2016 IEEE 24th International Conference on Network Protocols (ICNP)
Workshop on Machine Learning in Computer Networks (NetworkML 2016)

MS-LSTM: a Multi-Scale LSTM Model for BGP Anomaly Detection

Min Cheng[1], Qian Xu[1], Jianming Lv[2], Wenyin Liu[3]*, Qing Li[1]*and Jianping Wang[1]
[1]Department of Computer Science, City University of Hong Kong
[2]School of Computer Science and Engineering, South China University of Technology
[3]School of Computer Science and Technology, Guangdong University of Technology

ABSTRACT

Detecting anomalous Border Gateway Protocol (BGP) traffic is significantly important in improving both security and robustness of the Internet. Existing solutions apply classic classifiers to make real-time decision based on the traffic features of present moment. However, due to the frequently happening burst and noise in dynamic Internet traffic, the decision based on short-term features is not reliable. To address this problem, we propose MS-LSTM, a multi-scale Long Short-Term Memory (LSTM) model to consider the Internet flow as a multi-dimensional time sequence and learn the traffic pattern from historical features in a sliding time window. In addition, we find that adopting different time scale to preprocess the traffic flow has great impact on the performance of all classifiers. In this paper, comprehensive experiments are conducted and the results show that a proper time scale can improve about 10% accuracy of LSTM as well as all conventional machine learning methods. Particularly, MS-LSTM with optimal time scale 8 can achieve 99.5% accuracy in the best case.

1. INTRODUCTION

The Border Gateway Protocol (BGP) is protocol widely used in the Internet for Autonomous Systems (AS) to exchange routing and reachability information. To ensure the stability and security of the Internet, BGP update packets are closely monitored to detect anomalous events. An anomalous event of BGP may affect a large scale of users. For instance, on February 24, 2008, Pakistan Telecom (AS17557) started an unauthorized announcement of the prefix 208.65.153.0/24 and then its upstream providers, PCCW Global (AS3491), forwarded this announcement to the rest of the Internet, which resulted in the hijacking of YouTube traffic for two hours on a global scale. Apart from prefix hijacks, worms, misconfigurations, and electrical failures happen frequently. Capturing such events with high accuracy and low delay is the key to keep the Internet running.

Some previous works have been devoted to classify BGP traffic data into normal and anomalous. [1] designs signature-based detection and statistics-based method. [2] employs an instance-learning based framework using wavelet-transformation and clustering in pattern extraction to detect BGP-routing anomalies. More recently, [3] applies Support Vector Machine (SVM) and Hidden Markov Models (HMMs) algorithms with new selected features in classification and shows 81.5% accuracy in the optimal case.

Unfortunately, these existing anomaly detection methods have a number of limitations: (1) All the methods generally select the traffic features of present to make the decision regardless of the time series of the traffic data, where time series analysis can bring extra important information in identifying state changes. (2) Statistics-based techniques [2] assume the dataset follows a certain distribution and need domain knowledge such as threshold parameters. However, the regular traffic is random and it's difficult to decide fixed parameters of the traffic model. Thus, statistics-based techniques are not practical in real use.

Considering the above limitations, we propose to adopt long short-term memory (LSTM) model for BGP anomaly detection. Long Short-Term Memory is a recurrent neural network architecture proposed by Hochreiter [4] in 1997. LSTM network can handle long time series sequence data and outperforms alternative recurrent neural networks (RNN) and hidden markov models (HMMs) in numerous applications like handwriting recognition [5], speech recognition [6], and some other artificial intelligence cases. To the best of our knowledge, this is the first empirical study using LSTM to classify anomalies given multivariate time series extracted from BGP traffic.

Furthermore, we find that the time series of BGP traffic exhibit different distinct patterns in different time-scale. Selecting a proper time-scale may help the classification model to perform even better. Based on this observation, we integrate the time scale property into LSTM model, which is called the MS-LSTM model in this paper, and verify its performance in multiple time scale. Experiments show that MS-LSTM with the optimal time scale can achieve 99.5%of accuracy, which

*correspondence to Wenyin Liu and Qing Li at liuwy@gdut.edu.cn, itqli@cityu.edu.hk

978-1-5090-3281-5/16/$31.00 ©2016 IEEE

excels all the existed methods in BGP traffic anomaly detection.

