Synthetic Data Generation for Security Applications
Research Note

Shengzhe Xu [1] **Manish Marwah** [2] **Naren Ramakrishnan** [1]

Abstract

In terms of novelty of approach, here are some ideas to pursue to differentiate from prior work:

- Use of side or auxiliary variables, such as time of day, day of week, location if available, etc.

- Shape of conv layer filter based on this application

- While there are no RBG channels, variables along a row are heterogeneous, and need to be models appropriately.

- Size of the problem - some IP addresses/users may have very large number of flows per hour; how do you handle that in an tractable manner, e.g, perhaps some sort of multi-scale, and/or hierarchical modeling

1. Baseline Models

We start our research from the UGR16 dataset: A New Dataset for the Evaluation of Cyclostationarity-Based Network IDSs.

1.1. Data

We first want to analyze and model data of April Week3 day 1. We filter out the flows that both Source Address and Target Address are internal ip (starting with 42.219.*.*), which are our current research target.

As a result, which is quite different with our previous outgoing traffic data (only filter SA with the internal ip pattern but not for DA), we got 732 users and 206300 flows from day 1 data. As shown in Figure 1 and the data_stats.csv attachment file. In this data, 598 users have only 1 flow record. So we don't need to select users with medium #flow as we planned for outgoing traffic data, but selected top 12 users. Figure 2 shows the #flow of the 12 users in each hour of the day. Note that, top 2 users have much more #flow than the other users, which may because they are server or port scanning something like we discussed last week. Thus, we tried to use the top 3-12 users to trained our model.

(**Todo:** After filtering the flows in this way, the amount of the flows are not enough for the future deep learning. We can expand to data of multiple days for CNN model.)

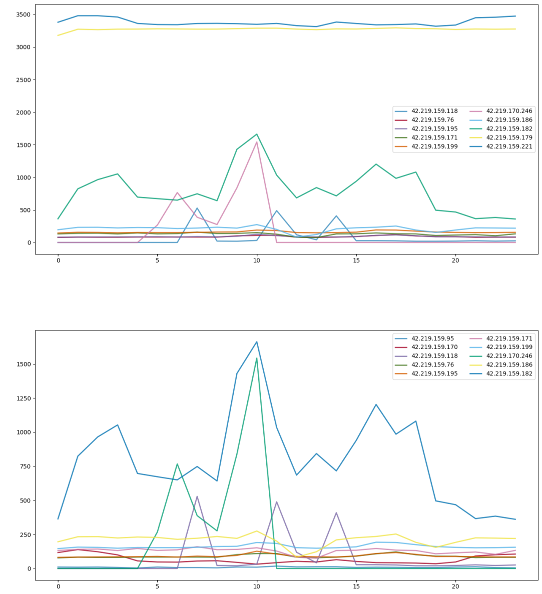

Figure 1. top: hourly data for top 1st to 10th users; bottom: hourly data for top 3rd to 12th users

(**Special Finding:** The #byte of all flows of some users, like 42.219.170.246, are all 64)

1.2. Baseline1: Independently model each column

To describe.

1.3. Baseline2: Naive Bayesian

$argmax_B P(B|T,(B-1)), where P(B|T,(B-1)) = P(T|B) * P(B-1|B) * P(B)$

the bar problem is a trade off. Actually the p(x) can be expanded to p(x in bar) * p(bar in whole data).

To decide the bar, there are 2 potential solutions. First, use same-width bars. Second, use same-sample bars, which means we evenly put equal amount of sample in one bar. Both the 2 solutions ask us to decide the number of bars. May be we can enumerate and try different number of bar value on the training set, then select the best one.

2. Result and Validation

We generate a group of data and try to evaluate them.

#user	#days	#flows generated	model
1	10	49648 rows	baseline1
1	10	49229 rows	baseline2

We want to use the cross-validation to evaluate the model we build. Like we should build the model from the raw data training set and then test the metrics between the raw data testing set and the model distribution. We make the use of the K-L divergence as well as the likelihood to be as our evaluation metrics.

2.1. K-L divergence

p(x) denote the distribution of the raw data and.

$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \ln \left(\frac{P(i)}{Q(i)} \right).$

all day data validation. 1. KL divergence of all day data.

each hour data

2.2. Likelihood

Since we model the data with the help of the Maximum Likelihood Estimate.

3. Problems to be Dealt With

The following link provides a theory that about the possible that p(i) and q(i) = 0.

https://math.stackexchange.com/questions/1228408/kullback-leibler-divergence-when-the-q-distribution-has-zero-values

https://stats.stackexchange.com/questions/275033/how-do-i-calculate-kl-divergence-between-two-multidimensional-distributions