UC BERKELEY IEOR CAPSTONE PROJECT: SLAC NATIONAL ACCELERATOR LABORATORY

Rob Holbrook, Michael Lee, Ramon Lim, Vikram Singh, Stephanie Zhu
Spring 2018

EXECUTIVE SUMMARY

BACKGROUND

SLAC National Accelerator Laboratory (SLAC) is a United States Department of Energy National Laboratory operated by Stanford University under the direction of the U.S. Department of Energy Office of Science. SLAC is a multi-facility laboratory supporting groundbreaking scientific experiments from researchers around the world. One of those facilities is the Linac Coherent Light Source (LINAC), the world's first X-ray Free Electron Laser driven by a 1 km electron linear accelerator. Fundamental to the physics of the LINAC is the acceleration of electron bunches to nearly the speed of light which emits intense X-rays for studying biological, chemical, and material science on molecular time and space scales. RF stations across the LINAC emit bursts of extremely high voltage necessary for these electrons to achieve near speed of light velocities.[1]

To ensure optimal operation of the accelerator, SLAC uses an open-source software environment called EPICs, Experimental Physics and Industrial Control System,[2] to monitor and operate all devices within the accelerators including RF stations. Real-time and historical operational data for RF stations, called Process Variables (PVs), are constantly logged and archived for each RF station. Each measurement type captured for the RF stations are called attributes. PVs are RF station-attribute specific and contain the following data: time of measurement, numerical measured value at that time for the specific attribute, alarm value, severity value.

Because RF stations involve multiple components requiring regular tuning and maintenance, automated detection, predicting maintenance needs or components failures would be of extreme value to scientific research. This capstone focuses on specific RF stations and their attribute data to develop machine learning models that will provide insight into improving RF station operations. Specifically, this capstone analyzes PV data for RF stations LI24-61, LI22-31, and LI28-31 and attributes, MOD_THY_RESV, MOD_THY_HTR, MKBVFTPJASIGMA (SIGMA), MOD_HV_V, SWRD_MOD, MOD_DQI_I, WNDW_I, and PJTN.

TECHNICAL ACCOMPLISHMENTS

- UNIX scripts to mine SLAC's EPICs system to retrieve PV data
- Conversion of EPICs data output to appropriately formatted csv files for processing
- Creating a cloud-based, SLAC-dedicated data warehouse for data querying and analytics
- Code to:
    - Convert time series data to data sets that can be modeled using supervised-learning classification approaches
    - Feature extraction, selection, ranking
    - Supervised-Learning machine learning models – Random Forest, Boosting, Logistic Regression
    - Unsupervised-learning models - K-Means Clustering, Hierarchical Clustering

SUMMARY OF INDIVIDUAL REPORTS

RAMON LIM

Applying Machine Learning to SLAC RF Station Operational Data (Process Variables) To Predict Hardware Issues

Description: The goal of this study was to understand the correlation that PV data has on predicting hardware issues tracked in the CATER system.

Summary: The project analyzed PV data for the RF stations and time frame in scope. CATER data was also analyzed for hardware (HW) problem jobs created for those specific stations. Three years of time-series PV data for each station were analyzed in 5 day "windows" with a window marked as a "positive observation" if a CATER HW problem job was created in that time frame, otherwise marked a "negative observation". Additionally, the average, standard deviation, range, skew, rate of change, and mean of the fast fourier transform(FFT) frequencies were calculated based on the process variables value data in each window to determine the features used to approach this study in a supervised-learning classification manner.

Logistic regression, random forest, and boosting algorithms were applied with general AUCs in the 0.7 to 0.9 range for better performing models. Key insights include: skew and FFT are the most relevant features, and MOD_THY_RESV and PJTN are the most relevant attributes in predicting CATER HW problems. The recommendation is to expand this study to include more data (stations, attributes, years), use feature extraction packages to select features without bias, expand the window of time observed before a CATER event, and investigate the use of survival analysis and deep learning/neural networks to strengthen predictions.

STEPHANIE ZHU

Analyzing the severity status for each attribute in a radio frequency (RF) station to predict hardware issues

Description: The goal of this subproject is to identify if changes in severity status can anticipate hardware issues tracked in the CATER system.

Summary: This is a time series classification project that views the severity status of an attribute in a RF station as the primary independent variable in predicting hardware issues. Three fiscal years' worth of data for selected attributes in RF stations was extracted and merged with instances of CATER events one week out. Additional features extracted from the severity status, such as rolling mean, rolling standard deviation, and presence of a severity alert in the past week were derived from the existing data. Due to an imbalanced dataset in which CATER events appeared fairly rarely given the breadth of time series data, the majority class was undersampled in order to match the number of the minority class and improve recall.

Random forest and logistic regression were applied with general AUCs in the 0.5 to 0.7 range. Key insights include: thyratron attributes (MOD_THY_RESV, MOD_THY_HTR) appear to be more powerful predictors than other attributes in predicting CATER events. The recommendation is to include additional attributes and RF stations not originally included in the primary dataset, generate additional features not already included in this report, and implement alternative time series models to better capture and anticipate future events.

ROB HOLBROOK

Using Unsupervised Learning to Identify Different Types of Jitter Failures

Description: The goal of this subproject is to identify different types of Jitter (PJTN) events within the targeted Klystrons.

