Network Performance Behavior Detection with Multivariate Clusters in Time Windows

Tyler Leibengood
Youngstown State University
Youngstown, Ohio
tjleibengood@student.ysu.edu

ABSTRACT

This project analyzes Tstat logs collected from Lawrence Berkeley National Laboratory's ESNet data transfer nodes to obtain insights on data transfer characteristics and behavior. Detecting anomalies in network transfers at the package level will provide solutions for improving network transfer rate. Several feature subsets from the Tstat logs were identify as good predictors for low network throughput. Dimensionality reduction was used to reduce the number of features and to select several sets of prominent features. K-means clustering provided a way to group data transfers by transfer quality. T-SNE was used to visualize and verify multi-variate clustering results in two-dimensional scatter plots. The results indicate that there is high correlation between the percentage of the smallest cluster of transfers and average throughput per time window for low throughput.

ACM Reference Format:
Tyler Leibengood. 2018. Network Performance Behavior Detection with Multivariate Clusters in Time Windows . In *Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'18).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Network traffic management plays a vital role in maintaining healthy operations within all varieties of computer networks. Online traffic monitoring can be used to predict traffic behavior and unexpected events in real-time. Transmission Control Protocols (TCPs) are some of the most basic protocols that verify the transmission and reception of data through a network. Network devices that record TCP logs can be used in correspondence with data analysis and prediction tools to monitor and anticipate traffic behavior in the network. Data analysis and machine learning methods can be employed to identify bottlenecks and explain the status of network traffic using data from network measurements.

2 MATERIALS AND METHODS

Large scientific facilities use Science DMZ, which includes several dedicated data transfer nodes, and high performance data movement tools to attain high network transfers for high-performance scientific applications. Tstat is one available tool for monitoring the network traffic. It computes over 80 different performance statistics at both the IP and TCP layers. Recently, machine learning techniques have been proposed to classify network traffic in order to detect any traffic patterns or anomalies. The main idea for this project is to first extract essential statistical properties of the packets that are flowing through the network, and second, group them together in clusters based on their similarities.

The proposed method can be accomplished using Principal Component Analysis (PCA) as a dimensionality reduction method and k-means as the clustering algorithm. The well-known dimensionality reduction method called Principal Component Analysis (PCA) [4] uses orthogonal transformations to determine principal components of the set in order of highest variance to lowest variance. PCA is used to select notable features that best describe the variance of the data. K-Means [2] is a partitioning clustering algorithm that requires the amount of clusters as an input parameter.

After clustering selected features, the data will undergo dimensionality reduction to visualize the success of the clustering. Van der Maaten in [3] proposed a versatile dimensionality reduction algorithm commonly used for visualization of complex patterns in high dimensional data called t-SNE.

3 RESULTS

For this project we used Tstat data containing 104 features and over ten million points from the start of January 1, 2017 until June 28, 2017 at 10:36 AM. The dataset is then ordered by the time of the first packet sent in each transfer. From the original 104 features, combinations of features were categorized into several subsets. Scalable features, scalable features without including port numbers, scalable ONTIC [1] chosen features, C2S (client to server) features, S2C (server to client) features, an intersect of the ONTIC chosen

Feat.	*S2C*	*C2S_relevant*	*S2C Recommended*	*no_ports*
1	s_port:16	c_cwin_max:70	s_mss_min:89	c_mss:64
2	s_mss:87	c_mss_max:65	s_pkts_dup:99	c_mss_max:65
3	s_rst_cnt:18	con_t:42	s_cwin_min:94	s_mss:87
4	s_win_scl:84	c_win_min:68	s_mss_max:88	c_win_scl:61
5	s_mss_min:89	c_rst_cnt:4	s_win_min:91	s_rst_cnt:18

Table 1: The five prominent features in the most successful subsets are shown above.

SC'18, November 2018, Dallas, Texas USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

