 (which does not utilize TR and TT) while for SMOReg, the best score is in setting 6 (which is just the combination of week number, TT and TR). When all features are utilized, the best score is obtained, through SMOReg. Although setting 10 is not the best in MLP, because its performance for NL case is poor, its achievement for NC prediction is the top. In considering the performance of the prediction model for both NL and NC, setting 5 is the best (obtained by ranking the summation of the RMSE scores for both prediction targets) when modelled with MLP. As compared to the baseline which is in Setting 1, the improvement of Setting 10 through MLP is 10.88% for NL and 22.7% for NC. Meanwhile, for SMOReg, NL prediction is improved by 13.25% for NL and 4.14% for NC. Results in MLP model also indicate that when the temporal trend abstraction is utilized, as in Setting 5, the performance of the prediction is higher compared to numerical based values in TMP and RF as used in Setting 3 and TMP, RF and monsoon as in Setting 4.

The results from both dengue cases datasets indicate that the temporal trend analysis mining approach provides advantage over the baseline (Setting 1). The combinations of the some of the features with temporal abstraction features also achieve higher results although inconsistently. Between MLP and SMOReg, the MLP is the better model for both datasets. Comparing between the results of Dataset 1 and Dataset 2, the performance of the regression model is better in dataset 1. This could be because the number of cases in Dataset 2 is bigger, and the values fluctuation rate is higher which suggests relations to the weather instability. This is because the dengue morphology is affected by the weather (in Dataset 2 the total of temperature is bigger than in Dataset 1, while the total rainfall is less).

4.2 Network Threat Prediction

The objectives of this experiment are to (i) investigate the performance of the temporal abstractions features, (ii) identify the impact of various training and testing split, and (iii) compare the performance of the regression models. The purpose of the model is so that the public university could be warned against the possible threat occurrence that it would experience prior to receiving the attack, specifically the non-completed multistage attacks that repeat their attempts. The output of the model is used as rules to be embedded in the sensor so it could alert the operation team when certain threat threshold is met. The idea is that if the organization can predict the threat severity level, it can handle these attacks by blocking those network connections at the switch level. For example, during a critical operational time, a stakeholder could restrict the access of potential attackers by blocking the entire subnet.

This experiment involves historical logs of network intrusion by the IPS subscribed in a public university which are unfiltered (i.e., logs of network access that may poses threat but the signature of the threat does not match with those in the IPS so the threat is assigned with label either ‘Allowed’, meaning no suspect, ‘Permit’, for logs from source that has been recorded as attacker previously, ‘Blocked’, which means the threat is totally not permitted to access the targeted destination). The IPS logs, (examples as shown in TABLE V), were collected within three months (total of 2161 instances). The distribution of the values inside the table indicates the difficulty to observe the pattern of the threat logs.

After analyzing the network access, the IPS labels it according to its severity level such as ‘Low’, ‘Info’, and ‘High’ (sorted according to increasing level of severity). The ex-

Setting	MLP		SMOReg	
	NL	NC	NL	NC
1	.4476	.3868	1.0972	2.8253
2	.3759	.6081	.7824	2.5188
3	.3305	.3828	.5725	1.5075
4	.3135	.2957	.6748	.8259
5	.2925	.2735	1.3781	1.5148
6	.3262	.3121	.3864	.6674
7	.2462	.4718	.4789	1.145
8	.7409	.2302	.4116	.4381
9	.5088	.2939	.4165	.981
10	.4346	.291	.3841	.4821

Table 4. Results of Dengue Outbreak Prediction Model for Hulu Langat for Year 2013 and 2014 Based on 70% Training and 30% Testing Split

Day	Hour = 0				Hour = 1			
	High	Info	Low	Tot	High	Info	Low	Tot
1	16	3	0	19	3	0	4	7
2	5	0	6	11	10	0	4	14
3	0	0	6	6	0	0	0	0
4	0	0	0	0	174	8	0	182
5	1	0	0	1	0	0	4	4

Table 5. Examples of Number of IPS Threat Logs Hourly For 5 Days According to Severity Level