In summary, the contributions of this work are organized as follows:

- We propose to adopt MS-LSTM, a multi-scale LSTM model, for BGP anomaly detection. The proposed MS-LSTM model can achieve 99.5% accuracy in BGP anomaly detection with much lower false alarm rate compared with the traditional methods.
- We empirically show that applying optimal and time scale to the existing classification model in BGP anomaly detection can improve their performance by 10%. This will provide a new perspective to improve the traditional classification methods when handling dataset with temporal information.

The rest of this paper is organized as follows: In section 2, we present a brief literature review which is related to our work. In section 3, we analyze the time series of BGP data. In section 4, we introduce our proposed model MS-LSTM for anomaly detection. The experiments are presented in Section 5. Finally, in section 6, we conclude our work.

2. RELATED WORK

Various methods have been proposed to identify the anomalies by analyzing traffic patterns. One of early and common methods is developing traffic behavior model based on statistics pattern and signal processing techniques such as cumulative sum over a time window [7], where the anomalies are identified as correlated abrupt changes occurring in the underlying distribution. However, the disadvantage is that it has difficulty in estimating the dimension distributions with all possible cases. Another widely used method is rule-based method, which is applying Internet Routing Forensics (IRF) to classify anomalies [8]. And the drawback is that it requires priori knowledge and high degree of computations.

Recently, many machine-learning methods have been employed to build traffic classification models and predict anomaly. Both unsupervised and supervised machine learning models are built to detect anomalies. [9] identifies anomaly with non-stationary traffic in networks. In [10] they propose one-class neighbor machine algorithm and recursive kernel based online anomaly detection method [11] to detect anomalous network dynamics. The Naive Bayes (NB) estimators are used to categorize the traffic flows [12], and in [3] they employ Support Vector Machine (SVM), Hidden Markov Model (HMM) and features selection methods [13] to detect the BGP anomalies. Those machine-learning models have achieved desirable performance. However, they only treat the input instances independently without considering the sequence nature of traffic data. In reality, the traffic data are multi-variant time series and the anomaly patterns vary gradually with the temporal information. Besides, those traditional machine learning methods are not designed for sequence classification and not suitable for anomaly detection in time series.

Recurrent Neural Networks (RNNs) is a popular deep learning model often used for sequence classification. RNNs, especially those based on Long Short-Term Memory (LSTM), achieves start-of-art performance in many sequence classification scenarios such as language processing, handwriting recognition, and image captioning. The most recent work applying LSTM in anomaly detection includes [14] [15], but those works are preliminary without considering much more information of characteristic and requirements in time series.

3. TIME SERIES ANALYSIS

In this section, we analyze the time series of real BGP traffic data to show the importance of detecting anomaly based on historical features rather than the data at present moment. We also illustrate how the time scale affects the pattern recognition while processing the time series.

As reported in previous research [16], the Internet traffic is highly dynamic and full of bursts and noises, which is consistent with our observation on BGP traffic. As illustrated in Figure 1, we select samples which are near the time when anomaly happens and display how the traffic features, maximum AS path length, changes along with the time. In both of the regular and anomaly portions, fluctuations are ubiquitous. For example, in the point A of the regular portion, there is a sudden burst with a relatively high value which seems to be anomaly. The traditional anomaly detection algorithms [3] [18], which make the decision only based on the state of present moment, may judge the point A as a anomaly event with high probability. On the other hand, if we consider more about the historical traffic, we can find that most of the values in the past are in a relatively regular scope and bursts are also very frequent before. Thus we may judge that point A may be only a sudden burst. Integrating the historical information into the classifier can make the decision more cautious and more accurate. Motivated by this consideration, we propose MS-LSTM in this paper to model the traffic flow as a time sequence and learn the pattern from historical traffic in a sliding time window.

