Precipitation Prediction on Cloud Distributed System

Group 7: David Hoxie (Daniels), Mashiur Chowdhury, Guimu Guo

July 2, 2018

1 Backgound

Machine Learning has been applied in many areas. However, most researchers run Machine Learning algorithm on the single machine which takes a lot of time. Even though the cloud computing is very mature, there is rare researcher put their Machine Learning study on the cloud. Today, Deep Learning achieve a breakthrough and speed up the process of modeling. But there is not sufficient comparison between deep learning and traditional Machine Learning Algorithm on the distributed system.

2 Problem Introduction

We intend to show that the application of Machine Learning deployed on a cloud based architecture can offer the benefits of decreased computational times, increased usability, and the reduction of high economical expenditures that are associated with the processing and storage of big-data.

Meteorological systems are difficult to analytically solve do to their non linearity, many systems tend to be chaotic in nature. This difficulty of finding analytic solutions presents an opportunity of applying machine learning to be able to describe the nature of the system in question.

We intend to demonstrate that Machine Learning can be used to accurately predict the amount of local precipitation, given temperature, Relative Humidity, and atmospheric pressure, as parameters.

The data for the project comes from Kaggle. `https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region/home`. The data set covers hourly weather data obtained from 122 various weather stations located in the southeast region (Brazil). This region includes the states of Rio de Janeiro, So Paulo, Minas Gerais e Espirito Santo.

3 System Architecture

The intent of this project is to build distributed system on cloud servers and run Machine Learning experiments on the cloud to forecast the amount of local Precipitation, given .

First, some raw servers from AWS (Amazon Web Services) are needed to build the distributed system. AWS is a subsidiary of Amazon.com that provides on-demand cloud computing platforms to individuals. To manage the storage of distributed servers, Hadoop needs to be installed beyond the operating system. It's collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. To run the Machine Learning Algorithm on Hadoop, Yarn and Spark are necessary to make the communication easily. Then utilizing the API of Spark, the Machine Learning Algorithms can be implemented.

4 Costs

The cost of deployment will very for the project depending on the number of computational cores that will be required for a given wall time of the computation. A rough early assuming a computation time of 48hrs estimate using amazon AWS EC2 using 4 of the 52.2xlarge image with 8 cores and 32GiB of memory would be $71.27 for the the system to be computed. Sotrage of the data on amazons S3 service would be less than $1.00 however the cost of data transfers need to be investigated further.

The scaling of the cost per user should not change the over all cost. The prediction of the data provided by the ML Algorithm would only need to be computed once. However if the platform is extended such that users can access the forecasts, then the data transfer of the forecast data needs to be accounted for.

5 Competition