Machine Learning Predictions for estimation of Job wait time on HPC (High Performance Computing) System

Deenadayalan Dasarathan
Master of Science in Data Science, Khoury College of Computer Sciences,
Northeastern University, Boston, Massachusetts
MLPredictionsOnJobWaitTime_GitHubRepo

Abstract. The HPC cluster typically hosts a variety of compute nodes and GPUs which are consumed by the users for their resource intensive tasks. This inevitably creates resource contention. Hence, most contemporary HPC systems use a scheduler system to provide fair use of the shared resources. On Discovery, a high-performance computing resource for the Northeastern University research computing, The SLURM is an open-source resource management and job scheduling system manages the HPC cluster resources. When a job is submitted to the cluster, the SLURM schedules the job in the queue and allocates the requested resources for the prescribed time window. Often the submitted jobs undergo some wait time in the queue to get the requested resource (no of cpus, no of nodes, requested memory, gpu node, etc.). The queue wait time is defined as the number of seconds between the time a job gets submitted to the resource manager (SLURM) and the time that the job starts executing. In many cases, there is a trade-off here, in the sense that the more resources are requested, the faster the computing job will be finished; however, it will take more time for the resource manager to collect the required resources. There can be many situations in which the knowledge of the queue wait time is desirable. The users, for example, are interested in the queue wait time because it counts towards the total time between submission of a job and retrieving its results (time to completion). To most users, the time spent waiting in the queue is no different from the time spent waiting until their simulation has finished executing. Therefore, it is of interest to the user to have an estimate of how long their job will spend queuing before submitting it. Having this information, they may tweak the number of resources they request to get the results sooner. They may also consider submitting to another, less crowded, system. It can also be beneficial in urgent computing. The proper selection of resources and assignment of tasks to them is particularly important for better productivity of HPC. On the other hand, the estimates of queue wait time benefits the HPC administrators as the SLURM scheduler can be modified to fit the specific workloads running on the cluster. In situations where many users are requesting the same resources, SLURM parameters can be tweaked to accommodate the job and thus reduce wait times.

Currently, it is possible to obtain an estimated queue wait time from the job scheduler but only after the job has been submitted. Additionally, the job scheduler works as a black box without showing a relationship between queue times and requested resources.

In this study, we propose a data-driven approach for predicting job wait time on HPC systems prior to submitting the job. Here, “data-driven” means that our approach actively observes, analyzes, and logs jobs collected on Discovery cluster. Supervised machine learning algorithms are applied to predict the queue wait time for the job which is dependent on several features like requested resource, resource availability, number of active jobs, number of pending jobs at that time.

Proposed Design: The machine learning model will have the following two regression algorithms,

- RandomForestRegressor
- XGBRegressor

RandomForestRegrssor is one of the best ensemble methods as it adds randomness (random subset of sample with replacement) to the model, while growing trees. And the aggregation on individual decision trees reduces the variance of forest, thus prevents overfitting of training data.

XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that enables estimation in a fast and accurate way.

The cross_val_score will be calculated for each estimator with different hyper parameters and the best model will be used to predict the queue wait time.

Data Set: The data set is extracted from Discovery HPC cluster SLURM Database using the sacct command. The extracted data set is then stored as a pickle file.

In the data set, we analyze 6 million+ records (submitted jobs/samples) and each record is associated with 87 features generated by 1300+ Discovery cluster users to build a machine learning model that establishes a relationship between queue time and requested resources and predicts the queue time based on those resources.

The data preprocessing steps will include,

1. Handling unknown values in the data set.
2. Converting the timestamp data (string) into pandas datatime type.
3. Dropping samples with missing values.
4. Convert categorical date into numerical data using One-hot encoding.

Libraries and Tools: The project will utilize the following tools and libraries.

- Jupyter Notebook
- Python 3.7
- Numpy
- Pandas
- Seaborn
- Sklearn
- Pickle
- Matplotlib
- Datetime
- Scipy

Outcome of this Project: The trained model will predict queue wait time of the job before submitting it to the scheduler, as well as provide comprehensive understanding of current workloads on the cluster. This model also yields insights into how to optimize resources requested for a job that users can reduce wait times.