Setting	time	TH_s=1	TH_s=2	TH_s=11	TH_s=12	TI_s=1	TI_s=2	TI_s=11	TIS_s=12	TL_s=1	TL_s=2	TL_s=11	TL_s=12
1.	Y												
2.	Y			Y				Y				Y	
3.	Y				Y				Y				Y
4.	Y		Y				Y				Y		
5.	Y	Y				Y				Y			
6.	Y	Y		Y		Y		Y		Y		Y	
7.	Y	Y	Y			Y	Y			Y	Y		
8.	Y	Y	Y	Y		Y	Y	Y		Y	Y	Y	
9.	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
10.	Y												
11.	Y			Y				Y				Y	
12.	Y				Y				Y				Y
13.	Y	Y				Y				Y			
14.	Y	Y	Y			Y	Y			Y	Y		

Table 6.Attributes Combination for IPS Threat Prediction

amples of threat categories labeled by the IPS under 'High' are 'HTTP_CSH-Apache-HTTP-Server-Mod_rpaf-X-Forwarded-For-Denial-Of-Service', 'Generic_MySQL-MaxDB-WebDBM-BOF', 'HTTP_CRL-D-Link-Unauthenticated-Remote-Command-Execution' and 'File-Filtering-Policy_Buffering-Limit-Exceeded'. The decision of the IPS to block, permit or allow the access is not according to the severity level. Thus, allowing an access could expose the organization to damage and predicting the possibility of the occurrence of the intrusion according to the severity level is essential, and addressed in this experiment. Although among the preliminary consideration for the experiment is to focus on the source IP of the intrusion, since the nature of IP spoofing and masking therefore the setting is flattened to severity level according to the asset.

Besides using different dataset and domain, the setting of this experiment is distinguished by its focus. In the previous experiment, only setting 70-30 is executed. In this experiment, the impact of the training and testing ratio is observed, besides monitoring the performance of the regression models when various values of *s* are used.

The features of the predictions models are as shown in TABLE 6. The regression models are developed for the prediction of number of firewall threat with type High (nH), IPS threat with type Info (nI) and firewall threat with type Low (nL). In the development for the IPS data using Dataset 3, 4, 5 and 6, *op* is set as the frequency of the threat cases and *l*=hourly while two values for *s* are used as data aggregation with several combinations of features.

For Dataset 3, $s = 1$ means total number of each threat type occurrence and its trend of previously 1 hour of the same day, $s = 2$ means total number of occurrence and trend of previously 2 hour of the same day, while $\mathrm{s}=11$ indicates trend of the threat type at the same hour, the day before and $s = 12$ indicates trend of the threat type at the same hour, two days before. TH denotes the trend of firewall threat with $\mathrm{severity} = \mathrm{High}$, TI denotes the trend of firewall threat with $\mathrm{severity} = \mathrm{Info}$ and TL denotes the trend of firewall threat with $\mathrm{severity} = \mathrm{Low}$. Setting 1 until 9 is used in the experiment with 70% training and 30% percentage testing data split while setting 10 until 14 is used in the experiment with 30-70, 20-80 and 10-90 data split.

The results for the prediction of nH, nI and nL as shown in TABLE 7 indicate that the combination of all features (Setting 9) achieve the best performance when modeled using SMOReg. Meanwhile, in MLP Setting 9 is at the third-best score. The results also indicate that for the prediction of nH using MLP, the best results is obtained by Setting 8 (combination of trend of previous 1 and 2 hours, and the trend comparing to the same hour during previous day) while for SMOReg, Setting 9 is the best. Meanwhile, for nI, the best model is through Setting 7 (combination of trend from previous 1 and 2 hours on the same day) in MLP and Setting 8 in SMOReg. The best regression model for nL is through Setting 5 (trend of previous 1 hour on the same day) using MLP and Setting 5 or 9 using SMOReg These results indicate that Setting 8 and 9 are the best feature representation, when the model is developed using either SMOReg or MLP. However, comparing between 8 and 9, since the total of the RMSE score for all three targets is lower when setting 9 is used, this indicates that exposure of the prediction model for longer duration would be more useful to ensure high accuracy.

