Predicting TCP Throughput Using Machine Learning for Improving Network Resource Allocation

Shahbaz Ali Khan, Zanxiang Yin
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA

I. Introduction

TCP protocol is a very important part of network protocol. The main function of TCP protocol is to ensure the reliability and order of data transmission between network devices. TCP is widely used to browse web sites or download files. It follows that the metrics that determine TCP network performance are critical to the impact of user experience. The concept of throughput fully indicates the amount of data that TCP receives and sends. In combination, TCP throughput reflects the transmission capacity of a TCP/IP network between two or more devices. Accurate prediction of TCP throughput can improve the efficiency of many applications that rely on selection of servers or peer-to-peer applications, such as video streaming or file sharing [4]. With the optimization of such applications, the allocation of network resources can be more reasonable to achieve the purpose of optimizing the network environment.

This study hopes to use machine learning to predict TCP throughput to improve network resource allocation. Packet loss and round-trip time are two important network performance indicators that affect TCP connection throughput. This paper will mainly achieve the prediction of TCP throughput through the training and prediction of packet loss and round-trip time.

II. Related Work

The field of research in TCP predictive throughput has been thoroughly studied in various directions over the last few years and covers the extensive use of machine learning techniques as well as mathematical models, and historic information. Two research studies in particular have shown impressive progress in this field.

First, Mirza et al. 's research "A Machine Learning Approach to TCP Throughput Prediction" concentrates on the role of machine learning for TCP throughput forecasting [3]. Through a deep dive into the various algorithms used in machine learning, they attempt to discover the benefits and drawbacks of these methods for analyzing and forecasting TCP performance. The research examines the possibility of using machine learning models for TCP performance prediction, for first time. It will provide the opportunity to conduct deeper research into the field in the near future.

Furthermore, On the Predictability of large-scale transfer TCP throughput, by Qi He, Constantine Dovrolis Qi He, Constantine Dovrolis Mostafa Ammar and others [2]. It focuses on the ability to predict TCP throughput during huge-scale transmission of data. Through a thorough examination of the structure and the empirical assessment of TCP throughput predictors, the authors provide crucial insight into the capabilities of TCP protocols for dealing with large-scale data. By using mathematical models and data from the past scientists can make the forecast of data transfer with large amounts more precise. Researchers also discuss problems associated with massive data transfer in complicated network settings, specifically when there are emergencies and other abnormal circumstances.

Two studies provide multiple research concepts. Mirza et al. concentrate on the possibilities of machines learning models as an attempt to create new strategies to improve TCP predictions of performance. The work done by Qi He et al. is more focused on analyzing the efficiency of the TCP protocol for large-scale scenario of data transmission, offering profound insights into practical application situations.

Together, these research studies give a new perspective on the subject of TCP throughput prediction. The research of Mirza and co. expands the scope of prediction techniques, whereas Qi He et al. provide a concrete example of the behavior of TCP protocols in various situations. This has the potential for further research as well as practical applications. Further research is required to integrate various approaches to better comprehend and overcome the issues with TCP predictions of throughput.

III. Methodology

This paper aims to predict TCP throughput by using machine learning technology to improve network resource allocation. Data integration and analysis of TCP throughput is the top priority of this study. This study will select two important characteristics that affect TCP throughput: packet loss and cycle time. Through these two important indicators, quantitative analysis is carried out.

I. Data Collection

Collecting data on TCP throughput is a very large effort. The network topology is quite complex, which means that it is complicated to accurately measure end-to-end TCP throughput in multiple subnets and routers. The first step in collecting TCP throughput data is to capture packets over the network, which can be done by using software such as WireShark, but can be somewhat mixed. At the same time, if you want to grab a certain amount of network parcels, there are also certain permission restrictions.

Knowing the above difficulties in collecting TCP throughput data, we decided to use the available network traffic data. This paper uses the experimental data of MIT Resilient Overlay Networks (RON) [1]. The elastic overlay network itself was developed to allow end hosts and applications to collaborate to obtain higher reliability and performance from the Internet. This paper uses its experimental data, which contains millions of delay and loss samples, as well as thousands of throughput samples collected on the RON testbed. In the RON dataset, the bandwidth dataset contains 3,237 datapoints and the latency and loss dataset contains 2,595,172 datapoints. This large sample size is enough to support our training and prediction of TCP throughput.

The bandwidth dataset helps us determine the throughput of TCP in the network. Using the byte count/duration formula, TCP throughput marker data can be derived. By studying the delay and loss dataset, round-trip delay information is extracted from this dataset to add network delay features while predicting TCP throughput. Finally, the preprocessing of TCP throughput data set is completed by connecting and merging the two data sets according to the IP addresses of the source and destination.

