Real-Time Machine Learning SVM Classification of Microearthquakes

David Crowe[1,2], Martin Schoenball[1], Zhao Hao[1]

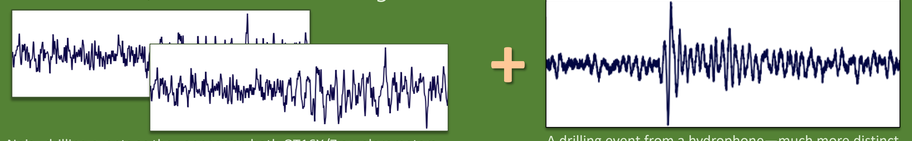
[1] Lawrence Berkeley National Laboratory, [2] Texas State University

Abstract

4,850 feet deep at the Sanford Underground Research Facility (SURF) in South Dakota, a dense 3D sensor array consisting of 12 hydrophones and 18 accelerometers deployed in six boreholes produce around 2 TB of seismic data a day accumulated from several sources. **Existing trigger-detection algorithms can detect seismic events, however an expert must still parse through these signals to search for happenings of interest like microearthquakes.** To support in this endeavor, a machine learning classifier was trained on thousands of labeled seismic samples from one microearthquake (MEQ) and three "noise" categories. Using 3 out of 60 channels from each Stream object, **the model correctly classified 94% of 1,127 microearthquakes** and a collective 90% from all four classes.

Data Optimization & Channel Selection

- Each Stream object for an event contains waveforms from 60 channels (1-6Mb ea.)
- Three channels: PDB03 (hydrophone) and OT16X & OT16Z (accelerometers) were used to **reduce the size of the training data by a factor of 20** in many cases (1-3kb) by exploiting the unique responses from each type of sensor.
- By using multiple channels, events which are not distinct in one channel may be revealed in the other, as in the case of drilling:



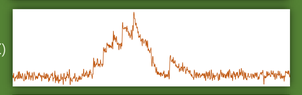
Noisy drilling events as they appear on both OT16X/Z accelerometers.

A drilling event from a hydrophone—much more distinct.

Examples of Labeled Waveforms

MEQ (OT16Z)
Micro-Earthquake seismic signature

ERT (OT16Z)
Electric Resistivity Tomography
Tool for understanding soil and rock composition underground

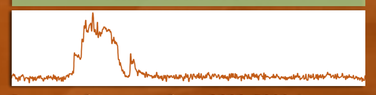


Drilling (PDB03)
From nearby borehole drilling

CASSM (OT16Z)
Continuous Active Seismic Source Monitoring
Used to infer dynamic properties of sub-surface environments

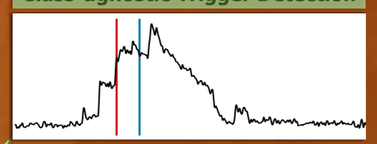
Seismic Signal to SVM Feature Space

Raw Waveform



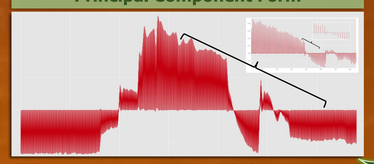
Raw waveforms vary from 3000-5000 data points, but SVMs work using samples of the same size. . .

Class-agnostic Trigger Detection



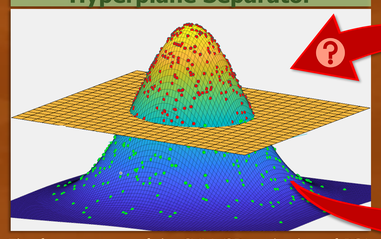
A recursive STA/LTA algorithm finds trigger or start location of an event to determine suitable window.

Principal Component Form



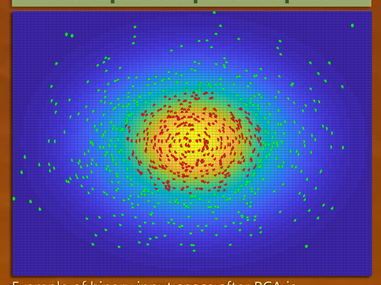
ERT waveform represented by 849 principal components.

Hyperplane Separator



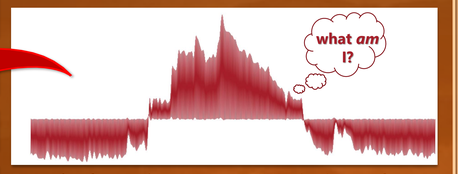
The feature space of the SVM (2D projection of PCA space) with the hyperplane separator defining the clusters of classes.

Principal Component Space



Example of binary input space after PCA is performed on waveform which shows how the data is not naturally linearly separable.

New Data Classification



Example of an "unknown" post-PCA seismic waveform ready for classification.

Confidence probability of a (correct) seismic signal classification.

ERT	99.2%
MEQ	0.41%
CASSM	0.38%
Drilling	0.01%

```
# Trains SVM classifier on channel data from all available streams
def create_classifier(do_grid_search):
      channel_data = []

      # extracts relevant channel PCA wiggles for training
      for t, trace in enumerate(total_data):
            channel_data.extend(trace)

      # 70% Training / 30% internal testing
      X_train, X_test, y_train, y_test = train_test_split(total_data,
                                                          class_labels,
                                                          test_size=0.3)

      X_train_2D = transform3Dto2D(np.array(X_train))
      X_test_2D = transform3Dto2D(np.array(X_test))

      # Using Scikit-learn's stellar SVM library
      classifier = svm.SVC(C=20, _size=200, class_weight=None, coef0=0.0,
      decision_function_shape='ovr', cachedegree=3, gamma=1, kernel='rbf',
      max_iter=-1, probability=True, random_state=None, shrinking=True,
      tol=0.001, verbose=False)

      # Find separator configuration
      classifier.fit(X_train_2D, y_train)

return X_train_2D, X_test_2D, y_train, y_test, classifier
```

Code responsible for crafting the support vector machine.

