Predicting Internet Path Dynamics and Performance with Machine Leanring

Zhenghui Wang
Shanghai Jiao Tong University
Shanghai, China
felixwzh@outlook.com

Hao Wang
Shanghai Jiao Tong University
Shanghai, China
?@?.com

Yuheng Zhi
Shanghai Jiao Tong University
Shanghai, China
?@?.com

Shukai Liu
Shanghai Jiao Tong University
Shanghai, China
?@?.com

ABSTRACT

We study the problem of predicting internet path dynamics and performance. We use `traceroute` measurement and machine learning models.

KEYWORDS

TODO

ACM Reference Format:
Zhenghui Wang, Hao Wang, Yuheng Zhi, and Shukai Liu. 2017. Predicting Internet Path Dynamics and Performance with Machine Leanring. In *Proceedings of SJTU Computer Network Workshop (CNW).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[[Zhenghui says "I think we should focus on the process not the final result, which is also important."]]

1 INTRODUCTION

[[Hao says "I have finished all of the introduction"]]

Internet paths change frequently due to inter/intra-domain routing changes, load balancing, and even misconfigurations and failures. Some of these changes can seriously disrupt performance, causing longer round-trip times, congestion, or even loss of connectivity. Thus, it is of great significant to predict Internet path dynamics and performance.

In the original paper, the authors focus on the problem of predicting Internet path changes and path performance using `traceroute` measurements. They uses the recent route information of paths (route age for paths, route changes in past, average RTT, etc) to do some predictions of paths. Their are three predict targets: (*i*) the remaining life time of a path (i.e., the time before a path changes), (*ii*) the number of path changes in a future timeslot, and (*iii*) the average RTT of a path in the next `traceroute` measurement.

To achieve these goals, it introduces a NETPerfTrace system, which relies on a standard random forest model for prediction. Moreover it uses extensive evaluation on the impact of different input features by studying the correlations between the inputs and the prediction targets, as well as target selection techniques.

The dataset it provides with is a full week of Paris `traceroute` measurements performed through the M-Lab open Internet measurement initiative. The author observes more than 450,000 different paths sampled through Paris-traceroute measurements from more than 180 geo-distributed servers. However, most of the paths are not periodically sampled during this week. And only 2,346 paths have at least 100 `traceroute` measurements during the analyzed week. So the dataset only contains corresponding infomation of these 2,346 paths.

We find some drawbacks in their data process. The author processes the data and calculates some statistics for the whole seven days and then use the data to get the ramdon forest model.

[[Hao says "maybe we can put the detail discussion of their data process problems in data analysis part and delete the above sentence only leave the first one."]]

And in our own work,firstly we reprocess the provided data in more resonable ways and use the same method (random forest) mentioned by the author to compare the results. Secondly, two other machine learning methods , xgboost and LSTM, are applied to get better performance.

2 RELATED WORKS

1. introduce the original paper

[[Hao says "I have finished this part(2.1)"]]

In the original paper, the problem they solve is to use random forest model to predict three labels for one path: (*i*) the remaining life time of a path (i.e., the time before a path changes), (*ii*) the number of path changes in a future timeslot, and (*iii*) the average RTT of a path in the next `traceroute` measurement.

And to achieve these three targets, 69 input features are used to describe the statistical properties of route dynamics and path latency. For the first route dynamic target, the first group of 11 features,referred to as F_A, is chosen. It describes the statistical properties (average, minimum, maximum, and percentiles) of the route duration observed for each path. And for the second target, the second group of 14 features, as F_B, are relevant to the prediction of number of route changes. F_B features take into account the statistical properties of route changes. In addition to that, F_B contains information about the number of route changes observed for a

CNW, December 2017, Shanghai, China

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

path so far and a binary feature indicating whether a route change occurred for a path in the current time slot. And the last group of 44 features ,referred to as F_C decribes statistical properties of path latency, which is relavent to the prediction of the next `traceroute` measurement. They calculate the staistical properties of four RTT metrics (average, minimum, maximum and standard deviation) reported from each `traceroute` measurement.

Then they study the correlation among the input features and the targets and apply feature selection techniques to select the best features for prediction. When using as input the full set of 69 input features $F_A \cup F_B \cup F_C$, and perform wrapper-based feature selection on top of this full set. The result shows that the top important features for three targets are not necessarily the ones in corresponding F_A, F_B or F_C.

And the authors use three different ways of selecting input features: (i) use all of the 69 features to predict each target, (ii) when predicting target X, use corresponding feature set F_X, (iii) after wrapper-based feature selection, use the top features for each target. And finally, what they get is the third way of choosing input features can achieve better results (though there is only minor differences in output results).

2. we can introduce some other machine learning methods applied in computer network scenarios. **[[Zhenghui says "I'll take this part"]]**

3. very briefly introduce xgboost and lstm

3 DATA ANALYSIS

[[Zhenghui says "I think this is a important part, who will take this part?"]]

1. We can plot some figures of the statistics of data, like the distribution of the route duration and avgRTT.

2. We can further discuss the relation between routes in one path or in different paths.

3. Then we could discuss why the authors of the paper process the data in a wrong way

4. We show our solution for data process. three ways for random forest models.

4 EXPERIMENT

4.1 Classic Models

We show the experiment results of the 3 different data we obtained, namely **K&fix**,**K&update**,**timeslot&update**. We need to find some difference between our 3 data processing methods and the authors', i.e.,**origin**.

The experiments we need to conduct are as follows:

- 1. **K&fix**+RF
- 2. **K&update**+RF
- 3. **timeslot&update**+RF
- 4. **origin**+RF
- 5. **K&fix**+xgBoost
- 6. **K&update**+xgBoost
- 7. **timeslot&update**+xgBoost
- 8. **origin**+xgBoost

[[Zhenghui says "Note"]]We first predict the route duration instead of resLife. Further study on the difference between these two predicting objective could be conduct if we have time. Currently, we predict the route duration for task 1, because we can only predict this when we use LSTM.

4.2 Deep Models

There are two kinds of input for the LSTM at each timestep, (i) one simple scalar (ii) a vector. The experiments we need to conduct are as follows:

- scalar input + LSTM
- vector input + LSTM

5 CONCLUSIONS

TODO

ACKNOWLEDGMENTS

TODO