International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

A Correlation and Clustering Algorithm for Time Series Monitoring of Network Flows

Aneesh C Rao, Shobha G

Abstract*: Over the last few decades the advent of machine learning has found applications in large number of domains. The ability to learn and identify patterns in data and make predictions on deviations from these patterns has found large scope in several fields. In computer networks, this problem can be applied to understanding the behavior of a network at any point of time and predicting when the behavior may change over time. This paper discusses an approach that uses a statistical machine learning algorithm for the time series behavior analysis of computer networks. The proposed algorithm makes use of unsupervised learning and statistical data analysis methods over the flows in the network. The novel aspect being explored is the analysis of the inter-dependencies between flows, in addition to monitoring anomalies of individual flows. The results presented, provide waveform representations of the correlation and clustering patterns of some sample flows based on packet sizes.

Keywords*: Statistical machine learning, Anomaly detection, Clustering, Correlation, Real-time monitoring

I. INTRODUCTION

Computer networks, in the real world, see the transfer of a large number of packets over multiple hops, at almost every second. The preservation of information is crucial to the success of the network. It is important to ensure quality-of-service (QoS) in such networks to ensure efficient and effective transfers of packets through flows. Over the years, several methods have been developed for maintaining quality-of-service. In this paper, an algorithm based on statistical machine learning is discussed for the time series monitoring of network flows in complex computer networks such as industrial networks, large data centres, etc. The quality-of-service checks are generally incorporated at intermediate and end nodes in the network, between incoming and outgoing flows (ingress and egress). Machine learning will be used to identify patterns in the packet flows, which will then be used to monitor their behaviour over time. The challenge is to ensure a real-time implementation, as flows in the network are dynamic in nature.

In several past implementations, historical data about flow parameters and patterns have been used as datasets for learning and predictions are made based on this available data. Most dynamic networks require real-time counterparts to these static implementations. Realtime implementations in the past use time series forecasting methods such as AR (auto-regressive), ARMA (auto-regressive moving average) and ARIMA (auto-regressive integrated moving-average). In this paper, a real-time implementation is proposed using Pearson's correlation algorithm and clustering based on the correlation.

Revised Manuscript Received on May 25, 2020.
* Correspondence Author
Aneesh C Rao*, B.E. Computer Science and Engineering, R.V. College of Engineering, Bangalore, India. E-mail: aneeshcrao.cs16@rvce.edu.in
Dr. Shobha G, Professor, Computer Science and Engineering, R.V. College of Engineering, Bangalore, India. E-mail: shohbag@rvce.edu.in

II. LITERATURE SURVEY

In [1] the author has presented a statistical technique for the analysis of the regularity of changes in the intensity of telecommunications traffic for the different periods of time. The data was collected over a period of five years of traffic intensities and an AR model was constructed for the monitoring.

In [2] a network traffic prediction model that uses real traces from a server to train a Long Short Term Memory (LSTM) neural network and generate predictions at short time scales is presented. As a pre-processing step, a feature-based clustering is performed to group similar time series together. The resulting model was able to predict network traffic with low prediction errors.

Alberto Mozo in [3] suggests the use of convolutional neural networks (CNNs) to forecast short-term changes in the amount of traffic crossing a network. Although this outperformed ARIMA, the behaviour at a higher scale is chaotic.

[4] addresses the challenge of real-time monitoring in high intensity traffic situations in a passive monitoring scheme. This problem is overcome by selectively sampling the network data so as to learn and relearn in real-time rather than having a passive pre-learning, which may not hold good over the entire lifetime of the network. The sampling can be done at intervals which can help in identifying new patterns due to additions of flows, flows that exit, etc. This model will be used in the project as it is the most appropriate in terms of real-time usage.

The model in [5] applies time series prediction on a routing protocol as a decision factor for managing traffic dynamically using AR, ARMA and ARIMA along with stationary assumptions obtained using ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function).

The paper [6] proposes and compares two proactive anomaly detection schemes based on statistical procedures – Principal Component Analysis and Ant Colony Optimisation metaheuristic.

These are used to generate a traffic profile which functions as the signature of the traffic, which is compared with real-time network traffic in the monitoring phase.

In [7] a survey is presented, where applications for time-series statistical analysis of computer network data, specifically for performance and anomaly detection are covered. This system employed a probe which monitored the network flows. Inside the probe, a Statistical Engine uses Finite Automata techniques to detect flow behaviour.

[8] covers an interesting improvement of the basic traffic prediction. The model being worked upon the paper combines wavelet techniques and the time-series ARIMA model. Wavelet decomposition of flow is used to get details coefficients and approximation coefficients. On the detail coefficients, the stationary series model is applied, and on the approximation coefficients, the difference method is applied; and finally, the ARIMA model is used to make predictions.