Classifier Construction & Data Parameterization

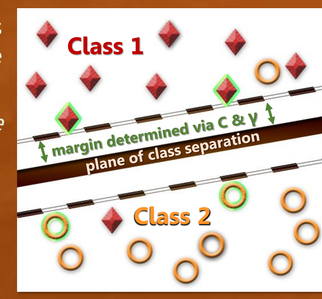
Building the Support Vector Machine

To perform classification, a support vector machine (SVM) classifier was fed thousands of principal components from waveforms of each of the four classes.

The classifier works by finding a separator that maximizes the distance between the points closest to the hyperplane and is tuned via:

C (penalty): determines how heavily samples on the wrong side of the separator affects the model's total error

Gamma (γ): distance of influence a single sample has on separator (regulates model overfitting)



An example hyperplane separator with support vectors highlighted.

Waveform Data Preparation

Because all samples must have the same size in training & classification stages, the following parameters were used to standardize the seismic signals:

- block size: number of "chunks" to split waveform data for analysis (always set to 2 in our case)
- number of components: resulting size of waveforms in data points
- trace window size: truncated portion of data to greatly expedite training & classification
  - This helps reduce the amount of non-characteristic noise present in all seismic signals
- trigger offset: allows for imperfect trigger detection by specifying buffer region to consider

Multi-channel Model Performance

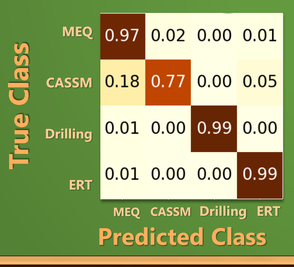
Results

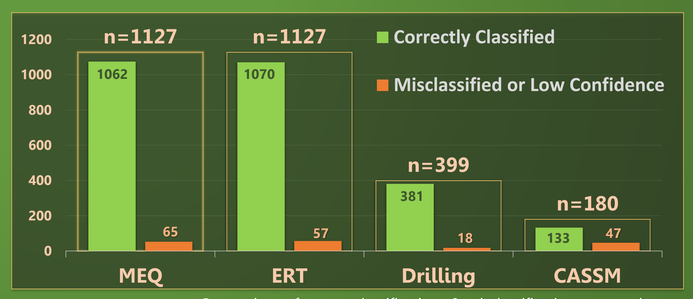
- Out of 1127 samples reserved for validation, 94% of MEQs were correctly classified with a certain degree of confidence.
- The confidence thresholds for "confident" and "semi-confident" were 70% & 50.1% respectively.
- The model was shown to suitably classify waveforms from all classes in near real-time **(~0.5s per sample)**
- Using data reserved for training, the SVM classifier produced a confusion matrix which reveals internal performance.

	Confident Predictions	Semi-Confident Predictions	Misclassified Seismic Signals	% Correctly Classified
MEQ	1003	59	65	94%
CASSM	119	14	47	74%
ERT	1054	16	57	96%
Drilling	376	5	18	95%

Breakdown of classification performed on batches of samples isolated from training environment.

Confusion matrix highlighting the model's internal incongruities.





Proportions of correct classifications & misclassifications across classes.

Discussion & Future Direction

Conclusions

- Ultimately, the model is effective with over **90% classification accuracy of 2833 waveforms across all four classes**. The classifier likely suffers from cross-class contamination which introduces a source of discretionary ambiguity, however carefully chosen SVM and PCA parameters combat this.
- Though intended for microseismic filtering, this model could certainly be applied to other similar waveform classification tasks.

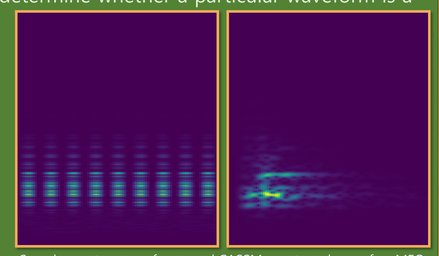
Implementation

- Once the SVM model (~25 MB) is loaded into memory, it can classify any given Obspy stream object in about 0.1-0.5 seconds.
- Though no events will be discarded, this classifier introduces a mechanism to reliably determine whether a particular waveform is a microearthquake or from another class.

Future Direction

Though the model performed suitably, ways to improve the overall performance include:

i. Increasing the total size of labeled data for model training (especially CASSM).
ii. Identifying additional channels to better represent microseismic events.
iii. Run more comprehensive grid searches to identify a more robust set of SVM parameters across different combinations of PCA representations.
iv. Alternatively, a CNN deep neural network model may prove fruitful by transforming the problem into an image classification task using waveform spectrograms.



Sample spectrograms for several CASSM events and one of an MEQ. This suggests great potential for use in a deep neural network.

Acknowledgements

This work was supported, in part, by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. Furthermore, this material was based upon work supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), Office of Technology Development, Geothermal Technologies Office, under Award Number DE-AC02-05CH11231 with LBNL. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The research supporting this work took place in whole or in part at the Sanford Underground Research Facility in Lead, South Dakota. The assistance of the Sanford Underground Research Facility and its personnel in providing physical access and general logistical and technical support is acknowledged. Additionally, background texture images of Liesegang banding are courtesy of Christopher David Benda (illinoisbotanizer.blogspot.com/2019/01/liesegang-banding-in-southern-illinois.html).