LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-754521

Dynamic Networks Experiment FY18 - Quality Control Focus Area Efforts

S. A. Magana-Zook

July 10, 2018

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Dynamic Networks Experiment (2018) - Quality Control Focus Area Efforts at LLNL

Motivation

The availability of data whose quality is fittingly labeled is of tantamount importance to signal detection and all other downstream signal processing operations. Without identification, the goals of low yield nuclear monitoring cannot reach their full potential. A fundamental question that remains is the knowledge of the capability gaps that prevent us from using open network data. With these new channels, we will need better understanding of the frequency, type, severity, and impact of artifacts when compared to conventional networks. Moreover, beyond a simple quality label, thresholds for acceptable use of the data will need to be established, likely on a use case by use case basis. Finally, orders of magnitude more data with an at-best static analyst pool size means that quality artifacts must be identified with as little human interaction as possible.

Automated methodologies must be developed to account for the scale of the problem when ad hoc networks are candidates for processing. At the same time that "data wealth" provides more opportunity for signal exploitation, the introduction of ad hoc networks introduces an increased risk quality issues in both sensor value and metadata issues. The artifacts shown below are examples of some sensor-based quality issues that impair downstream processing. The machine learning model used in this project was able to identify the artifacts that produce these spurious correlations, significantly improving analyst productivity.