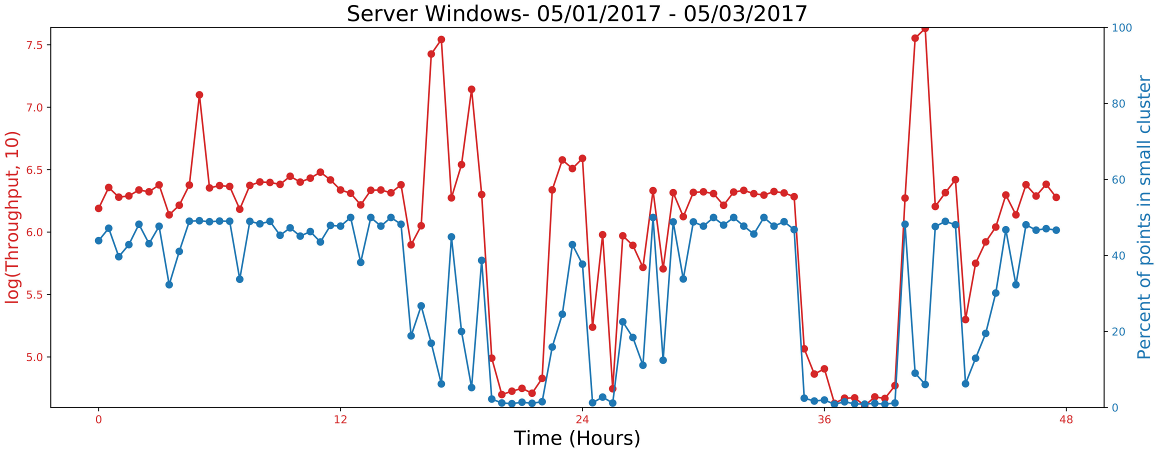


Figure 1: When throughput is low, the model detects throughput accurately. Because our algorithm doesn't allow percent to go over 50%, cases when throughputs are high are represented poorly.

features with C2S and S2C independently, and important features recommended by previous research. The names of these subsets is *scalable*, *no_ports*, *relevant*, *C2S*, *S2C*, *C2S_relevant*, *S2C_relevant*, and *Recommended* respectively.

To select the most prominent five features from each subset, a procedure based on PCA was used. The prominent features were identified using the highest variance in the first primary component of the PCA result. Table 1 contains prominent features in our most successful subsets.

The data is normalized, grouped in windows of 30 minutes, then grouped into two clusters using k-means. The percent of transfers in the smaller cluster and each window's average throughput are calculated.

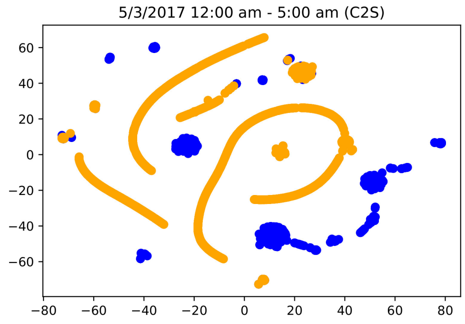
When comparing percent of points in the smaller of the two clusters with average throughput, the most successful subsets were *S2C*, *S2C Recommended*, *C2S_relevant*, and *no_ports*. Figure 1 shows an overlay of the percent of points in the minority cluster for different windows compared to the log of average throughput for corresponding windows of the *S2C* subset. The figure spans the first 2 days of May 2017.

T-SNE plots (Figure 2) showed the success of the clustering on an example of a subset that perform well and a subset that performed poorly to verify their distinction. Overlapping points with different labels implies that the k-means clustering is inaccurate.

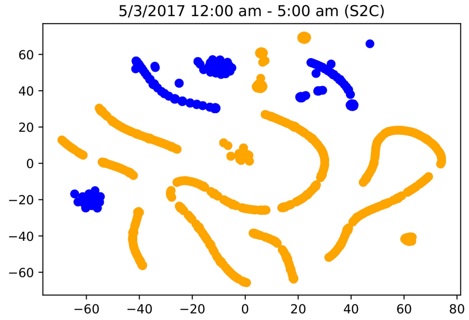
To quantitatively evaluate the results presented in Figure 1, root-mean-square error (RMSE) between the average throughput and the percentage of the smallest cluster was calculated only for data

	S2C	*C2S_relevant*	*S2C Recommended*	*no_ports*
RMSE	0.2057	0.2661	0.1962	0.2035

Table 2: Root Mean Squared Error test on best performing subsets



(a) Overlapping points of different labels are visible in the t-SNE mapping of the *C2S* data. This implies inaccurate k-means clustering.



(b) Points of different labels are visibly separate in the t-SNE mapping of the *S2C* data. This implies accurate k-means clustering.

Figure 2: T-SNE mapping of 5 hours of clustered *C2S* and *S2C* data.

transfers less than 1 Mb/s. Table 2 clearly shows that features extracted from the S2C Recommended subset perform the best in terms of the correlation with the average throughput.

4 CONCLUSIONS

Using clustering combined with dimensionality reduction to analyze Tstat logs shows a clear correlation between throughput and percentages of normal transfers. Out of all the feature subset combinations tested, it was found that four of them make accurate detections of windows with low throughput. This new method to detect network data transfer performance behavior can accurately and consistently cluster normal and anomalous network transfers and detect abnormally low throughput. A known deficiency of this approach is that it cannot determine which cluster is normal and which is anomalous. In the future, we plan to update the proposed method to additionally identify the cluster of normal transfers.

ACKNOWLEDGMENTS

I would like to thank my advisors, Alina Lazar, Alex Sim and John Wu for their advice and guidance. This work was prepared in partial fulfillment of the requirements of the Berkeley Lab Undergraduate Faculty Fellowship (BLUFF) Program, managed by Workforce Development & Education at Berkeley Lab. This work was supported by the U.S. Department of Energy, under Contract No. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center.

REFERENCES

[1] Juliette Dromard, Philippe Owezarski, Alberto Mozo Velasco, Bruno Ordozgoiti, and Stanislav Vakaruk. [n. d.]. Online Network Traffic Characterization. ([n. d.]).

[2] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[3] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[4] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572.