CSC 357 - Final Project Proposal

Dr. Mendes

Brandon Liang & Binwei Xu

Data Analytics Using Unsupervised Machine Learning: Spark vs MPI

Motivation

Current trends of technology led by Silicon Valley clearly includes big data analytics in all industries. Ever since we entered the information age, the explosion of data generated every second and accumulated ever since has driven machine learning and big data analysis to a new level of importance. As we continue to learn more about parallelism in database algorithm, it has come to us that distributed computing is capable of solving one of the biggest challenges in data analytics, the cost of inter-machine communication and computation time. Therefore, inspired by Reyes-Ortiz, Oneto and Anguita's paper over big data analytics, we propose a practical research to compare the performances of using Unsupervised Clustering (K-Nearest Neighbor algorithm) with a selected dataset on Spark vs MPI. We will then discuss the differences between the performances and reasons behind.

Data and Approach

In the project, we are going to use data generated from NBA play-by-play data for the past 10 seasons. NBA publishes detailed play-by-play data for each game in the regular seasons over the past decade on its official website. From those data, we will generate features based on team-specific metrics for each team in each season to produce my final dataset. There are 30 teams each season, so there will be 300 team-season pairs (we are assuming the same team in different seasons is independent from each other). The following table illustrates the potential features for a data point:

Season-Team	Percentage of Wins	Percentage of Times the Team Cuts Down the Deficit When the Team is within a margin of 10 on the Next Score Change	.....

1516-Warriors	0.89	.....	.....
.....	.....	.....	.....

The complete list of features will be determined and the data processing will be done efficiently. As you can see, the columns starting from the second column essentially make up the features for each team-season pair. The goal of unsupervised clustering is to identify similarities among the 300 team-season pairs; we expect the result to indicate which cluster/group of teams tend perform similarly as other teams of the same cluster/group as opposed to the rest. 300 data points may not be reflect a big number, but the more dimensions the features have, the more expensive the computation cost is throughout the algorithm. Moreover, unsupervised clustering requires the computation of distances between every pair of the entire data points. In this case, the dataset will not be an easy task to process without parallelism.

To implement parallelism and improve computation time, we plan to utilize Google Cloud Platform service to create virtual machine clusters, where we will run the algorithm with our dataset. We will explore the MPI to examine multi-machine infrastructures, as well as Spark with in-memory computing (Tachyon). Since MPI is high-performance oriented and has much thinner layers of implementation, we expect MPI to outperform Spark in terms of performance time.