Summary: The project analyzed 3 years of Klystron operational data for 3 separate Klystrons to segment out PJTN events. An event was began when the median PJTN value exceeded 0.2 for one minute and ended when the median PJTN value for a minute returned to below 0.2. These time windows were identified and data from additional attributes (e.g., RESV, SWRD) was added into the dataset for analysis.

Once the dataset was populated, several features were built based on the values for the identified attributes. Features were reduced due to data inadequacy in targeted time windows based on algorithm sensitivity to missing data. K-means, principal component analysis, hierarchical clustering, and gaussian mixture models were applied to identify potential clusters within the data. These naturally occurring clusters will be used to help identify different-in-kind PJTN events for both managing SLAC response to those events as well as improving the accuracy for future forecasting events.

MICHAEL LEE

Using Supervised Learning for Multiclass Prediction of Jitter Frequency

Description: The goal of this subproject is to analyze whether the frequency of Jitter (PJTN) events within an RF station can be predicted.

Summary: Attributes believed to be related to Jitter Events were developed into features. Jitter event data (developed in separate workstream), including the count and severity of events, was processed into multiclass response data. Tree-based models were trained for Jitter Event frequency prediction.

VIKRAM SINGH

Using Supervised Learning for Binary Prediction of Jitter Events

Description: The goal of this subproject is to predict whether a Jitter (PJTN) event will occur within any given calendar day the frequency of Jitter (PJTN) events within an RF station.

Summary: PJTN data was explored with the goal of appropriately defining a 'Jitter event'. Jitter events were processed into binary response data for each calendar day. Feature data (developed in separate workstream) was merged with response data. Linear Regression, Random Forest, and CART models were trained for prediction.

CONTEXT SUMMARY

Over the past several years, accelerator facilities are shutting down before new ones are opening, and thus restricting the opportunity for experiments and creating great uncertainty about future funding.[3] SLAC's linear accelerator has been, and is, integral to thousands of ground breaking, innovative experiments annually. With the goal of improving the LINAC's operational efficiency, our capstone project is aimed at predicting events and features that could potentially lead to accelerator failure and idleness. Our overarching project results are aimed at improving apparatus uptime to increase the number of research experiments and reduce operational expense. In later sections, we will discuss issues concerning our team's technology marketing strategy, the particle accelerator industry, and the LINAC's societal impact.

REFERENCES

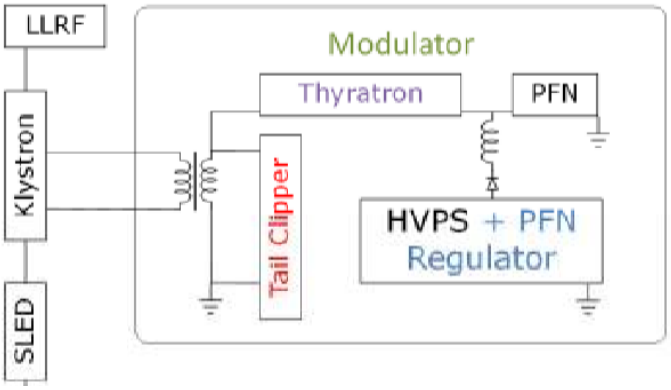
1. Vretenar, Maurizio - CERN AB/RF (2008). Introduction to RF Linear Accelerators. Retrieved from https://cas.web.cern.ch/sites/cas.web.cern.ch/files/lectures/frascati-2008/vretenar.pdf

2. Experimental Physics and Industrial Control System (2011). About EPICS. Retrieved from https://epics.anl.gov/about.php

3. Richter, Burton (2007). Charting the Course for Elementary Particle Physics, Lecture Given in San Francisco, California. Retrieved from https://arxiv.org/pdf/hep-ex/0702026.pdf

APPENDIX

RF Station and Components



Attributes and Descriptions

Attribute	Description
MOD_THY_RESV	Thyratron reservoir anode voltage for triggering modulator pulse.
MOD_THY_HTR	A readback value of the thyratron heater current.
MKBVFTPJASIGMA (SIGMA)	Klystron beam voltage amplitude jitter from the modulator. As with PJTN, unstable performance is both a problem of its own and and indicator of needed maintenance.
MOD_HV_V	Measured readback of the modulator high voltage output to the klystron. Should remain stable and preferably kept high for maximum beam acceleration. May be reduced in the case of hardware failure, or manually if needed to ameliorate problems seen when

	driving the thyratron or klystron too hard for stable performance.
SWRD_MOD	(0 or 1) A single logical bit of the station's status WoRD . If true, it indicates a Modulator Fault of some kind has occurred causing the station to be deactivated.
MOD_DQI_I	Modulator de-Qing current. Value used to adjust the regulation of charge stored in the modulator PFN (see "Regulation Circuits). Largely "set and forget," however improper de-Qing can lead to unstable modulator output.
WNDW_I	Temperature sensor mounted on the RF waveguide windows (part of the plumbing used to transmit the RF power down into the tunnel from upstairs). These get hot but heating up too much is bad. In some cases may also indicate increasing reflected power from downstairs or other faults. Used for monitoring and interlocks.
PJTN	High-power RF RMS phase jitter. As an RF station performance degrades, a first indicator is poor RF phase stability (larger PJTN).