Reflection for assignment 0

Shanglin Zou Oct,2018

Our group of two were looking for some techniques that we are interested in to analyze cyber security data.

We first tried to find some statistics or machine learning techniques that are useful for analyzing data, what I focused is time series and Dan is interested in K-Nearest-Neigbbours(K-NN) approaching. We both worked with Python, and working to combine our results together.

We actually choose the different sets of data. Since this project is not mentioned clearly that we need to choose the same one, the most important thing we think is to learn from each other's work, to understand the techniques that partner used.

My data set is "http bro log" dataset. This dataset focuses on request/response pairs and all relevant metadata. What I focused is "request_body_len" and "response_body_len", they are represents for actual uncompressed content size of the data transferred from the server/client. So, what I want to do is to fit these data to time series model.

There are some challenges that I need to overcome:

1. Need to know well about time series knowledge. Time series is a very big part of Statistics and also a good machine learning toolkit.
2. I am very new to the python. It is very hard for me to code with python. As I starting this, I need to looking for examples of time series analysis using python which can help me a lot for continue my work.

So first, I went to review the knowledge of time series and how to model a time series data. Resampling, differencing, using autoregressive function (ACF) plot and partial autoregressive function (PACF) plot to determine the order of ARIMA model, fitting and predicting are the steps that we model a time series data. Then, we went to find some good examples that using python to analyze cyber data. An example[1] I found is very understandable and useful for basic time series analytics, especially the ACF and PACF plot part. Before finding this example, I stuck for a long time, and got a really bad plot for that. For modelling section, I also met some problems, for example, when I try to predict the future value, it shows the same value for next several months. Then I try to remodel the data into different model and that works this time. Furthermore, I learned how the code works and fit it into my dataset, and get the prediction of my model. However, I still have a problem that I cannot figure it out, the fitted value is not close to my test data.

After I finish my work, I started to learn about Dan's work, he used Zeus Malware data. He made a model to predict which is Zeus and which is not. He used confusion matrix to analyze the effectiveness of the model and also can calculated sensitivity and specificity rate. Those two rates can show how well the model is.

Overall, from this project, I learned two techniques that are useful to analyze cyber security data and hope I can use them in the future.

Reference

[1]: AARSHAY JAIN. 2018. *Guide to Create a Time Series Forecast (with Codes in Python)*. [ONLINE] Available at: https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/. [Accessed 25 October 2018].