Effectiveness of Graph-Based Clustering in Hadoop MapReduce for Network Intrusion Detection

Achuachua Tesoh-Snowsel. Computer Science Department. University of Maryland Baltimore County. Baltimore, Maryland 21250. atesoh1@umbc.edu

Dr. Enyue Lu. Department of Computer Science. Salisbury University. Salisbury, Maryland 21801. ealu@salisbury.edu

ABSTRACT

Network security is becoming increasingly important with the growth in computer usage. This growth has brought about the need for better network intrusion detection systems capable of differentiating between bad and normal connections. In order to process large volumes of network traffic data quickly and detect intrusions, parallel computation frameworks for Network Intrusion Detection Systems (NIDS) have been used recently. Graph clustering algorithms have been implemented in Hadoop MapReduce and have proven to be an effective method for anomaly detection. Previous work [1] indicated that a K-Nearest Neighbor Graph (KNNG) can be used to achieve 92% intrusion detection accuracy for network records using K value of 50. In this study, we investigated the accuracy of barycentric clustering algorithm based on K-Nearest Neighbor Graph (KNNG) using different K values and try to discover the optimum value of K for the best accuracy of anomaly detection for the KDD Cup 1999 Dataset.

INTRODUCTION

Distributed Graph-Based Clustering framework have been employed in the detection of network intrusion. The challenge remains to find efficient and reliable algorithms especially in detecting network intrusion in large volumes of network data. We examined the efficiency and effectiveness of different K-values for the KNNG which will help optimize the barycentric clustering algorithm in Hadoop MapReduce. Hadoop is open source and it is distributed under apache license. Hadoop is a framework of tools which support the running of application big data.

NETWORK DATA

We will be using the KDD-99 data set for our investigations. The KDD Cup dataset [2] simulates Local Area Network traffic [3] is a 4 gigabytes of compressed binary TCP dump data that is made up of about five million connection records were each of the connection is either labelled “normal” or “attack”.

The datasets contain a total of 24 subtypes in the training set and additional 14 types in the test data only. Each attack fall into four main groups which include:
•DOS: Denial-Of-Service, that involves an attacker flooding the targeted resource with a lot of request in an attempt to prevent some or all legitimate requests from being fulfilled;
•R2L: Remote-2-Local, attacker tries to gain access to a machine by guessing password;
•U2R: User-2-Root, attacker that has access to user’s machines tries to gain access to super user privileges;
•Probing: surveillance and other probing, when attacker tries to gain information about a host.

Our main goal is to identify anomalies using effective graph clustering algorithms on a real-world dataset such as the KDD Cup 1999 Dataset.

GRAPH CREATION

We will model network traffic by constructing a similarity graph in which vertices represent network records and the edges are formed as a result of the similarity between two vertices.

For our study, we use the radial basis function(RBF) to calculate the edge weight.

$$w_{i,j} = \frac{1}{\exp(\|\vec{v}_i - \vec{v}_j\|)}$$

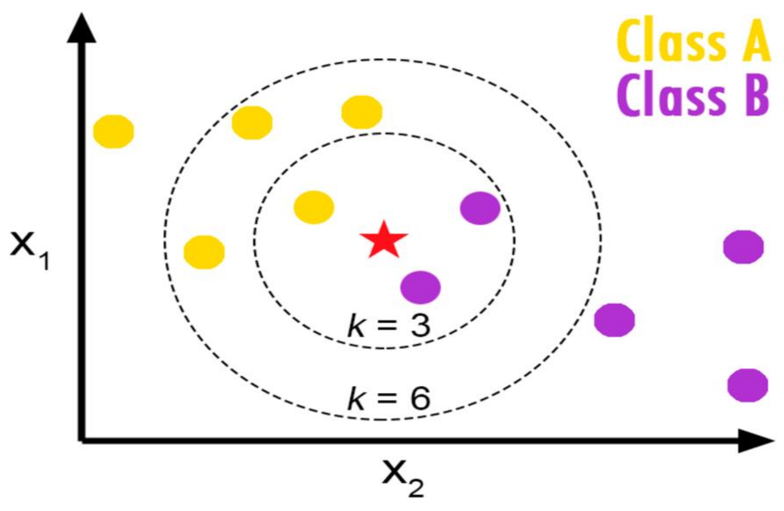
Where $\vec{v}_x$ is the data vector associated with the x^{th} network record represented as the x^{th} graph vertex. The idea for the RBF is that every point in the dataset $(x_n, y_n) \epsilon D$ will influence the value of the hypothesis at every point x:

$$Each\ (x_n, y_n) \in D\ influence\ h(x)$$

This influence is measure is inspired by the Euclidian distance between points.

CLUSTERING

Clustering involves the identification of similar groups in a data set. In order to group homogenous network records based on their similarity measure, we implemented the barycentric clustering algorithm [4]. In barycentric algorithm, each graph vertex is assign random positions which are iteratively adjusted based on their weights until vertex positions reach equilibrium. And once this happen, edges with longer than average length are deleted .



The above figure shows an example of two K-values 3 and 6 for KNNG.
Class A represents a cluster of yellow balls.
Class B represents a cluster of purple balls.
The smaller circle which encloses 1 yellow ball, 1 red star and 2 purple balls represents the red star and its 3 closest neighbors.
The larger circle which encloses 4 yellow balls, 2 purple balls is a representation of the red star and its 6 closest neighbors.

K Nearest Neighbor Graph (KNNG)

In order to improve the efficiency of the barycentric clustering algorithm, a KNN graph was used. In a k nearest neighbor graph, an edge connects two vertices if and only if the distance between x and y is among the k-th smallest distances from x to other objects from x [5]. K is some constant. One of the challenges of the KNN graph is choosing a good K-value. We tested different K values to find one which was most convenient for our data set and to produce the best results as possible.

DETECTION ACCURACY

After the clusters were formed, the accuracy of our clustering algorithm was calculated using the formula:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

- TP = True Positive, which represents the percentage of attacks records which were correctly identified as intrusion
- TN = True Negative, which is the percentage of normal records which were correctly identified as normal
- FP = False Positive, which is the percentage of normal records which were incorrectly identified as intrusions
- FN = False Negative, which represented the percentage of attack records which were incorrectly identified as normal records
- TPR = True Positive Ration, measures the proportion of attacks records which are correctly identified as the intrusion to all the attack records

CONCLUSION AND FUTURE WORK

From the simulation, we find that the accuracy increases with the increase of the K value but remains constant from 20 onwards for barycentric clustering algorithm. Future works will include the investigation of the accuracy of the k-truss clustering algorithms for different K values and running clustering algorithms on large real-world network data.

Reference

[1] C. McNeill, E. Lu, and M. Gobbert “Distributed Graph-Based Clustering for Network Intrusion Detection”, Extended Abstract, Companion of SC: IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SuperComputing), 2016
[2] KDD Cup 1999 Data, available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[3] S. D. Bay, D. F. Kibler, M. J. Pazzani and P. Smyth, “The uci kdd archive of large data sets for data mining research and experimentation,” 2000
[4] J.D. Cohen, “Barycentric Graph Clustering,” 2008.

ACKNOWLEDGEMENT: This work is funded by the National Science Foundation under the Research Experience for Undergraduates Program.