In [9] Linear, non-linear and hybrid time series modelling techniques for network traffic prediction are discussed. The paper also discusses the collection of data (DARPA dataset, NSL-KDD dataset and CAIDA dataset) and also some preprocessing techniques. The methods discussed, however, are not suitable for long-range network traffic.

The paper [10] presents a traffic prediction model for wireless mesh networks to ensure quality of service for users. The model pro non-linear and hybrid time series modelling techniques for network traffic prediction are discussed. The paper also discusses the collection of data (DARPA dataset, NSL-KDD dataset and CAIDA dataset) and also some pre-processing techniques. The methods discussed, however, are not suitable for long-range network traffic.

In the project under consideration, a model is design the adapt to the monitoring of long-range network traffic in industrial networks and can be ported onto switches and routers. The model is to be built using a correlation and clustering approach by dynamically analysing patterns in the network flows at operation time.

Cmputer networks, in the real world, see the transfer of a large number of packets over multiple hops, at almost every second. The preservation of information is crucial to the success of the network. It is important to ensure quality-of-service (QoS) in such networks to ensure efficient and effective transfers of packets through flows. Over the years, several methods have been developed for maintaining quality-of-service. In this paper, an algorithm based on statistical machine learning is discussed for the time series monitoring of network flows in complex computer networks such as industrial networks, large data centres, etc. The quality-of-service checks are generally incorporated at intermediate and end nodes in the network, between incoming and outgoing flows (ingress and egress). Machine learning will be used to identify

III. EXPERIMENTAL SETUP

For the construction and validation of the model, the model was ported as a software tool onto a switch/router and test under laboratory conditions. Routers can be connected over a path so as to test for multiple hops of packets. The time series monitoring is performed at each of the nodes. The training phase or the learning phase is not exclusive from the monitoring or real-time phase. The training and setting of parameters are done at run time as packets are flowing through the routers in the network. Once learning is completed, the system transitions into the monitoring phase, where flows are checked for deviations from ideal patterns. Packets can be sent through a signal generator connected physically to a switch by running a script. During testing, scripts can be written to introduce anomalies into the network by increasing or decreasing packet sizes, delaying packets and so on.

IV. EXPERIMENTAL ANALYSIS

The model describes a real-time approach to performing time series monitoring of flows in the network. The analysis is done in three phases – signature phase, learning phase and monitoring phase. Two main modules constitute the system – correlation and clustering. The correlation stage consists of auto-correlation, that is used for monitoring individual flows, and cross correlation that is used for monitoring interdependencies between flows.

In the signature phase, the shortest delta time (inter-packet gap) of the flows is calculated for the correlation granularity. Other parameters like the learning period, monitoring period, and sampling duration are all calculated in this phase. The signature phase may need to be repeated if a new flow is added. For example, if a slower flow is added, then the correlation granularity and smallest delta time will change, thus all parameters will need recalculation. Another instance is when a flow exits; when the slowest flow exits, the parameters will have to be set based on the next slowest flow, thus again requiring a recalculation. In this phase, packet timeout is also taken care, i.e. if a stray flow is created and dies out without any packets, then it is marked as timed out.

The learning phase is started once all the flows have finished their signature phase and all the parameters have been calculated. In the learning phase, in real time, packets are received and checked for sampling duration. Once enough packets are available, sampling is started and auto-correlation is performed. The auto-correlation is performed by sliding window technique, and the distance (time period) between two successive correlations is marked as the periodicity. Deviations from the periodicity during the learning phase are overlooked. The periodicity is used in the monitoring phase to detect anomalies in the flow pattern. Once all flows have completed their auto-correlations, they are cross correlated between each other to identify similar flows. The flow parameters used for these calculations are delta time, arrival time, packet size, etc. These initial cross correlations are used to determine the ideal behavior (inter-dependency) between flows.

The monitoring phase is started once all the flows have finished their learning phase and the ideal behavior has been identified. In this phase, real time monitoring is performed; packets are received and sampled for periodicity. The periodicity is compared with the ideal learning periodicity checked for variations (anomalies). Once all the packets are collected, cross correlation is performed between all pairs of flows for clustering.

The new clustering is compared with the learning (ideal) clusters for anomaly detection.

For the correlation between flows, the algorithm Pearson's correlation is used.

The advantage of using correlation as means of comparison is that it provides an effect size information. Different statistics can be combined and converted to represent a single effect coefficient. A correlation coefficient between two data series represents the similarity in structure (waveform) between the two series. Pearson's correlation coefficient gives information about the magnitude of association, as well as the direction of the relationship.