The next experiment explores on the impact of the prediction score in different training and testing setting. Only selected features setting are observed, which are identified based on their good performance in the prediction through 70:30 training:test ratio model development.

TABLE 8 shows the RMSE scores of the IPS threat predictions using the settings that have achieved high score in the previous experiment. Results indicate that for nH, the best prediction performance is through Setting 14 (previous 1 and 2 hour on the same day) using MLP. Meanwhile, for nI, Setting 10 (baseline, using time feature only) is the best, through SMOReg. The NL prediction modeling is the best through Setting 14 using MLP. However, considering the multi-prediction performance of nH, nI and nL, Setting 12 (trend comparing to the same hour on previous 2 days) is the best, through SMOReg. These observations indicate that, when the model is developed

Setting	MLP			SMOReg		
	nH	nI	nL	nH	nI	nL
1	.1423	.0887	.0143	.0709	.0637	.047
2	.0791	.0771	.0143	.0702	.063	.044
3	.0928	.144	.14646	.0699	.0632	.0446
4	.8408	3.1179	5.9046	.0696	.0628	.0436
5	.1095	.2576	.1763	.0696	.0623	.0432
6	.1299	.4968	.4055	.0704	.0631	.0446
7	.12	.1172	.5971	.0691	.0624	.0446
8	.1123	.1963	.6787	.0693	.0621	.0452
9	.1539	.266	.1861	.0683	.0623	.0432

Table 7. Results of IPS Threat prediction in setting 70-30 (Dataset 3)

with 30:70 ratio, the recent observations on the same day is more influential. The next experiment observes the performance of the regression model developed using ratio 20:80 (training:testing).

The results of experiment on Dataset 5 as shown in TABLE IX, indicate that Setting 13 (previous 1 hour same day) is the best features combination for multi-class prediction of nH, nI and nL, using both regression models. For nH, the best model is developed using Setting 13 and MLP, while for nI, Setting 12 (same hour, last two days) is the best using SMOReg. Setting 13 also returns the best achievement for nL prediction, using SMOReg. These suggest that for Dataset 5, utilizing the trends of the threat on recent one hour gives the best prediction performance. This is also applicable for multi-class prediction, using Setting 13 and MLP.

The results of the experiment on Dataset 6 unsurprisingly confirm that bigger exposure training exposure is needed for the regression model development. The results also suggest that Setting 12 in both regression models is the best setting. This setting also has a huge gap compared to the other settings. This indicates that the features in this setting are very distinguishable to the regression models. These results also suggest that the worst performance is obtained when the combination of all features are used (Setting 14), especially when modeled using MLP.

The experiment using dataset 3 until 6 results suggest that the temporal based mining approach has a superior performance against the baseline approaches. Suitable *s* value should be explored according to the dataset requirements and setting. Based on this experiment, observation of the previous 1 and 2 hours on the same day compared to earlier trend's prediction provides influential impact on the regression models performance. The training exposure ratio should also be bigger than the testing portion. This hints that the incoming IPS threat is unpredictable and early warning could not be consistently successful just by utilizing the suggested features.

5. Conclusion

Temporal data mining aims at finding interesting correlations or sequential patterns in sets of data stream. In this paper, we present two applications occurring in the fields of medical and cybersecurity. The given examples share one aspect in common: dealing with different models of incorporating the temporal aspects. It is not only the fact to find patterns inside data volumes but also to identify them based on their temporal behaviour.