Meanwhile, like most other time sequences, the Internet traffic has multi-scale property. That means in different time scales, the sequence can exhibit different distinct patterns. Figure 2 shows one BGP traffic feature in different time scale. The sequence at scale 1 is the original one, which has the finest granularity. As mentioned before, in this micro-scale, there are a lot of bursts, which may have serious negative impact

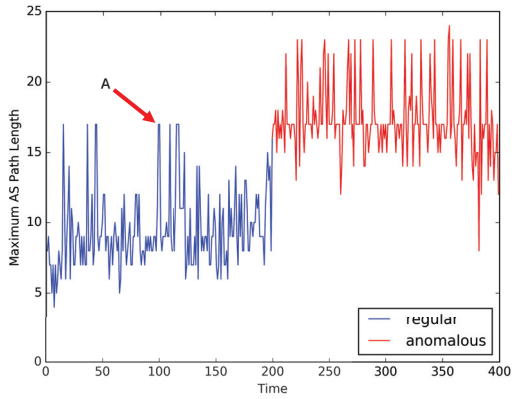

Figure 1: A time series of maximum AS path length, which is a feature of BGP traffic.

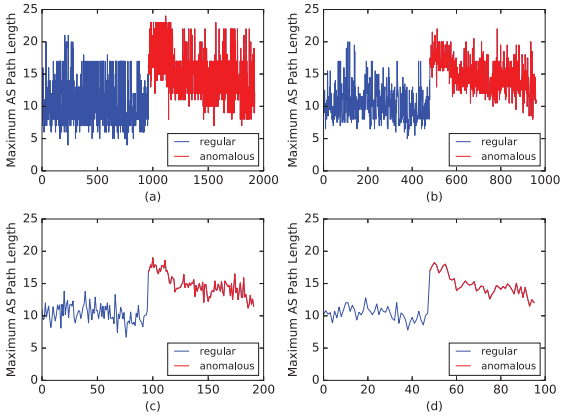

Figure 2: Time series in different time scale. (a) Time scale 1. (b) Time scale 5. (c) Time scale 10. (d) Time scale 20.

on classifiers. As we increase the time scale to $x(x = 5, 10, 20)$, which means we average the values in every x time points of the original sequence, a more smooth curve can be obtained. In a larger time scale, the global trend of the time sequence is easier to be captured, but it becomes harder to sense a local change. It is important to select a proper scale to process time series to achieve optimal prediction accuracy. In this paper, we look into the impact of different time-scales on the MS-LSTM model and achieve the best scale through comprehensive experiments on real BGP traffic data.

4. MS-LSTM MODEL

In this section, we firstly give definition of our problem and then introduce the preprocessing steps and MS-LSTM model. Finally, we give an overview of the whole pipeline.

4.1 Problem Definition

BGP update packets contain a lot of data which reflect the health of the network. In order to detect the anomalous traffic of the network, we need to analyze the pattern of historical data and train a classification model for the current identification.

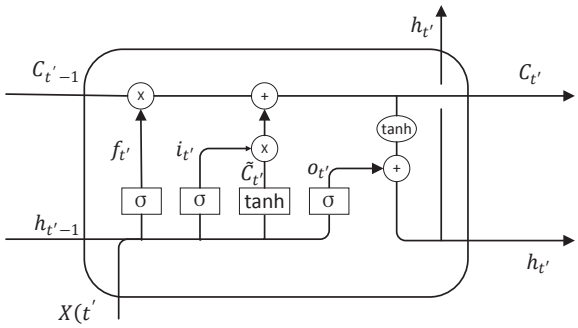

Figure 3: Structure of LSTM memory cell.

Given previous traffic features $x_{t_1}, x_{t_2}, ..., x_{t_n}$, our goal is to identify the current state of traffic $x_{t_{n+1}}$ using MS-LSTM model. Furthermore, we aim to find the optimal size of sliding sequence window e and time scale p to preprocessing the training data to achieve the best performance.

