HPC Job Outcome Prediction: System Log Feature Extraction and Importance

Alexandra DeLucia
Ultrascale Systems Research Center
Los Alamos National Laboratory

Elisabeth Moore
Ultrascale Systems Research Center
Los Alamos National Laboratory

Introduction

• Semi-supervised application of machine learning to the monitoring of high performance computing jobs
• Predicting the outcome of jobs using features from system log produced from the job

Research Questions
1. How accurately can syslogs predict job outcome?
2. Which features from syslogs are most informative?

Background

Job Log
• Jobs are recorded by the job scheduler (e.g. Moab, Slurm) in a job log file

Job Log Entry Format:

<table>
<thead>
<tr>
<th>Job State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANCELLED*</td>
<td>Jobs that are not used in this experiment.</td>
</tr>
<tr>
<td>COMPLETED</td>
<td>Job completed successfully</td>
</tr>
<tr>
<td>FAILED</td>
<td>Job did not complete for some reason (e.g. program bug)</td>
</tr>
<tr>
<td>NODE_FAIL</td>
<td>Job did not finish in the allocated time limit</td>
</tr>
</tbody>
</table>

System Log (Syslog)
• Syslogs give insight to process activities and are crucial for failure analysis

Approach

Data Set
A sample of 5,000 jobs from Los Alamos National Laboratory cluster Wolf over the month of June 2018

Feature Engineering and Extraction

Overview

Text Content Analysis
• Analyzed the text content using different methods from the fields of systems, natural language processing, and graph analysis

Feature Importance
• For each job we analyze the group of syslogs associated with the job
• Syslogs contain an inhomogeneous mixture of numerical, temporal, and text content
• Separated the numerical, temporal, and text content for individual analysis

Selected Text Clusters

Infolmop Clusters

Term Frequency-Inverse Document Frequency (TFIDF)
• Distribution of terms across the syslog message for all jobs
• Identifies unique words by giving rare words more weight

Latent Dirichlet Allocation (LDA)
• Generative statistical model that finds latent "topics" across documents
• Distribution of topics across syslog group

Process Tag Distribution
• Distribution of process tags across each syslog group

Experiment and Results

Experimental Setup
• Trained and tested a Random Forest model on all feature sets to predict job outcome (state)
• The model was evaluated on three prediction tasks

Results Summary
• All feature sets performed best on the Okay vs. Problem task
• The combined feature sets performed better than alone
• Best performing feature sets across all tasks:
  - Infolmop and Temporal & Numerical
  - LDA and Temporal & Numerical
  - Tag and Temporal & Numerical

Feature Importance
• Temporal features were important across all feature sets
• Numerical features, specifically decimal features, were also important

Acknowledgements

We thank Sean Blanchard and Nathan DeBardeleben, both at the Ultrascale Systems Research Center, Los Alamos National Laboratory, for domain expert input to this work. We also thank Hugh Greenberg for providing easy log access.

References