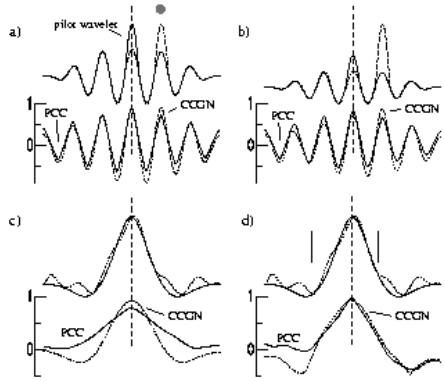

Fig 1. Cross correlation waveforms

Mathematical model for correlations

The Pearson's correlation between two time series **x** and **y** is given by

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The periodicity of the flow is given by

$$\rho = t_{r_1} - t_{r_2}$$

where t_{r1} and t_{r2} are the consecutive timestamps when $r > \alpha$

If r is the correlation coefficient between the data series of two flows, then the clustering is performed as follows

$$if\ r > \alpha, then\ flow\ A\ and\ B\ are\ added\ to\ the\ same\ cluster,$$
$$if\ r < \alpha, then\ flow\ A\ and\ B\ are\ moved\ into\ separate\ clusters$$

This model is followed during the learning as well as the monitoring phase.

Pearson's correlation is employed during the auto-correlation and cross-correlation stages. In auto-correlation, a single flow is used. Data statistics of a single flow series (delta-time, packet size, etc.) are taken in samples and correlated with each other over a sliding window. The correlation coefficient follows a waveform and the time period of this waveform is called the periodicity of the flow. This calculation is performed once during the learning phase and is taken to be the ideal periodicity. This parameter is calculated again during monitoring, but only to check for deviations from ideal behaviour. During the cross correlation stage of learning, all pairs of flows are taken and Pearson's correlation coefficient is calculated between their time series. If two flows correlate ($r > \alpha$, where α is a correlation threshold), they are clustered together, else, they are put into separate clusters. During the monitoring phase, this cross correlation is performed in real-time in phases and the clusters generated in each phase are compared with the ideal stage clusters. Deviations in periodicity and in clustering are classified as anomalies.

Further analysis can be performed based on the correlation parameter being used; for example, if the delta time periodicity of a flow changes, but the size periodicity remains same, then we can classify the anomaly as a packet delay anomaly. Similarly, in the case of cross correlation, if two flows that used to correlate, fail to correlate during a phase monitoring while considering the packet size parameter, then it can be certified that one of the flows (or both) received a sudden surge of large or smaller packets.

V. RESULTS

The model provides a very simple and real-time solution to the time series analysis problem. The system can be driven into learning at any point of time without any change in the core algorithm. Further, this provides a method for analysing the inter-dependencies between flows, which has not been explored before. The waveforms below denote the process of identifying correlation mismatches and anomalies in periodicity and clustering. An example of the auto-correlation as well as cross-correlation is shown. The anomalies can be immediately identified and reported in the same time series.

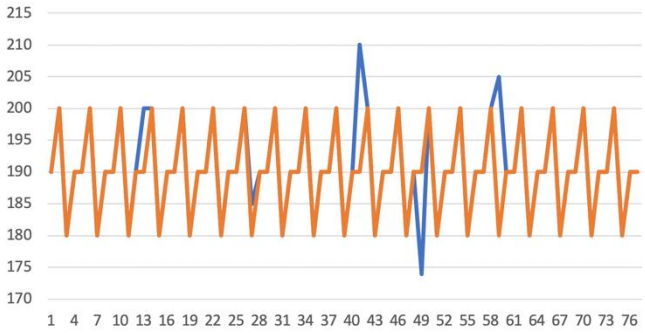

Fig 2. Orange - flow during ideal (learning). Blue - flow during monitoring (periodicity has changed)

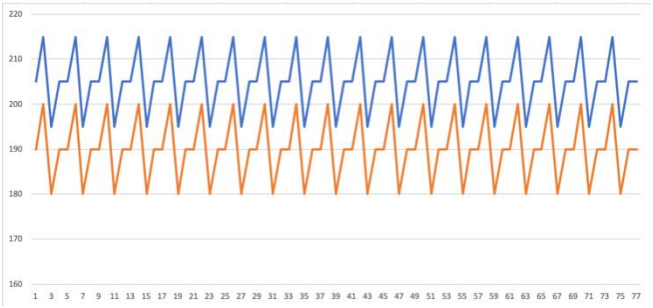

Fig 3. Two flows that correlate (same cluster)