4.2 Preprocessing

Suppose that the BGP traffic data is a timely sequence $x_{t_1}, x_{t_2}, ..., x_{t_n}$ collected in n time points with the interval to be one minute. Each element is a 33 dimension vector since we extract 33 features from traffic. Assuming the size of window is e, state of x_{t_n} is related with a subsequence $S_n = (x_{t_{n-e+1}}, x_{t_{n-e+2}}, ..., x_{t_n})$. Next, each subsequence S_n is compressed with time scale p, $S_n = (d_1, d_2, ..., d_{e/p})$. The new element d is the average value of p samples in S_n, e.g. $d_1 = 1/p(x_{t_{n-e+1}} + x_{t_{n-e+2}} + ...x_{t_{n-e+p}})$. In this way, we get $n - e + 1$ sets of training data S. And the label of each set is same with the state of last vector, $L(S_n) = L(x_{t_n})$

4.3 MS-LSTM Model

Our proposed model, multi scale long-short term memory (MS-LSTM) model, is the combination of preprocessing steps and Long Short-Term Memory (LSTM) network. The key part of our model is the LSTM network. Thus, we will introduce the detail structure of LSTM model and give an overview of MS-LSTM model.

Long Short-Term Memory (LSTM) is first proposed by Hochreiter [4] in 1997 as a special form of recurrent neural networks. The tradition recurrent neural networks suffer the problem of vanishing gradient and exploding gradient during the gradient back-propagation phase when gradient signals multiply a large number of times. LSTM network overcomes the long term dependency and is suitable for the BGP traffic anomaly detection.

To model LSTM, we let S be the input of the memory cell, $C_{t'}$ be the cell state, and $h_{t'}$ be the value of output gate at time t'. $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o, V_o$ are weight matrix and b_i, b_f, b_o are bias vectors.

The key to LSTM networks is the the ability to add or

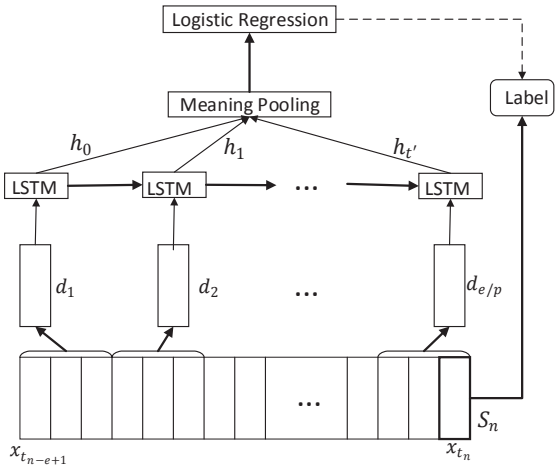

Figure 4: MS-LSTM classification model.

remove information to the cell state which is regulated by the structure called gates. Gates act like filters to let optional information through. They are composed of a sigmoid neural layer of which the output is between 0 to 1. The gates are represented as pointwise multiplication operations $\oplus$ and $\otimes$. From Figure 3, we can observe that an LSTM has three of these gates to control the information. Through an LSTM the information is transferred, filtered, combined, and finally output a overall message to the next memory cell. The detail steps of the transmission are listed below.

Step1: Throw away redundant old information.

$$f_t = \sigma(W_f \cdot [h_{t'-1}, X'_t] + b_f)$$

Step2: Store new useful information.

$$i_t = \sigma(W_i \cdot [h_{t'-1}, X'_t] + b_i)$$
$$\widetilde{C'_t} = tanh((W_c \cdot [h_{t'-1}, X'_t] + b_c))$$

Step3: Update the cell state.

$$C_t = f_t * C'_t - 1 + i_t * \widetilde{C'_t}$$

Step4: Output for next memory cell.

$$o_t = \sigma(W_o \cdot [h_{t'-1}, X'_t] + b_0),\ h_t = o_t * tanh(C'_t)$$

The LSTM model used in our experiment consists of a single LSTM layer, a mean polling layer, and a logistic regression layer. The output gate values $h_1, h_2, ..., h'_t$ are averaged in mean pooling layer and result a new h. The logistic regression layer is a binary classification layer which trains a cost function of h and label. Combined with the preprocessing of a selected S, the overview of the MS-LSTM network is showed in Figure 4. $n - e + 1$ sets of training data will be put in the MS-LSTM model to learn the pattern of continuous traffic features.

5. EXPERIMENT AND DISCUSSIONS

This section is mainly focus on the result of our experiment. We will first introduce the experiment setup, then the baseline algorithms to be compared with our method. At last, we report the evaluated results of the proposed MS-LSTM method.

