Weekly Progress Report

Name: Pratyay Amrit
Reg. No.: 140953430
Date: 23rd February 2018

1. Tasks Done:

Day/Date	Tasks Done
Monday	Found weaknesses in KDD'99 data set: • Massive redundancy (348435 rows in a 10% subset of the dataset) • Outdated data format and attributes • Poorly choosing training and testing data sets may lead to inaccurate classification.
Tuesday	Found another, more modern data set called UNSW-NB15, also used for IDS performance testing. The data set is much larger and carefully chosen training and testing data sets are available at their website. The data set is modern and goes well with the present standards. It tackles most of the issues of KDD'99, and has proven to be a better performer according to the 2015 published "The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems".
Wednesday	Reading through the above mentioned paper.
Thursday	Found detailed description of every attribute in the data set in the paper. The paper also mentions association rules among the different features for different types of data (normal, DoS, fuzzers, etc.)
Friday	Started working on the new data using python. About 1GB of data is available, training and testing sets are directly available. Nominal data has to be converted to numeric for kmeans.

2. Plans for upcoming week:

- Complete processing of the UNSW-NB15 data set.
- Apply kmeans and test it on the testing data set.
- Use the following performance parameters to analyze kmeans:
  TP: True Positive (correctly classified attacks)
  TN: True Negative (correctly classified normal data)
  FP: False Positive (incorrectly classified attack)
  FN: False Negative (incorrectly classified normal data)
  accuracy = (TP + TN) / (TP + TN + FP + FN)
  efficiency = accuracy / execution_time
- Condition data set for MLP
- Apply MLP