CS 204, Spring 2017 Project Proposal
Abhishek Kumar SRIVASTAVA May 15, 2017
Student ID: 861307778

Objective: The main motivation of my project is to do Network Anomaly detection using Co-clustering method. Using a data set and classify each connection as "normal" or "attack" connection.

Overview: In this project I will apply the 2 different co-clustering algorithms(soft co-clustering and hard co-clustering) to cluster connections and mark which connection reading(connection feature) gives strong indication of the cluster being anomalous with other clusters.

The reason for choosing co-clustering is because it can be good method for separating normal connections from anomalous ones. It is an attempt to isolate set of connections which stand out from the normal user behavior which is also called as "Network Intrusion Detection". Another reason for using co-clustering is in certain type of attacks they are correlated with only subset of connection parameters. We have to find subsets of parameters associated with set of connections that make such set anomalous in set of normal connections.

Co-clustering is a technique in which given data matrix, we try to find subset of row that are correlated to subset of its columns. It is different from traditional clustering in the sense that in that rows have to be correlated across all the columns. Co-clustering is an unsupervised learning and using this method have benefit such as analysis will not rely on any particular labeling which can yield biased results therefore avoiding fine tuning the algorithm for a specific dataset, also co-clustering can help identify the attacking connections which goes unnoticed.

Data description : The data set which i am using was used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99.

I use a labeled portion of the dataset, consisting of 494020 connections; 97277 out of them are 'normal' and 396743 are 'attacks'. There are 22 different types of attacks in the labeled data set. For each connection there are 37 recorded measurements(features). Few features were such as: 'duration', 'protocol_type', 'service', 'flag', 'src_bytes', 'dst_bytes', 'land', 'wrong_fragment', 'urgent' and etc.

Evaluations: Implementing the co-clustering techniques and use them to classify each connection between "normal connection" or "bad connection". Will apply other techniques to prove usefulness of co-clustering and find if any other classification methods works better.

References:

Network Anomaly Detection using Co-clustering. Evangelos E. Papalexakis, Alex Beutel, Peter Steenkiste, ASONAM'12
MultiAspectForensics: Pattern Mining on Large-scale Heterogeneous Networks with Tensor Analysis. Koji Maruhashi, Fan Guo, Christos Faloutsos, ASONAM'11