Table 1: Sample of a BGP Update Packet

Field	Value
TIMESTAMP	*1128124817(2005-10-01 08:00:17)*
LEHGTH	*85*
TYPE	*UPDATE*
PEER AS	*5511*
LOCAL AS	*12654*
PEER IP	*195.66.224.83*
LOCAL IP	*195.66.225.241*
ORIGIN	*0(IGP)*
AS PATH	*5511 1239 701 702 4637 4755 9829*
NLOGREGI	*61.0.192.0/18 61.0.64.0/18*
NEXT_HOP	*195.66.224.83*

5.1 Experiment Setup

Over years, many large-scale BGP security events have been reported, among which we collect BGP traffic of three misconfiguration events and three other anomaly events from RIPE [17] to train our detection model. For each event there are several hours of abnormal records and the rest are regular. We collect the BGP update message that originated from AS 1853, 12793, 13237 (rrc05, Vienna) and AS 513(rrc04, Geneva). We develop certain tools with python to extract dynamic traffics of network data and convert the MRT format to ASCII. An sample of BGP update message with ASCII format is shown in Table1.

5.2 Baseline

We choose 3 traditional machine learning algorithms which are widely used in classification as the baseline of our experiment. Support vector machines (SVM) and Naive Bayes Classifier (NB) are used in Nabil's work [3] while Adaptive Boosting (Ada.Boost) is used in [18]. Support vector machines (SVMs) are effective supervised learning models in classification and regression analysis. Different with SVM, Naive Bayes Classifier (NB) is a probabilistic classifier based on Bayes' theorem. And Adaptive Boosting (Ada.Boost) is an iteration algorithm used to improve the performance of learning algorithm because the misclassification is retrained. These three methods are used as baselines in our later comparison section.

In our experiment, we use SVM, NB, and Boosting modules in python machine learning package scikit-learn.

5.3 Evaluation and Comparisons

To demonstrate the effectiveness of the proposed model, we firstly compare the accuracy of different methods with sequence window. After that, we evaluate the performance by using different sizes of sequence window with other variables controlled. Similarly, the impact of the size of time scale is evaluated with sequence window fixed. Then, we use cross validation to show our method is effective regardless of the type of training sets. Finally, we compare the false alarm rate and missing alarm rate of different detection models. In practice, these two measurements are much more

Table 2: Accuracy of different BGP anomaly detection models with sequence window

Method	Accuracy	
	Traditional	Window Size(40)
SVM	76.7	84.9
NB	74.9	81.4
Ada.Boost	77.5	87
MS-LSTM	-	**91.5**

important than accuracy.

5.3.1 *Improvement of adding sequence window*

The traditional algorithms in Table 2 only take selected features sets at one time point as input. The accuracy of the left part is reported from [3] and [18]. We can observe that adding sequence window can improve the accuracy of all classification methods by almost 10%. The results show that LSTM achieves the best performance than the rest, and applying sequence window is useful for classification since it takes the temporal information into consideration compared with the traditional methods.

5.3.2 *Optimal Sequence Window Size*

After learning that adding sequence window has positive effects on the performance of classification model, we try to find the optimal size of sequence window to achieve the best performance. Figure 5 is the classification accuracy with various sequence window sizes ranging from 10 to 60. The size of sequence window determines how many former samples are considered in the classification model. For different window sizes, MS-LSTM always performs better than the rest. With the increase of window size, the accuracy of all methods firstly increases and then decreases, which means there exists an optimal sequence window size. Moreover, longer temporal information does not lead to better performance. In our experiment, the optimal size is 40. Though it's not a general conclusion, we can use this as initial size in other situations.

5.3.3 *Optimal Time Scale*

With the analysis in Section 3, sample rate affects the wave motion of each feature. We use fixed sequence window size $e = 40$ and time scale value $p = 1, 2, 5, 8, 10, 20, 40$ to test the classification performance. In Figure 6, we observe that MS-LSTM reaches the best performance on the time scale 8. In our experiment, the original input is sampled in one-minute interval, thus the optimal time scale is eight minutes. When the time scale is over eight minutes, the performance of classification will decrease. On contrary, SVM, Ada.Boost, and NB achieve their best performance respectively at one-minute time scale, and the accuracy is reduced with the time scale increasing.

