An Analysis of Datasets Using Big Data Techniques

Harry Rybacki

University of North Carolina at Greensboro
hrybacki@gmail.com

Abstract

This paper focuses on the analysis of two datasets: The NSL-KDD dataset and the UCI Forest Covertype dataset. After a brief overview of the statistics surrounding each dataset and an overview of relevant background information, big data and machine learning analysis techniques are used with Hadoop and Python to extract pertinent information. Specifically, the Stochastic Gradient Decent (SGD) algorithm is used in collusion with feature hashing to allow for similar scaling of both the size of the dataset and the computational complexity of the analysis.

I. Introduction

TBA

II. Background

Managing big data is difficult. As the number of features grows, the rate at which observations are stored, and the data varies more the storage and processing becomes increasingly more complex[2]. As a result, a combination of technologies designed to assist in the storage and processing as well as a careful choice of analysis algorithms is essential to getting any use from the data.

In his paper, Suthaharan recommends the use of Hadoop Distributed File Systems (HDFS) and Cloud Technologies to assist in the storing of data as well has the communication infrastructure of the analysis network.

As mentioned by Dalessandro, as datasets grow linearly the computational complexity of standard statistical analysis algorithms grows exponentially[1]. However, using the Stochastic Gradient Decent (SGD) algorithm can alleviate much of the computational complexity gains. Despite being less optimal on smaller datasets, SGD takes advantage of sparsity within datasets and searches for min/max of individual data points making it scale linearly with the dataset.

Furthermore, the use of feature hashing lessens the aforementioned problems surrounding high feature dimensionality. Although feature hashing degrades the quality of the data it allows for working with incredibly large, millions or billions, amounts of features[1] in a way that scales linearly.

III. Datasets

I. NSL-KDD Dataset

The NSL-KDD Dataset was constructed to resolve inherent problems in a similar dataset that was built in 1999[3]. The previous dataset was imbalanced and lead to bias toward more frequent attacks. This issue is however no longer present with the current dataset. Furthermore it is complete and there are no chunks of the dataset missing. We will have to assume that when this dataset was constructed the tools used to capture this information were accurate and the dataset was not tampered with in a way that would lead to inaccuracy.

With 125,973 observations, each of which has forty features, we can conclude that his is a high dimensional dataset. Furthermore, if an observer were to be collecting similar data in a live environment (an active network) the velocity of number of observations would increase at a non-trivial rate.

INSERT HISTOGRAM OF PROTOCOL TYPES

II. UCI Forest Covertype Dataset

The UCI Forest Covertype Dataset[4] was constructed for predictive modeling of forested lands the neighbor forested lands under the control of the original dataset owners. This dataset contains 581,012 observations each of which has fifty-four features. Relevant statistical information for each feature was easy to calculate. For example, feature 1, elevation, has a 581,012 observations, a mean of 2,959 meters, and a standard deviation of 280 meters.

INSERT HISTOGRAM OF ELEVATION

We can state that this dataset is not imbalanced as there are an equal number of , inaccurate, or incomplete.

IV. Computing Environment

TBA - Assignment 2

V. Machine Learning

TBA - Assignment 3

VI. Conclusion

TBA

VII. Acknowledgement

1. Dr. Shan Suthaharan, Associate Professor, University of North Carolina at Greensboro

References

[1] Dalessandro (2013). Bring The Noise: Embracing Randomness is the Key to Scaling Up Machine Learning Algorithms

[2] Suthaharan (2013). Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning

[3] ADD: NSL-KDD Dataset

[4] ADD: UCI Tree Cover Dataset