AI607: Graph Mining and Social Network Analysis (Fall 2020)

Term Project: TCP Connection Attack Prediction

Release: Sep 11, 2020
Progress Report: Nov 6, 2020, 11:59 pm
Final Report: Dec 4, 2020, 11:59 pm
Final Presentation: Dec 6, 2020, 11:59 pm

The ultimate goal of this project is a practice of data mining research by addressing the attack prediction problem in the TCP connection network. In this project, you will design, implement, and evaluate your approach for predicting the type of attacks in the network based on past TCP connection histories. Also, you will (a) write a progress report (b) write a final report, and (c) present your approach. While details of the following steps will be announced later, tentative schedules are as follows:

- Progress Report (Max Score: 20) - Nov 6, 2020, 11:59 pm
- Final Report (Max Score: 60) - Dec 4, 2020, 11:59 pm
- Presentation (Max Score: 20) - Dec 6, 2020, 11:59 pm.

This is a team project, and each team should consist of two or three members. You can find your teammates by all means (e.g., Classum), and one progress report should be submitted per team.

Your submission will be evaluated based on

- Presentation of your reports & presentation - 40%,
- Novelty of your proposed approach - 20%,
- validity of your proposed approach - 20%,
- **Accuracy - 20%**.

Note that accuracy is not our only concern. Instead of spending all your time optimizing the accuracy, we recommend spending more time developing novel and valid approaches and making your presentation clear and complete.

1 Problem: TCP connection attack prediction

1.1 Provided Data

Provided data contains TCP connection histories. Each TCP connection history consists of the source id, destination id, port, timestamp, and the type of connection. A detailed description of each is as follows.

- Source id: Id of the source IP address.
- Destination id: Id of the destination IP address.
- Port: Port number of the destination.
- Timestamp: Timestamp indicating the seconds elapsed from a particular date, which is unknown.
- Type of connection: There are 26 types of TCP connections which consist of one benign type and 25 attack types. Note that the benign type is denoted by "-", and each attack type is denoted by its name (e.g. apache2, back, dict, etc.).

All datasets consist of TCP connection histories for 30 minutes. Detailed information about the datasets is given below.

1. **Training Dataset** (`train_XXX.txt`)

   In the training dataset, all information about the TCP connection is included in each line of the file. More specifically, each line provides the source id, destination id, port, timestamp, and the type of connection, which are separated by a tab as follows:

   ```
   <SOURCE ID>\t<DESTINATION ID>\t<PORT>\t<TIMESTAMP>\t<CONNECTION>
   ```

2. **Validation Dataset** (`valid_query_XXX.txt, valid_answer_XXX.txt`)

   Validation datasets consist of query and answer files. In query files (`valid_query_XXX.txt`), each line of the file is in the following format:

   ```
   <SOURCE ID>\t<DESTINATION ID>\t<PORT>\t<TIMESTAMP>
   ```

   In answer files (`valid_answer_XXX.txt`), each file contains the tab-delimited list of attack types that the corresponding validation query dataset has. If the corresponding dataset does not contain any TCP connections corresponding to attacks, then the answer file will be empty.

3. **Test Dataset** (`test_query_XXX.txt`)

   Test datasets only contain query files. Each line of the file is in the following format:

   ```
   <SOURCE ID>\t<DESTINATION ID>\t<PORT>\t<TIMESTAMP>
   ```

Each file is located in a folder with its name of category. For example, The "`train_XXX.txt`" is located in the "`train`" folder.

1.2 Evaluation

You should submit `test_answer_XXX.txt`, which contains a tab-delimited set of the predicted attack types of `test_query_XXX.txt`. The **format of the file must be the same as** `valid_answer_XXX.txt`. If you predict that the network does not contain any attack, then leave the answer file empty.

Using this file, we will evaluate the performance of your approach. The performance will be evaluated using the weighted F1-score F_1^w, which is a weighted average of F1-score for each network:

$$F_1^w \coloneqq \frac{1}{2}\left[\frac{1}{|B|}\sum_{g\in B} F_1\left(g\right) + \frac{1}{|A|}\sum_{g\in A} F_1\left(g\right)\right]$$

where B represents the set of networks without any attacks, and A indicates the set of networks containing TCP attacks. We compute the F1-score on each network as follows:

$$F_1(g) = 2 \times \frac{prec(g) \cdot recall(g)}{prec(g) + recall(g)},$$

where

$$prec(g) = \frac{1 + \text{\# attack types actually included in } g \text{ among those predicted to be included in } g}{1 + \text{\# attack types predicted to be included in } g}$$

and

$$recall(g) = \frac{1 + \text{\# attack types predicted to be included in } g \text{ among those actually included in } g}{1 + \text{\# attack types actually included in } g}$$

Note that we may ask you to run your submitted code on another query set if your answer is suspiciously similar to any other group's answer.

1.3 Notes

- You may encounter some subtleties when it comes to implementation, please come up with your design and/or contact Hyeonsoo Jo (hsjo@kaist.ac.kr) and Inkyu Park (inkyupark@kaist.ac.kr) for discussion. Any ideas can be taken into consideration when grading if they are written in the *readme* file.
- Unlike the other assignments, you are allowed to use any programming language and any external libraries.

2 Presentation Video

The video should not be longer than 5 minutes, we recommend using PowerPoint to create a video. It should describe your approach with some intuition behind it, and it should discuss the accuracy of your approach (in terms of the weighted F1 score) on the validation set.

3 How to submit your project

3.1 Progress Report

Submit your progress report that is written using the attached template to KLMS by Nov 6, 2020, 11:59 pm. The file should be named report-[your student ids].pdf (e.g., report-20189000_20199000_20209000.pdf). Details will be announced soon.

3.2 Presentation Video

Submit your presentation video to KLMS by Dec 6, 2020, 11:59pm. The video should be named video-[your student ids].mp4 (e.g., video-20189000_20199000_20209000.mp4).

3.3 Final Submission

1. Submit project-[your student ids].tar.gz (e.g., project-20189000_20199000_20209000.tar.gz) to KLMS. Your submission should contain the following files:

    - **final_report.pdf**: a final report that is written using the attached template written in LaTeX
    - **slides.pdf**: slides used for final presentation
    - **test_answer.tar.gz**: this file should contain all the `test_answer_XXX.txt` files, which contain the predicted label of each network in the test dataset.
    - **readme.txt**: this file should contain the names of any individuals from whom you received help, and the nature of the help that you received. That includes help from friends, classmates, lab TAs, course staff members, etc. In this file, you are also welcome to write any comments that can help us grade your assignment better, your evaluation of this assignment, and your ideas. This file also should describe how to run your code.
    - **code.tar.gz**: your implementation

2. Make sure that no other files are included in the tar.gz file.