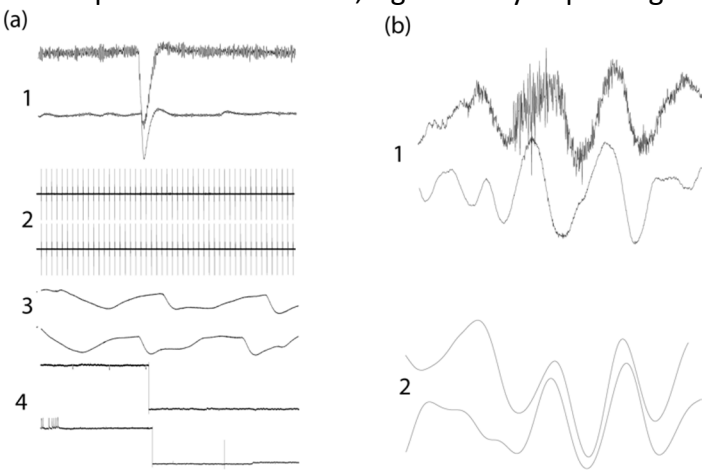


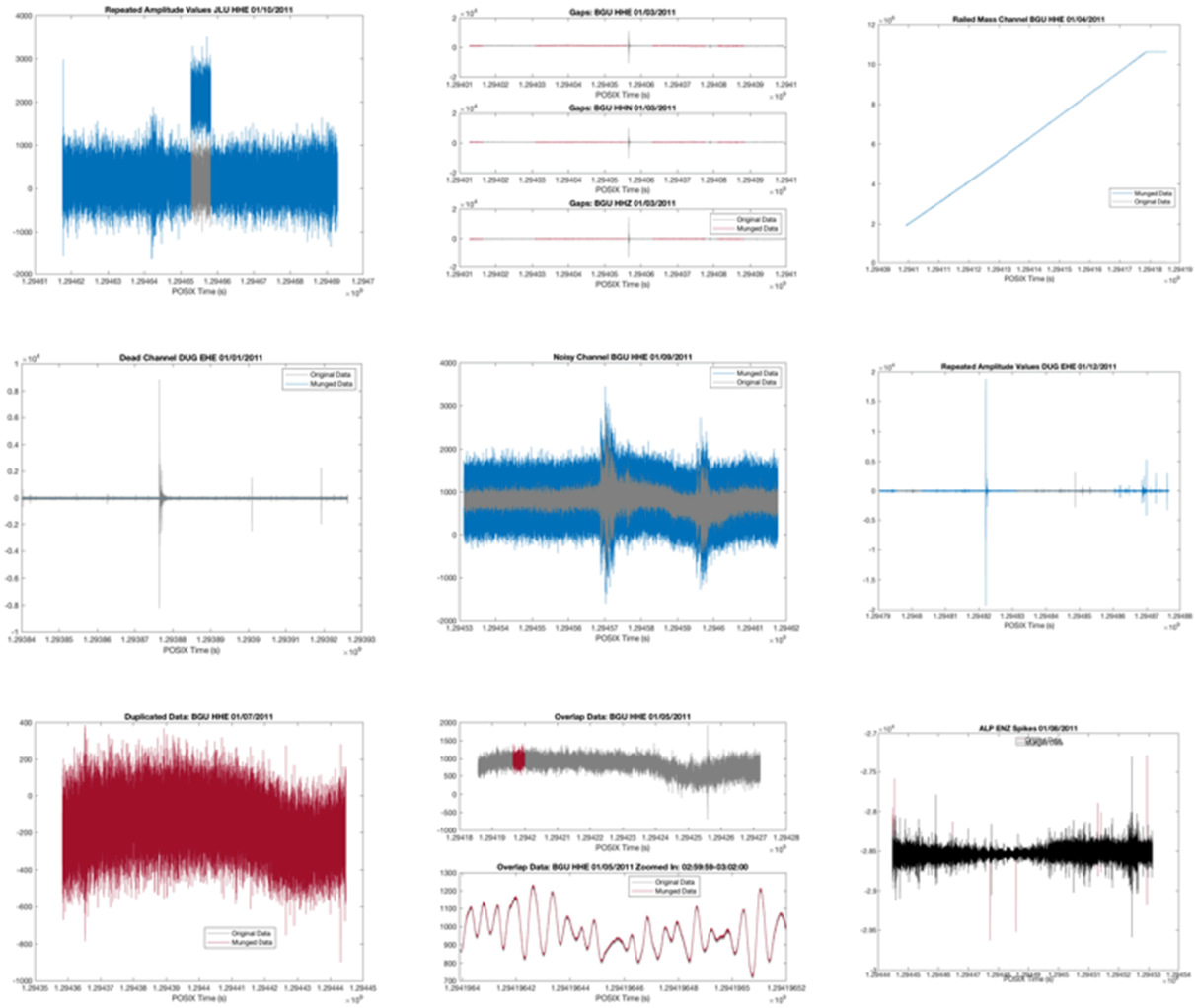
Figure 1 Sensor-based quality issue examples (Dodge and Walter)

Data

Nine types of synthetic sensor-based artifacts (examples shown in the figure below) were introduced to real data from the University of Utah network. These nine do not account for the full list of quality issues to be addressed in future experiments. Presently, the goal of this experiment was to establish the baseline capability for our tool(s) to be able to identify these

Prepared by LLNL under Contract DE-AC52-07NA27344.

synthetics. Beyond those artifacts displayed below, sensor response metadata was also altered in some cases to simulate impaired processing.



Related work at LLNL

The approach taken at LLNL to solve the problem of automated quality issue detection at scale has been to use machine learning. Machine learning offers promise in that, given the right set of training features, models can find relationships between features previously unknown to the analyst. Trained models also have a constant size in memory, making them the perfect candidates for scalability. For robustness and explainability we chose the Random Forest model.

Random Forests belong to the class of supervised learning machine learning models. In supervised learning, an analyst assigns a class label to each seismogram that will be used to train and test the model. This is a time consuming up-front cost that is of critical importance to the performance of the resultant model.

Prepared by LLNL under Contract DE-AC52-07NA27344.

Feature Importance

Machine learning models are not magic, they will only work well if they are given features with predictive power behind them. The Random Forest we trained was given 16 statistical features spread across 18,000 instances over which to learn the difference between a signal and an artifact. For each decision tree that made up the random forest, a random subset of these features and instances would be selected to be trained with.