The temporal methods offer intuitive methods of guiding the user through the process of data and model inspection and assist in drawing conclusions. The results of the

Setting	MLP			SMOReg		
	nH	nI	nL	nH	nI	nL
10	.113	.061	.1517	.0748	.0494	.03237
11	.6596	1.4145	1.9812	.0754	.162	.1998
12	.0911	.0922	.3822	.064	.058	.152
13	.0722	.1728	.2379	.0493	.0595	.1506
14	.0471	.0946	.1044	.1049	.2927	.3205

Table 8. Results of IPS Threat prediction in setting 30-70 (Dataset 4)

Setting	MLP			SMOReg		
	nH	nI	nL	nH	nI	nL
10	.2795	.4192	.1593	.3175	.2546	.3092
11	.1126	.2125	.3033	.1341	.1947	.454
12	.1028	.3067	.3245	.2427	.0608	.0461
13	.0669	.084	.1739	.1299	.0998	.107
14	.1221	3.2491	9.9939	.1049	.2927	.3205

Table 9. Attributes Combination For IPS Threat Prediction in setting 20-80 (Dataset 5)

Setting	MLP			SMOReg		
	nH	nI	nL	nH	nI	nL
10	12.6969	1.4144	8.6919	10.2593	4.079	3.5335
11	20.2908	5.8897	6.2472	5.3184	2.3196	1.2217
12	.2378	.0788	.21	1.2063	1.7623	.7617
13	1.518	.1406	.3505	4.0036	2.2652	.425
14	4.8734	46.8043	16.3847	2.2785	1.4711	2.5823

Table 10. Attributes Combination For IPS Threat Prediction in setting 20-80 (Dataset 6)

experiments indicate that the window sliding size influence the performance of the model, and specific configuration with regards to the dataset domain should be performed. The results on various training and testing ratio also indicate that more portions for the training would help the regression model to perform more reliably.

The future works that could be considered are to study the performance of the approach to more comprehensive dengue and intrusion logs. Whilst in this work the relations between the observational attributes are exploited, another approach that could be considered is to study the sequential relations series such as more details on (i) multiple events temporality such as the combination of day and hour to observe whether there exist a similar pattern of occurrence between a specific hour across several days and (ii) multiple event order such as overlapping, start-after and end-by, and so on.

Acknowledgment

The authors would like to thank several individuals who have contributed indirectly and indirectly to the paper such as Dr. Maslina Zolkepli for the discussion on possibility to develop prediction model for water quality preservation. Another note of appreciation goes to Mrs. Norhashimah Mat Sejani and Dr. Mohd Taufik Abdullah for supporting with the intrusion data preparation and discussion. A special thank is dedicated to the cyber threat profiling research team for various meetups which has inspired the approach introduced in this paper.

We would like to thank Causal Productions for permits to use and revise the template provided by Causal Productions. Original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com).

References

[1] Allen, J. F., Ferguson, G. (1994). Actions and Events in Interval Temporal Logic, *J. Log. Comput.*, 4 (5).

[2] Mörchen, F. (2006). A better tool than Allen's relations for expressing temporal knowledge in interval data, *In:* Proceedings the Tweleveth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[3] Sidi, S. R., Daud, F., Ahmad, M., Zainuddin, S., Abdullah, N.,Jabar, S. A., Affendey, M. A.,Ishak, L. S., Sharef, I.,Zolkepli, N. M., Nordin, M., Sejani, F. N. M., Hairani, H. A. (2017). Towards an Enhancement of Organizational Information Security through Threat Factor Profiling (TFP) Model, *Towar. an Enhanc. Organ. Inf. Secur. through Threat Factor Profiling Model*, 892 (2017) 1–8, 2017.

[4] Sidi, R. A., Marzanah, F., Affendey, A.J.,Ishak, L.S., Sharef, I., Zolkepli, N.M., Ming, M., Mokthi, T.M., Daud, M.F.A., Zainuddin, M., Hamid, N.B. (2017). A comparative analysis study on information security threat models: A propose for threat factor profiling, *J. Eng. Appl.Sci.*, 12 (3) 548–554.

[5] Husin, N. A., Mustapha, N., Sulaiman, N., Yaakob, R. (2012). "A Hybrid Model using Genetic Algorithm and Neural Network for Predicting Dengue Outbreak, *In*: 2012 4th Conference on Data Mining and Optimization (DMO), 2012, no. September, pp. 23–27.

[6] Husin, N. A., Salim, N., Ahmad, A. R. (2008). "Modeling of Dengue Outbreak Prediction in Malaysia: A Comparison of Neural Network and Nonlinear Regression Model, *In*: ITSim 2008. International Symposium on Information Technology, 2008.