5.3.4 *Cross Validation*

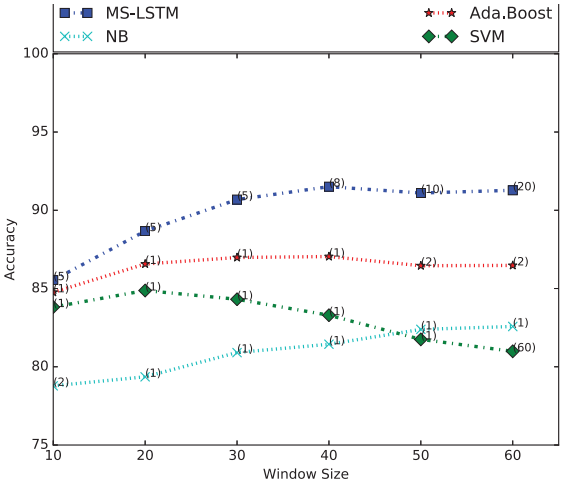

Figure 5: The accuracy with different window size.

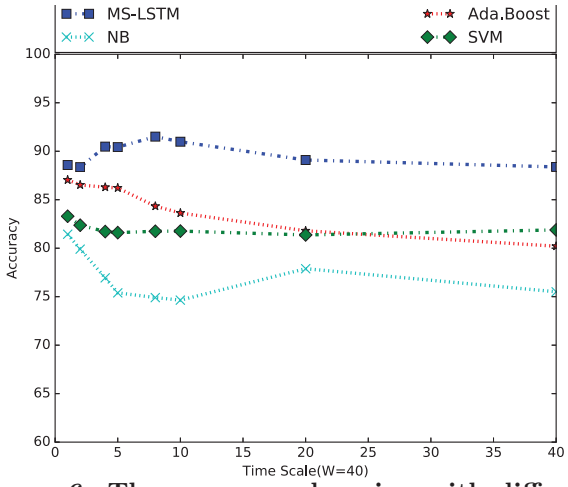

Figure 6: The accuracy changing with different time scale.

In real uses, there exist various types of anomalies and the training dataset may only contain several types. We simplify the problem by choosing three different types of worm: Code Red I, Nimda, and Slammer as whole set. We use two types of worm as training dataset and the other as the testing data set. Method 1 uses Code Red I , Method 2 uses Nimda while Method 3 uses Slammer as the testing dataset respectively. The result is showed in Table 3 and MS-LSTM performs best in almost all cases.

5.3.5 *False and Missing Alarm Rate*

In practical use, the false alarm rate as well as the missing alarm rate are also the concerns of accuracy. In our experiment, we use 38.7 hours (one-minute interval, 2320 samples) Nimda data to compare the predicted results with true labels. Figure 7 contains four subfigures, plotting (a) true labels, (b) predicted labels of MS-LSTM model, (c) predicted labels of SVM, and (d)

Table 3: The performance of different validation methods

	Accuracy	
Method	W=10	W=30
SVM1	78.2(73.4)	76.9(73.4)
NB1	51.8(55.2)	51.9(55.2)
AdaBoost1	82.3(76.0)	83.8(76.0)
MS-LSTM1	**90.4**	**86.8**
SVM2	72.5(68.8)	72.2(68.8)
NB2	51.2 (51.2)	56.7(51.2)
AdaBoost2	73.3(68.0)	67.7(68.0)
MS-LSTM2	**81.5**	**82.5**
SVM3	97.1(86.5)	96.3(86.5)
NB3	88.3 (53.2)	97.0(53.2)
AdaBoost3	**97.8(89.9)**	97.4(89.9)
MS-LSTM3	95.4	**99.5**