II. ML Coding and Debugging

Another component studied in this paper is the application of machine learning models. After comparing several machine learning models, we decided to use support vector machine (SVM) for TCP throughput prediction. SVM has excellent performance in nonlinear relationship model. Since the relationship of TCP throughput is likely to be non-linear, we give up the basic linear regression algorithm for prediction and training and choose SVM instead.

At the same time, the ability of SVM to resist overfitting is also a point we value very much. SVM can improve the generalization performance of the model by adjusting parameters when training some data sets and prevent over-expansion of training data. This is very tolerant of some impurities and outliers in the TCP throughput data.

Because in network performance testing, especially for TCP throughput testing, there are often some impurities and exceptions. The SVM model has less influence on one outlier, which helps to improve the stability of the model.

III. Evaluating Model performance

In the model evaluation stage, this paper uses different evaluation indicators, hoping to have a comprehensive understanding of the performance of the model. In view of the characteristics of SVM algorithm, this paper mainly uses the mean square error (MSE), mean absolute error (MAE) and R-square value to evaluate the model.

The advantage of MSE estimation is that it is more sensitive to large errors because the errors are squared. This helps us understand TCP throughput prediction errors and ensures that the model can be detected in time when major errors occur.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MAE is the opposite of MSE, and MAE is not sensitive to outliers, which provides considerable inclusiveness for TCP throughput models. In data where outliers are present, MAE is able to provide a coarser assessment of performance. The absolute value of the error is easier to understand and indicates the size of the average error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|$$

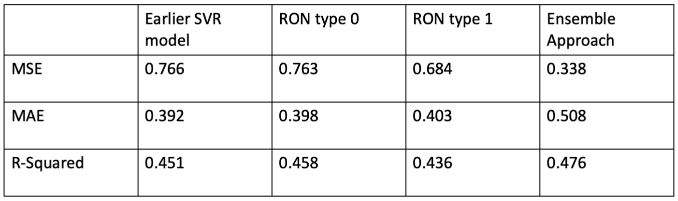
The R-squared value is a useful measure when the explanatory power of the model needs to be evaluated, or when the model needs to be compared with other models. R-squared values are standardized and range between 0 and 1 to facilitate comparisons between models. The closer to 1, the better the degree of fit of the model.

$$R^2 = 1- \frac{\sum_{i=1}^{n}(yi-\widehat{yi})^2}{\sum_{i=1}^{n}(yi-\overline{yi})^2}$$

IV. Results

By experimenting with the design and integration methods of SVR models for different RON data types, we have achieved a series of encouraging results. Compared with the SVR model alone, the ensemble method has achieved significant improvement in MSE and MAE, indicating that the overall prediction accuracy of the model has been improved. In particular, it is worth noting that in terms of R-Squared, the integrated approach achieves 0.436, further confirming the superiority of the model in fitting TCP throughput.

Figure 1. SVR model results



	Earlier SVR model	RON type 0	RON type 1	Ensemble Approach
MSE	0.766	0.763	0.684	0.338
MAE	0.392	0.398	0.403	0.508
R-Squared	0.451	0.458	0.436	0.476

This study provides useful implications for future research in related fields. First, we can consider further extending the classification of data types to cover a wider range of network environments and conditions. This will help improve the generalization ability of the model and make it robust to predicting TCP throughput in different scenarios.

Secondly, attempts can be made to introduce more complex ensemble methods or deep learning models to further improve the performance of the models. The success of ensemble approaches suggests that synergies between models can lead to clear improvements, and that deep learning models may have an advantage in capturing higher-level patterns and associativity.

In addition, further research on the interpretability of the model would be an interesting direction. Understanding how models make predictions is critical for network engineers and decision makers, so a deeper look at the decision-making process within models will help improve the overall understanding of TCP throughput behavior.

Overall, we have achieved satisfactory results by investigating the design and integration methods of different SVR models, but this also provides more directions and challenges for future research. In the ever-changing network environment, improving the accuracy and adaptability of TCP throughput prediction models will be an important topic for future research.

V. References

[1] Andersen, D. G., Balakrishnan, H., Kaashoek, M. F., & Rao, R. N. http://nms. csail. mit. edu/ron/ronweb.

[2] He, Q., Dovrolis, C., & Ammar, M. (2005). On the predictability of large transfer TCP throughput. *ACM SIGCOMM Computer Communication Review*, *35*(4), 145-156.

[3] Mirza, M., Sommers, J., Barford, P., & Zhu, X. (2007). A machine learning approach to TCP throughput prediction. *ACM SIGMETRICS Performance Evaluation Review*, *35*(1), 97-108.

[4] Tian, Y., Xu, K., & Ansari, N. (2005). TCP in wireless environments: problems and solutions. *IEEE Communications Magazine*, *43*(3), S27-S32.

[5] Jain, M., & Dovrolis, C. (2003). End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput. *IEEE/ACM Transactions on networking*, *11*(4), 537-549.