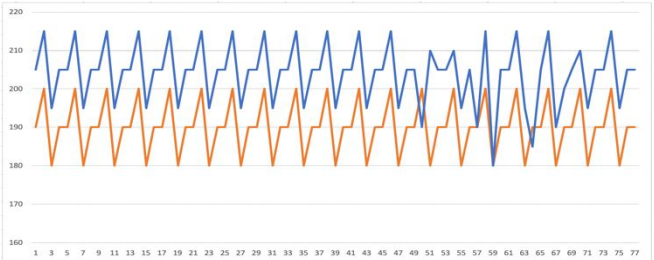

Fig 4. One of two correlating flows stops correlation

VI. CONCLUSION

Network traffic has become one of the increasing concerns in computer networks. The need for a simple and effective method to detect sudden traffic and anomalies in network flows is needed to ensure safe transmission of data over a link. The proposed solution of an algorithm for the time series monitoring can be integrated into a software tool to be used in the operating system of end nodes in a network such as routers. The simplistic implementation guarantees fast anomaly detection and does not add to the latency of packet processing at the node. A limitation of the model used is the relearning phase which needs to occur every time a new flow enters or leaves the network. This is because the correlation granularity may change if the slowest flow changes. This can be improved by performing various checks on the slowest flow in the network to avoid unnecessary relearning. Also, in some networks, the ideal patterns may itself change over time. In addition, this algorithm provides an effective solution for relearning the new ideal patterns without any alteration in the algorithm.

model provides a very simple and real-time solution to the time series analysis problem. The system can be driven into learning at any point of time without any change in the core algorithm. Further, this provides a method for analysing the inter-dependencies between flows, which has not been explored before. The waveforms below denote the process of identifying correlation.

ACKNOWLEDGEMENT

I would like to extend my gratitude towards my guide Dr. Shobha G for her continued support and guidance on my work in this project and paper. I would also like to sincerely thank the Department of Computer Science and Engineering, RV College of Engineering for giving me the chance to write this research paper on my final year project.

REFERENCES

1. Olga Malyeyeva, Yurii Davydovskyi and Viktor Kosenko, “Statistical analysis of data on the traffic intensity of Internet networks for the different periods of time”, CEUR-WS/Vol-2353, 2017
2. Aggelos Lazaris and Viktor K. Prasanna, “An LSTM Framework For Modeling Network Traffic”, University of Southern California, Vol 3, 2018
3. Alberto Mozo, Bruno Ordozgoiti B and Gómez-Canaval S, “Forecasting short-term data center network traffic load with convolutional neural networks”, PLOS ONE, Vol 12, Feb 2018
4. Thanasis Vafeiadis, Alexandros Papanikolaou, Christos Ilioudis and Stefanos Charchalakis, “Real-Time Network Data Analysis Using Time Series Models”, Simulation Modelling and Practice Theory 2019:173-180
5. Sangjoon Jung, Chonggun Kim and Younky Chung, “A Prediction Method of Network Traffic Using Time Series Models”, International Conference on Computational Science and Its Applications, 2016, pp 234-24
6. Gilberto Fernandes Jr. Luiz F. Carvalho Joel J.P.C. Rodriguesa Mario Lemes Proença Jr., “Network Anomaly Detection using IP flows with Principal Component Analysis and Ant Colony Optimization”, Journal of Network and Computer Applications, Vol 64, April 2016
7. Marcos Portnoi, Priscilla Santos Moraes and Martin Swany, “Time-Series Analysis for Performance Monitoring and Anomaly Detection in Computer Networks”, University of Delaware, 2014
8. LI Jing fei, SHEN Lei and Tong Yong an, “Prediction Of Network Flow Based On Wavelet Analysis And ARIMA Model”, International Conference on Wireless Networks and Information Systems, Vol 34, 2009
9. Manish R. Joshi and Theyazn Hassn Hadi, “A Review of Network Traffic Analysis and Prediction Techniques”, School of Computer Sciences, North Maharashtra University, 2017
10. Nie, L., Jiang, D., Yu, S., & Song, H. (2017). Network Traffic Prediction Based on Deep Belief Network in Wireless Mesh Backbone Networks. 2017 IEEE Wireless Communications and Networking Conference (WCNC). doi:10.1109/wcnc.2017.7925498

AUTHORS PROFILE

Aneesh C Rao is a 4th year B.E. Computer Science and Engineering student at R.V. College of Engineering. He has worked on several projects involving Machine Learning, Database concepts and Computer Networks. He has two conference publications at CSITSS IEEE conference, RVCE and National Satellite Technology Day, ISRO. His interests include Natural Language Processing, Operating Systems Machine Learning and Computer Networks.

Dr. Shobha G is a Professor in R.V. College of Engineering Bangalore. She has over 25 years of experience in teaching and over 14 years of experience in research. Her primary interests lie in Data Mining, Image Processing and Networking. She has 123 publications in International journals and conferences. She has also filed 4 patents and reviewed several books. She has published a chapter on ‘Machine Learning and book of Statistics’ She was the former Head of Department of Computer Science Department of RVCE.