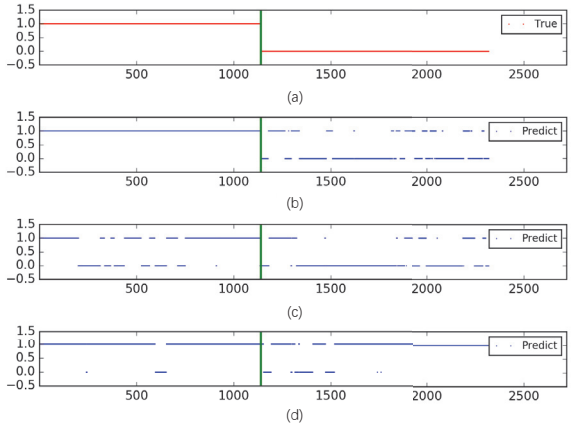

Figure 7: The comparison of true label and predict label of Nimda testing sequence.

predicted results of NB respectively. From Figure 7, our MS-LSTM model has no false alarm and can detect most anomalous samples compared with NB and SVM.

6. CONCLUSION

BGP anomaly detection is an important task since anomalies in one single router may affect the connectivity and stability of the whole network. To classify anomalies from normal ones, we propose a multi-scale LSTM model (MS-LSTM) for the detection of BGP anomalies of several typical real-world events. We compare MS-LSTM with different state-of art classification models on the real BGP datasets and observe that our method can achieve higher accuracy as well as lower false alarm rate. MS-LSTM can learn long dependency in temporal pattern with optimal time scale. In addition, we find that the selection of time scale has great impact on the performance of most classification models for BGP anomaly detection including LSTM.

7. ACKNOWLEDGEMENTS

The work described in this paper was supported by the grants from National Natural Science Foundation of China (No.61472337, No.61300221),Fundamental Research Funds for the Central Universities(No.2014ZZ0038) and was partially supported by NSFC-Guangdong Joint Fund under project U1501254 and Hong Kong Research Grant Council under CRF C7036-15G.

8. REFERENCES

[1] K. Zhang, A. Yen, X. Zhao, D. Massey, S. F. Wu, and L. Zhang. On detection of anomalous routing dynamics in BGP. In *International Conference on Research in Networking*, pages 259–270. Springer, 2004.
[2] J. Zhang, J. Rexford, and J. Feigenbaum. Learning-based anomaly detection in BGP updates. In *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, pages 219–220. ACM, 2005.
[3] N. M. Al-Rousan and L. Trajković. Machine learning models for classification of BGP anomalies. In *2012 IEEE 13th International Conference on High Performance Switching and Routing*, pages 103–108. IEEE, 2012.
[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
[5] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2009.
[6] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.
[7] M. Basseville, I. V. Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.
[8] J. Li, D. Dou, Z. Wu, S. Kim, and V. Agarwal. An internet routing forensics framework for discovering rules of abnormal BGP events. *ACM SIGCOMM Computer Communication Review*, 35(5):55–66, 2005.
[9] A. Dainotti, A. Pescape, and K. C. Claffy. Issues and future directions in traffic classification. *IEEE network*, 26(1):35–40, 2012.
[10] A. Munoz and J. M. Moguerza. Estimation of high-density regions using one-class neighbor machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):476–480, 2006.
[11] T. Ahmed, M. Coates, and A. Lakhina. Multivariate online anomaly detection using kernel recursive least squares. In *IEEE INFOCOM*, pages 625–633. IEEE, 2007.
[12] A. W. Moore and D. Zuev. Internet traffic classification using bayesian analysis techniques. In *ACM SIGMETRICS Performance Evaluation Review*, volume 33, pages 50–60. ACM, 2005.
[13] N. Al-Rousan, S. Haeri, and L. Trajković. Feature selection for classification of BGP anomalies using bayesian models. In *2012 International Conference on Machine Learning and Cybernetics*, volume 1, pages 140–147. IEEE, 2012.
[14] S. Chauhan and L. Vig. Anomaly detection in ECG time signals via deep long short-term memory networks. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–7. IEEE, 2015.
[15] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings*, page 89. Presses universitaires de Louvain, 2015.
[16] S. Basu, A. Mukherjee, and S. Klivansky. Time series models for internet traffic. In *INFOCOM'96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, volume 2, pages 611–620. IEEE, 1996.
[17] RIPE RIS raw data [`https://www.ripe.net/`].
[18] W. Hu, W. Hu, and S. Maybank. Adaboost-based algorithm for network intrusion detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(2):577–583, 2008.