[7] Husin, N. A., Mustapha, N., Sulaiman, M. N., Yaacob, R., Hamdan, H., Hussin, M. (2016). Performance of hybrid GANN in comparison with other standalone models on dengue outbreak prediction, *J. Comput. Sci.*, 12 (6) 300–306.

[8] Batal, I., Valizadegan, H., Cooper, G. (2011). A Pattern Mining Approach for Classifying Multivariate Temporal Data, *(Bibm), 2011 IEEE*, p. 1–21, 2011.

[9] Batal, I., Valizadegan, H., Cooper, G. F., Hauskrecht, M. (2013). A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data, *ACM Trans. Intell. Syst. Technol.*, 70 (4) 646–656.

[10] Batal, I., Cooper, G. F., Fradkin, D., Harrison, J., Moerchen, F., Hauskrecht, M. (2016). An Efficient Pattern Mining Approach for Event Detection in Multivariate Temporal Data, *Knowl. Inf. Syst.*, 46 (1) 115–150, 2016.

[11] Shahar, Y. (1999). Timing is Everything: Temporal Reasoning and Temporal Data Maintenance in Medicine, *In*: Proceedings Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99, 1999, p. 30–46.

[12] Shahar, Y. (1997). A framework for knowledge-based temporal abstraction, *Artif. Intell.*, 90 (1–2) 79–133.

[13] Daemen, A., *et al.* (2013), Modeling precision treatment of breast cancer., *Genome Biol.*, 14 (10) R110, Oct. 2013.

[14] Jakkula, V. R., Cook,D. J. (2007). Learning Temporal trend analysis in Smart Home Data, *In*: Proc. Second Int. Conf. Technol. Aging.

[15] Jære, M., Aamodt, A., Skalle, P. (2002).Representing temporal knowledge for case-based prediction, *In*: European Conference on Case-Based Reasoning, 2002, p. 174–188.

[16] Jabez, J., Muthukumar, B. (2015). Intrusion detection system (ids): Anomaly detection using outlier detection approach, *Procedia Comput. Sci.*, 48, no. C, pp. 338–346, 2015.

[17] Wu, S. X., Banzhaf, W. (2010). The use of computa tional intelligence in intrusion detection systems: A review, *Appl. Soft Comput. J.*, 10 (1) 1–35, 2010.

[18] Palanivel. (2014). Multi Scale Time Series Prediction for Intrusion Detection, *Am. J. Appl. Sci.*, 11(8) 1405–1411, 2014.

[19] Kholidy, H. A., Erradi, A., Abdelwahed, S. (2014). Attack Prediction Models for Cloud Intrusion Detection Systems, *In*: Proc. - 2nd Int. Conf. Artif. Intell. Model. Simulation, AIMS 2014, no. 1, p. 270–275, 2014.

[20] Qin, X., Lee, W. Attack Plan Recognition and Prediction Using Causal Networks.

[21] Nanda, S., Zafari, F., Decusatis, C., Wedaa, B. Yang, (2017). Predicting network attack patterns in SDN using machine learning approach, *In*:2016 IEEE Conf. Netw. Funct. Virtualization Softw. Defin. Networks, NFV-SDN 2016, p. 167–172, 2017.

[22] Liu, Y., Zhang, J.,Sarabi, A., Liu, M., Karir, M. Bailey, M. (2015). Predicting Cyber Security Incidents Using Feature-Based Characterization of Network-Level Malicious Activities, *In*: Proc. 2015 ACM Int. Work. Int. Work. Secur. Priv. Anal. - IWSPA '15, p. 3–9, 2015.

[23] Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K. (1999). Improvements to the SMO Algorithm for SVM Regression, *IEEE Trans. Neural Networks*, 1999.

[24] Husin, N. A., Salim, N., and Ahmad, A. R. (2008). Modeling of dengue outbreak prediction in Malaysia: A comparison of neural network and nonlinear regression model,*In:* Proc. - Int. Symp. Inf. Technol. 2008, ITSim, vol. 4, p. 6–9, 2008.