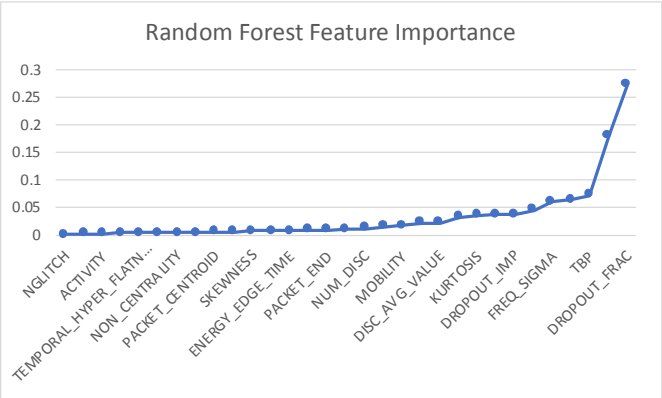


Figure 2 plot of each feature's importance to predicting the label of the label of the segment.

In the figure above, and the table below, no one feature provided the overwhelming predictive power. Rather, six features provided roughly 70% of the entropy, with the remainder spread among the other ten. This result suggests that feature reduction methodologies would likely add value to this pipeline.

Feature	Importance
Number of glitches	0.00030172
TEMPORAL_KURTOSIS	0.00223085
ACTIVITY	0.00252851
TEMPORAL_HYPER_KURTOSIS	0.00299533
TEMPORAL_HYPER_FLATNESS	0.00300932
TEMPORAL_SKEWNESS	0.00321737
NON_CENTRALITY	0.00348043
Extreme statistic	0.00434551
PACKET_CENTROID	0.00441469
TIME_CENTROID	0.00491872
SKEWNESS	0.00661037
Signal-to-noise ratio	0.00694195
ENERGY_EDGE_TIME	0.0072597
TIME_SIGMA	0.0084726
PACKET_END	0.00900621
PACKET_SIGMA	0.01086358
NUM_DISCONTINUTIES	0.01164575
AMPLITUDE	0.01530545
MOBILITY	0.01741838

Prepared by LLNL under Contract DE-AC52-07NA27344.

NUMBER_OF_MODES	0.02110537
DISC_AVG_VALUE	0.02245336
SPECTRAL_EDGE_FREQ	0.03198181
KURTOSIS	0.03505331
COMPLEXITY	0.03619735
DROPOUT_IMP	0.03629349
DISC_MAX_VALUE	**0.04433493**
FREQ_SIGMA	**0.06057438**
PACKET_TBP	**0.0633975**
TIME-BANDWIDTH-PRODUCT	**0.07181007**
DISTINCT_VAL_RATIO	**0.17955247**
DROPOUT_FRAC	**0.27227953**

Model Results

The random forest model produced an accuracy of 99.09%, a precision of 99.52%, and a recall of 99.15% as shown in the confusion matrix below. While the accuracy is undoubtable high, in the context of more than a billion waveforms, every percent counts. 0.5% of the examples were predicted to be signals when an analyst labeled them artifacts, an annoyance that interferes with processing of correlation and other downstream operations. The 21 signals that were labeled as artifacts are potentially more serious as this could result in data loss from a filtering operation to remove artifact segments.

N= 6392	**Predicted = signal**	**Predicted = artifact**
Actual = signal	TN = 4339	FP = 21
Actual = artifact	FN = 37	TP = 1995

Table 1 Random forest confusion matrix

Challenges and Future Work

While this approach shows promise, work remains in being able to detect even the subset of QC issues identified with DNE-18. For example, the 16 features trained on are more successful at catching the onset of DC-offset change than they are if the segment is in the middle of an existing offset. Additionally, these timeseries artifact features do not provide any visibility into metadata problems. Taken together, the thrust of future work lies in feature engineering.

Prepared by LLNL under Contract DE-AC52-07NA27344.